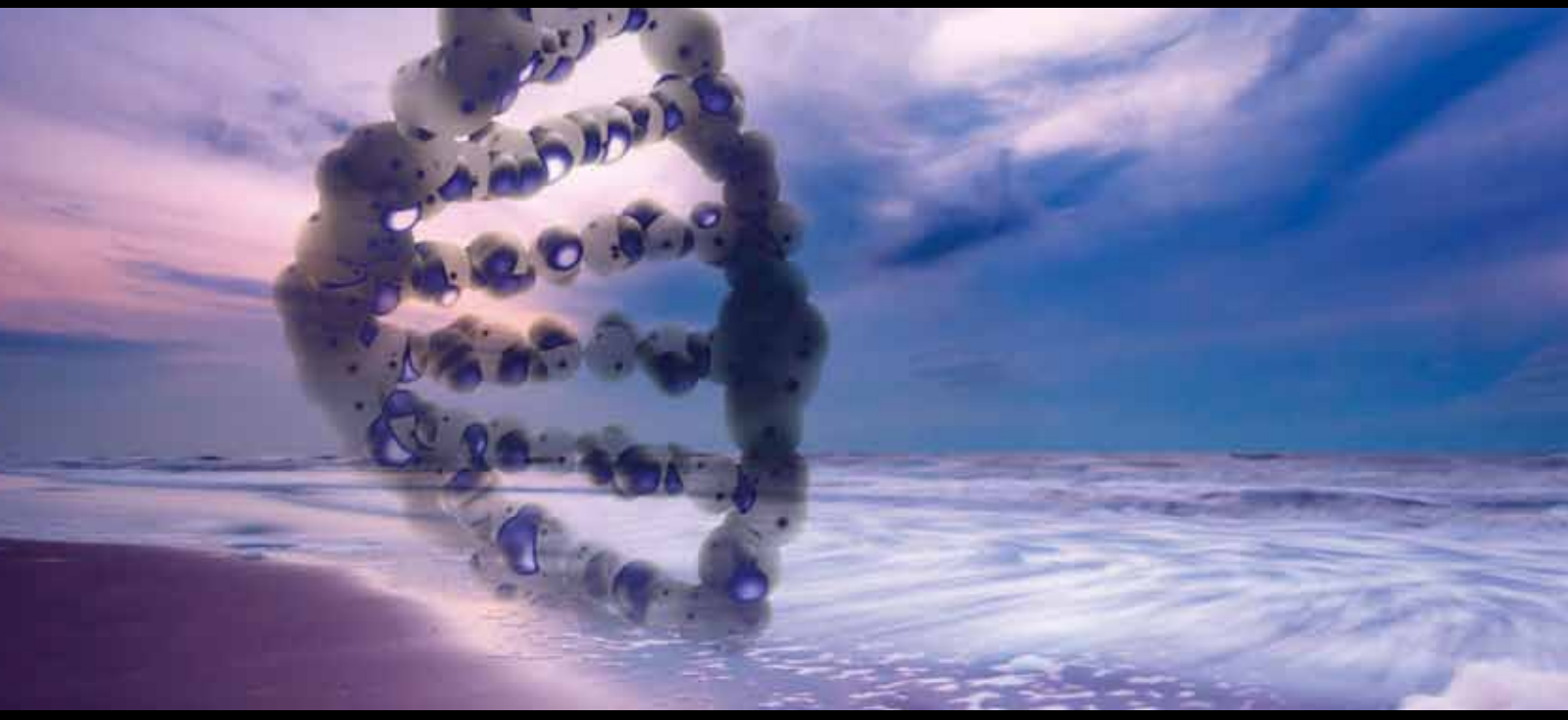
International Journal of Evolutionary Biology

Molecular Evolutionary Routes That Lead to Innovations

Guest Editors: Frédéric Brunet, Hideki Innan, Ben-Yang Liao, and Wen Wang



Molecular Evolutionary Routes That Lead to Innovations

International Journal of Evolutionary Biology

Molecular Evolutionary Routes That Lead to Innovations

Guest Editors: Frédéric Brunet, Hideki Innan, Ben-Yang Liao, and Wen Wang

Editorial Board

Contents

Hindawi Publishing Corporation
International Journal of Evolutionary Biology
Volume 2012, Article ID 483176, 2 pages
doi:10.1155/2012/483176

Editorial

Molecular Evolutionary Routes That Lead to Innovations

Frédéric Brunet,[1] Hideki Innan,[2] Ben-Yang Liao,[3] and Wen Wang[4]

[1] *Institut de Génomique Fonctionnelle de Lyon, Ecole Normale Supérieure de Lyon, 32-34 Avenue Tony Garnier, 69007 Lyon, France*
[2] *Hayama Center for Advanced Studies, The Graduate University for Advanced Studies, Kanagawa 240-0193, Japan*
[3] *Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, 35, Keyan Road, Zhunan, Miaoli County 350, Taiwan*
[4] *Max-Planck Junior Scientist Group on Evolutionary Genomics, Kunming Institute of Zoology, Chinese Academy of Sciences, Yunnan, Kunming 650223, China*

Correspondence should be addressed to Frédéric Brunet, frederic.brunet@ens-lyon.fr

Received 9 September 2012; Accepted 9 September 2012

In biology, it is always fascinating to observe the continuous changes occurring at different levels in living organisms. In demonstration of this interest, one question we regularly ask is the following:

> "*What are the events that contribute to the evolution of new functions and adaptive evolutionary innovations?*"

Deep down at the molecular level, we are referring to the various genetic and molecular mechanisms that occasionally either duplicate preexisting genes, or lead to the origination of new sequences. Altogether, these new and duplicated genes represent a substantial fraction of every genome sequenced. They are of multiple origins ranging from whole-genome duplications, which are documented in many eukaryotes, to various other modes of duplication—mostly single full or partial gene duplications—by DNA-based or retroposition events.

Each month, new and exciting articles dealing with these topics are published in different journals, and it has now been several years since a special issue collecting this research into a single volume has been published. Our belief is that it is now the time for the publication of a new special issue that would bring together researchers working in this field, and provide up-to-date research on the subject. For this special issue, we allowed these authors the liberty to offer either review papers or pure research articles. Others were also encouraged to use a rather unconventional format, bringing together a review of their published work, but updated with new insights into their latest discoveries. Experts in the field performed in-depth reviews of these submissions to ensure the articles would be of the highest quality. Among the 18 articles that went through this reviewing process, 17 were accepted. We feel that this represents the high quality of the work offered to us by the invited authors.

Above all, we are immensely grateful to all authors and reviewers who contributed to this special issue. We would like to take the chance to express our sincerest gratitude to each and every person involved in this project. The effort they have put forth is an expression of the dedicated passion we all share for this field of research. Additionally, we deeply hope that everyone will enjoy reading each and every one of these excellent articles as much as we did.

D. Kordiš and J. Kokosar examined the intron dynamics in the domesticated genes of eutherians, showing some gained with positional bias.

K. S. Swithers et al. reviewed the functional role of reticulate evolution, focusing on horizontal gene transfers, in creating novelties and complexities.

R. Ponce et al. reviewed novel genes, mostly in *Drosophila* and primates, providing details of the chimeric gene *Sdic*.

L. Chen et al. reviewed and discussed the important role alternative splicing may have in generating transcriptome and organism complexity during eukaryotic evolution.

R. Dubruille et al. reviewed and presented their latest results on the evolution of the hiphop/K81 telomere capping genes in *Drosophila*.

L. Huminiecki and G. C. Conant demonstrated the rise to complex innovations in cellular networks that can only be

evolved from whole-genome duplications using examples in fungi and vertebrates.

G. V. Markov and R. J. Sommer reviewed the evolution of novelties among conserved gene families in insects and nematodes genomes.

P. C. Sharma et al. provided a nice example of the potential evolution of microsatellites in duplicated regions of the rice genome.

E. Dubois et al. discussed the impact the domestication of the transposase of a piggyBac had on the spread of Tc1/mariner elements throughout the germline of a Paramecium.

G. Diss et al. provided some evidence that posttranslational regulatory control on a function might influence the divergence between paralogous genes.

O. M. Ramos and D. Ferrier gave an overview on the terminology, mechanisms, and relative frequencies of gene duplications at different scales in shaping animal genome architecture.

S. Good et al. gave force details of the complex evolution of the relationship of the relaxin family genes and their receptors after the fish-specific whole-genome duplication.

E. Betrán et al. considered the evolution of Y chromosome palindromes giving emphasis on the evolutionary role of gene conversion.

M. P. Francino presented very nicely the novelties brought by duplications and horizontal gene transfers and how this can impact a bacterial community.

S. Oulion et al. reconsidered the complex scenarios of duplications and loss events of the FGF family genes during the evolution of the Metazoa.

V. Katju brought a multidimensional consideration of the gene duplication process.

D. Chalopin et al. provided nice examples of genetic innovation in vertebrates by domestication of transposable elements.

Frédéric Brunet
Hideki Innan
Ben-Yang Liao
Wen Wang

Hindawi Publishing Corporation
International Journal of Evolutionary Biology
Volume 2012, Article ID 970920, 8 pages
doi:10.1155/2012/970920

Research Article

Purifying Selection Bias against Microsatellites in Gene Rich Segmental Duplications in the Rice Genome

P. C. Sharma,[1] Manish Roorkiwal,[1,2] and Atul Grover[1,3]

[1] *University School of Biotechnology, Guru Gobind Singh Indraprastha University, Sector 16C, Dwarka, New Delhi 110078, India*
[2] *Centre of Excellence in Genomics, International Crops Research Institute for the Semi-Arid Tropics, Patancheru, Hyderabad 502324, India*
[3] *Biotechnology Division, Defence Institute of Bio-Energy Research, Goraparao, Haldwani 263139, India*

Correspondence should be addressed to P. C. Sharma, prof.pcsharma@gmail.com

Received 20 March 2012; Revised 11 June 2012; Accepted 5 July 2012

Academic Editor: Frédéric Brunet

Little data is available on microsatellite dynamics in the duplicated regions of the rice genome, even though efforts have been made in the past to align genome sequences of its two sub-species. Based on the coordinates of duplicated sequences in the *indica* genome as available in the public domain, we identified microsatellites in these regions. CCG and GAAAA repeats occurred most frequently. In all, 259 microsatellites could be identified in the duplicated sequences using the criteria of minimum 90% alignability spread over a minimum of 1 Kb sequence. More than 25% of the repeats in duplicated regions occurred in the genic sequences. Only 45 (17%) of these 259 microsatellites were found conserved in the duplicated paralogues. Among these repeats, 40% maintained both sequence and length conservation. The effect of mutability of nearby regions could also be clearly seen in microsatellite regions. The overall purpose of this study was to investigate, whether microsatellites follow an independent course of evolutionary dynamics subsequent to events like genome reshuffling that simply drives these elements to different locations in the genome. To the best of our knowledge, this is the first comprehensive analysis of microsatellite conservation in the duplicated regions of any genome.

1. Introduction

Microsatellites represent a class of tandem DNA repeats with 1 to 6 bp long repeat units. These sequences occur in almost all the organisms and frequently constitute the hypervariable regions of the genome. No specific functions have been assigned to most of the microsatellites till date. However, in some cases at least, microsatellite alleles provide protective or adaptive advantage to the host [1]. In many cases, occurrence of different alleles has been found associated with different phenotypes [2]. Microsatellites are not expected to be conserved for long evolutionary periods either, as argued by Buschiazzo and Gemmell [3]. Nevertheless, models of microsatellite mutational dynamics have been developed based on comparison of orthologous microsatellite loci in related taxa [4–7]. However, whether these models also describe microsatellites at paralogous loci created by segmental changes within a genome remains to be investigated.

Availability of whole-genome sequences for rice (*Oryza sativa* L.) allows analysis of noncoding DNA also within the segmentally duplicated regions in addition to the gene order, tandemly arranged genes (TAGs) and gene functions. A collective look emerging from different reports on mapping of duplicated regions in rice genome [8–10] reflects that these studies primarily focused on the analysis of genes in these regions. The strategy commonly used involved making blocks of genes, and mapping them elsewhere in the genome. In a way, the noncoding DNA, particularly, the repetitive DNA has been ignored due to nonemployment of methods suitable for this kind of mapping. Nevertheless, to understand the complete mechanism of speciation and genome evolution, the characterization of conserved noncoding DNA is equally important [11]. No information, to date, is available on the fate of microsatellites in newly duplicated locations. Signatures of ancient duplications, in terms of sequence similarity of genes, and their genomic

order on chromosomes in rice, are widely available, as mapped by Yu et al. [10]. Using the same information as a reference, we have attempted to outline the dynamics of microsatellite DNA within the segmentally duplicated regions of the rice genome to enlighten the patterns of conservation and divergence of these sequences. The overall objective of this study was to investigate whether there is any participation of microsatellites in genome reshuffling or they are simply being carried over. We were also interested to know if after duplication the paralogous microsatellites (we call as "microsatellite twins") follow independent dynamics as both the sites are now different or similar dynamics as the neighbouring environment is still essentially the same. The latter point is important to understand whether microsatellite hypermutability is random or directional.

2. Methods

2.1. Sequence Resources. Whole-genome sequence of *Oryza sativa* subspecies *indica* was downloaded from http://rise.genomics.org.cn/rice/index2.jsp (BGI release 2003-08-01) in FASTA format. Based on the coordinates of duplicated sequences as provided by Yu et al. [10], the sequences of duplicated regions were retrieved from the whole-genome sequence in a text editor and were used as plain text files. The first set of sequences described by Yu et al. [10] has been referred here as group I sequences, and their paralogous duplicated sequences have been designated as group II sequences. These sequences were further split into 2.0 Mb bins for further analysis.

2.2. Analysis of Duplicated Sequences. Repeatmasker (http://www.repeatmasker.org/) with WU-blast [12] search engine was used with default sensitivity and rice as "DNA source" for mining of microsatellite repeats, which were subsequently aligned using glocal algorithm [13] in Vista Genome Browser (http://pipeline.lbl.gov/cgi-bin/gateway2) [14] following the method described earlier by Roorkiwal et al. [7]. A simple sequence with repeat motif length of 1–6 bp spanning a minimal length of 20 bp was considered as a microsatellite. Genes were predicted using MolQuest ver. 1.6.2 (Softberry; http://www.molquest.com/). Following analysis of the aligned map, segmental duplications were identified by the criteria of similarity >90% and length ≥1 Kb [15] and analysed for microsatellites and coordinates of the predicted genes.

2.3. Statistical Analysis. The data generated by mining of duplicated sequences and associated microsatellites were subjected to statistical analysis using χ^2 test and correlation test. The expected values were derived from the published reports [5, 7, 10].

3. Results and Discussion

Microsatellites constitute nearly 1% of the eukaryotic genomes, though in some organisms like *Plasmodium* they may be overrepresented [16]. Their biological significance to the host genomes has been a topic of debate in recent years. Moreover, little knowledge is available about their mutational dynamics [17, 18], primarily derived from the limited genomewide studies in model organisms [4, 5, 7]. Comprehensive surveys on microsatellite conservation across the species and within duplicated sequences of the same genome are, therefore, required to expand our understanding regarding their genomic significance. In the following sections, we present some points emerging from our study justifying our opinion that at least in part such a conservation and maintenance of microsatellites in segmentally duplicated sequences are visible in the rice genome.

3.1. Alignability of Duplicated Regions. Evidences exist for genome duplications in rice that occurred between 53 and 94 mya sometime prior to divergence of the cereal genomes [9, 10]. Further, a segmental duplication event between chromosomes 11 and 12 occurred around 5 mya is also well documented [19], in addition to numerous other individual gene duplications [1, 9]. In totality, the duplicated sequences in rice span 295 Mb, representing nearly two-third of the entire genome including 47% of the genic regions [10]. It is believed that duplication events are followed by several genomic changes including loss of gene functions, and in certain cases, loss of entire genes also [9].

Based on the data presented earlier by Yu et al. [10], we delimited total duplicated regions as 141 Mb of group I sequences and 154 Mb of group II sequences. However, the actual traceable duplicated segments meeting the criteria of >90% similarity and minimum of 1 Kb [15] length covered merely 3.8 Mb genome. The first and second groups of sequences spanned 1.89 and 1.90 Mb of the genome, respectively. Thus, the actual portion of the rice genome studied here came out to be merely 1% (~3.79 Mb). Maximum duplication events were observed on chromosome 2 (~0.34 Mb) and minimum on chromosome 7 spanning little lesser than 0.1 Mb (Table 1). Their distribution was obviously non random with $P(\chi^2) < 0.001$. Further, no correlation was observed between the size of duplicated segments and the length of chromosomes. Average length of bins was found highest on chromosome 5, and minimum on chromosome 6.

The size of the aligned pair and the alignment scores between two segments are generally in inverse relationship to their divergence time. However, in the present case, such a relationship has not been observed, as the most recent pair of duplicated sequences on chromosome 11 and 12 [19] was not the longest one (Table 1). Nevertheless, the mean similarity between the duplicated bins on chromosome 11 and 12 (Figure 1) was little higher at 94%, compared to mean similarity of 93.5% between duplicated bins of chromosome 2 and 4.

3.2. Microsatellite Abundance in Duplicated Regions. We earlier reported 45,782 microsatellites in 374.5 Mb of rice genome [7] using the same criteria and the tools used in the present study. Accordingly, 1% of the genome should

Table 1: Occurrence of genes and microsatellite repeats in duplicated regions of the rice genome.

	Duplicated segments (length covered in bp)	Intergenic	Genic		Gene frequency (bp/gene)	Repeat frequency (bp/repeat)
			Exon	Intron		
Chromosome 1 corresponding chromosome 5						
Segment 1.1	58 (81685)	46	9	3	6807.08	16337
Segment 5A1.1	50 (75106)	21	20	9	2589.86	8345.11
Segment 1.2	6 (9866)	2	3	1	2466.5	0
Segment 5A1.2	4 (5909)	2	1	1	203.76	0.00
Segment 1.3	163 (244228)	98	42	23	3757.35	9769.12
Segment 5A1.3	169 (268072)	110	41	18	9243.86	9928.59
Segment 1.4	3 (4300)	2	0	1	4300	0
Segment 5A1.4	1 (2247)	1	0	0	77.48	2247.00
Chromosome 2 corresponding chromosomes 4 and 6						
Segment 2.1	342 (518707)	200	89	53	3652.87	17886.45
Segment 4A2.1	347 (522868)	199	97	51	18029.93	13071.70
Segment 2.2	102 (149764)	49	43	10	2825.74	29952.8
Segment 6A2.2	105 (146574)	56	35	14	5054.28	24429.00
Segment 2.3	81 (124845)	50	14	17	4027.26	15605.63
Segment 6A2.3	77 (114157)	45	22	10	3936.45	16308.14
Chromosome 3 corresponding chromosomes 7, 10, and 12						
Segment 3.1	29 (42425)	14	11	4	2828.33	21212.5
Segment 7A3.1	31 (47154)	21	9	1	1626.00	23577.00
Segment 3.2	29 (41410)	14	12	3	2760.67	20705
Segment 7A3.2	36 (49456)	22	9	5	1705.38	16485.33
Segment 3.3	37 (59771)	26	5	6	5433.73	11954.2
Segment 10A3.3	42 (66214)	24	9	9	2283.24	16553.50
Segment 3.4	23 (28749)	15	1	7	3593.63	28749
Segment 10A3.4	28 (39198)	16	6	6	1351.66	19599.00
Segment 3.5	29 (41024)	15	10	4	2930.29	41024
Segment 12A3.5	24 (37014)	13	8	3	1276.34	18507.00
Chromosome 4 corresponding chromosomes 8 and 10						
Segment 4.1	17 (26044)	7	6	4	2604.4	8681.33
Segment 8A4.1	16 (21129)	11	4	1	728.59	0.00
Segment 4.2	40 (62581)	22	12	6	3476.72	15645.25
Segment 10A4.2	40 (59065)	22	11	7	2036.72	59065.00
Chromosome 8 corresponding chromosome 9						
Segment 8.1	28 (36632)	21	4	3	5233.14	6105.33
Segment 9A8.1	33 (42741)	28	3	2	1473.83	8548.20
Segment 8.2	130 (191894)	73	47	10	3366.56	31982.33
Segment 9A8.2	122 (180824)	72	41	9	6235.31	12054.93
Chromosome 11 corresponding chromosome 12						
Segment 11.1	111 (168247)	51	40	20	2804.12	12017.64
Segment 12A11.1	101 (158798)	45	39	17	5475.79	14436.18
Segment 11.2	43 (59793)	25	15	3	3321.83	8541.86
Segment 12A11.2	47 (65442)	20	18	9	2256.62	21814.00

have carried 458 microsatellites, had they been randomly distributed throughout the genome. However, only 259 microsatellites could be identified in this set of sequences, with 121 sequences identified in shorter set of 1.89 Mb, with an average frequency of one repeat locus per 16,453 bp, and 138 in group II of 1.9 Mb with average frequency of one repeat locus per 15,831 bp (Figure 2). When the frequency of specific microsatellite motifs in duplicated regions were

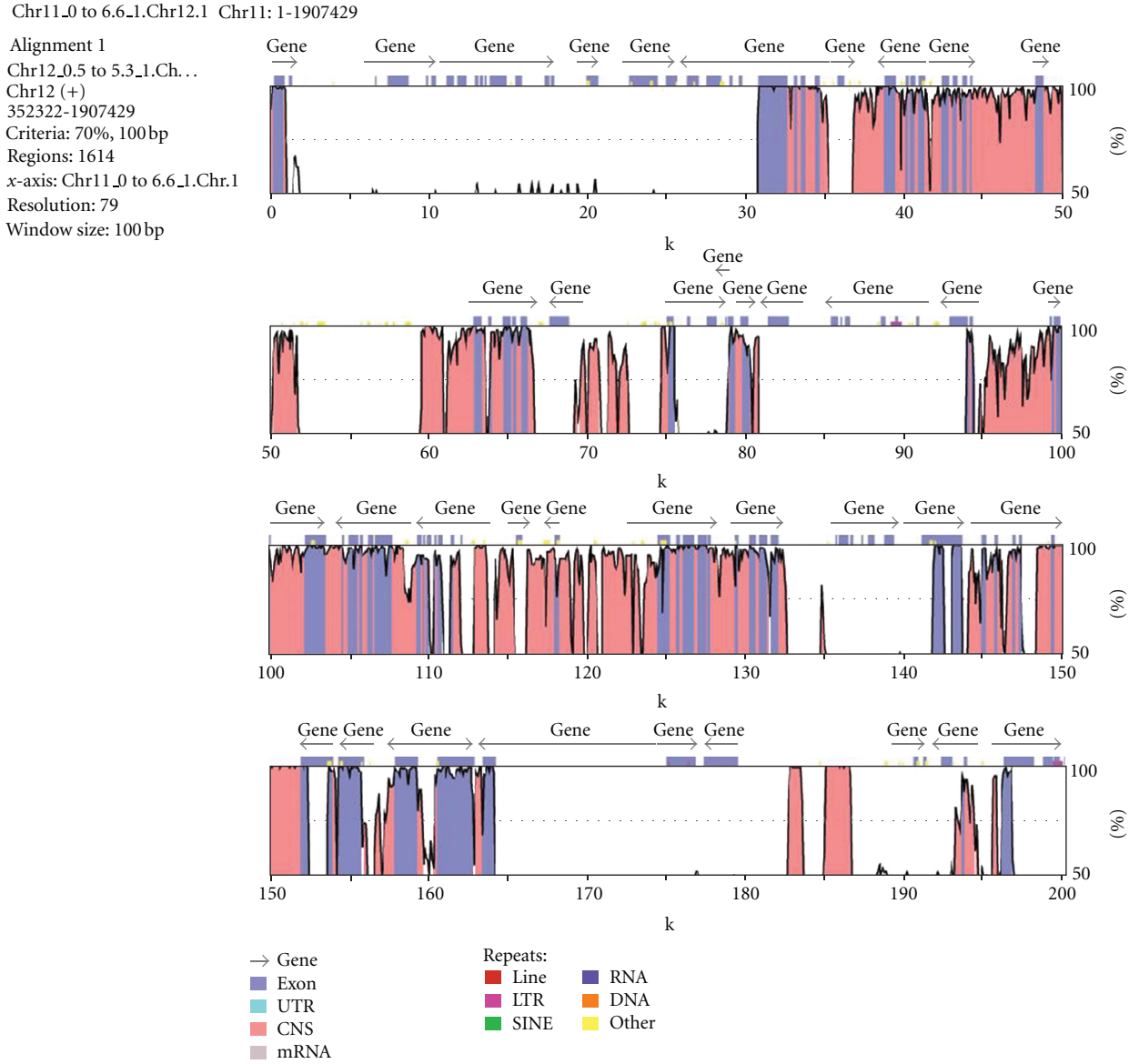


Figure 1: A representative figure of a duplicated segment mapped between chromosomes 11 and 12.

plotted against the expected values, based on previous studies [5, 7], frequency of most of the microsatellites were found much lower $P(\chi^2) < 0.001$, except for motifs like AAT, AGC, and CCG for which observed values corresponded to expected values. Clearly, there is certain level of purifying selection against the microsatellites in these duplicated regions of the rice genome.

CCG repeats (and direct and reverse complementary permutations thereof) were found most abundant in either set of sequences in consistency with the earlier reports [5, 7]. GAAAA repeats (and their permutations), known to be most abundant in rice genome among the penta-nucleotide repeats [5], were found the second most abundant and least mutable repeats (Table 2) among the duplicated sequences. Other repeats like A, AT, and so forth, otherwise abundant in rice genome, were not found preferentially distributed in duplicated regions (Figure 3). Relative abundance of each of the repeat motif in both of the sets of sequences was fairly comparable. Quite expectedly, majority of the microsatellites occurred in the intergenic sequences (Table 2), and least in the exonic sequences. Consistent with the previous findings [7], CCG repeats most frequently occurred in exonic sequences. As suggested earlier by some researchers [17, 18], intrinsic factors specific to the host genome and microsatellite themselves like repeat length, repeat sequence, neighboring genomic sequences, and so forth, are responsible for differential occurrence and conservation of microsatellites. Importantly, while the duplicated sequences have shown a higher frequency of genes, they have particularly shown a bias against the microsatellites (Figure 2).

Table 2: Traceability of microsatellites originating from group I sequences into group II sequences.

Motif	Region	Length (bp) in group I sequences	Traceability in group II sequences		
			Equal	Short	Long
Chromosome 1 corresponding chromosome 5			9	2	2
(CCG)n	Intergenic	58	√		
(CCG)n	Intergenic	78		√	
(CGG)n	Intergenic	60			√
(CGG)n	Intergenic	60			√
(GAAAA)n	Intergenic	26	√		
(GAAAA)n	Intergenic	33		√	
(TTTTC)n	Intergenic	26	√		
(TTTTC)n	Intergenic	26	√		
(TTTTC)n	Intergenic	26	√		
(TTTTC)n	Intron	26	√		
(TTTTC)n	Intron	26	√		
(TTTTC)n	Intergenic	22	√		
(TTTTC)n	Intergenic	26	√		
Chromosome 2 corresponding chromosome 4			6	2	4
(CCG)n	Intron	174		√	
(CGA)n	Intron	150			√
(CGA)n	Intron	150			√
(CGG)n	Intergenic	58	√		
(CGG)n	Intergenic	58	√		
(CGG)n	Intergenic	211			√
(CGG)n	Intergenic	126			√
(CGG)n	Intergenic	211		√	
(GAAAA)n	Intergenic	28	√		
(TTTTC)n	Intron	22	√		
(TTTTC)n	Intergenic	28	√		
(TTTTC)n	Intergenic	27	√		
Chromosome 2 corresponding chromosome 6			2	1	2
(CCG)n	Intron	74			√
(CCG)n	Intergenic	123			√
(CCG)n	Intergenic	75		√	
(TTTTC)n	Intron	27	√		
(TTTTC)n	Intergenic	27	√		
Chromosome 3 corresponding chromosomes 7, 10, and 12			0	0	2
(CGG)n	Intergenic	59			√
(GAAAA)n	Intergenic	22			√
Chromosome 4 corresponding chromosomes 8 and 10			0	0	0
Chromosome 8 corresponding chromosome 9			1	2	1
(CCG)n	Intergenic	72			√
(CCG)n	Intergenic	155		√	
(CCG)n	Intergenic	199		√	
(TAA)n	Intergenic	29	√		

Table 2: Continued.

Motif	Region	Length (bp) in group I sequences	Traceability in group II sequences		
			Equal	Short	Long
Chromosome 11 corresponding chromosome 12			1	2	1
(CCG)n	Exon	76			√
(CCG)n	Exon	154		√	
(CGG)n	Intergenic	147		√	
(TCG)n	Exon	70	√		

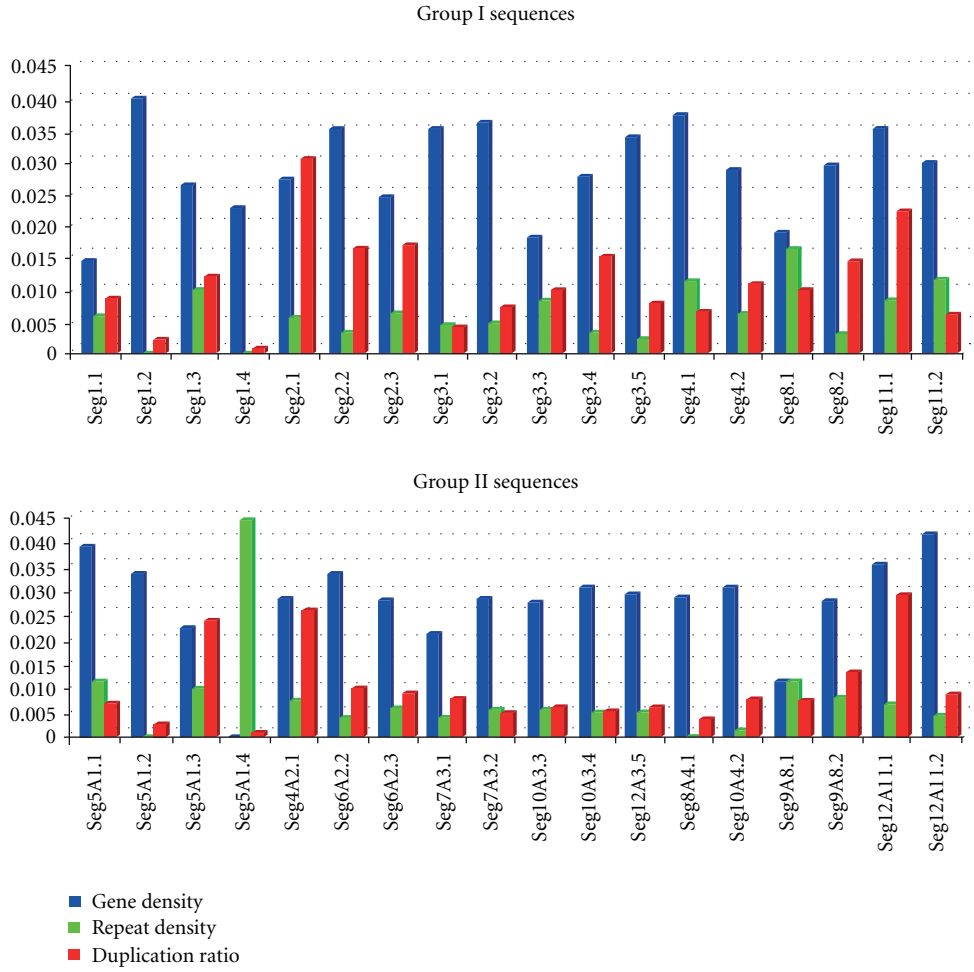


Figure 2: Gene versus repeat density on the entire duplicated segments in the rice genome. Duplication ratio refers to the ratio of the segment reported duplicated by Yu et al. [10], and the length of the fragments that we found aligning with >90% similarity for a minimum length of 1 Kb.

3.3. Microsatellite Conservation within the Duplicated Sequences. Out of the 259 microsatellites existing in the duplicated sequences, only 45 (17%) were found conserved in the paralogous sequences. Considering the mutability of microsatellites per locus per generation in rice, as described by Grover et al. [5], a microsatellite of 20 bp length may entirely be lost in around 2 million years provided all the mutations are unidirectional, targeting the shortening of the microsatellite. Thus, conservation of 17% of microsatellites in duplicated regions, with the average age of duplication around 56 mya, is especially significant as only 1% of the entire duplication blocks is identifiable today (discussed above). Interestingly, 42% of these repeats have their length conserved, which is significantly lesser than the global average in rice observed earlier [7], but clearly indicating that these alleles have been fixed in duplicated segments, most probably due to the vitality of their spatial occurrence [18]. Differences in the lengths of at least two paralogous microsatellites (with CCG motif) falling in exonic sequences on duplicated blocks on chromosome 11 and 12 indicate

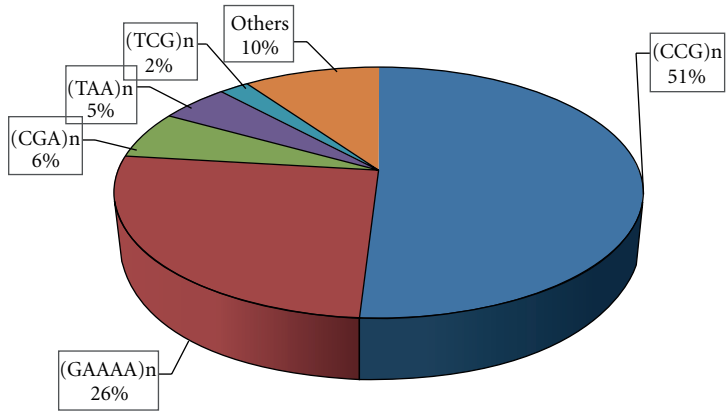


Figure 3: Abundance of microsatellite motifs in duplicated regions of the rice genome.

Table 3: Description of paralogous loci where microsatellite motif has been found altered either by splitting and integrating, or replaced with another motif.

Duplication pair	Motif at group I site	Motif at group II site
DP 1A5	(CCG)n	(TCC)n
	(TTAA)n	(CCG)n
	(CGG)n	(CCG)n
	(CGA)n	
DP 2A4	(CGG)n	(CGA)n
DP 8A9	(CCG)n	(CCG)n
	(TCG)n	
	(TAA)n	(CGA)n
DP 11A12	(CCG)n	(CCG)n
		(CCG)n
	(CGG)n	(CGA)n
	(CCG)n	(CCG)n
		(CCG)n
	(CCG)n	(CCG)n
		(CCG)n

the relative advantage of repeatability and hypermutability of microsatellites in genes, as has been suggested earlier as well [1, 3, 20–22].

It was also interesting to note that at some of the genomic positions a single microsatellite repeat corresponded to two microsatellite repeats with the same motif (Table 3). This is possible due to recurring splitting and expansion events at microsatellite loci [18]. Of all the paralogous microsatellites observed, 40% maintained both sequence and length characteristics. Majority of these microsatellites were located on duplicated segments of chromosomes 1 and 5. It is quite possible, that these loci might have been fixed. However, we do not overrule the possibility that one or both of the sequences have undergone a number of mutations purely in stochastic manner and eventually arriving to the same lengths simultaneously, now seen as conserved alleles. Out of these two possibilities, it is the first one that generates more interest, as microsatellites associated with important regions in the genome will display lower variability during genetic drift and selective sweeps [18, 23]. Consequently, lesser activity will be observed on a microsatellite locus that is lying next to a genomic region adapted to a given environment [24]. Therefore, we do not overrule the possibility that the microsatellites that show sequence as well as length conservation represent important "evolutionary chronometers" [25] and might have been tightly linked to genomic regions of significance [18]. Microsatellites located in mutationally constrained regions are expected to be maintained passively. Highly conserved microsatellites are often associated with other conserved genomic elements and show a stronger negative relationship with single nucleotide polymorphisms (SNPs) density [26]. Interestingly enough, in five instances, a particular microsatellite motif has given way to another motif, precisely at the same site (Table 3). Grover and Sharma [18] explained such events by calling them as "metamorphosis" at microsatellite sites. Apparently, in three of the five cases, the new microsatellites appeared originally by a single site substitution, which later expanded possibly by "polymerase slippage" to mature into a fully grown microsatellite. Evidently, both the abundance and conservation of microsatellites had a heterogeneous pattern across the rice chromosomes. However, the distribution of sequence motifs across the chromosomes and across the blocks and segments of duplications more or less remained the same. Conserved microsatellites within the duplicated regions of the genome are desired candidates to study the overall significance of microsatellite conservation in different genomes.

3.4. Microsatellites versus Genes in Segmentally Duplicated Regions. Out of 259, only 68 (26.25%) microsatellites were found to be associated with genes. Out of these genic microsatellites, 17 (25%) were present in exonic regions and remaining 51 (75%) were located in the intronic regions. Interestingly, 18 of the repeats and their counterparts were located to different genomic entities. For example, while one locus was located in the intergenic region, its paralgoue occurred in the genic region. Such spatial distribution can occur due to homologous recombination [27] or some other

minor genomic rearrangements due to retrotransposition, local genomic reorganization and reshuffling. Thus, such microsatellites can be considered as "genomic fossils," which can help in retracing the evolutionary events in the genome.

Acknowledgments

The authors' microsatellite research has been supported by Indian Council of Agricultural Research (ICAR), Department of Science and Technology (DST) and Defence Research and Development Organization (DRDO). M. Roorkiwal acknowledges research fellowship from University Grants Commission (UGC).

References

[1] Y. Kashi and D. G. King, "Simple sequence repeats as advantageous mutators in evolution," *Trends in Genetics*, vol. 22, no. 5, pp. 253–259, 2006.

[2] J. B. W. Wolf, C. Harrod, S. Brunner, S. Salazar, F. Trillmich, and D. Tautz, "Tracing early stages of species differentiation: ecological, morphological and genetic divergence of Galápagos sea lion populations," *BMC Evolutionary Biology*, vol. 8, article 150, 2008.

[3] E. Buschiazzo and N. J. Gemmell, "Conservation of human microsatellites across 450 million years of evolution," *Genome Biology and Evolution*, vol. 2, pp. 153–165, 2010.

[4] T. Barbará, C. Palma-Silva, G. M. Paggi, F. Bered, M. F. Fay, and C. Lexer, "Cross-species transfer of nuclear microsatellite markers: potential and limitations," *Molecular Ecology*, vol. 16, no. 18, pp. 3759–3767, 2007.

[5] A. Grover, V. Aishwarya, and P. C. Sharma, "Biased distribution of microsatellite motifs in the rice genome," *Molecular Genetics and Genomics*, vol. 277, no. 5, pp. 469–480, 2007.

[6] A. Grover, B. Ramesh, and P. C. Sharma, "Development of microsatellite markers in potato and their transferability in some members of Solanaceae," *Physiology and Molecular Biology of Plants*, vol. 15, no. 4, pp. 343–358, 2009.

[7] M. Roorkiwal, A. Grover, and P. C. Sharma, "Genome-wide analysis of conservation and divergence of microsatellites in rice," *Molecular Genetics and Genomics*, vol. 282, no. 2, pp. 205–215, 2009.

[8] R. Guyot and B. Keller, "Ancestral genome duplication in rice," *Genome*, vol. 47, no. 3, pp. 610–614, 2004.

[9] X. Wang, X. Zhao, J. Zhu, and W. Wu, "Genome-wide investigation of intron length polymorphisms and their potential as molecular markers in rice (*Oryza sativa* L.)," *DNA Research*, vol. 12, no. 6, pp. 417–427, 2005.

[10] J. Yu, J. Wang, W. Lin et al., "The genomes of *Oryza sativa*: a history of duplications," *PLoS Biology*, vol. 3, no. 2, article e38, pp. 266–281, 2005.

[11] D. Retelska, E. Beaudoing, C. Notredame, C. V. Jongeneel, and P. Bucher, "Vertebrate conserved non coding DNA regions have a high persistence length and a short persistence time," *BMC Genomics*, vol. 8, article 398, 2007.

[12] S. F. Altschul, T. L. Madden, A. A. Schäffer et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[13] M. Brudno, S. Malde, A. Poliakov et al., "Glocal alignment: finding rearrangements during alignment," *Bioinformatics*, vol. 19, supplement 1, pp. i54–i62, 2003.

[14] K. A. Frazer, L. Pachter, A. Poliakov, E. M. Rubin, and I. Dubchak, "VISTA: computational tools for comparative genomics," *Nucleic Acids Research*, vol. 32, supplement 2, pp. W273–W279, 2004.

[15] L. Zhang, H. H. S. Lu, W. Y. Chung, J. Yang, and W. H. Li, "Patterns of segmental duplication in the human genome," *Molecular Biology and Evolution*, vol. 22, no. 1, pp. 135–141, 2005.

[16] P. C. Sharma, A. Grover, and G. Kahl, "Mining microsatellites in eukaryotic genomes," *Trends in Biotechnology*, vol. 25, no. 11, pp. 490–498, 2007.

[17] A. Bhargava and F. F. Fuentes, "Mutational dynamics of microsatellites," *Molecular Biotechnology*, vol. 44, no. 3, pp. 250–266, 2010.

[18] A. Grover and P. C. Sharma, "Is spatial occurrence of microsatellites in the genome a determinant of their function and dynamics contributing to genome evolution?" *Current Science*, vol. 100, no. 6, pp. 859–869, 2011.

[19] The Rice Chromosomes 11 and 12 Sequencing Consortia, "The sequence of rice chromosomes 11 and 12, rich in disease resistance genes and recent gene duplications," *BMC Biology*, vol. 3, article 20, 2005.

[20] J. W. Fondon III and H. R. Garner, "Molecular origins of rapid and continuous morphological evolution," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 52, pp. 18058–18063, 2004.

[21] J. M. Hancock and M. Simon, "Simple sequence repeats in proteins and their significance for network evolution," *Gene*, vol. 345, no. 1, pp. 113–118, 2005.

[22] D. E. Riley and J. N. Krieger, "UTR dinucleotide simple sequence repeat evolution exhibits recurring patterns including regulatory sequence motif replacements," *Gene*, vol. 429, no. 1-2, pp. 80–86, 2009.

[23] C. Schlötterer, "Hitchhiking mapping—functional genomics from the population genetics perspective," *Trends in Genetics*, vol. 19, no. 1, pp. 32–38, 2003.

[24] C. Schlötterer, M. Kauer, and D. Dieringer, "Allele excess at neutrally evolving microsatellites and the implications for tests of neutrality," *Proceedings of the Royal Society B Biological Science*, vol. 271, no. 1541, pp. 869–874, 2004.

[25] C. A. Driscoll, M. Menotti-Raymond, G. Nelson, D. Goldstein, and S. J. O'Brien, "Genomic microsatellites as evolutionary chronometers: a test in wild cats," *Genome Research*, vol. 12, no. 3, pp. 414–423, 2002.

[26] M. Brandström and H. Ellegren, "Genome-wide analysis of microsatellite polymorphism in chicken circumventing the ascertainment bias," *Genome Research*, vol. 18, no. 6, pp. 881–887, 2008.

[27] M. Brandström, A. T. Bagshaw, N. J. Gemmell, and H. Ellegren, "The relationship between microsatellite polymorphism and recombination hot spots in the human genome," *Molecular Biology and Evolution*, vol. 25, no. 12, pp. 2579–2587, 2008.

Hindawi Publishing Corporation
International Journal of Evolutionary Biology
Volume 2012, Article ID 310278, 14 pages
doi:10.1155/2012/310278

Research Article

New Insights into Ligand-Receptor Pairing and Coevolution of Relaxin Family Peptides and Their Receptors in Teleosts

Sara Good,[1] Sergey Yegorov,[1] Joran Martijn,[2] Jens Franck,[1] and Jan Bogerd[2]

[1] *Department of Biology, University of Winnipeg, Winnipeg, MB, Canada R3B 2E9*
[2] *Department of Biology, Faculty of Science, University of Utrecht, 3584 CH Utrecht, The Netherlands*

Correspondence should be addressed to Sara Good, s.good@uwinnipeg.ca and Jan Bogerd, j.bogerd@uu.nl

Received 30 March 2012; Revised 7 June 2012; Accepted 15 June 2012

Academic Editor: Frédéric Brunet

Relaxin-like peptides (RLN/INSL) play diverse roles in reproductive and neuroendocrine processes in placental mammals and are functionally associated with two distinct types of receptors (RXFP) for each respective function. The diversification of RLN/INSL and RXFP gene families in vertebrates was predominantly driven by whole genome duplications (2R and 3R). Teleosts preferentially retained duplicates of genes putatively involved in neuroendocrine regulation, harboring a total of 10-11 receptors and 6 ligand genes, while most mammals have equal numbers of ligands and receptors. To date, the ligand-receptor relationships of teleost Rln/Insl peptides and their receptors have largely remained unexplored. Here, we use selection analyses based on sequence data from 5 teleosts and qPCR expression data from zebrafish to explore possible ligand-receptor pairings in teleosts. We find support for the hypothesis that, with the exception of RLN, which has undergone strong positive selection in mammalian lineages, the ligand and receptor genes shared between mammals and teleosts appear to have similar pairings. On the other hand, the teleost-specific receptors show evidence of subfunctionalization. Overall, this study underscores the complexity of RLN/INSL and RXFP ligand-receptor interactions in teleosts and establishes theoretical background for further experimental work in nonmammals.

1. Introduction

Relaxin-like peptides are members of the insulin superfamily and, like insulin and insulin-like growth factors (IGF), are small peptides (~60 amino acids) that share a common two-domain structure (A and B domains) in their mature form [1]. Functionally, however, relaxin family peptides are different from insulin and IGF: they bind to unrelated receptors and play diverse roles in reproduction and neuroendocrine regulation as opposed to carbohydrate/fat metabolism and growth. Four relaxin family peptide-encoding genes (*RLN*, *RLN3, INSL3*, and *INSL5*) originated early in vertebrate history and are shared by most vertebrates [2]. The receptors for the RLN/INSL peptides belong to two distinct groups of G protein-coupled receptors (GPCR), collectively named the relaxin family peptide receptors (RXFP) [3].

In mammals, there are four known receptors, RXFP1–4, associated with the four relaxin family ligands. RXFP1 and RXFP2 are evolutionarily related to glycoprotein hormone receptors (e.g., luteinizing and follicle-stimulating hormone receptors), containing a large extracellular domain made up of ten leucine-rich repeats (LRR) and a low-density lipoprotein receptor type A (LDLa) module; they are the cognate receptors for the ligands RLN and INSL3 in humans, both of which primarily have reproductive actions [3]. On the other hand, RXFP3 and RXFP4 are classic type I peptide GPCRs with short N-terminal domains; they are evolutionarily related to somatostatin and angiotensin receptors and, in humans, are the cognate receptors for RLN3 and INSL5, both of which are associated with neuroendocrine signaling [3].

The two hormones with reproductive functions in mammals, RLN and INSL3, are the best understood. The hormone RLN is well known for its role in parturition, where it softens connective tissues of the reproductive tract via tissue remodeling and prepares the mammary glands for lactation, but it has numerous other physiological actions as well [1]; its receptor (RXFP1) also exhibits a wide distribution suggesting endocrine action in mammals [7] (Table S1, see supplementary materials available online at

doi: 10.1155/2012/310278). In teleosts, the peptide sequence of Rln is highly similar to that of Rln3 [8]; although its function remains unknown, the *rln* gene exhibits substantial overlap in expression with *rln3*, both being highly expressed in brain, although teleost *rln* is also significantly expressed in gonads [9]. While mammalian and teleost *RLNs* differ somewhat in their expression patterns, *INSL3* has a more similar expression pattern in the two lineages; it is highly expressed in Leydig cells in both mammals [10] and teleosts [8], and at lower levels in other tissues (see Table S1). In mammals, the receptor for INSL3, RXFP2, is also highly expressed in testes suggesting paracrine action [6], but lower levels of RXFP2 expression are observed in a wide array of tissues [7]. The receptor has been, until now, unstudied in teleosts.

The peptides RLN3 and INSL5 exert their influence primarily through the hypothalamic-pituitary-gonadal (HPG) axis [11, 12]. RLN3 is the most conserved member of the family; it is predominantly expressed in the nucleus incertus (NI) in mammalian brain [13] and its homologous region in teleosts [14]. Ascending RLN3-producing projections from the NI innervate a broad range of RXFP3-expressing regions of the forebrain in mammals, including the hypothalamus and it is implicated in the acute stress response and regulation of food intake [12, 15]. Collectively, these lines of evidence suggest that RLN3 acts through the HPG axis and may play a dual role linking nutritional status to reproductive function [12]. Lastly, INSL5 is the least well understood member of the family, but in humans its primary sites of expression are rectum, colon, and uterus [16, 17] (see Table S1). The receptor for INSL5 in mammals, RXFP4, has a wide distribution being found in colon, placenta, testis, thymus, prostate, kidney, and brain in human [18], strongly suggesting endocrine action.

Despite the evolutionary distance separating RXFP1/2 and RXFP3/4-type receptors, experimental studies have shown that some RLN/INSL peptides can bind additional (secondary) receptors at lower affinity [5]. For example, in addition to RXFP3, RLN3 can bind to and activate RXFP1 and RXFP4, RLN can bind to RXFP2 in addition to RXFP1 [19], and INSL5 can bind to (but activate only weakly) RXFP3 in addition to its primary receptor RXFP4. Such "primary" and "secondary" ligand-receptor interactions have been demonstrated for human RLN/INSL-RXFP pairs, but analogous pairings in other vertebrates, such as teleosts, in which relaxin family peptide-receptor signaling and diversification have taken an evolutionary pathway distinct from that in mammals [2], remain to be established.

Recent evolutionary analyses revealed that vertebrate *RLN/INSL* genes and their receptors primarily diversified through the two rounds (2R) of whole genome duplication (WGD) that occurred in early vertebrate evolution and, in teleosts, during the teleost fish-specific WGD (3R) (Figure 1). To summarize, mammals retained 4 ligand and 4 receptor genes following 2R, while teleosts have 10 (most teleosts) or 11 (zebrafish) receptor and 6 ligand genes following 3R (Figure 1) and after-3R local duplications (Figure 2, Table S2). Many of the genes retained in duplicate in fish (*rln3*-, *insl5*-, and *rxfp3*-type genes) are hypothetically involved in neuroendocrine regulation (Figure 3). But due to a lack of understanding of the evolutionary history of *rln/insl* and *rxfp* genes in teleosts, the ligand-receptor pairings in teleosts are virtually unknown.

One of the interesting aspects of the evolution of RLN/INSL peptides is how a set of relatively closely related ligands signals via two unrelated types of receptors. Yegorov and Good [2] hypothesized that this dual-functioning arose in the ancestral pre-2R RLN/INSL peptide that had roles in both reproductive (via RXFP1/2-receptor) and neuroendocrine (via RXFP3/4) regulations in primitive vertebrates (Figure 3). As a result of the WGDs, the ancestral tripartite system gave rise to two distinct parties of RLN/INSL-RXFP ligand-receptor pairs (Figure 3). Curiously, it can be observed that, with the exception of the RXFP1 receptor and its ligand RLN, each of the duplication events resulted in a single ligand that potentially could function with two related receptors (Figure 3). In most mammals, this tripartite model became reduced to a 1 : 1 relationship for ligands and receptors after the divergence of tetrapods from the gnathostome ancestor (as described above), but in teleosts, there are multiple receptors for some ligands, which may have occurred through receptor subfunctionalization (Figure 3). Based on the evolutionary history of duplication, and the ligand-receptor pairings in mammals, we developed hypotheses concerning which ligand-receptor pairings we expect in teleosts (Figure 4). The primary goal of this paper is to test our hypotheses about the Rln/Insl-Rxfp ligand-receptor pairs in teleosts using selection analyses and experimental qPCR data from zebrafish.

2. Results

2.1. Selection Analyses. We performed two kinds of molecular evolutionary analyses to (1) hypothesize which ligand-receptor pairings may occur in teleosts and (2) examine differences in selection among mammalian and teleost genes.

(1) Previous studies have used the correlation of evolutionary distances between putative ligand-receptor pairs as evidence of cofunctioning [20, 21]. Here, we employed a similar correlation approach, but rather than comparing the mean evolutionary distances among gene pairs, we compared the proportion of sites under different forms of selection (purifying, neutral, or positive) in pairs of teleost genes to the "primary" ligand-receptor pairs known to exist in mammals, *rln-rxfp1*, *insl3-rxfp2*, and *rln3-rxfp3-1*. If the genes coding for the ligands and receptors coevolve, we expect a correlation in the rates and types of selection on ligand-receptor pairs. This would correspond to values falling along the (0, 0 : 1, 1) plane of the XY-plot. On the other hand, a similar [X,Y]-value for the same ligand-receptor pair in mammals and teleosts would suggest that the pair plays a similar role in the two lineages.

(2) We tested for evidence of (a) codon-specific positive selection in mammalian and teleost ligand and receptor genes and (b) codon-specific positive selection in mammalian versus teleost genes using the branch-site model of

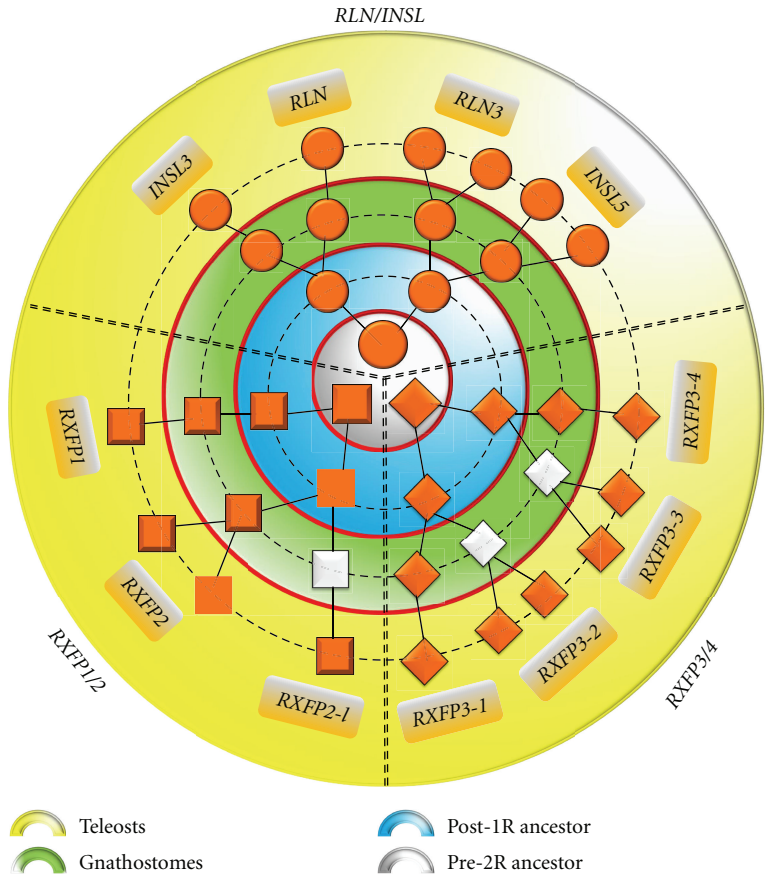


Figure 1: The role of Whole Genome Duplications (WGD) in the expansion of the *RLN/INSL* and *RXFP* genes in vertebrates. The three gene families (one ligand (circles) and two receptor (squares) families) arose as a result of WGDs (1R, 2R, and 3R) from three ancestral genes. *RLN/INSL* peptides: following 1R, there were two *RLN/INSL*-like genes, following 2R one of these gave rise to *RLN3* and *INSL5*, and the other to *RLN* and *INSL3*. After the teleost fish-specific WGD (3R), the duplicates of *rln3* and *insl5* were retained bringing the total number of *rln/insl* genes in teleosts to 6. *RXFP3/4* receptors: four *RXFP3/4*-type receptor genes were generated from a single-ancestral gene during 2R; all four of these genes were retained in teleosts, but in tetrapods, only two receptors, *RXFP3* (termed *RXFP3-1*) and *RXFP4* (*RXFP3-4*), were retained. After 3R, the duplicates of *rxfp3-2* and *rxfp3-3* were retained. *RXFP1/2* receptors: most vertebrates have only a single copy of *RXFP1* and *RXFP2,* a few (opossum, frog, reptiles, and zebrafish) have *RXFP2-like*. Layers coloured in four distinct colors indicate ancestral stages (legend below); WGDs are depicted as red lines surrounding these ancestral stages. White shapes indicate genes lost in most (*RXFP2-l*, *RXFP3-3*) or all (*RXFP3-2*) tetrapod lineages. *RXFP2-l*= *RXFP2-like*. Based on Yegorov and Good [2].

positive selection. While the first analysis (a) tests whether specific codons have been positively selected within lineages, the second (b) looks for evidence that codons have been differentially selected in mammalian versus teleost lineages.

(1) Evidence for Ligand-Receptor Coevolution for Mammalian and Teleost Orthologs. Between 70 and 93% of the sites across all genes, and in both mammals and teleosts, have been subject to purifying selection (Figure 5(a)). Additionally, the extent of purifying selection was symmetric for the ligand-receptor pairs *rln3-rxfp3* and *insl3-rxfp2* suggesting close coevolution, while for the remaining two pairs, *rln-rxfp1* and *insl5-rxfp4*, the proportion of sites under purifying selection was higher for the receptor genes (between 0.7 and 0.92) than for the ligands (ranging from 0.4–0.95), suggesting a more diffuse coevolution (or no coevolution), and more relaxed evolution on the ligand.

On the other hand, there are significantly fewer sites which are evolving neutrally (Figure 5(b)) or are subject to positive selection (Figure 5(c)). For the receptor genes, from 3 to 20% of the sites were found to be evolving neutrally (Figure 5(b)), and from 2 to 13% were subject to positive selection; *rxfp3* exhibits the fewest neutral or positively selected sites, *rxfp4* has the highest proportion of sites under neutral evolution and *rxfp2* exhibits the highest proportion of sites under positive selection. Largely due to the anomalous nature of asymmetric selection on the *rln-rxfp1* ligand-receptor system in mammals, the extent of neutral and positive selection among ligand genes varied widely between mammals and teleosts, primarily because teleost *rln* was found to have a large number of sites evolving neutrally, whereas mammalian *RLN* has a large proportion

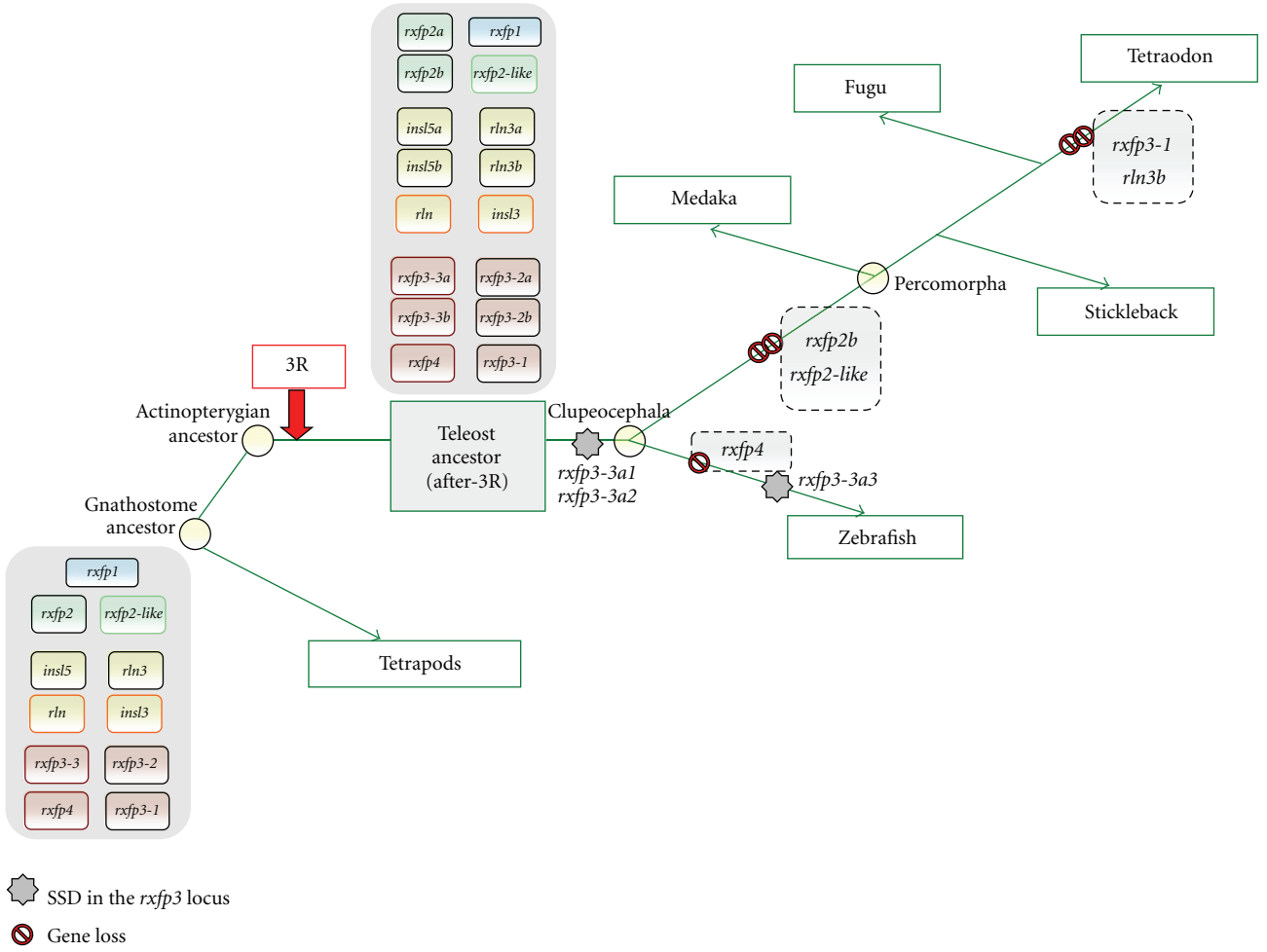


Figure 2: Post 3R gene loss and gain in five teleost fish species. Following 3R, teleosts start with a gene set composed of 10 receptors and 6 ligands. Prior to divergence of zebrafish, *rxfp3-3a* is locally duplicated, generating tandem genes *rxfp3-3a1* and *rxfp3-3a2*. Zebrafish retains most of the genes, except *rxfp4*, but gains an additional copy of the *rxfp3-3* gene, *rxfp3-3a3*, through SSD. Other teleosts lose *rxfp2-like* and also the 3R-duplicate *rxfp2b*. SSD: small-scale (local) duplication. 3R: fish-specific WGD. Data from Yegorov and Good [2]. Phylogeny and classification of fish adapted from Kinoshita et al. [4].

of sites subject to positive selection (Figures 5(b) and 5(c), resp.).

The Selection Analysis Supports Our Hypothesis for Many Ligand-Receptor Pairs in Teleosts, but the Receptors for the Two Insl5 Paralogs Remain Unclear. Given the presence of additional ligand and receptor genes in teleosts for which no ortholog was present in mammals, the correlation approach could not be used for the additional ligand-receptor genes in teleosts because there was no reference comparison in mammals and too many possible pairs to consider. Thus, to examine the possible pairings of these additional genes, we simply plotted the proportion of sites subject to each form of selection in teleosts for visual comparison (Figure 6). This revealed that the gene coding for Rln has a higher number of neutrally evolving sites than the gene of its proposed receptor, Rxfp1, although this may be an artifact of the comparison to mammalian RLN. On the other hand, the numbers of selected sites in the genes of the proposed ligand-receptor pairs *insl3-rxfp2* (as demonstrated above), *rln3a-rxfp3-2a/rxfp3-2b,* and *rln3b-rxfp3-1* were similar, supporting possible cofunctioning, although *rxfp3-2a* shows a higher fraction of positively selected sites than either of the *rln3* ligand genes. Lastly, however, there was also a poor correlation in the expected selection profile of *insl5* compared with its proposed receptor genes: both teleost *insl5a* and *insl5b* evolve relatively neutrally but none of their proposed receptors do, with the exception of *rxfp4*, which has a slightly higher rate of neutral and positive selection. The remaining three *rxfp3-3* receptor genes are very conserved (Figure 6). Thus, although teleost *insl5* and *rxfp4* genes had similar selection profiles to those of mammals (see above), suggesting a conserved function between the two lineages, the other three proposed receptors for the *insl5* paralogs (i.e., *rxfp3-3a1*, *rxfp3-3a2* and *rxfp3-3b*) exhibited strong

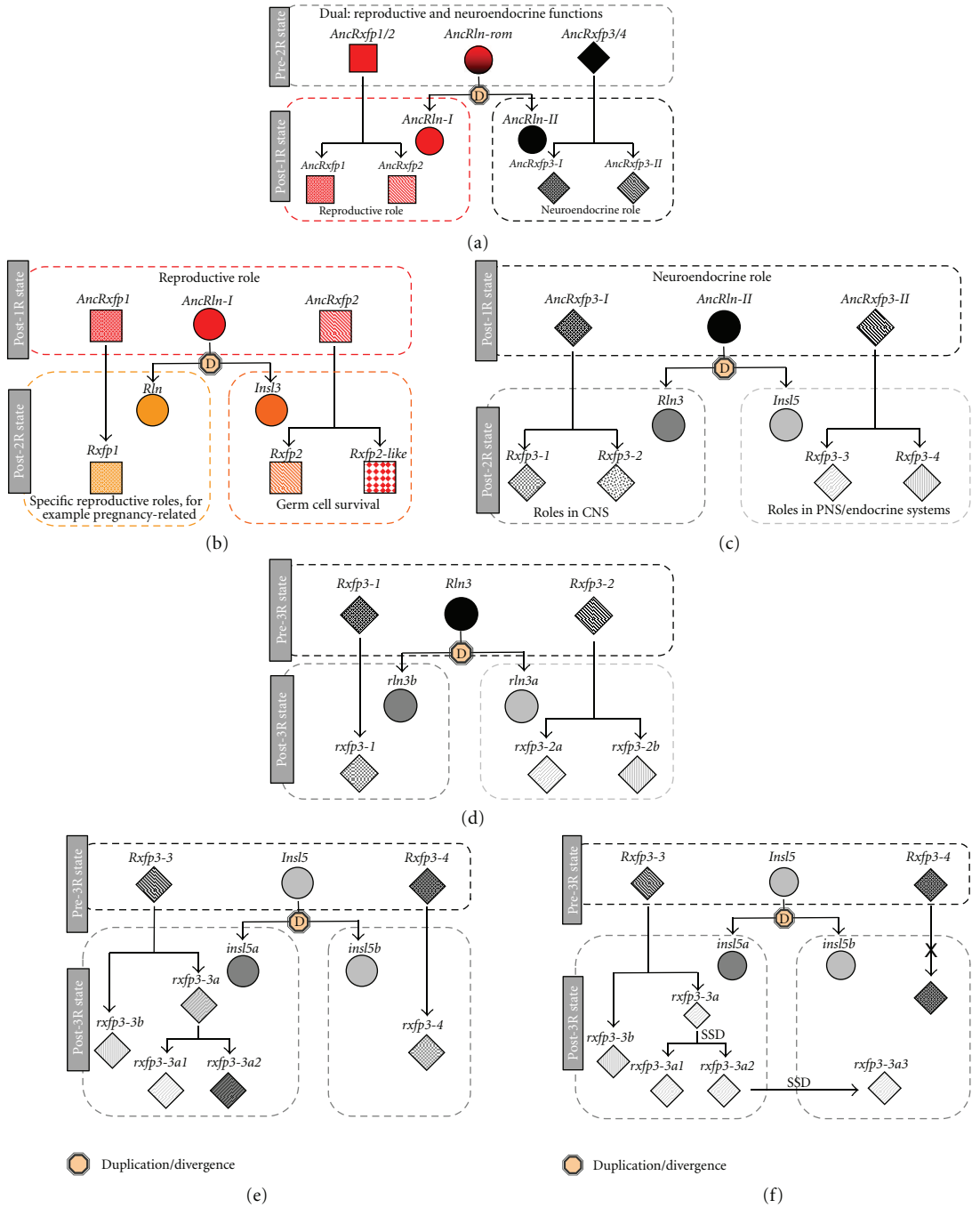


Figure 3: The hypothesized functional diversification of the *rln*/*insl* and *rxfp* genes in the gnathostome ancestors (a, b, and c) and in teleosts (d, e, and f). (a) The pre-1R three-gene system gave rise to two ligand genes and two pairs of receptor genes following 1R. After 1R, both ligands and receptors are structurally and functionally identical, which is favorable for promiscuous ligand-receptor interactions, in combination with selective pressures promoting a division of reproductive and neuroendocrine systems, leading to the establishment of novel ligand-receptor pairs. (b) Duplication and divergence of the *rln-rxfp1* and *insl3-rxfp2* ancestor genes. On the basis of the proposed relatedness of *rxfp2-like* to *rxfp2*, we hypothesize that Rxfp-like, at least immediately after 2R, functioned as a receptor for Insl3. (c) Duplication and divergence of the genes ancestral to *rln3* and *insl5* and their rxfp3/4-type receptor genes. Since all tetrapods lost *rxfp3-2* and most of them also lost *rxfp3-3*, their ligand-receptor pairs lost their ancestral three-component nature and became two-component, that is, Rln3-Rxfp3-1 and Insl5-Rxfp4. (d) Teleosts retained all after-2R rxfp3/4 receptor genes and seem to have experienced further subfunctionalization with the formation of complex ligand-receptor relationships. We hypothesize a functional specialization of the two *rln3* paralogs to work with *rxfp3-1* (*rln3a*) and two *rxfp3-2* genes. (e) Diversification of *rxfp3-3* and *rxfp3-4* genes in percomorpha (f) Zebrafish has lost its *rxfp3-4* (i.e., *rxfp4*) gene but has an extra copy of *rxfp3-3a3*, which may imply that the receptor of Insl5b is Rxfp3-3a3. Note that in (b) and (c) *insl5* paralogs are chosen arbitrarily and the interaction of the peptide with the receptors can be reversed; that is, Insl5a may function with Rxfp3-4 and Insl5b may interact with Rxfp3-3 receptors SSD = small scale duplication.

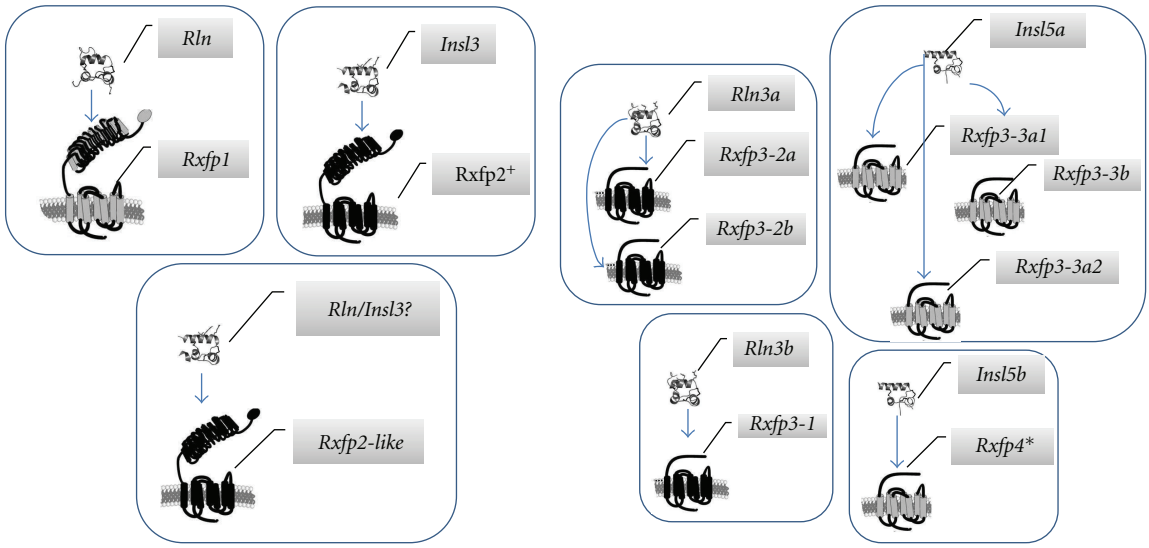


Figure 4: Ligand-receptor pairings of the Rln/Insl peptides and their Rxfp receptors putatively associated with (a) reproductive and (b) neuroendocrine processes in teleosts as hypothesized based on mammalian pairings and on their gene duplication history (see Figures 1 and 3). Few tetrapods (reptiles, frog, and opposum) and zebrafish have been found to possess the receptor *rxfp2-like*, which is phylogenetically more closely related to *rxfp2* than *rxfp1*, but still of ancient origin [2]. However, the lack of *insl3* in the reptiles, that also harbour *rxfp2-like* (data from [2]), suggests that Rln may be an alternate ligand. $^{+}$Zebrafish retained two 3R paralogs of *rxfp2*, *rxfp2a*, and *rxfp 2b*, while the remaining teleosts appear to have lost one copy. *In zebrafish, the *rxfp4* gene was lost and possibly replaced by *rxfp3-3a3* (see Figures 3 and 4). Images of receptors and peptides adopted with permission from the publisher for Halls et al. [5] and Kong et al. [6].

purifying selection and did not closely parallel the selection profile of either candidate ligands.

(2a) Evidence for Codon-Specific Positive Selection in Mammalian and Teleost Ligand and Receptor Genes. To look for evidence of codon-specific positive selection in mammalian and teleost lineages, we compared models 7 (purifying selection), 8 (positive selection), and 8a (relaxation of purifying selection) using maximum likelihood-based comparisons [22] in mammals and teleosts. Genes are considered to be under positive selection if the support for model 8 is greater than model 7, but also model 8a. For genes that exhibited evidence of positive selection, determination of the amino acid sites estimated to be under selection was tested using Bayesian Empirical Bayes (BEB). We found evidence of positive selection for mammalian *INSL5* and mammalian *RLN*; however, the hypothesis that the positive selection found in mammalian *INSL5* is actually caused by a relaxation of purifying selection (i.e., tested by comparing model 8a versus model 8) could not be rejected. The extent of positive selection on mammalian *RLN* is extensive; however, in total, 12 amino acid positions were identified as having a BEB probability > 0.9 that $\omega > 1.0$ (i.e., to be under positive selection) and another five had a probability > 0.8 that $\omega > 1.0$ (Table S3). This suggests the presence of strong diversifying selection on mammalian *RLN*. In teleosts, only *insl3* showed evidence of having codons subject to positive selection at two sites (Table S3).

There was some, but limited, evidence of positive selection on the receptor genes within mammalian or teleost lineages. Only one codon was found to exhibit strong evidence of positive selection in mammalian *RXFP1*, and two for *RXFP2*, while three codons showed evidence of positive selection in fish *rxfp2*, but the latter hypothesis was more likely attributed to a relaxation of purifying selection. Additionally, a few codons were found to have evidence of positive selection in mammalian *RXFP3* and teleost *rxfp4* (stronger evidence). Although mammalian *RXFP4* also showed evidence of positive selection (model 8 was preferred over models 7 and 8a); no specific codons had a BEB probability of being under strong positive selection. Overall, this suggests similar patterns of selection on ligand-receptor pairs, with the notable exception of *RLN-RXFP1* in mammals for which strong evidence of positive selection exists for the ligand, but no strong evidence of positive selection on the mammalian receptor gene, *RXFP1*.

(2b) Evidence for Differential Selection on Teleost Versus Mammalian Lineages for Orthologous Receptors. Although the above analyses suggested that only mammalian *RLN* has experienced high levels of codon-specific positive selection, using the branch-site model of codon-specific positive selection, we tested whether mammalian and teleost lineages have been subject to lineage-specific positive selection, that is whether they have been selected to be fixed for different amino acids (Table S5). This analysis revealed considerable evidence of lineage-specific selection indicating that mammalian and teleost lineages have evolved in different ways, and it also highlighted some important differences in the regions of the receptors that have been subject to

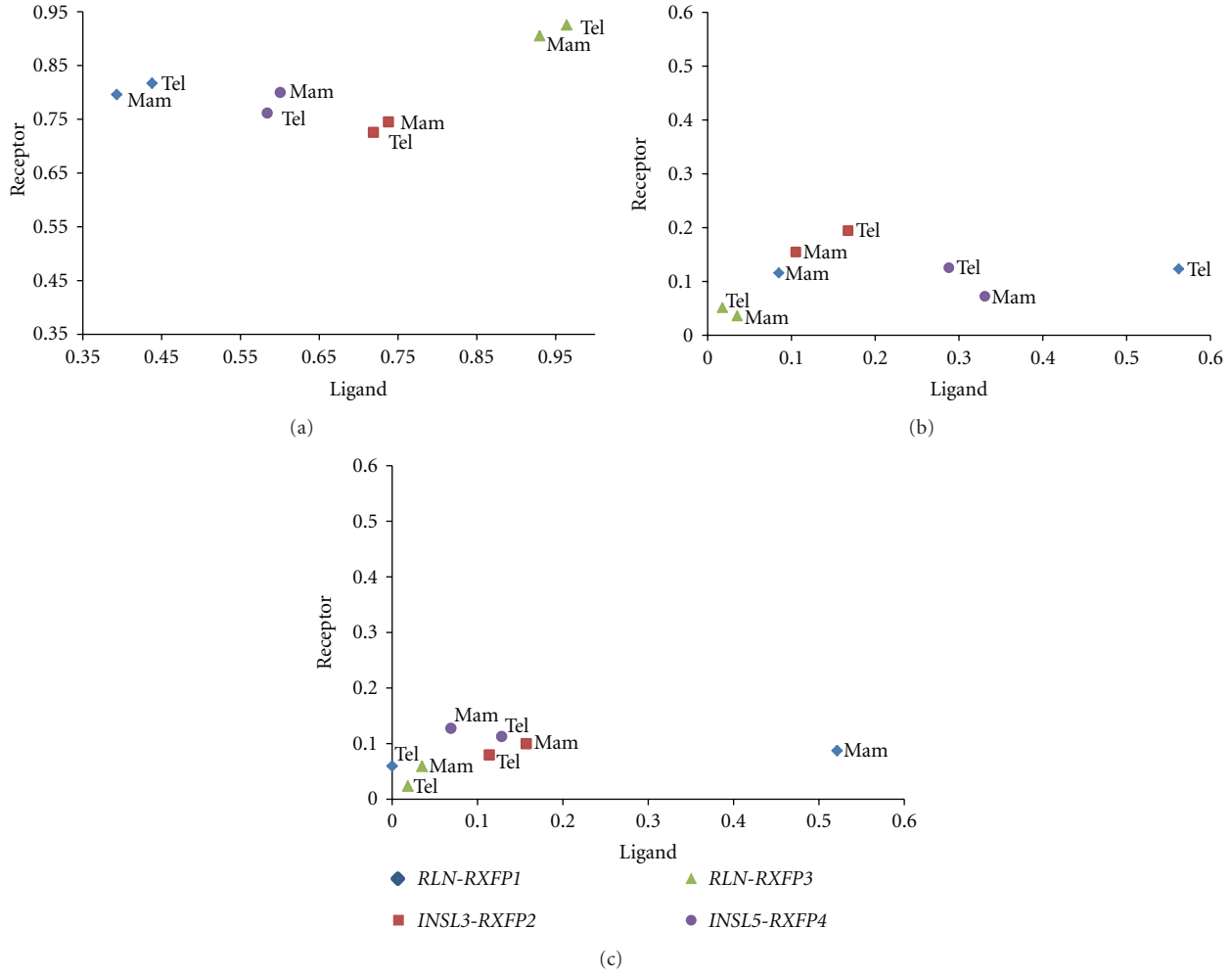


Figure 5: Estimated proportion of sites in the ligand (*x*-axis) and receptor (*y*-axis) genes evolving under (a) purifying, (b) neutral, and (c) positive selection in the genes of the putative ligand-receptor pairs of the RLN/INSL-RXFP system in mammals and teleosts.

positive selection. By mapping, codons were found to have evidence of positive selection to their position in the mature proteins; we find that (1) the low-density lipoprotein/leucine rich repeat (LDL/LRR) region of *RXFP1/2*-type genes is an important region of diversification among lineages; (2) for the 7 transmembrane (7TM) region shared between the two receptor types, all regions have more selected sites in *RXFP3/4*- than in *RXFP1/2*-type genes, except extracellular loop 2 (ECL2), and (3) intracellular loops 1 (ICL1) and 3 (ICL3) have many positively selected sites for *RXFP3/4* genes while ICL3 also has many amino acids selected for *RXFP1/2* type genes (Figure 7).

Closer examination of the sites that were selected in mammalian versus teleost lineages revealed somewhat different regions of selection in teleosts versus mammals. For *RXFP1*, mammals had more selection on the first few domains of the LDLa/LRR region, while teleosts exhibit greater selection on the terminal LRR domains. Additionally, in general mammalian, *RXFP1* genes were found to have more selected sites in the ICLs (ICL1 and ICL3), while teleosts exhibit more selection in the ECLs (ECL1 and ECL3) (Figure S1). This suggests that while the overall patterns of selection are similar among mammalian and teleost putative ligand-receptor orthologs, divergent selection has operated in both lineages for all genes, and some of this selection could be associated with intra- versus extracellular signaling (Figure S1).

Quantitative Expression of All Ligand and Receptor Genes in Zebrafish across Multiple Tissues. To infer functional ligand-receptor relationships, we assessed the expression of both ligand and receptor genes in male and female zebrafish heart, intestine, gonads, muscle, gills, brain, and eyes using real-time, quantitative PCR. Overall, the fold increase of the target to housekeeping genes, especially the receptors, was similar for both sexes in all tissues (except gonad) confirming the reliability of the data (Figures S2 and S3). To allow comparison of the relative amounts of mRNAs produced per tissue, the relative mRNA expression levels were normalized to the total amount of RNA isolated per tissue (Figure 8). This revealed that for all tissues studied, the expression levels

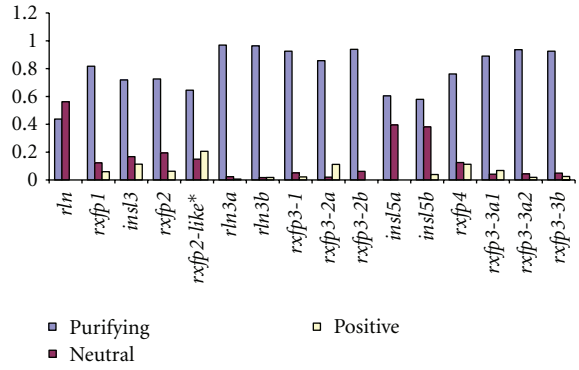


Figure 6: The proportion of sites in ligand and receptor genes subjected to different kinds of selection in teleosts. Selection types: purifying (light purple), neutral (dark purple), and positive (yellow). Hypothesized ligand-receptor pairs are placed in consecutive order. *Only present in zebrafish.

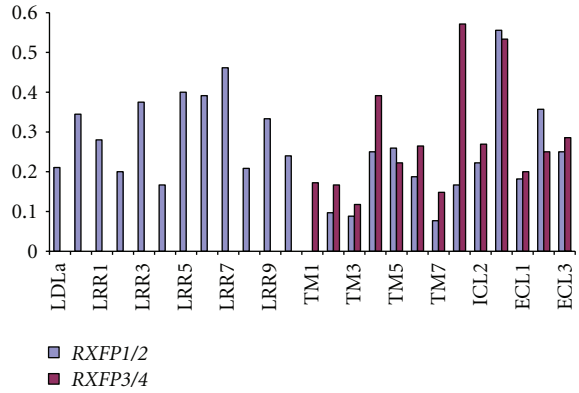


Figure 7: The proportion of amino acids selected per region in *RXFP1/2* and *RXFP3/4* receptor genes that showed evidence of positive selection in mammalian or teleost lineages according to the branch-site test of positive selection. LDLa—low-density lipoprotein module A, LRR—leucine rich repeat, TM—transmembrane domain, ICL—intracellular loop, and ECL—extracellular loop.

of all *rxfp* genes appeared to be higher than the expression levels of all *rln*/*insl* genes, except for the very high expression levels of *insl3* in testis tissue (Figure 8).

The ligand *rln* was most abundantly expressed in gonads and male intestine (Figures S2 and 8); its primary hypothesized receptor, *rxfp1*, was also highly expressed in gonads, as was a potential secondary candidate receptor, *rxfp2a* (Figure 8). The *rxfp1* transcript was also detected in male heart and brain, while *rxfp2b* expression was found in brain and eyes. Expression of the zebrafish-specific *rxfp2-like* transcript, a candidate receptor for Rln and Insl3, was only found in brain at high levels. Very high expression of *insl3* mRNA was found in testes and somewhat lower levels in ovaries and eyes. The primary candidate receptors for Insl3 are Rxfp2a and Rxfp2b, and high expression of both *rxfp2a* and *rxfp2b* was observed in gonads, while *rxfp2-like* was not detected in testes or ovaries. As expected, *rln3a* and *rln3b* expression was found predominantly in brain and gonad, but we also identified *rln3a* expression in heart (Figures S2 and 8). On the other hand, all of the *rxfp3-1*, *rxfp3-2*, and *rxfp3-3* genes showed a similar expression pattern: high expression in brain with lower levels in testes and eye, only *rxfp3-3a3* exhibited relatively low expression in brain. Relatively high levels of *insl5a* and *insl5b* mRNA were found in intestine, but additionally *insl5a* expression was found in gonads and brain. Our hypothesized candidate receptors for Insl5a are Rxfp3-3a1, Rxfp3-3a2, and Rxfp3-3b and for Insl5b is Rxfp3a3 (Figure 4): of the genes coding for these receptors, only *rxfp3-3b* showed high expression in the intestine (Figure 8).

3. Discussion

The main goal of this paper was to explore possible ligand-receptor pairings for the *rln*/*insl-rxfp* genes in teleosts. Based on previous bioinformatic analyses, we describe how teleosts preferentially retained 2R- and 3R-derived paralogs of genes putatively involved in neuroendocrine functions (*rln3*/*insl5-rxfp3/4*), ultimately leading to a greater number (10-11) of receptor genes than ligands (6). Given that the ligand-receptor pairings in teleosts are largely unknown, we employed selection and expression analyses to explore the possible ligand-receptor pairings. Overall, the selection analyses showed that (1) the extent of purifying, neutral, and positive selection acting on the four *RLN-RXFP* orthologs was highly similar between mammalian and teleost genes suggesting that, with the exception of mammalian *RLN*, ligands and receptors have the same binding relationships in both lineages and (2) the ligand-receptor pairs *RLN3-RXFP3* and *INSL3-RXFP2* exhibited highly similar selection profiles suggesting close coevolution, while the pair *INSL5-RXFP4* exhibited a more diffuse coevolution, and *RLN-RXFP1* exhibited much faster evolution of the ligand in mammals than in teleosts. The overall similarity between the genes in teleosts and mammals is supported by the observation that all of the teleost ligand genes exhibit predominant expression in the same tissues as their orthologs in mammals: rln and insl3—gonad, rln3—brain and insl5—intestine. However, even if the binding relationships are the same, it does not mean that the gene pairs have the same function in mammals and teleosts; indeed, the branch-site test of positive selection suggests that differentiation in function has occurred between the two groups. Secondly, although the binding relationships of the genes with orthologs in mammals and teleosts may be the same, it was difficult to resolve the ligand-receptor pairing relationships for the additional genes found in teleosts, but not in mammals.

3.1. The Highly Conserved Pair RLN3-RXFP3 Expanded through Gene Duplication and Possible Subfunctionalization in Teleosts. The *RLN3-RXFP3* system shows strong evidence of ligand-receptor coevolution with almost all amino acids being subject to purifying selection for both genes, and exhibiting a nearly perfect correlation in both mammals and teleosts. These findings are in accordance with previous studies and further support hypotheses about the highly

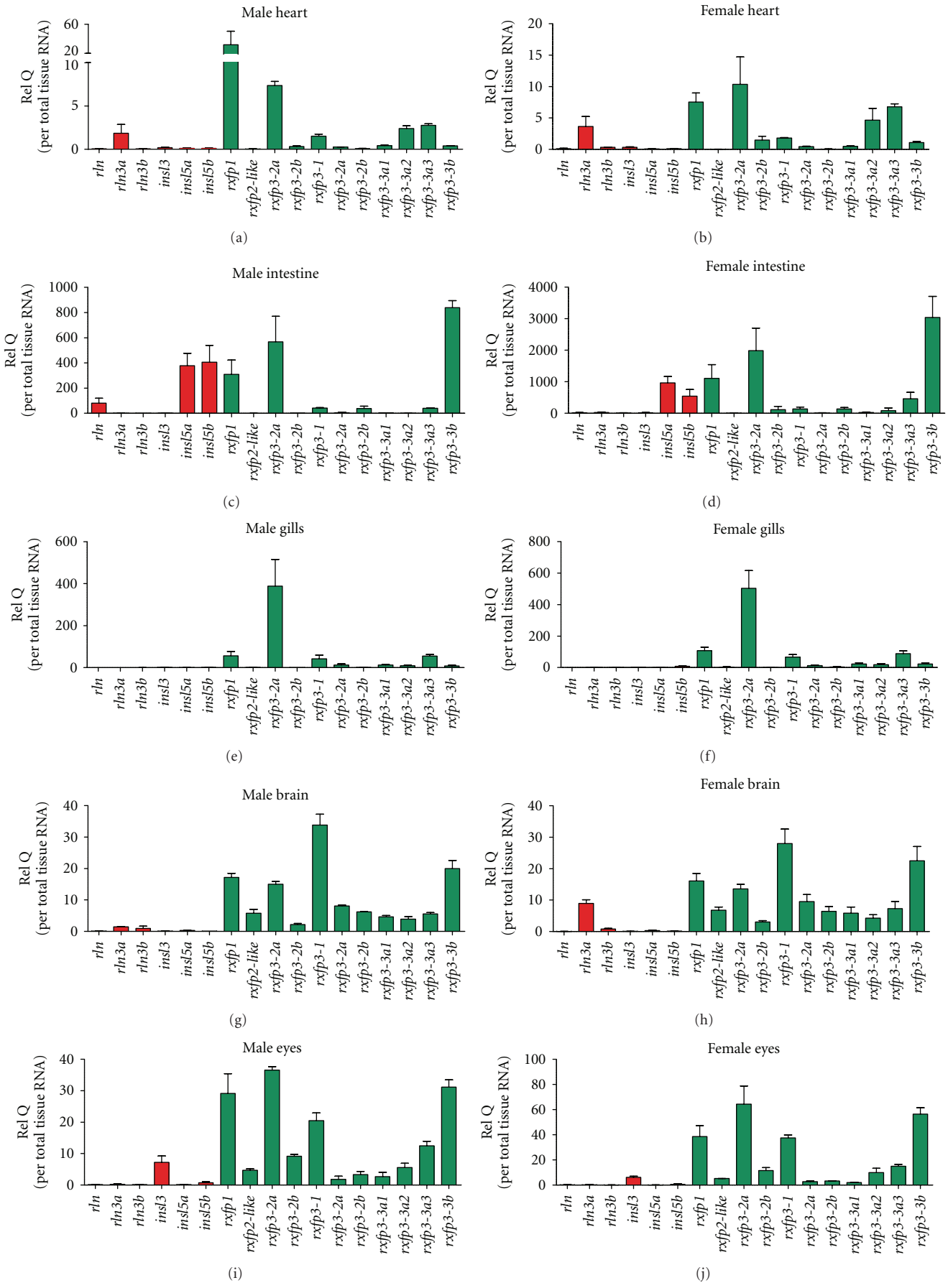


Figure 8: Continued.

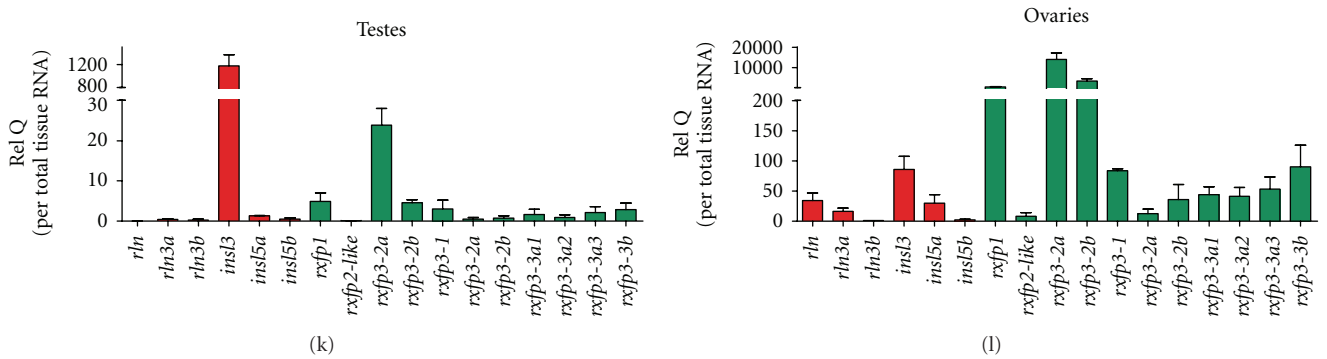


FIGURE 8: Relative expression of *rln/insl* and *rxfp* genes in zebrafish tissues. The expression of a gene relative to the average expression across all genes in a given tissue of males and females is shown. Red and green bars indicate the relative expression of the ligand and receptor genes, respectively. Three biological replicates were used to determine the standard errors on the relative expression.

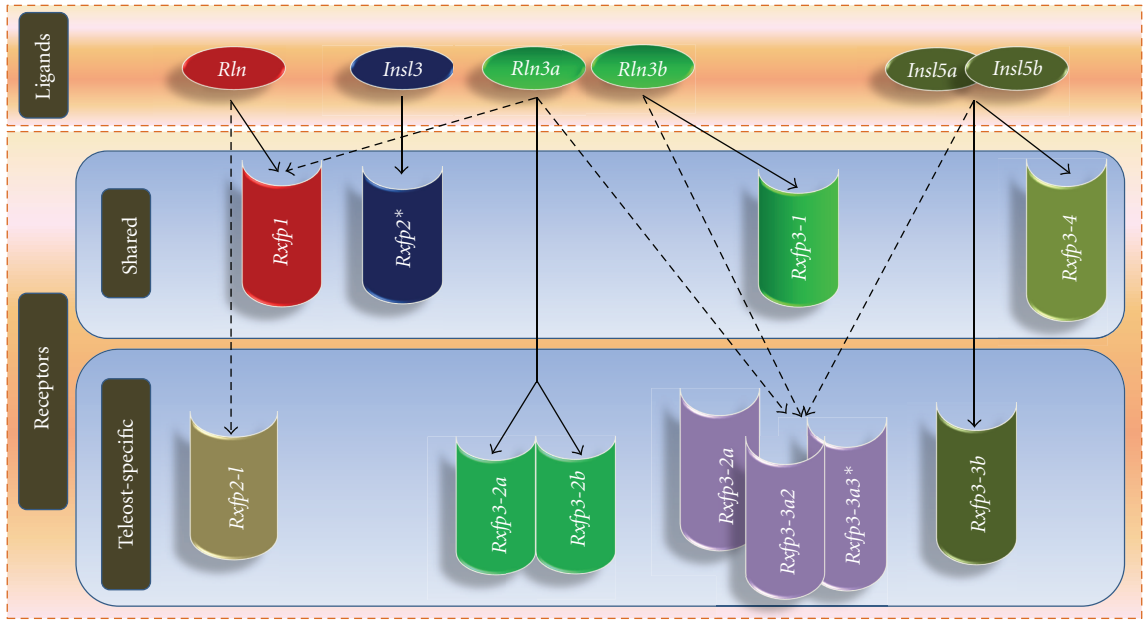


FIGURE 9: Proposed Rln/Insl-Rxfp ligand-receptor pairings based on previous genomic data (see Figure 4) and the analyses presented here. Solid lines represent potential ligand-receptor relationships that are well supported, while dashed lines represent uncertain pairings. There are four receptor orthologs between mammals and teleosts, and the data support the same ligand-receptor pairings for these genes in teleosts. For the seven teleost-specific receptors, strong ligand-receptor pairings were supported for three (solid lines), while the other three Rxfp3-3 receptors have unclear pairings (dotted lines), and Rxfp2-like is a probable secondary receptor for Rln.

conserved nature of the *RLN3-RXFP3* genes, and their probable parallel function across most vertebrates [23]. However, teleosts possess two 3R-derived *rln3* paralogs (*rln3a* and *rln3b*) and multiple *rxfp3*-type genes, not all of which are orthologous to mammalian *RXFP3*. Based on the duplication history of the genes [2], we proposed that the Rln3 peptide together with Rxfp3-1 and Rxfp3-2 receptors formed a tripartite ancestral teleost ligand-receptor signaling system (Figure 3), and hypothesized that the after-3R subfunctionalization of the *rln3* paralogs would be associated with subfunctionalization of the *rxfp3-1* and *rxfp3-2* receptor genes (Figures 3 and 4). Taking into account that in *Tetraodon nigroviridis* the loss of *rln3b* coincides with the pseudogenization of *rxfp3-1* (Figure 3), we further propose that Rln3b is a cognate ligand of Rxfp3-1, while Rln3a has specialized to function with two receptors, namely, Rxfp3-2a and Rxfp3-2b (Figure 9). This hypothesis is supported by experimental data presented here and elsewhere. For example, experimental studies performed in zebrafish [14] and eel [9] indicate that the expression of the *rln3* paralogs in fish shows strong homology to mammalian *RLN3*, where they are predominantly expressed in the periaqueductal grey, a region homologous to NI in mammals. Additionally, it is known that *rln3a* is expressed in a broader range of tissues (including gonad) than *rln3b*, indicating that *rln3a* and *rln3b* exhibit spatial (and perhaps temporal) subfunctionalization [8, 9, 14]. Our expression analyses indicate both coexpression of the *rln3* paralogs with *rxfp3-1* and *rxfp3-2* genes, and

also possible subfunctionalization of the receptor since all of the *rxfp3-1* and *rxfp3-2* (and even *rxfp3-3*) genes are highly expressed in brain, while *rxfp3-2a* and *rxfp3-2b* are additionally expressed in the ovary, but at lower levels, mimicking the expression pattern of its candidate ligand, *rln3a*.

3.2. Receptors for the Insl5 Paralogs in Teleosts Are Difficult to Resolve. Resolving the ligand-receptor pairings for the Insl5-Rxfp4 system in teleosts is more difficult. We hypothesized that the Rxfp3-3 and Rxfp3-4 descendents (Figure 1) are the potential receptors for Insl5a and Insl5b (Figure 2, supplementary Figure S3). Specifically, we hypothesized that, in teleosts, Rxfp3-3a1, Rxfp3-3a2, and Rxfp3-3b are candidate receptors for Insl5a while Rxfp3-4 (aka Rxfp4) is the receptor for Insl5b; in zebrafish, the loss of *rxfp3-4* was compensated by the gain of *rxfp3-3a3* (Figure 2), and the latter could serve as the receptor for Insl5b (Figures 3 and 4). Despite this prediction, the selection and expression data provided little evidence for which receptors may bind to the two teleost *insl5* paralogs (Figure 9). The selection profile of teleost *rxfp4* is the best match for that of both *insl5a* and *insl5b*, but all three *rxfp3-3*-type receptors are dominated by purifying selection and have selection profiles similar to those of *rln3*. On the other hand, the experimental data in zebrafish (which lacks *rxfp4*) indicate that *insl5a* is expressed in intestine and gonads and *insl5b* is expressed predominantly in intestine, and both paralogs exhibit low but significant expression in brain. This is consistent with the pattern in mammals, but the only receptor expressed at high levels in intestine was *rxfp3-3b*. The failure to find stronger evidence of coexpression of additional receptors for the Insl5 paralogs may be caused, in part, by the endocrine action of Insl5 and its expression in peripheral tissues [18, 24], many of which were not examined here, or possibly by developmental regulation of one or both of the *insl5* paralogs. Three of the other Rxfp3-3 receptor genes, *rxfp3-3a1*, *rxfp3-3a2*, and *rxfp3-3a3*, were all additionally expressed in brain and male gonads, therefore if Insl5a is a ligand for these receptors, teleosts may have expanded and subfunctionalized the role of the Insl5 peptides involved in the HPG axis. Further experimental work, including *in situ* hybridization, should be performed on *insl5* and *rxfp3-3* receptors in teleosts to thoroughly assess this hypothesis. Furthermore, the coexpression of *insl5*- and *rxfp3/4*-type genes in a teleost species other than zebrafish should be performed since zebrafish possesses a slightly unique suite of genes (Table S2), which did not allow for qPCR analyses of *rxfp4*.

3.3. The INSL3-RXFP2 System Exhibits Similar Expression Patterns in Mammals and Zebrafish. While teleosts exhibit a clear expansion of the *rln/insl* and *rxfp* genes involved in neuroendocrine pathways, the 3R duplicates of *rln* and *insl3* and their corresponding *rxfp1/2*-type receptors expanded minimally. We find good support for the hypothesis that Insl3-Rxfp2 are ligand-receptor pairs in teleosts: their selection profiles are highly similar and, in zebrafish, which contain two *rxfp2* paralogs (*rxfp2a* and *rxfp2b*), both receptor genes are highly expressed in gonads, although *rxfp2b* is additionally quite highly expressed in brain. Previously, it was shown that *insl3* expression in zebrafish shows strong parallels to that in mammals: *in situ* and qPCR analyses on male gonads reveal that it is expressed predominantly in Leydig cells [8], and the more thorough qPCR analyses presented here further demonstrate that it is very abundantly expressed in male gonads, but also in female ovaries. Current *in situ* analysis (underway in our laboratory) has also revealed the specificity of *rxfp2a* and *rxfp2b* expression in Leydig cells (unpublished data). On the other hand, although *rxfp2-like* (which among teleosts is only present in zebrafish) has a similar selection profile to *insl3*, we found it to be predominantly expressed in brain, rendering interpretation difficulty, and we favor the hypothesis that Rxfp2-like is an alternate receptor for Rln (see Figures 4 and 9).

3.4. RLN-RXFP1 System in Placental Mammals and Teleosts: Conserved Receptor but Rapidly Evolving Ligand in Mammals. The only ligand-receptor pair for which there was a poor correlation in the nature of selection was RLN-RXFP1 in mammals. While *RXFP1* genes in mammals and teleosts have evolved in similar ways, the gene coding for the hormone relaxin, *rln,* has been subject to purifying and neutral evolution in teleosts, but has been the target of strong positive selection in mammals (see Figure 5(c), Table S3). In accordance with two recent studies showing the strong role of selection on the relaxin locus [25, 26], we find that approximately 50% of the codons in mammalian *RLN* show evidence of positive selection, whereas no sites in teleost *rln* do. Additionally, the qPCR expression pattern of *rxfp1* in zebrafish shows broad but low levels of expression across multiple tissues, including gonad and brain. Using RT-PCR and *in situ* analyses in zebrafish, Donizetti et al. [27] showed that expression of *rxfp1* in zebrafish brain begins early in development and shows strong overlap with that of *RXFP1* in humans. Based on the similar amino acid sequence of Rln and Rln3 in teleosts, they propose that Rxfp1 could be an additional receptor for Rln3a and/or Rln3b in teleosts. A study comparing the expression of *rln3a, rln3b*, and *rln* in eel using *in situ* and qPCR analyses [9] found that the expression of teleost *rln* is similar to that of *rln3*, but with lower expression in brain and higher in gonads, similar to that observed in which expression was predominantly found in gonad. This pattern is supported by our hypothesis for the evolution of the system in which the ancestral ligand molecule is hypothesized to have functioned in both reproductive and neuroendocrine pathways (Figure 3).

3.5. Evidence for Differential Selection in Teleost Versus Mammalian rln/insl-rxfp Genes Suggests Functional Divergence of the Ligand-Receptor Coding Sequences. Although we have focused on the similarities in the evolution of mammalian and teleost *RLN*/*INSL-RXFP* genes, the analysis of codon-specific positive selection revealed that mammalian and teleost genes have been subject to differential selection and that some receptor domains are the targets of more selection

than others. For this analysis, sites were deemed to be subject to codon-specific selection if, when comparing a particular branch of the phylogenetic tree, there was evidence that certain amino acids were selected to be different from those in the "background" lineage for the same gene. By analyzing the genes in this way, we found that for the *RXFP1/2*-type genes, the LDLa-LRR region generally showed high levels of selection, not surprisingly, since they are involved in receptor-ligand signaling [5]. Functional studies have shown that the LRR region is important for the binding of the cognate ligand; the LDLa module is essential for cAMP accumulation which takes place after the ligand is recognized and bound [5]. Apart from these regions, the only other two regions which were identified as having more than 20% of the sites subject to selection for *RXFP1/2* genes were ICL3 and ECL2.

In general, lineage-specific selection was higher for the *RXFP3/4*-type genes: all domains were found to have more than 20% of the amino acids subject to positive selection except for four regions of the transmembrane domain (TM1, TM2, TM3, andTM7) and ECL1. Of particular interest is the fact that for the *RXFP3/4*-type genes, ICL1 is equally important as ICL3 in terms of selection. The finding that ICL3 (both receptor types) and ICL1 (RXFP3/4-type receptors) are targets of selection suggests that a major component of selection for the RXFP receptors concerns downstream receptor signaling rather than selection for ligand binding *per se*.

4. Conclusions

Although the majority of the relaxin family genes originated prior to the divergence of osteichthyans, the fate of the family in teleosts and mammals is markedly different owing to the differential retention and diversification of genes in each lineage. Earlier studies suggested that teleosts only possessed *relaxin 3*- and *rxfp3*-like genes and proposed that RLN and INSL3 were neurohormones that recruited their RXFP1/2-type receptors after the divergence of mammals [28], a view that is inconsistent with the data presented here and elsewhere [2, 8, 29, 30]. The goal of this study was to establish a theoretical background for further experimental work on the *rln/insl-rxpf* systems in teleosts. Although the study was limited because its methodology relied on the known ligand-receptor pairings and expression data from mammals as a reference, our analyses suggest that the orthologs of the four 2R-derived ligand genes (*RLN*, *INSL3*, *RLN3*, and *INSL5*) have similar ligand-receptor pairings in teleosts and mammals (with the exception of the unusual situation with Rln-Rxfp1). Despite these similar patterns, there is also evidence of differential selection on specific amino acids in mammalian versus teleost lineages, suggesting functional divergence in the two lineages.

It is interesting that the RLN/INSL peptides diversified their reproductive functions in mammals, owing to local duplications at the relaxin locus [23, 25, 26, 29], while teleosts underwent a massive diversification of the genes believed to be involved in neuroendocrine regulation (*rln3*/*insl5-rxpf3/4*). Overall, we find evidence that many of these "additional" receptor genes in teleosts have characteristics of the RLN3-RXFP3 system, that is, slow evolution and predominant expression in the brain, while the primary receptors for the two Insl5 paralogs in teleosts remain obscure. Nevertheless, we find that teleosts greatly expanded and probably subfunctionalized the role of the rxfp3-2- and rxfp3-3-derived receptors; their cognate ligands and their physiological functions should be the focus of future experimental work.

5. Materials and Methods

5.1. Selection Profiles of Candidate Ligand-Receptor Pairs. We obtained sequences and performed an alignment based on the coding sequence for the RLN/INSL-RXFP genes from 5 teleosts (zebrafish, medaka, fugu, tetraodon, and stickleback) and 11 placental mammals (human, rhesus, cow, pig, horse, dog, guinea pig, mouse, rat, rabbit, and elephant) as described previously [2]. The accession numbers of all genes are listed in Tables S4 and S7 in Yegorov and Good [2], and the alignment is available upon request.

We calculated the proportion of codons in ligand and receptor pairs estimated to be subject to purifying, neutral, or positive selection using the sites model in PAML [22]. Next, to assess whether teleost ligand or receptor genes have been subject to adaptive divergent selection, we used several methods that examine the ratio of nonsynonymous to synonymous (d_N/d_S) substitutions. Because d_S provides an approximation of the neutral rate of substitution, $\omega = d_N/d_S$ ratios are used to determine selection pressure on genes or codon positions, with $\omega > 1$ indicative of positive Darwinian selection [31].

Site Models. We employed models that allow ω to vary among sites and tested a series of models to look for evidence of positive selection. First, we compared model M7 (beta) versus M8 (beta + ω) to test for evidence of positive selection and then compared model 8 versus model 8a to assess whether the evidence for positive selection was actually caused by a relaxation of purifying selection (or true positive selection); for both comparisons we used the site model tests in PAML [32]. Likelihood ratio tests (LRTs) were constructed to compare model M7 versus M8 and M8a versus M8. Twice the log likelihood difference between models was compared with a chi-square distribution with number of degrees of freedom (df) calculated as the difference in the number of estimated parameters between models. Model M8 was additionally used to identify codon sites under positive selection using a Bayes Empirical Bayes (BEB) criterion.

Branch-Site Models. We hypothesized that at least some of the receptor genes may have experienced lineage-specific positive selection in mammals versus teleosts. To examine this we used the branch-site model A of Zhang et al. [22], which tests whether the members of a user-defined clade (branch) on a phylogenetic tree exhibit evidence of codon-specific selection relative to the remaining (background)

lineages. Tests of positive selection were made by comparing the branch-site model A in which $(d_N/d_S) > 1$ (alternative hypothesis) to the model A in which $d_N/d_S = 1$ fixed (null hypothesis) and by setting the foreground branch to the base of the clade containing the relaxin family ortholog in teleosts and the background to the same ortholog in mammals or tetrapods (depending on the tree structure) or vice versa. Analysis of the branch-site model A was done using CODEML from the PAML package (PAML v. 4.2); models were compared using the likelihood ratio test with 1 degree of freedom and, where significant, the posterior probability that a codon was under positive selection was estimated using the Bayes Empirical Bayes (BEB) procedure [22].

5.2. Quantitative Expression Analysis in Zebrafish Tissues

Animals. Sexually mature male and female zebrafish (*Danio rerio*) from the Tübingen AB strain were used. Animal housing [33] and experimentation were consistent with Dutch national regulations and were approved by the Utrecht University Animal Use and Care Committee.

RNA Isolation and cDNA Synthesis. Various tissues (heart, intestine, testis, ovary, muscle, gill, brain, and eye) were dissected from male and female adult zebrafish and immediately flash frozen in liquid nitrogen.Tissue samples from 3 individual zebrafish, for each gender, were combined for each replicate and the RNA was isolated using the FastRNA Pro Green kit (Bio 101 Systems), according to the manufacturer's recommendations. Three independent RNA isolations (biological replicates), each containing pooled tissues from 3 individual fish, were performed for each tissue per sex. Possible genomic DNA contamination was removed from each total RNA fraction with the RNAse-free DNase Treatment & Removal kit (Ambion), which includes a final step to remove the DNAse I from the reaction. Next, cDNA synthesis was performed with 2 μg of each total RNA samples, as described previously [34].

Real-Time, Quantitative PCR. Primers (Table S6) for real-time, quantitative PCR (qPCR) to detect zebrafish *rln/insl* and *rxfp* mRNAs were designed and validated for specificity and amplification efficiency on serial dilutions of testis cDNA [35] using SYBR Green-based assays (Applied Biosystems, Foster City, CA, USA). All primers were designed on different exons, except for the primers detecting the *rxfp3* cDNAs, since all *rxfp3* genes are single-exon genes. Moreover, each qPCR run was followed by a melt curve analyses to exclude potential PCR amplifications from genomic DNA contamination. To normalize the data, a TaqMan Gene Expression Assay was acquired to detect the endogenous control RNA, eukaryotic *18S ribosomal* RNA (Applied Biosystems). To examine the relative expression of genes across tissues, the relative fold change of the genes of interest was normalized to the *18S ribosomal RNA* reference gene and to a calibrator (calculated as the mean expression of all genes) (supplementary Figures S2 (ligands) and S3 (receptors). All qPCRs and calculations (using the $\Delta\Delta C_T$ method) were performed as described previously [35–37]. To compare the expression levels of all relaxin family peptide and receptor genes in *whole* zebrafish tissues, expression levels were additionally corrected for the total RNA yield per tissue per sex (Figure 8).

Appendix

See Supplementary Tables S1, S2, S3, S4, S5, and S6 and Figures S1, S2, and S3 (see Supplementary materials available online at doi:10.1155/2012/310278).

Acknowledgments

The authors thank Murray Wiegand and Kevin Campbell for comments on an earlier version of this work. The authors also thank two anonymous reviewers whose comments improved this paper. Parts of this research were included in the M.S. theses of S. Yegorov and J. Martijn. This research was funded by a discovery grant from the National Science and Engineering Research Council (NSERC) to S. Good and by a Manitoba Graduate Scholarship and University of Winnipeg Graduate Award to S. Yegorov.

References

[1] O. D. Sherwood, "Relaxin's physiological roles and other diverse actions," *Endocrine Reviews*, vol. 25, no. 2, pp. 205–234, 2004.

[2] S. Yegorov and S. Good, "Using paleogenomics to study the evolution of gene families: origin and duplication history of the relaxin family hormones and their receptors," *PLoS ONE*, vol. 7, no. 3, Article ID e32923, 2012.

[3] M. L. Halls, R. A. D. Bathgate, and R. J. Summers, "Relaxin family peptide receptors RXFP1 and RXFP2 modulate cAMP signaling by distinct mechanisms," *Molecular Pharmacology*, vol. 70, no. 1, pp. 214–226, 2006.

[4] M. Kinoshita, K. Murata, K. Naruse et al., *Medaka: Biology, Management, and Experimental 1 Protocols*, Wiley-Blackwell, Iowa City, IA, Iowa, USA, 2009.

[5] M. L. Halls, E. T. Van Der Westhuizen, R. A. D. Bathgate, and R. J. Summers, "Relaxin family peptide receptors - Former orphans reunite with their parent ligands to activate multiple signalling pathways," *British Journal of Pharmacology*, vol. 150, no. 6, pp. 677–691, 2007.

[6] R. C. K. Kong, P. J. Shilling, D. K. Lobb, P. R. Gooley, and R. A. D. Bathgate, "Membrane receptors: structure and function of the relaxin family peptide receptors," *Molecular and Cellular Endocrinology*, vol. 320, no. 1-2, pp. 1–15, 2010.

[7] S. Y. Hsu, K. Nakabayashi, S. Nishi et al., "Activation of orphan receptors by the hormone relaxin," *Science*, vol. 295, no. 5555, pp. 671–674, 2002.

[8] S. V. Good-Avila, S. Yegorov, S. Harron et al., "Relaxin gene family in teleosts: phylogeny, syntenic mapping, selective constraint, and expression analysis," *BMC Evolutionary Biology*, vol. 9, no. 1, article 293, 2009.

[9] G. B. Hu, M. Kusakabe, and Y. Takei, "Localization of diversified relaxin gene transcripts in the brain of eels," *General and Comparative Endocrinology*, vol. 172, no. 3, pp. 430–439, 2011.

[10] K. Kawamura, J. Kumagai, S. Sudo et al., "Paracrine regulation of mammalian oocyte maturation and male germ cell survival," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 19, pp. 7323–7328, 2004.

[11] S. L. Dun, E. Brailoiu, Y. Wang et al., "Insulin-like peptide 5: expression in the mouse brain and mobilization of calcium," *Endocrinology*, vol. 147, no. 7, pp. 3243–3248, 2006.

[12] B. M. McGowan, S. A. Stanley, J. Donovan et al., "Relaxin-3 stimulates the hypothalamic-pituitary-gonadal axis," *American Journal of Physiology*, vol. 295, no. 2, pp. E278–E286, 2008.

[13] C. M. Smith, P. J. Shen, A. Banerjee et al., "Distribution of relaxin-3 and RXFP3 within arousal, stress, affective, and cognitive circuits of mouse brain," *Journal of Comparative Neurology*, vol. 518, no. 19, pp. 4016–4045, 2010.

[14] A. Donizetti, M. Fiengo, S. Minucci, and F. Aniello, "Duplicated zebrafish relaxin-3 gene shows a different expression pattern from that of the co-orthologue gene," *Development Growth and Differentiation*, vol. 51, no. 8, pp. 715–722, 2009.

[15] Y. Watanabe, Y. Miyamoto, T. Matsuda, and M. Tanaka, "Relaxin-3/INSL7 regulates the stress-response system in the rat hypothalamus," *Journal of Molecular Neuroscience*, vol. 43, no. 2, pp. 169–174, 2011.

[16] D. Conklin, C. E. Lofton-Day, B. A. Haldeman et al., "Identification of INSL5, a new member of the insulin superfamily," *Genomics*, vol. 60, no. 1, pp. 50–56, 1999.

[17] S. Y. Hsu, "Cloning of two novel mammalian paralogs of relaxin/insulin family proteins and their expression in testis and kidney," *Molecular Endocrinology*, vol. 13, no. 12, pp. 2163–2174, 1999.

[18] C. Liu, C. Kuei, S. Sutton et al., "INSL5 is a high affinity specific agonist for GPCR142 (GPR100)," *Journal of Biological Chemistry*, vol. 280, no. 1, pp. 292–300, 2005.

[19] M. L. Halls, C. P. Bond, S. Sudo et al., "Multiple binding sites revealed by interaction of relaxin family peptides with native and chimeric relaxin family peptide receptors 1 and 2 (LGR7 and LGR8)," *Journal of Pharmacology and Experimental Therapeutics*, vol. 313, no. 2, pp. 677–687, 2005.

[20] D. Cyranoski, "Two by two," *Nature*, vol. 458, no. 7240, pp. 826–829, 2009.

[21] A. B. Prasad, M. W. Allard, and E. D. Green, "Confirming the phylogeny of mammals by use of large comparative sequence data sets," *Molecular Biology and Evolution*, vol. 25, no. 9, pp. 1795–1808, 2008.

[22] J. Zhang, R. Nielsen, and Z. Yang, "Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level," *Molecular Biology and Evolution*, vol. 22, no. 12, pp. 2472–2479, 2005.

[23] T. N. Wilkinson, T. P. Speed, G. W. Tregear et al., "Evolution of the relaxin-like peptide family," *BMC Evolutionary Biology*, vol. 5, article 14, 2005.

[24] C. Liu and T. W. Lovenberg, "Relaxin-3, INSL5, and their receptors," *Results and Problems in Cell Differentiation*, vol. 46, pp. 213–237, 2008.

[25] J. I. Arroyo, F. G. Hoffman, and J. C. Opazo, "Gene duplication and positive selection explains unusual physiological roles of the relaxin gene in the European rabbit," *Journal of Molecular Evolution*, vol. 74, no. 1-2, pp. 52–60, 2012.

[26] J. I. Arroyo, F. G. Hoffmann, and J. C. Opazo, "Gene turnover and differential retention in the relaxin/insulin-like gene family in primates," *Molecular Phylogenetics and Evolution*, vol. 63, no. 3, pp. 768–776, 2012.

[27] A. Donizetti, M. Fiengo, R. Del Gaudio, R. Di Giaimo, S. Minucci, and F. Aniello, "Characterization and developmental expression pattern of the relaxin receptor rxfp1 gene in zebrafish," *Development Growth and Differentiation*, vol. 52, no. 9, pp. 799–806, 2010.

[28] R. Ivell, M. Kotula-Balak, D. Glynn, K. Heng, and R. Anand-Ivell, "Relaxin family peptides in the male reproductive system-a critical appraisal," *Molecular Human Reproduction*, vol. 17, no. 2, pp. 71–84, 2011.

[29] F. G. Hoffmann and J. C. Opazo, "Evolution of the relaxin/insulin-like gene family in placental mammals: implications for its early evolution," *Journal of Molecular Evolution*, vol. 72, no. 1, pp. 72–79, 2011.

[30] J. I. Park, J. Semyonov, L. C. Chia, W. Yi, W. Warren, and S. Y. T. Hsu, "Origin of INSL3-mediated testicular descent in therian mammals," *Genome Research*, vol. 18, no. 6, pp. 974–985, 2008.

[31] Z. Yang and J. P. Bielawski, "Statistical methods for detecting molecular adaptation," *Trends in Ecology & Evolution*, vol. 15, no. 12, pp. 496–503, 2000.

[32] R. Nielsen and Z. Yang, "Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene," *Genetics*, vol. 148, no. 3, pp. 929–936, 1998.

[33] M. Westerfield, *The Zebrafish Book: A Guide for the Laboratory Use of Zebrafish (Danio rerio)*, University of Oregon Press, Eugene, Ore, USA, 2000.

[34] F. E. M. Rebers, C. P. Tensen, R. W. Schulz, H. J. T. Goos, and J. Bogerd, "Modulation of glycoprotein hormone α- and gonadotropin IIß-subunit mRNA levels in the pituitary gland of mature male African catfish, *Clarias gariepinus*," *Fish Physiology and Biochemistry*, vol. 17, no. 1–6, pp. 99–108, 1997.

[35] J. Bogerd, M. Blomenröhr, E. Andersson et al., "Discrepancy between molecular structure and ligand selectivity of a testicular follicle-stimulating hormone receptor of the African catfish (*Clarias gariepinus*)," *Biology of Reproduction*, vol. 64, no. 6, pp. 1633–1643.

[36] P. P. de Waal, D. S. Wang, W. A. Nijenhuis, R. W. Schulz, and J. Bogerd, "Functional characterization and expression analysis of the androgen receptor in zebrafish (*Danio rerio*) testis," *Reproduction*, vol. 136, no. 2, pp. 225–234, 2008.

[37] Á. García-López, J. Bogerd, J. C. M. Granneman et al., "Leydig cells express follicle-stimulating hormone receptors in African catfish," *Endocrinology*, vol. 150, no. 1, pp. 357–365, 2009.

Hindawi Publishing Corporation
International Journal of Evolutionary Biology
Volume 2012, Article ID 341932, 24 pages
doi:10.1155/2012/341932

Review Article

In with the Old, in with the New: The Promiscuity of the Duplication Process Engenders Diverse Pathways for Novel Gene Creation

Vaishali Katju

Department of Biology, University of New Mexico, Albuquerque, NM 87131, USA

Correspondence should be addressed to Vaishali Katju, vkatju@unm.edu

Received 18 May 2012; Accepted 3 June 2012

Academic Editor: Frédéric Brunet

The gene duplication process has exhibited far greater promiscuity in the creation of paralogs with novel exon-intron structures than anticipated even by Ohno. In this paper I explore the history of the field, from the neo-Darwinian synthesis through Ohno's formulation of the canonical model for the evolution of gene duplicates and culminating in the present genomic era. I delineate the major tenets of Ohno's model and discuss its failure to encapsulate the full complexity of the duplication process as revealed in the era of genomics. I discuss the diverse classes of paralogs originating from both DNA- and RNA-mediated duplication events and their evolutionary potential for assuming radically altered functions, as well as the degree to which they can function unconstrained from the pressure of gene conversion. Lastly, I explore theoretical population-genetic considerations of how the effective population size (N_e) of a species may influence the probability of emergence of genes with radically altered functions.

1. Introduction

A recognition of the significance of novel traits for the origin of biological complexity and diversity is not new. Darwin himself spelled out the link between the evolution of novel traits and origin of new life forms, despite an agonizing lack of awareness of the genetic nature of variation and heredity. Armed with knowledge of the molecular basis of biological inheritance stemming from the rediscovery of Mendelian genetics and the first cytological glimpses of duplication events [1, 2], neo-Darwinists and early geneticists were swift to recognize the evolutionary potential of gene duplication as a means of exapting ancestral genes for novel functions [3–5]. The evolutionary advantage of gene duplication was appreciated well in advance of the discovery of DNA and was well surmised by Huxley in 1942—"... small repeats of this type ... constituted the chief method by which the number of genes is increased, thus providing duplicate factors and the opportunity for slight divergent specialization of homologous genes, giving great delicacy of adjustment" [6]. Gene duplication research in the 1940s through 1950s was decidedly cytological in flavour, often employing mutagenic treatments to accelerate mutation rates for the purpose of identifying the frequency, chromosomal location, and breakpoints of duplications and other structural variants. ([7–12], among others). Commencing in the early 1960s, experimental studies of gene duplication took on more of an evolutionary perspective, with greater efforts being directed at the nature of molecular evolutionary change due to alterations of the base composition, and the identification of different types of duplications leading to novel genes with radically altered reading frames [13–15]. Notably, a few studies specifically identified the creation of *partial* gene duplicates by incomplete duplication of the progenitor copy's open reading frame as in the case of human haptoglobins [13], human hemoglobins [14], and protamines in the Pacific herring *Clupea pallasi* [15], among others. Indeed, in their article, Smithies et al. [13] succinctly detailed the evolutionary potential of such radically altered gene duplicates—"We suggest that proteins with radically changed properties can be formed as a consequence of the single genetic event of a chromosomal rearrangement involving non-integral numbers of genes. Chromosomal rearrangements of this type appear to provide a mechanism for achieving more rapid

and extensive changes in protein structure in evolution than are possible by point mutations even when preceded by gene duplication." However, a true recognition of the role of gene duplication in the creation of radically altered structures would not be forthcoming until the advent of the genomic revolution.

Susumu Ohno is largely credited with formalizing and instigating the study of gene duplication into the burgeoning field it is today with the publication of his treatise titled *Evolution by Gene Duplication* [16]. In his book, Ohno hypothesized that the vertebrate lineage had undergone two rounds of whole-genome duplication; variations of his idea are now collectively referred to as the "two rounds" (2R) hypothesis (e.g., [17–19]). Although modest in size and somewhat simplistic and narrow in its depiction of the plausible pathways of gene duplication, *Evolution by Gene Duplication* has certainly earned its keep as the first book entirely devoted to the subject of gene and genome duplication. It also provided the first theoretical framework for the evolution of novel gene function by one copy following gene duplication. Ohno postulated that single-copy genes with essential functions are actively policed by purifying natural selection that serves to eliminate newly-acquired "forbidden" mutations that may compromise the ancestral gene function. This active removal of new mutations by single-copy genes in turn precludes them from exploring new evolutionary space (and gain of novel functions). The gene duplication process, by creating a redundant locus, simultaneously (i) permits the uninterrupted maintenance of the ancestral function by one copy and (ii) enables the extra, initially redundant copy to accumulate mutations that facilitate its rebirth as a new gene with a "hitherto non-existent function" (neofunctionalization) or hasten its degeneration into a "nonsense, DNA base sequence" [16, 20] or pseudogene (nonfunctionalization).

Analyses of entire populations of young gene duplicates identified from whole-genome sequence data have established that the duplication process shows little respect for gene boundaries and can spawn remarkably diverse sets of duplication products with varying degrees of structural resemblance to the ancestral copy. At one end of the spectrum, small-scale duplication (SSD henceforth) events faithfully duplicate the entire ancestral open reading frame (ORF) and possibly large stretches of upstream and downstream flanking regions, thereby capturing important ancestral *cis*-regulatory elements such as promoters. At the opposing end of the spectrum, other SSD events can display immense promiscuity by fashioning novel ORFs from both coding and noncoding genetic material ([21–24], among others). Furthermore, the recent discovery of the creation of *de novo* genes in entirety from noncoding DNA [25–31], although not duplicative in nature, completely turn Müller's [5] and Ohno's dictum [16] of "every gene from a pre-existing gene" on its head.

In this paper, I focus on the diversity of the gene duplication process whereby new genes are created by incorporating genetic tracts from previously existing genes as well as noncoding DNA (intergenic and intronic), and the evolutionary consequences of this promiscuity inherent in the gene duplication process. First, I describe the canonical model of gene duplicate evolution as envisioned by Ohno and delineate its major tenets as well as its failure to encapsulate the full complexity of the gene duplication process as revealed by whole-genome sequence data. Second, I discuss the various flavours of gene duplicates originating from both DNA- and RNA-mediated mutational events and explore their respective potential for the creation of evolutionary innovations and biological diversity. Third, I explore the various scenarios under which gene paralogs can escape homogenization by ectopic gene conversion, rendering them free to evolve along novel evolutionary trajectories and assume divergent functions. Lastly, I explore theoretical population-genetic considerations of how the effective population size (N_e) of a species may influence the probability of emergence of genes with radically altered functions.

2. Ohno's Canonical Model of Gene Duplicate Evolution

Ohno's overly restrictive view of the gene duplication process is related to his implicit assumption that the gene duplication process yields an extra copy that is fully redundant to the ancestral copy, both at the functional and sequence level. For this requirement to hold, the entire ancestral repertoire of coding sequence and regulatory elements would have had to be replicated during the duplication process. This supposed complete redundancy between duplicate copies then necessitates the implicit prediction that either copy, the ancestral or the derived, is capable of assuming (i) the ancestral function or (ii) becoming neofunctionalized/nonfunctionalized (see Figure 1). As such, the evolutionary fate of a duplicate copy under Ohno's model then rests on chance or stochastic events; the first gene copy to be hit by mutations, be they degenerating or neofunctionalizing, will be more prone to an altered evolutionary fate. We now know that gene duplicates, especially those stemming from SSD events often do not meet this assumption of functional equivalency to the ancestral copy at birth. In Ohno's defense, experimentally determined sequence data in the 1960s for newly originated gene duplicates from SSD events was scant at best, in direct contrast to the more abundant chromosomal complement data. For example, the high diploid chromosome numbers in modern vertebrate lineages and the vast differences in their chromosomal complements led Ohno and his colleagues to conclude that gene duplication by polyploidization constituted an "obligatory evolutionary requirement" for the evolutionary diversification of vertebrates [32–36]. It appears that Ohno himself was acutely aware that his model of gene duplicate evolution could possibly fall short of encapsulating the full complexity of the process, given the paucity of experimental data in his time. In the Introduction section of his 1970 treatise, he acknowledges: "In this golden age of biology, a book faces the danger of becoming obsolete before its publication. It is my belief that in order to avoid early obsolescence, the author, judging on the basis of the scant evidence available, is obliged to anticipate future

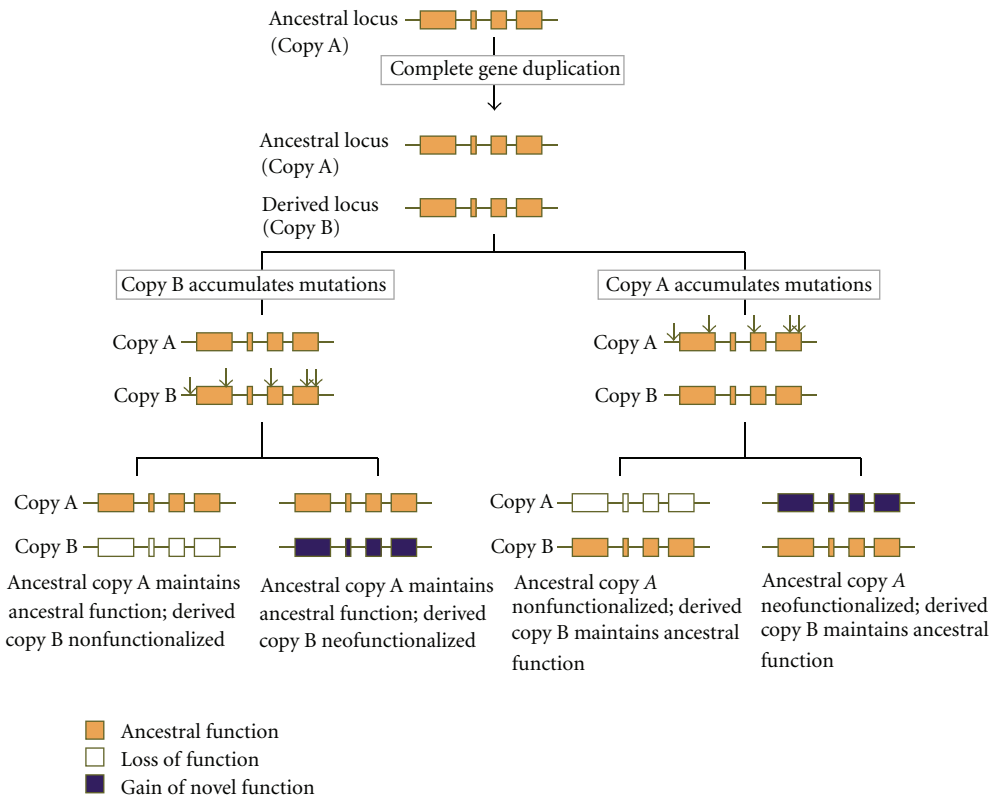


Figure 1: Different evolutionary fates of gene paralogs under Ohno's canonical model of gene duplicate evolution [16]. Ohno postulated that gene duplicates are born redundant in sequence and function to the ancestral copy. As such, either copy (ancestral or derived) has the potential to assume the identity of the locus maintaining the ancestral function leaving the other paralog free to accumulate mutations that may in turn engender vastly different evolutionary outcomes, namely loss of function (nonfunctionalization) or gain of a new function (neofunctionalization). A third evolutionary outcome not displayed in the schematic is the conservation of the ancestral function by both paralogs, in instances where natural selection favours increased gene dosage for increased levels of the ancestral gene product.

developments and paint a picture with broad strokes of his brush. This I have done rather freely in this book."

Ohno's model of functional equivalency of gene duplicates at birth was likely influenced by his views on the seminal role of polyploidization in evolution. Whole-genome duplication (WGD) or polyploidization offers several inherent advantages over SSD events such as tandem gene duplications, each of which were discussed in detail by Ohno [16]. First, gene duplication by polyploidization does not disrupt ancestral gene dosage ratios between functionally related genes. Second, the simultaneous duplication of all genes within an ancestral genome "in one fell swoop" via polyploidization would appear to have far greater evolutionary potential for functional diversification than SSD events creating one or a few duplicate copies at a time. However, rapid advancement of molecular techniques that enable genome-wide analyses of DNA content in conjunction with the use of experimental lines maintained under strict bottlenecking conditions to permit the accumulation of mutations under relaxed selective constraints have now provided evidence for astoundingly high genome-wide rates of spontaneous gene duplication via SSD events in the yeast, *Saccharomyces cerevisiae* [37] and the nematode, *Caenorhabditis elegans* [38] that exceed the base substitution rate by several orders of magnitude. These high per-locus duplication rates directly contribute to the immense copy-number variation being observed in various species [39–55]. Third, polyploidization also entails the coordinated duplication of the structural gene and associated *cis*- and *trans*-regulatory elements, thereby reducing the frequency of gene duplicates that are already nonfunctionalized at inception ("dead at birth") owing to the incomplete duplication of their regulatory systems. These advantageous characteristics of WGD-originated gene duplicates likely influenced Ohno's views on what comprised the most evolutionarily successful class of gene duplicates stemming from SSD events—essentially, *complete* gene duplication events where the ancestral coding sequence and the entire ancestral repertoire of regulatory elements were inherited intact in the derived paralog. In other words, Ohno's canonical model of gene duplicate evolution only focused on one particular class of gene duplicates arising from SSD events (*complete* gene duplicates) that, at inception, most resembled paralogs derived from polyploidization events. It is not that *partial* gene duplicates produced by incomplete duplication events were unknown to science; in fact, commencing in the early 1960s, a handful of studies had already established their existence [13–15] and discussed the implications of the creation of such genes with drastically

altered reading frames for the origin of evolutionary novelties [13]. It is not clear why Ohno overlooked entire classes of SSD duplicates derived from either incomplete duplication of a (i) single locus or (ii) multiple loci when arriving at his model of functional diversification of gene duplicates. Perhaps he believed that the majority of these duplicates were likely rendered nonfunctionalized at birth and as such, were rapidly eliminated from the population with minor or no evolutionary potential.

3. DNA-Mediated Duplication Events

3.1. Mechanisms of Mutation. DNA-mediated duplication (and deletion) events can originate via three mechanisms, namely (i) nonallelic homologous recombination (NAHR), (ii) nonhomologous end joining (NHEJ) and (iii) replication slippage.

Non-allelic homologous recombination (NAHR henceforth), also known as ectopic homologous recombination, refers to meiotic recombination between nonallelic but highly similar paralogous tracts of DNA that are already present in the genome. These extant paralogs in the genome are also referred to as low-copy repeats (LCRs) in the medical genetics literature [56, 57]. The NAHR pathway is by far the most precise means to repair double-strand breaks with no or minor loss of genetic information because it replicates the missing information from one homologous chromosome to another (interchromosomal) or between sister chromatids (interchromatid) or within the same chromatid (intrachromatid) [58, 59]. NAHR amongst paralogs in direct transcriptional orientation simultaneously leads to a duplication and deletion product each whereas NAHR among inverted paralogs results in inversions [59]. Unequal crossing-over events contributing to duplications are but NAHR events between paralogs in genomic proximity [60]. NAHR also requires extensive DNA sequence identity between paralogs in order to proceed (reviewed in [57]), approximately 50 bp in *E. coli* [61] and up to 300 bp in mammals [62, 63]. In contrast to NHEJ, which is facilitated by small DNA tracts of microhomology or no homology, paralogous DNA segments (or LCRs) facilitating NAHR are deemed lengthier with 95–97% sequence identity. Stankiewicz and Lupski [56] initially defined LCRs as ranging from 10–400 kb in size but a more recent paper by Hastings and colleagues has refined the criteria to encompass any paralogous segments >1 kb in size and with >95% sequence identity [57]. NAHR serves as a major contributor to genomic rearrangements such as duplications and deletions; Kidd et al. [64] inferred NAHR as having the greatest contribution to the formation of duplicates (~42%; 41/98 events) and 38% (49/129 events) of all deletion events in their analysis of structural variants in eight human genomes.

Nonhomologous end joining (NHEJ henceforth) is another important contributor of duplications and deletions and like NAHR, is a recombination repair pathway for double-strand breaks in multicellular eukaryotes [65]. However, NHEJ differs from NAHR in that it requires little or no sequence homology. Hence, the NHEJ pathway is often described as being homology-independent. NHEJ works to modify the two broken ends of a double-strand break, rendering them compatible and capable of rejoining but with concomitant loss of genetic information; as such it is a far more imprecise repair mechanism. The fact that it is commonly employed for DNA repair in multicellular eukaryotes despite its imprecise nature has been somewhat of a puzzle. Lieber et al. [58] have proposed that NHEJ is far more efficient in effecting DNA repairs in highly repetitive regions of the genome compared to the NAHR pathway, hence its common deployment in multicellular eukaryotes whose genomes comprise a substantial fraction of repetitive DNA elements. While NAHR is restricted to late S or G2 of the cell cycle, the NHEJ pathway is ubiquitous and can function through all phases of the cell cycle [65]. NHEJ events can lead to complex structural changes simultaneously involving duplications, deletions and inversions with microhomology junctions ([66, 67]; reviewed in [57]) and were found to contribute to the origin of ~30% and ~45% of the characterized duplications and deletions, respectively, in the human genome [64].

Slipped-strand mispairing or replication slippage is a third mechanism mediating duplications and deletions of DNA fragments. By the 1980s, multiple independent studies had already reported on the existence of minisatellites or VNTRs and the hypervariable genetic variation associated with their occurrence [68–70]. This was rapidly followed by the discovery of microsatellites or short tandem repeats (STRs) or simple sequence repeats (SSRs) [71]. Replication slippage or slipped-strand mispairing can lead to both duplications and deletions of genomic regions associated with these STRs [72–74]. The proximity of the repeat sequences and their high degree of sequence homology are expected to have a saltatory effect on gene family expansion and shrinkage [75].

3.2. Complete Gene Duplicates. *Complete* gene duplications are characterized by the duplication of an entire gene (Figure 2(a)). A strict adherence to Ohno's model of gene duplication then necessitates that for a *complete* duplicate to be redundant in both sequence and function to the ancestral copy, the entire ancestral coding region and regulatory elements would have to be inherited by the duplicate copy. Because *cis*-regulatory elements are poorly annotated in most genomes, efforts aimed at *complete* duplicate identification have relied entirely on a direct comparison of the ORF nucleotide sequences of the two paralogs. The paralogs exhibiting nucleotide sequence homology between their initiation and termination codons (including introns, when present) have traditionally been classified as *complete* duplicates. Therefore, some unknown proportion of the *complete* gene duplications identified in the manner described above likely had the ancestral ORF sequence duplicated without the concomitant duplication of the ancestral repertoire of regulatory elements, which may induce a divergent evolutionary trajectory for the newly created paralog at conception itself. As such, a subset of putative *complete* duplicates would fail to meet Ohno's strict definition of derived copies

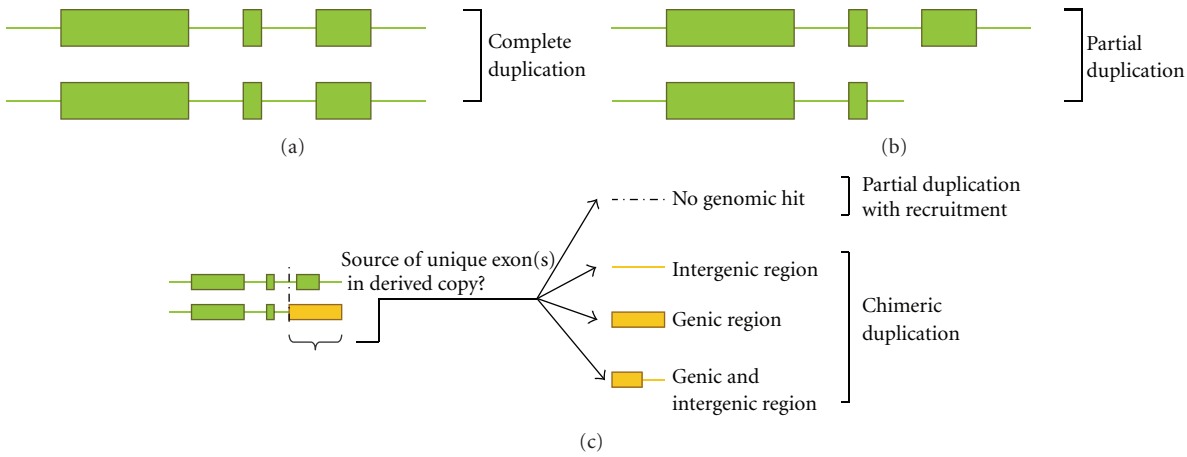


Figure 2: Small-scale DNA-mediated duplication events can yield daughter loci with varying degrees of structural resemblance to the ancestral copy depending on the extent of the duplication span and the location of the duplication breakpoints. Rectangles represent exons and solid horizontal lines through exons denote introns and flanking region sequences. Corresponding colours between the ancestral (top) and the derived (bottom) locus denote sequence homology. (a) *Complete* gene duplication wherein the duplication event spans, at a minimum, the entire ORF of the ancestral locus from the initiation codon to the termination codon. The duplication event may or may not encompass upstream and downstream flanking region sequences. In the schematic, *complete* duplication of an ancestral locus yields a derived copy (bottom) comprising three exons and intervening introns as well as some 5′ and 3′ flanking region sequences. (b) *Partial* gene duplication wherein only a portion of the ancestral ORF is duplicated. In the schematic, the downstream duplication breakpoint occurred within intron 2 of the ancestral copy (top), yielding a truncated derived copy (bottom) comprising some of the ancestral 5′ flanking region sequence, exons 1 and 2 and part of intron 2. (c) *Chimeric* gene duplication or *partial* gene duplication with *recruitment*. In instances wherein the derived locus (bottom) has unique exon(s) to the exclusion of the ancestral copy, the type of duplication event depends on the genomic source(s) of the unique coding region(s). In the schematic, the derived copy (bottom) has a unique exon 3 (yellow) to the exclusion of the ancestral locus (top). If a BLAST query of the unique exonic sequence yields no hits in the genome, it is suggestive of a *partial* gene duplication event and subsequent recruitment of intergenic sequence by the duplicate copy from its new genomic neighbourhood to yield an intact ORF. To qualify as a *chimeric* duplicate, this unique coding region of the derived locus must also exhibit evidence of duplication from another genomic source, be it intergenic, genic or a combination of genic and intergenic regions. The creation of such fusion genes could have occurred as a single evolutionary event (a single duplication event encompassing portions of two adjacent genes) or may represent independent duplication events and subsequent fusion of these fragments via shuffling.

being created functionally identical to the ancestral locus. The actual proportion of *complete* duplicates that display functional divergence from the ancestral locus at or close to birth despite inheriting the ancestral copy's full coding sequence remain to be characterized via detailed functional assays.

3.2.1. Contrasting Proportions of Complete Duplicates between Unicellular and Multicellular Eukaryotes. Only a handful of studies have identified the relative frequencies of *complete* duplicates relative to other structural categories of duplicates (*partial* and *chimeric*) in an entire genome or a particular age-cohort of duplicates within a genome (Table 1). For reasons identified in the preceding paragraph, the number of true *complete* duplicates bearing complete functional redundancy to the ancestral locus may be significantly fewer in frequency than reported in these studies. A study of putative evolutionarily young gene duplicates (synonymous divergence $dS < 0.1$) comprising small multigene families in the *Caenorhabditis elegans* genome found that *complete* duplicates comprised only ~40% (114/290) of all duplicate pairs considered [76]. *Drosophila* genomes tend to have remarkably similar proportions of *complete* duplicates to *C. elegans* within their cohort of evolutionarily young gene duplicates; ~41% in *Drosophila melanogaster* [24] and ~44% in *D. pseudoobscura* [77]. Contrastingly, the proportion of extant *complete* duplicates stemming from SSD events in the sequenced *Saccharomyces cerevisiae* genome is significantly greater at 82% (18/22 pairs) [78].

The contrasting proportion of *complete* duplicates in multicellular eukaryotes (*C. elegans*, *D. melanogaster*, and *D. pseudoobscura*) relative to the unicellular eukaryote *S. cerevisiae* begs further evaluation. Clearly, more genomes within diverse taxonomic groups and across kingdoms will have to be evaluated to generate robust sample sizes and greater phylogenetic independence in order to enable strong generalizations. However, if prokaryotes and unicellular eukaryotes are indeed verified as having a majority of *complete* duplicates (among genes of SSD-origin only) and multicellular eukaryotes a minority, there are several scenarios that may, individually or in concert, contribute to this pattern. First, assuming that DNA-replication errors and double-strand breaks are likely to yield similar distributions of duplication spans across diverse genomes, the probability of encompassing an entire gene during duplication is greater in compact genomes with smaller average gene lengths (shorter exon length, fewer introns, and/or shorter intron

Table 1: Extant percentage of *complete*, *partial* and *chimeric* duplicates originating from small-scale DNA-mediated duplication events in sequenced genomes, mutation accumulation lines, or natural isolates of model organisms.

Phylogenetic distribution	% Gene duplicates from SSD-events			Comments	References
	Complete	*Partial*	*Chimeric*		
Caenorhabditis elegans					
Sequenced genome of N2 strain	39	21	40	290 duplicate pairs with $dS \leq 0.10$	[76]
Mutation accumulation (MA) lines	37	63	—	30 spontaneous duplication events across ten MA lines bottlenecked for an average of 432 generations	[38]
Drosophila melanogaster					
Sequenced genome	41	27	32	72 duplicate pairs with >80% sequence identity	[24]
Natural isofemale lines	14	76	10	928 CNVs across 15 natural isolates	[45]
Drosophila pseudoobscura					
Sequenced genome	44	56	—	101 duplicate pairs derived from DNA-mediated duplication events; >80% sequence identity	[77]
Saccharomyces cerevisiae					
Sequenced genome	89	7	4	47 duplicate pairs derived from small-scale duplication events; $dS \leq 0.35$	[78]

length). Conversely, in species with lengthier average gene lengths due to the elongation of genic coding regions and/or the presence of numerous lengthy introns, gene duplication is less likely to capture an ORF in its entirety, yielding a greater number of partial gene fragments. The average length of coding sequences in eukaryotes was found to be 445 bp longer than their prokaryotic counterparts [79]. Coding sequence differences may therefore contribute marginally to average gene length differences between the two kingdoms. However, average intron lengths and intron densities should have a far greater contribution to variation in average gene length between lineages given that (i) they can be highly variable across diverse taxa [80–85] as well as between closely-related taxa [86] and (ii) duplication across introns is part and parcel of DNA-mediated duplication events. The median gene lengths within the *S. cerevisiae* and *C. elegans* genomes do not appear starkly different (1.1 and 1.4 kb, resp.) and yet the frequencies of *complete* duplicates in their genomes are substantially so (82% and 40% in *S. cerevisiae* and *C. elegans,* resp.) [78]. This suggests that some other factor(s) may be implicated in the observed differences in the frequencies of extant *complete* duplicates within their genomes.

A second possibility is that structurally heterogeneous duplicates such as *partial* and *chimeric* duplicates may have deleterious effects at birth and may be more effectively eliminated by purifying natural selection in some genomes relative to others, given that the efficacy of natural selection is greater in genomes of species with larger effective population sizes (N_e). This concept is discussed in greater detail in Section 6 below. The first sequenced genomes to become available were often of individual(s) derived from laboratory strains or natural populations. As such, the source population of the sequenced individual(s) may have already been subject to selection (or genetic drift) and characterization of all current duplications would not represent the entire spectrum of spontaneous duplications that may have arisen within the genome in the recent past. In other words, some duplications with deleterious effects may have already been purged from the genome in their infancy prior to their identification in whole-genome sequence data, leading to underestimates of the spontaneous duplication rate as well as a skewed pool of gene duplicates with lower rates of loss. Extending this reasoning to the observed frequencies of *complete* duplicates in various sequenced genomes, it is not possible to distinguish whether the high frequency of extant *complete* duplicates, say in the initial whole-genome sequence of *S. cerevisiae* [87], is due to a higher spontaneous rate of *complete* gene duplication or greater purifying selection acting to weed out *partial* and *chimeric* duplicates. The spontaneous rate of gene duplication and the spectrum of different classes of gene duplications have been characterized in long-term *C. elegans* mutation accumulation lines with severely diminished efficacy of natural selection [38]. Lipinski et al. [38] characterized 30 duplicated genes via DNA-mediated duplication events, of which ~37% (11/30) were *complete* duplicates which is in accord with the observed frequency of 40% *complete* duplicates in the originally sequenced genome of the *C. elegans* N2 laboratory strain [76]. The picture is far from clear for genomes of prokaryotes and unicellular eukaryotes with large N_e. The initially sequenced genome of *S. cerevisiae* has relatively few extant gene duplicates created from SSD-events [78, 88]. While this may initially give the appearance of a low spontaneous rate of gene duplication, whole-genome sequencing of yeast mutation accumulation lines has revealed, to the contrary, an extremely elevated rate of spontaneous duplication and deletion far exceeding the base substitution rate [37]. This suggests that most duplication/deletion events in yeast are selectively purged from the genome in their infancy. The relative frequencies of spontaneously arising *complete*, *partial*, and *chimeric* duplicates in yeast has not yet been experimentally determined. It may be possible in the future to conclude whether the higher frequency of *complete* duplicates in the initially

sequenced *S. cerevisiae* genome is owing to (i) a higher rate of *complete* gene duplication, or (ii) a greater efficacy of selection against structurally heterogeneous gene duplicates (*partial* and *chimeric*), or (iii) a combination of scenarios (i) and (ii).

3.2.2. Are Complete Duplicates More Limited in their Ability to Explore Evolutionary Space and Assume Radically Novel Functions? Complete duplicates as defined under Ohno's classical model of gene duplicate evolution commence their evolutionary life structurally and functionally redundant to the ancestral copy from which they are derived. If an altered environmental regime induces a selective pressure for amplification of an ancestral gene product, *complete* duplicates are the best poised relative to other structural classes of duplicates (such as *partial* and *chimeric* duplicates) to assist the ancestral copy in responding to cellular needs for "more of the same." Gene amplification involving segmental duplications of gene clusters yielding mostly *complete* duplicates is known to enable bacterial growth in carbon-limited environments [89–91] and the evolution of antibiotic resistance [92]. A similar pattern is observed in multicellular eukaryotes. Gene amplification is implicated in copper resistance by yeast via tandemly arrayed duplications of the *CUP1* locus [93], insecticide resistance [94, 95], and heavy metal tolerance by insects [96], in the recruitment of lysozyme as a major stomach enzyme in cows [97] and in protozoan resistance to drugs [98]. The initial preservation of the duplicate copy under selection for increased dosage does not preclude eventual functional diversification of the two paralogs via refinement of their secondary functions as envisioned under the IAD (innovation, amplification, divergence) model of Bergthorsson et al. [99].

Ohno [16] believed that the extra copy first accumulates debilitating "forbidden" mutations leading to a loss of function, which in turn instigates its evolution along an altered evolutionary trajectory under a regime of relaxed selective constraints. While the majority of the accumulating mutations in the extra copy are expected to be nonfunctionalizing, a rare beneficial mutation may arise and impart a novel function, thereby facilitating its resurrection. This has also been referred to as the *mutation during nonfunctionality* (MDN) model [100]. If sequence and functional divergence between paralogs is largely a consequence of point mutations in the postduplication period, how might this influence the evolutionary potential of *complete* duplicates? There are instances of *complete* duplicates whose evolution proceeds along the trajectory envisioned by Ohno, such as visual pigment proteins in catarrhine primates [101] and pancreatic ribonuclease paralogs in colobine primates [102]. Catarrhine primates (Old World monkeys, apes, and humans) have trichromatic vision due to an evolutionarily recent X-linked duplication yielding the red and green opsins, each of which comprises six homologous exons in humans, pygmy chimpanzee, gorilla and orangutan [103]. Hence, red and green opsins represent a *complete* gene duplication event. The encoded opsin proteins differ intraspecifically by 12–18 amino acids [103], of which as few as three residues (positions 180, 230 and 285) have been shown to account for the spectral difference of ≈30 nm in peak absorption values between the red (L opsin) and green (M opsin) photopigments [104–106]. Interestingly, one species of Platyrrhine monkeys, the howler monkey, displays convergent evolution via duplication of the X-linked L opsin to yield a new M opsin independently of the Catarrhine primates ([107] reviewed in [108]). In both instances, acquisition of a novel function, the evolution of a novel green photoreceptor encoded by the M opsin via duplication of the L opsin (encoding the red photoreceptor) was effected by the accumulation of point mutations in the coding regions. Alternatively, the accumulation of point mutations in regulatory regions of a *complete* duplicate has the potential to alter tissue specificity or temporal expression patterns relative to the ancestral copy. However, I propose that, collectively speaking, *complete* duplicates have markedly less potential to assume "radically" novel functions from their progenitor copy. The argument that the creation of structurally heterogeneous duplicates with a radical refashioning of ancestral exon-intron structure can lead to immediate acquisition of drastically novel function conferring a great selective advantage has been presented before [13, 23, 75]. A gradual increase in sequence divergence between *complete* duplicates via point mutations can lead to minor tinkering of the ancestral function or an alteration of expression patterns. Successive mutational hits by base substitutions in one paralog could lead to incremental changes in its function but it might require substantial evolutionary time to evolve a drastically altered novel function if other alterations to exon-intron structure via indels or shuffling events are absent. Despite 35–40 million years of evolutionary divergence from its progenitor, the green opsin gene of Old World primates is still very much a photoreceptor gene as is its ancestral paralog. The origin of radically altered functions might be more the domain of *partial* and *chimeric* duplicates with extensive changes to their exon-intron structure.

3.3. Partial Gene Duplicates. Partial gene duplications are also referred to as incomplete, intragenic, or internal gene duplications and are characterized by the incomplete duplication of an ancestral gene (Figure 2(b)). In some instances, *partial* gene duplication is additionally associated with *de novo* internal amplification of short DNA tracts. The most accurate means of identifying *partial* duplicates would be to align the ORFs of a known ancestral and derived copy within a duplicate pair and visually determine if the derived copy is a truncated version of the ancestral ORF sequence. In some instances, the derived copy's ORF may appear superficially attenuated in length relative to that of the ancestral copy but its 5′ and/or 3′ flanking sequence will exhibit complete sequence homology to the ancestral ORF. These cases should be treated as *complete* duplication events because they are suggestive of a role of postduplication events in the alteration of the derived copy's ORF. For example, base substitutions or frameshift mutations can lead to the conversion of an ancestral sense codon to a premature stop codon in the derived copy's ORF. Conversely, the derived

paralog's ORF may appear superficially lengthier than its ancestral counterpart, but sequence analysis may reveal a true *partial* duplication associated with massive internal gene amplification of a small ancestral DNA sequence tract. A further complicating factor in the accurate identification of *partial* duplicates may be posed by the presence of unique coding sequence in the derived copy's ORF, to the exclusion of the ancestral copy [23]. This may lead to the erroneous conclusion that the derived copy is a *chimeric* duplicate derived from the duplication of multiple genomic sources. However, if the unique sequence of the derived copy fails to generate any valid genomic hits for potential donor sequence(s) via a BLAST search, it is suggestive of a *partial duplication with recruitment* (see top panel of Figure 2(c)). In other words, the derived paralog was formed via a *partial* duplication of the ancestral copy and additionally recruited neighbourhood sequence from its new genomic location to complete its ORF. For reasons outlined above, accurate identification of structurally heterogeneous classes of duplicates such as *partials* and *chimerics* are more challenging than *complete* duplication events.

Determining the ancestral and derived copy within duplicate pairs showing *partial* structural resemblance is best facilitated by comparing the exon-intron structure of the two paralogs within the focal genome to that of a single-copy ortholog in a closely-related outgroup genome. The paralog within the focal genome displaying greater similarity in exon-intron structure to the single-copy ortholog is taken to represent the ancestral paralog. However, in the absence of outgroup genomic sequences, the earliest studies tentatively classified paralogs as *partial* duplicates when direct examination of their ORF sequences revealed that one paralog had unique coding sequence to the exclusion of the other paralog (e.g., [23]). That is, the shorter paralog had a truncated ORF and displayed no sequence homology with the lengthier paralog beyond the putative duplication breakpoint within its ORF. The disadvantage of this approach to *partial* duplicate identification is the possibility that the shorter copy is actually the ancestral paralog and the lengthier copy may have resulted from a *complete* gene duplication of the ancestral ORF and the addition of extra sequence via recruitment of noncoding DNA or shuffling events involving other genic regions. In a subsequent study, Katju and Lynch [23] utilized the *C. briggsae* genome as an outgroup to assign ancestral versus derived copy status to paralogs within a subset of 14 *C. elegans* duplicate pairs that had been previously classified as *partial* duplicates [76]. The authors found that 13/14 (93%) of these were indeed *partial* duplications in that the lengthier *C. elegans* paralog was the ancestral copy given its exon-intron structure resembled that of the single-copy *C. briggsae* ortholog. In only one instance, the shorter paralog was the ancestral copy with the lengthier paralog representing a novel chimeric gene derived from multiple genic and intergenic sources. More interestingly, 43% of a subset of *C. elegans* 23 duplicate pairs previously identified as *chimeric* duplicates (both paralogs had unique coding sequence to the exclusion of the other copy) were actually cases of *partial duplication with recruitment* (see Figure 2(c)). As such, their original study [76] likely underestimated and overestimated the number of *partial* and *chimeric* duplicates, respectively.

Given that a mere comparison of exon-intron structures of two paralogs may belie the actual type of duplication contributing to the creation of the extra copy, I propose a stricter definition of what comprises a true *partial* gene duplication event. Most crucially, an accurate assignment will require independent verification of the identities of the ancestral versus derived copy via comparative genomic approaches. A *partial* duplicate should only be derived from the partial duplication of a single ancestral genic source. In some instances, the *partial* duplicate may possess unique ORF sequence to the exclusion of the ancestral copy, leading to the erroneous conclusion that it is a *chimeric* duplicate. If this unique ORF sequence of the derived copy fails to generate any hits in the genome (exonic, intronic or intergenic), it should be classified as originating from a *partial duplication event with recruitment* wherein additional neighbouring sequence from the derived paralog's new insertion site were utilized to fashion novel exon(s) and/or intron(s) (see first panel in Figure 2(c)). Because these novel exon(s) and/or intron(s) in the duplicated copy are not derived from independent duplication events involving subsequent shuffling and fusion, as such the derived paralog is still very much a *partial* duplicate despite its superficial chimeric appearance. For example, the *Hun* gene in *Drosophila* is thought to have arisen from a *partial* duplication of the *Bällchen* gene with additional recruitment and exonization of flanking intergenic sequence [109]. Hence, *Hun* ought to be classified as a *partial duplicate with recruitment,* not a *chimeric* duplicate, as its novel ORF sequence (comprising ~33 amino acids) is not derived from an independent duplication event.

3.3.1. High Genomic Abundance of Partial Duplicates in Multicellular Eukaryotes. Relatively high frequencies of *partial* duplicates have been identified in the sequenced genomes of *C. elegans*, *D. melanogaster*, *D. pseudoobscura*, and *S. cerevisiae* and directly falsify one of the major tenets of Ohno's model [16] that gene duplicates are created structurally redundant to their ancestral counterparts. Katju and Lynch's 2003 structural analysis of *C. elegans* paralogs [76] lacked a reference outgroup genome for comparison as the *C. briggsae* sequenced genome had yet to be released. *C. elegans* duplicate pairs with paralogs of differing amino acid lengths wherein the entire ORF of the shorter paralog was homologous to the lengthier paralog's ORF but the latter had unique ORF sequence to the exclusion of the shorter paralog were classified as putative *partial* duplicates. This class of *partial* duplicates comprised 21% of 290 duplicate pairs with synonymous divergence per synonymous site (dS) ranging from 0 to 0.1 (Table 1). As discussed in the previous section, this is likely an underestimate as a subsequent analysis by the authors using the *C. briggsae* genome as an outgroup showed that 43% of a subset of 23 *chimeric* duplicate pairs in reality represented *partial duplicates with recruitment* [23]. If the results generated from the subset of 37 *C. elegans*

duplicate pairs can be generalized to the entire data set of 290 duplicate pairs, *partial* duplicates may represent ~38% of young *C. elegans* duplicates (rather than 21%) and *chimeric* duplicates only 23% (instead of 40%). These predictions await further experimental validation via a comparative genomic approach identifying the ancestral versus derived paralog for all duplicate pairs in *C. elegans*. Zhou et al. [24] reported *partial* duplicates as comprising 27% of newly originated genes in *D. melanogaster* (Table 1). Meisel's [77] study utilized a single-copy *D. melanogaster* ortholog as an outgroup to determine the directionality of structural alterations, if any, within *D. pseudoobscura* paralogs. This is a far superior approach as it enables accurate classification of the different structural classes of paralogs within a genome. The majority of *D. pseudoobscura* duplicate pairs (53%) were classified as *partial* duplicates (Table 1). The unicellular eukaryote, *S. cerevisiae*, has an extremely low percentage of identifiable *partial* duplicates, at about 7% (Table 1) even though older cohorts of gene duplicates with $dS \leq 0.35$ were included in the analyses [78]. Furthermore, some fraction of these yeast duplicates may be evolutionarily older than that suggested by their *dS* values given the high degree of paralog homogenization in this genome due to the concerted action of codon usage bias selection and ectopic gene conversion. As discussed in Section 3.2.1, the paucity of structurally heterogeneous gene duplicates (*partial* and *chimeric*) in *S. cerevisiae* may be due to (i) a higher probability of *complete* duplicate origin given the compact nature of the genome and/or (ii) stronger purifying selection against structurally heterogeneous paralogs.

Two studies, one of experimentally evolved laboratory lines and the other of natural isolates report the highest fraction of *partial* duplicates thus far (Table 1). Lipinski et al.'s [38] analysis of long-term *C. elegans* mutation accumulation lines found *partial* duplicates accounting for 63% of detectable spontaneous gene duplication events. As with the Lipinski et al. study [38], Meisel [77] found zero frequency of *chimeric* duplicates in *D. pseudoobscura*. These results suggest the tantalizing hypothesis that while *chimeric* duplicates can be created in one fell swoop by duplication across two adjacent genes, the majority of them may owe their creation to secondary fusion events involving the recombination of previously duplicated *partial* duplicate fragments. A population-genomic study of evolutionarily young genes still segregating as copy-number variants (CNVs) in 15 *D. melanogaster* natural isofemale lines identified 76% of all duplications to be *partial* duplicates [45]. It is not clear if CNVs associated with these *partial* duplicates are indicative of insufficient evolutionary time for fixation or represent the incipient stages of eventual loss from the genome.

3.3.2. Partial Duplicates Can Encode Drastically Novel Functions Relative to the Ancestral Copy. *Partial* duplicates were most certainly overlooked by Ohno as having much evolutionary potential. Even in the current genomic era, the common trend is to lump together *partial* duplicates as pseudogenes given that their ORFs show signatures of disruption to the ancestral reading frame and the presence of premature stop codons. While many *partial* duplicates may indeed be evolutionary dead ends, their high rates of origin and potential for evolution of radically novel functions due to their drastically altered exon-intron structure relative to the ancestral copy urges some measure of caution against, in common parlance, throwing the baby out with the bathwater. Furthermore, *partial* duplicates may remain nonfunctional in a genome for some evolutionary time, but may be resurrected by exapting to novel functions under altered environmental regimes. In the paragraph below, I highlight some intriguing examples of acquisition of radically altered function by *partial* duplicates (listed in Table 2).

Freezing avoidance by various polar and subpolar species of fishes highlight the remarkable nature of adaptation in living organisms owing to the presence of certain antifreeze proteins that evolved independently from functionally unrelated ancestral genes. The molecular origins of these antifreeze proteins are testament to the evolutionary potential of *partial duplication* events in conjunction with *de novo* internal amplification of short sequences. The antifreeze glycoprotein (AFGP) of Antarctic notothenioid fish was likely created by a *partial duplication* of an ancestral pancreatic enzyme trypsinogen wherein a small portion of its 5′ untranslated region, E1, I1, a small fragment of E2, terminal E6, and a portion of the 3′ untranslated region were initially duplicated. This was likely followed by the internal amplification of a 9 bp sequence straddling the first intron-second exon that eventually encoded the repetitive tripeptide backbone of the AFGP which directly contributes to the novel protein's ice-binding capacity and inhibitory effect on the growth of ice crystals [21]. Deng et al. [110] recently elucidated the independent evolutionary origin of another class of antifreeze proteins, the AFPIII (Type III antifreeze protein) in an Antarctic zoarcid fish from an ancestral sialic acid synthase (SAS) gene unrelated to trypsinogen. The two-exon AFPIII gene was derived from a *partial* duplication of the ancestral SAS gene, encompassing a portion of both its 5′ flanking region sequence and E1, I5, terminal E6, and portion of the 3′ flanking region sequence [110]. The AFPIII locus comprises >30 AFPIII genes arrayed in ~8 kb repeats with one AFPIII gene per repeat [110, 111]. E2 of AFPIII imparts the antifreeze property of the molecule and is derived wholly from the ancestral SAS terminal E6 exon. E1 of AFPIII encodes for a signal peptide for extracellular export of the mature antifreeze protein and was created *de novo* by combining 54 nt of 5′ flanking region upstream from the translation start site and inclusion of the first six codons of the ancestral E1 of SAS. Indeed, exonization of ancestral noncoding sequence, as is implicated in the origin of the first exon of the novel AFPIII gene, may have the greatest contribution to new domain gains in animal proteins relative to retrotransposition and recombination-mediated intronic insertion events [112].

Partial duplications with recruitment bear immense potential to generate radically novel functions due to exonization of ancestral noncoding sequences leading to the possible emergence of novel protein domains. Species in the genus *Caenorhabditis* employ one of two modes of reproduction. Nine of the first 11 species to be cultured display

Table 2: Some examples of *partial* gene duplicates conferring novel function.

Phylogenetic distribution	Partial duplicate	Ancestral locus	Type of partial duplication	Comments	References
Antarctic Notothenioid fish *Dissostichus mawsoni*	*AFGP*	Trypsinogen	Partial duplication with internal amplification	Creation of a novel antifreeze glycoprotein from an ancestral pancreatic enzyme	[21]
Antarctic eelpout *Lycodichthys dearborni*	*AFPIII*	Sialic acid synthase	Partial duplication in tandem array	Creation of a novel antifreeze protein from an ancestral cytoplasmic enzyme	[110]
Caenorhabditis elegans	*fog-2*	*ftr-1*	Partial duplication with recruitment	Creation of a novel gene implicated in hermaphrodite spermatogenesis from an ancestral gene of unknown function; evolution of hermaphroditism	[113]
Common ancestor, of *Drosophila simulans*, *D. mauritiana, D. sechellia*	*Hun*	*Bällchen*	Partial duplication with recruitment	Creation of a novel gene with testis-specific expression from an ancestral kinase gene	[109]
Homo sapiens	*SRGAP2C*	*SRGAP2*	Partial duplication	Novel gene unique to humans; linked to increased cognitive ability in the *Homo* lineage	[114, 115]
Xenopus laevis	*DM-W*	*DMRT-1*	Partial duplication	Creation of a novel female sex-determination gene	[116–118]

a gonochoristic obligate female/male outcrossing mode of reproduction. Two species, *C. elegans* and *C. briggsae,* have an androdioecious breeding system with populations composed of self-fertile hermaphrodites and males at a low frequency (<0.1%) [119]. Two independent lines of evidence suggest convergent evolution of hermaphroditism within *C. elegans* and *C. briggsae* as follows: (i) these two hermaphroditic species are phylogenetically separated by two gonochoristic species [120] and (ii) the sperm production pathway in the hermaphrodites of these two species involves different genes [121]. The evolution of hermaphroditism in *C. elegans* may have been specifically promoted by the appearance of a novel gene, *fog-2*, via a *partial* gene duplication of *ftr-1*, a gene of unknown function [113]. The appearance of *fog-2* conferred *C. elegans* hermaphrodites with a limited ability to perform spermatogenesis [122, 123]. The ancestral locus, *ftr-1* comprises four exons encoding 314 amino acids (aa). The exon-intron structure of *fog-2*, comprising five exons (327 aa) exhibits both similarities and dissimilarities relative to *ftr-1*. *fog-2* was created by the duplication of *ftr-* 1′s E1-E3 and part of E4. The C-terminal region of *fog-2* encompassing the latter half of E4, I5 and E5, as well as its 3′ flanking region bear no obvious sequence homology to *ftr-1*, nor do they generate any sequence hits in the *C. elegans* genome. Hence, *fog-2* was created by a *partial duplication with recruitment* event involving exonization of noncoding sequence from its new genomic neighbourhood to complete its open reading frame [113, 124]. Notably, the recruitment and subsequent exonization of this unique noncoding sequence in *fog-2*′s 3′ end may have facilitated neofunctionalization after duplication, given that this novel region is implicated in binding with a translation repressor GLD-1 that represses feminization and promotes hermaphrodite spermatogenesis [122]. The X-linked *Hun* gene in three *Drosophila* species represents another example of a sex-specific gene created by *partial duplication with recruitment* [109]. *Hun* was created by a *partial* duplication of the autosomal gene *Bällchen*, a kinase involved in germ cell development, with subsequent recruitment of new neighbourhood sequence into its terminal exon, leading to a novel testis-specific expression.

More recently, the origin of an extra *SRGAP2* gene in the lineage leading to modern humans via a *partial duplication* event is being credited with major evolutionary changes related to brain development and advancement of cognitive abilities in the human lineage and its divergence from primate relatives [114, 115]. The ancestral *SRGAP2* gene comprising 22 exons (encoding 1071 aa) underwent several independent duplication events in the human-lineage leading to the creation of *partial* duplicates *SRGAP2B-SRGAP2D* spanning the first nine ancestral exons [115]. One of the *partial* paralogs, *SRGAP2C,* encoding a truncated version of the ancestral SRGAP2 protein product comprising 458 aa residues as well as 7 unique residues at the carboxyl terminus, is thought to dimerize with the ancestral protein product leading to a dominant negative interaction essentially involving the knockout of the ancestral gene function and facilitating (i) rapid neuron migration and (ii) the development of greater spine extensions on neuronal surfaces which in turn are thought to facilitate greater connections between neurons [114].

Partial duplicates have also been implicated in the regulation of the ancestral paralogs from which they are derived. The putative *partial* duplicate of the nitric oxide synthase (NOS) expressed in the central nervous system of the snail *Lymnaea stagnalis* appears to function as a translational regulator to inhibit the translation of the ancestral neuronal NOS protein [125]. The mouse *Makorin1-p1* is a truncated version of the *Makorin-1* gene and presumably arose via a *partial* gene duplication spanning 700 bp of the

5′ region of the ancestral, 2600 bp long *Makorin-1* gene [126]. The authors proposed that the mRNA expression of the *partial* duplicate *Makorin1-p1* inhibited degradation of the ancestral locus *Makorin*-1′s mRNA, thereby enhancing and stabilizing the ancestral gene's expression. Furthermore, evolutionary analyses suggested that *Makorin1-p1* was not evolving under relaxed selective constraints as would be expected of a neutral locus [127]. However, the authenticity of Hirotsune et al.'s [126] conclusions were subsequently debated by Gray et al. [128] who argued that *Makorin1-p1* is not transcribed and the mRNA attributed to it was an alternatively spliced variant of the ancestral *Makorin-1* locus. Another intriguing example of a *partial* duplicate acting in a regulatory role is the *DM-W* gene in the African clawed frog, *Xenopus laevis*. *DM-W* functions in female sex-determination and appears to have originated from the *partial* duplication of the first four exons of the male-specific, six-exon autosomal gene *DMRT1β* [116, 117]. *DM-W* initiates primary ovary formation in female gonads by antagonizing the activation of male-specific genes by *DMRT1* via transcriptional repression [116, 118]. Likewise, targeted knockdown of *ABCC6P1*, a putative *partial* duplicate of the human ACB transporter genes *ABCC6* leads to a reduction in the mRNA expression of *ABCC6* [129]. Finally, the role of *partial* duplicates in the formation of small noncoding RNAs with novel regulatory functions has barely been touched upon [130]. For example, Guo et al. [131] identified 22,956 DNA-mediated "pseudogenes" in the rice genome with a subset of them being strong candidates for assuming novel regulatory functions as small RNAs. However, the exact structural nature of these putative "pseudogenized" paralogs has yet to be elucidated in detail.

3.4. Chimeric Gene Duplicates. Current literature is replete with examples of putative *chimeric* duplicates. The diversity of mechanisms that can lead to the formation of novel genes exhibiting a mosaic or chimeric appearance certainly adds to the confusion that abounds with respect to their classification. As discussed by Cardoso-Moreira and Long [132], a multitude of genomic rearrangements following gene duplication (such as deletions, inversions, and translocations) can lead to a chimeric gene structure. The difference between a chimeric appearing gene and a true *chimeric* duplicate is subtle but cannot be relegated to pure semantics, and hence ought not to be ignored. In their review article [132], Cardoso-Moreira and Long offer the following definition: "A new gene is considered chimeric if it recruits novel sequence from nearby regions." But if its creation involved gene duplication, what class of gene duplicate would it represent? As I have highlighted in the *fog-2* example in Section 3.3.2, a *partial duplication with recruitment* event created the chimeric/mosaic structure of *fog-2*. A partial fragment of the ancestral *ftr-1* gene's ORF was duplicated in conjunction with exonization of new neighbourhood noncoding sequence to render an intact ORF. Because *fog-2* is derived from the duplication of only one ancestral source, it qualifies as a *partial* duplicate, albeit with *recruitment*.

I propose that *chimeric* gene duplicates be classified as paralogs derived from the duplication of two or more ancestral donor sequences, with at least one donor sequence required to be of genic origin (hence the classification as a "gene" duplicate). This definition can accommodate a variety of DNA-mediated mutational events leading to the formation of *chimeric* duplicates. A single duplication event partially encompassing two adjacent genes can instantaneously create a *chimeric* duplicate derived from the juxtaposition of two partial ancestral gene fragments. This ought to be a common mechanism of *chimeric* duplicate creation, as the gene duplication process appears to have little respect for gene boundaries [76, 133]. *Chimeric* duplicates can also be created via shuffling events that fuse together partially duplicated fragments of disparate ancestral origins (exonic, intronic and intergenic). Of course, to qualify as a *chimeric* gene duplicate, at least one of the ancestral donor sequences would have to be derived from a genic source. Figure 2(c) graphically represents the various types of *chimeric* duplicates derived from DNA-mediated duplication events. Figure 2(c) displays two duplicate copies (upper and lower gene copies are the ancestral and derived paralog, resp.) with sequence homology across exons 1 and 2 and terminating in intron 2 (shaded in green). The lower derived copy has a unique nonhomologous, terminal exon 3 (shaded in yellow). While the derived copy exhibits a superficial chimeric gene structure, a final classification is dependent on whether or not the unique terminal exon is derived from a duplication event. If the unique exonic sequence of the derived copy fails to generate any valid Blast hits in the genome (i.e., fails to identify a potential ancestral donor sequence), the duplicate should be classified as a *partial duplicate with recruitment* (top panel of Figure 2(c)). Alternatively, any significant hits to (i) intergenic, (ii) genic (exonic and/or intronic), or (iii) combination of intergenic and genic sequences in the genome would constitute evidence for its classification as a *chimeric* duplicate (lower three panels of Figure 2(c)).

3.4.1. Abundance of Chimeric Gene Duplicates in Genomes of Multicellular Eukaryotes. Among the three structural classes of gene duplication, *chimeric* gene duplicates are possibly the most challenging to classify accurately. *Partial duplicates with recruitment* superficially resemble *chimeric* duplicates and distinguishing between these two categories requires comparisons of exon-intron structure of both paralogs with a single-copy ortholog as well as additional investigations to further determine the existence of potential ancestral donor sequences for unique sequence tracts in the derived copy. In the absence of genome sequences of closely-related outgroup species to enable a comparative genomic approach to duplicate classification, early studies directly compared paralogous ORF sequences to indirectly estimate the frequency of *chimeric* duplicates. Katju and Lynch's study [76] of evolutionarily young *C. elegans* gene duplicates initially classified *chimeric* duplicates as comprising two paralogs of differing amino acid sequence length wherein sequence homology between the two copies was disrupted within the ORFs of

both copies, such that both had unique ORF sequence to the exclusion of the other copy. However, this approach will fail to distinguish *partial duplicates with recruitment* from *chimeric duplicates*. Indeed, in a subsequent study using a comparative genomic approach, 43% of a subset of 23 *C. elegans* gene duplicates previously characterized as *chimeric* duplicates in the absence of an outgroup sequence were found to constitute *partial duplicates with recruitment* [23]. In their study of *Drosophila melanogaster* gene duplicates, Zhou et al. [24] classified a derived paralog as a *chimeric* duplicate if it possessed a >50 bp nonhomologous sequence to the exclusion of the ancestral copy. This approach too suffers from the inability to distinguish *partial duplicates with recruitment* from *chimeric duplicates*. As such, some unknown fraction of *D. melanogaster chimeric* duplicates as identified by Zhou et al. [24] likely represent *partial duplicates with recruitment*.

Table 1 reports the percentage of *chimeric* duplicates in four species. It is highly likely that measures of 40% (116/290 duplicate pairs) *chimeric* duplicates within *C. elegans* [76] and 32% within *D. melanogaster* [24] are overestimates due to the misclassification of *partial duplicates with recruitment* as *chimeric* duplicates. Emerson et al.'s [45] calculation of 10% *chimeric* duplicates in natural isolates of *D. melanogaster* are derived from direct observation of *partial* duplication events across two adjacent genes leading to the formation of *chimeric* duplicates, and as such represent a *bona fide* conservative estimate of the frequency of *chimeric* duplicates within these genomes. Irrespective, these measures of *chimeric* and *partial* duplicates taken together underscore the widespread existence of structurally heterogeneous duplicates within evolutionarily young cohorts of gene duplicates in multicellular eukaryotic genomes. More specifically, they directly contradict Ohno's assumption that gene duplicates commence their evolutionary life redundant in sequence and function to their ancestral counterparts.

As was the case with *partial* duplicates, *chimeric* duplicates in *S. cerevisiae* are observed in extremely low frequency (4%) [78]. Taken together, structurally heterogeneous duplicates only comprise 11% of yeast duplicates derived from small-scale, DNA-mediated duplication events. This is in direct contrast to the genomes of multicellular eukaryotic species like *C. elegans* and *D. melanogaster* wherein structurally heterogeneous duplicates (*partials* and *chimerics*) comprise 56–86% of all duplicates (Table 1).

3.4.2. Evolutionary Potential of Chimeric Duplicates. A recognition of the evolutionary potential of chimeric genes is not new [134]. As is the case with *partial* duplicates, *chimeric* duplicates derived from the fusion of multiple duplicated frames of diverse genomic origins can play a significant role in the origin of evolutionary novelties. In *C. elegans*, *chimeric* duplicates were found to possess novel exons fashioned from diverse genomic sources including repetitive elements as well as exonic, intronic, and intergenic sequences [23]. Shuffling of fragments or domains can alter the regulation and functionality of the novel gene and facilitate its fixation at the species-level if the new function offers a selective advantage at the point of conception [135]. Because numerous examples of *chimeric* duplicates exist and have been extensively reviewed in preceding publications [132, 136, 137], for logistic purposes I restrict my discussion to a few cases.

Several *chimeric* duplicates have originated from a single duplication event that partially overlapped two adjacent genes. Incomplete duplication across two ORFs would appear to entail a high probability of creating a degenerated novel ORF marked for a nonfunctionalizing fate. However, it appears that *chimeric* duplicates in *Drosophila*, a genus in which they are particularly well-studied, have appreciably lower rates of origin (~11 duplicates/my) relative to other DNA-mediated duplications (~80 duplicates/my) but are equally liable to be preserved in the genome [138]. It is also of considerable biological interest to elucidate if such *chimeric* duplicates created by the fusion of two ancestral genes are (i) equally divergent in function from either ancestral gene, or (ii) possess a function that resembles that of both ancestral genes, or (iii) disproportionately resemble one ancestral gene's function.

The *Sdic* cluster represents an interesting example of a *chimeric* duplicate in *D. melanogaster* [22]. Although derived from an incomplete duplication event across two independent ancestral genes, the mutational events altering its genomic organization subsequent to its formation resemble that of the *partial* duplicate AFPIII cluster [110]. The *Sdic* gene was created by an incomplete duplication event spanning the latter half of an upstream cytoplasmic dynein gene *Cdic* and the N-terminal region of its adjacent downstream neighbour *AnnX* which encodes for an annexin protein. Several internal deletions subsequent to the duplication event refashioned a novel ORF which was duplicated multiple times to form a tandem array of ~10 copies. Interestingly, *Sdic*'s function resembles that of its *Cdic* progenitor in that it too encodes for a dynein, except one whose expression profile has been substantially narrowed and altered to be testis-specific [22]. The *Qtzl* gene in *D. melanogaster* represents a recently derived *chimeric* duplicate created via incomplete duplication across adjacent ancestral genes *CG12264* and *escl* [139]. While the exact function of *Qtzl* remains to be ascertained, it exhibits a strong molecular signature of preservation by natural selection, namely a drastically reduced level of genetic diversity in its genomic location and rapid fixation across 35 natural isolates of *D. melanogaster* despite its recent evolutionary origin [139]. It is also interesting to note that the expression profile of *Qtzl* disproportionately resembles that of its *escl* parent despite the observation that the ancestral *escl* donor sequence was inherited out of frame. Therefore, it appears that both *Sdic* and *Qtzl* described above display (i) substantial narrowing of their spatial expression profiles and (ii) appear to have functional roles that disproportionately resemble one of the two parental genes contributing to their chimeric origin.

Opazo et al. [140] report on an intriguing example of a *chimeric* duplicate derived from the fusion of two ancestral globin genes that has proceeded to functionally supplant one of its parental genes. A proto β-globin gene duplicated in the common ancestor of eutherian mammals following

divergence from marsupials to generate the adjacent paralogs HBB (β-globin) and HBD (δ-globin) [141–143]. Most eutherian lineages have seen the deletion or degeneration of the HBD paralog with functional haemoglobin products encoded for by one or more HBB genes [143]. Paenungulate mammals comprising the three orders of Proboscidea (elephants), Sirenia (dugongs and manatees), and Hyracoidea (hyraxes) possess a chimeric HBB/HBD (β/δ) globin gene derived from a duplication event across the HBB and HBD paralogs. The parental HBB and HBD genes have both been pseudogenized by a N-terminal deletion and 3.2 kb insertion, respectively. The *chimeric* HBB/HBD duplicate in paenungulate mammals has assumed the functional role of its ancestral HBB gene and encodes for the β-chain subunits of adult haemoglobin.

4. RNA-Mediated Duplication Events

Retrotransposition is another dominant mechanism facilitating the creation of gene duplicates. Such RNA-mediated duplication events, also referred to as retroduplications, occur when spliced messenger RNA of an ancestral locus is reverse transcribed into cDNA and then reinserted into a novel genomic position. Gene duplication by retrotransposition instantaneously creates a duplicate gene with diverged characteristics from its progenitor locus [144–146]. First, retroduplication typically creates a single-exon gene duplicate from a multiexonic ancestral gene. Second, because retroduplication only encompasses transcribed sequences, the duplicate copy inherently lacks the ancestral repertoire of regulatory elements that control the expression of its progenitor locus. Preservation of a functional retrocopy (often referred to as "retrogenes" or "processed genes") is then dependent on the retrocopy's ability to fortuitously recruit a novel promoter and other key *cis*-regulatory elements. Third, retrocopies are randomly inserted into novel genomic locations and as such inherit a genomic environment characterized by a complete disruption of ancestral synteny and the gain of new neighbourhood genes. These drastic alterations to the ancestral gene structure and genomic environment can engender the evolution of a radically novel gene if the retrocopy can escape the associated high risk of pseudogenization.

The typical outcome of duplication via retrotransposition is thought to be the creation of a single-exon gene duplicate from a multiexonic ancestral gene (Figure 3(a)). Duplication by retrotransposition is implicated by the presence of several diagnostic features in a processed retrocopy, namely (i) the lack of ancestral introns, (ii) an absence of the ancestral upstream promoter region, (iii) coincident boundaries with the ancestral transcribed regions, (iv), a polyadenylation signal followed by a short poly(A) tail at the 3′ end, (v) the presence of flanking direct repeats, and (vi) a novel genomic location. In some instances, retrotransposition of a partially processed pre-mRNA transcript leads to a semiprocessed retrocopy wherein some ancestral introns and flanking region sequence are left intact (Figure 3(b)). Most importantly, these retrocopies can recombine with

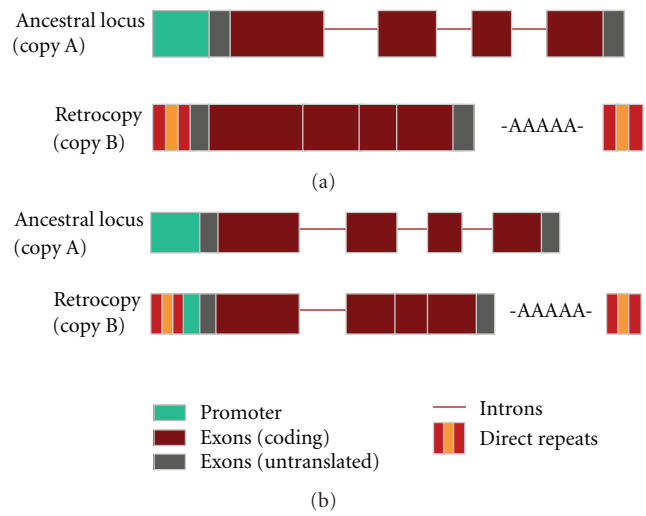


Figure 3: The creation of two classes of retrocopies via RNA-mediated duplication events. (a) Processed copies are created when a spliced mRNA of an ancestral locus is reverse transcribed into cDNA, leading to the creation of a single-exon duplicate from a multiexonic ancestral gene. The processed copy lacks all ancestral introns and the ancestral promoter, while possessing a poly(A) tail at the 3′ end and flanking direct repeats at both the 5′ and 3′ end. (b) A semiprocessed copy can be formed via retrotransposition of a partially-processed pre-mRNA transcript leading to the inheritance of some ancestral introns and flanking region sequence. In the schematic, the retrocopy possesses a poly(A) tail and direct flanking repeats. It lacks introns 2 and 3 but has inherited the ancestral intron 1 and a portion of the ancestral promoter.

other duplicated fragments derived from DNA-mediated duplication events or exonize flanking region sequences to create even more drastically altered ORFs. For logistic reasons, I classify these hybrid genes derived from both DNA- and RNA-mediated duplication events as retrocopies though they may need to be categorized as a different class of gene duplicates in the future.

4.1. Genomes Vary in the Extent of RNA-Mediated Duplications. RNA-mediated duplication events are certainly common in fly, mammalian, marsupial, and grass genomes as evidenced by the presence of a multitude of retrocopies ([147–152] among others) but are less frequent in birds. *C. elegans* and monotremes [23, 153, 154]. This variation in the frequency of retrogenes among phylogenetically diverse lineages is thought to be contingent on the presence/absence of key enzymes involved in retrotransposition and their activity in the germline in order to enable fidelity of inheritance [154]. A study of 290 evolutionarily young *C. elegans* gene duplicates identified a mere three duplicate pairs wherein one paralog lacked intron(s) relative to the other copy [76]. In a subsequent study, two of these three cases were confirmed as having originated via retrotransposition whereas one case represented intron gain by the derived copy [23], suggesting that 99.3 and 0.7% of paralogs belonging to small gene families in *C. elegans* owe their origins to

DNA- and RNA-mediated duplication events, respectively. In *D. pseudoobscura*, 37% of *complete* duplicates were classified as possibly originating via retrotransposition or ambiguous events [77]. RNA-mediated duplication events were implicated in the creation of approximately 7% of evolutionarily young *D. melanogaster* paralogs, and in ~ 10% of all duplication events in the *D. melanogaster* species complex [24].

4.2. Rapidly Emerging Data Highlights the Immense Evolutionary Potential of Retrocopies. A recognition of the evolutionary potential of retrogenes derived from RNA-mediated duplication events has been slow in coming despite the fact that they, akin to structurally heterogeneous DNA duplicates, can facilitate the creation of novel genes with radically altered ORFs. Lead proponents who have championed the importance of DNA-mediated *partial* and *chimeric* duplications in generating raw material for the origin of novel biochemical functions have been far more skeptical about the importance of retrocopies in evolution. Patthy [75] professed it very unlikely that processed genes (retrocopies) have much to contribute to the origin of novel genes, given that their chance of survival is severely diminished due to the drastic loss of regulatory features at birth. As such, retrocopies were systematically referred to as processed pseudogenes or retropseudogenes [155, 156].

The last decade has seen the identification of numerous functional retrogenes in diverse lineages. Detailed sequence and functional analyses of these retrogenes have demonstrated them as remarkably adept at recruiting novel regulatory elements and other genic fragments to emerge as mosaic genes conferring a myriad of novel functions. The function of these retrogenes has been best studied in the *Drosophila* lineage and there are several excellent reviews that provide detailed information about their origins and trajectories leading to functional diversification [132, 136, 154, 157, 158]. Rather, I will highlight a few examples that represent the diversity of mechanisms that enable retrocopies to persevere in genomes and evolve novel functions. Retrogenes exhibit a spectrum in the degree of functional divergence from their ancestral gene sources. At one end of the spectrum, some retrogenes evolve to function in a capacity similar to the ancestral gene, with relatively minor modifications to their spatial and temporal expression patterns despite gain of novel exons and promoters from their new genomic environment. At the opposing end of the spectrum and more in line with biological expectations, certain retrogenes gain drastically altered biological functions that appear wholly unrelated to that of their ancestral counterparts.

Two examples presented below represent retrogenes created "functional on arrival" due to the fortuitous inheritance of ancestral promoters during retrotransposition. The murine preproinsulin I gene is derived from the ancestral preproinsulin II which houses one intron in the 5′ untranslated region and another within the coding region. Preproinsulin I is an example of a semiprocessed retrocopy (Figure 3(b)) derived from a partially processed pre-mRNA of preproinsulin II that included the intron in the 5′ untranslated region and ancestral upstream regulatory sequence which enabled its expression following integration into a novel genomic location [159]. Another intriguing example is exemplified by the origin of *PGK-2*. The human *PGK-1* (phosphoglycerate kinase) is an ancestral X-linked gene comprising 11 exons and 10 introns that encodes for an enzyme involved in the metabolism of glucose to pyruvate. Its autosomal paralog *PGK-2* is a retrogene lacking all ancestral introns that shows testis-specific expression in the late stages of spermatogenesis [160, 161]. The X-linked *PGK-1* is inactivated in spermatogenic cells prior to meiosis. However, mature spermatozoa need significant amounts of phosphoglycerate kinase to metabolize fructose present in semen. The inactivation of the single X-chromosome in spermatogenic cells before meiosis is thought to have created the need for a functional autosomal gene copy with a capacity for expression in the testis where the X is inactivated, a role that was fulfilled by the random creation of the *PGK-2* retrocopy. Most interestingly, the *PGK-2* retrocopy was born functional given that it initially included a copy of the ancestral promoter; only later did it evolve a testis-specific promoter [144, 145]. The preservation of *PGK-2* was favoured by selection for a compensatory response to the inactivation of its progenitor copy. This study has instigated widespread research into what appears to be a common phenomenon in mammals [152, 162] and *Drosophila* [163]—the migratory pattern of X-linked housekeeping genes to autosomal locations via retrotransposition in a bid to escape transcriptional inactivation of X-linked genes in the male germline during meiosis under the influence of natural selection [164].

The gain of novel functions by retrocopies is also facilitated by their commonly observed fusion with existent gene duplicates derived from DNA-mediated duplication events [165, 166]. The chimeric retrogene *jingwei* in *Drosophila tessieri* and *D. yakuba* was created by retrotransposition of the *Adh* gene, with subsequent insertion of the *Adh* retrosequence into the third intron of a duplicate gene *Yande* (derived from a DNA-mediated *complete* duplication of the *Yellow emperor* gene). The insertion of the single-exon *Adh* retrosequence into *Yande* led to the degeneration of *Yande*'s nine terminal exons, and the origin of the novel gene *jingwei* comprising three *Yande* exons and the single-exonic *adh* retrocopy. This new gene functions as a novel dehydrogenase with increased specificity for long-chain alcohol substrates and a narrowed breadth of expression pattern relative to its ancestor *Adh* [167]. *Adh-Twain* in *D. guanche*, *D. madeirensis,* and *D. subobscura* represents another independent evolutionary formation of a novel gene derived from the fusion of an *Adh* retrocopy and an existing paralog of the *GAPDH* gene labeled as *CG9010* [168]. Unlike *jingwei*, *Adh-Twain* does not appear to have had a major shift in its expression pattern, instead displaying a broad expression pattern similar to its ancestor *Adh* [169].

Retrocopy insertion into a novel genomic location and subsequent exonization of noncoding sequence from its new genomic neighbourhood can yield an ORF with the potential to bestow radically novel functions. The formation of the *Rps23* retrogene in mice via this mechanism has conferred

increased resistance to the progression of Alzheimer-causing amyloid plaques, a function quite divergent from the ribosomal protein role of its progenitor copy [170].

5. Escaping the Tether of Gene Conversion

Ohno's canonical model of gene duplicate evolution [16] posits that gene duplicates bearing *complete* sequence and functional redundancy gradually accumulate mutations leading to alternative fates of neofunctionalization or nonfunctionalization with eventual loss from the genome. This model of paralog evolution is overly simplistic because we know paralogs to be capable of nonreciprocal recombination with each other via ectopic (interlocus) gene conversion. Gene conversion is a form of concerted evolution wherein a donor sequence converts a homologous recipient sequence over some length of its tract leading to increased sequence homogeneity between the two paralogs. Hence, gene conversion acts as an effective tether constraining sequence and predictably, functional diversification between paralogs. The evolutionary trajectories of gene duplicates subsequent to their formation is thus governed by two opposing forces; sequence divergence by new mutations and repeated erosion of this achieved sequence heterogeneity via gene conversion [135, 171, 172]. Gene conversion is a ubiquitous process leading to sequence homogenization of paralogs across virtually all organisms that have been subject to detailed enquiries, from microbes to vertebrates [103, 124, 172–179].

Gene conversion has substantial bearing on the functional fate of gene duplicates. Although we currently lack accurate experimental estimates of the rate of spontaneous ectopic gene conversion between paralogs from mutation accumulation lines that are severely bottlenecked each generation to reduce the efficacy of natural selection, the frequent and independent origin of phenotypes associated with gene conversion events in experimentally evolved lines [113] and detectable signatures of gene conversion among genome-wide studies of paralogs [178] certainly implicate a high rate of ectopic gene conversion. Under environmental regimes where an increased gene dosage of an ancestral protein product is beneficial, natural selection is expected to favour the maintenance of a *complete* structural resemblance and sequence homogeneity between paralogs via gene conversion [180–182]. On the flip side, if spontaneous gene conversion events between paralogs occur at an appreciable frequency, how are paralogs able to escape the evolutionary tether of sequence homogenization by gene conversion to achieve neofunctionalized states? We know that with increasing sequence divergence, the frequency of gene conversion between paralogs is expected to taper off, thereby increasing the probability of functional divergence between paralogs [171]. However, how is this threshold of sequence divergence between paralogs ever achieved in the first place under the constant onslaught of gene conversion? This is especially pertinent for duplicates residing in genomic proximity, given substantial evidence that closely-spaced paralogs experience a higher frequency of gene conversion events [177, 178, 183–185].

Walsh [171] was the first to theoretically explore the conundrum of gene duplicate neofunctionalization in the face of gene conversion pressure. He suggested that "terminator mutations" such as large indels, mobile element insertion and translocation of one paralog to a novel genomic location via retrotransposition may provide the necessary break in sequence homology between paralogs to retard the frequency of gene conversion between them. It is apparent from several studies that the movement of one paralog to another chromosome promotes sequence divergence between the two copies [146] though it is not clear whether this is derived from reduced gene conversion pressure or the inheritance of a novel genomic environment by the paralog. More recently, Innan explored the role of diversifying natural selection in the maintenance of paralog sequence diversity under the pressure of gene conversion [172]. The patterns of DNA variation in human antigen-coding paralogs *RHCE* and *RHD* appear consistent with a model of selection maintaining antigen diversity despite frequent gene conversion, although the strength of selection required to counterbalance homogenization by gene conversion was inferred to be extremely high. Deeb et al. [103] found that despite frequent gene conversion between the X-linked red and green opsin paralogs of Old World primates, certain codons coding for amino acid residues implicated in the separation of peak absorbance between the two pigments were left intact within each paralog thereby implying a role of natural selection in counterbalancing gene conversion.

I additionally suggest that structural heterogeneity among paralogs inherited at birth (as in *partial* and *chimeric* duplicates and retrocopies) plays a very important role in restricting complete homogenization of paralogs via gene conversion, thereby promoting neofunctionalization in addition to the fact that these novel sequences encode novel amino acids. If the unique coding regions in one or both paralog(s) encode novel functional domains, neofunctionalization could be promoted despite ongoing gene conversion in their homologous regions (Figure 4). As such, the creation of a structurally heterogeneous paralog by gene duplication immediately confers on the derived copy a "terminator mutation" as envisioned by Walsh [171] that serves to diminish the homogenizing effects of gene conversion. As a case and point, I revisit the creation of the *fog-2* gene in *C. elegans* from an ancestral gene of unknown function, *ftr-1* (discussed earlier in Section 3.3.2). *fog-2*, implicated in the origin of hermaphroditism in *C. elegans* likely originated from a *partial duplication with recruitment* event resulting from the incomplete duplication of *ftr-1* that prematurely terminated in the terminal exon of *ftr-1*, and subsequently exonized noncoding sequence from its new genomic neighborhood to complete its ORF [113]. Intriguingly, the recruitment of this unique sequence in the 3′ end of *fog-2* likely facilitated its neofunctionalization after duplication [124]. Frequent gene conversions of *fog-2* by *ftr-1* in both experimentally evolved and wild *C. elegans* populations fail to diminish or compromise the function of *fog-2* in hermaphrodite spermatogenesis, given that the neofunctionalized sequence tract in *fog-2* was created by

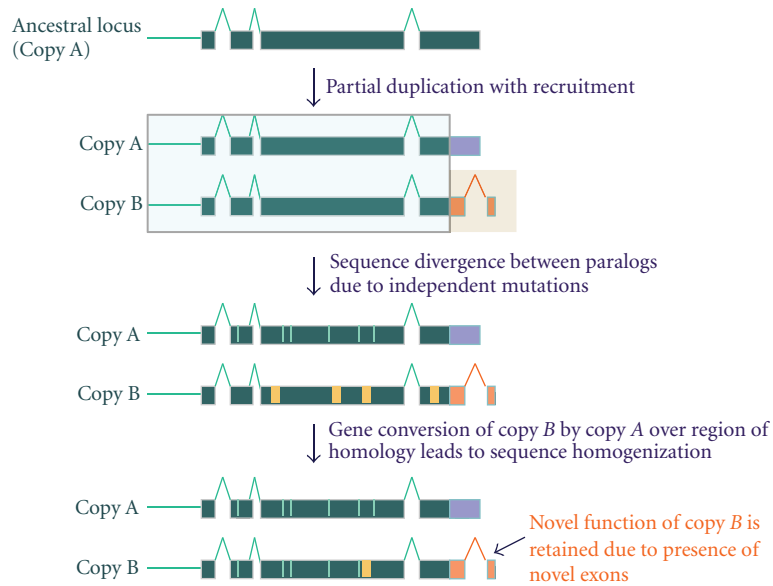


Figure 4: Schematic outlining how structurally heterogeneous duplicates can neofunctionalize by escaping gene conversion based on the *fog-2/ftr-1* case in *C. elegans* [113, 124, 185]. Consider an ancestral gene (Copy A) comprising four exons (top panel) that is partially duplicated. Small shaded rectangles represent exons. The duplication event is initiated in the 5′ flanking region and terminates within exon 4. The large transparent blue rectangle highlights the region of homology between the paralogs as do like coloured exons and introns. The derived copy B additionally recruits noncoding sequence from its genomic neighbourhood to fashion a novel partial exon 4, intron 4 and exon 5 that bears no sequence homology to ancestral copy A (depicted by orange rectangles and highlighted by a large transparent cream rectangle). This novel recruited region also imparts a novel function to copy B. The two paralogs diverge by gradual accumulation of mutations (horizontal narrow lines within exons delineate point mutations). Recurrent episodes of gene conversion of copy B (recipient) by copy A (donor) constrains sequence divergence across the region of homology. Despite high gene conversion pressure, copy B is able to preserve its neofunctionalized state due to the presence of nonhomologous coding sequence that remains unconstrained by gene conversion.

the exonization of novel noncoding sequence and bears no homology to the *ftr-1* sequence.

6. The Influence of Effective Population Size (N_e)

Population-genetic theory predicts that the ultimate fate of a mutation (in our case, duplication), be it eventual fixation or loss in a population, depends on the efficacy of natural selection. The efficacy of natural selection (or selection intensity), in turn, depends on the product of (i) the selection coefficient (s) of the mutation (also known as the fitness effect of a mutation) and (ii) the effective population size (N_e) of the species. Therefore, the intensity of effectiveness of selection is expressed as $N_e s$ [186–189]. A decreased intensity of selection could therefore result from either a smaller s or a decreased N_e. As such, the genomes of prokaryotes and unicellular eukaryotes with extremely large N_e are expected to experience far greater efficacy of selection than those of multicellular eukaryotes with significantly smaller N_e. The disparity in N_e across the transitions from prokaryotes to unicellular eukaryotes to multicellular eukaryotes can span several orders of magnitude, from $>10^7$ for prokaryotic species and $\sim 10^4$ for larger vertebrates (Table 3).

Lynch and colleagues have posited the provocative hypothesis that the historically lower N_e of multicellular eukaryotes with their concomitant reduced efficacy of selection have provided a permissive environment for the accumulation of certain key elements of genomic architecture that would otherwise be eliminated in genomes more effectively patrolled by purifying selection [81, 190, 191]. Extending this argument, it may be hypothesized that the longer persistence time for such initially nonadaptive genetic elements in eukaryotic species with small N_e could enhance the probability of future exaptation to novel biochemical functions at a later evolutionary stage, leading to the emergence of biological complexity from initially nonadaptive processes.

Given the accumulating evidence that structurally heterogeneous gene duplicates (*partial* and *chimeric*) as well as retrogenes can confer radically novel functions, their near absence in the sequenced genomes of species with large N_e such as *S. cerevisiae* remains a puzzle. Undoubtedly, partially duplicated fragments are less likely to originate in small, compact genomes with shorter genes and fewer, smaller introns. However, spontaneous segmental duplications do originate frequently in yeast [37] and because the gene duplication process appears largely irreverent to gene boundaries, terminal loci within a segmental duplicate fragment should have a higher probability of being partially duplicated. Most gene duplicates may be slightly deleterious when born and likely confer a slight penalty on the fitness of their carriers by creating a minor dosage imbalance. Given the large N_e in microorganisms and unicellular eukaryotes, gene duplicates

Table 3: Estimates of effective population size (N_e) for a sampling of species.

Species	N_e	References
Prokaryotes		
Escherichia coli	25,000,000	[192]
Unicellular Eukaryotes		
Paramecium species	25,000,000–75,000,000	[193]
Plasmodium falciparum	210,000–300,000	[194]
Saccharomyces paradoxus	10,000,000	[195]
Multicellular Eukaryotes		
Invertebrates		
Caenorhabditis elegans	80,000	[196]
Caenorhabditis remanei	1,600,000	[197]
Drosophila melanogaster	1,150,000	[198]
Drosophila simulans	2,600,000	[199]
Plants		
Arabidopsis lyrata	138,000	[200]
Arabidopsis thaliana	127,000	[200]
Capsella grandiflora	500,000	[201]
Helianthus annuus	832,000	[200]
Helianthus petiolaris	733,000	[200]
European aspen *Populus tremula*	118,000–500,000	[202, 203]
Zea mays	590,000	[200]
Vertebrates		
Mus domesticus	161,000	[199]
Mus castaneus	500,000	[204]
Bonobos	12,300	[205]
Chimpanzee	21,300	[205]
Human	10,400	[205]
Gray Whale	34,410	[206]

bearing even slightly negative selective coefficients may be efficiently purged from these genomes. *Partial* duplicates may initially be at a greater selective disadvantage than *complete* duplicates given that the majority of them likely originate lacking function, and are therefore more prone to eradication in these genomes. And might their efficient eradication impose limits to future phenotypic evolution in these species?

These nonadaptive hypotheses certainly warrant further testing as a null model before invoking the ubiquitous guidance of natural selection in the origin of adaptive phenotypes via gene duplication. This is not to say that gene duplicates bearing a great selective advantage at birth are not existent. Like any other class of mutation, gene duplicates can be born advantageous, neutral or deleterious. However, collectively speaking, do different structural classes vary with respect to their fitness effects and how might this, in conjunction with the species N_e, impinge on their future evolutionary trajectories? As a first step, mutation accumulation (MA) lines subjected to attenuated selection via repeated bottlenecking provide the best means to investigate the spontaneous rates of occurrence of different structural classes of gene duplicates within phylogenetically diverse genomes and infer the evolutionary forces that govern their subsequent preservation or demise. As discussed earlier, the paucity of SSD-originated *partial* and *chimeric* duplicates in the first yeast genome to be sequenced could be due to lower rates of origin and/or higher probabilities of eradication via natural selection. The characterization of gene duplicates arising in yeast MA lines evolved under conditions of reduced efficacy of selection would enable an accurate determination of the spontaneous rates of origin of different structural classes of gene duplicates. If *partial* and *chimeric* duplicates occur at a significantly higher frequency in the MA lines relative to sequenced genomes not subjected to MA treatment, it would provide evidence for eradication of such duplicates via purifying selection in the latter. We could additionally infer that structurally heterogeneous classes of duplicates, collectively speaking, are more likely to be deleterious relative to *complete* duplicates. In the case of *C. elegans*, the frequency spectrum of structurally homogeneous (*complete* duplicates) and heterogeneous (pooled *partial* and *chimeric*) duplicates in the genomes of MA lines is remarkably concordant with that of the originally sequenced N2 strain [38, 76]. There was an absolute absence of detectable *chimeric* duplicates in the MA lines but this likely reflects the limited diagnostic ability of array Comparative Genomic Hybridization (aCGH) techniques to detect *chimeric* duplicates. The concordant frequencies of these structural classes in the MA lines and the N2 strain strongly suggests that *partial*/*chimeric* duplicates are not subject to greater purifying selection in the *C. elegans* genome, given the relatively low N_e for this species (Table 3).

7. Conclusions

The plethora of genomic sequence data has facilitated tremendous advances in our understanding of the gene duplication process. The high frequencies of structurally heterogeneous gene duplicates in many lineages bear direct testament to the inherent promiscuity of the gene duplication process and contribute directly to its potential for rapidly generating novel genes implicated in the emergence of biological innovations. The identification of these structurally heterogeneous duplicates with known novel functions additionally demonstrates that Ohno's canonical model of gene duplicate evolution only represents one of multiple routes that can be assumed by gene duplicates during their evolution. Future investigations should focus on elucidating the relative roles of selection versus random genetic drift in the evolution of new genes via duplication. More importantly, we need to further investigate how the degree of structural resemblance between duplicates and their progenitors impinges on their evolutionary constraints and opportunities in evolution. *Complete* duplicates by virtue of their structural similarity to ancestral genes may be bound to function within the phenotypic bounds of their ancestral counterparts. In contrast, retrocopies and *partial* and *chimeric* duplicates, although more likely to be

nonfunctional at birth, may bear greater potential to assume radically novel functions due to the inheritance of novel coding and regulatory elements.

The detailed structural characterization of extant paralogs across phylogenetically diverse genomes would serve to elucidate (i) the various mutational mechanisms responsible for the creation of gene duplicates, (ii) the relative abundance of different structural classes of gene duplicates, (iii) the relative contribution of diverse genomic sequences to the creation of novel genes, (iv) the relative survivorship of different classes of gene duplicates across different age-cohorts of gene duplicates and in different genomic backgrounds, and (v) whether these patterns vary across taxa or display phylogenetic independence.

Acknowledgments

This paper was improved by valuable suggestions from an anonymous reviewer. The author is additionally grateful to Margarida Cardoso-Moreira and J. J. Emerson for their help with data interpretation at short notice and to Ulfar Bergthorsson for insightful comments. The author is especially indebted to the Editor Frédéric Brunet for the supreme patience he has shown and for the encouragement. This work was supported by National Science Foundation Grant DEB-0952342.

References

[1] A. H. Sturtevant, "The effects of unequal crossing over at the bar locus in *Drosophila*," *Genetics*, vol. 10, no. 2, pp. 117–147, 1925.

[2] C. B. Bridges, "Salivary chromosome maps: with a key to the banding of the chromosomes of *Drosophila melanogaster*," *Journal of Heredity*, vol. 26, no. 2, pp. 60–64, 1935.

[3] J. B. S. Haldane, "The part played by recurrent mutation in evolution," *American Naturalist*, vol. 67, no. 708, pp. 5–19, 1933.

[4] H. J. Müller, "The origination of chromatin deficiencies as minute deletions subject to insertion elsewhere," *Genetica*, vol. 17, no. 3-4, pp. 237–252, 1935.

[5] H. J. Müller, "Bar duplication," *Science*, vol. 83, no. 2161, pp. 528–530, 1936.

[6] J. Huxley, *Evolution: The Modern Synthesis*, Allen and Unwin, London, UK, 1942.

[7] L. J. Stadler and H. Roman, "The effect of X-rays upon mutation of the gene-A in maize," *Genetics*, vol. 33, no. 3, pp. 273–303, 1948.

[8] O. H. Frankel, "A self-propagating structural change in *Triticum*: I. Duplication and crossing-over," *Heredity*, vol. 3, pp. 163–194, 1949.

[9] M. Y. Menzel and M. S. Brown, "Viable deficiency-duplications from a translocation in *Gossypium hirsutum*," *Genetics*, vol. 37, no. 6, pp. 678–692, 1952.

[10] F. J. Ratty, "Gene action and position effect in duplications in *Drosophila melanogaster*," *Genetics*, vol. 39, no. 4, pp. 513–528, 1954.

[11] J. W. Lesley and M. M. Lesley, "Effect of seed treatments with X-ray and phosphorus-32 on tomato plants of 1st, 2nd, and 3rd generations," *Genetics*, vol. 41, no. 4, pp. 575–588, 1956.

[12] B. H. Judd, "Formation of duplication-deficiency products by asymmetrical exchange within a complex locus of *Drosophila melanogaster*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 47, pp. 545–550, 1961.

[13] O. Smithies, G. E. Connell, and G. H. Dixon, "Chromosomal rearrangements and the evolution of haptoglobin genes," *Nature*, vol. 196, no. 4851, pp. 232–236, 1962.

[14] W. M. Fitch, "Evidence suggesting a partial, internal duplication in the ancestral gene for heme-containing globins," *Journal of Molecular Biology*, vol. 16, no. 1, pp. 9–16, 1966.

[15] J. A. Black and G. H. Dixon, "Evolution of protamine: a further example of partial gene duplication," *Nature*, vol. 216, no. 5111, pp. 152–154, 1967.

[16] S. Ohno, *Evolution by Gene Duplication*, Springer, Berlin, Germany, 1970.

[17] R. F. Furlong and P. W. H. Holland, "Polyploidy in vertebrate ancestry: Ohno and beyond," *Biological Journal of the Linnean Society*, vol. 82, no. 4, pp. 425–430, 2004.

[18] P. Dehal and J. L. Boore, "Two rounds of whole genome duplication in the ancestral vertebrate," *PLoS Biology*, vol. 3, no. 10, Article ID e314, 2005.

[19] M. Kasahara, "The 2R hypothesis: an update," *Current Opinion in Immunology*, vol. 19, no. 5, pp. 547–552, 2007.

[20] S. Ohno, "So much "junk" DNA in our genome," *Brookhaven Symposia in Biology*, vol. 23, pp. 366–370, 1972.

[21] L. Chen, A. L. DeVries, and C.-H. C. Cheng, "Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 8, pp. 3811–3816, 1997.

[22] D. I. Nurminsky, M. V. Nurminskaya, D. de Aguiar, and D. L. Hartl, "Selective sweep of a newly evolved sperm-specific gene in *Drosophila*," *Nature*, vol. 396, no. 6711, pp. 572–575, 1998.

[23] V. Katju and M. Lynch, "On the formation of novel genes by duplication in the *Caenorhabditis elegans* genome," *Molecular Biology and Evolution*, vol. 23, no. 5, pp. 1056–1067, 2006.

[24] Q. Zhou, G. Zhang, Y. Zhang et al., "On the origin of new genes in *Drosophila*," *Genome Research*, vol. 18, no. 9, pp. 1446–1455, 2008.

[25] M. T. Levine, C. D. Jones, A. D. Kern, H. A. Lindfors, and D. J. Begun, "Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 26, pp. 9935–9939, 2006.

[26] J. Cai, R. Zhao, H. Jiang, and W. Wang, "De novo origination of a new protein-coding gene in *Saccharomyces cerevisiae*," *Genetics*, vol. 179, no. 1, pp. 487–496, 2008.

[27] D. G. Knowles and A. McLysaght, "Recent de novo origin of human protein-coding genes," *Genome Research*, vol. 19, no. 10, pp. 1752–1759, 2009.

[28] M. Toll-Riera, N. Bosch, N. Bellora et al., "Origin of primate orphan genes: a comparative genomics approach," *Molecular Biology and Evolution*, vol. 26, no. 3, pp. 603–612, 2009.

[29] M. T. Donoghue, C. Keshavaiah, S. H. Swamidatta, and C. Spillane, "Evolutionary origins of Brassicaceae specific genes in *Arabidopsis thaliana*," *BMC Evolutionary Biology*, vol. 11, no. 2, article 47, 2011.

[30] M. P. Eichenlaub and L. Ettweiler, "De novo genesis of enhancers in vertebrates," *PLoS Biology*, vol. 9, no. 11, Article ID e1001188, 2011.

[31] D. D. Wu, D. M. Irwin, and Y. P. Zhang, "De novo origin of human protein-coding genes," *PLoS Genetics*, vol. 7, no. 11, Article ID e1002379, 2011.

[32] S. Ohno, J. Muramoto, L. Christian, and N. B. Atkin, "Diploid-tetraploid relationship among old-world members of the fish family Cyprinidae," *Chromosoma*, vol. 23, no. 1, pp. 1–9, 1967.

[33] K. Bender and S. Ohno, "Duplication of the autosomally inherited 6-phosphogluconate dehydrogenase gene locus in tetraploid species of cyprinid fish," *Biochemical Genetics*, vol. 2, no. 2, pp. 101–107, 1968.

[34] J. Klose, U. Wolf, H. Hitzeroth, H. Ritter, N. B. Atkin, and S. Ohno, "Duplication of the LDH gene loci by polyploidization in the fish order Clupeiformes," *Human Genetics*, vol. 5, no. 3, pp. 190–196, 1968.

[35] S. Ohno, U. Wolf, and N. B. Atkin, "Evolution from fish to mammals by gene duplication," *Hereditas*, vol. 59, no. 1, pp. 169–187, 1968.

[36] S. Ohno, "The role of gene duplication in vertebrate evolution," in *The Biological Basis of Medicine*, E. D. Bittar and N. Bittar, Eds., vol. 4, chapter 4, pp. 109–132, London Academic Press, 1969.

[37] M. Lynch, W. Sung, K. Morris et al., "A genome-wide view of the spectrum of spontaneous mutations in yeast," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 27, pp. 9272–9277, 2008.

[38] K. J. Lipinski, J. C. Farslow, K. A. Fitzpatrick, M. Lynch, V. Katju, and U. Bergthorsson, "High spontaneous rate of gene duplication in *Caenorhabditis elegans*," *Current Biology*, vol. 21, no. 4, pp. 306–310, 2011.

[39] J. Sebat, B. Lakshmi, J. Troge et al., "Large-scale copy number polymorphism in the human genome," *Science*, vol. 305, no. 5683, pp. 525–528, 2004.

[40] G. H. Perry, J. Tchinda, S. D. McGrath et al., "Hotspots for copy number variation in chimpanzees and humans," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 21, pp. 8006–8011, 2006.

[41] R. Redon, S. Ishikawa, K. R. Fitch et al., "Global variation in copy number in the human genome," *Nature*, vol. 444, no. 7118, pp. 444–454, 2006.

[42] E. B. Dopman and D. L. Hartl, "A portrait of copy-number polymorphism in *Drosophila melanogaster*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 50, pp. 19920–19925, 2007.

[43] T. A. Graubert, P. Cahan, D. Edwin et al., "A high-resolution map of segmental DNA copy number variation in the mouse genome," *PLoS Genetics*, vol. 3, no. 1, Article ID e3, 2007.

[44] L. Carreto, M. F. Eiriz, A. C. Gomes, P. M. Pereira, D. Schuller, and M. A. S. Santos, "Comparative genomics of wild type yeast strains unveils important genome diversity," *BMC Genomics*, vol. 9, article 524, 2008.

[45] J. J. Emerson, M. Cardoso-Moreira, J. O. Borevitz, and M. Long, "Natural selection shapes genome-wide patterns of copy-number polymorphism in *Drosophila melanogaster*," *Science*, vol. 320, no. 5883, pp. 1629–1631, 2008.

[46] A. S. Lee, M. Gutiérrez-Arcelus, G. H. Perry et al., "Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies," *Human Molecular Genetics*, vol. 17, no. 8, pp. 1127–1136, 2008.

[47] G. H. Perry, F. Yang, T. Marques-Bonet et al., "Copy number variation and evolution in humans and chimpanzees," *Genome Research*, vol. 18, no. 11, pp. 1698–1710, 2008.

[48] X. She, Z. Cheng, S. Zöllner, D. M. Church, and E. E. Eichler, "Mouse segmental duplication and copy number variation," *Nature Genetics*, vol. 40, no. 7, pp. 909–914, 2008.

[49] W. K. Chen, J. D. Swartz, L. J. Rush, and C. E. Alvarez, "Mapping DNA structural variation in dogs," *Genome Research*, vol. 19, no. 3, pp. 500–509, 2009.

[50] T. J. Nicholas, Z. Cheng, M. Ventura, K. Mealey, E. E. Eichler, and J. M. Akey, "The genomic architecture of segmental duplications and associated copy number variants in dogs," *Genome Research*, vol. 19, no. 3, pp. 491–499, 2009.

[51] N. M. Springer, K. Ying, Y. Fu et al., "Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content," *PLoS Genetics*, vol. 5, no. 11, Article ID e1000734, 2009.

[52] G. E. Liu, Y. Hou, B. Zhu et al., "Analysis of copy number variations among diverse cattle breeds," *Genome Research*, vol. 20, no. 5, pp. 693–703, 2010.

[53] J. S. Maydan, A. Lorch, M. L. Edgley, S. Flibotte, and D. G. Moerman, "Copy number variation in the genomes of twelve natural isolates of *Caenorhabditis elegans*," *BMC Genomics*, vol. 11, no. 1, article 62, 2010.

[54] G. Gianuzzi, P. D'Addabbo, M. Gasparro et al., "Analysis of high-identity segmental duplications in the grapevine genome," *BMC Genomics*, vol. 12, no. 8, Article ID 436, 2011.

[55] M. B. Rogers, J. D. Hilley, N. J. Dickens et al., "Chromosome and gene copy number variation allow major structural change between species and strains of *Leishmania*," *Genome Research*, vol. 21, no. 12, pp. 2129–2142, 2011.

[56] P. Stankiewicz and J. R. Lupski, "Molecular-evolutionary mechanisms for genomic disorders," *Current Opinion in Genetics and Development*, vol. 12, no. 3, pp. 312–319, 2002.

[57] P. J. Hastings, J. R. Lupski, S. M. Rosenberg, and G. Ira, "Mechanisms of change in gene copy number," *Nature Reviews Genetics*, vol. 10, no. 8, pp. 551–564, 2009.

[58] M. R. Lieber, Y. Ma, U. Pannicke, and K. Schwarz, "Mechanism and regulation of human non-homologous DNA end-joining," *Nature Reviews Molecular Cell Biology*, vol. 4, no. 9, pp. 712–720, 2003.

[59] D. J. Turner, M. Miretti, D. Rajan et al., "Germline rates of de novo meiotic deletions and duplications causing several genomic disorders," *Nature Genetics*, vol. 40, no. 1, pp. 90–95, 2008.

[60] D. R. Schrider and M. W. Hahn, "Gene copy-number polymorphism in nature," *Proceedings of the Royal Society B*, vol. 277, no. 1698, pp. 3213–3221, 2010.

[61] S. T. Lovett, R. L. Hurley, V. A. Sutera, R. H. Aubuchon, and M. A. Lebedeva, "Crossing over between regions of limited homology in *Escherichia coli*: RecA-dependent and RecA-independent pathways," *Genetics*, vol. 160, no. 3, pp. 851–859, 2002.

[62] R. M. Liskay, A. Letsou, and J. L. Stachelek, "Homology requirement for efficient gene conversion between duplicated chromosomal sequences in mammalian cells," *Genetics*, vol. 115, no. 1, pp. 161–167, 1987.

[63] L. T. Reiter, P. J. Hastings, E. Nelis, P. de Jonghe, C. van Broeckhoven, and J. R. Lupski, "Human meiotic recombination products revealed by sequencing a hotspot for homologous strand exchange in multiple HNPP deletion patients," *The American Journal of Human Genetics*, vol. 62, no. 5, pp. 1023–1033, 1998.

[64] J. M. Kidd, G. M. Cooper, W. F. Donahue et al., "Mapping and sequencing of structural variation from eight human genomes," *Nature*, vol. 453, no. 7191, pp. 56–64, 2008.

[65] J. Gu and M. R. Lieber, "Mechanistic flexibility as a conserved theme across 3 billion years of nonhomologous DNA end-joining," *Genes and Development*, vol. 22, no. 4, pp. 411–415, 2008.

[66] P. Stankiewicz, C. J. Shaw, J. D. Dapper et al., "Genome architecture catalyzes nonrecurrent chromosomal rearrangements," *The American Journal of Human Genetics*, vol. 72, no. 5, pp. 1101–1116, 2003.

[67] F. Zhang, M. Khajavi, A. M. Connolly, C. F. Towne, S. D. Batish, and J. R. Lupski, "The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans," *Nature Genetics*, vol. 41, no. 7, pp. 849–853, 2009.

[68] A. J. Jeffreys, V. Wilson, and S. L. Thein, "Hypervariable "minisatellite" regions in human DNA," *Nature*, vol. 314, no. 6006, pp. 67–73, 1985.

[69] A. J. Jeffreys, V. Wilson, and S. L. Thein, "Individual-specific "fingerprints" of human DNA," *Nature*, vol. 316, no. 6023, pp. 76–79, 1985.

[70] Y. Nakamura, M. Leppert, and P. O'Connell, "Variable number of tandem repeat (VNTR) markers for human gene mapping," *Science*, vol. 235, no. 4796, pp. 1616–1622, 1987.

[71] D. Tautz, "Hypervariability of simple sequences as a general source for polymorphic DNA markers," *Nucleic Acids Research*, vol. 17, no. 16, pp. 6463–6471, 1989.

[72] G. Levinson and G. A. Gutman, "Slipped-strand mispairing: a major mechanism for DNA sequence evolution," *Molecular Biology and Evolution*, vol. 4, no. 3, pp. 203–221, 1987.

[73] M. Bzymek and S. T. Lovett, "Instability of repetitive DNA sequences: the role of replication in multiple mechanisms," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 15, pp. 8319–8325, 2001.

[74] J. M. Chen, N. Chuzhanova, P. D. Stenson, C. Férec, and D. N. Cooper, "Meta-analysis of gross insertions causing human genetic disease: novel mutational mechanisms and the role of replication slippage," *Human Mutation*, vol. 25, no. 2, pp. 207–221, 2005.

[75] L. Patthy, *Protein Evolution*, Blackwell Science, 1999.

[76] V. Katju and M. Lynch, "The structure and early evolution of recently arisen gene duplicates in the *Caenorhabditis elegans* genome," *Genetics*, vol. 165, no. 4, pp. 1793–1803, 2003.

[77] R. P. Meisel, "Evolutionary dynamics of recently duplicated genes: selective constraints on diverging paralogs in the *Drosophila pseudoobscura* genome," *Journal of Molecular Evolution*, vol. 69, no. 1, pp. 81–93, 2009.

[78] V. Katju, J. C. Farslow, and U. Bergthorsson, "Variation in gene duplicates with low synonymous divergence in *Saccharomyces cerevisiae* relative to *Caenorhabditis elegans*," *Genome Biology*, vol. 10, no. 7, article R75, 2009.

[79] L. Xu, H. Chen, X. Hu, R. Zhang, Z. Zhang, and Z. W. Luo, "Average gene length is highly conserved in prokaryotes and eukaryotes and diverges only between the two kingdoms," *Molecular Biology and Evolution*, vol. 23, no. 6, pp. 1107–1108, 2006.

[80] M. Deutsch and M. Long, "Intron-exon structures of eukaryotic model organisms," *Nucleic Acids Research*, vol. 27, no. 15, pp. 3219–3228, 1999.

[81] M. Lynch and J. S. Conery, "The origins of genome complexity," *Science*, vol. 302, no. 5649, pp. 1401–1404, 2003.

[82] S. W. Roy and W. Gilbert, "The evolution of spliceosomal introns: patterns, puzzles and progress," *Nature Reviews Genetics*, vol. 7, no. 3, pp. 211–221, 2006.

[83] M. Yandell, C. J. Mungall, C. Smith et al., "Large-scale trends in the evolution of gene structures within 11 animal genomes," *PLoS Computational Biology*, vol. 2, no. 3, Article ID e15, pp. 113–125, 2006.

[84] E. Gazave, T. Marqués-Bonet, O. Fernando, B. Charlesworth, and A. Navarro, "Patterns and rates of intron divergence between humans and chimpanzees," *Genome Biology*, vol. 8, no. 2, article R21, 2007.

[85] L. Zhu, Y. Zhang, W. Zhang, S. Yang, J. Q. Chen, and D. Tian, "Patterns of exon-intron architecture variation of genes in eukaryotic genomes," *BMC Genomics*, vol. 10, no. 1, article 47, 2009.

[86] S. P. Moss, D. A. Joyce, S. Humphries, K. J. Tindall, and D. L. Hunt, "Comparative analysis of teleost genome sequences reveals an ancient intron size expansion in the zebrafish lineage," *Genome Biology and Evolution*, vol. 3, pp. 1187–1196, 2011.

[87] A. Goffeau, G. Barrell, H. Bussey et al., "Life with 6000 genes," *Science*, vol. 274, no. 5287, pp. 546–567, 1996.

[88] L. Z. Gao and H. Innan, "Very low gene duplication rate in the yeast genome," *Science*, vol. 306, no. 5700, pp. 1367–1370, 2004.

[89] R. V. Sonti and J. R. Roth, "Role of gene duplications in the adaptation of *Salmonella typhimurium* to growth on limiting carbon sources," *Genetics*, vol. 123, no. 1, pp. 19–28, 1989.

[90] A. B. Reams and E. L. Neidle, "Genome plasticity in *Acinetobacter*: new degradative capabilities acquired by the spontaneous amplification of large chromosomal segments," *Molecular Microbiology*, vol. 47, no. 5, pp. 1291–1304, 2003.

[91] S. Zhong, A. Khodursky, D. E. Dykhuizen, and A. M. Dean, "Evolutionary genomics of ecological specialization," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 32, pp. 11719–11724, 2004.

[92] S. Sun, O. G. Berg, J. R. Roth, and D. I. Andersson, "Contribution of gene amplification to evolution of increased antibiotic resistance in *Salmonella typhimurium*," *Genetics*, vol. 182, no. 4, pp. 1183–1195, 2009.

[93] S. Fogel and J. W. Welch, "Tandem gene amplification mediates copper resistance in yeast," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 79, no. 17, pp. 5342–5346, 1982.

[94] T. Lenormand, T. Guillemaud, D. Bourguet, and M. Raymond, "Appearance and sweep of a gene duplication: adaptive response and potential for new functions in the mosquito *Culex pipiens*," *Evolution*, vol. 52, no. 6, pp. 1705–1712, 1998.

[95] R. D. Newcomb, D. M. Gleeson, C. G. Yong, R. J. Russell, and J. G. Oakeshott, "Multiple mutations and gene duplications conferring organophosphorus insecticide resistance have been selected at the *Rop-1* locus of the sheep blowfly, *Lucilia cuprina*," *Journal of Molecular Evolution*, vol. 60, no. 2, pp. 207–220, 2005.

[96] G. Maroni, J. Wise, J. E. Young, and E. Otto, "Metallothionein gene duplications and metal tolerance in natural populations of *Drosophila melanogaster*," *Genetics*, vol. 117, no. 4, pp. 739–744, 1987.

[97] D. M. Irwin and A. C. Wilson, "Multiple cDNA sequences and the evolution of bovine stomach lysozyme," *Journal of Biological Chemistry*, vol. 264, no. 19, pp. 11387–11393, 1989.

[98] M. Ouellette, E. Hettema, D. Wust, F. Fase-Fowler, and P. Borst, "Direct and inverted DNA repeats associated with P-glycoprotein gene amplification in drug resistant *Leishmania*," *The EMBO Journal*, vol. 10, no. 4, pp. 1009–1016, 1991.

[99] U. Bergthorsson, D. I. Andersson, and J. R. Roth, "Ohno's dilemma: evolution of new genes under continuous selection," *Proceedings of the National Academy of Sciences of the*

United States of America, vol. 104, no. 43, pp. 17004–17009, 2007.

[100] A. L. Hughes, "The evolution of functionally novel proteins after gene duplication," *Proceedings of the Royal Society B*, vol. 256, no. 1346, pp. 119–124, 1994.

[101] S. Yokoyama and R. Yokoyama, "Molecular evolution of visual pigment proteins and other G-protein coupled receptor genes," in *Population Biology of Genes*, N. Takahata and J. F. Crow, Eds., pp. 307–322, Baifukan, Tokyo, Japan, 1990.

[102] J. Zhang, H. F. Rosenberg, and M. Nei, "Positive Darwinian selection after gene duplication in primate ribonuclease genes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 7, pp. 3708–3713, 1998.

[103] S. S. Deeb, A. L. Jorgensen, L. Battisti, L. Iwasaki, and A. G. Motulsky, "Sequence divergence of the red and green visual pigments in great apes and humans," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 91, no. 15, pp. 7262–7266, 1994.

[104] M. Neitz, J. Neitz, and G. H. Jacobs, "Spectral tuning of pigments underlying red-green color vision," *Science*, vol. 252, no. 5008, pp. 971–974, 1991.

[105] R. E. Ibbotson, D. M. Hunt, J. K. Bowmaker, and J. D. Mollon, "Sequence divergence and copy number of the middle- and long-wave photopigment genes in Old World monkeys," *Proceedings of the Royal Society B*, vol. 247, no. 1319, pp. 145–154, 1992.

[106] S. L. Merbs and J. Nathans, "Absorption spectra of human cone pigments," *Nature*, vol. 356, no. 6368, pp. 433–435, 1992.

[107] G. H. Jacobs, M. Neitz, J. F. Deegan, and J. Neitz, "Trichromatic colour vision in New World monkeys," *Nature*, vol. 382, no. 6587, pp. 156–158, 1996.

[108] A. K. Surridge, D. Osorio, and N. I. Mundy, "Evolution and selection of trichromatic vision in primates," *Trends in Ecology and Evolution*, vol. 18, no. 4, pp. 198–205, 2003.

[109] J. R. Arguello, Y. Chen, S. Yang, W. Wang, and M. Long, "Origination of an X-linked testes chimeric gene by illegitimate recombination in *Drosophila*," *PLoS Genetics*, vol. 2, no. 5, p. e77, 2006.

[110] C. Deng, C. H. C. Cheng, H. Ye, X. He, and L. Chen, "Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 50, pp. 21593–21598, 2010.

[111] Wang Xin, A. L. DeVries, and C. H. C. Cheng, "Genomic basis for antifreeze peptide heterogeneity and abundance in an Antarctic eel pout—gene structures and organization," *Molecular Marine Biology and Biotechnology*, vol. 4, no. 2, pp. 135–147, 1995.

[112] M. Buljan, A. Frankish, and A. Bateman, "Quantifying the mechanisms of domain gain in animal proteins," *Genome Biology*, vol. 11, no. 7, article R74, 2010.

[113] V. Katju, E. M. LaBeau, K. J. Lipinski, and U. Bergthorsson, "Sex change by gene conversion in a *Caenorhabditis elegans fog-2* mutant," *Genetics*, vol. 180, no. 1, pp. 669–672, 2008.

[114] C. Charrier, K. Joshi, J. Coutinho-Budd et al., "Inhibition of *SRGAP2* function by its human-specific paralogs induces neoteny during spine maturation," *Cell*, vol. 149, no. 4, pp. 923–935, 2012.

[115] M. Y. Dennis, X. Nuttle, P. H. Sudmant et al., "Evolution of human-specific neural *SRGAP2* genes by incomplete segmental duplication," *Cell*, vol. 149, no. 4, pp. 912–922, 2012.

[116] S. Yoshimoto, E. Okada, H. Umemoto et al., "A W-linked DM-domain gene, *DM-W*, participates in primary ovary development in *Xenopus laevis*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 7, pp. 2469–2474, 2008.

[117] A. J. Bewick, D. W. Anderson, and B. J. Evans, "Evolution of the closely related, sex-related genes *DM-W* and *DMRT1* in African clawed frogs (*Xenopus*)," *Evolution*, vol. 65, no. 3, pp. 698–712, 2011.

[118] S. Yoshimoto, N. Ikeda, Y. Izutsu, T. Shiba, N. Takamatsu, and M. Ito, "Opposite roles of *DMRT1* and its W-linked paralogue, *DM-W*, in sexual dimorphism of *Xenopus laevis*: implications of a ZZ/ZW-type sex-determining system," *Development*, vol. 137, no. 15, pp. 2519–2526, 2010.

[119] J. Hodgkin and T. Doniach, "Natural variation and copulatory plug formation in *Caenorhabditis elegans*," *Genetics*, vol. 146, no. 1, pp. 149–164, 1997.

[120] K. Kiontke, N. P. Gavin, Y. Raynes, C. Roehrig, F. Piano, and D. H. A. Fitch, "*Caenorhabditis* phylogeny predicts convergence of hermaphroditism and extensive intron loss," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 24, pp. 9003–9008, 2004.

[121] R. C. Hill, C. E. de Carvalho, J. Salogiannis, B. Schlager, D. Pilgrim, and E. S. Haag, "Genetic flexibility in the convergent evolution of hermaphroditism in *Caenorhabditis* nematodes," *Developmental Cell*, vol. 10, no. 4, pp. 531–538, 2006.

[122] R. Clifford, M. H. Lee, S. Nayak, M. Ohmachi, F. Giorgini, and T. Schedl, "FOG-2, a novel F-box containing protien, associates with the GLD-1 RNA binding protein and directs male sex determination in the *C. elegans* hermaphrodite germline," *Development*, vol. 127, no. 24, pp. 5265–5276, 2000.

[123] S. Nayak, J. Goree, and T. Schedl, "*fog-2* and the evolution of self-fertile hermaphroditism in *Caenorhabditis*," *PLoS Biology*, vol. 3, no. 1, Article ID e6, 2005.

[124] H. S. Rane, J. M. Smith, U. Bergthorsson, and V. Katju, "Gene conversion and DNA sequence polymorphism in the sex-determination gene *fog-2* and its paralog *ftr-1* in *Caenorhabditis elegans*," *Molecular Biology and Evolution*, vol. 27, no. 7, pp. 1561–1569, 2010.

[125] S. A. Korneev, J. H. Park, and M. O'Shea, "Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene," *Journal of Neuroscience*, vol. 19, no. 18, pp. 7711–7720, 1999.

[126] S. Hirotsune, N. Yoshida, A. Chen et al., "An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene," *Nature*, vol. 423, no. 6935, pp. 91–96, 2003.

[127] O. Podlaha and J. Zhang, "Nonneutral evolution of the transcribed pseudogene *Makorin1-p1* in mice," *Molecular Biology and Evolution*, vol. 21, no. 12, pp. 2202–2209, 2004.

[128] T. A. Gray, A. Wilson, P. J. Fortin, and R. D. Nicholls, "The putatively functional *Mkrn1-p1* pseudogene is neither expressed nor imprinted, nor does it regulate its source gene in *trans*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 32, pp. 12039–12044, 2006.

[129] A. P. Piehler, M. Hellum, J. J. Wenzel et al., "The human ABC transporter pseudogene family: evidence for transcription and gene-pseudogene interference," *BMC Genomics*, vol. 9, article 165, 2008.

[130] R. Sasidharan and M. Gerstein, "Genomics: protein fossils live on as RNA," *Nature*, vol. 453, no. 7196, pp. 729–731, 2008.

[131] X. Guo, Z. Zhang, M. B. Gerstein, and D. Zheng, "Small RNAs originated from pseudogenes: *cis*- or *trans*-acting?" *PLoS Computational Biology*, vol. 5, no. 7, Article ID e1000449, 2009.

[132] M. Cardoso-Moreira and M. Long, "The origin and evolution of new genes," in *Evolutionary Genomics: Statistical and Computational Methods, Methods in Molecular Biology*, M. Anisimova, Ed., vol. 856, chapter 7, pp. 161–186, Springer, 2012.

[133] M. Lynch and V. Katju, "The altered evolutionary trajectories of gene duplicates," *Trends in Genetics*, vol. 20, no. 11, pp. 544–549, 2004.

[134] R. L. Watts and D. C. Watts, "Gene duplication and the evolution of enzymes," *Nature*, vol. 217, no. 5134, pp. 1125–1130, 1968.

[135] H. Innan and F. Kondrashov, "The evolution of gene duplications: classifying and distinguishing between models," *Nature Reviews Genetics*, vol. 11, no. 2, pp. 97–108, 2010.

[136] M. Long, E. Betrán, K. Thornton, and W. Wang, "The origin of new genes: glimpses from the young and old," *Nature Reviews Genetics*, vol. 4, no. 11, pp. 865–875, 2003.

[137] M. W. Hahn, "Distinguishing among evolutionary models for the maintenance of gene duplicates," *Journal of Heredity*, vol. 100, no. 5, pp. 605–617, 2009.

[138] R. L. Rogers, T. Bedford, and D. L. Hartl, "Formation and longevity of chimeric and duplicate genes in *Drosophila melanogaster*," *Genetics*, vol. 181, no. 1, pp. 313–322, 2009.

[139] R. L. Rogers, T. Bedford, A. M. Lyons, and D. L. Hartl, "Adaptive impact of the chimeric gene *Quetzalcoatl* in *Drosophila melanogaster*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 24, pp. 10943–10948, 2010.

[140] J. C. Opazo, A. M. Sloan, K. L. Campbell, and J. F. Storz, "Origin and ascendancy of a chimeric fusion gene: the β/δ-globin gene of paenungulate mammals," *Molecular Biology and Evolution*, vol. 26, no. 7, pp. 1469–1478, 2009.

[141] M. Goodman, B. F. Koop, J. Czelusniak, and L. Weiss, "The η-globin gene—its long evolutionary history in the β-globin gene family of mammals," *Journal of Molecular Biology*, vol. 180, no. 4, pp. 803–823, 1984.

[142] J. C. Opazo, F. G. Hoffmann, and J. F. Storz, "Genomic evidence for independent origins of β-like globin genes in monotremes and therian mammals," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 5, pp. 1590–1595, 2008.

[143] J. C. Opazo, F. G. Hoffmann, and J. F. Storz, "Differential loss of embryonic globin genes during the radiation of placental mammals," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 35, pp. 12950–12955, 2008.

[144] M. Bento Soares, E. Schon, A. Henderson et al., "RNA-mediated gene duplication—the rat preproinsulin I gene is a functional retroposon," *Molecular and Cellular Biology*, vol. 5, no. 8, pp. 2090–2103, 1985.

[145] P. H. Boer, C. N. Adra, Y. F. Lau, and M. W. McBurney, "The testis-specific phosphoglycerate kinase gene *pgk-2* is a recruited retroposon," *Molecular and Cellular Biology*, vol. 7, no. 9, pp. 3107–3112, 1987.

[146] B. P. Cusack and K. H. Wolfe, "Not born equal: increased rate asymmetry in relocated and retrotransposed rodent gene duplicates," *Molecular Biology and Evolution*, vol. 24, no. 3, pp. 679–686, 2007.

[147] Z. Zhang, N. Carriero, and M. Gerstein, "Comparative analysis of processed pseudogenes in the mouse and human genomes," *Trends in Genetics*, vol. 20, no. 2, pp. 62–67, 2004.

[148] A. C. Marques, I. Dupanloup, N. Vinckenbosch, A. Reymond, and H. Kaessmann, "Emergence of young human genes after a burst of retroposition in primates," *PLoS Biology*, vol. 3, no. 11, article e357, pp. 1970–1979, 2005.

[149] W. Wang, H. Zheng, C. Fan et al., "High rate of chimeric gene origination by retroposition in plant genomes," *Plant Cell*, vol. 18, no. 8, pp. 1791–1802, 2006.

[150] Y. Bai, C. Casola, C. Feschotte, and E. Betrán, "Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in *Drosophila*," *Genome Biology*, vol. 8, no. 1, article R11, 2007.

[151] D. Pan and L. Zhang, "Quantifying the major mechanisms of recent gene duplications in the human and mouse genomes: a novel strategy to estimate gene duplication rates," *Genome Biology*, vol. 8, no. 8, article R158, 2007.

[152] L. Potrzebowski, N. Vinckenbosch, A. C. Marques, F. Chalmel, B. Jégou, and H. Kaessmann, "Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes," *PLoS Biology*, vol. 6, no. 4, pp. 709–716, 2008.

[153] L. W. Hillier, W. Miller, E. Birney et al., "Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution," *Nature*, vol. 432, no. 7018, pp. 695–716, 2004.

[154] H. Kaessmann, N. Vinckenbosch, and M. Long, "RNA-based gene duplication: mechanistic and evolutionary insights," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 19–31, 2009.

[155] J. Maestre, T. Tchénio, O. Dhellin, and T. Heidmann, "mRNA retroposition in human cells: processed pseudogene formation," *The EMBO Journal*, vol. 14, no. 24, pp. 6333–6338, 1995.

[156] A. J. Mighell, N. R. Smith, P. A. Robinson, and A. F. Markham, "Vertebrate pseudogenes," *FEBS Letters*, vol. 468, no. 2-3, pp. 109–114, 2000.

[157] M. Long, M. Deutsch, W. Wang, E. Betrán, F. G. Brunet, and J. Zhang, "Origin of new genes: evidence from experimental and computational analyses," *Genetica*, vol. 118, no. 2-3, pp. 171–182, 2003.

[158] Q. Zhou and W. Wang, "On the origin and evolution of new genes—a genomic and experimental perspective," *Journal of Genetics and Genomics*, vol. 35, no. 11, pp. 639–648, 2008.

[159] F. Perler, A. Efstratiadis, and P. Lomedico, "The evolution of genes—the chicken preproinsulin gene," *Cell*, vol. 20, no. 2, pp. 555–566, 1980.

[160] J. R. McCarrey and K. Thomas, "Human testis-specific *PGK* gene lacks introns and possesses characteristics of a processed gene," *Nature*, vol. 326, no. 6112, pp. 501–505, 1987.

[161] J. R. McCarrey, "Molecular evolution of the human *Pgk-2* retroposon," *Nucleic Acids Research*, vol. 18, no. 4, pp. 949–955, 1990.

[162] J. Bradley, A. Baltus, H. Skaletsky, M. Royce-Tolland, K. Dewar, and D. C. Page, "An X-to-autosome retrogene is required for spermatogenesis in mice," *Nature Genetics*, vol. 36, no. 8, pp. 872–876, 2004.

[163] E. Betrán, K. Thornton, and M. Long, "Retroposed new genes out of the X in *Drosophila*," *Genome Research*, vol. 12, no. 12, pp. 1854–1859, 2002.

[164] D. R. Schrider, K. Stevens, C. M. Cardeno, C. H. Langley, and M. W. Hahn, "Genome-wide analysis of retrogene polymorphisms in *Drosophila melanogaster*," *Genome Research*, vol. 21, no. 12, pp. 2087–2095, 2011.

[165] M. Long, W. Wang, and J. Zhang, "Origin of new genes and source for N-terminal domain of the chimerical gene, *Jingwei*, in *Drosophila*," *Gene*, vol. 238, no. 1, pp. 135–141, 1999.

[166] W. Wang, J. Zhang, C. Alvarez, A. Llopart, and M. Long, "The origin of the *Jingwei* gene and the complex modular structure of its parental gene, *yellow emperor*, in *Drosophila melanogaster*," *Molecular Biology and Evolution*, vol. 17, no. 9, pp. 1294–1301, 2000.

[167] J. Zhang, A. M. Dean, F. Brunet, and M. Long, "Evolving protein functional diversity in new genes of *Drosophila*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 46, pp. 16246–16250, 2004.

[168] C. D. Jones, A. W. Custer, and D. J. Begun, "Origin and evolution of a chimeric fusion gene in *Drosophila subobscura*, *D. madeirensis*, and *D. guanche*," *Genetics*, vol. 170, no. 1, pp. 207–219, 2005.

[169] C. D. Jones, D. J. Begun, and T. Ohta, "Parallel evolution of chimeric fusion genes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 32, pp. 11373–11378, 2005.

[170] Y. W. Zhang, S. Liu, X. Zhang et al., "A functional mouse retroposed gene *Rps23r1* reduces Alzheimer's β-amyloid levels and tau phosphorylation," *Neuron*, vol. 64, no. 3, pp. 328–340, 2009.

[171] J. B. Walsh, "Sequence-dependent gene conversion—can duplicated genes diverge fast enough to escape conversion?" *Genetics*, vol. 117, no. 3, pp. 543–557, 1987.

[172] H. Innan, "A two-locus gene conversion model with selection and its application to the human *RHCE* and *RHD* genes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 15, pp. 8793–8798, 2003.

[173] S. A. Liebhaber, M. Goossens, and Y. W. Kan, "Homology and concerted evolution at the α1 and α2 loci of human α-globin," *Nature*, vol. 290, no. 5801, pp. 26–29, 1981.

[174] A. J. L. Brown and D. Ish-Horowicz, "Evolution of the 87A and 87C heat-shock loci in *Drosophila*," *Nature*, vol. 290, no. 5808, pp. 677–682, 1981.

[175] R. Ollo and F. Rougeon, "Gene conversion and polymorphism: generation of mouse immunoglobulin γ2a chain alleles by differential gene conversion by γ2b chain gene," *Cell*, vol. 32, no. 2, pp. 515–523, 1983.

[176] K. Iatrou, S. G. Tsitilou, and F. C. Kafatos, "DNA sequence transfer between two high-cysteine chorion gene families in the silkmoth *Bombyx mori*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 81, no. 14 I, pp. 4452–4456, 1984.

[177] T. D. Petes and C. W. Hill, "Recombination between repeated genes in microorganisms," *Annual Review of Genetics*, vol. 22, pp. 147–168, 1988.

[178] C. Semple and K. H. Wolfe, "Gene duplication and gene conversion in the *Caenorhabditis elegans* genome," *Journal of Molecular Evolution*, vol. 48, no. 5, pp. 555–564, 1999.

[179] G. Santoyo and D. Romero, "Gene conversion and concerted evolution in bacterial genomes," *FEMS Microbiology Reviews*, vol. 29, no. 2, pp. 169–183, 2005.

[180] T. Ohta, "The mutational load of a multigene family with uniform members," *Genetical Research*, vol. 53, no. 2, pp. 141–145, 1989.

[181] L. D. Hurst and N. G. C. Smith, "The evolution of concerted evolution," *Proceedings of the Royal Society B*, vol. 265, no. 1391, pp. 121–127, 1998.

[182] R. P. Sugino and H. Innan, "Selection for more of the same product as a force to enhance concerted evolution of duplicated genes," *Trends in Genetics*, vol. 22, no. 12, pp. 642–644, 2006.

[183] W. R. Engels, C. R. Preston, and D. M. Johnson-Schlitz, "Long-range *cis* preference in DNA homology search over the length of a *Drosophila* chromosome," *Science*, vol. 263, no. 5153, pp. 1623–1625, 1994.

[184] J. R. Murti, M. Bumbulis, and J. C. Schimenti, "Gene conversion between unlinked sequences in the germline of mice," *Genetics*, vol. 137, no. 3, pp. 837–843, 1994.

[185] V. Katju and U. Bergthorsson, "Genomic and population-level effects of gene conversion in *Caenorhabditis* paralogs," *Genes*, vol. 1, no. 3, pp. 452–468, 2010.

[186] M. Kimura, "Diffusion models in population genetics," *Journal of Applied Probability*, vol. 1, no. 2, pp. 177–132, 1964.

[187] J. F. Crow and M. Kimura, *An Introduction to Population Genetics Theory*, Harper and Row, New York, NY, USA, 1970.

[188] T. Ohta, "Slightly deleterious mutant substitutions in evolution," *Nature*, vol. 246, no. 5428, pp. 96–98, 1973.

[189] M. Kimura, *The Neutral Theory of Molecular Evolution*, Cambridge University Press, Cambridge, UK, 1983.

[190] M. Lynch, B. Koskella, and S. Schaack, "Mutation pressure and the evolution of organelle genomic architecture," *Science*, vol. 311, no. 5768, pp. 1727–1730, 2006.

[191] M. Lynch, *The Origins of Genome Architecture*, Sinauer Associates, Sunderland, Mass, USA, 2007.

[192] J. Charlesworth and A. Eyre-Walker, "The rate of adaptive evolution in enteric bacteria," *Molecular Biology and Evolution*, vol. 23, no. 7, pp. 1348–1356, 2006.

[193] M. S. Snoke, T. U. Berendonk, D. Barth, and M. Lynch, "Large global effective population sizes in *Paramecium*," *Molecular Biology and Evolution*, vol. 23, no. 12, pp. 2474–2479, 2006.

[194] J. Mu, J. Duan, K. D. Makova et al., "Chromosome-wide SNPs reveal an ancient origin for *Plasmodium falciparum*," *Nature*, vol. 418, no. 6895, pp. 323–326, 2002.

[195] I. J. Tsai, D. Bensasson, A. Burt, and V. Koufopanou, "Population genomics of the wild yeast *Saccharomyces paradoxus*: quantifying the life cycle," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 12, pp. 4957–4962, 2008.

[196] A. D. Cutter, "Nucleotide polymorphism and linkage disequilibrium in wild populations of the partial selfer *Caenorhabditis elegans*," *Genetics*, vol. 172, no. 1, pp. 171–184, 2006.

[197] A. D. Cutter, S. E. Baird, and D. Charlesworth, "High nucleotide polymorphism and rapid decay of linkage disequilibrium in wild populations of *Caenorhabditis remanei*," *Genetics*, vol. 174, no. 2, pp. 901–913, 2006.

[198] J. A. Shapiro, W. Huang, C. Zhang et al., "Adaptive genic evolution in the *Drosophila genomes*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 7, pp. 2271–2276, 2007.

[199] A. Eyre-Walker, P. D. Keightley, N. G. C. Smith, and D. Gaffney, "Quantifying the slightly deleterious mutation model of molecular evolution," *Molecular Biology and Evolution*, vol. 19, no. 12, pp. 2142–2149, 2002.

[200] T. I. Gossmann, B. H. Song, A. J. Windsor et al., "Genome wide analyses reveal little evidence for adaptive evolution in many plant species," *Molecular Biology and Evolution*, vol. 27, no. 8, pp. 1822–1832, 2010.

[201] T. Slotte, J. P. Foxe, K. M. Hazzouri, and S. I. Wright, "Genome-wide evidence for efficient positive and purifying selection in *Capsella grandiflora*, a plant species with a large effective population size," *Molecular Biology and Evolution*, vol. 27, no. 8, pp. 1813–1821, 2010.

[202] P. K. Ingvarsson, "Multilocus patterns of nucleotide polymorphism and the demographic history of *Populus tremula*," *Genetics*, vol. 180, no. 1, pp. 329–340, 2008.

[203] P. K. Ingvarsson, "Natural selection on synonymous and nonsynonymous mutations shapes patterns of polymorphism in *Populus tremula*," *Molecular Biology and Evolution*, vol. 27, no. 3, pp. 650–660, 2010.

[204] D. L. Halligan, F. Oliver, A. Eyre-Walker, B. Harr, and P. D. Keightley, "Evidence for pervasive adaptive protein evolution in wild mice," *PLoS Genetics*, vol. 6, no. 1, Article ID e1000825, 2010.

[205] N. Yu, M. I. Jensen-Seaman, L. Chemnick, O. Ryder, and W. H. Li, "Nucleotide diversity in gorillas," *Genetics*, vol. 166, no. 3, pp. 1375–1383, 2004.

[206] S. E. Alter, E. Rynes, and S. R. Palumbi, "DNA evidence for historic population size and past ecosystem impacts of gray whales," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 38, pp. 15162–15167, 2007.

Hindawi Publishing Corporation
International Journal of Evolutionary Biology
Volume 2012, Article ID 724519, 11 pages
doi:10.1155/2012/724519

Research Article

Genetic Innovation in Vertebrates: Gypsy Integrase Genes and Other Genes Derived from Transposable Elements

Domitille Chalopin, Delphine Galiana, and Jean-Nicolas Volff

Institut de Génomique Fonctionnelle de Lyon, Université de Lyon, Ecole Normale Supérieure de Lyon, CNRS, Université Lyon 1, 69364 Lyon Cedex 07, France

Correspondence should be addressed to Domitille Chalopin, domitille.chalopin@ens-lyon.fr

Received 25 May 2012; Accepted 15 July 2012

Academic Editor: Wen Wang

Due to their ability to drive DNA rearrangements and to serve as a source of new coding and regulatory sequences, transposable elements (TEs) are considered as powerful evolutionary agents within genomes. In this paper, we review the mechanism of molecular domestication, which corresponds to the formation of new genes derived from TE sequences. Many genes derived from retroelements and DNA transposons have been identified in mammals and other vertebrates, some of them fulfilling essential functions for the development and survival of their host organisms. We will particularly focus on the evolution and expression of Gypsy integrase (*GIN*) genes, which have been formed from ancient event(s) of molecular domestication and have evolved differentially in some vertebrate sublineages. What we describe here is probably only the tip of the evolutionary iceberg, and future genome analyses will certainly uncover new TE-derived genes and biological functions driving genetic innovation in vertebrates and other organisms.

1. Introduction

For a long time, transposable elements (TEs) have been considered as pure selfish and junk elements parasiting the genome of living organism [1, 2]. These sequences are able to "move", that is, to insert into new locations within genomes. This phenomenon is called transposition. Retroelements use retrotransposition, that is, the reverse transcription of an RNA intermediate and integration of the cDNA molecule produced, to generate new copies of themselves within genomes (copy-and-paste mechanism). This mechanism directly increases the copy number of the element. Among protein-coding autonomous retroelements, distinction is generally made between elements with long terminal repeats (LTRs: LTR retrotransposons and retroviruses) and retroelements without LTRs (non-LTR retrotransposons or LINE elements). Retroviruses and LTR retrotransposons are mainly distinguished by the presence versus absence of an envelope gene, which encodes a protein necessary for virus entry into the target cell. After germ line infection, reverse-transcribed retrovirus genomes can be integrated into the host genome and transmitted through vertical inheritance to the host progeny [3]. Such sequences, called endogenous retroviruses, are generally inactivated by mutations. Gain or loss of the envelope gene can transform a retrotransposon into a retrovirus, and *vice versa* [4, 5]. The second large category of TEs, DNA transposons, generally excises from their original insertion site and reintegrate into a new location (cut-and-paste mechanism). For most DNA transposons, transposition is catalyzed by an enzyme called transposase [6]. Finally, noncoding nonautonomous elements using for their transposition proteins encoded by autonomous sequences exist for both retroelements and DNA transposons.

Despite the deep-rooted vision of junk DNA, there is growing evidence that TEs are more than simple genome parasites. Particularly, they have been shown to serve as a genomic reservoir for new regulatory and coding sequences allowing genetic innovation and organismal evolution. A fascinating facet of the roles of TEs in evolution is their ability to be "molecularly domesticated" to form new cellular protein-coding genes [7, 8]. TE-encoded proteins have properties

that can be of interest for host cellular pathways. They can bind, copy, cut, process, and recombine nucleic acids, as well as modify and interact with host proteins. There are many cases of TE-derived genes fulfilling important functions in plants, fungi, and animals, including vertebrates (for review, [8, 9]). We will present here several prominent examples of vertebrate genes formed from TE-coding sequences during evolution, with more emphasis on Gypsy integrase (*GIN*) genes that we have analyzed in different fish species.

2. Genes Derived from Retroelements

2.1. Gag-Derived Genes. Several multigenic families have been formed from different events of molecular domestication of the *gag* gene of Ty3/Gypsy elements, a super family of LTR retrotransposons active in fish and amphibians but extinct in mammals [9, 10]. The *gag* gene encodes a structural protein with three functional regions: the matrix (MA) domain playing a role in targeting cellular membranes, the capsid (CA) domain involved in interactions with other proteins during particle assembly, and the nucleocapsid (NC), which binds to viral RNA genomes through zinc fingers.

One *gag*-related gene family is called *Mart*. This gene family is mammal specific and constituted by 12 genes in human [11]. Most *Mart* genes are found on mammalian X chromosome, suggesting an initial event of molecular domestication on the X, followed by serial local duplication events that subsequently extended this gene family. All *Mart* genes have retained from the original *gag* sequence an intronless open reading frame. Some of them still encode the ancestral Gag zinc finger, suggesting nucleic acid binding properties for the protein. Two autosomal *Mart* genes, *PEG10* (*Mart2)* and *PEG11/Rtl1* (*Mart1*), are subject to genomic imprinting and are expressed from the paternal allele [12, 13]. This epigenetic regulation has been proposed to be derived from a defence mechanism repressing the activity of the ancestral retrotransposon before domestication [14]. At least two *Mart* genes, *PEG11/Rtl1* (*Mart1*) and *PEG10* (*Mart2)*, have essential but nonredundant functions in placenta development in the mouse [15, 16]. *PEG10* and other *Mart* genes might also control cell proliferation and apoptosis, with possible involvement in cancer ([8] and references therein).

Another mammalian gene family derived from a LTR retrotransposon *gag* gene is called *Ma* or *Pnma* (paraneoplastic Ma antigens) [17]. Fifteen *Ma/Pnma* genes are present in the human genome, most of them being located on the X chromosome as observed for *Mart* genes. Some Ma proteins are expressed by patients with paraneoplastic neurological disorders and might be targeted by autoimmune response leading to progressive neurological damage [18]. Several Ma proteins are also involved in apoptosis, including Ma4 (Pnma4/Map1/Maop1) and Ma1/Pnma1 [19, 20].

A third family is the SCAN domain family. This family is constituted of DNA binding proteins with an N-terminus region called the SCAN domain, which is derived from the Gag protein of a Gmr1-like Gypsy/Ty3 retrotransposon [21–24]. The SCAN family is vertebrate specific, with approximately 70 and 40 members in human and mouse, respectively. Several SCAN proteins have been shown to be transcription factors regulating diverse biological processes such as hematopoiesis, stem cell properties, or cell proliferation and apoptosis (for review [21]).

Finally, other *gag*-related genes are present in mammalian genomes [10]. One of them, *Fv1*, is of retroviral origin and controls replication of the murine leukaemia virus in the mouse [25].

2.2. Envelope-Derived Genes. During mammalian evolution, retroviral envelope genes have been domesticated several times independently to generate genes involved in placenta development [26]. These genes, derived from endogenous retroviruses, encode proteins called syncytins. Syncytins mediate the fusion of trophoblast cells to form the syncytiotrophoblast layer, a continuous structure with microvillar surfaces forming the outermost foetal component of the placenta [27]. Two syncytin genes of independent origins encoding placenta-specific fusogenic proteins are present in human and other simians (Syncytin-1 and -2, [28]) as well as in rodents (Syncytin-A and Syncytin-B, [29]). Independent Syncytin genes are also found in rabbit [30], guinea pig [31], and Carnivora [32], indicating multiple convergent domestication of *env*-derived Syncytin genes in different mammalian sublineages. Some Syncytins might be involved in other biological processes. For example, human Syncytin-1 plays a role in osteoclast fusion, neuroinflammation, and possibly multiple sclerosis [33, 34].

Other retroviral *env*-derived open reading frames are present in vertebrate genomes; but intensive work is required to determine their functions. Some of them might confer resistance to viral infection, as shown for the *Fv-4* locus. This locus, containing an entire ecotropic murine leukemia virus (MuLV) *env* gene, controls susceptibility to infection by MuLV [35].

2.3. Other Retroelement-Derived Genes. In mammals, a gene called *CGIN1* is partially derived from the integrase gene of an endogenous retrovirus. The integrase gene has been fused 125–180 million years ago to a duplicate of the cellular gene *KIAA0323*. A role of *CGIN1* in resistance against retroviruses has been proposed [36].

Several genes with homology to retroelement aspartyl protease genes are present in vertebrate genomes. One of them, a gene encoding a protein called SASPase, is necessary for the texture and hydration of the stratum corneum, the outermost layer of the epidermis [37].

Finally, the telomerase, the reverse transcriptase extending the ends of linear chromosomes in vertebrates and other eukaryotes, might be derived from a retroelement [38].

3. Genes Derived from DNA Transposons

Many examples of genes derived from transposase genes from diverse subfamilies of DNA transposons have been described in vertebrates and other organisms [8, 39, 40].

One well-studied example is the recombination-activating protein Rag1, which together with Rag2 catalyzes the V(D)J somatic site-specific recombination responsible for the formation and diversity of genes encoding immunoglobulins and T-cell receptors in jawed vertebrates. *Rag1* has been formed from the transposase of a Transib DNA transposon, and the V(D)J recombination signal sequences recognized by Rag1 might be derived from the transposon ends bound by the ancestral transposase [41].

The mammal-specific gene *CENP-B* encodes a Pogo transposase-derived protein that controls centromere formation depending on the chromatin context [42]. Interestingly, an independent event of molecular domestication of Pogo transposase also led to the formation of centromeric proteins in fission yeast [43]. In yeast, CENP-B-like proteins restrict the activity of retrotransposons and promote replication progression at forks paused by retrotransposon LTRs [44, 45]. Other genes are derived from Pogo-like transposons in mammals [46]. One example is the *Jerky* gene, which encodes a brain-specific mRNA-binding protein that may regulate mRNA use in neurons [47].

Similarly, several examples of genes derived from hAT transposases have been found in mammals, some of them having been fused to zinc finger domains [46]. Some hAT transposase-related proteins work as transcription factors. One of them, ZEBD6/MGR, negatively regulates *IGF2* expression and muscle growth. Indeed, it has been shown that mutation in a regulatory sequence prohibiting ZEBD6/MGR binding leads to *IGF2* upregulation and enhanced muscle growth in commercially bred pigs [48, 49].

In primates, the gene encoding the Metnase/SETMAR protein has been formed through fusion of the transposase gene of a Mariner transposon with a SET histone methyltransferase gene. Metnase/SETMAR is a DNA binding protein with endonuclease activity that promotes DNA double-strand break repair through nonhomologous end joining (NHEJ) [50, 51].

Several genes derived from PiggyBac-like transposons have been detected in human and other vertebrates [52]. One of them, *PGBD3*, serves as an alternative 3′ terminal exon for the Cockayne Syndrome B (CSB) gene, leading to the expression of a CSB-transposase fusion protein [53]. At least one Harbinger transposon-derived gene, *HARBI1*, encoding a predicted nuclease, is present in mammals, birds, amphibians, and fish [54]. Likewise, genes derived from a new type of DNA transposon called Zisupton have been identified in fish and other vertebrates [55]. Finally, mammalian and bird genomes possess at least one gene clearly derived from a P transposon; additional vertebrate genes like *THAP9* encoding proteins with a THAP domain might be also related to P-like transposases [8, 40, 56–61].

4. Gypsy Integrase Genes: Data from Fish

Two vertebrate genes with unknown functions, *GIN1* and *GIN2 (Gypsy Integrase 1 and 2),* encode proteins showing significant homologies to integrases encoded by LTR retrotransposons [62, 63]. Further analyses showed that both genes have been formed from GIN transposons, a new family of metazoan DNA transposons with a transposase that shows strong similarities with LTR retrotransposon integrases [64]. *GIN1,* which shows similarities with GINO transposons from *Hydra magnipapillata*, is present in mammals, birds, and reptiles, suggesting a molecular domestication event at the base of the Amniota ca. 300 million years ago. Mammalian GIN1 proteins have conserved amino-acid residues necessary for integrase activity. Using our own analyses, we will now particularly focus on the *GIN2* gene. We provide here updated *GIN2* structural and phylogenetic analyses using new vertebrate sequences and present first expression data for this gene in fish.

GIN2 is present in several fish species, as well as in cartilaginous fish (elephant shark), coelacanth, amphibians, birds, reptiles, and marsupials, but neither in monotremes nor in placental mammals [63] (Figures 1, 2, and 3). Furthermore, *GIN2* was not detected in lamprey. Hence, the molecular domestication event having led to the formation of *GIN2* might have taken place before the divergence between tetrapods/bony fish and cartilaginous fish around 500 million years ago, with subsequent loss in monotremes and placental mammals. The formation of *GIN2* might even be older, since potentially domesticated GIN-like sequences related to *GIN2* have been detected in the urochordates *Ciona savignyi* and *C. intestinalis* [63]. Phylogenetic analysis suggests that *GIN2* is derived from GINA transposons, which are *bona fide* transposable elements in *Hydra magnipapillata* (Figure 1). This suggests that *GIN1* and *GIN2* have been formed through two independent molecular domestication events, one at the base of Amniota and the other in a more ancient vertebrate ancestor (Figure 3).

After domestication, the HHCC zinc finger present in the ancestral integrase has been maintained, suggesting ability to bind to DNA or RNA (Figure 2). Conservation of the important catalytic triad (DDE, aspartic acid/aspartic acid/glutamic acid) of the integrase is less obvious. While this motif has been proposed to be conserved in GIN1, this is not the case for GIN2 based on a published alignment with sequences from GIN-related transposases [63] (Figure 2). As shown in Figure 2, the first aspartic acid residue is present in most species but absent from amphibians and birds. However, multiple sequence alignment revealed an aspartate conserved in all GIN2 and GIN1 sequences ca. 20 amino-acids downstream. The second aspartic acid residue is not found in GIN2 but an aspartate is conserved four amino acids away in all GIN2 sequences except for opossum. Finally, the glutamic acid residue is found only in several species and substituted by an aspartate in fish; but a conserved glutamate is detected 16 amino acids away. Hence, the question of the functionality of GIN2 as an integrase remains open and should be definitely answered through functional analyses. A third domain with unknown function called GPY/F [64, 67] is also detected in GIN proteins, but in some cases the phenylalanine residue is replaced by a leucin. *GIN2* contains eight protein-coding exons, with an exon-intron structure well conserved in fish and other vertebrates (Figure 4). Some introns might be derived from the ancestral transposon; others might be the result of events of intronization after molecular domestication. *GIN2* is located in the same

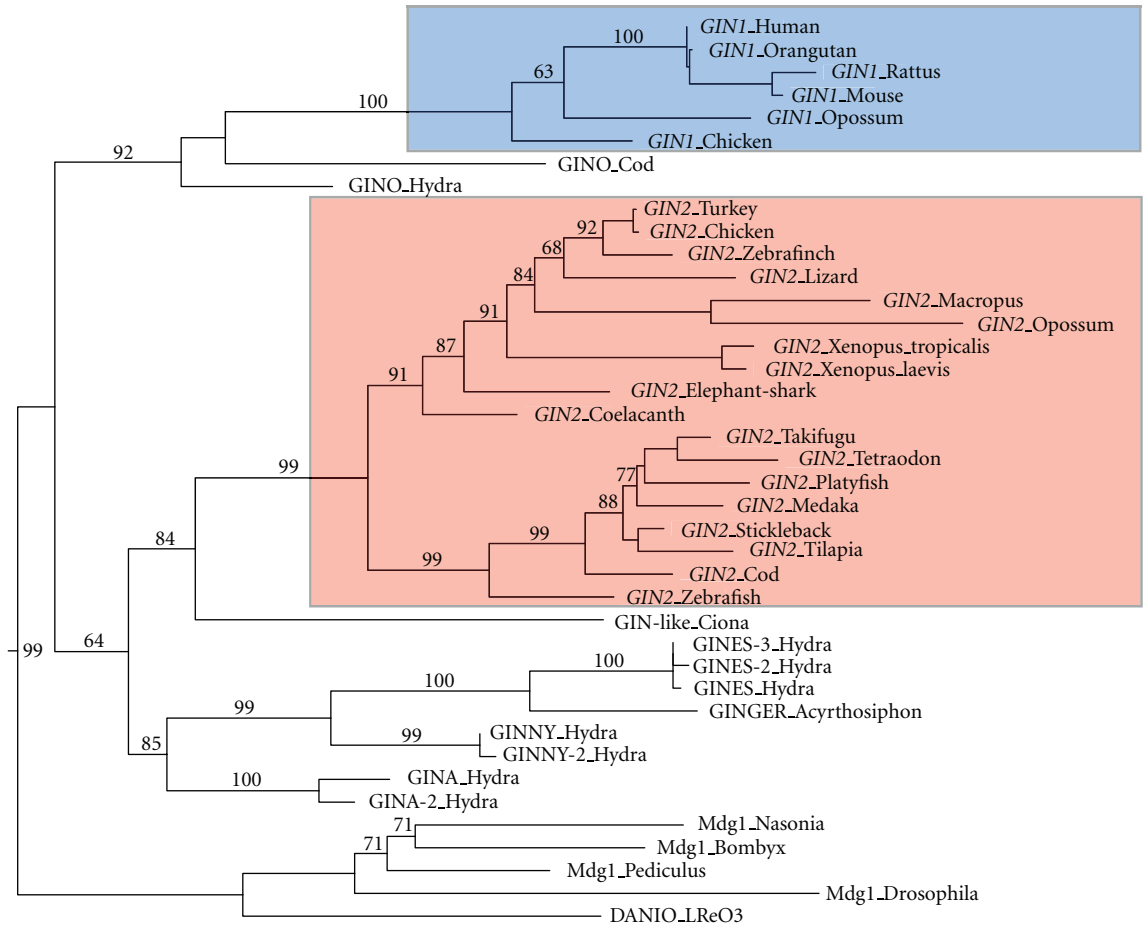


Figure 1: Molecular phylogeny of GIN proteins. Phylogenetic tree based on a 352 amino-acid integrase alignment. Protein sequences were aligned with clustalW and phylogenetic tree was constructed using maximum likelihood from PhyML package (optimized default bootstrap) [65]. Sequences were recovered from NCBI and Ensembl or predicted from genome sequences. Accession numbers and sequence alignments are available upon request.

orthologous genomic region between *OGFOD2* and *ABCB9* in marsupials, birds, reptiles, and fish, confirming that this gene does not correspond to a mobile sequence (Figure 5).

Expressed sequence tag (EST) analysis indicated that *GIN2* is expressed in different adult tissues and developmental stages in chicken: brain (accession number: CN219658), liver (BG713188), head (BU225420), embryonic tissue (BU210425), limb (BU256599), small intestine (BU297502), muscle (BU437928), and ovary (BU447634). Only ESTs from the whole body are available for *Xenopus*. Few ESTs are also found in zebrafish: muscle (CT684014), gills (EB908574), reproductive system (BI867074), and eye (BI879358).

To determine more precisely *GIN2* expression pattern in fish, quantitative real-time PCR was performed on different embryonic developmental stages in zebrafish (*Danio rerio*), as well as on adult tissues from zebrafish and platyfish (*Xiphophorus maculatus*) (Figure 6). During zebrafish embryogenesis, *GIN2* expression level strongly increases from the dome stage and progressively decreases until the end of somite stages. This result suggests that *GIN2* possibly plays a role during gastrulation. Gastrulation, which is characterized by morphologic movements of involution and extension, starts at the beginning of the epiboly to finish at bud stage [68]. In adult zebrafish, the higher level of expression for *GIN2* was observed in brain, followed by gonads and eyes. In contrast, *GIN2* expression was maximal in gonads in the platyfish (Figure 6).

To conclude, our analysis integrates data from several newly sequenced vertebrate genomes, particularly teleostean and cartilaginous fishes as well as coelacanth, in order to better understand the distribution and evolutionary history of *GIN* genes. Since *GIN2* is apparently not present in lamprey, we propose that *GIN2* was formed before the divergence between cartilaginous and ray-finned fish about 500 million years ago (Figure 3). We also provide the first expression data for *GIN2* in fish particularly supporting a function in gastrulation during zebrafish embryogenesis.

5. Conclusion

At first glance, transposable elements were considered as "junk" DNA with no important functions for genomes and

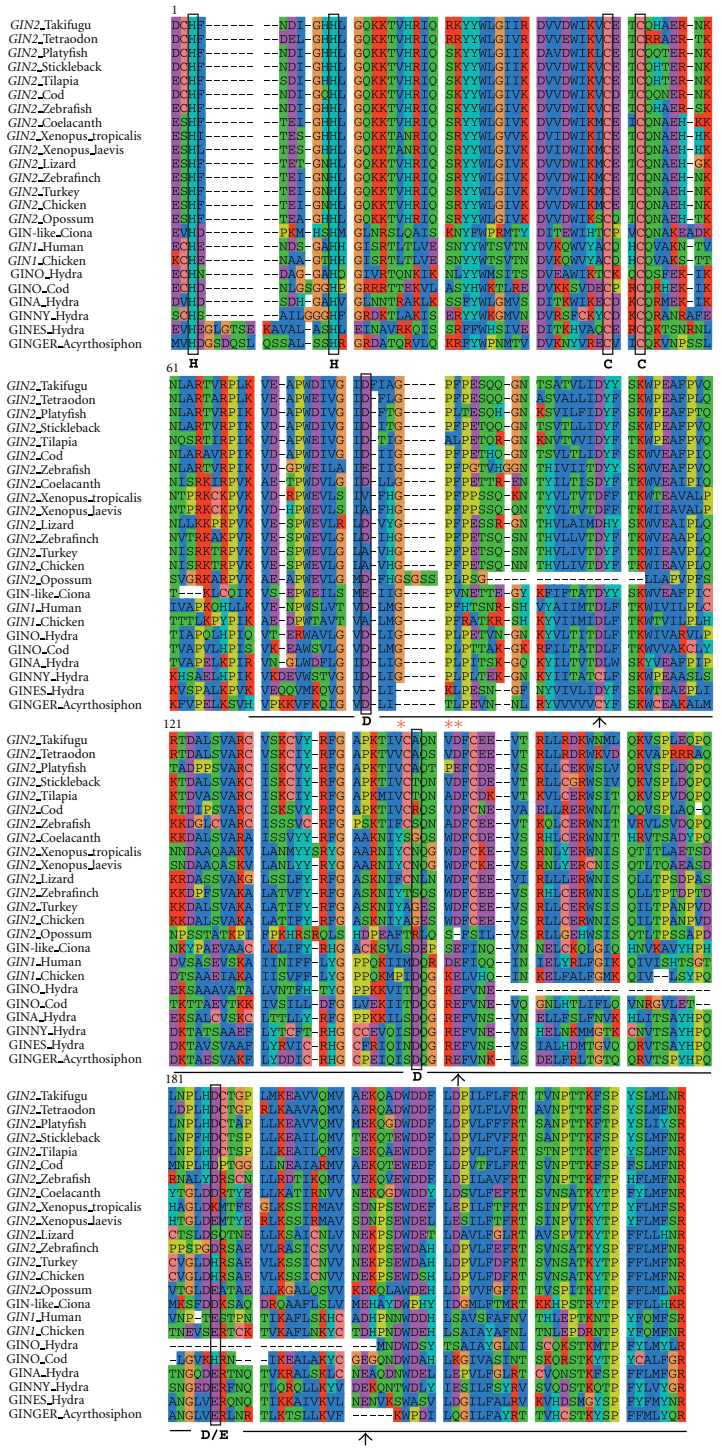


FIGURE 2: Sequence alignment of predicted GIN2-related proteins. HHCC zinc finger and integrase-like domain of *GIN2* were aligned using clustalW [66]. The black line indicates the position of the integrase-like domain. HHCC and DDE motifs are shown in black boxes and GPF motif is highlighted by red asterisks. Arrows indicate alternative conserved D/E residues in GIN2 sequences.

organisms. Today, nobody can deny the importance of transposable elements during evolution in terms of innovation power, particularly through molecular domestication events. Domesticated elements are *bona fide* cellular genes derived from transposable element sequences encoding for example integrases, transposases, Gag proteins, or envelopes. After domestication, TE-derived genes have lost their ability to transpose through the elimination of sequences such as long terminal repeats, terminal-inverted repeats, or other open reading frames and protein domains essential for

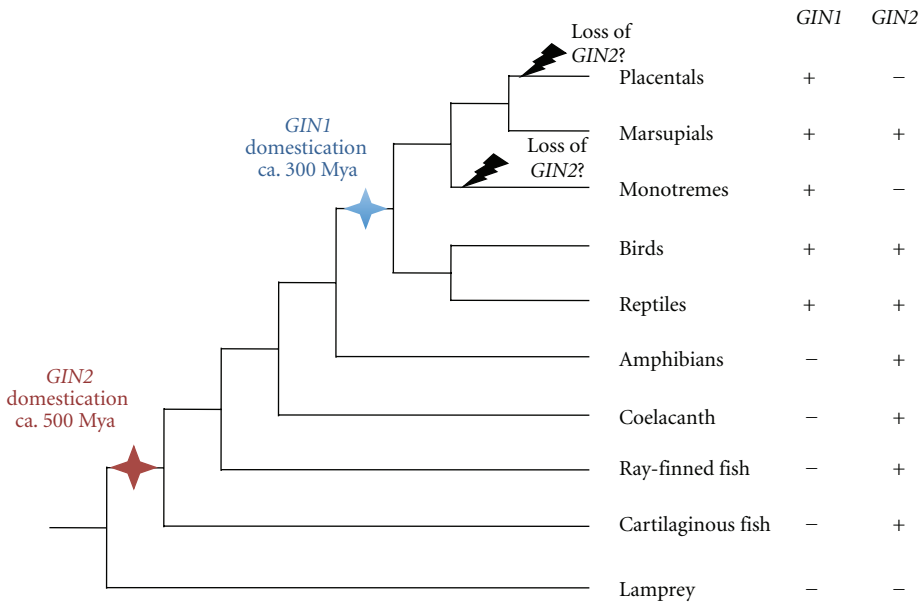


Figure 3: One possible scenario for the evolution of *GIN* genes in vertebrates. The two molecular domestication events are highlighted by blue and red stars, having led to the formation of *GIN1* and *GIN2*, respectively. Presence (+) or absence (−) of *GIN1* and *GIN2* in the different lineages is indicated.

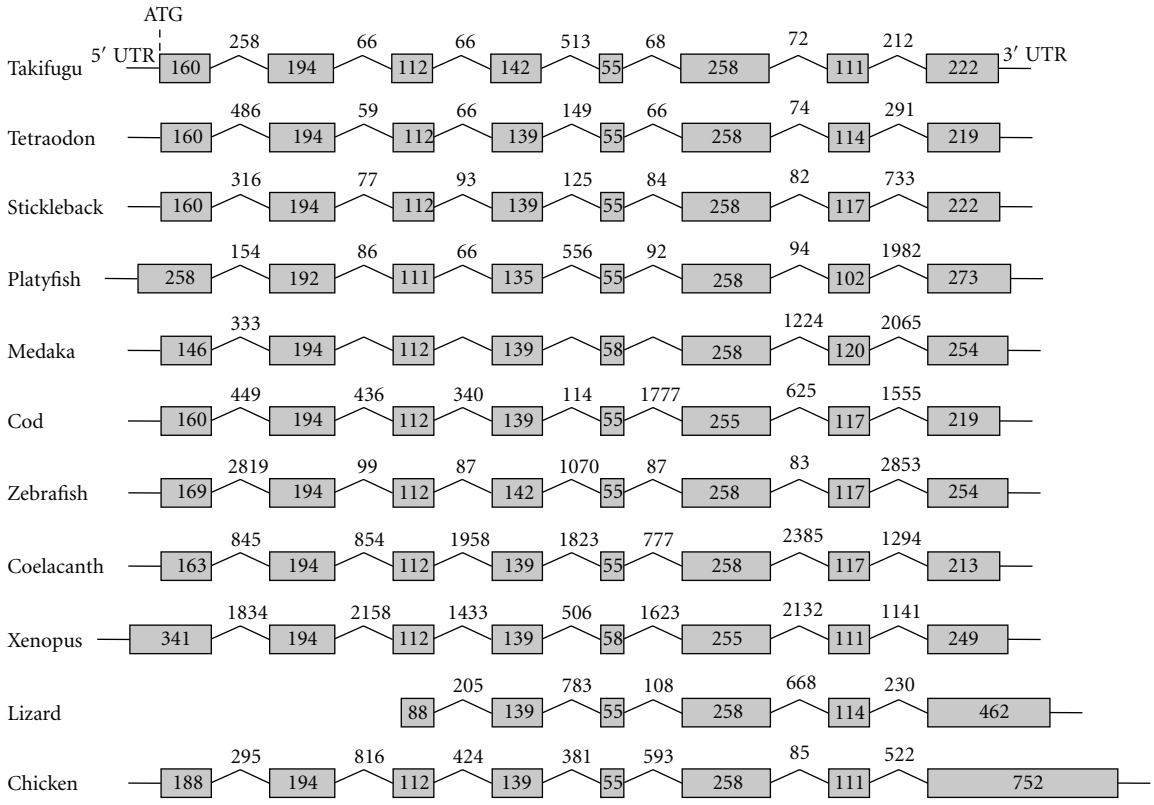


Figure 4: Exon-intron structure of *GIN2* genes in fish and other vertebrates. Exons are represented by grey boxes, introns by broken lines. Exon/intron sizes are given as base pairs.

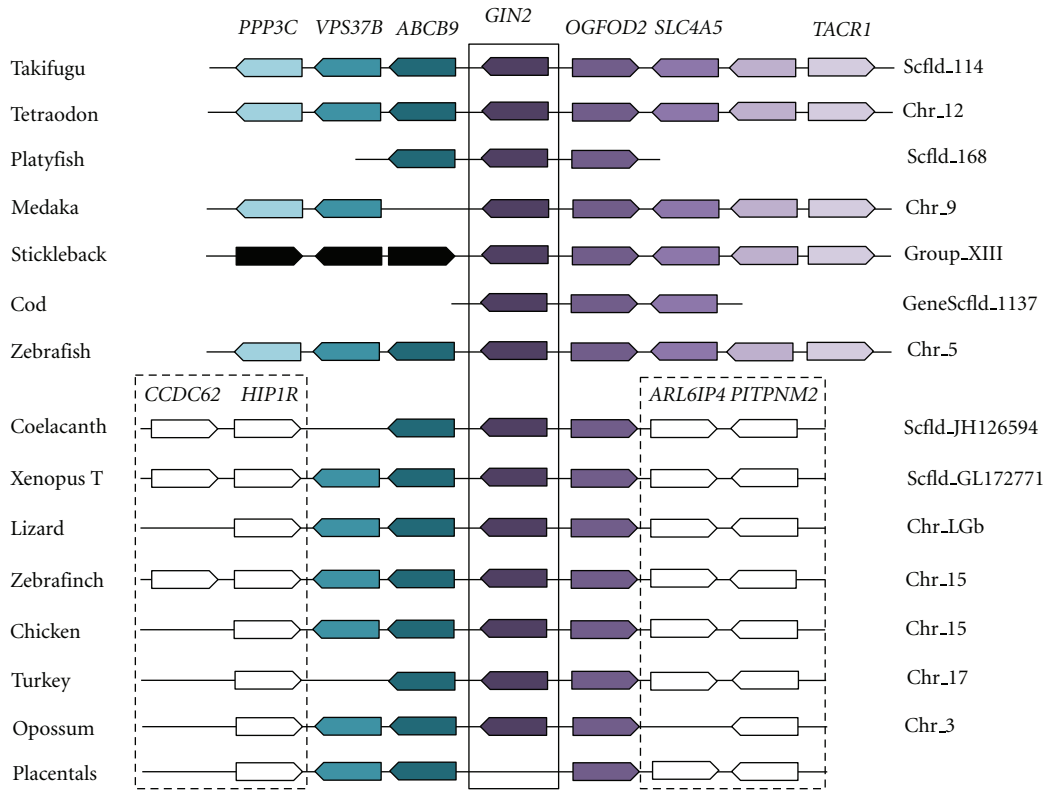


Figure 5: Comparison of *GIN2*-containing genomic regions in vertebrates. Synteny analysis was performed using Ensembl (http://www.ensembl.org/index.html), Genomicus (http://www.dyogen.ens.fr/genomicus-66.01/cgi-bin/search.pl), and BLAST analysis (http://blast.ncbi.nlm.nih.gov/Blast.cgi). Genes represented by white boxes are not found in this region in fish. Black boxes from stickleback represent different genes from ABCB9, VPS37B, and PPP3C in other fish. Their Ensembl accession numbers are from right to left: ENSGACG00000013652, ENSGACG00000013659, and ENSGACG13660.

transposition. Elimination of such sequences might occur by genetic drift or might even be selected for transposition or retrotransposition of a domesticated sequence might change its copy number and pattern of expression. Many domesticated sequences have important functions, for example in cell proliferation. Transposition of such a gene might have strongly deleterious consequences for the host, for instance cancer. It might, therefore, be important to immobilize TE-derived genes at fixed position within a genome to control their expression.

In vertebrates, many TE-derived genes are mammal specific, suggesting that molecular domestication probably played an important role in the evolution of this specific sublineage. Accordingly, many domesticated sequences are involved in placenta formation. Other TE-derived genes like *GIN2* are present in some vertebrate sublineages but absent from mammals. In birds, reptiles, amphibians, and fish, domesticated sequences might be more difficult to identify due to the concomitant presence of active TEs within genomes. Availability of additional genome sequences will probably allow the identification of many TE-derived genes specific of these sublineages that contribute to diversification within vertebrates.

We focused on *GIN* genes, a pair of ancient vertebrate domesticated genes for which no function has been identified so far. Both *GIN1* and *GIN2* are derived from GIN transposons that themselves gained their transposase from the integrase of LTR retrotransposons.

GIN1 was detected in mammals, birds, and reptiles, indicating that it was formed in a common ancestor of Amniota ca. 300 million years ago [63]. *GIN2* might be even older, since it was detected in tetrapods, bony fish, and sharks, and possibly in urochordates. The presence of both genes over such long periods of evolution is suggestive of important, so far unknown conserved functions in vertebrates. *GIN2* was lost in a common ancestor of monotremes and placental mammals, suggesting that either GIN2 function was not essential anymore, or that this function is fulfilled now by GIN1 in these sublineages.

The evolutionary scenario having led to the formation of *GIN1* and *GIN2* remains unclear. Presence of conserved intron positions [63] suggests a unique origin followed by duplication and intron gain in a common ancestor of *GIN1* and *GIN2* (paralogy). In this case, *GIN1* would have been lost among others in fish. Alternatively, *GIN1* and *GIN2* might have been generated from two independent

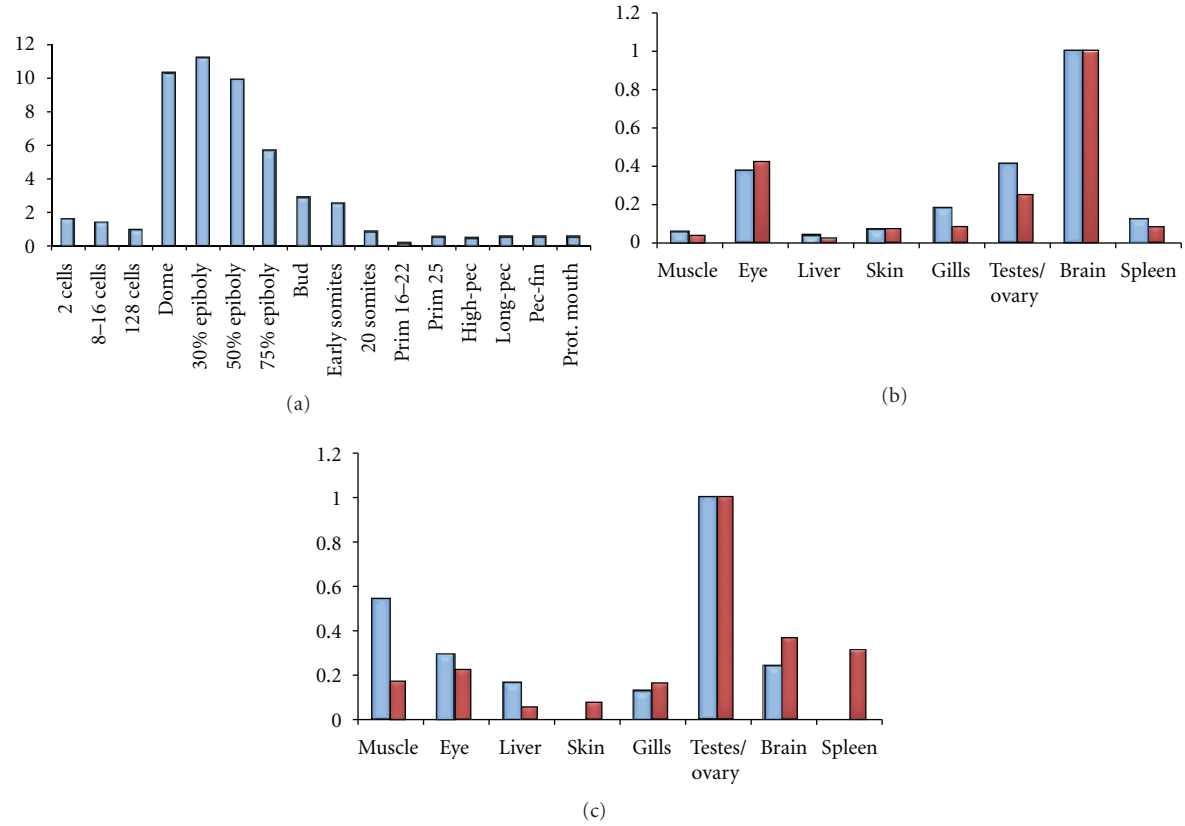


Figure 6: qPCR expression analysis of *GIN2* in zebrafish and platyfish. (a) Expression pattern of *GIN2* during embryonic development in zebrafish. (b) Expression pattern of *GIN2* in adult organs of zebrafish. (c) Expression pattern of *GIN2* in adult organs of platyfish. Multiple RNA extractions using different individuals were performed leading to independent sets of cDNA. Two independent sets and three independent sets of cDNA were tested for embryonic stages and adult organs, respectively. For all sets and for each sample of cDNA, qPCR reaction was done three times (triplicate). One representative experiment is shown with blue bars for male samples and red bars for female samples. *GIN2* expression was normalized using three housekeeping genes: *RPL7*, *beta-actin* and *EF1-alpha*. Analyses were done using the ΔΔCt method [55]. mRNA extractions were done using Trizol and reverse transcription steps were carried out using Fermentas kit. Finally, qPCR was performed using a Bio-Rad kit at the following step: 40 cycles of 94°C and 55°C. Primer sequences are available upon request.

events of molecular domestication, as suggested by the close phylogenetic relationship of *bona fide* GIN transposons with each of both genes (Figure 1). Presence of introns at conserved positions might in this case reflect intron conservation between ancestral GIN transposons at the origin of both molecular domestication events.

GIN1 and GIN2 functions might be related to the binding to DNA or RNA, since both proteins have conserved the HHCC zinc finger present in the ancestral integrase. Conservation of the integrase activity appears possible but must be tested through functional assays. In fish, *GIN2* is particularly expressed in brain and gonads; its expression pattern during zebrafish embryogenesis suggests a role during gastrulation. Functional analysis in fish will provide important insights into the biological function of *GIN2* in vertebrates.

Taken together, data on *GIN* and other TE-derived genes support the important role of molecular domestication as a driver of genetic innovation during evolution. What we have presented here probably only represents the tip of the evolutionary iceberg. There is no doubt that future genome comparisons and functional gene analyses will uncover new domesticated genes and novel biological functions essential for the diversification of vertebrates and other living organisms.

Acknowledgments

The authors' work is supported by grants from the Agence Nationale de la Recherche (ANR).

References

[1] W. F. Doolittle and C. Sapienza, "Selfish genes, the phenotype paradigm and genome evolution," *Nature*, vol. 284, no. 5757, pp. 601–603, 1980.

[2] L. E. Orgel and F. H. C. Crick, "Selfish DNA: the ultimate parasite," *Nature*, vol. 284, no. 5757, pp. 604–607, 1980.

[3] C. Feschotte and C. Gilbert, "Endogenous viruses: insights into viral evolution and impact on host biology," *Nature Reviews Genetics*, vol. 13, no. 4, pp. 283–296, 2012.

[4] H. S. Malik, S. Henikoff, and T. H. Eickbush, "Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses," *Genome Research*, vol. 10, no. 9, pp. 1307–1318, 2000.

[5] D. Ribet, F. Harper, A. Dupressoir, M. Dewannieux, G. Pierron, and T. Heidmann, "An infectious progenitor for the murine IAP retrotransposon: emergence of an intracellular genetic parasite from an ancient retrovirus," *Genome Research*, vol. 18, no. 4, pp. 597–609, 2008.

[6] M. J. Curcio and K. M. Derbyshire, "The outs and ins of transposition: from MU to kangaroo," *Nature Reviews Molecular Cell Biology*, vol. 4, no. 11, pp. 865–877, 2003.

[7] H. Kaessmann, "Origins, evolution, and phenotypic impact of new genes," *Genome Research*, vol. 20, no. 10, pp. 1313–1326, 2010.

[8] J. N. Volff, "Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes," *BioEssays*, vol. 28, no. 9, pp. 913–922, 2006.

[9] E. M. Zdobnov, M. Campillos, E. D. Harrington, D. Torrents, and P. Bork, "Protein coding potential of retroviruses and other transposable elements in vertebrate genomes," *Nucleic Acids Research*, vol. 33, no. 3, pp. 946–954, 2005.

[10] M. Campillos, T. Doerks, P. K. Shah, and P. Bork, "Computational characterization of multiple Gag-like human proteins," *Trends in Genetics*, vol. 22, no. 11, pp. 585–589, 2006.

[11] J. Brandt, S. Schrauth, A. M. Veith et al., "Transposable elements as a source of genetic innovation: expression and evolution of a family of retrotransposon-derived neogenes in mammals," *Gene*, vol. 345, no. 1, pp. 101–111, 2005.

[12] C. Charlier, K. Segers, L. Karim et al., "The callipyge mutation enhances the expression of coregulated imprinted genes in cis without affecting their imprinting status," *Nature Genetics*, vol. 27, no. 4, pp. 367–369, 2001.

[13] R. Ono, S. Kobayashi, H. Wagatsuma et al., "A retrotransposon-derived gene, PEG10, is a novel imprinted gene located on human chromosome 7q21," *Genomics*, vol. 73, no. 2, pp. 232–237, 2001.

[14] S. Suzuki, R. Ono, T. Narita et al., "Retrotransposon silencing by DNA methylation can drive mammalian genomic imprinting," *PLoS Genetics*, vol. 3, no. 4, article e55, 2007.

[15] R. Ono, K. Nakamura, K. Inoue et al., "Deletion of Peg10, an imprinted gene acquired from a retrotransposon, causes early embryonic lethality," *Nature Genetics*, vol. 38, no. 1, pp. 101–106, 2006.

[16] Y. Sekita, H. Wagatsuma, K. Nakamura et al., "Role of retrotransposon-derived imprinted gene, Rtl1, in the feto-maternal interface of mouse placenta," *Nature Genetics*, vol. 40, no. 2, pp. 243–248, 2008.

[17] M. Schüller, D. Jenne, and R. Voltz, "The human PNMA family: novel neuronal proteins implicated in paraneoplastic neurological disease," *Journal of Neuroimmunology*, vol. 169, no. 1-2, pp. 172–176, 2005.

[18] J. Dalmau, S. H. Gultekin, R. Voltz et al., "Ma1, a novel neuron- and testis-specific protein, is recognized by the serum of patients with paraneoplastic neurological disorders," *Brain*, vol. 122, pp. 27–39, 1999.

[19] H. L. Chen and S. R. D'Mello, "Induction of neuronal cell death by paraneoplastic Ma1 antigen," *Journal of Neuroscience Research*, vol. 88, no. 16, pp. 3508–3519, 2010.

[20] K. O. Tan, N. Y. Fu, S. K. Sukumaran et al., "MAP-1 is a mitochondrial effector of Bax," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 41, pp. 14623–14628, 2005.

[21] L. C. Edelstein and T. Collins, "The SCAN domain family of zinc finger transcription factors," *Gene*, vol. 359, no. 1-2, pp. 1–17, 2005.

[22] R. O. Emerson and J. H. Thomas, "Gypsy and the birth of the SCAN domain," *Journal of Virology*, vol. 85, no. 22, pp. 12043–12052, 2011.

[23] D. Ivanov, J. R. Stone, J. L. Maki, T. Collins, and G. Wagner, "Mammalian SCAN domain dimer is a domain-swapped homolog of the HIV capsid C-terminal domain," *Molecular Cell*, vol. 17, no. 1, pp. 137–143, 2005.

[24] T. L. Sander, K. F. Stringer, J. L. Maki, P. Szauter, J. R. Stone, and T. Collins, "The SCAN domain defines a large family of zinc finger transcription factors," *Gene*, vol. 310, no. 1-2, pp. 29–38, 2003.

[25] S. Best, P. L. Tissier, G. Towers, and J. P. Stoye, "Positional cloning of the mouse retrovirus restriction gene Fv1," *Nature*, vol. 382, no. 6594, pp. 826–829, 1996.

[26] H. S. Malik, "Retroviruses push the envelope for mammalian placentation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 7, pp. 2184–2185, 2012.

[27] M. Sha, X. Lee, X. P. Li et al., "Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis," *Nature*, vol. 403, no. 6771, pp. 785–789, 2000.

[28] S. Blaise, N. De Parseval, L. Bénit, and T. Heidmann, "Genomewide screening for fusogenic human endogenous retrovirus envelopes identifies syncytin 2, a gene conserved on primate evolution," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 22, pp. 13013–13018, 2003.

[29] A. Dupressoir, G. Marceau, C. Vernochet et al., "Syncytin-A and syncytin-B, two fusogenic placenta-specific murine envelope genes of retroviral origin conserved in Muridae," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 3, pp. 725–730, 2005.

[30] O. Heidmann, C. Vernochet, A. Dupressoir, and T. Heidmann, "Identification of an endogenous retroviral envelope gene with fusogenic activity and placenta-specific expression in the rabbit: a new "syncytin" in a third order of mammals," *Retrovirology*, vol. 6, article 107, 2009.

[31] C. Vernochet, O. Heidmann, A. Dupressoir et al., "A syncytin-like endogenous retrovirus envelope gene of the guinea pig specifically expressed in the placenta junctional zone and conserved in Caviomorpha," *Placenta*, vol. 32, no. 11, pp. 885–892, 2011.

[32] G. Cornelis, O. Heidmann, S. Bernard-Stoecklin et al., "Ancestral capture of syncytin-Car1, a fusogenic endogenous retroviral envelope gene involved in placentation and conserved in Carnivora," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 7, pp. E432–E441, 2012.

[33] J. M. Antony, K. K. Ellestad, R. Hammond et al., "The human endogenous retrovirus envelope glycoprotein, syncytin-1, regulates neuroinflammation and its receptor expression in multiple sclerosis: a role for endoplasmic reticulum chaperones in astrocytes," *Journal of Immunology*, vol. 179, no. 2, pp. 1210–1224, 2007.

[34] K. Søe, T. L. Andersen, A. S. Hobolt-Pedersen, B. Bjerregaard Bolette, L. I. Larsson, and J. M. Delaissé, "Involvement of human endogenous retroviral syncytin-1 in human osteoclast fusion," *Bone*, vol. 48, no. 4, pp. 837–846, 2011.

[35] H. Ikeda and H. Sugimura, "Fv-4 resistance gene: a truncated endogenous murine leukemia virus with ecotropic interference properties," *Journal of Virology*, vol. 63, no. 12, pp. 5405–5412, 1989.

[36] A. Marco and I. Marín, "CGIN1: a retroviral contribution to mammalian genomes," *Molecular Biology and Evolution*, vol. 26, no. 10, pp. 2167–2170, 2009.

[37] T. Matsui, K. Miyamoto, A. Kubo et al., "SASPase regulates stratum corneum hydration through profilaggrin-to-filaggrin processing," *EMBO Molecular Medicine*, vol. 3, no. 6, pp. 320–333, 2011.

[38] T. H. Eickbush, "Telomerase and retrotransposons: which came first?" *Science*, vol. 277, no. 5328, pp. 911–912, 1997.

[39] A. Böhne, F. Brunet, D. Galiana-Arnoux, C. Schultheis, and J. N. Volff, "Transposable elements as drivers of genomic and biological diversity in vertebrates," *Chromosome Research*, vol. 16, no. 1, pp. 203–215, 2008.

[40] C. Feschotte and E. J. Pritham, "DNA transposons and the evolution of eukaryotic genomes," *Annual Review of Genetics*, vol. 41, pp. 331–368, 2007.

[41] V. V. Kapitonov and J. Jurka, "RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons," *PLoS Biology*, vol. 3, no. 6, article e181, 2005.

[42] T. Okada, J. I. Ohzeki, M. Nakano et al., "CENP-B controls centromere formation depending on the chromatin context," *Cell*, vol. 131, no. 7, pp. 1287–1300, 2007.

[43] C. Casola, D. Hucks, and C. Feschotte, "Convergent domestication of pogo-like transposases into centromere-binding proteins in fission yeast and mammals," *Molecular Biology and Evolution*, vol. 25, no. 1, pp. 29–41, 2008.

[44] H. P. Cam, K. I. Noma, H. Ebina, H. L. Levin, and S. I. S. Grewal, "Host genome surveillance for retrotransposons by transposon-derived proteins," *Nature*, vol. 451, no. 7177, pp. 431–436, 2008.

[45] M. Zaratiegui, M. W. Vaughn, D. V. Irvine et al., "CENP-B preserves genome integrity at replication forks paused by retrotransposon LTR," *Nature*, vol. 469, no. 7328, pp. 112–115, 2011.

[46] A. F. A. Smit and A. D. Riggs, "Tiggers and other DNA transposon fossils in the human genome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 4, pp. 1443–1448, 1996.

[47] W. Liu, J. Seto, G. Donovan, and M. Toth, "Jerky, a protein deficient in a mouse epilepsy model, is associate with translationally inactive mRNA in neurons," *Journal of Neuroscience*, vol. 22, no. 1, pp. 176–182, 2002.

[48] F. Butter, D. Kappei, F. Buchholz, M. Vermeulen, and M. Mann, "A domesticated transposon mediates the effects of a single-nucleotide polymorphism responsible for enhanced muscle growth," *EMBO Reports*, vol. 11, no. 4, pp. 305–311, 2010.

[49] E. Markljung, L. Jiang, J. D. Jaffe et al., "ZBED6, a novel transcription factor derived from a domesticated DNA transposon regulates IGF2 expression and muscle growth," *PLoS Biology*, vol. 7, no. 12, Article ID e1000256, 2009.

[50] R. Cordaux, S. Udit, M. A. Batzer, and C. Feschotte, "Birth of a chimeric primate gene by capture of the transposase gene from a mobile element," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 21, pp. 8101–8106, 2006.

[51] M. Shaheen, E. Williamson, J. Nickoloff, S. H. Lee, and R. Hromas, "Metnase/SETMAR: a domesticated primate transposase that enhances DNA repair, replication, and decatenation," *Genetica*, vol. 138, no. 5, pp. 559–566, 2010.

[52] A. Sarkar, C. Sim, Y. S. Hong et al., "Molecular evolutionary analysis of the widespread piggyBac transposon family and related "domesticated" sequences," *Molecular Genetics and Genomics*, vol. 270, no. 2, pp. 173–180, 2003.

[53] J. C. Newman, A. D. Bailey, H. Y. Fan, T. Pavelitz, and A. M. Weiner, "An abundant evolutionarily conserved CSB-PiggyBac fusion protein expressed in cockayne syndrome," *PLoS Genetics*, vol. 4, no. 3, Article ID e1000031, 2008.

[54] V. V. Kapitonov and J. Jurka, "Harbinger transposons and an ancient HARBI1 gene derived from a transposase," *DNA and Cell Biology*, vol. 23, no. 5, pp. 311–324, 2004.

[55] A. Böhne, Q. Zhou, A. Darras et al., "Zisupton—a novel superfamily of DNA transposable elements recently active in fish," *Molecular Biology and Evolution*, vol. 29, no. 2, pp. 631–645, 2012.

[56] S. E. Hammer, S. Strehl, and S. Hagemann, "Homologs of *Drosophila* P transposons were mobile in zebrafish but have been domesticated in a common ancestor of chicken and human," *Molecular Biology and Evolution*, vol. 22, no. 4, pp. 833–844, 2005.

[57] T. Clouaire, M. Roussigne, V. Ecochard, C. Mathe, F. Amalric, and J. P. Girard, "The THAP domain of THAP1 is a large C2CH module with zinc-dependent sequence-specific DNA-binding activity," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 19, pp. 6907–6912, 2005.

[58] J. B. Parker, S. Palchaudhuri, H. Yin, J. Wei, and D. Chakravarti, "A transcriptional regulatory role of the THAP11-HCF-1 complex in colon cancer cell function," *Molecular and Cellular Biology*, vol. 32, no. 9, pp. 1654–1670, 2012.

[59] M. Roussigne, C. Cayrol, T. Clouaire, F. Amalric, and J. P. Girard, "THAP1 is a nuclear proapoptotic factor that links prostate-apoptosis-response-4 (Par-4) to PML nuclear bodies," *Oncogene*, vol. 22, no. 16, pp. 2432–2442, 2003.

[60] M. Roussigne, S. Kossida, A. C. Lavigne et al., "The THAP domain: a novel protein motif with similarity to the DNA-binding domain of P element transposase," *Trends in Biochemical Sciences*, vol. 28, no. 2, pp. 66–69, 2003.

[61] A. Sabogal, A. Y. Lyubimov, J. E. Corn, J. M. Berger, and D. C. Rio, "THAP proteins target specific DNA sites through bipartite recognition of adjacent major and minor grooves," *Nature Structural & Molecular Biology*, vol. 17, no. 1, pp. 117–123, 2010.

[62] C. Lloréns and I. Marín, "A mammalian gene evolved from the integrase domain of an LTR retrotransposon," *Molecular Biology and Evolution*, vol. 18, no. 8, pp. 1597–1600, 2001.

[63] I. Marín, "GIN transposons: genetic elements linking retrotransposons and genes," *Molecular Biology and Evolution*, vol. 27, no. 8, pp. 1903–1911, 2010.

[64] W. Bao, V. V. Kapitonov, and J. Jurka, "Ginger DNA transposons in eukaryotes and their evolutionary relationships with long terminal repeat retrotransposons," *Mobile DNA*, vol. 1, no. 1, article 3, 2010.

[65] S. Guindon, J. F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel, "New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0," *Systematic Biology*, vol. 59, no. 3, pp. 307–321, 2010.

[66] M. A. Larkin, G. Blackshields, N. P. Brown et al., "Clustal W and clustal X version 2.0," *Bioinformatics*, vol. 23, no. 21, pp. 2947–2948, 2007.

[67] H. S. Malik and T. H. Eickbush, "Modular evolution of the integrase domain in the Ty3/Gypsy class of LTR retrotransposons," *Journal of Virology*, vol. 73, no. 6, pp. 5186–5190, 1999.

[68] C. B. Kimmel, W. W. Ballard, S. R. Kimmel, B. Ullmann, and T. F. Schilling, "Stages of embryonic development of the zebrafish," *Developmental Dynamics*, vol. 203, no. 3, pp. 253–310, 1995.

Hindawi Publishing Corporation
International Journal of Evolutionary Biology
Volume 2012, Article ID 298147, 12 pages
doi:10.1155/2012/298147

Research Article

Evolution of the FGF Gene Family

Silvan Oulion, Stephanie Bertrand, and Hector Escriva

CNRS, UMR 7232, BIOM, Université Pierre et Marie Curie Paris 06, Observatoire Océanologique, 66650 Banyuls-sur-Mer, France

Correspondence should be addressed to Hector Escriva, hescriva@obs-banyuls.fr

Received 27 April 2012; Accepted 6 June 2012

Academic Editor: Frédéric Brunet

Fibroblast Growth Factors (FGFs) are small proteins generally secreted, acting through binding to transmembrane tyrosine kinase receptors (FGFRs). Activation of FGFRs triggers several cytoplasmic cascades leading to the modification of cell behavior. FGFs play critical roles in a variety of developmental and physiological processes. Since their discovery in mammals, FGFs have been found in many metazoans and some arthropod viruses. Efforts have been previously made to decipher the evolutionary history of this family but conclusions were limited due to a poor taxonomic coverage. We took advantage of the availability of many new sequences from diverse metazoan lineages to further explore the possible evolutionary scenarios explaining the diversity of the FGF gene family. Our analyses, based on phylogenetics and synteny conservation approaches, allow us to propose a new classification of FGF genes into eight subfamilies, and to draw hypotheses for the evolutionary events leading to the present diversity of this gene family.

1. Introduction

Fibroblast growth factors (FGFs) form a family of generally extracellular signaling peptides, which are key regulators of many biological processes ranging from cell proliferation to the control of embryonic development in metazoans. Ever since the mitogenic activity of FGF-like factors was first observed in 1939 [1] and the first FGF factor was isolated in the 1970s [2], a large number of members of this gene family have been isolated and characterized in different metazoans.

FGFs are small proteins (between 17 and 34 kDa) characterized by a relatively well conserved central domain of 120 to 130 amino acids. This domain is organized into 12 antiparallel β sheets forming a triangular structure called beta trefoil. In general, FGFs function through binding to a tyrosine kinase receptor (FGFR) on the surface of the cell membrane. Two FGF ligands bind a dimeric receptor in the presence of heparan sulphate proteoglycan (HSPG) allowing the transphosphorylation and activation of the intracellular tyrosine kinase domain of the receptor. Binding to FGFRs usually activates several intracellular cascades (i.e., Ras/MAPK, PI3K/Akt, and PLCγ/PKC) which may regulate the transcription of different target genes. Through the activation of these cytoplasmic pathways, the FGF signal controls several major cellular functions such as cell proliferation, migration, differentiation, or survival. An intracellular mode of action has also been described in the case of FGF1 but it is poorly documented [3].

Concerning the evolutionary history of the FGF gene family, several studies using molecular phylogenetics as well as synteny conservation analyses have been performed [4–8]. The first phylogeny-based classifications of the gene family were proposed before the whole complement of FGF genes was described in mammals which led to incomplete conclusions [5, 8]. The first phylogenetic studies including all the mammalian FGFs proposed a division of the gene family into six [9] or seven [6] subfamilies. In 2005, Popovici and collaborators performed the first study including both protostome and deuterostome FGFs as well as FGFs from baculoviruses, an arthropod-specific group of viruses [4]. They proposed to divide the FGF gene family into eight subfamilies: subfamily A (including orthologs of FGF 1 and 2), subfamily B (orthologs of FGF 3, 7, 10, and 22), subfamily C (orthologs of FGF 4, 5, and 6), subfamily D (orthologs of FGF 8, 17, 18, and 24 from vertebrates but also of EGL-17, PYR, and THS from protostomes), subfamily E (orthologs of FGF 9, 16, and 20 but also of LET-756 from nematodes), subfamily F (orthologs of FGF 11, 12, 13, and 14), subfamily G (orthologs of FGF 15/19, 21, and 23), and subfamily H which is specific of arthropod FGFs (i.e., BNL) and of FGFs

found in arthropod-specific viruses [4]. This classification is widely accepted today, however, the phylogenetic position of FGF3 and FGF5 is not completely solved, which calls into question the constitution of the two subfamilies B and C. Moreover, the description of FGF genes in the sea anemone *Nematostella vectensis* now raises the question of the timing of the appearance and diversification of the FGF gene family.

In this study we take advantage of the exponential increase of publicly available genomic sequences to present an update of the FGF gene content in different evolutionary lineages. Phylogenetic approaches, together with synteny conservation analyses of these data, allow us to propose a new classification of the FGF gene family which (i) confirms the paralogy relationships of the FGF4/5/6 subfamily members and (ii) suggest that orthologs of the mammalian FGF3 form a new subfamily.

2. The FGF Gene Content Varies among Different Metazoan Lineages

The recent development of high throughput sequencing techniques has generated a large number of sequences available in different public databases. Among them we have searched for FGF domain coding sequences within the major metazoan phyla, in order to clarify the evolutionary history of this family. We have limited our study to the analysis of amino acid sequences deposited in the Genbank, the Ensembl, and the JGI databases for cnidarians, lophotrochozoans, ecdysozoans, and deuterostomes, although many ESTs sequences putatively coding for FGF proteins might also be found.

2.1. FGF Genes in Diploblastic Metazoans. FGF genes were previously described in two anthozoan species: *Nematostella vectensis* and *Acropora millepora* [10, 11]. In *Nematostella*, 13 genes encoding FGF ligands were predicted from the genome sequence [11] but their phylogenetic relationships with bilaterian FGFs are not fully established. Four of these genes group with the FGF8/17/18/24 subfamily and six group with the FGF1/2 subfamily with low support. In the hydrozoan *Hydra magnipapillata* we have found 4 predicted genes coding for FGFs (see Table 1). Among them, one (called FGF24) belongs to the FGF8/17/18/24 subfamily. Another one groups with several *Nematostella* FGF genes whose position is not robustly supported but might belong to the FGF1/2 subfamily (see Figure S1 in supplementary material available online at doi:10.1155/2012/298147). For the other two, no clear relationship with either *Nematostella* or bilaterian FGFs can be proposed according to phylogenetic reconstructions. We also looked for ctenophore EST sequences putatively encoding FGF domains but we failed to find any in public databases.

2.2. FGF Genes in Protostomes. In protostomes, FGF genes have only been described in ecdysozoans, particularly in arthropods. Three genes have been characterized in the model organism *Drosophila melanogaster* [12, 13], called *Branchless* (*Bnl*), *Thisbe* (*Ths*), and *Pyramus* (*Pyr*). In the coleopteran *Tribolium castaneum*, four FGF genes called *Tc-FGF1a, Tc-FGF1b, Tc-FGF8,* and *Tc-Bnl* [14] have also been identified. *Ths* and *Pyr* from *Drosophila*, as well as *Tc-FGF8* from *Tribolium*, were shown to belong to the FGF8/17/18/24 subfamily, whereas *Tc-FGF1a*, and *TcFGF1b* belong to the FGF1/2 subfamily. On the other hand, *Branchless* orthologs from both species show no clear evolutionary relationships with any of the vertebrates FGF gene subfamilies leading Popovici and collaborators to propose a new subfamily including *Bnl* from arthropods and baculovirus-specific FGF genes [4]. In the genome of the nematode *Caenorhabditis elegans* two FGF genes are found called *let-756* (lethal protein 756) and *egl-17* (egg laying defective 17) [4, 15], which are members of the FGF9/16/20 and FGF8/17/18/24 subfamilies, respectively [4].

In order to obtain a more complete picture of the diversity of the FGF gene family in ecdysozoans, we searched other available sequences (see Table 1). Thus, in different nematode species we only found orthologs of the two known *C. elegans* genes (Figure S2). In arthropods, we found FGF coding genes in the crustacean *Daphnia pulex*, in the chelicerate *Ixodes scapularis*, and in insects from different classes such as *Apis mellifera*, *Harpegnathos saltator,* or *Pediculus humanus* (see Table 1). The orthology relationships of the two FGF genes we found in *Daphnia* cannot be clearly determined, whereas for all the other arthropods the different genes we found always belong to the *Bnl*, FGF1/2, or FGF8/17/18/24 subfamilies (Figure S2).

No study of the FGF gene set in lophotrochozoans has been published yet so we searched for lophotrochozoan FGF coding sequences in Genbank and in the complete genome sequences of the mollusc *Lottia gigantea* and of the annelids *Helobdella robusta* and *Capitella teleta*. We found only one gene in *Capitella* whose position in the FGF phylogenetic tree is not robustly supported, but probably belongs to the FGF8/17/18/24 subfamily. In *Lottia gigantea*, two FGF genes are present in the complete genome, and again their evolutionary relationship with the different subfamilies cannot be clearly determined even if the best blast hit results for these genes are always orthologs of the FGF8/17/18/24 and FGF9/16/20 subfamilies (see Table 1). Taken together, these data demonstrate (i) that lophotrochozoans also possess some FGF coding genes, although quite divergent from the other protostome genes, and (ii) that members of only four subfamilies, FGF1/2, FGF8/17/18/24, FGF9/16/20, and *Bnl*, can be clearly found in protostomes.

2.3. FGF Genes in Deuterostomes. Deuterostomes comprise vertebrates, the related invertebrate chordates (urochordates and cephalochordates) and three other invertebrate taxa: hemichordates and echinoderms, which form the Ambulacraria group, and the recently described phylum of Xenoturbellida [16]. Nothing is known concerning the FGF gene content in Xenoturbella and we did not find any FGF coding sequence for this group. Conversely, recent studies have shown that one FGF gene exists in the sea urchin *Strongylocentrotus purpuratus* (i.e., echinoderm) [17], and we have identified in the databases six FGF genes in the hemichordate *Saccoglossus kowalevskii* of which one gene can

Table 1: FGF domain containing protein sequences used in this study. For each species the accession number of all the proteins found are given, as well as orthology when well supported in phylogenetic reconstructions. Best blastP hit are given when no clear orthologu relationships was found.

Species	Accession number	Description	Database	Best blastP hit accession	Best blastP hit name	Orthology
Hydra magnipapillata	XP_002165496.1	Predicted: similar to fibroblast growth factor homologous factor 4	Genbank	NP_001180935.1	Fibroblast growth factor 12 (*Macaca mulatta*)	
	XP_002164870.1	Predicted: similar to fibroblast growth factor 24	Genbank			FGF8/17/18
	XP_002166704.1	Predicted: similar to fibroblast growth factor 1B, partial	Genbank			FGF1/2
	XP_002170051.1	Predicted: similar to Fibroblast growth factor 14 (*Hydra magnipapillata*)	Genbank	XP_001094679.1	Predicted: fibroblast growth factor 20 (*Macaca mulatta*)	
Lottia gigantea	fgenesh2_pg.C_sca_110000014		JGI	XP_002643284.1	EGL-17 (*Caenorhabditis briggsae*)	
	fgenesh2_pg.C_sca_16000265		JGI	XP_003455818.1	Predicted: fibroblast growth factor 20-like (*Oreochromis niloticus*)	
Capitella teleta	fgeneshl_pg.C_scaffold_1200000001		JGI	XP_002922927.1	Predicted: fibroblast growth factor 18-like (*Ailuropoda melanoleuca*)	
Trichinella spiralis	XP_003370033	Fibroblast growth factor 20	Genbank	NP_001098209.1	Fibroblast growth factor 20a (*Oryzias latipes*)	
	EFV50493.1	Fibroblast growth factor 18	Genbank			FGF8/17/18
Brugia malayi	XP_001894505.1	Fibroblast growth factor family protein	Genbank			FGF9/16/20
	XP_001899322.1	Fibroblast growth factor family protein	Genbank			FGF8/17/18
Apis mellifera	XP_623927.2	Predicted: hypothetical protein LOC551529	Genbank			FGF1/2
	XP_001120331.2	Predicted: hypothetical protein LOC724469	Genbank			BNL
	XP_003695580.1	Predicted: fibroblast growth factor 18-like	Genbank			FGF8/17/18
Harpegnathos saltator	EFN80858.1	Hypothetical protein EAI_11890	Genbank	XP_003399646.1	Predicted: hypothetical protein LOCI00646960 (*Bombus terrestris*)	
	EFN81752.1	Fibroblast growth factor 18	Genbank			FGF8/17/18
	EFN88402.1	Heparin-binding growth factor 1	Genbank			FGF1/2

Table 1: Continued.

Species	Accession number	Description	Database	Best blastP hit accession	Best blastP hit name	Orthology
Pediculus humanus subsp. corporis	EEB17861.1	Fibroblast growth factor, putative	Genbank	XP_003243356.1	Predicted: hypothetical protein LOC100572243 (*Acyrthosiphon pisum*)	
	EEB19433.1	Heparin-binding growth factor 1 precursor, putative	Genbank			FGF1/2
	EEB18362.1	Conserved hypothetical protein	Genbank	XP_002431100.1	Predicted: hypothetical protein LOC 100569010 (*Acyrthosiphon pisum*)	
Ixodes scapularis	XP_002433492.1	Hypothetical protein IscW_ISCW015993	Genbank	XP_003203489.1	Predicted: glia-activating factor-like (*Meleagris gallopavo*)	
	XP_002400933.1	Heparin-binding growth factor, putative	Genbank			FGF1/2
Daphnia pulex	EFX75093.1	Hypothetical protein DAPPUDRAFT_108237	Genbank	XP_003243356.1	Predicted: hypothetical protein LOC 100572243 (*Acyrthosiphon pisum*)	
	EFX86332.1	Hypothetical protein DAPPUDKAFT_98099	Genbank	XP_001635198.1	Predicted protein (*Nematostella vectensis*)	
Saccoglossus kowalevskii	ADB22412.1	Fibroblast growth factor 8/17/18 protein	Genbank			FGF8/17/18
	ADB22409.1	Hypothetical protein	Genbank	XP_799351.2		
	ACY92516.1	Fgf-Sk1 protein	Genbank	NP_001233192.1		
	ACY92517.1	FGF9-like protein	Genbank			FGF9/16/20
	ACY92515.1	FGF13-like protein	Genbank			FGF9/16/20
	ADB22411.1	Fibroblast growth factor 20-like protein	Genbank			FGF9/16/20
Oikopleura dioica	CBY43668.1	Unnamed protein product	Genbank	XP_003441021.1	Predicted: fibroblast growth factor 14-like (*Oreochromis niloticus*)	
	CBY37156.1	Unnamed protein product	Genbank			FGF11/12/13/14
	CBY40156.1	Unnamed protein product	Genbank			FGF11/12/13/14
	CBY12333.1	Unnamed protein product	Genbank			FGF9/16/20
	CBY34733.1	Unnamed protein product	Genbank	NP_001007762.1	Keratinocyte growth factor precursor (*Danio rerio*)	
	CBY23701.1	Unnamed protein product	Genbank	XP_002594626.1	Hypothetical protein BRAFLDRAFT_149779 (*Branchiostoma floridae*)	

be clearly assigned to the FGF8/17/18/24 subfamily. Three other genes are orthologs of the FGF9/16/20 subfamily, indicating that an hemichordate-specific duplication occurred for this gene; another one has been previously shown to be ortholog of the FGF19/21/23 [18]; the sixth gene shows no clear orthology relationships with any FGF gene subfamily (see Table 1) [18].

In chordates, the FGF gene content is also different among the three subphyla. In cephalochordates, eight FGF genes have been found and orthology relationships using phylogenetics or conservation of synteny approaches have been suggested for six of them (i.e., FGF1/2, FGF8/17/18, FGF9/16/20, FGFA ortholog of FGF3/7/10/22, FGFB ortholog of FGF4/5/6, and FGFC ortholog of FGF19/21/23) [19]. In the urochordate *Ciona intestinalis*, six genes encoding FGF ligands have been described [20], and we identified one more gene in databases, called FGF-NA1, bringing the total FGF gene content to seven. Of them, only two were shown to be clear orthologs of the FGF8/17/18/24 and FGF11/12/13/14 subfamilies [20]. In another urochordate, the larvacean *Oikopleura dioica*, we found six FGF coding genes, among which two can be assigned to the FGF11/12/13/14 subfamily, and one to the FGF9/16/20 subfamily (see Table 1 and Figure S4). In vertebrates, an explosion in the number of genes encoding FGFs occurred and we can find between 19 and 27 FGF genes depending on the species. This explosion is not specific to the FGF gene family and is linked to the two rounds of genome duplication (three rounds in teleosts) that occurred in this lineage as previously demonstrated [4, 21]. In sarcopterygians we identified 19 FGF genes in the chicken and 23 in the coelacanth, whereas 22 FGF genes (FGF 1–23) have been characterized in mouse and human (the mouse FGF15 is the ortholog of the human FGF19). These 22 mammalian genes were previously used to reconstruct the evolutionary history of the family [4, 6], which led to the classification of FGFs into seven paralogy groups. However, in teleosts, an additional round of genome duplication (3R hypothesis) occurred [22], which, together with a high number of FGF gene losses, produced 27 FGF genes in the zebrafish [23].

3. The FGF Gene Family Is Composed by Eight Subfamilies

Due to the low sequence conservation of most of the FGF genes found in early divergent metazoan lineages, and the short length of the FGF domain, we have based our phylogenetic study on vertebrate FGFs, as in previous studies [4, 6]. However, the new FGF sequence data, particularly within chordates, allow us to suggest a new classification of the FGF gene family in metazoans, which is divided into 8 subfamilies instead of 7 (in addition to the arthropod + baculoviruses—specific family proposed by Popovici et al. [4]). These families are the FGF1/2, FGF3, FGF4/5/6, FGF7/10/22, FGF8/17/18/24, FGF9/16/20, FGF11/12/13/14 and FGF19/21/23 (Figures 1 and S5).

In all the studies performed so far, the vertebrate FGF3 always grouped into either the subfamily FGF3/7/10/22 or the subfamily FGF3/4/6 [4, 6, 8]. In fact, the correct classification of FGF3 is still debated and assignment to one or another subfamily depends on the methods used. Therefore, most of the phylogenetic analyses published grouped FGF3 with FGF7, FGF10, and FGF22, but with very low node robustness. Other studies, using the genomic locations of this gene, grouped it with FGF4 and FGF6 and it has even been suggested that the FGF3/4/6 and FGF19/21/23 subfamilies can be assembled into a single subfamily FGF3/4/6/19/21/23 (with FGF5 grouping in this case with the FGF1/2 subfamily) [7]. Here, based particularly on results obtained through the study of gene content, phylogenetic distribution, and conservation of synteny between amphioxus and vertebrates [19], we propose a new evolutionary scenario in which FGF3 forms a new subfamily (Figures 1, 2, and S5). This scenario could reconcile the different evolutionary hypotheses suggested in previous studies.

In our hypothesis, an ancestral FGF gene (named FGF3/4/5/6) was duplicated in tandem before chordate diversification. Such duplication might have occurred before eumetazoan diversification or specifically in the chordate ancestor. Thus, the putative ancestor (either eumetazoan or chordate ancestor) had two FGF genes maintained in cluster: FGF3 and FGF4/5/6. This situation can still be observed in the cephalochordate *Branchiostoma floridae* in which FGFB and FGFE are clustered in a genomic region showing synteny conservation with the vertebrate locus containing the FGFs 3, 4 and 6 [19] (Figure 3). This hypothesis implies a loss of FGF3 in different lineages, the number of lineages that lost FGF3 depends on the timepoint at which this gene appeared (i.e., in urochordates in one hypothesis (Figures 2(b) and 5), or in urochordates, ambulacrarians, protostomes, and cnidarians in the other hypothesis, see Figure 5). According to this scenario the origin of FGF3 would be ancient (i.e., at least prior to chordates diversification) and not due to the vertebrate-specific genome duplications.

Another FGF gene whose phylogenetic position is debated is FGF5. Indeed, depending on the phylogenetic approach and on the gene set used for the phylogenetic reconstruction, it clusters either with FGF4/6 or with FGF1/2 [4, 23]. Moreover, conservation of synteny also suggests the paralogy of FGF1, 2, and 5 [7]. However, a deeper synteny analysis of the human FGF5 locus shows conservation of this locus with both the FGF1/2 and FGF4/6 loci (Figure 3). This mixed syntenic conservation, together with our phylogenetic analyses supporting the FGF4/5/6 subfamily (Figure 1), suggests that FGF5 is a real paralog of FGF4 and 6. The partial synteny conservation with the FGF1 and 2 loci might be explained by a genomic translocation of the FGF5 locus (including its neighbouring genes BMP3, PAQR3) close to the ANXA3 locus (Figures 2(a) and 3).

4. The Evolutionary History of the FGF Gene Family Is Characterized by Gene Duplications and Gene Losses

Phylogenetic reconstructions using FGF sequences from all metazoan phyla often fail to completely solve the orthology relationship between the different members of this family

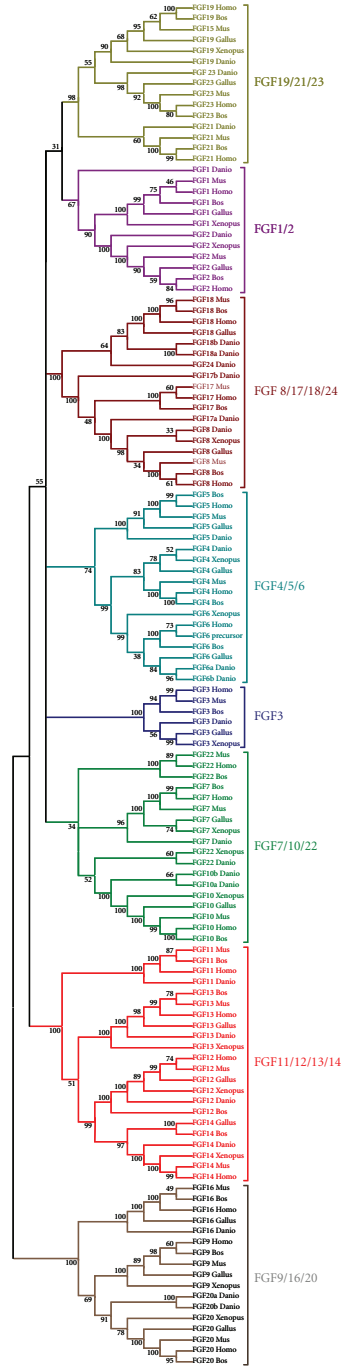


Figure 1: FGF phylogeny in vertebrates. Neighbor-joining tree showing the classification into eight subfamilies of the different vertebrate FGF genes (i.e., FGF1/2, FGF3, FGF4/5/6, FGF7/10/22, FGF8/17/18/24, FGF9/16/20, FGF11/12/13/14, and FGF19/21/23). Sequences of *Homo sapiens*, *Mus musculus*, *Bos taurus*, *Gallus gallus*, *Xenopus tropicalis*, and *Danio rerio* were used to perform the phylogeny.

mainly because of the reduced size of the FGF domain and because of the high divergence of the sequences between the different lineages. However, using the phylogenetic distribution of FGF genes into eight subfamilies, we can propose evolutionary scenarios accounting for the FGF gene content found in the different metazoan lineages. Several hypotheses can be drawn explaining such a distribution of FGF orthologs. Here we focus mainly on two of these hypotheses: a first hypothesis where the eight FGF subfamilies are chordate-specific (Figures 4 and 5, hypothesis 1) and a second hypothesis where the eight subfamilies were ancestral to all eumetazoans (Figure 5, hypothesis 2). In both hypotheses, the evolutionary history of the FGF gene content in chordates is the same (Figure 4), but depending on the hypothesis, it changes for the other metazoan lineages (Figure 5).

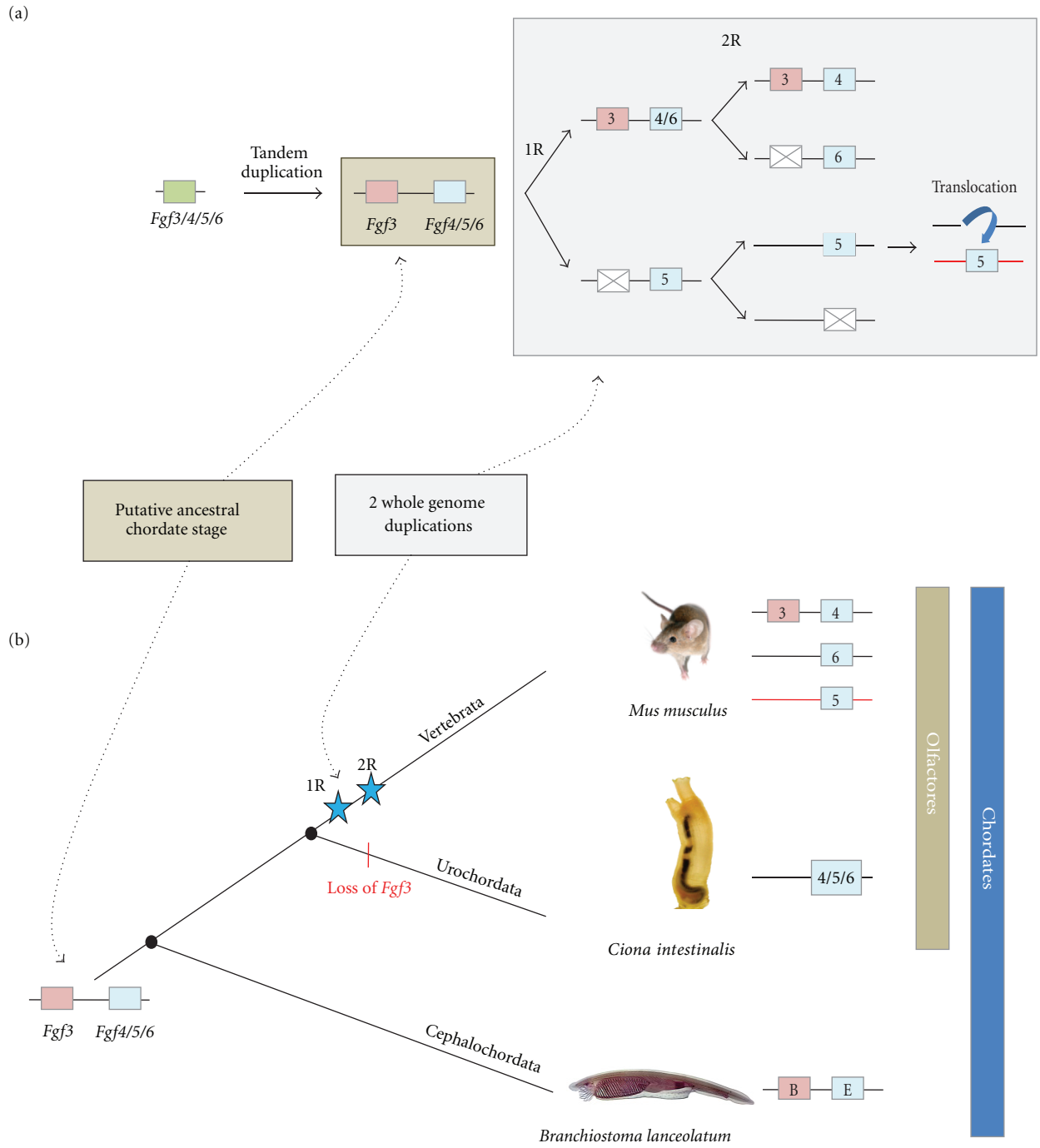


Figure 2: Evolutionary scenario of the FGF3 subfamily. (a) Genomic events leading to the birth of the FGF3 subfamily. From a single *FGF3/4/5/6* gene, a tandem duplication occurred before the chordate diversification giving rise to an *FGF3* and an *FGF4/5/6* gene (brown box). The two rounds of whole genome duplication, followed by several gene losses and by a specific translocation of the chromosome region containing *FGF5* (grey box) conducted to the gene content currently found in vertebrates. (b) Evolutionary relationships between FGFs 3, 4, 5, and 6 in chordates. Here, the chordate ancestor had both *FGF3* and *FGF4/5/6*. This gene content was kept in amphioxus, whereas *FGF3* was lost in urochordates and different gene losses account in vertebrates for the presence of a single FGF3 gene and three genes of the FGF4/5/6 paralogy group. This implies that in amphioxus *FGF3* and *FGFB* are orthologs, as well as *FGF4/5/6* and *FGFE*.

As we have shown, in cnidarians (diploblastic metazoans) we found the presence of, at least, orthologs of the FGF8/17/18 and probably FGF1/2 subfamilies. Thus, we can suggest that the eumetazoan ancestor possessed at least one ortholog of these two subfamilies.

Our analyses suggest that the arthropod ancestor already possessed at least three FGF genes belonging to the FG1/2, FGF8/17/18 and *Bnl* subfamilies (Figure 5). *Bnl* is specific to arthropods and arthropod viruses and its origin is still unknown. Two possible evolutionary scenarios can be drawn

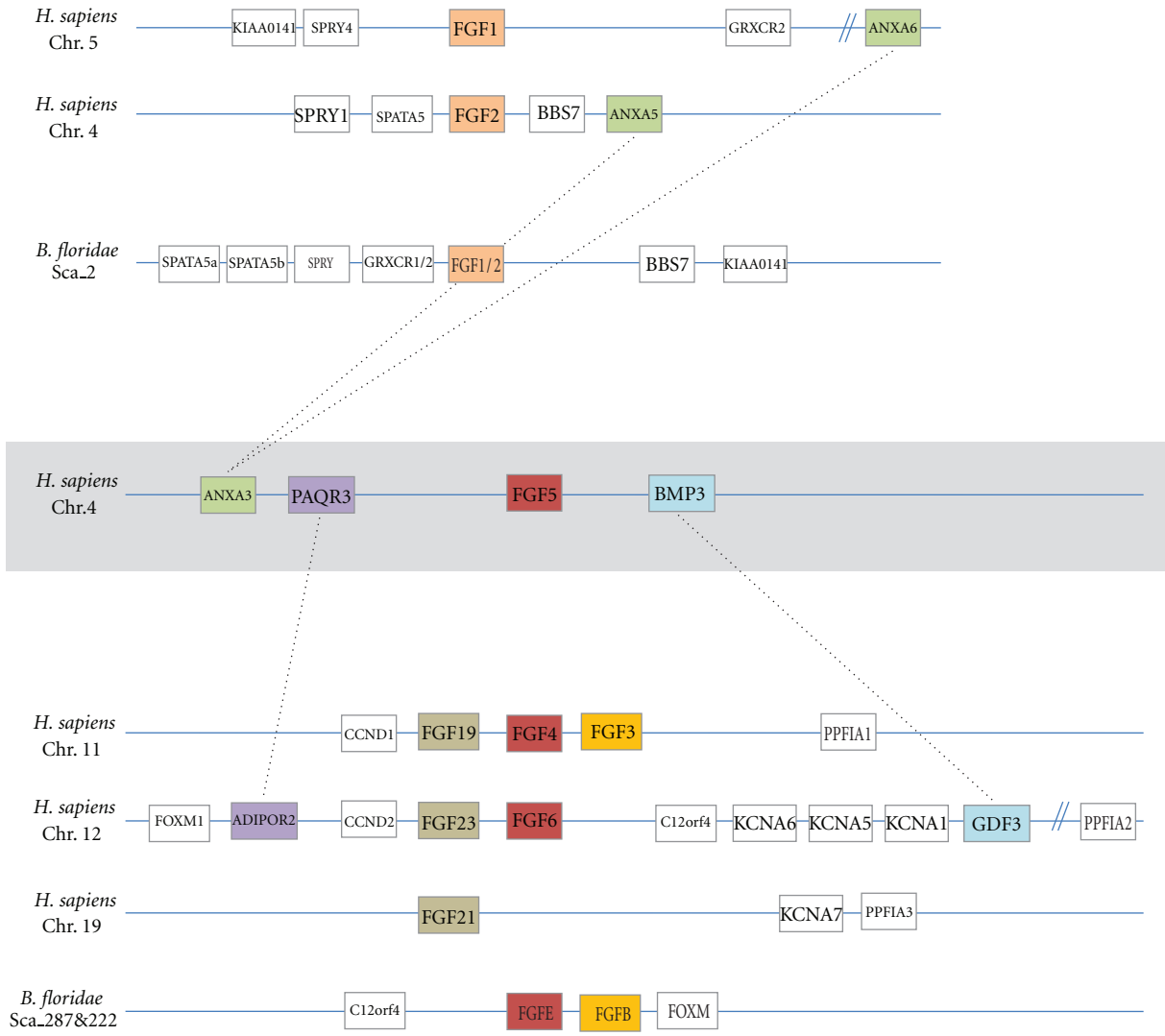


Figure 3: Chromosomal maps of human and amphioxus FGF1/2 and FGF4/5/6 genes loci. Synteny is well conserved among vertebrates and amphioxus for FGF1/2 (orange—upper part) and for FGF4/6 (red), which are also syntenic with FGFs 19/21/23 (brown) and with FGF3 (yellow—lower part). The synteny of FGF5 with BMP3, PAQR3, and ANXA3 suggests that this gene belongs to the FGF4/5/6 subfamily, but was probably secondarily translocated with his neighboring genes (BMP3, PAQR3, etc.) close to ANXA3.

for *Bnl* genes. In the first scenario, a *Bnl* ortholog might have existed ancestrally and then been lost in all metazoan lineages except arthropods. Then this gene was captured by baculoviruses after the arthropod radiation [4]. In a second scenario, an arthropod FGF gene was translocated into baculoviruses and, following a period of fast evolution leading to the loss of any phylogenetic signal, reintegrated into the arthropod genome. In the ancestor of nematodes, two FGF genes, orthologs of the FGF9/16/20 and FGF8/17/18/24 families were present. Taking these results into account, we can propose the existence of a minimal FGF gene set of three genes in the ancestor of ecdysozoans (orthologs of FGF1/2, FGF8/17/18/24 and FGF9/16/20). The few data obtained in lophotrochozoans do not allow us to clearly conclude on the FGF gene set of the protostome ancestor. However, we can suggest the presence of at least members of the FGF1/2, FGF8/17/18, and FGF9/16/20 subfamilies.

The two hypotheses proposed here for the evolutionary history of the FGF gene family (Figure 5) suggest that a single paralogous gene for each subfamily was kept in cephalochordates and that specific gene duplications or losses did not occur during evolution in this lineage (Figure 4). In fact, genetic conservation in amphioxus is not restricted to FGFs since different studies have shown that gene content in amphioxus tends to be associated with very few gene losses [24–28]. Concerning other chordates, even if the phylogenetic distribution of the seven urochordate FGF genes is not strongly supported (see Figure S4), we can assume that *C. intestinalis* has orthologs of the FGF4/5/6, FGF7/10/22, FGF8/17/18, FGF9/16/20, FGF11/12/13/14, and FGF19/21/23 subfamilies but that it lost the orthologs of the FGF1/2 and FGF3 subfamilies (Figure 4). Moreover, the seventh gene (Ci-FGFL), as proposed by Popovici et al., could be a specific duplication of FGF7/10/22 [4]. In sarcopterygian

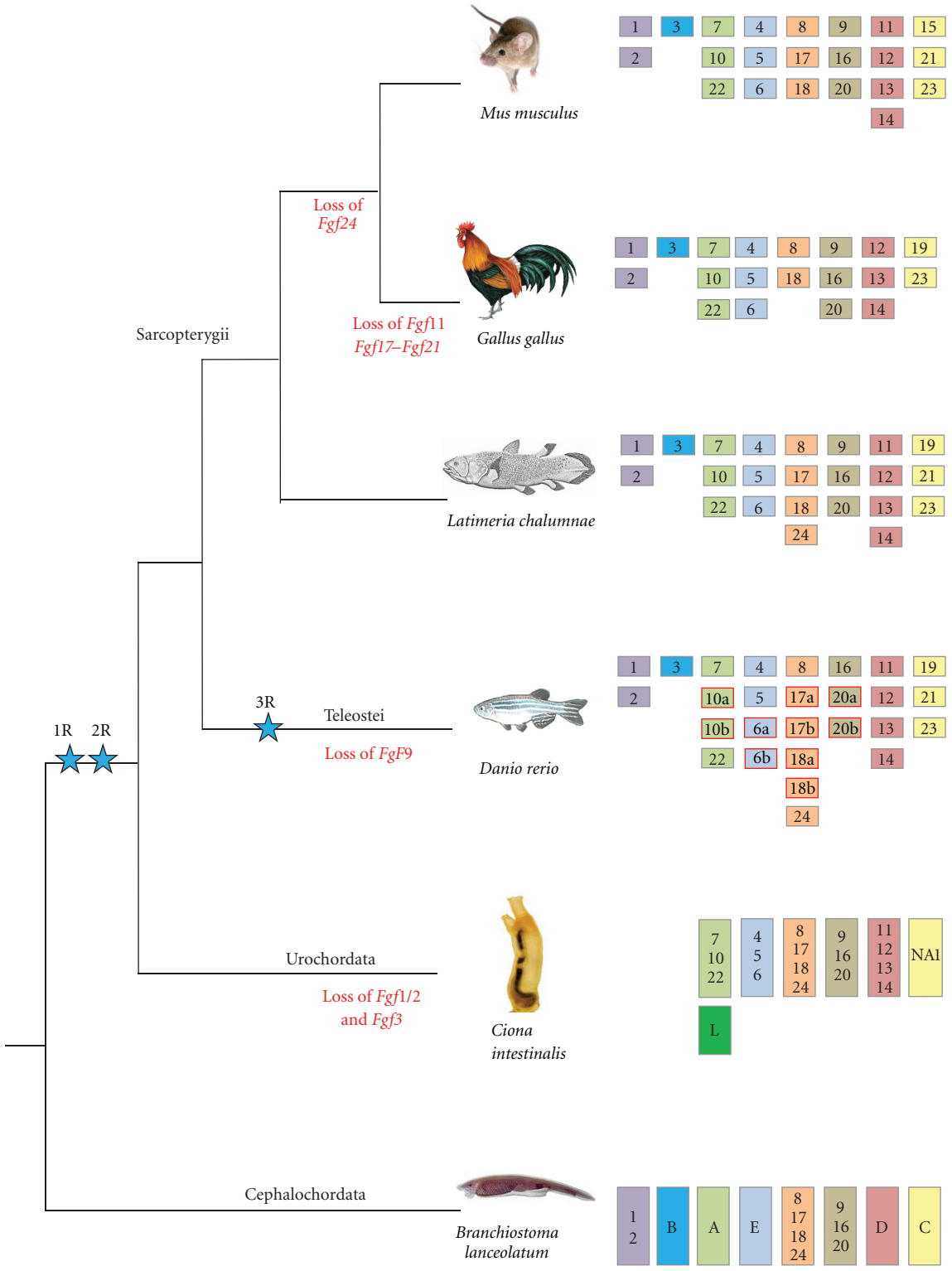


Figure 4: FGF gene content in chordates. Each of the eight FGF paralogy groups is represented by one color. Gene losses are indicated under the tree branches and specific teleost duplications are outlined in red. The urochordate FGFL which is considered as a specific duplication of FGF7/10/22 in this group is colored in dark green. Blue stars represent genome duplications.

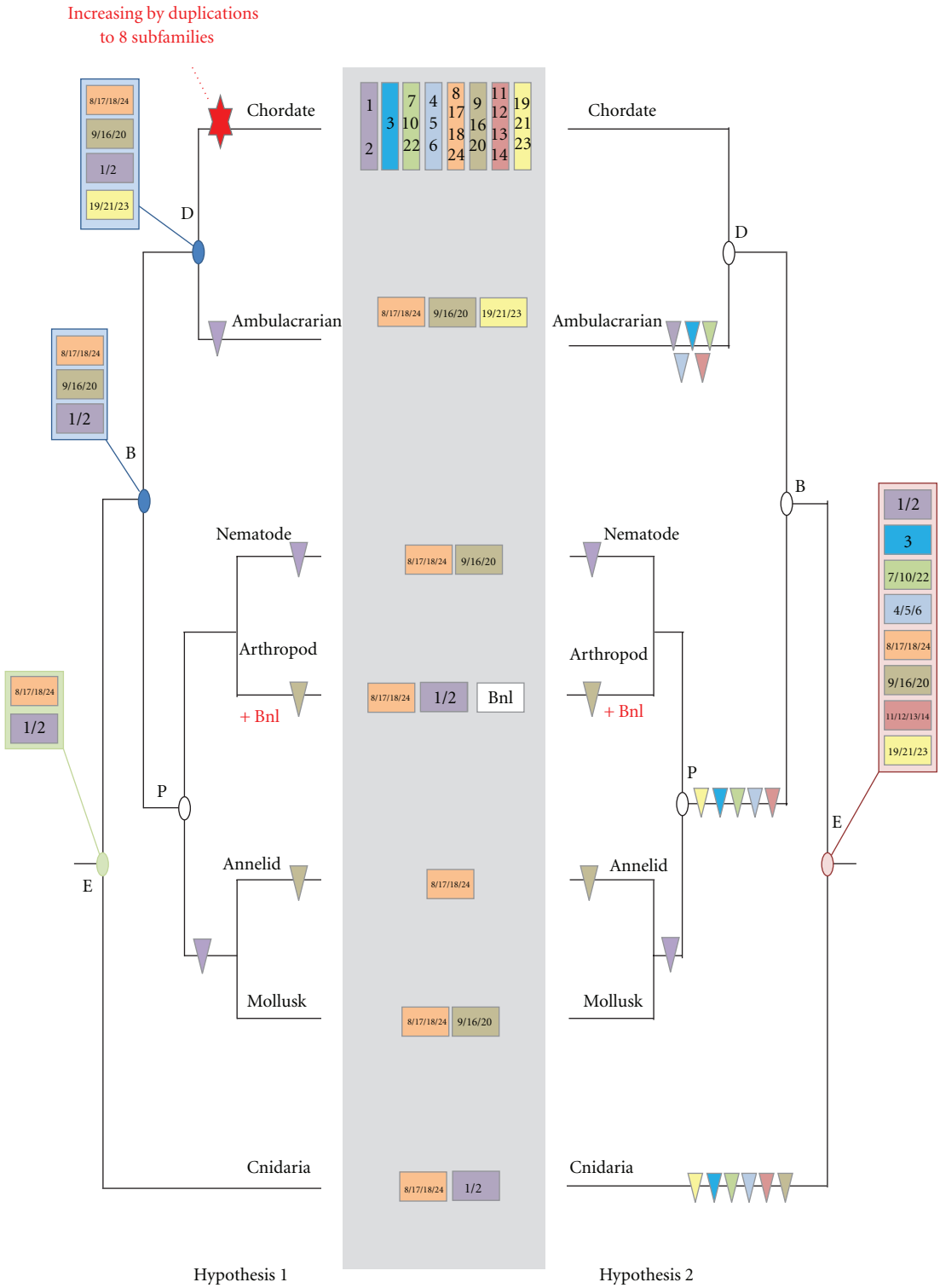


Figure 5: Evolutionary scenarios for FGF evolution in eumetazoans. The minimal gene content of each eumetazoan lineage (chordates, ambulacrarians, nematodes, arthropods, annelids, mollusks, and cnidarians) is mentioned in the center (grey box). Two evolutionary hypotheses are proposed: on the left side (hypothesis 1), starting from a minimum gene set of two genes (green box) in the eumetazoan ancestor, diversity of the subfamily is acquired through chordate-specific duplications; on the right side (hypothesis 2), diversity of the subfamily was acquired very early in metazoan evolution, with 8 subfamilies in the eumetazoan ancestor (red box) and then numerous gene losses in the different lineages occurred. Gene losses are represented by triangles. E: eumetazoan ancestor; P: protostome ancestor; D: deuterostome ancestor and B: Bilaterian ancestor.

vertebrates, the gene set of the different species suggests that numerous gene losses occurred following the two rounds of genome duplication (from eight ancestral genes, after two rounds of duplication, we should find 32 genes, but depending on the species we find between 19 and 23 genes—Figure 4). Moreover, some lineage-specific gene losses also occurred in sarcopterygians; for example, the loss of FGF24 in tetrapods and losses of FGF11, 17, and 21 in chicken. In teleosts, gene losses were even more important, since instead of 46 genes (i.e., a duplication of the 23 FGF genes present in the osteichthyan ancestor [22]) we only find 27 in zebrafish [23]. Indeed, duplicated copies generated by this third genome duplication were only retained for FGF10, FGF6, FGF17, FGF18, and FGF20 (Figure 4).

In non-chordate deuterostomes, the only FGF gene found in the sea urchin cannot be assigned to any FGF subfamily using phylogenetic reconstructions, whereas five of the six genes found in *S. kowalevskii* belong to the FGF8/17/18/24, FGF9/16/20, and FGF19/21/23 subfamilies (Figure S3) [18]. The remaining gene does not show clear phylogenetic relationships with the different FGF subfamilies. Therefore, whatever the evolutionary hypothesis (i.e., chordate-specific duplications versus early duplication giving rise to eight subfamilies in the ancestral eumetazoan), we can propose that there were at least three FGF genes in the ambulacrarian ancestor (i.e., orthologs of FGF8/17/18/24, FGF9/16/20, and FGF19/21/23) (Figure 5). This result suggests that the deuterostome ancestor had probably at least these three genes plus FGF1/2 which is present in chordates and in protostomes but seems to be lost in the Ambulacraria. At this stage of the analysis it is difficult to say if specific chordate duplications led to the eight chordate FGFs (hypothesis 1, Figure 5), or if there was already eight genes in the deuterostome ancestor, several of them having being lost in Ambulacraria (hypothesis 2, Figure 5).

Here, for simplicity, we showed two extreme scenarios, one starting from the minimum gene set in the eumetazoan ancestor (only two genes) and the second starting from the maximum (eight genes). However, many other intermediate scenarios can be imagined. These two major evolutionary scenarios (Figure 5) imply different duplication/loss evolutionary histories. The first hypothesis implies two main points: (i) the ancestral eumetazoan had an FGF gene set of at least two genes (orthologs of FGF1/2 and FGF8/17/18/24) and (ii) important chordate-specific duplications occurred generating the present diversity of the FGF gene family observed in this lineage, which is divided into eight subfamilies (hypothesis 1, Figure 5). The second scenario implies a high degree of gene losses during metazoan evolution. Thus, from eight ancestral FGF gene families already present in the eumetazoan ancestor, six gene losses occurred in cnidarians, five in protostomes and five in ambulacrarians (hypothesis 2, Figure 5). Moreover, both hypotheses require lineage-specific duplications. The second hypothesis is less parsimonious than the first, but no matter which is correct, what seems clear is that the evolutionary history of the FGF gene family required numerous events of gene duplication and gene loss at different times and in different evolutionary lineages. The next question we should address in the near future is which are the implications of this complicated evolutionary history of the FGF gene family on the functional evolution of this signal and in the morphological evolution of metazoans.

5. Materials and Methods

5.1. Identification of FGF Sequences. FGF sequences were identified using BLASTP search in the NCBI and JGI [25] databases using all known FGF domain amino acid sequences. We also browsed the Pfam database [29] for entries possessing an FGF domain. Sequence accession numbers of FGF sequences identified in this study are shown in Table 1.

5.2. Phylogenetic Analyses of Vertebrate FGFs. FGF amino acid sequences were aligned using clustalX [30] and regions of ambiguous homology were removed. Neighbour-Joining tree was generated using *MEGA* version 5 [31] with a Poisson model and a discrete gamma-distribution model with four rate categories. Maximum Likelihood (ML) tree was built using PHYML3.0 [32] with a JTT model as proposed by ProtTest2.4 [33]. The node robustness of both trees was estimated by a bootstrap test (100 replicates).

5.3. Phylogenetic Analyses of Nonvertebrate FGFs. The FGF domain coding region of retrieved sequences was aligned with known FGF sequences from metazoans using T-Coffee [34]. The resulting alignment was manually corrected in SeaView [32]. Maximum Likelihood (ML) trees were generated using PHYML3.0 [32] with a LG+G model as proposed by ProtTest2.4 [33]. The robustness of the tree nodes was estimated using aLRT.

Acknowledgments

The laboratory of H.Escriva is supported by the Agence Nationale de la Recherche Grants ANR-2010-BLAN-1716 01 and ANR-2010-BLAN-1234 02. The authors thank also Peter Mills who kindly revised the English of an earlier version of the paper.

References

[1] O. A. Trowell and E. N. Willmer, "Studies on the Growth of Tissues in vitro," *The Journal of Experimental Biology*, vol. 16, pp. 60–70, 1939.

[2] D. Gospodarowicz, K. L. Jones, and G. Sato, "Purification of a growth factor for ovarian cells from bovine pituitary glands," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 71, no. 6, pp. 2295–2299, 1974.

[3] E. Kolpakova, A. Wiedlocha, H. Stenmark, O. Klingenberg, P. O. Falnes, and S. Olsnes, "Cloning of an intracellular protein that binds selectively to mitogenic acidic fibroblast growth factor," *Biochemical Journal*, vol. 336, part 1, pp. 213–222, 1998.

[4] C. Popovici, R. Roubin, F. Coulier, and D. Birnbaum, "An evolutionary history of the FGF superfamily," *BioEssays*, vol. 27, no. 8, pp. 849–857, 2005.

[5] D. M. Ornitz and N. Itoh, "Fibroblast growth factors," *Genome Biology*, vol. 2, no. 3, article 3005, 2001.

[6] N. Itoh and D. M. Ornitz, "Evolution of the Fgf and Fgfr gene families," *Trends in Genetics*, vol. 20, no. 11, pp. 563–569, 2004.

[7] N. Itoh, "The Fgf families in humans, mice, and zebrafish: their evolutional processes and roles in development, metabolism, and disease," *Biological and Pharmaceutical Bulletin*, vol. 30, no. 10, pp. 1819–1825, 2007.

[8] F. Coulier, P. Pontarotti, R. Roubin, H. Hartung, M. Goldfarb, and D. Birnbaum, "Of worms and men: an evolutionary perspective on the fibroblast growth factor (FGF) and FGF receptor families," *Journal of Molecular Evolution*, vol. 44, no. 1, pp. 43–56, 1997.

[9] H. S. Kim, "The human FGF gene family: chromosome location and phylogenetic analysis," *Cytogenetics and Cell Genetics*, vol. 93, no. 1-2, pp. 131–132, 2001.

[10] U. Technau, S. Rudd, P. Maxwell et al., "Maintenance of ancestral complexity and non-metazoan genes in two basal cnidarians," *Trends in Genetics*, vol. 21, no. 12, pp. 633–639, 2005.

[11] D. Q. Matus, G. H. Thomsen, and M. Q. Martindale, "FGF signaling in gastrulation and neural development in Nematostella vectensis, an anthozoan cnidarian," *Development Genes and Evolution*, vol. 217, no. 2, pp. 137–148, 2007.

[12] D. Sutherland, C. Samakovlis, and M. A. Krasnow, "*branchless* encodes a Drosophila FGF homolog that controls tracheal cell migration and the pattern of branching," *Cell*, vol. 87, no. 6, pp. 1091–1101, 1996.

[13] A. Stathopoulos, B. Tam, M. Ronshaugen, M. Frasch, and M. Levine, "Pyramus and thisbe: FGF genes that pattern the mesoderm of Drosophila embryos," *Genes and Development*, vol. 18, no. 6, pp. 687–699, 2004.

[14] A. Beermann and R. Schröder, "Sites of Fgf signalling and perception during embryogenesis of the beetle *Tribolium castaneum*," *Development Genes and Evolution*, vol. 218, no. 3-4, pp. 153–167, 2008.

[15] R. D. Burdine, E. B. Chen, S. F. Kwok, and M. J. Stern, "egl-17 encodes an invertebrate fibroblast growth factor family member required specifically for sex myoblast migration in *Caenorhabditis* elegans," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 6, pp. 2433–2437, 1997.

[16] S. J. Bourlat, T. Juliusdottir, C. J. Lowe et al., "Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida," *Nature*, vol. 444, no. 7115, pp. 85–88, 2006.

[17] F. Lapraz, E. Röttinger, V. Duboc et al., "RTK and TGF-β signaling pathways genes in the sea urchin genome," *Developmental Biology*, vol. 300, no. 1, pp. 132–152, 2006.

[18] A. M. Pani, E. E. Mullarkey, J. Aronowicz, S. Assimacopoulos, E. A. Grove, and C. J. Lowe, "Ancient deuterostome origins of vertebrate brain signalling centres," *Nature*, vol. 483, no. 7389, pp. 289–294, 2012.

[19] S. Bertrand, A. Camasses, I. Somorjai et al., "Amphioxus FGF signaling predicts the acquisition of vertebrate morphological traits," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 22, pp. 9160–9165, 2011.

[20] Y. Satou, K. S. Imai, and N. Satoh, "Fgf genes in the basal chordate Ciona intestinalis," *Development Genes and Evolution*, vol. 212, no. 9, pp. 432–438, 2002.

[21] P. Dehal and J. L. Boore, "Two rounds of whole genome duplication in the ancestral vertebrate," *PLoS Biology*, vol. 3, no. 10, p. e314, 2005.

[22] O. Jatllon, J. M. Aury, F. Brunet et al., "Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype," *Nature*, vol. 431, no. 7011, pp. 946–957, 2004.

[23] N. Itoh and M. Konishi, "The zebrafish FgF family," *Zebrafish*, vol. 4, no. 3, pp. 179–186, 2007.

[24] N. Takatori, T. Butts, S. Candiani et al., "Comprehensive survey and classification of homeobox genes in the genome of amphioxus, *Branchiostoma* floridae," *Development Genes and Evolution*, vol. 218, no. 11-12, pp. 579–590, 2008.

[25] N. H. Putnam, T. Butts, D. E. K. Ferrier et al., "The amphioxus genome and the evolution of the chordate karyotype," *Nature*, vol. 453, no. 7198, pp. 1064–1071, 2008.

[26] S. Huang, S. Yuan, L. Guo et al., "Genomic analysis of the immune gene repertoire of amphioxus reveals extraordinary innate complexity and diversity," *Genome Research*, vol. 18, no. 7, pp. 1112–1126, 2008.

[27] L. Z. Holland, R. Albalat, K. Azumi et al., "The amphioxus genome illuminates vertebrate origins and cephalochordate biology," *Genome Research*, vol. 18, no. 7, pp. 1100–1111, 2008.

[28] S. D'Aniello, M. Irimia, I. Maeso et al., "Gene expansion and retention leads to a diverse tyrosine kinase superfamily in amphioxus," *Molecular Biology and Evolution*, vol. 25, no. 9, pp. 1841–1854, 2008.

[29] M. Punta, P. C. Coggill, R. Y. Eberhardt et al., "The Pfam protein families database," *Nucleic Acids Research*, vol. 40, pp. D290–D301, 2012.

[30] J. D. Thompson, T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins, "The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools," *Nucleic Acids Research*, vol. 25, no. 24, pp. 4876–4882, 1997.

[31] K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar, "MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods," *Molecular Biology and Evolution*, vol. 28, no. 10, pp. 2731–2739, 2011.

[32] M. Gouy, S. Guindon, and O. Gascuel, "Sea view version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building," *Molecular Biology and Evolution*, vol. 27, no. 2, pp. 221–224, 2010.

[33] F. Abascal, R. Zardoya, and D. Posada, "ProtTest: selection of best-fit models of protein evolution," *Bioinformatics*, vol. 21, no. 9, pp. 2104–2105, 2005.

[34] C. Notredame, D. G. Higgins, and J. Heringa, "T-coffee: a novel method for fast and accurate multiple sequence alignment," *Journal of Molecular Biology*, vol. 302, no. 1, pp. 205–217, 2000.

Hindawi Publishing Corporation
International Journal of Evolutionary Biology
Volume 2012, Article ID 846421, 10 pages
doi:10.1155/2012/846421

Review Article

Mechanisms of Gene Duplication and Translocation and Progress towards Understanding Their Relative Contributions to Animal Genome Evolution

Olivia Mendivil Ramos and David E. K. Ferrier

The Scottish Oceans Institute, School of Biology, University of St Andrews, East Sands, Fife KY16 8LB, UK

Correspondence should be addressed to David E. K. Ferrier, dekf@st-andrews.ac.uk

Received 26 March 2012; Revised 30 May 2012; Accepted 27 June 2012

Academic Editor: Ben-Yang Liao

Duplication of genetic material is clearly a major route to genetic change, with consequences for both evolution and disease. A variety of forms and mechanisms of duplication are recognised, operating across the scales of a few base pairs upto entire genomes. With the ever-increasing amounts of gene and genome sequence data that are becoming available, our understanding of the extent of duplication is greatly improving, both in terms of the scales of duplication events as well as their rates of occurrence. An accurate understanding of these processes is vital if we are to properly understand important events in evolution as well as mechanisms operating at the level of genome organisation. Here we will focus on duplication in animal genomes and how the duplicated sequences are distributed, with the aim of maintaining a focus on principles of evolution and organisation that are most directly applicable to the shaping of our own genome.

1. Introduction

New genes constitute some of the major raw material for the evolution of biodiversity. They do not arise out of thin air. Some instances of new gene evolution from previously non-coding sequence have now been discovered [1, 2]. Also, new genes can be formed by shuffling of pre-existing nucleotide sequences. The relatively recent discovery of large numbers of taxonomically restricted genes also demands a closer investigation of their mode(s) of origin [3]. Nevertheless, a major mechanism for the generation of new genes is via duplication. Such duplicates are called paralogues, to reflect their homologous relationship being due to a duplication event rather than a speciation event (see Figure 1).

Since the first animal whole genome sequence of the nematode *Caenorhabditis elegans* [8], the number of animal whole genome sequences has been increasing at an impressive rate. It should, however, be kept in mind that there is a high level of variability in the "quality" of these genome sequences; "quality" here referring to the depth of sequence coverage of the genome, levels of effort to fill gaps in the sequence, and amount of independent mapping data to inform and confirm the assembly. As a result, many of the animal whole genome sequences that are available must be handled with caution when estimating the extent and nature of duplication events. Furthermore, most animal genome sequences can only be assembled to a subchromosomal scale, with genomic scaffolds covering only fragments of chromosomes. This becomes important when trying to assess duplication and translocation mechanisms and distinguishing intra- and interchromosomal events. Inevitably, the organisms with the largest research communities and the most intensively studied genomes tend to have the highest quality genome assemblies and annotations. Most studies of gene and genome duplications, and hypotheses about mechanisms, stem from analyses of such organisms as vertebrates (including humans, other mammals, and fish) and insect and nematode model systems, as will become clear below.

Here we review the current terminology used for duplicated genes and then discuss the role of whole genome duplication, particularly within the context of vertebrate

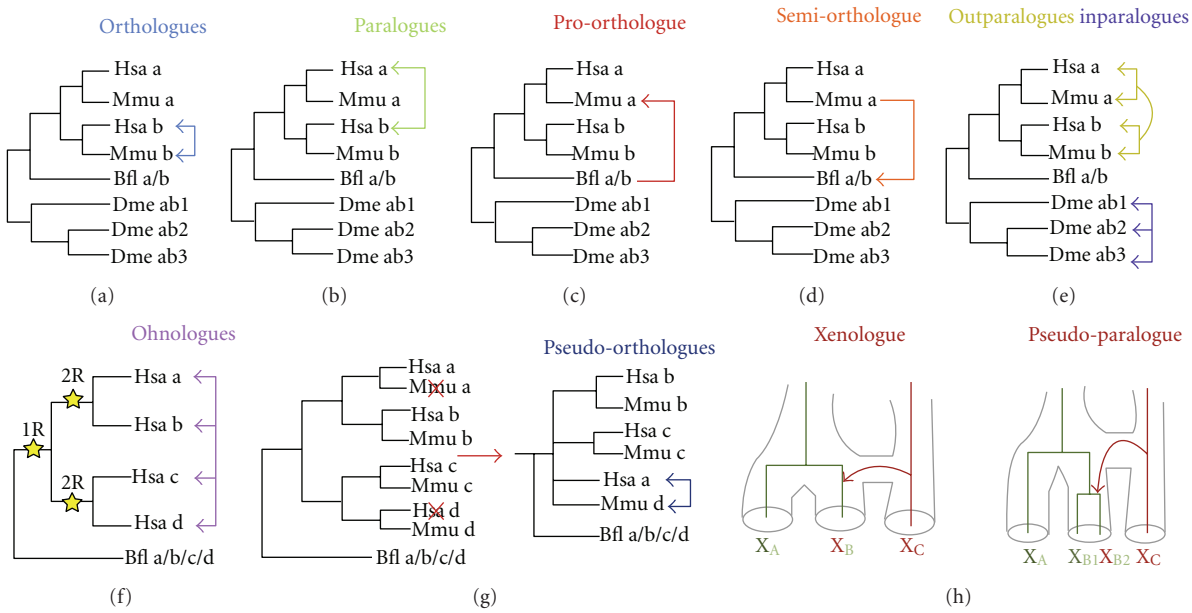


Figure 1: Overview of the current terminology. The different panels represent term(s) for duplicated genes. (a) Orthologues. The square blue arrows represent an orthologous relationship between the two genes. (b) Paralogues. The square green arrows represent paralogous relationships between the genes. (c) Proto-orthologue. The square red arrow represents the pro-orthologue relationship of gene a/b from *Branchiostoma floridae* to gene a from *Mus musculus*. (d) Semi-orthologue. The square orange arrow represents the semi-orthologous relationship of gene a of *Mus musculus* to gene a/b from *Branchiostoma floridae*. (e) Inparalogues and Outparalogues. The square yellow arrows represent the outparalogous relationship in which human and mouse a genes are outparalogous to human and mouse b genes. As a set, genes a and b from mouse and human represents coorthologues. The square purple arrows represent the inparalogous relationship between the genes which duplicated within this lineage. (f) Ohnologues. The square pink arrows delimit all the paralogues coming from WGD and the stars represent the duplication events. (g) Pseudo-orthologues. The square navy arrows represent the pseudo-orthologues. The red Xs represent lineage-specific gene losses. (h) Xenologues and Pseudo-paralogues. Species are represented by subindices A, B, and C, and the Xs represent the orthologous genes with their colouring designating the species of origin. All of the figures are adapted from [4–6]. Bfl: *Branchiostoma floridae*, Dme: *Drosophila melanogaster,* Hsa: *Homo sapiens*, and Mmu: *Mus musculus*.

evolution, and review the current understanding of modes of subchromosomal duplications and recent data on mechanisms for distribution of these duplicated sequences around the genome.

2. Terminology: Beware Overlap, Synonyms, and Ambiguity (and Use with Care)

The terminology used to define the evolutionary relationships between duplicated genes has become increasingly detailed. The precise inference of the evolutionary relationships between duplicated genes is fundamental for most comparative genomic studies, but it can be complicated because duplication is often combined with speciation and subsequent gene loss [4].

The most widely used terms for describing evolutionary relationships between genes are homologous, orthologous, and paralogous. Fitch [9] defined homologous genes as those that share a common ancestor. A subset of homologous genes are orthologous, these being the genes separated only by speciation and not by a duplication event (Figure 1(a)). Another subset of homologous genes are paralogous, which are those resulting from a duplication event (Figure 1(b)). Sharman [4] defined additional terms to describe the relationships amongst paralogues. Pro-orthology denotes the relationship of a gene to one of the descendants of its orthologue after duplication of that orthologue (Figure 1(c)). Conversely, semi-orthology is the relationship of one of a set of duplicated genes to a gene that is orthologous to the ancestor of the whole set (Figure 1(d)). Sharman [4] also proposed the term trans-homology to describe members of the same gene family descendant from an ancestral gene via two independent gene duplication events. A further important term connected with paralogy is the one proposed by Wolfe [10], who coined the term ohnologue for those paralogues stemming from a whole genome duplication (Figure 1(f)). Two years later, Sonnhammer and Koonin [5] highlighted that the definition of a paralogous relationship can be related to a speciation event. Thus, they coined the terms inparalogues and outparalogues. Inparalogues are paralogues in a given lineage that all evolved by gene duplications that happened after a speciation event that separated the given lineage from the other lineage under consideration (Figure 1(e)). Outparalogues are paralogues in a given lineage that evolved by gene duplications that happened before a speciation event (Figure 1(e)). Careful consideration must be taken when using the terms such as inparalogues, outparalogues, and ohnologues. The

specification of the relation of the duplication event to the speciation event must be included when these terms are used, otherwise evolutionary interpretations and use of terminology can easily be confused. Finally, a new umbrella term, duplogs [11], has been thrown into the duplication terminology pool to define intraspecies paralogues. This term amalgamates all the types of paralogues within a species, including inparalogues, outparalogues, and ohnologues.

Sonnhammer and Koonin [5] also defined co-orthologues, which are synonymous with Sharman's [4] definition of trans-homologues, and are inparalogues of one lineage which are homologous to another set of inparalogues in a second lineage. Artifacts stemming from phylogenetic inference, such as lineage-specific gene loss, can mislead the deduction of the evolutionary relationship of genes. For this purpose, Koonin [6] devised the term pseudo-orthologue to accommodate those genes that are essentially paralogues but appear to be orthologues due to differential, lineage-specific gene loss (Figure 1(g)). Further useful terms are xenologue and pseudo-paralogue. Xenologues are homologues acquired through horizontal gene transfer by one or both species that are being compared, but appearing to be orthologues when pairwise comparison of the genomes is performed (Figure 1(h)) [6]. Pseudo-paralogues are homologues that through the analysis in a single genome are interpreted as paralogues; however, these homologues originated by a combination of vertical inheritance and horizontal gene transfer (Figure 1(h)) [6].

Recently a new term, toporthology, has been specified, which aims to include another aspect of the concept of orthology, that of positional orthology [12]. Toporthology describes the evolutionary relationship of orthologues that retain their ancestral genomic positions. In the context of gene duplications, a duplication event is said to be "symmetric" if deletion of either of the copies of the duplicated sequences would return the gene order to the original, ancestral state. Thus, tandem duplicates and whole-chromosome/genome duplication are symmetrical duplications. A duplication event is "asymmetric" if deleting only one of the copies could return the gene order to its original, ancestral state. Consequently, dispersed segmental duplications and retrotranspositions are asymmetrical duplications. From these definitions two genes are positionally homologous, topohomologous, if they are homologous and neither gene comes from an asymmetric duplication since the time of their common ancestor. The contrast to this case is atopohomologous. The topo- and atopo- prefixes can similarly be applied to orthologues and paralogues.

The term toporthology and its associated derivations need to be used with extreme caution [12]. The value, and aim, of distinguishing toporthologues/topoparalogues is to distinguish those genes (which are not necessarily one-to-one orthologues) that are most comparable in terms of their evolutionary history. However, being able to distinguish toporthology obviously requires reliable, accurate genome assemblies and hinges on distinguishing parent/source locations from daughter/target locations of duplicated regions. Also, the distinction of toporthology can obviously be complicated by genomic rearrangements that occur after the duplication event and which can obscure whether a duplication was symmetric or asymmetric. Currently, the complications introduced by such postduplication genomic rearrangements lead to some counterintuitive uses of the terminology. One might assume that toporthology/topoparalogy simply refers to orthologues/paralogues that are both in the ancestral locations, and conversely that atoporthology/atopoparalogy simply describes the situation in which at least one of the genes is no longer in the ancestral location. The use of the terminology is not so straight-forward, however, as can be seen by a close inspection of Figure 2 in [12], in which YA1 and YA2 are topoparalogues rather than atopoparalogues despite YA2 no longer being in the ancestral location. The classification of YA1 and YA2 as topoparalogues arises because they were not produced by an asymmetric duplication, but then the subsequent change of position of YA2 has obscured this. Consequently the precision of the data (taxonomic sampling and quality of genome assembly) severely compromises the utility of this terminology. Despite the apparent use of the terms to reflect relationships relative to ancestral locations within the genome, in fact the movement of genes to new, nonancestral locations subsequent to the duplication event is not accommodated. Consequently toporthologues/topoparalogues are not necessarily both in the ancestral genomic position. This terminology thus risks being counterintuitive and confusing in its present form.

The above summary of duplicate terminology serves to illustrate two things. Firstly, there is the complexity of the evolutionary processes involved in production of duplicates and the care that must thus be exercised when comparing genes between species. Secondly, there is currently an overabundance of terminology, some of which is redundant and some of which is counterintuitive. It is to be hoped that with time the terminology will settle on a consensus of selected terms and those that are impractical or potentially misleading will be abandoned. We now turn from the terminology of gene duplication to the biological processes and evolutionary events.

3. Whole Genome Duplications (WGDs): Origin of Vertebrates and 2R

One of the most striking features of the human genome, which is shared with the other members of our subphylum, the Vertebrata, is the extensive occurrence of paralogons: homologous regions of chromosomes that are related via duplication events rather than speciation events [13]. This observation is usually attributed to the occurrence of two rounds of whole genome duplication at the origin of the vertebrates (the so-called 2R hypothesis), because of the preponderance of four paralogons for each region of the human genome being considered. Thus, one copy of the diploid genome duplicated to give two copies, and this tetraploid state then duplicated a second time to effectively give an octoploid state [14], which with time has been "diploidized" again but with the remnants of the octoploid state being detectable from analyses of the paralogons. The 2R events were inevitably followed by extensive gene loss,

as would be expected given the inevitable high levels of genetic redundancy that would ensue from such large-scale duplications, such that less than 30% of the 2R paralogous genes are estimated to remain [15]. This means that 2R paralogue families now consist of between two to four members [16], thus providing a significant pool of extra genes that have made a significant contribution to the evolution and diversification of the vertebrates.

This 2R hypothesis has its roots in the ideas of Susumu Ohno, and it then began to gain increasing support from molecular genetic work, principally from the invertebrate chordate amphioxus. For example, amphioxus has a single Hox gene cluster whilst humans have four [17, 18]. The 2R hypothesis was not universally accepted at first [19], largely on the grounds of differing interpretations of molecular phylogenetic trees and the assessment of branching topologies within different gene families and amongst paralogues. The topology argument that formed the basis for challenging the 2R hypothesis [19] requires the trees to be interpreted in a very restricted fashion, with the four paralogues adopting a symmetrical topology of ((A, B)(C, D)). This was supposed to represent the first WGD producing two paralogues, which were the precursors to AB and CD, followed by the second WGD producing the A and B as well as C and D genes. However, it is far from clear that duplicated genes always behave in the expected post-duplication way, with daughters evolving at equal rates post-duplication. In fact there is increasing evidence for asymmetric evolution of duplicated genes [20], often with disruptions to tree topology that tend to arise from Long Branch Attraction [21]. Also, as analyses progressed to genome-scale data the controversy has largely subsided with the ever-increasing evidence in favour of 2R. This is typified by the sequencing of the whole genome of the American amphioxus, *Branchiostoma floridae*, and analyses not just of paralogue phylogenies but also patterns of gene synteny across chordates. The trend for a single locus in amphioxus matching four loci in humans (and other vertebrates, with some notable exceptions mentioned below), which was originally developed from work on the Hox gene cluster(s) [22] was found to extend to large-scale, genome-wide Quadruple Conserved Synteny [23].

There have still been one or two dissenting voices, such as [24] arguing instead for segmental duplications occurring at different times rather than whole genome duplications (and hence simultaneous origins of paralogons). However, we note that the interpretation of the molecular phylogenies in [24] contains a number of errors, including deductions based on support values at inappropriate nodes as well as nodes that do not have significant support values. Questionable rooting strategies are employed in several of the trees in [24] and incomplete datasets are used for some genes, such as the Sp transcription factors [25]. The analyses of Abbasi [24] in fact do not challenge the 2R hypothesis, but in fact often support it as soon as one accepts that some gene loss occurred after 2R. That gene loss is a common phenomenon is now without doubt [15, 26–30]. Also, since both WGD events occurred close together in time, and via autotetraploidy in both cases, then it is to be expected that the phylogenies of the paralogues do not in fact adopt the ((A, B)(C, D)) topology, as explained by Furlong and Holland [14]. Tree topologies should thus not still be being used as a test of 2R with the view that divergence from the ((A, B)(C, D)) topology is in conflict with 2R. Furthermore, the 2R hypothesis no longer relies solely upon the topology of individual gene trees, but instead gains its most convincing support from conserved synteny arrangements that cover over 90% of the human genome and extends to the genomes of birds and fish (including chicken, stickleback, and puffer fish) [23]. Therefore, we hold the view that the 2R hypothesis (with subsequent gene loss) is definitely the most parsimonious explanation for the origin and evolution of vertebrate genomes.

The plausibility of the 2R hypothesis is further strengthened by the discoveries of whole genome duplications elsewhere in the animal kingdom, thus demonstrating that the process can certainly occur, and do so with reasonable frequency (see Table 1) [31, 32]. For example, the origin of the teleost fish coincides with another WGD, the 3R event. Again, this hypothesis is strongly supported by the patterns of synteny relative to other vertebrates and the existence of extensive paralogons matching the topology expected for a 3R event [33]. Whole genome duplications and polyploidization events are constantly coming to light within the animal kingdom, and are clearly a significant mode of duplication that has shaped animal evolution. Duplications also occur on a smaller scale, at the subchromosomal level.

4. Subchromosomal Duplications: Variable Sizes, Rates, and Mechanisms

Duplications that encompass sections of DNA smaller than whole chromosomes are given the generic name of segmental duplications (SDs). These can vary enormously in size, from a few base pairs up to many megabases, and may or may not contain intact, functional genes. They can also be found in several different arrangements, which are important for considerations as to how these SDs might form. SDs can be adjacent (tandem duplications), separated, or interspersed along a particular chromosome (intrachromosomal) or on distinct chromosomes (interchromosomal). The detection of SDs in these different categories obviously depends upon the quality of a genome sequence assembly, but the prevalence of SDs in the human genome, for example, tend to be estimated at about 5-6% (for SDs $\geq$1 kb, with $\geq$90% sequence identity, and filtered for transposable elements and other high-copy repeats) [62]. Estimates of SD prevalence in other mammals tends to produce slightly lower levels than in humans, although in the case of mouse that has recently been revised upwards to almost 5% and hence is now thought to be comparable to the levels in humans [62, 63]. A striking aspect of the comparisons between rates and distributions of SDs in various mammalian genome sequences is that tandem duplications are by far the most prevalent category of SD, comprising 75–90% of SDs in the cow for example [64]. This preponderance of tandem duplicates in mammals as diverse as cows, rodents, and dogs does not, however, reflect the situation in humans, in which SDs are much more frequently interspersed [64–67]. The interspersed distribution

Table 1: Examples of species undergoing whole genome duplication or polyploidisation events. Adapted from [31, 32].

Species/Taxon (Common name)	References
Xenopus laevis (African clawed frog)	Morin et al. [34]
Tympanoctomys barrerae (red viscacha rat)	Gallardo et al. [35]
Daphnia pulex (water flea)	Vergilino et al. [36]
Schmidtea polychroa (planarian flatworm)	D'Souza et al. [37]
Acipenser brevirostrum (shortnose sturgeon)	Fontana et al. [38]
Scaphirhynchus platorynchus (shovelnose sturgeon)	Schultz [39]
Polyodon spathula (american paddlefish)	Schultz [39]
Menidia sp. (atlantic silverside)	Echelle and Mosier [40]
Barbatula barbatula (stone loach)	Collares-Pereira et al. [41]
Catostomidae (suckers)	Schultz [39]
Botia spp. (pakistani loach)	Yu et al. [42], Rishi and Shashikala Rishi [43]
Cobitis spp. (loach)	Schultz [39], Vrijenhoek et al. [44], Janko et al. [45]
Misgurnus anguillicaudatus (dojo loach)	Arai et al. [46]
Misgurnus fossilis (european weather loach)	Raicu and Taisescu [47]
Barbodes spp. (tinfoil)	Chenuil et al. [48]
Barbus spp. (barb)	Suzuki and Taki [49]
Acrossocheilus sumatranus (large-scale barb)	Suzuki and Taki [49]
Aulopyge hugelii (dalmatian barbelgudgeon)	Mazik et al. [50]
Cyprinus carpio (carp)	Wang et al. [51]
Carassius auratus (goldfish)	Schultz [39], Yu et al. [42], Shimizu et al. [52]
Schizothorax spp. (snowtrouts)	Mazik et al. [50]
Synocyclocheilus spp. (barbels)	Yu et al. [42], Rishi and Shashikala Rishi [43]
Tor spp. (mahseer)	J. Gui et al. [53]
Zacco platypus (freshwater minnow)	Yu et al. [42], Mazik et al. [50]
Poecilia spp. (guppy)	Schultz [39], Vriejenhoek et al. [44]
Poeciliopsis spp. (desert minnows)	Schultz [39]
Protopterus dolloi (slender lungfish)	Vervoort [54]
Lepisosteus oculatus (spotted gar)	Schultz [39]
Stizostedion vitreum (walleye)	Ewing et al. [55]
Salmonidae (salmons)	Allendorf and Thorgaard [56]
Clarias batrachus (walking catfish)	Pandey and Lakra [57]
Heteropneustes fossilis (indian catfish)	Pandian and Koteeswaran [58]
Hyla versicolor (grey treefrog)	Ptacek et al. [59], Mable and Bogart [60]
Neobatrachus spp. (burrowing frogs)	Mable and Roberts [61]

of human SDs is possibly the result of an expansion of Alu transposable elements within primates [62, 68]. Moving outside of the mammals, the fruit fly *Drosophila melanogaster* has the majority of its SDs in the intrachromosomal category (86%), and of these most are situated close together in the genome (50% and <14 kb apart) [69].

The different categories of SDs (tandem, interspersed intrachromosomal, and interchromosomal) may well reflect different mechanisms of DNA-based duplication. Non-homologous end-joining (NHEJ) is more likely to account for adjacent duplications [70–72] with the repair of DNA breaks being more likely to occur between ends in close proximity. The alternative of nonallelic homologous recombination (NAHR) is likely mediated via repetitive sequences dispersed around the genome and hence is a route to interspersed duplications. This process has been given the name duplication-dependent strand annealing (DDSA) by Fiston-Lavier et al. [69], who also noted that in *D. melanogaster* the mean size of intrachromosomal events is larger than the average size of interchromosomal events (3.1 kb versus 2.1 kb, respectively). This contrasts with the average size of SDs in humans being approximately 18.6 kb and 14.8 kb for the intrachromosomal and interchromosomal categories respectively [73].

In addition to this observation that intrachromosomal SDs tend to be longer than interchromosomal SDs possibly reflecting different mechanisms being the cause of their origin, it is striking that the size of SDs varies in different species. A further "data point" is provided by the nematode *Caenorhabditis elegans*, in which the average size of SDs is only 1.4 kb [74]. This implies that the size of duplication is not necessarily determined by physical properties of the DNA or possibly the duplication mechanism (unless mechanisms

differ between the taxa thus far examined), but instead is likely to relate to the structure and organization of the genome. Density and distribution of repetitive sequences will be one factor, and these vary across different species. In addition, strong selective pressures are likely to come into operation when genes are duplicated within SDs, often disrupting genetic networks and pathways if a gene is duplicated and then expressed (e.g., via dosage imbalance [75]). Thus there will tend to be selective pressure against duplications that encompass genes (and their regulatory elements), thus reducing the average size of segmental duplicates in taxa with smaller, more compact genes.

Alongside consideration of the duplication mechanisms within the context of determining the organisation of duplicated genes, it follows that one must also consider processes by which segments of DNA or genes can be translocated around the genome. Although these mechanisms are not necessarily leading to generation of duplications (and in fact often are not) they are still crucial in understanding the subsequent distribution of genes, which in the present context happen to be duplicates. Retrotransposition is one of the duplication mechanisms that does not necessarily lead to generation of functional duplicated genes, but is crucial in distributing duplicated single genes, especially in an inter-chromosomal fashion [76–79]. Inversions are very common and help to scatter duplicated genes along a particular chromosome arm [71, 80]. Also large-scale events such as inversions between arms involving the centromere or chromosome fusions and fissions are also known to play a prominent role in karyotype evolution, and reciprocal translocations between chromosome arms are very common. Surprisingly high rates of reciprocal translocations occur in humans, with estimates of around one in 500 newborns carrying such large-scale rearrangements [81–84]. This is not necessarily unusual to humans, as cattle reciprocal translocations have been estimated to occur at a rate of 1.4 per 1000 animals [85]. These high rates of translocations are thought to be mediated via NAHR using duplicated or repetitive segments located in different chromosomes, that is interchromosomal low-copy repeats (LCRs) [86]. Ou et al. [86] characterized several hundred interchromosomal LCRs in the human genome, ranging in size from 5kb to over 50kb, all of which they suggest can act as the substrates for reciprocal translocations. In addition, Hermetz et al. [87] described a translocation occurring via homologous recombination between HERV elements on different chromosomes.

In combination all of these routes to rearrangement of genome organisation often make it difficult to accurately determine between likely mechanisms of duplicate origin. This is because it is difficult to determine whether the locations of any two duplicated sequences reflect their organisation at their point of origin, or instead is the end point of originating by a process such as tandem duplication and then subsequently being dispersed. Attempts to address this problem have involved estimating the age of duplicates by calculating the rates of synonymous substitutions (K_s). This has led to observations that younger genes tend to be closer together in the genome, particularly being more highly represented in the intrachromosomal category of duplicates relative to the interchromosomal category [74, 88]. However, such estimates of gene age can be confounded by the process of gene conversion, which can homogenise gene sequence after the origin of the duplicates [89, 90]. Since gene conversion is more likely to occur between genes that are in close proximity then there will be a degree of misjudging the age of duplicates as inappropriately young, and this effect will be most pronounced in the categories of closely linked genes such as tandem duplicates. Furthermore, the positive correlation between age and dispersal in the genome has recently been questioned with the proposal of a process named drift duplication [11]. Ezawa and colleagues' [11] comparisons of duplicate age and genomic location in human, mouse, zebrafish, *C. elegans*, *D. melanogaster*, and *Drosophila pseudoobscura* suggest that interspersed intrachromosomal duplications can be generated at once, rather than originating as tandem duplicates which are subsequently relocated away from each other, and this can happen at comparable rates to tandem duplication [11].

The precise mechanism leading to drift duplication is not specified by Ezawa et al. [11], and is likely to involve a combination of processes. One of these could well be the recently discovered process of duplication via circular DNA-based translocation. Durkin et al. [7] recently found that in "lineback" or "witrik" cows a translocation of 492 kb occurred which was then followed by a repatriation of a 575 kb segment, including the KIT gene that is involved in the pigmentation patterning of the cows and their distinctive "lineback" phenotype. The intriguing aspect to these translocations is the order of sequences within the translocated segment, which is consistent with translocation via a circular DNA intermediate which is opened up for reinsertion at a different point in the circle from the boundaries of the original excision (Figure 2). Also, since the repatriated segment was larger than the originally translocated segment then some sequence duplication results (Figure 2). Further examples of duplications via circular DNA intermediates are being found, such as the vasa genes of *Tilapia* [91]. The difference between the cow and *Tilapia* examples however is that the cow circular DNA intermediate is repatriated into an ancestral locus, presumably due to homologous recombination, whereas the *Tilapia* vasa duplicates that arose via circular intermediates have gone to new locations. The *Tilapia* vasa example is thus more reminiscent of drift duplication, but it remains to be seen how prevalent such circular DNA translocation events are and how the reintegration sites are selected.

Given the range of genomic rearrangement mechanisms and their apparent frequencies, it is perhaps surprising that syntenic arrangements can be conserved for vast evolutionary timespans, for example, from humans to the origin of chordates [23] and beyond, to even some basal lineages of animals such as the cnidarian *Nematostella vectensis* and the placozoan *Trichoplax adhaerens* [92, 93]. What is also striking is that this phenomenon of long-term general synteny conservation is not detected uniformly across the animal kingdom. Some lineages and groups of animals seem to have particularly derived genome organisations relative to other animals (e.g., *Oikopleura* and urochordates in general;

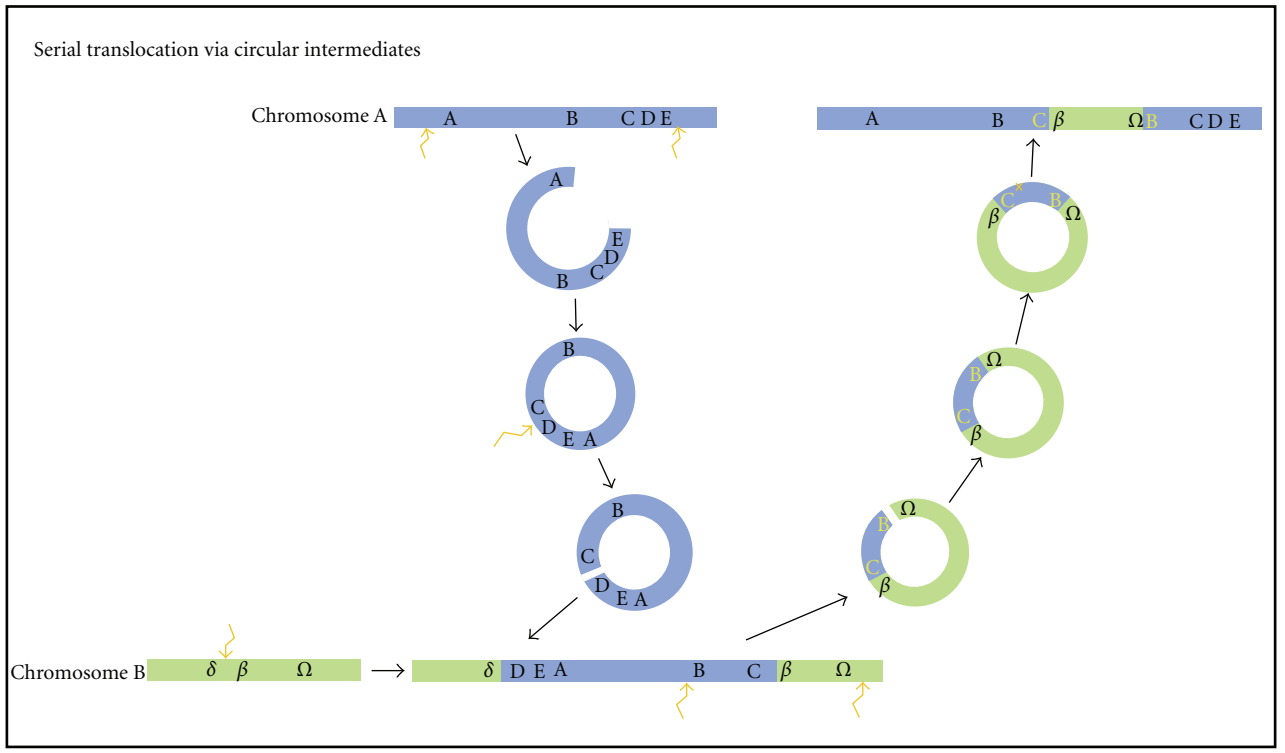


Figure 2: Scheme of a serial translocation via circular DNA intermediates. Two excisions create a fragment of chromosome A, delimited by genes A and E. This fragment circularizes. At reinsertion into a new genomic location, the circle is linearized by being opened between C and D and inserts between genes ∂ and β of chromosome B. The subsequent translocation involves an excision delimited by genes B and Ω. The fragment created circularizes and has sequence identity to the region on chromosome A between the C and B genes. This region of homology allows a repatriation of the segment of original genes from chromosome A, creating a duplication as well as translocating genes from chromosome B. Blue and green lines represent fragments of two different chromosomes. The capital and Greek letters represent genes within the chromosomes. The yellow capital letters denote the genes translocated from chromosome B (green line). The angled orange arrows represent excision points in the DNA. The orange cross represents a homologous recombination site. Adapted from [7].

Drosophila and other Diptera; nematodes like *C.elegans* [8, 94, 95]). One could speculate that this might reflect different abundances of repetitive elements, for example, which can have a role in facilitating genomic rearrangements. Another possibility is that gene sizes, and perhaps more importantly gene densities within the chromosomes, vary significantly across the animal kingdom. This variation might not just be the number of nucleotides spanned by the coding sequence, but also by the regulatory elements, which will influence how frequently rearrangement mutations can occur that are still compatible with organismal viability. Regardless of this, some animal genomes seem to be more tolerant of, or prone to, rearrangements than others. With the burgeoning amounts of human genome sequence data, particularly in relation to disease and cancer genomics, a new phenomenon involving a catastrophic rearrangement of the genome has recently been described: chromothripsis [96, 97]. Perhaps the process of chromothripsis has a relevance beyond the realms of cancer and disease biology and may be comparable to processes whereby some animal genomes become extensively rearranged relative to other lineages.

5. Conclusion

Gene and genome duplication constitute major forces in evolutionary innovation. The variety of mechanisms by which such duplications occur, as well as the various means by which the duplicated segments are subsequently rearranged (and sometimes partially lost), requires careful analysis and consistent use of biologically informed terminology. Obviously a major goal for the future will be to expand the taxonomic coverage of high-quality genome assemblies to enable the deduction of more accurate and more widely applicable, general conclusions about such phenomena as gene and genome duplications. This should be complemented by the continued development of in silico tools and models to estimate duplication and rearrangement rates. Such tools then need to be applied across an increased range of genomes in order to distinguish general mechanisms and principles from lineage-specific oddities, such as lack of synteny between urochordates and vertebrates or the paucity of tandem duplications in humans relative to other mammals.

References

[1] W. Wang, H. Zheng, S. Yang et al., "Origin and evolution of new exons in rodents," *Genome Research*, vol. 15, no. 9, pp. 1258–1264, 2005.

[2] Q. Zhou, G. Zhang, Y. Zhang et al., "On the origin of new genes in *Drosophila*," *Genome Research*, vol. 18, no. 9, pp. 1446–1455, 2008.

[3] K. Khalturin, G. Hemmrich, S. Fraune, R. Augustin, and T. C. G. Bosch, "More than just orphans: are taxonomically-restricted genes important in evolution?" *Trends in Genetics*, vol. 25, no. 9, pp. 404–413, 2009.

[4] A. C. Sharman, "Some new terms for duplicated genes," *Seminars in Cell and Developmental Biology*, vol. 10, no. 5, pp. 561–563, 1999.

[5] E. L. L. Sonnhammer and E. V. Koonin, "Orthology, paralogy and proposed classification for paralog subtypes," *Trends in Genetics*, vol. 18, no. 12, pp. 619–620, 2002.

[6] E. V. Koonin, "Orthologs, paralogs, and evolutionary genomics," *Annual Review of Genetics*, vol. 39, pp. 309–338, 2005.

[7] K. Durkin, W. Coppieters, C. Drögüller et al., "Serial translocation by means of circular intermediates underlies colour sidedness in cattle," *Nature*, vol. 482, no. 7383, pp. 81–84, 2012.

[8] R. Waterston and J. Sulston, "The genome of Caenorhabditis elegans," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 24, pp. 10836–10840, 1995.

[9] W. M. Fitch, "Distinguishing homologous from analogous proteins," *Systematic Zoology*, vol. 19, no. 2, pp. 99–113, 1970.

[10] K. Wolfe, "Robustness—it's not where you think it is," *Nature Genetics*, vol. 25, no. 1, pp. 3–4, 2000.

[11] K. Ezawa, K. Ikeo, T. Gojobori, and N. Saitou, "Evolutionary patterns of recently emerged animal duplogs," *Genome Biology and Evolution*, vol. 3, no. 1, pp. 1119–1135, 2011.

[12] C. N. Dewey, "Positional orthology: putting genomic evolutionary relationships into context," *Briefings in Bioinformatics*, vol. 12, no. 5, Article ID bbr040, pp. 401–412, 2011.

[13] Y. Nakatani, H. Takeda, Y. Kohara, and S. Morishita, "Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates," *Genome Research*, vol. 17, no. 9, pp. 1254–1265, 2007.

[14] R. F. Furlong and P. W. H. Holland, "Were vertebrates octoploid?" *Philosophical Transactions of the Royal Society B*, vol. 357, no. 1420, pp. 531–544, 2002.

[15] T. Makino and A. McLysaght, "Ohnologs in the human genome are dosage balanced and frequently associated with disease," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 20, pp. 9270–9274, 2010.

[16] R. F. Furlong and P. W. H. Holland, "Polyploidy in vertebrate ancestry: ohno and beyond," *Biological Journal of the Linnean Society*, vol. 82, no. 4, pp. 425–430, 2004.

[17] J. Garcia-Fernandez and P. W. H. Holland, "Archetypal organization of the amphioxus Hox gene cluster," *Nature*, vol. 370, no. 6490, pp. 563–566, 1994.

[18] L. Z. Holland, R. Albalat, K. Azumi et al., "The amphioxus genome illuminates vertebrate origins and cephalochordate biology," *Genome Research*, vol. 18, no. 7, pp. 1100–1111, 2008.

[19] A. L. Hughes, "Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history," *Journal of Molecular Evolution*, vol. 48, no. 5, pp. 565–576, 1999.

[20] G. C. Conant and A. Wagner, "Asymmetric sequence divergence of duplicate genes," *Genome Research*, vol. 13, no. 9, pp. 2052–2058, 2003.

[21] V. J. Lynch and G. P. Wagner, "Multiple chromosomal rearrangements structured the ancestral vertebrate Hox-bearing protochromosomes," *PLoS Genetics*, vol. 5, no. 1, Article ID e1000349, 2009.

[22] P. W. Holland, J. Garcia-Fernàndez, N. A. Williams, and A. Sidow, "Gene duplications and the origins of vertebrate development," *Development*, pp. 125–133, 1994.

[23] N. H. Putnam, T. Butts, D. E. K. Ferrier et al., "The amphioxus genome and the evolution of the chordate karyotype," *Nature*, vol. 453, no. 7198, pp. 1064–1071, 2008.

[24] A. A. Abbasi, "Unraveling ancient segmental duplication events in human genome by phylogenetic analysis of multigene families residing on HOX-cluster paralogons," *Molecular Phylogenetics and Evolution*, vol. 57, no. 2, pp. 836–848, 2010.

[25] N. D. Schaeper, N. M. Prpic, and E. A. Wimmer, "A clustered set of three Sp-family genes is ancestral in the Metazoa: evidence from sequence analysis, protein domain structure, developmental expression patterns and chromosomal location," *BMC Evolutionary Biology*, vol. 10, no. 1, article 88, 2010.

[26] A. L. Hughes and R. Friedman, "Differential loss of ancestral gene families as a source of genomic divergence in animals," *Proceedings of the Royal Society B*, vol. 271, supplement 3, pp. S107–S109, 2004.

[27] E. G. J. Danchin, P. Gouret, and P. Pontarotti, "Eleven ancestral gene families lost in mammals and vertebrates while otherwise universally conserved in animals," *BMC Evolutionary Biology*, vol. 6, article 5, 2006.

[28] D. J. Miller, G. Hemmrich, E. E. Ball et al., "The innate immune repertoire in Cnidaria—ancestral complexity and stochastic gene loss," *Genome Biology*, vol. 8, no. 4, article R59, 2007.

[29] S. Wyder, E. V. Kriventseva, R. Schröder, T. Kadowaki, and E. M. Zdobnov, "Quantification of ortholog losses in insects and vertebrates," *Genome Biology*, vol. 8, no. 11, article R242, 2007.

[30] T. Takahashi, C. McDougall, J. Troscianko et al., "An EST screen from the annelid Pomatoceros lamarckii reveals patterns of gene loss and gain in animals," *BMC Evolutionary Biology*, vol. 9, no. 1, article 240, 2009.

[31] S. C. Le Comber and C. Smith, "Polyploidy in fishes: patterns and processes," *Biological Journal of the Linnean Society*, vol. 82, no. 4, pp. 431–442, 2004.

[32] B. K. Mable, ""Why polyploidy is rarer in animals than in plants": myths and mechanisms," *Biological Journal of the Linnean Society*, vol. 82, no. 4, pp. 453–466, 2004.

[33] O. Jatllon, J. M. Aury, F. Brunet et al., "Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype," *Nature*, vol. 431, no. 7011, pp. 946–957, 2004.

[34] R. D. Morin, E. Chang, A. Petrescu et al., "Sequencing and analysis of 10,967 full-length cDNA clones from Xenopus laevis and Xenopus tropicalis reveals post-tetraploidization transcriptome remodeling," *Genome Research*, vol. 16, no. 6, pp. 796–803, 2006.

[35] M. H. Gallardo, J. W. Bickham, R. L. Honeycutt, R. A. Ojeda, and N. Köhler, "Discovery of tetraploidy in a mammal," *Nature*, vol. 401, no. 6751, p. 341, 1999.

[36] R. Vergilino, C. Belzile, and F. Dufresne, "Genome size evolution and polyploidy in the *Daphnia pulex* complex (Cladocera: Daphniidae)," *Biological Journal of the Linnean Society*, vol. 97, no. 1, pp. 68–79, 2009.

[37] T. G. D'Souza, M. Storhas, H. Schulenburg, L. W. Beukeboom, and N. K. Michiels, "Occasional sex in an 'asexual' polyploid hermaphrodite," *Proceedings of the Royal Society B*, vol. 271, no. 1543, pp. 1001–1007, 2004.

[38] F. Fontana, L. Congiu, V. A. Mudrak et al., "Evidence of hexaploid karyotype in shortnose sturgeon," *Genome*, vol. 51, no. 2, pp. 113–119, 2008.

[39] R. J. Schultz, "Role of polyploidy in the evolution of fishes," in *Polyploidy: Biological Relevance*, W. H. Lewis, Ed., pp. 341–378, Plenum Press, New York, NY, USA, 1980.

[40] A. A. Echelle and D. T. Mosier, "All-female fish: a cryptic species of Menidia (Atherinidae)," *Science*, vol. 212, no. 4501, pp. 1411–1413, 1981.

[41] M. Collares-Pereira, J. Madeira, and P. Rab, "Spontaneous triploidy in the stone loach *Noemacheilus barbatulus* (Balitoridae)," *Copeia*, vol. 2, pp. 483–484, 1995.

[42] X. Yu, T. Zhou, K. Li, Y. Li, and M. Zhou, "On the karyosystematics of cyprinid fishes and a summary of fish chromosome studies in China," *Genetica*, vol. 72, no. 3, pp. 225–235, 1987.

[43] K. K. Rishi, Shashikala, and S. Rishi, "Karyotype study on six Indian hill-stream fishes," *Chromosome Science*, vol. 2, pp. 9–13, 1998.

[44] R. C. Vrijenhoek, R. M. Dawley, C. J. Cole, and J. P. Bogart, "A list of known unisexual vertebrates," in *Evolution and Cytology of Unisexual Vertebrates*, R. Dawley and J. Bogart, Eds., pp. 19–23, The State University of New York, New York, NY, USA, 1989.

[45] K. Janko, J. Bohlen, D. Lamatsch et al., "The gynogenetic reproduction of diploid and triploid hybrid spined loaches (Cobitis: Teleostei), and their ability to establish successful clonal lineages—on the evolution of polyploidy in asexual vertebrates," *Genetica*, vol. 131, no. 2, pp. 185–194, 2007.

[46] K. Arai, K. Matsubara, and R. Suzuki, "Production of polyploids and viable gynogens using spontaneously occurring tetraploid loach, *Misgurnus anguillicaudatus*," *Aquaculture*, vol. 117, no. 3-4, pp. 227–235, 1993.

[47] P. Raicu and E. Taisescu, "Misgurnus fossilis, a tetraploid fish species," *Journal of Heredity*, vol. 63, pp. 92–94, 1972.

[48] A. Chenuil, N. Galtier, and P. Berrebi, "A test of the hypothesis of an autopolyploid vs. allopolyploid origin for a tetraploid lineage: application to the genus *Barbus* (Cyprinidae)," *Heredity*, vol. 82, no. 4, pp. 373–380, 1999.

[49] A. Suzuki and Y. Taki, "Karyotype of tetraploid origin in a tropical Asian cyprinid, *Acrossocheilus sumatranus*," *Japanese Journal of Ichthyology*, vol. 28, pp. 173–176, 1981.

[50] E. Y. Mazik, A. T. Toktosunov, and P. Ráb, "Karyotype study of four species of the genus *Diptychus* (Pisces, Cyprinidae) with remarks on polyploidy of Scizothoracine fishes," *Folia Zoologica*, vol. 38, pp. 325–332, 1989.

[51] J.-T. Wang, J.-T. Li, X.-F. Zhang, and X.-W. Sun, "Transcriptome analysis reveals the time of the fourth round of genome duplication in common carp (Cyprinus carpio)," *BMC Genomics*, vol. 13, no. 1, article 96, 2012.

[52] Y. Shimuzu, T. Oshiro, and M. Sakaizumi, "Electrophoretic studies of diploid, triploid and tetraploid forms of the Japanese silver crucian carp, *Carassius auratus* langsdorfi," *Japanese Journal of Ichthyology*, vol. 40, pp. 65–75, 1993.

[53] J. Gui, Y. Li, K. Li, Y. Hong, and T. Zhou, "Studies on the karyotypes of Chinese cyprinid fishes: karyotypes of three tetraploid species in Barbinae and one tetraploid species in *Cyprininae*," *Acta Genetica Sinica*, vol. 12, pp. 202–208, 1985.

[54] A. Vervoort, "Tetraploidy in Protopterus (Dipnoi)," *Experientia*, vol. 36, no. 3, pp. 294–296, 1980.

[55] R. R. Ewing, C. G. Scalet, and D. P. Evenson, "Flow cytometric identification of larval triploid walleyes," *Progressive Fish Culturist*, vol. 53, pp. 177–180, 1991.

[56] F. W. Allendorf and G. H. Thorgaard, "Tetraploidy and the evolution of Salmonid fishes," in *Evolutionary Genetics of Fishes*, B. J. Turner, Ed., pp. 1–53, Plenum Press, New York, NY, USA, 1984.

[57] N. Pandey and W. S. Lakra, "Evidence of female heterogamety, B-chromosome and natural tetraploidy in the Asian catfish, *Clarias batrachus*, used in aquaculture," *Aquaculture*, vol. 149, no. 1-2, pp. 31–37, 1997.

[58] T. J. Pandian and R. Koteeswaran, "Natural occurrence of monoploids and polyploids in the Indian catfish, *Heteropneustes fossilis*," *Current Science*, vol. 76, no. 8, pp. 1134–1137, 1999.

[59] M. B. Ptacek, H. C. Gerhardt, and R. D. Sage, "Speciation by polyploidy in treefrogs: multiple origins of the tetraploid, *Hyla versicolor*," *Evolution*, vol. 48, no. 3, pp. 898–908, 1994.

[60] B. K. Mable and J. P. Bogart, "Hybridization between tetraploid and diploid species of treefrogs (genus Hyla)," *Journal of Heredity*, vol. 86, no. 6, pp. 432–440, 1995.

[61] B. K. Mable and J. D. Roberts, "Mitochondrial DNA evolution of tetraploids in the genus *Neobatrachus* (Anura: Myobatrachidae)," *Copeia*, no. 4, pp. 680–689, 1997.

[62] J. A. Bailey and E. E. Eichler, "Primate segmental duplications: crucibles of evolution, diversity and disease," *Nature Reviews Genetics*, vol. 7, no. 7, pp. 552–564, 2006.

[63] X. She, Z. Cheng, S. Zöllner, D. M. Church, and E. E. Eichler, "Mouse segmental duplication and copy number variation," *Nature Genetics*, vol. 40, no. 7, pp. 909–914, 2008.

[64] G. E. Liu, M. Ventura, A. Cellamare et al., "Analysis of recent segmental duplications in the bovine genome," *BMC Genomics*, vol. 10, article 571, 2009.

[65] E. Tuzun, J. A. Bailey, and E. E. Eichler, "Recent segmental duplications in the working draft assembly of the brown Norway rat," *Genome Research*, vol. 14, no. 4, pp. 493–506, 2004.

[66] I. C. G. S. Consortium, "Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution," *Nature*, vol. 432, no. 7018, pp. 695–777, 2004.

[67] T. J. Nicholas, Z. Cheng, M. Ventura, K. Mealey, E. E. Eichler, and J. M. Akey, "The genomic architecture of segmental duplications and associated copy number variants in dogs," *Genome Research*, vol. 19, no. 3, pp. 491–499, 2009.

[68] J. A. Bailey, G. Liu, and E. E. Eichler, "An Alu transposition model for the origin and expansion of human segmental duplications," *American Journal of Human Genetics*, vol. 73, no. 4, pp. 823–834, 2003.

[69] A. S. Fiston-Lavier, D. Anxolabehere, and H. Quesneville, "A model of segmental duplication formation in *Drosophila* melanogaster," *Genome Research*, vol. 17, no. 10, pp. 1458–1470, 2007.

[70] J. M. Ranz, F. Casals, and A. Ruiz, "How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus *Drosophila*," *Genome Research*, vol. 11, no. 2, pp. 230–239, 2001.

[71] J. M. Szamalek, D. N. Cooper, W. Schempp et al., "Polymorphic micro-inversions contribute to the genomic variability of humans and chimpanzees," *Human Genetics*, vol. 119, no. 1-2, pp. 103–112, 2006.

[72] R. P. Meisel, "Repeat mediated gene duplication in the *Drosophila* pseudoobscura genome," *Gene*, vol. 438, no. 1-2, pp. 1–7, 2009.

[73] L. Zhang, H. H. S. Lu, W.-Y. Chung, J. Yang, and W.-H. Li, "Patterns of segmental duplication in the human genome," *Molecular Biology and Evolution*, vol. 22, no. 1, pp. 135–141, 2005.

[74] V. Katju and M. Lynch, "The Structure and early evolution of recently Arisen gene duplicates in the Caenorhabditis elegans genome," *Genetics*, vol. 165, no. 4, pp. 1793–1803, 2003.

[75] R. A. Veitia, S. Bottani, and J. A. Birchler, "Cellular reactions to gene dosage imbalance: genomic, transcriptomic and proteomic effects," *Trends in Genetics*, vol. 24, no. 8, pp. 390–397, 2008.

[76] D. Pan and L. Zhang, "Quantifying the major mechanisms of recent gene duplications in the human and mouse genomes: a novel strategy to estimate gene duplication rates," *Genome Biology*, vol. 8, no. 8, article R158, 2007.

[77] A. Bhutkar, S. M. Russo, T. F. Smith, and W. M. Gelbart, "Genome-scale analysis of positionally relocated genes," *Genome Research*, vol. 17, no. 12, pp. 1880–1887, 2007.

[78] D. V. Babushok and H. H. Kazazian, "Progress in understanding the biology of the human mutagen LINE-1," *Human Mutation*, vol. 28, no. 6, pp. 527–539, 2007.

[79] M. D. Lorenzen, A. Gnirke, J. Margolis et al., "The maternal-effect, selfish genetic element Medea is associated with a composite Tc1 transposon," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 29, pp. 10085–10089, 2008.

[80] C. M. B. Carvalho, M. B. Ramocki, D. Pehlivan et al., "Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome," *Nature Genetics*, vol. 43, no. 11, pp. 1074–1081, 2011.

[81] M. Oliver-Bonet, J. Navarro, M. Carrera, J. Egozcue, and J. Benet, "Aneuploid and unbalanced sperm in two translocation carriers: evaluation of the genetic risk," *Molecular Human Reproduction*, vol. 8, no. 10, pp. 958–963, 2002.

[82] C. M. Ogilvie and P. N. Scriven, "Meiotic outcomes in reciprocal translocation carriers ascertained in 3-day human embryos," *European Journal of Human Genetics*, vol. 10, no. 12, pp. 801–806, 2002.

[83] E. M. Chang, J. E. Han, I. P. Kwak, W. S. Lee, T. K. Yoon, and S. H. Shim, "Preimplantation genetic diagnosis for couples with a Robertsonian translocation: practical information for genetic counseling," *Journal of Assisted Reproduction and Genetics*, vol. 29, no. 1, pp. 67–75, 2012.

[84] E. Anton, J. Blanco, J. Egozcue, and F. Vidal, "Sperm FISH studies in seven male carriers of Robertsonian translocation t(13;14)(q10;q10)," *Human Reproduction*, vol. 19, no. 6, pp. 1345–1351, 2004.

[85] L. De Lorenzi, P. Morando, J. Planas, M. Zannotti, L. Molteni, and P. Parma, "Reciprocal translocations in cattle: frequency estimation," *Journal of Animal Breeding and Genetics*. In press.

[86] Z. Ou, P. Stankiewicz, Z. Xia et al., "Observation and prediction of recurrent human translocations mediated by NAHR between nonhomologous chromosomes," *Genome Research*, vol. 21, no. 1, pp. 33–46, 2011.

[87] K. E. Hermetz, U. Surti, J. D. Cody, and M. K. Rudd, "A recurrent translocation is mediated by homologous recombination between HERV-H elements," *Molecular Cytogenetics*, vol. 5, no. 1, article 6, 2012.

[88] M. Lynch and J. S. Conery, "The evolutionary fate and consequences of duplicate genes," *Science*, vol. 290, no. 5494, pp. 1151–1155, 2000.

[89] K. Ezawa, S. OOta, and N. Saitou, "Genome-wide search of gene conversions in duplicated genes of mouse and rat," *Molecular Biology and Evolution*, vol. 23, no. 5, pp. 927–940, 2006.

[90] N. Osada and H. Innan, "Duplication and gene conversion in the *Drosophila* melanogaster genome," *PLoS Genetics*, vol. 4, no. 12, Article ID e1000305, 2008.

[91] K. Fujimura, M. A. Conte, and T. D. Kocher, "Circular DNA intermediate in the duplication of nile tilapia vasa genes," *PLoS ONE*, vol. 6, no. 12, Article ID e29477, 2011.

[92] N. H. Putnam, M. Srivastava, U. Hellsten et al., "Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization," *Science*, vol. 317, no. 5834, pp. 86–94, 2007.

[93] M. Srivastava, E. Begovic, J. Chapman et al., "The Trichoplax genome and the nature of placozoans," *Nature*, vol. 454, no. 7207, pp. 955–960, 2008.

[94] H. C. Seo, R. B. Edvardsen, A. D. Maeland et al., "Hox cluster disintegration with persistent anteroposterior order of expression in *Oikopleura dioica*," *Nature*, vol. 430, no. 7004, pp. 67–71, 2004.

[95] M. D. Adams, S. E. Celniker, R. A. Holt et al., "The genome sequence of *Drosophila* melanogaster," *Science*, vol. 287, no. 5461, pp. 2185–2195, 2000.

[96] P. J. Stephens, C. D. Greenman, B. Fu et al., "Massive genomic rearrangement acquired in a single catastrophic event during cancer development," *Cell*, vol. 144, no. 1, pp. 27–40, 2011.

[97] A. R. Quinlan and I. M. Hall, "Characterizing complex structural variation in germline and somatic genomes," *Trends in Genetics*, vol. 28, no. 1, pp. 43–53, 2012.

Hindawi Publishing Corporation
International Journal of Evolutionary Biology
Volume 2012, Article ID 394026, 14 pages
doi:10.1155/2012/394026

Review Article

The Ecology of Bacterial Genes and the Survival of the New

M. Pilar Francino[1,2]

[1] *Unitat Mixta d'Investigació en Genòmica i Salut, Centre Superior d'Investigació en Salut Pública i Institut Cavanilles de Biodiversitat i Biologia Evolutiva, 46020 València, Spain*
[2] *School of Natural Sciences, University of California, Merced, CA 95343, USA*

Correspondence should be addressed to M. Pilar Francino, francino_pil@gva.es

Received 21 April 2012; Accepted 26 June 2012

Academic Editor: Frédéric Brunet

Much of the observed variation among closely related bacterial genomes is attributable to gains and losses of genes that are acquired horizontally as well as to gene duplications and larger amplifications. The genomic flexibility that results from these mechanisms certainly contributes to the ability of bacteria to survive and adapt in varying environmental challenges. However, the duplicability and transferability of individual genes imply that natural selection should operate, not only at the organismal level, but also at the level of the gene. Genes can be considered semiautonomous entities that possess specific functional niches and evolutionary dynamics. The evolution of bacterial genes should respond both to selective pressures that favor competition, mostly among orthologs or paralogs that may occupy the same functional niches, and cooperation, with the majority of other genes coexisting in a given genome. The relative importance of either type of selection is likely to vary among different types of genes, based on the functional niches they cover and on the tightness of their association with specific organismal lineages. The frequent availability of new functional niches caused by environmental changes and biotic evolution should enable the constant diversification of gene families and the survival of new lineages of genes.

1. Introduction

Genomic science has brought about the possibility of disclosing the genetic underpinnings of entire organisms, and, for microbes, even of complex populations and communities. In doing so, it has unearthed a good number of surprising facts regarding the structure, organization, and variability of genomes, with strong evolutionary implications. In bacteria, the relatively small size of genomes has warranted the obtention of entire genomic sequences for a vast number of organisms, including numerous sets of sequences of closely related taxa. As genome sequences of closely related bacteria have accumulated, our view of bacterial genomes has radically changed.

Genome comparisons have demonstrated that little 16S rRNA sequence divergence can be accompanied by large differences in total gene repertoire [1–7], and that even populations of a single 16S rRNA species can be made up of vast numbers of genomic varieties [8]. Much of the observed variation among closely related bacterial genomes has been attributed to gains and losses of genes that are acquired horizontally, often by way of mobile genetic elements (MGEs) as well as to a variety of genomic rearrangements that include gene duplications and large gene amplifications [1, 2, 4–6, 9]. This wide extent of genomic variability speaks to flexible and dynamic genomes and undoubtedly contributes to the amazing ability of bacteria to survive and adapt to varying environmental challenges. On the other hand, from a gene-centric perspective, the duplicability and transferability of individual genes imply that natural selection should operate, not only at the organismal level, but also at the level of the gene. It has long been recognized that MGEs can evolve selfishly [10, 11]. However, because of HGT, the distinction between selfish, mobile DNA and the sedentary host genome is blurry and it can be argued that all genes are semiautonomous entities capable of responding to natural selection at different levels: selfishly, because they can "reproduce" by duplication and "migrate" by horizontal gene transfer (HGT), and cooperatively, because their long-term survival depends on their ability to be replicated as parts of integrated genomes [12].

2. Genes as Ecological Semiautonomous Entities: The Gene's Functional Niche

Given the ubiquity of MGEs and the numerous HGT mechanisms in operation in the prokaryotic world, a given gene may be transported to any number of organisms in a variety of environments. However, not all of these may represent suitable venues for the incoming gene to persist. To be stably maintained in the long term, a gene would need to encode a function that conferred a valuable adaptation to the organism. An incoming gene not providing a novel capability to the recipient would not be selected for, not spread in the recipient population and eventually be obliterated by the high substitution and deletion rates operating in prokaryotic genomes [13–15]. So, a gene can be considered to have a potential range of genomes in which it could survive, by means of providing a selectable function. The functional need covered by the gene can be considered its very own niche, the defined functional space that would allow it to survive in the range of genomes where such a need exists.

Following this analogy, genes will have, as do organismal species, fundamental and realized ecological niches. The fundamental niche of a gene can be defined as the entire range of genomes where such a gene might survive, in case it were capable of reaching them and they did not contain a similar gene realizing the same function. The realized niche will encompass the actual range of genomes where the gene has been established. In many cases, the realized niche of a gene will likely be substantially narrower than its fundamental niche. Besides physical limitations to gene dispersal, this could result from the fact that many populations of potential hosts may contain a better-adapted gene occupying or competing for the same functional niche, be it an ancestral gene or a different gene acquired by HGT. The following paragraphs explore the different types of gene niches and the likelihood that they are realized by horizontally acquired genes.

2.1. Housekeeping Ancestral Functions. Gene functions that are essential for life are usually covered by ancestral genes that have been associated with their host genomes throughout evolutionary history, being passed on through generations via vertical inheritance [16, 17]. Since they are already represented in every living organism, there is no empty niche available into which these genes might expand, and therefore their realized niche is identical to their fundamental niche. Moreover, the long-term permanence of these genes within their specific genomes should have enabled a high degree of coadaptation with other stable members of the genomic community, favoring the evolution of a well-integrated gene core. Therefore, the niches represented by these functions are likely to be stably occupied by finely adapted genes that should be almost impossible to displace by incoming, horizontally transferred genes evolved in different genomic backgrounds. So, given that empty niches are not available and that displacement of orthologs in different genomes should be unlikely, horizontal transfer events should rarely be successful for these genes.

2.2. Ecologically Restricted Essential Functions. Functions that are required for survival or that substantially improve fitness only for specific habitats and/or lifestyles represent niches that may be available to horizontally transferred genes under certain circumstances. Such functions are often first acquired by HGT [18] so that the level of coadaptation between the gene that occupies the functional niche and its genomic companions will depend on the time elapsed since the HGT event occurred and the organism in question adopted its current lifestyle. Genes that were recently acquired by an organism and enabled it to occupy a new niche may have become indispensable for the organism's survival, but may not be better suited to their role than genes sharing the same fundamental niche present in the genomes of other species. In such cases, competing incoming genes may have a selective advantage and be able to replace the original gene. In addition, nonhousekeeping functional needs, even when essential, may be more sensitive to varying environmental conditions and be altered as organisms adapt to a new environment or radiate into a range of similar lifestyles. These circumstances should also provide opportunities for displacement of the resident genes by newcomers transferred horizontally.

2.3. Nonessential Accessory Functions. A large fraction of the genes unveiled by sequencing projects is present across only some of the genomes of a given bacterial species. This gene fraction has been termed the dispensable or accessory genome and can amount to a large proportion of the total gene repertoire detected across the species, the pangenome [3]. It includes numerous functions that are not generally required for survival, but that may ameliorate adaptation to transient conditions or patchy environments. Shifting environmental conditions should often generate vacant functional niches, transiently made available to a variety of existing genes capable of fulfilling the required role after horizontal transfer. Many factors might come into play in determining which of the possibly numerous genes sharing the available fundamental niche will eventually occupy it. Such factors should include the physical proximity and phylogenetic relatedness among potential donor and recipient organisms, which increases the likelihood of a successful gene transfer [19, 20] as well as the likelihood that the gene might be associated with MGEs capable of reaching the potential recipient. If different genes were transferred, competition among them should ensue and the one better adapted to the available functional niche should spread more widely into the recipient population.

2.4. Selfish, Gene-Centric Functions. Some genes may not require a preexisting niche provided by a functional need at the organismal level. This is probably best exemplified by systems that behave as addiction modules, or poison/antidote systems, which commonly spread via plasmid-mediated HGT [21]. Such systems consist of two genes acting as a toxin and an antitoxin for the cell that carries them. The toxin kills cells if expressed above a certain level, and the antitoxin inactivates the toxin and/or regulates its expression, thereby

preventing cell killing. The toxin is more stable and long-lasting than the antitoxin so that constant (over)production of antitoxin is required for cell survival. Thus, toxin/antitoxin systems can be said to carve their own niche into the genome, as their permanence is ensured independently, and possibly in detriment, of organismal-level requirements [22, 23]. Although not mobile, these two-gene systems can be considered selfish in the sense that their primary function is self-preservation, similar to insertion sequences and other transposable elements [10, 11].

3. Creation of Functional Niches and Survival of the New

Most importantly, novel functional niches are likely to appear constantly, as environments change and species evolve, enabling the evolution of novel genes. Although radically new biochemistries and structural functions may be required more rarely, specialization of preexisting functions to meet modified organismal needs should often be advantageous. In addition, any existing functional niche could likely be subdivided into multiple narrower niches, which could be occupied by genes with higher degrees of functional specialization. Numerous cases of specialization of different homologs into related but distinct functional niches have been documented, including the evolution of affinity for different substrates or cofactors [24–28], of different enzymatic kinetic profiles [29, 30] and of different interaction patterns with other proteins [31]. Functional niches for homologous genes with different specializations can be available within single organisms, when their lifestyle encompasses variable environments [32], different developmental stages [33] or dynamic interactions with hosts [31]. In fact, antagonistic biotic interactions, such as those between pathogenic bacteria and their hosts or those between bacteria and their phages, can lead to arms races that permanently favor diversity [34]. Such processes should constantly generate new niches for bacterial genes, differing from existing ones by the novel constraints imposed by changes in the interacting organism. Clearly, similar but distinct gene niches can also be generated when different organisms specialize into different subdivisions of an organismal niche, as occurs during adaptive radiations. A beautiful example of concordant gene and organismal radiations has recently been described within soil archaeal ammonia oxidizers, where clades of *amo*A, a key functional gene of ammonia oxidation, dominate specific ranges of soil pH [35].

Functional niches likely to be substantially different from those covered by existing genes are also being created today by the plethora of human-made compounds released into the environment. Enzymatic pathways capable of degrading some of these compounds have readily evolved, including pathways for the degradation of the pesticides atrazine [36], pentachlorophenol [37, 38] and 1-3-dichloropropene [39], and of chloronitrobenzenes and dinitrotoluenes used in the production of industrial chemicals and pharmaceuticals [40, 41]. Most of these chemicals are highly toxic and mutagenic, which has likely represented a strong selective pressure for the rapid evolution of degradation pathways. However, many human-made compounds are highly recalcitrant to degradation and have only been present in the environment for a few decades so that evolution has had limited time to produce genes and enzymes to fill the novel niches they provide. For example, no naturally occurring microbes are known to completely mineralize the dielectric fluid PCB or the insecticide paraoxon although they can be partially detoxified. The products of extant reactions of detoxification or incomplete degradation of anthropogenic chemicals, such as the partially dechlorinated compounds resulting from reductive dehalogenation of PCB [42], can accumulate in the environment and provide suites of novel functional niches. Some anthropogenic compounds cannot be degraded by any single organism, but their degradation can be accomplished by enzymes from different microbes working together, as occurs with the explosive trinitrotoluene (TNT). In this case, none of the participant microbes carries out enough of the involved reactions to reap a metabolic benefit from TNT degradation [43]. Therefore, there still remains a niche, or niches, for the evolution of integrated functional pathways within single microbes that can utilize TNT as a novel source of carbon, nitrogen, or phosphorous.

A special case of human-generated selective pressure for the evolution of novel bacterial gene functions is provided by the widespread use of antibiotics. Although some antibiotic resistances may have preceded the use of antibiotics by humans, serving to defend bacteria from chemical warfare from other microbes, it is clear that the large variety of resistance genes existing today attests to a rampant diversification. Novel resistance genes evolve and spread rapidly, with the typical timeframe for worldwide dissemination of a newly emerged gene being under three years from the initial deployment of the antibiotic [44, 45].

How are the partition of functional gene niches and the occupation of novel ones achieved, often in record times? The capacity of proteins to alter their substrate ranges has been well documented during experimental evolution [46, 47]. Enzymes subject to directed laboratory selection under high rates of mutation and recombination rapidly increase their level of activity on new substrates by several orders of magnitude over the wild type. On the other hand, computational analyses of molecular dynamics indicate that the substrate range of an enzyme can just as easily be narrowed down. For instance, the alkaline phosphatase of *Escherichia coli* acts on both phosphomonoesters and phosphodiesters through a single-reaction mechanism, but simulation studies indicate that specialization for hydrolysis of mono- or di-esters, seen in other members of the alkaline phosphatase superfamily, depends on mutations that alter the nature or positioning of a single amino-acid residue [48]. Thus, niche partitioning may often result from the small alteration of existing gene functions, which could come about through a small number of mutational events. The occupation of new niches can also be enabled by small modifications that alter enzyme specificity. For instance, in *Pseudomonas diminuta*, a phosphotriesterase enzyme specialized in the utilization of pesticides is thought to have evolved recently from an amidohydrolase endowed with several promiscuous activities, and

experiments with a homologous amidohydrolase have shown that the substitution of a single amino-acid can radically change the specificity of the enzyme [29]. The evolution of novel functions that are more distinct from preexisting ones may necessitate a wider arsenal of genetic alterations. As an example, the newly evolved enzymes that degrade chlorinated aromatics in the environment bear in their sequences the hallmarks of complex genetic rearrangements, including recombinational events mediated by transposable elements [49].

The accumulation of mutations that generate novel functions usually occurs during divergence between homologous genes. The process of divergence can be initiated in paralogy, after gene duplication within a given genome, or in orthology, following organismal speciation. The following section explores the potential differences between these two modes of evolutionary divergence and their consequences for the evolution of new functions as well as the role of processes that reassort the evolutionary novelties generated during divergence.

4. Genes as Phylogenetic Semiautonomous Entities: Gene Lineages

4.1. Orthologous Gene Divergence. After organismal speciation, newly created pairs of orthologous genes will start to diverge, at a rate that will depend, among other factors, on the degree of recombination between the two daughter species at that specific gene locus. In bacteria, the capacity of genetic exchange across species boundaries implies that orthologous genes may continue to recombine for a certain time after two incipient species have formed, and even among closely related species, at rates dependant on the amount of sequence divergence at that locus and on the potential fitness decrease caused by recombinant gene products [50–52]. If the species remain in contact, recombination between orthologous genes may continue as long as the accumulated level of sequence divergence does not pose an impediment for homologous recombination (HR) mechanisms [53–55]. If the organismal niches occupied by the two species do not pose different selective pressures on a given gene locus, most of the divergence that will accumulate between orthologs should be neutral, and functional divergence should not occur, barring the potential fixation of deleterious changes due to drift. If no functional divergence occurs, recombinant gene sequences are likely to be equally fit as the parental ones, at least until species-specific patterns of codon usage or other sequence-level adaptations begin to take place.

In contrast, if the organismal niches of the daughter species represent different functional niches for a given gene, divergence should occur under positive selection on one or both orthologs. In this case, recombination events between orthologs may produce sequences that are less fit in their corresponding niches than the parental ones and should be selected against. This should facilitate further divergence and specialization of one or both orthologs for their specific functional niches. As a result, such orthologs may be considered different evolutionary entities, as they fill different functional niches and evolve as separate lineages not incurring genetic exchange.

4.2. Paralogous Gene Divergence. The processes of gene duplication or amplification will generate gene copies that are identical to each other and often clustered within a genome [56, 57]. Although genetically unstable due to loss by HR, such sets of identical paralogs may be maintained in the genome if the increase in the amount of gene product is beneficial to the organism [57–62]. Although the fitness effects of most duplications, as those of other mutations, will likely be deleterious or neutral, there is ample evidence that gene amplification and increased production of the encoded protein can occasionally provide specific selective advantages. Adaptation by means of gene duplications and larger amplifications has been repeatedly documented in bacteria. Gene amplifications have been implicated in enhanced virulence in pathogens as well as in increased production or fixation of host-required nutrients in symbionts. Gene amplification has also been demonstrated to underlie instances of bacterial resistance to antibiotics and heavy metals [63], as well as experimental adaptation to growth at high temperature [64], and on limiting or unusual carbon sources [65, 66]. Such sets of paralogs that continue to occupy the same functional niche will evolve more or less cohesively, depending on the level of recombination among the different copies. Divergence among them may actually be deleterious so that change-of-function mutations would be selected against while gene conversion events that maintained the original sequence would be selected for.

However, it seems unlikely that two or more identical genes would be maintained in long-term evolution for the purpose of increasing the amount of gene product. To that effect, alternative, probably more efficient strategies would be available to an organism, such as increases in the expression of one gene copy through regulatory changes. "Clonal" gene amplifications are, therefore, likely to be transient in nature. Analyses of gene family sizes indicate that the amplifications with the largest numbers of gene copies are usually very recent [61, 67–72], and calculations of the age of appearance of gene duplicates within bacterial clades indicate that most of them are young [9], implying that, overall, most paralogs do indeed disappear rapidly. If higher gene dosage is the only pressure that maintains gene amplification, the high deletion rates operating in bacteria [13–15] should favor the rapid elimination of superfluous gene copies once regulatory variants producing higher amounts of product appear, or if selection for higher dosage wanes. In such cases, most of the gene clones can be expected to eventually disappear, unless they are rescued by adaptive mutations that allow them to occupy a novel functional niche.

If there is formation of new, separate niches, new functions can evolve and genic lineage diversification will occur. Selection for higher gene dosage may actually often respond, not to a requirement for higher levels of a gene product's extant function, but to a novel functional need that can be partially accomplished by the amplification of an existing gene. The evolution of the new function may

start with the amplification of a gene having some level of preadaptation for that function. Gene amplification would provide the means to attain biologically significant levels of functionality, as the efficiency of the preadapted gene product for the new function would presumably be low [58, 61]. Amplification could be followed by positive selection to adapt the gene product to the novel requirements. A period of competition among the different evolving paralogs in the population might ensue, resulting in the preservation of the most effective variant and the likely pseudogenization and eventual loss of the rest [61]. As in the case of orthologs adapting to different niches, recombination among paralogs diverging in function should probably be selected against, facilitating their separation into independently evolving gene lineages. There is ample evidence that paralogous gene amplifications have resulted in diversified functions of high adaptive value, including expansions of metabolic and regulatory capabilities [28, 73, 74], sensory complexity [75], or antigenic variation [76].

4.3. Advantages of Paralogous Divergence for the Diversification of Gene Function. Both gene duplication and speciation are processes that enable cladogenesis, the generation from a common ancestor of new evolutionary entities that occupy different functional niches and evolve as separate lineages. In orthology, however, the new gene lineage created by the splitting-off of a new species may remain constrained by similar selective pressures if the functional need it served in the ancestor remains unchanged. In time, the new gene lineage will likely adapt to its specific genomic and environmental context, but significant functional divergence should occur only if the lifestyle of the novel species changes substantially and in a manner that alters the functional need served by the gene. In contrast, gene duplications and larger amplifications generate a number of replicas of the same gene within a single organism, so that each one of them may be free to engage in different evolutionary paths, including the retention of the original function in one or a subset of the gene copies.

In particular, the generation of new gene functions may be most facilitated in the context of large gene amplifications that are positively selected from their inception. As mentioned above, amplification of an existing gene with some level of preadaptation to the newly required function may be a first adaptive strategy when an organism is confronted with a new functional need. In this case, the duplicates would be maintained by natural selection, ensuring the permanence in the population of a large number of gene copies that could be the target of mutations with a potential ameliorating effect on the novel function. Moreover, the existence of multiple gene copies would enable the simultaneous exploration of different zones of the adaptive landscape, including fitness valleys that might allow them to transition into separate adaptive peaks, potentially leading to distinct functions. During this process, the existence of related gene sequences within the same genome, and likely in proximity to one another [62], would allow for recombination to occur, which, although potentially hindering the process of divergence, might also bring together beneficial mutations acquired in different paralogs or purge those that are detrimental. Finally, this mode of evolution might be reinforced by the sequential acquisition of beneficial mutations alternating with rounds of selected amplifications of the best-adapted paralogs at every step. In the long term, if new but related functional niches continue to appear due to changes in the environment or in the organism's lifestyle, families and superfamilies of paralogous genes may be generated in adaptive radiations analogous to those observed for species lineages [61].

4.4. Mixing It Up: Recombination Among Genes and HGT across Organismal Lineages. Once novel genes with distinct functional niches have evolved, in orthology or in paralogy, they may be available for filling those niches in other organisms through HGT. As evidenced through the sequencing of bacterial isolates and environmental metagenomes, the divergence processes just described have created gene families that encompass an enormous variety of related sequences. Functional characterization of these sequences lags way behind their discovery, but it is likely that many of them encode related but distinct protein functions. Consequently, when a novel functional niche opens within an organism, a large number of existing related genes may be able to fill it in similar, but probably not identical, manners. Genes with functions similar to the one required could be encoded within the organism's genome so that the niche could be filled by the processes of amplification and divergence developed above. Alternatively, the niche may also readily be filled via horizontal acquisition of a foreign gene that might already be well-suited to the novel need. Functional niches that originate as organisms undergo substantial lifestyle modifications should be more often filled by HGT, given the capacity of this process to bring in functions radically different from those already encoded by the organism. Accordingly, HGT has been repeatedly documented to be a prime contributor to the adaptation of bacteria to novel environments [4, 18, 77–82]. Similarly, the adoption of symbiotic or pathogenic lifestyles, as well as their diversification in terms of host range and tropism, is most often enabled by the acquisition of genes encoded within MGEs [78, 83–94]. Moreover, bacterial subpopulations that gain access to very different niches by HGT may readily become independent lineages, and thus HGT can be considered a motor of prokaryotic speciation and long-term diversification [18, 50, 77, 78, 95, 96]. Amazing examples of appearance of major lineages with specific biologies introduced by HGT include the emergence of bacterial methanotrophs via acquisition of archaeal genes [97] and that of cyanobacteria via the gain of a second photosystem allowing for oxygenic photosynthesis, possibly transferred from the Firmicutes [98].

Clearly, the existence of HGT enables the dissociation of gene and organismal lineages. By reassorting genes across organisms, HGT may constitute an important driver of further gene lineage diversification. Horizontally transferred genes will be exposed to different ecological and genomic

environments and their associated selective and mutational pressures, which should favor divergence from the genes in the donor population. Several studies have confirmed that recently acquired genes have an accelerated rate of evolution in comparison to that of ancestral genes in the same genome [2, 88, 99, 100]. This could respond to relaxed selective pressure in genes not conferring a significant advantage to the host, to neutral substitutions due to the host's mutational biases, or to positive selection for adjusting gene expression or protein function to the specific needs of the host. It has been documented that, in some cases, proteins may undergo significant shifts in function after HGT. For example, omptins are a family of outer membrane proteases that have spread horizontally through a large variety of Gram-negative bacteria infecting vertebrates and plants, and their functions have been substantially modified to adapt to these different lifestyles [101]. Genes that have been acquired by HGT have also been shown to undergo more duplication events than ancestral genes [102]. Gene duplication following transfer could compensate for suboptimal expression or activity of genes and proteins that are not adapted to their novel genomic and cellular backgrounds and could facilitate the appearance of better-adapted gene copies or the evolution of paralogs displaying more substantial functional shifts. So, although gene duplication and HGT independently contribute to fill in new functional niches, and may be best suited to do so under different circumstances, the two processes may sometimes be coupled in the course of adaptation.

Another factor likely to impact the evolution and diversification of horizontally transferred genes in a substantial manner is the fact that these genes may spend significant periods of time in association with MGEs. This association should expose them to mutational pressures different from those affecting chromosomal genes due to the specific replication modes employed by such elements. In fact, most bacteria contain large numbers of ORFans, that is, annotated genes that are restricted to a particular genome and that possess no known homologs in any other organisms, and the sequence characteristics of these genes, including short length, high AT content, and often phage-like dinucleotide frequencies, have been considered to be hallmarks of substantial periods of evolution within phage genomes [2].

Besides sorting out entire genes across organisms, the capacity of bacteria for genetic exchange can also diversify gene lineages through HR among related genes that have diverged under orthology or paralogy. Although normally strongly constrained by sequence differences, HR can operate on more divergent sequences when the SOS system is induced, or in mutants of the methyl-directed mismatch repair system (MMRS) [103]. When operating on closely related genes, HR will mostly serve to limit their divergence, but the occasional exchange between diverged genes that have reached different functionalities can create novel capacities not present in the parental sequences. This phenomenon has been documented in the generation of novel resistance phenotypes among antibiotic resistance genes and is thought to have been a major source of diversification for genes such as $bla_{\text{CTX-M}}$, *ampC*, and *qac* [45]. HR also generates functional diversity in gene families, such as the histidine kinases involved in signal transduction, with members that contain variable combinations of sequence domains, by enabling domain-shuffling among paralogs [74, 104]. In addition, MGEs encode more than a hundred different enzymes capable of recombining DNA at short specific nucleotide sequences, without the long stretches of homology required for HR [105], that could potentially contribute to the creation of novel genes by recombination of sequences from different origins. However, unlike the case of eukaryotes, where rearrangements promoted by MGEs contribute significantly to the generation of novel chimeric genes [106], this process should be rare in bacteria, as the uninterrupted nature of bacterial coding sequences must severely limit the chances of a foreign sequence being integrated without disrupting gene function. Nonhomologous recombination, though, may be an important driver of the evolution of new genes within bacteriophages and may contribute to the generation of the phage-derived ORFans that abound in bacterial genomes [107].

5. Emergent Genomes within Organismal Lineages

Within the described context of genes as ecological and phylogenetic semiautonomous entities that possess their own functional niches and undergo their own historical processes, the genome appears as an emergent gene community, a moving picture of gene associations, some of which will endure, while others will be transient in nature. The genome not only provides the material framework in which the genes are embedded, but also the information required to regulate their interactions. The result is a community of interdependent genes that operate as an integrated whole. This picture is similar to that of ecological communities of organisms, which are capable of maintaining regulated dynamics and species interactions even if their exact species composition fluctuates through time and space [12].

5.1. Variable Levels of Association between Gene and Organismal Lineages. Clearly, the length and tightness of the association between gene and organismal lineages can vary greatly, depending on the type of functional niche occupied by the gene. Genes that perform housekeeping ancestral functions are essential for life and usually remain associated with the same organismal lineage through vertical inheritance. These essential genes will evolve under strong selective constraints that will simultaneously penalize gene loss, accelerated sequence evolution rates, horizontal transfer and, for highly interactive proteins, gene duplication, in order to maintain a stable, host-specialized, and coadapted genomic core. Such constraints signify that the phylogenetic history of the genes of the universally essential genomic core should mostly parallel that of the organismal lineages where they reside. From the point of view of reconstructing the deep phylogenetic history of organismal lineages, these genes should represent the best available markers. Representative

genes of this evolutionary mode are those involved in transcription and translation, which encode highly interactive coadapted protein complexes and rarely undergo HGT [108–114].

Beyond the core of genes essential for life, many genes may be stably associated with specific organismal lineages, even though they may have been first acquired by HGT in an ancestral species. The numbers and types of genes that are conserved increase significantly within progressively shallower phylogenetic lineages [115]. For instance, whereas all prokaryotes almost certainly share less than 50 genes, over 100 genes are common to all bacteria [116], and 205 single-copy genes are present across all the γ-Proteobacteria, a large and ancient group that originated over 500 million years ago [117]. Within the γ-proteobacterial enteric family, the core genomes of the large and well-characterized species *E. coli* and *S. enterica* have been estimated at around 1000 and 2800 genes, respectively [118, 119]. Besides informational genes, these species-level cores mainly include genes involved in the biosynthesis of aminoacids, nucleotides, cofactors and proteins as well as in the metabolism of DNA, fatty acids, and phospholipids [120]. However, genes that are ubiquitous within restricted phylogenetic clades, such as species or genera, may not be stably associated with specific lineages within the clade, but rather may undergo frequent horizontal shuffling within these groups, depending on the population structure and level of ecological divergence of the different lineages, the level of functional niche specialization of the gene, and the capacity of the organisms to incorporate exogenous DNA.

Finally, a large fraction of genes present erratic patterns of presence across different organisms, including closely related strains, and phylogenetic relationships that are indicative of frequent HGT among organisms with varying degrees of relatedness. Some of these accessory genes are clearly associated with narrow but defined niches or lifestyles, most notably in the case of pathogens and symbionts, where genes acquired horizontally within specific strains can remain in their genomes for long periods of time. However, many other accessory genes encode functions that may ameliorate adaptation to transient conditions or patchy environments. As such, their long-term permanence in a given organismal lineage is not warranted. Phylogenetic analyses confirm that most horizontally acquired genes are eventually lost [121–123]. In fact, the high rate of horizontal gene acquisition that has been documented for many bacterial genomes must have been compensated by a similarly high rate of gene loss, otherwise genomes would continuously increase in size. Therefore, the makeup of the accessory component of genomes must continuously fluctuate through time. Moreover, analyses of genomic diversity at the population level indicate that the presence of accessory genes can be variable even among cooccurring cells within local populations [4]. This local population-level variability suggests that these accessory genes are not stably associated with subspecific lineages adapted to particular but defined environmental conditions, that is, ecotypes [124], but rather that their presence in a given genome may respond to transient selective pressures in a variable environment. Extensive temporal and geographic sampling of natural microbial communities at variable scales may be able to reveal whether specific accessory genes are stably or preferentially associated with certain genomic variants, or whether their distribution indicates a more nomadic existence with frequent transfers among variants and/or across different organismal lineages [12].

For genes that only maintain transient associations with specific genomes, long-term survival may be ensured by HGT across a variety of organisms that only occasionally require the function provided by the gene. The likelihood that gene lineages might survive by this strategy should be linked to (1) the fitness increase provided by the gene, (2) the pattern and scale of environmental variability, and (3) the transmissibility of the gene within and among organismal lineages. For instance, accessory genes that provide strong fitness advantages under transient conditions may temporarily reach high frequencies within local populations or communities, and their long-term maintenance may be ensured by their capacity to disperse to other microbial communities during such periods of high abundance. This may be the case of accessory genes that confer resistance to broad range antibiotics or other strong selective agents that affect a variety of species but have a patchy distribution. On the other extreme, accessory genes for which presence is neutral or even detrimental relative to organismal fitness may be able to survive via high levels of transmissibility. This scenario may be approximated by addiction modules, such as toxin-antitoxin systems, which have extensive horizontal mobility due to their associations with plasmids, phages, transposons, or integrons, while most likely presenting scarce benefits to organismal fitness [12, 22, 23], although potential roles for these systems in bacterial stress adaptation have been proposed [125, 126].

5.2. Enabling Gene Cooperation in Emergent Genomes. Whatever their origin, the different genes that coexist within a genome at a given time must operate within the larger context of this higher level of organization. The genomic framework in which the genes are embedded must encode the information required to orchestrate their function in order to face the demands of the environment. In addition, various structural organization constraints will affect the physical distribution of genes within the genome.

Incoming genes should not be able to integrate at any genomic location. Clearly, gene insertions within coding regions and other functional sequences will often be deleterious, but more subtle aspects of genome organization also need to be considered, as genes are not randomly distributed within the genome, but are rather organized at several different levels. Many genes are part of operons, and the disruption of operon structure may negatively impact the coordinated expression of the constituent genes. Larger organizational domains are also present in bacterial chromosomes, associated with structural constraints imposed by the processes of DNA replication and segregation at cell division, the disruption of which strongly affects fitness [127–130]. Also, patterns of DNA supercoiling along the chromosome have been implicated in coordinating the expression levels

of contiguous genes in a manner that varies along the bacterial growth cycle [131]. Therefore, the wide fluctuations in gene composition that characterize bacterial genomes need to preserve these fundamental structural properties. This may be accomplished in part by limiting changes to specific genomic regions. The observed variability among the genomes of closely related organisms is indeed often confined to a few genomic locations, while the overall genomic framework is tightly conserved [120]. In addition to the likely role of natural selection, sequences and/or genomic architectures that serve as hotspots for recombination can contribute to the generation of such a pattern. For instance, many genomic islands and phages integrate preferentially next to tRNA genes [132], and these sites often concentrate a substantial fraction of a genome's recently acquired genes.

Bacterial genomes are also well integrated in terms of coadaptation among their different components, such as the molecules that participate in complex multimeric enzymes or structures. Another amazing example of coadaptation is that between the complement of tRNAs in a genome and the codon usage bias of its coding sequences. In many bacteria, the most highly expressed genes utilize restricted sets of codons that display optimal interactions with the most abundant tRNA species for a given aminoacid present in the genome, allowing for fast and accurate translation of their mRNA [133–137]. Genes from exogenous origin will display codon usages that are distinct from those of the host genome, especially if they are incoming from a genome of different GC content or if they have spent substantial amounts of time associated to phages, which have very elevated AT frequencies. Expression of foreign genes with a codon usage that is not matched to the recipient cell will then be compromised [138]. Accordingly, it has been shown that most recently acquired genes have a codon usage similar to that of the recipient genome at the moment of introgression, indicating that codon usage compatibility is likely to increase the fixation probability of transferred genes [139]. Nevertheless, experimental evidence shows that genes with poorly matched codon usages can be retained if they confer a strong selective advantage, and that their expression level can be rapidly adjusted by regulatory changes [138]. Moreover, the mutational and selective processes particular to the host genome should in time modify their codon usage towards patterns typical of ancestral genes [2, 140, 141].

Most importantly, the expression of the assortment of genes present in the genome needs to be regulated and coordinated for meaningful biological function. Exogenous genes may arrive within the genome accompanied by cognate regulatory sequences and regulator genes. Examination of the evolutionary histories of transcription factors in *E. coli* indicates that many specific regulators were horizontally acquired along with the adjacent genes that they regulate, whereas global regulators are encoded by genes that evolved vertically within the γ-Proteobacteria. Also, horizontally transferred genes are often regulated by multiple regulators, with most of this complex regulation probably evolving after transfer [142], speaking to the likely existence of strong selective pressures to fine-tune their gene expression and integrate it within the context of global regulatory networks. As an example, the expression of virulence genes of the *Salmonella* SPI-1 and SPI-2 pathogenicity islands is controlled by a complex regulatory cascade involving several global regulatory systems as well as specific regulators [143, 144]. The integration of transferred genes into the host's regulatory network appears to occur gradually over evolutionary time, as genes resulting from increasingly ancient transfer events show increasing numbers of transcriptional regulators as well as improved coregulation with interacting proteins. In addition to the recruitment of existing transcription factors, increased integration is accomplished by sequence evolution of the cisregulatory regions and changes in the codon usage of the transferred genes [145].

The topology of the networks of gene and protein interactions may facilitate or hinder the incorporation or loss of accessory genes. Comparative analyses have shown that gene networks often contain a core of ancestral genes involved in large numbers of interactions (hubs) that are highly conserved across species, while genes that are progressively acquired during evolution encode less connected and less central proteins. Therefore, regulatory [146], metabolic [147, 148] and protein interaction networks [146, 149] appear to grow by acquiring genes in the periphery. This network topology and mode of growth clearly enable a flux of accessory genes onto a core genomic framework, allowing for niche exploration and adaptation to changing environments. Another common property of gene networks that facilitates the exchange of accessory components is modularity. Bacterial regulatory and metabolic networks are often organized in well-defined modules, sets of genes or proteins that are strongly interconnected and with a function that is separable from those of other modules. This property is believed to be one of the main contributors to the robustness and evolvability of biological networks [150]. Simulation analyses have shown that modularity allows for specialization in gene activity because it decreases interference between different groups of genes and facilitates cooption, the utilization of existing gene activity to build new functional patterns [151].

Clearly, genomes are shaped by structural and organizational properties that ensure their existence as coherent levels of organization in the face of the malleability conferred by gene duplicability and horizontal transfer. Such properties can be considered emergent properties [152] of the genome because they ensue from the relationships among its different components (sequences with coding, regulatory, or structural functions) and shape its global interaction with the environment. Emergent genome properties will be acted upon by natural selection at the organismal level, which will eliminate unfit combinations of genes or interactions. Moreover, organismal-level selection should favor the appearance, maintenance, and refinement of emergent genome properties, such as genomic architectures and gene network topologies, that enable the organized gain, loss, and reshaping of functional capacities to facilitate organismal adaptation, specially under conditions of frequent environmental change.

6. Concluding Remarks

Beyond gene and organismal selection, the exchange of genes among individuals from different species generates genetic relatedness between them and further diversifies the levels of organization at which natural selection may act. In particular, by increasing relatedness at transferred loci, HGT could favor the evolution of cooperation among gene-exchanging individuals [153–156], although differences in relatedness between mobile loci and the rest of the genome raise the possibility of conflict regarding what type of interaction with a given neighbour might be most advantageous. More generally, genes that undergo frequent transfers among organismal lineages represent a supraspecific gene pool that can increase the fitness of individual organisms belonging to different species. In a given microbial community, organisms from many different species may have access to the same supraspecific gene pool, which will be composed of transferable genes present in the genomes of the community's microbes and MGEs, or brought in by immigrants from other communities. For instance, metagenomic analyses of the human gut microbiome have evidenced that identical antibiotic resistance genes can be shared by bacteria belonging to different bacterial phyla within a single individual [157]. In addition, the MGEs present in the community will largely shape the type, rate, and directionality of the HGT processes that will distribute the gene pool. Under certain conditions, such as those involving frequent environmental fluctuations, the survival of entire communities may depend on their metagenomic pool of transferable, accessory genes, and on the type and rate of HGT processes that may distribute it among organisms facing similar selective pressures [12]. Therefore, the supraspecific pool of accessory genes and the community's capacity for HGT should be considered emergent community properties that may enable the operation of natural selection at the community level.

So, in the bacterial world, the capacity of genes to duplicate and to transfer among organismal lineages enables natural selection to proceed at several different levels, including that of the gene, the organism and, possibly, the community. Selective pressures operating at the gene and organism levels result in the diversification of gene lineages to track new functional niches and the capacity of well-organized genomes to accommodate new genes, while selection at the community level might contribute to the maintenance of MGEs that reassort supraspecific pools of accessory genes across organisms. Overall, these processes enable the astonishing diversity of the bacterial world.

References

[1] N. T. Perna, G. Plunkett, V. Burland et al., "Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7," *Nature*, vol. 409, no. 6819, pp. 529–533, 2001.

[2] V. Daubin and H. Ochman, "Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*," *Genome Research*, vol. 14, no. 6, pp. 1036–1042, 2004.

[3] H. Tettelin, V. Masignani, M. J. Cieslewicz et al., "Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial 'pan-genome'," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 39, pp. 13950–13955, 2005.

[4] M. L. Coleman, M. B. Sullivan, A. C. Martiny et al., "Genomic islands and the ecology and evolution of *Prochlorococcus*," *Science*, vol. 311, no. 5768, pp. 1768–1770, 2006.

[5] P. Normand, P. Lapierre, L. S. Tisa et al., "Genome characteristics of facultatively symbiotic *Frankia* sp. strains reflect host range and host plant biogeography," *Genome Research*, vol. 17, no. 1, pp. 7–15, 2007.

[6] M. W. J. Van Passel, P. R. Marri, and H. Ochman, "The emergence and fate of horizontally acquired genes in *Escherichia coli*," *PLoS Computational Biology*, vol. 4, no. 4, Article ID e1000059, 2008.

[7] A. Mira, A. B. Martín-Cuadrado, G. D'Auria, and F. Rodríguez-Valera, "The bacterial pan-genome: a new paradigm in microbiology," *International Microbiology*, vol. 13, no. 2, pp. 45–57, 2010.

[8] J. R. Thompson, S. Pacocha, C. Pharino et al., "Genotypic diversity within a natural coastal bacterioplankton population," *Science*, vol. 307, no. 5713, pp. 1311–1313, 2005.

[9] T. J. Treangen and E. P. C. Rocha, "Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes," *PLoS Genetics*, vol. 7, no. 1, Article ID e1001284, 2011.

[10] W. F. Doolittle and C. Sapienza, "Selfish genes, the phenotype paradigm and genome evolution," *Nature*, vol. 284, no. 5757, pp. 601–603, 1980.

[11] L. E. Orgel and F. H. C. Crick, "Selfish DNA: the ultimate parasite," *Nature*, vol. 284, no. 5757, pp. 604–607, 1980.

[12] M. P. Francino, "Gene survival in emergent genomes," in *Horizontal Gene Transfer in Microorganisms*, M. P. Francino, Ed., pp. 1–22, Caister Academic Press, Norfolk, UK, 2012.

[13] A. Mira, H. Ochman, and N. A. Moran, "Deletional bias and the evolution of bacterial genomes," *Trends in Genetics*, vol. 17, no. 10, pp. 589–596, 2001.

[14] L. Gómez-Valero, E. P. C. Rocha, A. Latorre, and F. J. Silva, "Reconstructing the ancestor of *Mycobacterium leprae*: the dynamics of gene loss and genome reduction," *Genome Research*, vol. 17, no. 8, pp. 1178–1185, 2007.

[15] C. H. Kuo and H. Ochman, "Deletional bias across the three domains of life," *Genome Biology and Evolution*, vol. 1, pp. 145–152, 2009.

[16] I. Comas, A. Moya, and F. González-Candelas, "Phylogenetic signal and functional categories in Proteobacteria genomes," *BMC Evolutionary Biology*, vol. 7, no. 1, supplement, article S7, 2007.

[17] A. Wellner, M. N. Lurie, and U. Gophna, "Complexity, connectivity, and duplicability as barriers to lateral gene transfer," *Genome Biology*, vol. 8, no. 8, article R156, 2007.

[18] J. G. Lawrence, "Selfish operons and speciation by gene transfer," *Trends in Microbiology*, vol. 5, no. 9, pp. 355–359, 1997.

[19] J. G. Lawrence and H. Hendrickson, "Lateral gene transfer: when will adolescence end?" *Molecular Microbiology*, vol. 50, no. 3, pp. 739–749, 2003.

[20] I. Comas and F. González-Candelas, "The evolution of horizontally transferred genes: a model for prokaryotes," in *Horizontal Gene Transfer in Microorganisms*, M. P. Francino, Ed., pp. 75–91, Caister Academic Press, Norfolk, UK, 2012.

[21] K. S. Makarova, Y. I. Wolf, and E. V. Koonin, "Comprehensive comparative-genomic analysis of Type 2 toxin-antitoxin systems and related mobile stress response systems in prokaryotes," *Biology Direct*, vol. 4, article 19, 2009.

[22] T. Naito, K. Kusano, and I. Kobayashi, "Selfish behavior of restriction-modification systems," *Science*, vol. 267, no. 5199, pp. 897–899, 1995.

[23] I. Kobayashi, "Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution," *Nucleic Acids Research*, vol. 29, no. 18, pp. 3742–3756, 2001.

[24] H. Mihara, T. Fujii, S. I. Kato, T. Kurihara, Y. Hata, and N. Esaki, "Structure of external aldimine of *Escherichia coli* CsdB, an IscS/NifS homolog: implications for its specificity toward selenocysteine," *Journal of Biochemistry*, vol. 131, no. 5, pp. 679–685, 2002.

[25] G. Michel, A. W. Roszak, V. Sauvé et al., "Structures of shikimate dehydrogenase AroE and its paralog YdiB: a common structural framework for different activities," *The Journal of Biological Chemistry*, vol. 278, no. 21, pp. 19463–19472, 2003.

[26] M. Blaise, H. D. Becker, J. Lapointe, C. Cambillau, R. Giegé, and D. Kern, "Glu-Q-tRNAAsp synthetase coded by the yadB gene, a new paralog of aminoacyl-tRNA synthetase that glutamylates tRNAAsp anticodon," *Biochimie*, vol. 87, no. 9-10, pp. 847–861, 2005.

[27] Y. Yin and J. F. Kirsch, "Identification of functional paralog shift mutations: conversion of *Escherichia coli* malate dehydrogenase to a lactate dehydrogenase," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 44, pp. 17353–17357, 2007.

[28] L. A. Nahum, S. Goswami, and M. H. Serres, "Protein families reflect the metabolic diversity of organisms and provide support for functional prediction," *Physiological Genomics*, vol. 38, no. 3, pp. 250–260, 2009.

[29] L. Mandrich and G. Manco, "Evolution in the amidohydrolase superfamily: substrate-assisted gain of function in the E183K mutant of a phosphotriesterase-like metal-carboxylesterase," *Biochemistry*, vol. 48, no. 24, pp. 5602–5612, 2009.

[30] A. Law and M. J. Boulanger, "Defining a structural and kinetic rationale for paralogous copies of phenylacetate-CoA ligases from the cystic fibrosis pathogen Burkholderia cenocepacia J2315," *The Journal of Biological Chemistry*, vol. 286, no. 17, pp. 15577–15585, 2011.

[31] D. McNally and M. A. Fares, "In silico identification of functional divergence between the multiple groEL gene paralogs in Chlamydiae," *BMC Evolutionary Biology*, vol. 7, article 81, 2007.

[32] G. Sanchez-Perez, A. Mira, G. Nyiro, L. Pašić, and F. Rodriguez-Valera, "Adapting to environmental changes using specialized paralogs," *Trends in Genetics*, vol. 24, no. 4, pp. 154–158, 2008.

[33] J. Li, Y. Wang, C. Y. Zhang et al., "*Myxococcus xanthus* viability depends on GroEL supplied by either of two genes, but the paralogs have different functions during heat shock, predation, and development," *Journal of Bacteriology*, vol. 192, no. 7, pp. 1875–1881, 2010.

[34] L. Van Valen, "Predation and species diversity," *Journal of Theoretical Biology*, vol. 44, no. 1, pp. 19–21, 1974.

[35] C. Gubry-Rangin, B. Hai, C. Quince et al., "Niche specialization of terrestrial archaeal ammonia oxidizers," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 52, pp. 21206–21211, 2011.

[36] N. Shapir, E. F. Mongodin, M. J. Sadowsky, S. C. Daugherty, K. E. Nelson, and L. P. Wackett, "Evolution of catabolic pathways: genomic insights into microbial s-triazine metabolism," *Journal of Bacteriology*, vol. 189, no. 3, pp. 674–682, 2007.

[37] M. Cai and L. Xun, "Organization and regulation of pentachlorophenol-degrading genes in *Sphingobium chlorophenolicum* ATCC 39723," *Journal of Bacteriology*, vol. 184, no. 17, pp. 4672–4680, 2002.

[38] M. H. Dai, J. B. Rogers, J. R. Warner, and S. D. Copley, "A previously unrecognized step in pentachlorophenol degradation in *Sphingobium chlorophenolicum* is catalyzed by tetrachlorobenzoquinone reductase (PcpD)," *Journal of Bacteriology*, vol. 185, no. 1, pp. 302–310, 2003.

[39] G. J. Poelarends, M. Wilkens, M. J. Larkin, J. D. Van Elsas, and D. B. Janssen, "Degradation of 1,3-dichloropropene by *Pseudomonas cichorii* 170," *Applied and Environmental Microbiology*, vol. 64, no. 8, pp. 2931–2936, 1998.

[40] K. S. Ju and R. E. Parales, "Application of nitroarene dioxygenases in the design of novel strains that degrade chloronitrobenzenes," *Microbial Biotechnology*, vol. 2, no. 2, pp. 241–252, 2009.

[41] M. Kivisaar, "Evolution of catabolic pathways and their regulatory systems in synthetic nitroaromatic compounds degrading bacteria," *Molecular Microbiology*, vol. 82, no. 2, pp. 265–268, 2011.

[42] D. H. Pieper and M. Seeger, "Bacterial metabolism of polychlorinated biphenyls," *Journal of Molecular Microbiology and Biotechnology*, vol. 15, no. 2-3, pp. 121–138, 2008.

[43] S. D. Copley, "Evolution of efficient pathways for degradation of anthropogenic chemicals," *Nature Chemical Biology*, vol. 5, no. 8, pp. 559–566, 2009.

[44] A. A. Medeiros, "Evolution and dissemination of β-lactamases accelerated by generations of β-lactam antibiotics," *Clinical Infectious Diseases*, vol. 24, no. 1, supplement, pp. S19–S45, 1997.

[45] M. Barlow, J. Caywood, S. Lai, J. Finley, and C. Swanlund, "Horizontal gene transfer and recombination in the evolution of antibiotic resistance genes," in *Horizontal Gene Transfer in Microorganisms*, M. P. Francino, Ed., pp. 165–176, Caister Academic Press, Norfolk, UK, 2012.

[46] B. G. Hall, "The EBG system of *E. coli*: origin and evolution of a novel β-galactosidase for the metabolism of lactose," *Genetica*, vol. 118, no. 2-3, pp. 143–156, 2003.

[47] A. Aharoni, L. Gaidukov, O. Khersonsky, S. M. Gould, C. Roodveldt, and D. S. Tawfik, "The 'evolvability' of promiscuous protein functions," *Nature Genetics*, vol. 37, no. 1, pp. 73–76, 2005.

[48] V. López-Canut, M. Roca, J. Bertrán, V. Moliner, and I. Tuñón, "Promiscuity in alkaline phosphatase superfamily. Unraveling evolution through molecular simulations," *Journal of the American Chemical Society*, vol. 133, no. 31, pp. 12050–12062, 2011.

[49] B. Frantz, T. Aldrich, and A. M. Chakrabarty, "Microbial degradation of synthetic recalcitrant compounds," *Biotechnology Advances*, vol. 5, no. 1, pp. 85–99, 1987.

[50] J. G. Lawrence, "Gene Transfer in Bacteria: speciation without species?" *Theoretical Population Biology*, vol. 61, no. 4, pp. 449–460, 2002.

[51] W. P. Hanage, C. Fraser, and B. G. Spratt, "Fuzzy species among recombinogenic bacteria," *BMC Biology*, vol. 3, article 6, 2005.

[52] A. C. Retchless and J. G. Lawrence, "Temporal fragmentation of speciation in bacteria," *Science*, vol. 317, no. 5841, pp. 1093–1096, 2007.

[53] P. Shen and H. V. Huang, "Homologous recombination in *Escherichia coli*: dependence on substrate length and homology," *Genetics*, vol. 112, no. 3, pp. 441–457, 1986.

[54] J. Majewski, P. Zawadzki, P. Pickerill, F. M. Cohan, and C. G. Dowson, "Barriers to genetic exchange between bacterial species: *Streptococcus pneumoniae* transformation," *Journal of Bacteriology*, vol. 182, no. 4, pp. 1016–1023, 2000.

[55] P. Meier and W. Wackernagel, "Impact of *mutS* inactivation on foreign DNA acquisition by natural transformation in *Pseudomonas stutzeri*," *Journal of Bacteriology*, vol. 187, no. 1, pp. 143–154, 2005.

[56] A. B. Reams and E. L. Neidle, "Selection for gene clustering by tandem duplication," *Annual Review of Microbiology*, vol. 58, pp. 119–142, 2004.

[57] A. B. Reams, E. Kofoid, M. Savageau, and J. R. Roth, "Duplication frequency in a population of *Salmonella enterica* rapidly approaches steady state with or without recombination," *Genetics*, vol. 184, no. 4, pp. 1077–1094, 2010.

[58] H. Hendrickson, E. S. Slechta, U. Bergthorsson, D. I. Andersson, and J. R. Roth, "Amplification-mutagenesis: evidence that "directed" adaptive mutation and general hypermutability result from growth with a selected gene amplification," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 4, pp. 2164–2169, 2002.

[59] F. A. Kondrashov, I. B. Rogozin, Y. I. Wolf, and E. V. Koonin, "Selection in the evolution of gene duplications," *Genome Biology*, vol. 3, no. 2, article RESEARCH0008, 2002.

[60] S. D. Hooper and O. G. Berg, "On the nature of gene innovation: duplication patterns in microbial genomes," *Molecular Biology and Evolution*, vol. 20, no. 6, pp. 945–954, 2003.

[61] M. P. Francino, "An adaptive radiation model for the origin of new gene functions," *Nature Genetics*, vol. 37, no. 6, pp. 573–577, 2005.

[62] D. I. Andersson and D. Hughes, "Gene amplification and adaptive evolution in bacteria," *Annual Review of Genetics*, vol. 43, pp. 167–195, 2009.

[63] D. Romero and R. Palacios, "Gene amplification and genomic plasticity in prokaryotes," *Annual Review of Genetics*, vol. 31, pp. 91–111, 1997.

[64] M. M. Riehle, A. F. Bennett, and A. D. Long, "Genetic architecture of thermal adaptation in *Escherichia coli*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 2, pp. 525–530, 2001.

[65] A. B. Reams and E. L. Neidle, "Genome plasticity in *Acinetobacter*: new degradative capabilities acquired by the spontaneous amplification of large chromosomal segments," *Molecular Microbiology*, vol. 47, no. 5, pp. 1291–1304, 2003.

[66] S. Zhong, A. Khodursky, D. E. Dykhuizen, and A. M. Dean, "Evolutionary genomics of ecological specialization," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 32, pp. 11719–11724, 2004.

[67] M. Gerstein, "A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure," *Journal of Molecular Biology*, vol. 274, no. 4, pp. 562–576, 1997.

[68] M. A. Huynen and E. Van Nimwegen, "The frequency distribution of gene family sizes in complete genomes," *Molecular Biology and Evolution*, vol. 15, no. 5, pp. 583–589, 1998.

[69] J. Qian, N. M. Luscombe, and M. Gerstein, "Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model," *Journal of Molecular Biology*, vol. 313, no. 4, pp. 673–681, 2001.

[70] P. M. Harrison and M. Gerstein, "Studying genomes through the aeons: protein families, pseudogenes and proteome evolution," *Journal of Molecular Biology*, vol. 318, no. 5, pp. 1155–1174, 2002.

[71] G. P. Karev, Y. I. Wolf, and E. V. Koonin, "Simple stochastic birth and death models of genome evolution: was there enough time for us to evolve?" *Bioinformatics*, vol. 19, no. 15, pp. 1889–1900, 2003.

[72] O. Lespinet, Y. I. Wolf, E. V. Koonin, and L. Aravind, "The role of lineage-specific gene family expansion in the evolution of eukaryotes," *Genome Research*, vol. 12, no. 7, pp. 1048–1059, 2002.

[73] M. H. Serres, A. R. W. Kerr, T. J. McCormack, and M. Riley, "Evolution by leaps: gene duplication in bacteria," *Biology Direct*, vol. 4, article 46, 2009.

[74] E. Alm, K. Huang, and A. Arkin, "The evolution of two-component systems in bacteria reveals different strategies for niche adaptation," *PLoS Computational Biology*, vol. 2, no. 11, article e143, pp. 1329–1342, 2006.

[75] B. S. Goldman, W. C. Nierman, D. Kaiser et al., "Evolution of sensory complexity recorded in a myxobacterial genome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 41, pp. 15200–15205, 2006.

[76] W. Viratyosin, L. A. Campbell, C. C. Kuo, and D. D. Rockey, "Intrastrain and interstrain genetic variation within a paralogous gene family in *Chlamydia pneumoniae*," *BMC Microbiology*, vol. 2, article 1, pp. 1–11, 2002.

[77] J. G. Lawrence, "Gene transfer, speciation, and the evolution of bacterial genomes," *Current Opinion in Microbiology*, vol. 2, no. 5, pp. 519–523, 1999.

[78] J. P. Gogarten, W. F. Doolittle, and J. G. Lawrence, "Prokaryotic evolution in light of gene transfer," *Molecular Biology and Evolution*, vol. 19, no. 12, pp. 2226–2238, 2002.

[79] L. D. Alcaraz, G. Olmedo, G. Bonilla et al., "The genome of *Bacillus coahuilensis* reveals adaptations essential for survival in the relic of an ancient marine environment," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 15, pp. 5803–5808, 2008.

[80] K. Penn, C. Jenkins, M. Nett et al., "Genomic islands link secondary metabolism to functional adaptation in marine Actinobacteria," *ISME Journal*, vol. 3, no. 10, pp. 1193–1203, 2009.

[81] J. Wiedenbeck and F. M. Cohan, "Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches," *FEMS Microbiology Reviews*, vol. 35, no. 5, pp. 957–976, 2011.

[82] X. Wu, S. Monchy, S. Taghavi, W. Zhu, J. Ramos, and D. van der Lelie, "Comparative genomics and functional analysis of niche-specific adaptation in *Pseudomonas putida*," *FEMS Microbiology Reviews*, vol. 35, no. 2, pp. 299–323, 2011.

[83] E. A. Groisman and H. Ochman, "Pathogenicity islands: bacterial evolution in quantum leaps," *Cell*, vol. 87, no. 5, pp. 791–794, 1996.

[84] H. Ochman and E. A. Groisman, "Distribution of pathogenicity islands in *Salmonella* spp.," *Infection and Immunity*, vol. 64, no. 12, pp. 5410–5412, 1996.

[85] E. F. Boyd and D. L. Hartl, "*Salmonella* virulence plasmid: modular acquisition of the *spv* virulence region by an F-plasmid in *Salmonella enterica* subspecies I and insertion into the chromosome of subspecies II, IIIa, IV and VII isolates," *Genetics*, vol. 149, no. 3, pp. 1183–1190, 1998.

[86] J. Hacker and J. B. Kaper, "Pathogenicity islands and the evolution of microbes," *Annual Review of Microbiology*, vol. 54, pp. 641–679, 2000.

[87] H. Ochman, J. G. Lawrence, and E. A. Grolsman, "Lateral gene transfer and the nature of bacterial innovation," *Nature*, vol. 405, no. 6784, pp. 299–304, 2000.

[88] H. Ochman and N. A. Moran, "Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis," *Science*, vol. 292, no. 5519, pp. 1096–1098, 2001.

[89] I. K. Toth, L. Pritchard, and P. R. J. Birch, "Comparative genomics reveals what makes an enterobacterial plant pathogen," *Annual Review of Phytopathology*, vol. 44, pp. 305–336, 2006.

[90] T. P. Stinear, T. Seemann, P. F. Harrison et al., "Insights from the complete genome sequence of *Mycobacterium marinum* on the evolution of *Mycobacterium tuberculosis*," *Genome Research*, vol. 18, no. 5, pp. 729–741, 2008.

[91] E. C. Berglund, A. C. Frank, A. Calteau et al., "Run-off replication of host-adaptability genes is associated with gene transfer agents in the genome of mouse-infecting *Bartonella grahamii*," *PLoS Genetics*, vol. 5, no. 7, Article ID e1000546, 2009.

[92] M. Marchetti, D. Capela, M. Glew et al., "Experimental evolution of a plant pathogen into a legume symbiont," *PLoS Biology*, vol. 8, no. 1, Article ID e1000280, 2010.

[93] M. W. Silby, C. Winstanley, S. A. Godfrey, S. B. Levy, and R. W. Jackson, "*Pseudomonas* genomes: diverse and adaptable," *FEMS Microbiology Reviews*, vol. 35, no. 4, pp. 652–680, 2011.

[94] M. A. Schmidt, "LEEways: tales of EPEC, ATEC and EHEC," *Cellular Microbiology*, vol. 12, no. 11, pp. 1544–1552, 2010.

[95] F. De la Cruz and J. Davies, "Horizontal gene transfer and the origin of species: lessons from bacteria," *Trends in Microbiology*, vol. 8, no. 3, pp. 128–133, 2000.

[96] F. M. Cohan, "Bacterial species and speciation," *Systematic Biology*, vol. 50, no. 4, pp. 513–524, 2001.

[97] L. Chistoserdova, J. A. Vorholt, R. K. Thauer, and M. E. Lidstrom, "C1 transfer enzymes and coenzymes linking methylotrophic bacteria and methanogenic archaea," *Science*, vol. 281, no. 5373, pp. 99–102, 1998.

[98] J. Xiong, K. Inoue, and C. E. Bauer, "Tracking molecular evolution of photosynthesis by characterization of a major photosynthesis gene cluster from *Heliobacillus mobilis*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 25, pp. 14851–14856, 1998.

[99] P. S. Novichkov, M. V. Omelchenko, M. S. Gelfand, A. A. Mironov, Y. I. Wolf, and E. V. Koonin, "Genome-wide molecular clock and horizontal gene transfer in bacterial evolution," *Journal of Bacteriology*, vol. 186, no. 19, pp. 6575–6585, 2004.

[100] W. Hao and G. Brian Golding, "The fate of laterally transferred genes: life in the fast lane to adaptation or death," *Genome Research*, vol. 16, no. 5, pp. 636–643, 2006.

[101] J. Haiko, L. Laakkonen, B. Westerlund-Wikström, and T. K. Korhonen, "Molecular adaptation of a plant-bacterium outer membrane protease towards plague virulence factor Pla," *BMC Evolutionary Biology*, vol. 11, no. 1, article 43, 2011.

[102] S. D. Hooper and O. G. Berg, "Duplication is more common among laterally transferred genes than among indigenous genes," *Genome Biology*, vol. 4, no. 8, article R48, 2003.

[103] I. Matic, C. Rayssiguier, and M. Radman, "Interspecies gene exchange in bacteria: the role of SOS and mismatch repair systems in evolution of species," *Cell*, vol. 80, no. 3, pp. 507–515, 1995.

[104] M. K. Ashby and J. Houmard, "Cyanobacterial two-component proteins: structure, diversity, distribution, and evolution," *Microbiology and Molecular Biology Reviews*, vol. 70, no. 2, pp. 472–509, 2006.

[105] M. Brigulla and W. Wackernagel, "Molecular aspects of gene transfer and foreign DNA acquisition in prokaryotes with regard to safety issues," *Applied Microbiology and Biotechnology*, vol. 86, no. 4, pp. 1027–1041, 2010.

[106] S. Yang, J. R. Arguello, X. Li et al., "Repetitive element-mediated recombination as a mechanism for new gene origination in *Drosophila*," *PLoS Genetics*, vol. 4, no. 1, article e3, pp. 0078–0087, 2008.

[107] V. Daubin, E. Lerat, and G. Perrière, "The source of laterally transferred genes in bacterial genomes," *Genome Biology*, vol. 4, no. 9, article R57, 2003.

[108] R. Jain, M. C. Rivera, and J. A. Lake, "Horizontal gene transfer among genomes: the complexity hypothesis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 7, pp. 3801–3806, 1999.

[109] C. Brochier, E. Bapteste, D. Moreira, and H. Philippe, "Eubacterial phylogeny based on translational apparatus proteins," *Trends in Genetics*, vol. 18, no. 1, pp. 1–5, 2002.

[110] E. W. Brown, J. E. LeClerc, B. Li, W. L. Payne, and T. A. Cebula, "Phylogenetic evidence for horizontal transfer of *mutS* alleles among naturally occurring *Escherichia coli* strains," *Journal of Bacteriology*, vol. 183, no. 5, pp. 1631–1644, 2001.

[111] Y. Nakamura, T. Itoh, H. Matsuda, and T. Gojobori, "Biased biological functions of horizontally-transferred genes in prokaryotic genomes," *Nature Genetics*, vol. 36, no. 7, pp. 760–766, 2004.

[112] F. D. Ciccarelli, T. Doerks, C. Von Mering, C. J. Creevey, B. Snel, and P. Bork, "Toward automatic reconstruction of a highly resolved tree of life," *Science*, vol. 311, no. 5765, pp. 1283–1287, 2006.

[113] R. Sorek, Y. Zhu, C. J. Creevey, M. P. Francino, P. Bork, and E. M. Rubin, "Genome-wide experimental determination of barriers to horizontal gene transfer," *Science*, vol. 318, no. 5855, pp. 1449–1452, 2007.

[114] S. S. Abby, E. Tannier, M. Gouy, and V. Daubin, "Lateral gene transfer as a support for the tree of life," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 13, pp. 4962–4967, 2012.

[115] R. L. Charlebois and W. F. Doolittle, "Computing prokaryotic gene ubiquity: rescuing the core from extinction," *Genome Research*, vol. 14, no. 12, pp. 2469–2477, 2004.

[116] S. R. Santos and H. Ochman, "Identification and phylogenetic sorting of bacterial lineages with universally conserved genes and proteins," *Environmental Microbiology*, vol. 6, no. 7, pp. 754–759, 2004.

[117] E. Lerat, V. Daubin, and N. A. Moran, "From gene trees to organismal phylogeny in prokaryotes: the case of the γ-Proteobacteria," *PLoS Biology*, vol. 1, no. 1, article E19, 2003.

[118] O. Lukjancenko, T. M. Wassenaar, and D. W. Ussery, "Comparison of 61 sequenced *Escherichia coli* genomes," *Microbial Ecology*, vol. 60, no. 4, pp. 708–720, 2010.

[119] A. Jacobsen, R. S. Hendriksen, F. M. Aaresturp, D. W. Ussery, and C. Friis, "The *Salmonella enterica* Pan-genome," *Microbial Ecology*, vol. 62, no. 3, pp. 487–504, 2011.

[120] M. Touchon, C. Hoede, O. Tenaillon et al., "Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths," *PLoS Genetics*, vol. 5, no. 1, Article ID e1000344, 2009.

[121] J. G. Lawrence and H. Ochman, "Molecular archaeology of the *Escherichia coli* genome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 16, pp. 9413–9417, 1998.

[122] E. Lerat, V. Daubin, H. Ochman, and N. A. Moran, "Evolutionary origins of genomic repertoires in bacteria," *PLoS Biology*, vol. 3, no. 5, article e130, 2005.

[123] C. H. Kuo and H. Ochman, "The fate of new bacterial genes," *FEMS Microbiology Reviews*, vol. 33, no. 1, pp. 38–43, 2009.

[124] F. M. Cohan, "Towards a conceptual and operational union of bacterial systematics, ecology, and evolution," *Philosophical Transactions of the Royal Society B*, vol. 361, no. 1475, pp. 1985–1996, 2006.

[125] E. Maisonneuve, L. J. Shakespeare, M. G. Jørgensen, and K. Gerdes, "Bacterial persistence by RNA endonucleases," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 32, pp. 13206–13211, 2011.

[126] O. Vesper, S. Amitai, M. Belitsky et al., "Selective translation of leaderless mRNAs by specialized ribosomes generated by MazF in *Escherichia coli*," *Cell*, vol. 147, no. 1, pp. 147–157, 2011.

[127] J. E. Rebollo, V. Francois, and J. M. Louarn, "Detection and possible role of two large nondivisible zones on the *Escherichia coli* chromosome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 85, no. 24, pp. 9391–9395, 1988.

[128] F. R. Blattner, G. Plunkett, C. A. Bloch et al., "The complete genome sequence of *Escherichia coli* K-12," *Science*, vol. 277, no. 5331, pp. 1453–1462, 1997.

[129] F. Boccard, E. Esnault, and M. Valens, "Spatial arrangement and macrodomain organization of bacterial chromosomes," *Molecular Microbiology*, vol. 57, no. 1, pp. 9–16, 2005.

[130] E. Esnault, M. Valens, O. Espéli, and F. Boccard, "Chromosome structuring limits genome plasticity in *Escherichia coli*," *PLoS genetics*, vol. 3, no. 12, article e226, 2007.

[131] P. Sobetzko, A. Travers, and G. Muskhelishvili, "Gene order and chromosome dynamics coordinate spatiotemporal gene expression during the bacterial growth cycle," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 2, pp. E42–E50, 2012.

[132] K. P. Williams, "Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies," *Nucleic Acids Research*, vol. 30, no. 4, pp. 866–875, 2002.

[133] T. Ikemura, "Codon usage and tRNA content in unicellular and multicellular organisms," *Molecular Biology and Evolution*, vol. 2, no. 1, pp. 13–34, 1985.

[134] P. M. Sharp and W. H. Li, "An evolutionary perspective on synonymous codon usage in unicellular organisms," *Journal of Molecular Evolution*, vol. 24, no. 1-2, pp. 28–38, 1986.

[135] M. Bulmer, "The selection-mutation-drift theory of synonymous codon usage," *Genetics*, vol. 129, no. 3, pp. 897–907, 1991.

[136] S. Kanaya, Y. Yamada, Y. Kudo, and T. Ikemura, "Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis," *Gene*, vol. 238, no. 1, pp. 143–155, 1999.

[137] E. P. C. Rocha, "Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization," *Genome Research*, vol. 14, no. 11, pp. 2279–2286, 2004.

[138] D. Amorós-Moya, S. Bedhomme, M. Hermann, and I. G. Bravo, "Evolution in regulatory regions rapidly compensates the cost of nonoptimal codon usage," *Molecular Biology and Evolution*, vol. 27, no. 9, pp. 2141–2151, 2010.

[139] A. Medrano-Soto, G. Moreno-Hagelsieb, P. Vinuesa, J. A. Christen, and J. Collado-Vides, "Successful lateral transfer requires codon usage compatibility between foreign genes and recipient genomes," *Molecular Biology and Evolution*, vol. 21, no. 10, pp. 1884–1894, 2004.

[140] J. G. Lawrence and H. Ochman, "Amelioration of bacterial genomes: rates of change and exchange," *Journal of Molecular Evolution*, vol. 44, no. 4, pp. 383–397, 1997.

[141] G. S. Vernikos, N. R. Thomson, and J. Parkhill, "Genetic flux over time in the *Salmonella* lineage," *Genome Biology*, vol. 8, no. 6, article R100, 2007.

[142] M. N. Price, P. S. Dehal, and A. P. Arkin, "Horizontal gene transfer and the evolution of transcriptional regulation in *Escherichia coli*," *Genome Biology*, vol. 9, no. 1, article R4, 2008.

[143] O. Gal-Mor, D. Elhadad, W. Deng, G. Rahav, and B. B. Finlay, "The *Salmonella enterica* PhoP directly activates the horizontally acquired SPI-2 gene *ssel* and is functionally different from a *S. bongori* ortholog," *PLoS ONE*, vol. 6, no. 5, Article ID e20024, 2011.

[144] L. C. Martínez, H. Yakhnin, M. I. Camacho et al., "Integration of a complex regulatory cascade involving the SirA/BarA and Csr global regulatory systems that controls expression of the *Salmonella* SPI-1 and SPI-2 virulence regulons through HilD," *Molecular Microbiology*, vol. 80, no. 6, pp. 1637–1656, 2011.

[145] M. J. Lercher and C. Pál, "Integration of horizontally transferred genes into regulatory interaction networks takes many million years," *Molecular Biology and Evolution*, vol. 25, no. 3, pp. 559–567, 2008.

[146] W. Davids and Z. Zhang, "The impact of horizontal gene transfer in shaping operons and protein interaction networks—direct evidence of preferential attachment," *BMC Evolutionary Biology*, vol. 8, no. 1, article 23, 2008.

[147] C. Pál, B. Papp, and M. J. Lercher, "Adaptive evolution of bacterial metabolic networks by horizontal gene transfer," *Nature Genetics*, vol. 37, no. 12, pp. 1372–1375, 2005.

[148] A. Kreimer, E. Borenstein, U. Gophna, and E. Ruppin, "The evolution of modularity in bacterial metabolic networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 19, pp. 6976–6981, 2008.

[149] M. D'Antonio and F. D. Ciccarelli, "Modification of gene duplicability during the evolution of protein interaction network," *PLoS Computational Biology*, vol. 7, no. 4, Article ID e1002029, 2011.

[150] A. Hintze and C. Adami, "Evolution of complex modular biological networks," *PLoS Computational Biology*, vol. 4, no. 2, article e23, 2008.

[151] C. Espinosa-Soto and A. Wagner, "Specialization can drive the evolution of modularity," *PLoS Computational Biology*, vol. 6, no. 3, article e1000719, 2010.

[152] Y. Bar-Yam, *Dynamics of Complex Systems*, Westview Press, Boulder, CO, USA, 1997.

[153] J. G. Lawrence, "Microbial evolution: enforcing cooperation by partial kin selection," *Current Biology*, vol. 19, no. 20, pp. R943–R945, 2009.

[154] T. Nogueira, D. J. Rankin, M. Touchon, F. Taddei, S. P. Brown, and E. P. C. Rocha, "Horizontal gene transfer of the secretome drives the evolution of bacterial cooperation and virulence," *Current Biology*, vol. 19, no. 20, pp. 1683–1691, 2009.

[155] S. E. Mc Ginty, D. J. Rankin, and S. P. Brown, "Horizontal gene transfer and the evolution of bacterial cooperation," *Evolution*, vol. 65, no. 1, pp. 21–32, 2011.

[156] D. J. Rankin, E. P. C. Rocha, and S. P. Brown, "What traits are carried on mobile genetic elements, and why," *Heredity*, vol. 106, no. 1, pp. 1–10, 2011.

[157] L. E. de Vries, Y. Vallès, Y. Agersø et al., "The gut as reservoir of antibiotic resistance: microbial diversity of tetracycline resistance in mother and infant," *PLoS ONE*, vol. 6, no. 6, Article ID e21644, 2011.

Hindawi Publishing Corporation
International Journal of Evolutionary Biology
Volume 2012, Article ID 292068, 12 pages
doi:10.1155/2012/292068

Review Article

Polyploidy and the Evolution of Complex Traits

Lukasz Huminiecki[1,2] and Gavin C. Conant[3,4]

[1] *CMB, Karolinska Institute, 17177 Stockholm, Sweden*
[2] *DBB, Stockholm University, 10691 Stockholm, Sweden*
[3] *MU Informatics Institute, University of Missouri, Columbia, MO 65211, USA*
[4] *Division of Animal Sciences, University of Missouri, Columbia, MO 65211-5300, USA*

Correspondence should be addressed to Lukasz Huminiecki, lukasz.huminiecki@scilifelab.se

Received 15 March 2012; Revised 29 May 2012; Accepted 5 June 2012

Academic Editor: Ben-Yang Liao

We explore how whole-genome duplications (WGDs) may have given rise to complex innovations in cellular networks, innovations that could not have evolved through sequential single-gene duplications. We focus on two classical WGD events, one in bakers' yeast and the other at the base of vertebrates (i.e., two rounds of whole-genome duplication: 2R-WGD). Two complex adaptations are discussed in detail: aerobic ethanol fermentation in yeast and the rewiring of the vertebrate developmental regulatory network through the 2R-WGD. These two examples, derived from diverged branches on the eukaryotic tree, boldly underline the evolutionary potential of WGD in facilitating major evolutionary transitions. We close by arguing that the evolutionary importance of WGD may require updating certain aspects of modern evolutionary theory, perhaps helping to synthesize a new evolutionary systems biology.

1. Introduction

Characteristic changes in karyotype number have allowed researchers to infer polyploidy events for many decades [1]. It was thus with a reasonably long history of research that Susumo Ohno was able to suggest that polyploidy was a vital route to evolutionary innovation [2]. Ohno was of course a forceful proponent of a general role for duplication in evolution: writing that "[if evolution occurred only through changes allele frequencies] ... from a bacterium only numerous forms of bacteria would have emerged [...B]ig leaps in evolution required the creation of new gene loci with previously nonexistent functions" [2]. What is less obvious on first reading is his distinction between the role played by WGD and that played by other, smaller scale, duplications (or SSDs). While the differences in the scales of these events are self-evident, there are at least two other features of WGD that are critical in giving rise to these differing roles. The first is that, as many authors have reported, particular functional classes of genes (e.g., transcription factors, kinases, ribosomal proteins, and cyclins) are duplicated by WGD more frequently than by SSD [3–8]. Ohno had in fact explored the most likely reason for this difference: "hub" genes with many interactions with other loci, be those interactions regulatory, protein interaction or metabolic, will tend to respond poorly to a change in copy number. As a result, they will tend to survive in duplicate after WGD but will not survive after smaller scale events [2, 5, 9–11]. This idea has now been termed the dosage balance hypothesis [12–14].

The second difference between single-gene and genome duplication is the kind of adaptations each may give rise to. Interest in gene duplication is intense in evolutionary biology circles because, as Haldane recognized [15], duplication is a powerful means for generating genetic material with the potential for innovation. There are many models of duplicate gene evolution [16]: probably the most discussed are neofunctionalization [1, 2, 16], whereby one copy of a duplicate gene pair acquires a new beneficial function *after* the duplication, and subfunctionalization, where multifunctioned genes have their functions subdivided by duplication [17–19]. Since some of these subfunctions might themselves be novel and suffer from antagonistic pleiotropy (e.g., one subfunction cannot be optimized without detrimentally altering the other; [17, 20]) subfunctionalization can

represent an important path to innovation. What genome duplication brings to this story is the potential for *multigene* novelties [21]: with a duplication of the entire genome to explore, evolution has more space to innovate. In this paper, we explore the evidence for multi-gene innovations in yeast and animals resulting from their respective WGDs [8, 22–25]. We then discuss in detail two key innovations that are associated with WGD: aerobic ethanol fermentation in yeast and increased complexity in the vertebrate developmental regulatory network. In so doing, we will remind ourselves of Francois Jacob's insight as to the mechanisms of evolution: the innovations produced are in keeping with the work of a tinkerer, not an engineer [26], and are contingent on their possessors' evolutionary history [27].

2. WGD and Single-Gene Innovations

The existence of neutral models of duplicate gene resolution [18, 19] and apparent examples of their action after WGD [28] means that, before pursuing multi-gene adaptations from WGD, it is worthwhile to pause and ask whether examples of single-gene innovations due to WGD are known. We do so even though those innovations may appear no different than what might be expected from an SSD event. As a matter of fact, there are good examples from yeast. For instance, consider the *S. cerevisiae* WGD-produced paralogs *GAL1* and *GAL3:* a sugar kinase and a regulator, respectively [29]. In the non-WGD *Kluyveromyces lactis,* the single ortholog of these two genes possesses both functions [30]. However, these two *ohnologs* [31] are not simply an example of neutral subfunctionalization: Hittinger and Carroll [20] have shown an adaptive conflict in the promoter of the *K. lactis* gene that was resolved by the gene duplication. In particular, it would be more "cost-effective" to have highly dynamic expression in the *K. lactis GAL1* gene, with strong repression in the absence of galactose. However, because this same locus also encodes the regulatory function performed by the Gal3 protein in *S. cerevisiae*, such strong repression would result in insufficient expression of *GAL1* to perform its regulatory function in the absence of galactose. Gene duplication allowed a decoupling of the expression levels of these two distinct functions. The WGD-produced duplication was thus exploited as the last step in the evolutionary development of a metabolic subsystem with a fine degree of transcriptional control.

3. Multigene Adaptations

The most unique potential impact of genes duplicated at WGD, however, is not in single-gene adaptations. Instead, it is the potential for correlated changes across multiple genes resulting in altered cellular networks, including signal transduction and transcriptional regulatory networks. That such changes occur is indirectly suggested by the observations that duplicates from the yeast WGD are more likely to be part of protein complexes and more likely to share protein interaction partners than SSD duplicates [10, 32]. The products of such retained duplicates are also enriched for proteins regulated by phosphorylation [33]. Both observations are in keeping with the expectations of the dosage balance hypothesis [12]. Similarly, we have shown an example of coherent changes in the coexpression networks of *S. cerevisiae*. To do so, we used an algorithm for detecting subdivided networks. This algorithm divides genes (connected by edges if they are coexpressed across multiple microarray experiments; [34]) into two columns, where each row consists of a pair of WGD-produced paralogs (Figure 1(a)). We then searched for the arrangement of genes that minimized the number of edges crossing between columns and compared that number to the number of such crossing edges seen in randomized networks. The relative paucity of crossing edges in the real network suggests *network* subfunctionalization, where groups of ohnologs are subdivided into two co-expression clusters [34].

3.1. WGD and the Crabtree Effect. While these global patterns of change after WGD suggest large-scale alterations, the best example of a change that can be at least provisionally tied to a phenotype is the evolution of the *Crabtree* effect. Baker's yeast is somewhat unusual in its metabolism: even when oxygen is available, it prefers to only partially oxidize glucose into ethanol rather than fully oxidize it into CO_2 and water (the Crabtree effect; [35, 36]). This fermentative lifestyle is odd insomuch as it is energetically less favorable than the complete conversion of sugars into carbon dioxide (e.g., respiration). However, there is a general association between whether or not a yeast species possesses the ancient WGD and the Crabtree effect [37].

One clue to the source of this apparent paradox can be found in a group of duplicated genes from the WGD, all involved in the early stages of glucose metabolism. These genes include two glucose sensors (*SNF3* and *RGT2*), two glucose transporters (*HXT6/HXT1*), and two duplicate enzymes that catalyze the initial step of glycolysis (e.g., the hexokinases *HXK1* and HXK2). Strikingly, in all three ohnolog pairs, one member acts when glucose concentrations are low and the other when they are high [38–40]. A second piece of the puzzle is due to theoretical work on resource competition among organisms inhabiting a large but ephemeral environmental resource. Such competition among cells can actually favor lineages that rapidly oxidize glucose relative to their more efficient but slower-growing competitors [41–43]. This *tragedy of the commons* [44] occurs because even though the efficient cells are able to convert more glucose into energy, they pay for this efficiency in reduced temporal growth rates, meaning that the fast, wasteful, cells can come to numerically dominate the resource patch.

Given these observations and expectations, we and others proposed that the yeast WGD had several effects on its patterns of glucose metabolism (Figure 2(a)). First, we proposed that the increase in gene copy number produced by the WGD gave rise (after some gene losses in other parts of the genome) to an increased flux through glycolysis [37, 47, 51, 52]. Second, because oxidative phosphorylation of pyruvate is constrained by oxygen concentrations and the spatial structure of the mitochondria, the WGD-possessing cells

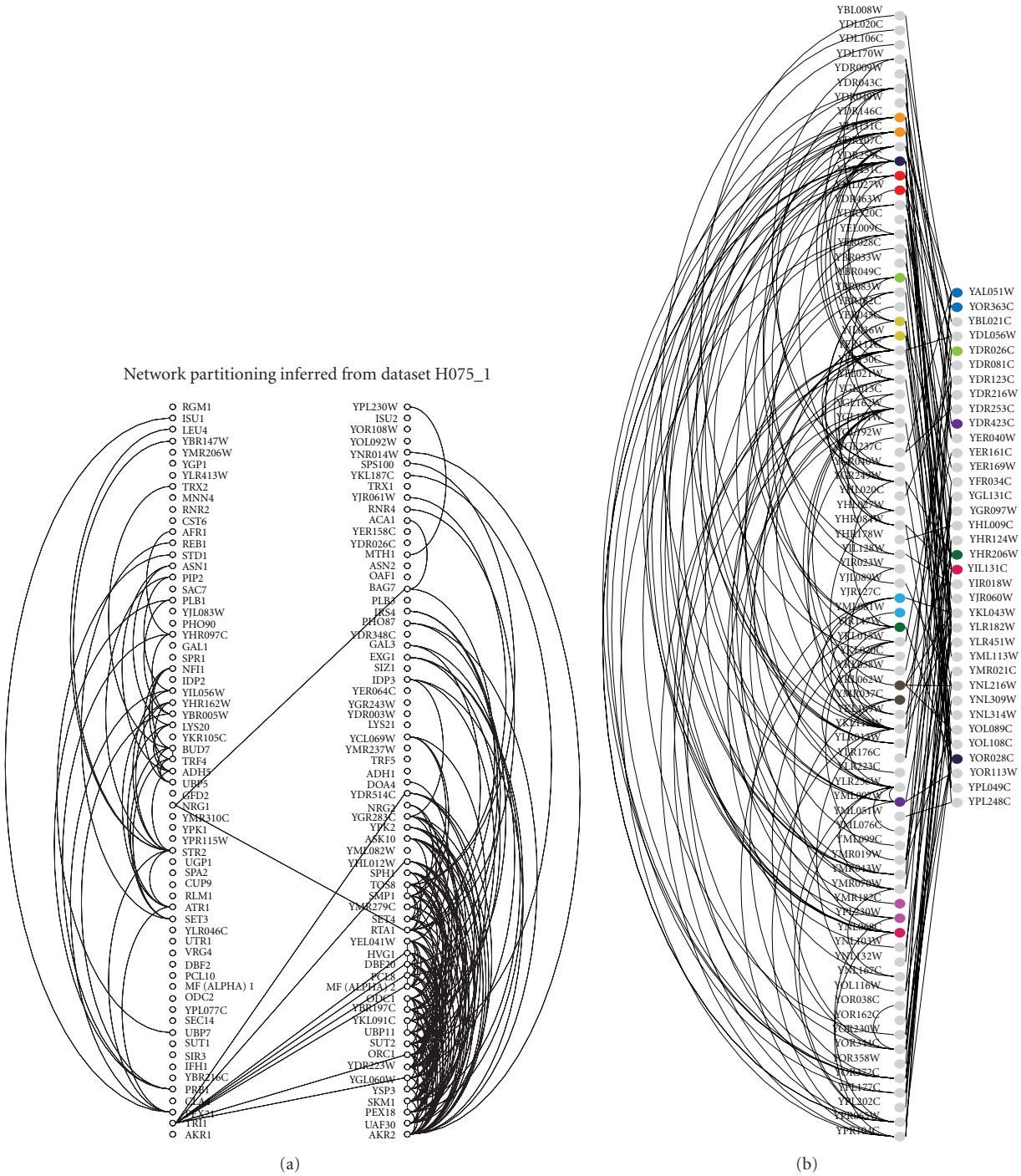


Figure 1: Network evolution after the yeast WGD. (a) The yeast coexpression networks show evidence of subfunctionalization after WGD. A co-expression network consisting of 65 pairs of WGD-produced paralogs (e.g., ohnologs) is illustrated. Each row contains a pair of ohnologs; edges join genes with co-expression correlation (Pearson's r) ≥ 0.75 across >200 microarray experiments. In each row, the position of the two ohnologs can be exchanged: we searched for the arrangement that minimized the number of interactions between the two columns (central diagonal edges). The number of such "crossing edges" is much smaller than what would be expected by chance (see [34]). (b) The above patterns are at least partly driven by changes in transcriptional regulation. We have previously shown that WGD-derived duplicated transcription factors have diverged considerably since WGD [45]. Here we show the relative lack of overlap between these duplicated regulators' functions. On the right are the transcription factors (TFs) that target other TFs but are not themselves targeted by a TF. On the left are the TFs that are regulated by other TFs. Duplicated TF pairs from WGD (e.g., ohnologs) are shown in the same color.

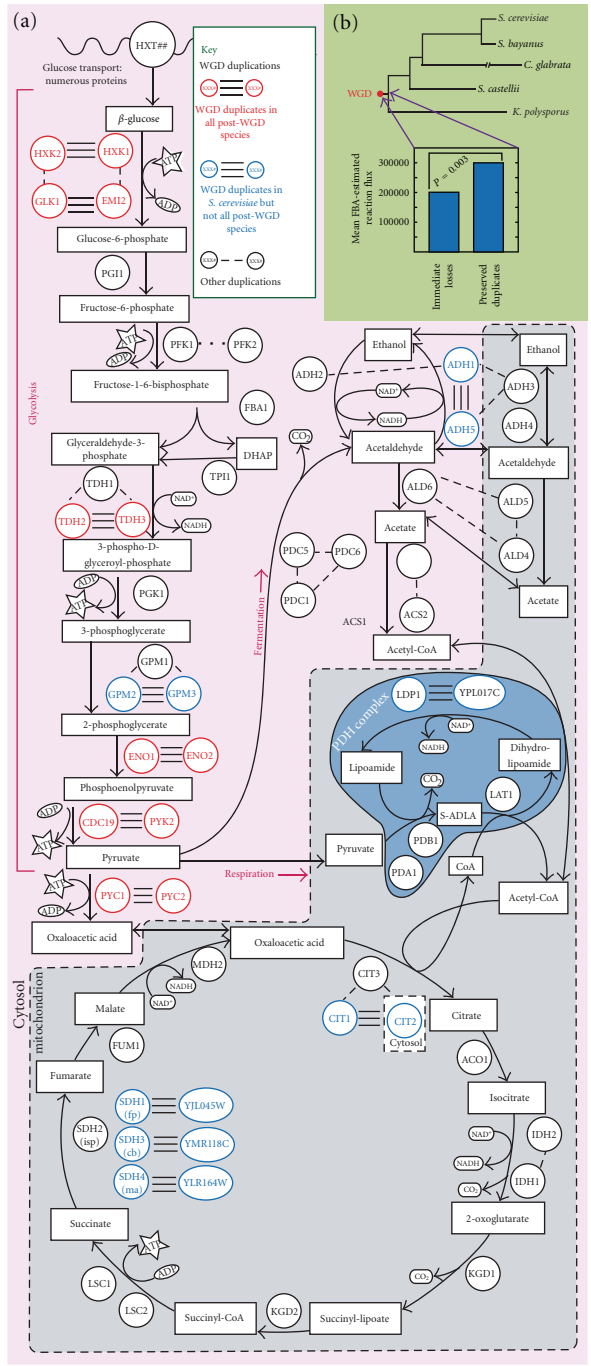


Figure 2: WGD and yeast carbon metabolism. (a) Illustrated are glycolysis, alcohol fermentation, and the mitochondrial TCA cycle. We denote the enzymes catalyzing a reaction with circled gene names. Products of SSD events are indicated by single lines joining the pair of enzymes. Enzymes duplicated at WGD are joined by three lines. WGD pairs in red are preserved in duplicate in four extant yeasts: *S. cerevisiae*, *S. bayanus*, *C. glabrata*, and *S. castellii*. Protein localization for *CIT*, *ADH*, and *ALD* is taken from Huh et al. [46]. We indicate the pyruvate dehydrogenase (PDH) multienzyme complex with a darker blue enclosure. There is a clear bias in where the duplicated enzymes lie, particularly if only those preserved in duplicate across four species are considered. From [47]. (b) Duplicated enzymes losses immediately after WGD were biased toward enzymes catalyzing low-flux reactions. The fluxes through all reactions in the yeast metabolic network [48] were computed under a variety of nutrient conditions as previously described [49]. Then, using our tool for estimating the timing of gene loss after WGD [50], we identified enzymes likely to have been lost along the short branch separating the WGD from the divergence of *K. polysporus* from the remaining four yeast species. We compared the fluxes of those enzymes to that of enzymes retained in duplicate along that same branch. The genes lost immediately after WGD were more likely to code for enzymes of low flux ($P = 0.003$, permutation analysis; unpublished data).

were required to redirect some of this increased glycolytic flux to the (previously anaerobic) fermentative pathways [47]. The result was likely to induce the sort of competitive situation between efficient and inefficient cells just described. A degree of independent confirmation to these ideas was provided by Van Hoek and Hogeweg [53], who were able to show computationally that similar WGD events modeled in modern *S. cerevisiae* could also be expected to result in over retention of glycolytic enzymes and increased glycolytic flux. Recent work in our lab also supports this contention, showing that duplicate losses immediately after WGD were biased toward genes coding for low-flux enzymes (Figure 2(b), unpublished data).

If the WGD was in fact a trigger for moving *S. cerevisiae* and its relatives down a path toward increasing Crabtree effect, we would expect it to have been followed by later evolutionary changes reinforcing this propensity. Indeed, at least two such post-WGD changes are known. First, in yeasts with the WGD, loss of *cis*-regulatory elements among the genes for the *mitochondrial* ribosomal proteins has decoupled the expression of the cytosolic and mitochondrial ribosomal proteins [54]. This change had an important effect: *S. cerevisiae* can now upregulate production of cytosolic ribosomes independently of the mitochondrial ones, an outcome that increases the efficiency of aerobic fermentation by avoiding unnecessary ribosome synthesis in the quiescent mitochondria. The second example is a post-WGD *SSD* event in the alcohol dehydrogenase family. The result of this event was two specialized ADH loci, one for the synthesis of ethanol and a second isoenzyme responsible for the back-conversion of ethanol to pyruvate (once glucose is exhausted Crabtree yeasts can reimport and respire the ethanol they previously produced; [55]). Such specialization likely would only have been beneficial in the context of a preexisting WGD-produced Crabtree adaptation.

3.2. Other Examples of Coordinated Evolution in Post-WGD Yeasts. There are at least two other cellular subsystems in *S. cerevisiae* that show evidence of large-scale changes after WGD, although the details are less well understood than is the case for metabolism. First, in the transcriptional regulatory network, pairs of transcription factors duplicated at WGD, while still showing detectable similarities in their targets inherited from the WGD, have diverged considerably (Figure 1(b); [45, 56]). More interestingly, the cytosolic ribosomal proteins in *S. cerevisiae* were highly over-retained after-WGD [6], representing roughly 10% of all retained duplicates, despite being less than 4% of the pre-WGD genome [57, 58]. These duplicates are extremely curious in that many of them have undergone considerable gene conversion, such that, despite their divergence at the ancient WGD, they have virtually identical protein sequences in modern bakers' yeast [23, 58]. At first blush, this result could be explained in terms of selection for high copy number [59] and the dosage balance hypothesis. The story became mysterious, however, with the discovery that several of these paralogs, while nearly identical in protein sequence, have distinctly different knockout phenotypes [60–62]. In keeping with the idea of coordinated evolution among multiple paralogs, a number of these duplicated pairs show asymmetric specialization of one of the two ohnologs to expression in the developing bud of the yeast cell [61, 62]. We speculate that these ribosomal proteins will represent another example of a system-level specialization induced by the WGD. In this view, the rampant gene conversion is a result of the highly interactive nature of the ribosome. Thus, both paralogs must "fit" exactly into the complex ribosomal structure and what differs is not their protein function but their expression domain.

3.3. WGD and Evolutionary Innovations in Plants. WGD is rampant in plant genomes, particularly those of angiosperms [63, 64]. The systems and network biology of these events have recently been extensively reviewed [65–68], and we will not attempt to do justice to the subject here. However, we do note that while the complexity of plant biology makes identifying precise evolutionary trajectories quite difficult, there are several suggestive coincidences of timing between the origins of new traits and the duplication of regulatory genes involved in those traits [66]. For example, glucosinolates are a class of secondary metabolites, the diversity of which has become expanded in the model plant *Arabidopsis thaliana* and its relatives. If one maps this expansion onto the phylogeny of these plants, it is curiously close to one of the *Arabidopsis* WGD events. Even more strikingly, several of the regulators and enzymes responsible for glucosinolate production in *Arabidopsis* have surviving duplications from that WGD [69]. More generally, we have recently shown [70] that the pattern of post-WGD duplicate retention in the *Arabidopsis* metabolic network seems to be driven by two different forces: a tendency to initially retain clusters of related enzymes (as would be expected under the dosage balance hypothesis) followed by a selective regime that appears to retain duplicates for reactions of high flux (similar to situation seen in *S. cerevisiae*).

3.4. 2R and the Remodeling of the Vertebrate Developmental and Signal Transduction Networks. Another example of WGD-induced functional innovation at the systems level concerns the vertebrate developmental toolkit and signal transduction engines. The metazoans, because they have bodies organized into distinct tissues, are clearly characterized by significant phenotypic complexity. They seem to have appeared about 640 million years ago and may have been preceded by other multicellular lineages of uncertain relationships [71]. On the basis of mitochondrial DNA sequence comparisons, the choanoflagellates have been identified as the closest single-celled animal relatives [72, 73] with the basal metazoan being either the placozoans [74–76] or the sponges [77, 78]. Although the role of WGD in metazoan evolution is not fully understood, several examples of WGDs among the vertebrates have been identified [21]. These include two rounds (2R) of genome duplication at the base of vertebrates (2R-WGD; [25]), the fish-specific genome duplication (FSGD; [4, 79, 80]), and WGDs in the genus *Xenopus* [81].

Despite their phenotypic complexity, animals' gene content is not vastly greater than that of other organisms [82, 83]. Part of the explanation for this relative paucity of extra genes is the nature of development, which occurs by sequential differentiation in bifurcating cell lineages rather than through entirely distinct differentiation programs for each tissue. Nonetheless, the transformation to multicellularity must have been accompanied by appearance of new genes coding for adhesion molecules, extracellular matrix proteins (such as collagen), and cell-to-cell communication. Indeed, considerable progress has been made in identifying the novel signaling pathways involved in control of development and body plan formation [84]. In keeping with the theme of relatively little genome expansion coupled to the appearance of the metazoans, only a small fraction of the genes in the genome contribute to the development of the body plan. However, these genes make up a developmental toolkit that is strongly conserved across the eumetazoans. Transcriptional factors of particular interest are homeobox genes (Hox, ParaHox, EHGbox, and NK-like); KLF, Osr and Sp1/Egr genes, tlx, Snail, and slug zinc-finger proteins; MASH, myoD, mef, hairy, and twist helix-loop-helix transcriptional factors; T-box transcriptional factors [85–87]. These transcriptional regulators interact with the outside world through signal transduction pathways, the most important of which are those employing transforming growth factor-β (TGF-β), Wnt, Notch, Hedgehog, Toll, tyrosine kinase receptors, the nuclear hormone receptors, and the G-protein-coupled receptors. The identification of the shared toolkit of signalling pathways underlying animal development is a key discovery of modern biology. Following this work, we have recently found that the vertebrate signal transduction engine was highly modified by the 2R-WGD [88], suggesting that some of the complexity of vertebrates may have required the innovative capacity of WGD [89].

3.5. Gene Duplications in the Transforming Growth Factor-β Pathway. Our initial study, focused on the TGF-β pathway, provided early evidence of the impact of the 2R-WGD on vertebrate signaling. This signaling pathway has been long recognized as one of the most fundamental and versatile in metazoans, with central roles in development, organogenesis, stem-cell control, immunity, and cancer [90]. After an investigation of 33 genomes, we showed that the evolution of the TGF-β pathway in animals can be best explained according to the 2R model, with additional duplications in teleost fishes [91]. The components of the core pathway (both receptors and Smads) expanded dramatically and permanently at the base of vertebrates as a result of the 2R-WGD. In particular, four ancestral Smads (an I-Smad, a Co-Smad, and two R-Smads of the BMP and TGF-β *sensu stricto* channels) gave rise to the eight known Smads of the human genome, classified as two TGF-β *sensu stricto* (Smad2,3) and three bone-morphogenetic-protein- (BMP-) type (Smad1,5,8) receptor-activated Smads (R-Smads), one common mediator Smad (Co-Smad; Smad4), and two inhibitory Smads (I-Smads; Smad6,7).

3.6. General Expansion of Signaling Pathways after 2R. In a more general analysis, we found that the 2R-WGD affected the overwhelming majority (three quarters) of human signaling genes, with the strongest effect on developmental pathways involving receptor tyrosine kinases, Wnt and TGF-β ligands, GPCRs, and the apoptosis pathway. Unlike genes deriving from recent tandem duplications, genes retained after 2R were enriched in protein interaction domains and multifunctional signaling modules of Ras and MAP-kinase cascades. The set of human 2R-ohnologs (2ROs), corresponding to 9,958 unique Entrez Genes, is enriched in many classic signaling domains (such as tyrosine and serine/threonine kinase domains, the seven-transmembrane receptor domains of the rhodopsin and secretin families, and the Ras family domain), as well as well-known protein interaction domains, including the SH2, SH3, PTB, and PDZ domains.

PDZ domains are particularly interesting as they are abundant in vertebrate neuronal synapses, serving as scaffolds for the assembly of large neurotransmission signaling complexes [92]. Thus, these results suggest that 2R may have provided evolutionary material for subsequent changes in vertebrate brain development. Further evidence for this contention came when we found that 2ROs are preferentially expressed in Gene Expression Atlas samples associated with brain and nervous tissue. These brain-expressed 2ROs are also enriched in Gene Ontology (GO) terms related to synaptic transmission. Studies in fly and mouse have shown that vertebrate synapses are more complex than those of invertebrates [93]: it is thus intriguing to speculate as to a role for 2R in inducing this phenomenon.

Another potential source of vertebrate neuronal complexity is their use of apoptosis to shape brain structures and compartments. We found that the apoptosis pathway was dramatically remodeled through 2R [88]. Figure 3 illustrates the complex topology of the human apoptosis signaling subnetwork created by 2R [88]. It is clear that coordinated duplications of caspases resulted in a substantial evolutionary novelty. Moreover, the complexity of the evolutionary changes introduced by 2R is best appreciated by examining the conservation of regulatory interactions (directed edges in the network). To better illustrate changes in network topology induced by the 2R-WGD, we subdivided the conserved edges into those originating from a shared regulator and acting on a pair of 2ROs (conserved incoming edges—CIEs), and those originating from a paralogous pair directed towards a shared target (conserved outgoing edges—COEs). CIEs suggest a common conserved regulator, located upstream in terms of information flow. In contrast, COEs indicate evolutionary conservation of a common regulatory target, located downstream (Figure 3).

Finally, while many genes for ancient cellular functions were not retained in duplicate after 2R, the genes of the cell cycle are an exception to this rule (an interesting link to the overretention of cyclins after the yeast WGD; [8]). Most cyclins, including key cell cycle-regulating groups A, B, and D, underwent diversification at the base of vertebrates and are represented by between two and four vertebrate-specific paralogs derived from the 2R-WGD [88].

"From" node	"To" node	Type of bridge	Conserved regulatory edges
BCL2L1	BCL2	Positive (and negative in direction)	2 positive CIEs (BAD and BAP31); 7 negative CIEs (CASP3, RAD9, Bmf, BNIP3, BNIP3L, Hrk and Puma); 8 negative COEs (BAD, BAK, BAX, BID, BIK, BIM, CYTOCHROME C, Noxa); 1 scrambled negative edge pair (p53 inhibits BCL2, and is itself inhibited by BCL2L1).
CASP3	CASP7	Positive	2 positive COEs (CAD and ICAD); 1 scrambled positive edge pair (CASP9); 4 negative CIEs (cIAP1, cIAP2, NAIP, Livin); 3 negative COEs (PARP, MEF2B, PROKR1); 1 scrambled negative edge pair (XIAP).
CASP8	CASP10	Negative	2 positive COEs (CASP3 and IAP); 1 positive CIE pair (FADD).

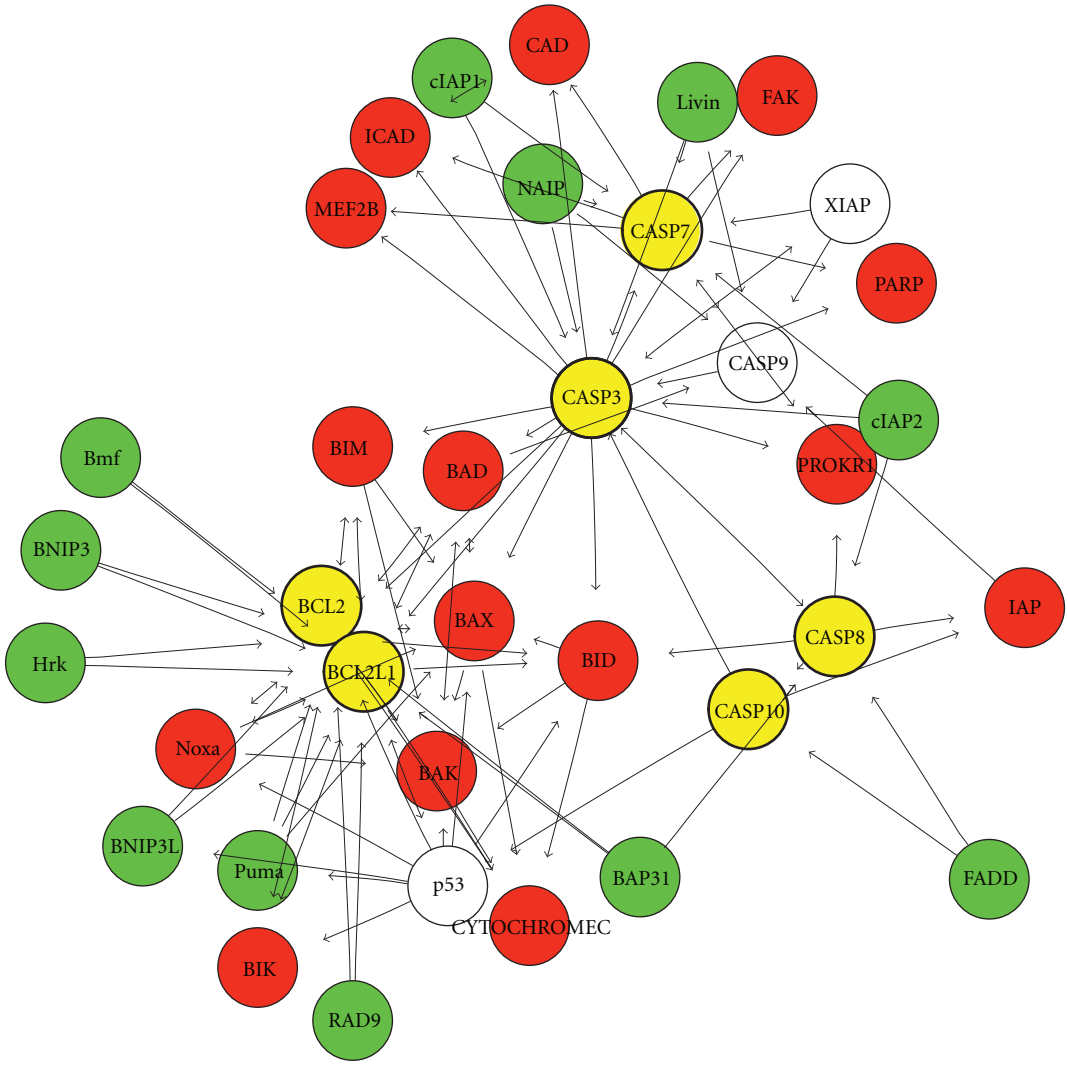


FIGURE 3: 2R-ohnologs in the apoptosis pathway. A network diagram of the vertebrate apoptosis pathway is shown with pairs of 2R-ohnologs (2ROs) highlighted in yellow. There are three 2RO pairs in the subnetwork: BCL2L1 and BCL2; CASP3 and CASP7; CASP8 and CASP10. Nodes are color coded: yellow signifies nodes mapping to 2ROs, while green and red signify those mapping to CIEs and COEs, respectively (see text). CASP8 and CASP10 are initiator caspases. CASP3 and CASP7 are executioner-caspases functioning downstream of these initiator caspases. IAPs are apoptosis inhibitors. The balance between antiapoptotic BCL2 and BCL2L1, and proapoptotic BAD, BAK, BAX, BID, BIM, and Puma, and Noxa determines the final activity of the intrinsic pathway of apoptosis. From [88].

Similarly, cyclin-dependent kinases, cyclin-dependent kinase inhibitors, and orthologs of the *S. pombe* WEE1 inactivator of the CDK/cyclin complex (Wee1 and Wee2) were also retained in duplicate after 2R [88]. Strikingly, cyclins D1-D3 respond to extracellular mitogens, cytokines, hormones, and juxtacrine ligands, providing an interface between signal transduction and the cell cycle. These cyclins then pair with CDKs 4 and 6, driving the transition to G1 [94–96]. It would be very informative to test if cyclins D1-D3 and CDK4/6 simply increase robustness of the cell cycle. If not, there may be functional differences between the 2R-derived cyclin D/CDK complexes in terms of the upstream signaling pathways they integrate or the downstream target genes they activate [88].

3.7. 2R and Vertebrate Complexity. In contrast to the predictions of the dosage balance hypothesis, vertebrate genes having developmental expression were more likely to revert to single copy after whole-genome duplication [97, 98]. However, this observation may be qualified by the fact that, after the FSGD, almost all retained duplicates have diverged in spatial and/or temporal expression during embryogenesis, and many were key developmental genes that function as transcription factors or signaling molecules during embryogenesis [99]. These general trends of retention and expression change, as well as the above functional analyses, clearly indicated that 2R fundamentally altered vertebrates' signaling pathways and cell cycles [88]. In consequence, it may have set the stage for the emergence of other key vertebrate evolutionary novelties (such as complex brains, the circulatory system, or heart, bone, cartilage, musculature, and adipose tissues; [71, 100]).

It should also be noted that the methodology used in these studies of the 2R-WGD [88] precluded an investigation of the amphioxus genome, as this genome was not included in release 6 of the TreeFam database. However, in other studies, the genome of the cephalochordate *Branchiostoma floridae* (e.g., amphioxus or lancelet) provided very strong evidence in support of the 2R hypothesis [101, 102]. Another strategically positioned pre-2R genome, that of sea urchin, is being developed as a developmental and systems biology model for understanding gene regulatory network evolution, which, together with the signal transduction pathways of this species, has been particularly well annotated [103–105]. Comparisons of sea urchin's developmental regulatory networks with those of vertebrates is likely to reveal further insights into the impact of the 2R-WGD.

4. Concluding Thoughts

The broader significance of these changes for our understanding of the forces and mechanisms driving the evolutionary process could well be extremely significant. Firstly, we propose that WGDs, like human technical innovations such as the railroad, greatly expand of genotypic and phenotypic space that might be explored by evolution. For example, the 2R quadrupling of components of the vertebrate signaling network not only immediately expanded the available space of signaling network states, but also kick-started rapid co-evolution of nodes into novel topologies during the subsequent "diploidization." We have also recently proposed that WGD has an important role in evolutionary transitions by relaxing epistatic constraints [70], effectively increasing the size of the neutral genetic space in which innovation can occur [106]. Secondly, an exciting possibility exists that at least some WGDs may be instantaneous speciations: if so, they would be evolutionary events whose occurrence is somewhat in contrast to an exclusively gradualist view of evolution. Early authors of modern synthesis, coming from background in population genetics, were perhaps overly wedded to gradualism, where natural selection acts on small variations in large populations. The molecular mechanism of WGDs is most likely auto- or allopolyploidy. WGDs could therefore be interpreted as saltations, that is, sudden evolutionary changes occurring within a single generation. However, population genetic processes are of course of central importance during subsequent re-diploidization. During that gradual process of duplicate loss over millions of years, there may be losses driven by natural selection acting to fix null mutants for duplicated loci, a process which fits well with Neo-Darwinian views.

In a related vein, gene duplications may have a role in enhancing robustness—the organism's resilience to genetic or environmental perturbations [107]. At the simplest level, duplication provides short-term robustness through genetic "backups." However, WGDs could also lead to an increase in distributed robustness, which is a consequence of the existence of multiple solutions to the same biological problem. A well-known example of this idea is the redundant paths through metabolic networks that confer robustness [48, 108]. It is fairly straightforward to envisage an analogous situation in signal transduction or the cell cycle: multiple regulatory mechanisms could in that case increase the level of control, allowing, for instance, the development of complex vertebrate embryos with many novel organs and tissue types.

Genome duplication might have also facilitated innovation in other ways. For instance, the establishment of crosstalk between signaling pathways [109] may have resulted from WGD. The post-WGD redundancy would have allowed the partial subdivision of duplicated pathways, resulting in a network of a higher degree of connectivity and robustness. Thus, it is striking that few novel signaling genes emerged through post-2R events [88], since SSD events lack the opportunity for this type of change. Another area for future investigation is the impact of 2R-WGD on non-coding genes [110]. Published studies and our own observations indicate that no preferential retention of miRNA genes can be attributed to 2R-WGD [111]. Instead, functional innovation in miRNA regulation appears to have occurred during the more recent mammalian diversification [112]. This suggests a model where major evolutionary transitions exploit expansions in different classes of genomics elements: protein-coding genes at the transition to vertebrates and miRNA genes during diversification of mammals.

It is tempting to hypothesize that gene duplication can initially promote redundancy of system parts, allowing evolutionary tinkering, while genome duplications are

correlated with an increase in distributed robustness. In the future, we propose testing this hypothesis by asking if more WGD-produced duplications are found in distinct signaling pathways when compared to SSD gene duplicates of similar age. More generally, both redundancy and robustness provide the evolutionary space for adaptations, and there are suggestions that WGD facilitated the colonization of novel environments and ecological niches [52, 113].

Genome duplication undoubtedly represents a tremendous evolutionary opportunity: the release of epistasis alone that results from WGD may have important implications [45]. However, as the examples described here suggest, the resulting innovations are unlikely to fit neatly into the neofunctionalization/subfunctionalization paradigm [16, 114], nor are they likely to be fully understood without a detailed knowledge of the cellular systems in which they are active.

Acknowledgments

The authors would like to thank Michaël Bekaert, Patrick Edger, Carl-Henrik Heldin, Corey Hudson, and Chris Pires for helpful discussions. L. Huminiecki is supported by ENFIN, a Network of Excellence funded by the European Commission FP6 Programme, under the thematic area "Life sciences, genomics and biotechnology for health," Contract no. LSHG-CT-2005-518254, the Swedish Research Council, and Swedish Bioinformatics Infrastructure for Life Sciences (BILS). G. C. Conant is supported by the Reproductive Biology Group of the Food for the 21st Century program at the University of Missouri.

References

[1] J. S. Taylor and J. Raes, "Duplication and divergence: the evolution of new genes and old ideas," *Annual Review of Genetics*, vol. 38, pp. 615–643, 2004.

[2] S. Ohno, *Evolution by Gene Duplication*, Springer, New York, NY, USA, 1970.

[3] J. M. Aury, O. Jaillon, L. Duret et al., "Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*," *Nature*, vol. 444, no. 7116, pp. 171–178, 2006.

[4] O. Jatllon, J. M. Aury, F. Brunet et al., "Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype," *Nature*, vol. 431, no. 7011, pp. 946–957, 2004.

[5] S. Maere, S. De Bodt, J. Raes et al., "Modeling gene and genome duplications in eukaryotes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 15, pp. 5454–5459, 2005.

[6] C. Seoighe and K. H. Wolfe, "Yeast genome evolution in the post-genome era," *Current Opinion in Microbiology*, vol. 2, no. 5, pp. 548–554, 1999.

[7] B. C. Thomas, B. Pedersen, and M. Freeling, "Following tetraploidy in an *Arabidopsis* ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes," *Genome Research*, vol. 16, no. 7, pp. 934–946, 2006.

[8] K. H. Wolfe and D. C. Shields, "Molecular evidence for an ancient duplication of the entire yeast genome," *Nature*, vol. 387, no. 6634, pp. 708–713, 1997.

[9] I. Wapinski, A. Pfeffer, N. Friedman, and A. Regev, "Natural history and evolutionary principles of gene duplication in fungi," *Nature*, vol. 449, no. 7158, pp. 54–61, 2007.

[10] L. Hakes, J. W. Pinney, S. C. Lovell, S. G. Oliver, and D. L. Robertson, "All duplicates are not equal: the difference between small-scale and genome duplication," *Genome Biology*, vol. 8, no. 10, article R209, 2007.

[11] M. Freeling, "Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition," *Annual Review of Plant Biology*, vol. 60, pp. 433–453, 2009.

[12] J. A. Birchler and R. A. Veitia, "The gene balance hypothesis: from classical genetics to modern genomics," *Plant Cell*, vol. 19, no. 2, pp. 395–402, 2007.

[13] M. Freeling and B. C. Thomas, "Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity," *Genome Research*, vol. 16, no. 7, pp. 805–814, 2006.

[14] B. Papp, C. Pál, and L. D. Hurst, "Dosage sensitivity and the evolution of gene families in yeast," *Nature*, vol. 424, no. 6945, pp. 194–197, 2003.

[15] J. B. S. Haldane, "The part played by recurrent mutation in evolution," *American Naturalist*, vol. 67, pp. 5–9, 1933.

[16] H. Innan and F. Kondrashov, "The evolution of gene duplications: classifying and distinguishing between models," *Nature Reviews Genetics*, vol. 11, no. 2, pp. 97–108, 2010.

[17] D. L. Des Marais and M. D. Rausher, "Escape from adaptive conflict after duplication in an anthocyanin pathway gene," *Nature*, vol. 454, no. 7205, pp. 762–765, 2008.

[18] A. Force, M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, and J. Postlethwait, "Preservation of duplicate genes by complementary, degenerative mutations," *Genetics*, vol. 151, no. 4, pp. 1531–1545, 1999.

[19] A. Stoltzfus, "On the possibility of constructive neutral evolution," *Journal of Molecular Evolution*, vol. 49, no. 2, pp. 169–181, 1999.

[20] C. T. Hittinger and S. B. Carroll, "Gene duplication and the adaptive evolution of a classic genetic switch," *Nature*, vol. 449, no. 7163, pp. 677–681, 2007.

[21] M. Sémon and K. H. Wolfe, "Consequences of genome duplication," *Current Opinion in Genetics and Development*, vol. 17, no. 6, pp. 505–512, 2007.

[22] B. Dujon, D. Sherman, G. Fischer et al., "Genome evolution in yeasts," *Nature*, vol. 430, pp. 35–44, 2004.

[23] M. Kellis, B. W. Birren, and E. S. Lander, "Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*," *Nature*, vol. 428, no. 6983, pp. 617–624, 2004.

[24] F. S. Dietrich, S. Voegeli, S. Brachat et al., "The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome," *Science*, vol. 304, no. 5668, pp. 304–307, 2004.

[25] M. Kasahara, "The 2R hypothesis: an update," *Current Opinion in Immunology*, vol. 19, no. 5, pp. 547–552, 2007.

[26] F. Jacob, "Evolution and tinkering," *Science*, vol. 196, no. 4295, pp. 1161–1166, 1977.

[27] U. Alon, "Biological networks: the tinkerer as an engineer," *Science*, vol. 301, no. 5641, pp. 1866–1867, 2003.

[28] A. Van Hoof, "Conserved functions of yeast genes support the duplication, degeneration and complementation model for gene duplication," *Genetics*, vol. 171, no. 4, pp. 1455–1461, 2005.

[29] P. J. Bhat and T. V. S. Murthy, "Transcriptional control of the GAL/MEL regulon of yeast *Saccharomyces cerevisiae*: mechanism of galactose-mediated signal transduction," *Molecular Microbiology*, vol. 40, no. 5, pp. 1059–1066, 2001.

[30] F. T. Zenke, R. Engels, V. Vollenbroich, J. Meyer, C. P. Hollenberg, and K. D. Breunig, "Activation of Gal4p by galactose-dependent interaction of galactokinase and Gal80p," *Science*, vol. 272, no. 5268, pp. 1662–1665, 1996.

[31] K. Wolfe, "Robustness—it's not where you think it is," *Nature Genetics*, vol. 25, no. 1, pp. 3–4, 2000.

[32] Y. Guan, M. J. Dunham, and O. G. Troyanskaya, "Functional analysis of gene duplications in *Saccharomyces cerevisiae*," *Genetics*, vol. 175, no. 2, pp. 933–943, 2007.

[33] G. D. Amoutzias, Y. He, J. Gordon, D. Mossialos, S. G. Oliver, and Y. Van De Peer, "Posttranslational regulation impacts the fate of duplicated genes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 7, pp. 2967–2971, 2010.

[34] G. C. Conant and K. H. Wolfe, "Functional partitioning of yeast co-expression networks after genome duplication," *PLoS Biology*, vol. 4, no. 4, article e109, 2006.

[35] R. Geladé, S. Van de Velde, P. V. Van Dijck, and J. M. Thevelein, "Multi-level response of the yeast genome to glucose," *Genome Biology*, vol. 4, no. 11, article 233, 2003.

[36] M. Johnston and J. H. Kim, "Glucose as a hormone: receptor-mediated glucose sensing in the yeast *Saccharomyces cerevisiae*," *Biochemical Society Transactions*, vol. 33, no. 1, pp. 247–252, 2005.

[37] A. Merico, P. Sulo, J. Piškur, and C. Compagno, "Fermentative lifestyle in yeasts belonging to the *Saccharomyces* complex," *FEBS Journal*, vol. 274, no. 4, pp. 976–989, 2007.

[38] P. Herrero, J. Galíndez, N. Ruiz, C. Martinez-Campa, and F. Moreno, "Transcriptional regulation of the *Saccharomyces cerevisiae HXK1, HXK2* and *GLK1* genes," *Yeast*, vol. 11, no. 2, pp. 137–144, 1995.

[39] A. Maier, B. Völker, E. Boles, and G. F. Fuhrmann, "Characterisation of glucose transport in *Saccharomyces cerevisiae* with plasma membrane vesicles (countertransport) and intact cells (initial uptake) with single Hxt1, Hxt2, Hxt3, Hxt4, Hxt6, Hxt7 or Gal2 transporters," *FEMS Yeast Research*, vol. 2, no. 4, pp. 539–550, 2002.

[40] S. Özcan and M. Johnston, "Function and regulation of yeast hexose transporters," *Microbiology and Molecular Biology Reviews*, vol. 63, no. 3, pp. 554–569, 1999.

[41] R. C. MacLean and I. Gudelj, "Resource competition and social conflict in experimental populations of yeast," *Nature*, vol. 441, no. 7092, pp. 498–501, 2006.

[42] T. Pfeiffer and S. Schuster, "Game-theoretical approaches to studying the evolution of biochemical systems," *Trends in Biochemical Sciences*, vol. 30, no. 1, pp. 20–25, 2005.

[43] T. Pfeiffer, S. Schuster, and S. Bonhoeffer, "Cooperation and competition in the evolution of ATP-producing pathways," *Science*, vol. 292, no. 5516, pp. 504–507, 2001.

[44] G. Hardin, "The tragedy of the commons," *Science*, vol. 162, no. 3859, pp. 1243–1248, 1968.

[45] G. C. Conant, "Rapid reorganization of the transcriptional regulatory network after genome duplication in yeast," *Proceedings of the Royal Society B*, vol. 277, no. 1683, pp. 869–876, 2010.

[46] W. K. Huh, J. V. Falvo, L. C. Gerke et al., "Global analysis of protein localization in budding yeast," *Nature*, vol. 425, no. 6959, pp. 686–691, 2003.

[47] G. C. Conant and K. H. Wolfe, "Increased glycolytic flux as an outcome of whole-genome duplication in yeast," *Molecular Systems Biology*, vol. 3, article 129, 2007.

[48] N. C. Duarte, M. J. Herrgård, and B. Ø. Palsson, "Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model," *Genome Research*, vol. 14, no. 7, pp. 1298–1309, 2004.

[49] Å. Pérez-Bercoff Å, A. McLysaght, and G. C. Conant, "Patterns of indirect protein interactions suggest a spatial organization to metabolism," *Molecular BioSystems*, vol. 7, pp. 3056–3064, 2011.

[50] G. C. Conant and K. H. Wolfe, "Probabilistic cross-species inference of orthologous genomic regions created by whole-genome duplication in yeast," *Genetics*, vol. 179, no. 3, pp. 1681–1692, 2008.

[51] L. M. Blank, F. Lehmbeck, and U. Sauer, "Metabolic-flux and network analysis in fourteen hemiascomycetous yeasts," *FEMS Yeast Research*, vol. 5, no. 6-7, pp. 545–558, 2005.

[52] J. Piškur, E. Rozpedowska, S. Polakova, A. Merico, and C. Compagno, "How did *Saccharomyces* evolve to become a good brewer?" *Trends in Genetics*, vol. 22, no. 4, pp. 183–186, 2006.

[53] M. J. A. Van Hoek and P. Hogeweg, "Metabolic adaptation after whole genome duplication," *Molecular Biology and Evolution*, vol. 26, no. 11, pp. 2441–2453, 2009.

[54] J. Ihmels, S. Bergmann, M. Gerami-Nejad et al., "Molecular biology: rewiring of the yeast transcriptional network through the evolution of motif usage," *Science*, vol. 309, no. 5736, pp. 938–940, 2005.

[55] J. M. Thomson, E. A. Gaucher, M. F. Burgan et al., "Resurrecting ancestral alcohol dehydrogenases from yeast," *Nature Genetics*, vol. 37, no. 6, pp. 630–635, 2005.

[56] D. Fusco, L. Grassi, B. Bassetti, M. Caselle, and M. Cosentino Lagomarsino, "Ordered structure of the transcription network inherited from the yeast whole-genome duplication," *BMC Systems Biology*, vol. 4, article 77, 2010.

[57] K. P. Byrne and K. H. Wolfe, "The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species," *Genome Research*, vol. 15, no. 10, pp. 1456–1461, 2005.

[58] A. M. Evangelisti and G. C. Conant, "Nonrandom survival of gene conversions among yeast ribosomal proteins duplicated through genome doubling," *Genome Biology and Evolution*, vol. 2, pp. 826–834, 2010.

[59] F. A. Kondrashov and A. S. Kondrashov, "Role of selection in fixation of gene duplications," *Journal of Theoretical Biology*, vol. 239, no. 2, pp. 141–151, 2006.

[60] T. Y. Kim, C. W. Ha, and W. K. Huh, "Differential subcellular localization of ribosomal protein L7 paralogs in *Saccharomyces cerevisiae*," *Molecules and Cells*, vol. 27, no. 5, pp. 539–546, 2009.

[61] S. Komili, N. G. Farny, F. P. Roth, and P. A. Silver, "Functional specificity among ribosomal proteins regulates gene expression," *Cell*, vol. 131, no. 3, pp. 557–571, 2007.

[62] L. Ni and M. Snyder, "A genomic study of the bipolar bud site selection pattern in *Saccharomyces cerevisiae*," *Molecular Biology of the Cell*, vol. 12, no. 7, pp. 2147–2170, 2001.

[63] Y. Van De Peer, "Computational approaches to unveiling ancient genome duplications," *Nature Reviews Genetics*, vol. 5, no. 10, pp. 752–763, 2004.

[64] D. E. Soltis, V. A. Albert, J. Leebens-Mack et al., "Polyploidy and angiosperm diversification," *American Journal of Botany*, vol. 96, no. 1, pp. 336–348, 2009.

[65] R. De Smet and Y. Van de Peer, "Redundancy and rewiring of genetic networks following genome-wide duplication events," *Current Opinion in Plant Biology*, vol. 15, pp. 168–176, 2012.

[66] M. E. Schranz, S. Mohammadin, and P. P. Edger, "Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model," *Current Opinion in Plant Biology*, vol. 15, pp. 147–153, 2012.

[67] L. Chae, I. Lee, J. Shin, and S. Y. Rhee, "Toward understanding how molecuar networks evolve in plants," *Current Opinion in Plant Biology*, vol. 15, pp. 177–184, 2012.

[68] L. M. Liberman, R. Sozzani, and P. N. Benfey, "Integrative systems biology: an attempt to describe a simple weed," *Current Opinion in Plant Biology*, vol. 15, pp. 162–167, 2012.

[69] M. E. Schranz, P. P. Edger, J. C. Pires, N. M. van Dam, and C. W. Wheat, "Comparative genomics in the Brassicales: ancient genome duplications, glucossinolate diversification and Pierinae herbivore radiation," in *Genetics, Genomics and Breeding of Oilseed Brassicas*, J. B. David Edwards, I. Parkin, and C. Kole, Eds., Science Publishers, Jersey, British Isles, UK, 2011.

[70] M. Bekaert, P. P. Edger, J. C. Pires, and G. C. Conant, "Two-phase resolution of polyploidy in the *Arabidopsis* metabolic network gives rise to relative and absolute dosage constraints," *Plant Cell*, vol. 23, no. 5, pp. 1719–1728, 2011.

[71] J. W. Valentine, *On the Origin of Phyla*, University of Chicago Press, Chicago, Ill, USA, 2004.

[72] N. King, M. J. Westbrook, S. L. Young et al., "The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans," *Nature*, vol. 451, no. 7180, pp. 783–788, 2008.

[73] B. F. Lang, C. O'Kelly, T. Nerad, M. W. Gray, and G. Burger, "The closest unicellular relatives of animals," *Current Biology*, vol. 12, no. 20, pp. 1773–1778, 2002.

[74] O. Voigt, A. G. Collins, V. B. Pearse et al., "Placozoa—no longer a phylum of one," *Current Biology*, vol. 14, no. 22, pp. R944–R945, 2004.

[75] B. Schierwater, "My favorite animal, *Trichoplax adhaerens*," *BioEssays*, vol. 27, no. 12, pp. 1294–1302, 2005.

[76] S. L. Dellaporta, A. Xu, S. Sagasser et al., "Mitochondrial genome of *Trichoplax adhaerens* supports Placozoa as the basal lower metazoan phylum," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 23, pp. 8751–8756, 2006.

[77] S. P. Leys, D. S. Rohksar, and B. M. Degnan, "Sponges," *Current Biology*, vol. 15, no. 4, pp. R114–R115, 2005.

[78] C. Nielsen, "Six major steps in animal evolution: are we derived sponge larvae?" *Evolution and Development*, vol. 10, no. 2, pp. 241–257, 2008.

[79] J. S. Taylor, Y. Van de Peer, I. Braasch, and A. Meyer, "Comparative genomics provides evidence for an ancient genome duplication event in fish," *Philosophical Transactions of the Royal Society B*, vol. 356, no. 1414, pp. 1661–1679, 2001.

[80] K. Vandepoele, W. De Vos, J. S. Taylor, A. Meyer, and Y. Van De Peer, "Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 6, pp. 1638–1643, 2004.

[81] F. J. Chain and B. J. Evans, "Multiple mechanisms promote the retained expression of gene duplicates in the tetraploid frog Xenopus laevis," *PLoS Genetics*, vol. 2, no. 4, article e56, 2006.

[82] E. S. Lander, L. M. Linton, B. Birren et al., "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, pp. 860–921, 2001.

[83] J. C. Venter, M. D. Adams, E. W. Myers et al., "The sequence of the human genome," *Science*, vol. 291, pp. 1304–1351, 2001.

[84] I. Ruiz-Trillo, G. Burger, P. W. H. Holland et al., "The origins of multicellularity: a multi-taxon genome initiative," *Trends in Genetics*, vol. 23, no. 3, pp. 113–118, 2007.

[85] A. Pires-daSilva and R. J. Sommer, "The evolution of signalling pathways in animal development," *Nature Reviews Genetics*, vol. 4, no. 1, pp. 39–49, 2003.

[86] E. M. De Robertis, "Evo-Devo: variations on ancestral themes," *Cell*, vol. 132, no. 2, pp. 185–195, 2008.

[87] S. B. Carroll, J. K. Grenier, and S. D. Weatherbee, *From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design*, Blackwell, Malden, Mass, USA, 2nd edition, 2005.

[88] L. Huminiecki and C. H. Heldin, "2R and remodeling of vertebrate signal transduction engine," *BMC Biology*, vol. 8, article 146, 2010.

[89] P. W. H. Holland, J. Garcia-Fernandez, N. A. Williams, and A. Sidow, "Gene duplications and the origins of vertebrate development," *Development*, vol. 120, pp. 125–133, 1994.

[90] P. ten Dijke and C. H. Heldin, *Smad Signal Transduction: Smads in Proliferation, Differentiation and Disease*, Springer, Dordrecht, The Netherlands, 2006.

[91] L. Huminiecki, L. Goldovsky, S. Freilich, A. Moustakas, C. Ouzounis, and C. H. Heldin, "Emergence, development and diversification of the TGF- signalling pathway within the animal kingdom," *BMC Evolutionary Biology*, vol. 9, no. 1, article 28, 2009.

[92] E. Kim and M. Sheng, "PDZ domain proteins of synapses," *Nature Reviews Neuroscience*, vol. 5, no. 10, pp. 771–781, 2004.

[93] R. D. Emes, A. J. Pocklington, C. N. G. Anderson et al., "Evolutionary expansion and anatomical specialization of synapse proteome complexity," *Nature Neuroscience*, vol. 11, no. 7, pp. 799–806, 2008.

[94] J. D. Watson, T. A. Baker, S. P. Bell, A. Gann, M. Levine, and R. Losick, *Molecular Biology of the Gene: International Edition*, Benjamin Cummings, San Francisco, Calif, USA, 6th edition, 2007.

[95] F. Marks, U. Klingüller, and K. Müller-Decker, *Cellular Signal Processing: An Introduction To the Molecular Mechanisms of Signal Transduction*, Garland Science, New York, NY, USA, 2009.

[96] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*, Garland Science, New York, NY, USA, 5th, edition, 2008.

[97] J. Roux and M. Robinson-Rechavi, "Developmental constraints on vertebrate genome evolution," *PLoS Genetics*, vol. 4, no. 12, article e1000311, 2008.

[98] J. Roux and M. Robinson-Rechavi, "Age-dependent gain of alternative splice forms and biased duplication explain the relation between splicing and duplication," *Genome Research*, vol. 21, no. 3, pp. 357–363, 2011.

[99] K. S. Kassahn, V. T. Dang, S. J. Wilkins, A. C. Perkins, and M. A. Ragan, "Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates," *Genome Research*, vol. 19, no. 8, pp. 1404–1418, 2009.

[100] C. P. Hickman Jr., L. S. Roberts, S. L. Keen, A. Larson, and D. Eisenhour, *Animal Diversity*, The McGraw-Hill, Columbus, Ohio, USA, 5th edition, 2008.

[101] N. H. Putnam, T. Butts, D. E. K. Ferrier et al., "The amphioxus genome and the evolution of the chordate karyotype," *Nature*, vol. 453, no. 7198, pp. 1064–1071, 2008.

[102] L. Z. Holland, R. Albalat, K. Azumi et al., "The amphioxus genome illuminates vertebrate origins and cephalochordate biology," *Genome Research*, vol. 18, pp. 1100–1111, 2008.

[103] A. Fernandez-Guerra, A. Aze, J. Morales et al., "The genomic repertoire for cell cycle control and DNA metabolism in *S. purpuratus*," *Developmental Biology*, vol. 300, no. 1, pp. 238–251, 2006.

[104] E. Sodergren, G. M. Weinstock, E. H. Davidson et al., "The genome of the sea urchin Strongylocentrotus purpuratus," *Science*, vol. 314, pp. 941–952, 2006.

[105] Q. Tu, C. T. Brown, E. H. Davidson, and P. Oliveri, "Sea urchin Forkhead gene family: phylogeny and embryonic expression," *Developmental Biology*, vol. 300, no. 1, pp. 49–62, 2006.

[106] A. Wagner, "Neutralism and selectionism: a network-based reconciliation," *Nature Reviews Genetics*, vol. 9, no. 12, pp. 965–974, 2008.

[107] A. Wagner, *Robustness and Evolvability in Living Systems*, Princeton University Press, Princeton, NJ, USA, 2005.

[108] J. S. Edwards and B. O. Palsson, "The *Escherichia coli* MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 10, pp. 5528–5533, 2000.

[109] I. Amit, R. Wides, and Y. Yarden, "Evolvable signaling networks of receptor tyrosine kinases: relevance of robustness to malignancy and to cancer therapy," *Molecular Systems Biology*, vol. 3, article 151, 2007.

[110] X. Dong, P. Navratilova, D. Fredman, Ø. Drivenes, T. S. Becker, and B. Lenhard, "Exonic remnants of whole-genome duplication reveal cis-regulatory function of coding exons," *Nucleic Acids Research*, vol. 38, no. 4, pp. 1071–1085, 2009.

[111] A. M. Heimberg, L. F. Sempere, V. N. Moy, P. C. J. Donoghue, and K. J. Peterson, "MicroRNAs and the advent of vertebrate morphological complexity," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 8, pp. 2946–2950, 2008.

[112] J. Li, G. Musso, and Z. Zhang, "Preferential regulation of duplicated genes by microRNAs in mammals," *Genome Biology*, vol. 9, no. 8, article R132, 2008.

[113] J. A. Fawcett, S. Maere, and Y. Van De Peer, "Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 14, pp. 5737–5742, 2009.

[114] X. He and J. Zhang, "Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution," *Genetics*, vol. 169, no. 2, pp. 1157–1164, 2005.

Hindawi Publishing Corporation
International Journal of Evolutionary Biology
Volume 2012, Article ID 436196, 13 pages
doi:10.1155/2012/436196

Review Article

Transposon Invasion of the *Paramecium* Germline Genome Countered by a Domesticated PiggyBac Transposase and the NHEJ Pathway

Emeline Dubois,[1,2,3] Julien Bischerour,[1,2,3] Antoine Marmignon,[1,2,3] Nathalie Mathy,[1,2,3] Vinciane Régnier,[1,2,3,4] and Mireille Bétermier[1,2,3]

[1] *CNRS, Centre de Génétique Moléculaire, UPR3404, 1 Avenue de la Terrasse, 91198 Gif-sur-Yvette Cedex, France*
[2] *CNRS, Centre de Recherches de Gif-sur-Yvette, FRC3115, 91198 Gif-sur-Yvette Cedex, France*
[3] *Université Paris-Sud, Département de Biologie, 91405 Orsay, France*
[4] *Université Paris Diderot, Sorbonne Paris Cité, Sciences du Vivant, 75205 Paris Cedex 13, France*

Correspondence should be addressed to Mireille Bétermier, mireille.betermier@cgm.cnrs-gif.fr

Received 20 March 2012; Accepted 7 May 2012

Academic Editor: Frédéric Brunet

Sequences related to transposons constitute a large fraction of extant genomes, but insertions within coding sequences have generally not been tolerated during evolution. Thanks to their unique nuclear dimorphism and to their original mechanism of programmed DNA elimination from their somatic nucleus (macronucleus), ciliates are emerging model organisms for the study of the impact of transposable elements on genomes. The germline genome of the ciliate *Paramecium*, located in its micronucleus, contains thousands of short intervening sequences, the IESs, which interrupt 47% of genes. Recent data provided support to the hypothesis that an evolutionary link exists between *Paramecium* IESs and *Tc1/mariner* transposons. During development of the macronucleus, IESs are excised precisely thanks to the coordinated action of PiggyMac, a domesticated *piggyBac* transposase, and of the NHEJ double-strand break repair pathway. A PiggyMac homolog is also required for developmentally programmed DNA elimination in another ciliate, *Tetrahymena*. Here, we present an overview of the life cycle of these unicellular eukaryotes and of the developmentally programmed genome rearrangements that take place at each sexual cycle. We discuss how ancient domestication of a *piggyBac* transposase might have allowed *Tc1/mariner* elements to spread throughout the germline genome of *Paramecium*, without strong counterselection against insertion within genes.

1. Introduction

Since the initial evidence for the existence of transposable elements (TEs) reported by McClintock [1], large-scale sequencing of the genome of a wide range of living organisms has highlighted the abundance of TE-derived sequences relative to the coding portion of genomes. Transposable elements, often considered as "selfish" or "parasitic" DNA, are mobile genetic elements that encode their own mobility enzymes and move from one genomic locus to another. Based on their transposition mechanisms, they can be classified into two main categories [2, 3]: class I elements transpose *via* the reverse transcription of an RNA molecule, while class II elements transpose *via* a DNA intermediate. Class II transposons, also called DNA transposons, are found in variable proportions among eukaryotic and prokaryotic genomes; for instance, they constitute the major fraction of resident TEs in bacteria (reviewed in [4]) but are underrepresented relative to class I elements in the human genome [5] and absent from the genome of the yeast *Saccharomyces cerevisiae* [6]. Among DNA transposons, cut-and-paste transposons move in two steps: (i) excision from the donor site, as a result of transposase-induced DNA cleavages at their ends and (ii) integration into the target site through strand transfer of their free 3′OH ends. The most widespread cut-and-paste transposons, which are found in

eukaryotes and in prokaryotes, encode a so-called "DDE transposase," an enzyme that bears a conserved triad of acidic residues (DDE or DDD) that catalyzes the excision and integration steps [2, 7]. Upon integration, these particular cut-and-paste transposons duplicate a short (2 to 15 bp) target sequence (TSD or target site duplication) on each side of the newly integrated copy (reviewed in [8]). Upon excision, staggered double-strand cleavages at their ends generally leave the two copies of the TSD at the donor site and the resulting double-strand break can be repaired by end joining [9–11]. This generates a "footprint" at the excision site, formed by two copies of the TSD flanking a few remaining bp from the transposon. However, if transposition takes places during replication, homologous recombination with the sister chromatid may restore the initial transposon copy at the donor locus, which leads to a net increase in transposon copy number in the genome [11–13].

Increase in transposon copy number may have detrimental effects on host fitness, since insertions can disrupt coding regions, modify the expression of adjacent cellular genes, or trigger ectopic recombination between distant transposon copies (reviewed in [8]). Several defense strategies have been developed by the host to inactivate transposition. In eukaryotes, posttranscriptional inactivation of TEs is mediated by homologous small RNAs, which may also induce histone and DNA methylation to inactivate the transcription of transposon genes (reviewed in [14]). In some hosts, like filamentous fungi, heavy mutagenesis of repeated sequences, a phenomenon named repeat-induced point mutation (or RIP), was also reported [15], but the involvement of small RNAs in this process has not been clearly demonstrated. As a result of these defense responses, many extant genomes harbor large numbers of defective copies of transposable elements that have lost their ability to transpose. Transposons, however, may also provide novel and advantageous functions to the host, as illustrated by the growing list of entire or truncated transposase genes found in eukaryotic genomes, either isolated or fused to genes encoding unrelated protein domains, which still appear to be expressed and encode proteins, but are not embedded within a mobile element anymore [16]. According to several criteria described in [8, 17], these genes have been domesticated to become cellular genes, but, in most cases, their function has not been elucidated. Most often, only putative transposase domains involved in nucleic acid binding have been conserved and may play a role in cellular DNA or RNA metabolism. Intriguingly, most domesticated transposases have lost their characteristic DDE (or DDD) signature and only for very few of them have evidence been obtained for *in vivo* or *in vitro* DNA cleavage activity. Remarkable examples of catalytically active domesticated DDE transposases were reported in different organisms: the RAG1 endonuclease, related to *Transib* transposases [18], catalyzes V(D)J recombination of immunoglobulin genes during the differentiation of lymphocytes in vertebrates (reviewed in [19]); alpha3, a domesticated *mutator*-like transposase, is involved in mating-type switching in the yeast *Kluyveromyces lactis* [20]; finally, SETMAR, identified in the human genome, carries a histone methyltransferase domain fused to the partially active catalytic site of a *mariner* transposase [21] and is thought to participate in the repair of DNA double-strand breaks, in the restart of stalled DNA replication forks and in chromosome decatenation ([22], reviewed in [23]).

Evidence for a role of domesticated *piggyBac* transposases in programmed genome rearrangements was reported recently in ciliates [24, 25]. Thanks to their unique nuclear dimorphism and to their original mechanism of programmed DNA elimination from their somatic nucleus, ciliates, and most specifically *Paramecium*, have emerged as novel model organisms for the study of the impact of transposable elements on genomes (for recent reviews, see [26, 27]). Here, we will present an overview of the life cycle of these unicellular eukaryotes and of the massive and developmentally programmed genome rearrangements that take place at each sexual cycle. We will discuss how the domestication of an ancient *piggyBac* cut-and-paste transposase in *Paramecium* might subsequently have allowed *Tc1*/*mariner* elements to spread throughout the germline genome, without strong counterselection against insertion within genes.

2. Developmentally Programmed Elimination of Germline Transposons and Related Sequences in Ciliates

2.1. Nuclear Dimorphism in Ciliates. Ciliates form a deeply branching monophyletic group in the eukaryote tree [28]. These unicellular organisms are characterized by the coexistence, in their cytoplasm, of two functionally distinct types of nuclei (Figure 1). The diploid germline micronucleus (MIC) is transcriptionally silent during vegetative growth, but harbors the genetic information that is transmitted to the next sexual generation. According to ciliate species, the number of MICs per cell may vary (one in *Tetrahymena thermophila* and two in *Paramecium tetraurelia*, e.g.). Gene expression is carried out from the highly polyploid somatic macronucleus (MAC: ~800n in *Paramecium*), which is therefore essential for cell survival at all stages. During sexual events (conjugation between compatible mating types or, for some species, self-fertilization also called autogamy), MIC meiosis leads to the formation of haploid nuclei, one of which divides once to yield two identical gametic nuclei. The fusion of two gametic nuclei (reciprocally exchanged between mating partners during conjugation or originating from the same cell during autogamy) gives rise to the zygotic nucleus. In the meantime, the MAC is progressively degraded and is ultimately lost. New MICs and MACs differentiate from mitotic copies of the zygotic nucleus. Throughout this developmental process, the old MAC ensures all gene transcription and is progressively replaced by the new MAC [29]. Therefore, the development of a functional new MAC is essential for the survival of sexual progeny, once the old MAC has disappeared from the cell.

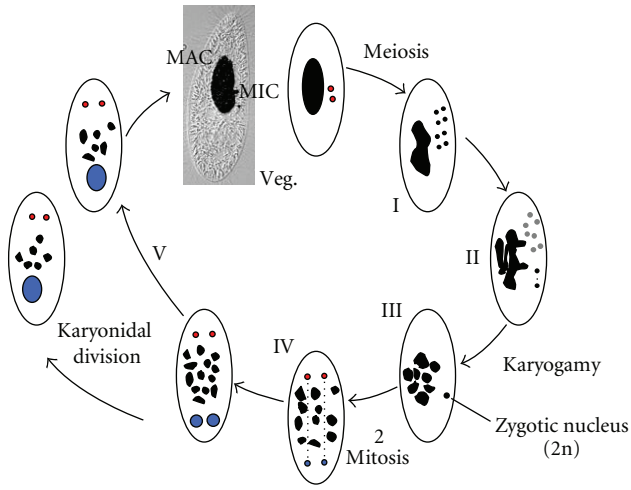


Figure 1: Nuclear dimorphism and the sexual cycle in *Paramecium*. The merged picture on top shows a vegetative *Paramecium* cell (Veg.), with its DAPI-stained nuclei (in black). The figure represents the major steps of the sexual cycle observed during autogamy, a self-fertilization process triggered by starvation. Upon starvation, the two germline diploid MICs (red) undergo meiosis to give rise to eight haploid nuclei (I), a single of which migrates to a specialized cell compartment, where it divides once to give two identical gametic nuclei (II). Meanwhile, the remaining seven meiotic products are degraded (grey dots in II) and the old MAC (black) gets fragmented into ~30 pieces. During karyogamy, two gametic nuclei fuse to form the diploid zygotic nucleus (III). The zygotic nucleus then undergoes two successive mitotic divisions (IV): after the second division, the nuclei which migrate to the anterior cellular pole become the new MICs of the sexual progeny (red), while those that localize to the posterior pole differentiate into new developing MACs (blue) and undergo programmed genome rearrangements. At the first cell division (or karyonidal division), the new MICs divide by mitosis and each of the two developing new MACs segregates into a daughter cell (V), where it continues to amplify the rearranged somatic genome to a final ploidy of ~800n. During conjugation (not shown), meiosis is triggered by the mating of two compatible sexual partners, which undergo reciprocal exchange of their haploid gametic nuclei. As a result, the zygotic nucleus in each partner is formed by the fusion of a resident and a migratory haploid nucleus. Exconjugants separate between the first and second divisions of the zygotic nucleus, and MAC development takes place as described for autogamous cells.

2.2. MIC and MAC Genomes Have Different Structures. MAC chromosomes of ciliates are shorter than their MIC chromosomes and apparently do not carry centromeres, which is consistent with the observation that the MAC divides through an amitotic process, with no chromosome condensation (reviewed in [30]). In contrast, the MIC undergoes mitosis and meiosis. Moreover, early studies of the complexity of MIC and MAC genomes pointed out that the two nuclei do not harbor the same DNA content, although they derive from the same zygotic nucleus. Indeed, the MIC genome contains additional sequences that are removed from the somatic genome during MAC development (reviewed in [31, 32]).

Pulse field electrophoresis analyses indicated that the size of *Paramecium* MAC chromosomes varies between 50 kb and 1 Mb [33]. The study of particular MAC chromosome ends in *P. primaurelia* [34] and *P. tetraurelia* [35, 36] revealed that they are capped by a mixture of G_3T_3 or G_4T_2 telomeric repeats added at heterogeneous positions by a single, error-prone telomerase [37, 38]; several telomere-addition regions distant of several kbp have been identified for some MAC chromosomes, each one extending over ~1 kb. The MAC genome sequence of *Paramecium tetraurelia* was obtained in 2006 by a consortium of European labs [39]. This study provided a global view of the structure of some 150 acentromeric MAC chromosomes and highlighted the fact that the somatic genome is streamlined for gene expression, with a very high gene density (78% coding) and essentially no repeated sequences. Even more striking, ~40,000 genes were annotated in the MAC genome, as a consequence of at least three successive whole genome duplications (WGD) during evolution of the *Paramecium aurelia* group of sibling species, to which *P. tetraurelia* belongs.

In contrast to somatic "chromosomes," only limited knowledge of the number and structure of *Paramecium* germline chromosomes is available (Figure 2). Early microscopy studies proposed that 35 to 50 pairs of 1 to 7 Mb chromosomes harbor the genetic content of the MIC in *P. tetraurelia* [40]. Molecular analyses of a couple of germline regions encompassing MAC chromosome ends revealed no conserved nucleotide sequence motif for chromosome fragmentation in *Paramecium* [35]. This situation is quite different from that observed in other ciliates, in which consensus chromosome breakage sequences (CBS) were found at fragmentation sites (reviewed in [31]). Instead, the fragmentation of *Paramecium* MAC chromosomes seems to be associated with heterogeneous elimination of repeated germline sequences (minisatellites, germline transposons, etc.) located downstream of telomere addition sites [34, 41]. Southern blot hybridization experiments confirmed that known germline transposons are eliminated from the somatic genome during MAC development (O. Garnier, unpublished and [24, 42]). On a genome-wide scale, more work is clearly needed to gain full insight into the DNA content (and more specifically their TE landscape) of large germline regions that are eliminated from the MAC in association with chromosome fragmentation. In contrast, along chromosomes, the availability of a λ phage library of *P. tetraurelia* MIC DNA constructed by Preer et al. in 1992—which has represented a technical *tour de force* [43]—made it possible to compare the nucleotide sequence of particular MAC and collinear MIC loci; these studies led to the identification of short, noncoding sequences called IESs (internal eliminated sequences) that interrupt both coding and noncoding regions in the germline genome and are excised precisely from MAC chromosomes ([44], reviewed in [45, 46]). Recent genome-wide sequencing of a set of 45,000 IESs confirmed the early description of these sequences [41]. *Paramecium* IESs are very short (93% are shorter than 150 bp long and one-third are within the 26–30 bp size range) and each one appears to be single copy in the genome. Their exquisitely precise excision is essential

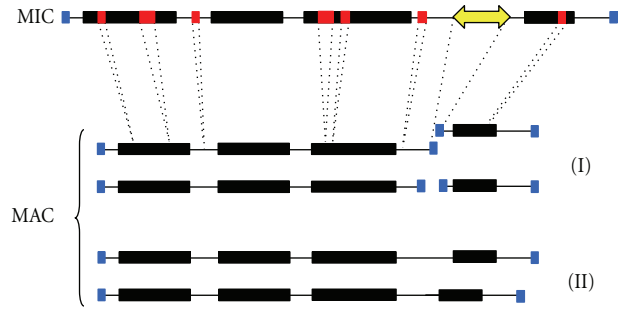


Figure 2: Schematic representation of the structure of the MIC and MAC genomes in *Paramecium*. On the MIC chromosome displayed on top, genes (black boxes) are interrupted by short internal eliminated sequences (IESs, in red), some of which are also found in noncoding regions (thin line). Repeated germline sequences (transposons, minisatellites, etc.) are symbolized by a yellow double-headed arrow. During MAC development, each MIC chromosome is amplified ~400-fold and gives rise to a population of heterogeneous MAC chromosomes. Indeed, imprecise elimination of repeated DNA is associated with alternative rearrangements: (I) chromosome fragmentation is observed and telomeres (blue squares) are added to new MAC chromosome ends; (II) the two chromosome arms that flank the eliminated germline region can be joined in an imprecise manner to generate internal deletions of heterogeneous sizes.

for the recovery of a functional new MAC, since 47% of genes are interrupted by at least one of these intervening sequences in the germline genome [41]. Their only absolutely conserved feature is the presence of one flanking 5′-TA-3′ at each end, while a single TA is retained at their excision site on mature MAC chromosomes. Because of this conservation, *Paramecium* IESs have defined the family of the so-called "TA-IESs" [47]. In other ciliates, IESs are also eliminated during MAC development, but their structure varies from one species to the other. The existence of short TA-IESs has been reported in *Euplotes crassus* and *Oxytricha fallax*, while IESs in *Tetrahymena thermophila* are larger and are generally not flanked by TA repeats (reviewed in [31]).

2.3. Paramecium IESs Are Likely Remnants of Tc1/mariner Transposons. Statistical analysis of the nucleotide sequence of the ends of ~20 IESs from different *Paramecium aurelia* species was performed by Klobutcher and Herrick, who identified a degenerate 8 bp consensus (5′**TA**(C/T)AG (C/T)N(A/G)3′) that defines a loosely conserved terminal inverted repeat (TIR) at IES ends [48]. This consensus sequence, which includes the flanking TA, was confirmed by all the following analyses of increasing numbers of IESs [41, 46, 49]. Interestingly, it also matches the ends of the short TA-IESs found in *Euplotes* and *Oxytricha* [47] and also of *Tc1/mariner*-related transposons Tec1 and Tec2 present in high copy numbers in the germline genome of *Euplotes crassus* [48]. Based on the observation that TA-IESs and Tec transposons coexist in *Euplotes*, Klobutcher and Herrick proposed their IBAF model (invasion/bloom/abdicate/fade), according to which TA-IESs within a given ciliate species have evolved from *Tc1/mariner* transposons, which would have invaded the MIC genome and accumulated internal substitutions/deletions during evolution. Therefore, transposon remnants would have lost their coding capacity while being kept under strong selection pressure for their elimination from the somatic genome [50]. Interestingly, a common feature of *Tc1/mariner* transposons is their preference for TA dinucleotides as integration targets, which they duplicate upon insertion; thus, the conserved TAs at the boundaries of TA-IESs would simply be the TSDs generated by integration of ancestral *Tc1/mariner*-related TEs.

The IBAF model for the evolutionary origin of IESs has recently obtained further support in *Paramecium*. Transposon-like sequences were indeed identified in the heterogeneously eliminated fraction of the germline genome of different *P. aurelia* strains [34, 41]. Sequence alignment of these elements has led to the establishment of a consensus for each family (*Tennessee* in *P. primaurelia*, *Sardine* and *Thon* in *P. tetraurelia*) and to the unambiguous identification of open reading frames encoding putative transposases harboring the characteristic DD(35)E triad of *Tc1/mariner*/IS*630* transposons. Reminiscent of the situation described in *E. crassus*, the six outward terminal nucleotides (including the flanking TA) of the long (500 to 700 bp) and complex TIRs of the *Sardine* and *Thon* elements of *P. tetraurelia* match the consensus of IES ends (Figure 3(a)). Analysis of the three WGDs that took place during the evolutionary history of *P. tetraurelia* provided evidence that IESs have appeared continuously in the germline genome and that their size tends to shorten over time [41]. Among the largest IESs (>500 bp), a few closely related IESs are inserted at nonhomologous germline loci. Some of them exhibit significant sequence similarities with the long TIRs of known TEs from the *Thon* family and are excised from MAC chromosomes just like any other IES. The existence of these "solo TIRs" provides further support to the notion that some IESs at least have derived from recently mobile TEs from the *Tc1/mariner* family. Genetic evidence indicates that the conserved TAs, which are supposed to represent the TSD created by integration of the ancestral *Tc1/mariner*, are essential for the developmentally programmed excision of IESs [51–55].

3. IES Excision in *Paramecium*: A Cut-and-Close Reaction Mediated by a Domesticated Transposase

3.1. IES Excision Is Related to Cut-and-Paste Transposition of piggyBac, Not of Tc1/mariner Elements. One of the assumptions of the IBAF model is that, at first, ancestral invading germline *Tc1/mariner* transposons were eliminated from the somatic genome during MAC development, thanks to the action of their own transposase [50]; thus, programmed genome rearrangements have allowed TEs to proliferate in the MIC, with little or no effect on the phenotype of the cell, as long as they are correctly excised from the MAC. Then,

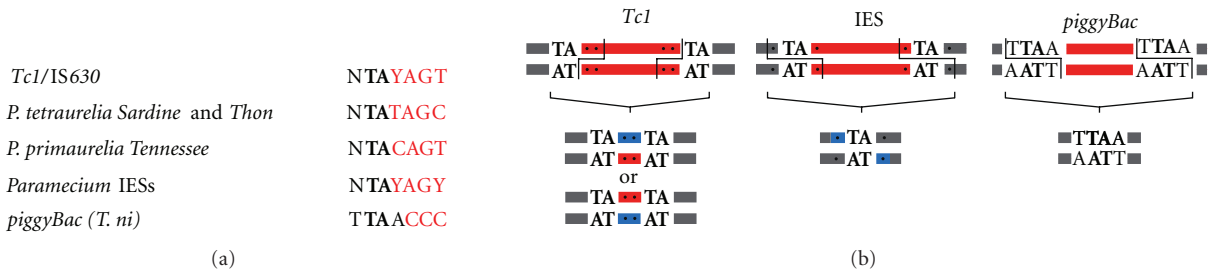


Figure 3: *Paramecium* IES excision: a comparison with *Tc1/mariner* and *piggyBac* transposons. (a) Nucleotide sequence alignment of the ends of *Paramecium* IESs with the termini of *Tc1* transposons (general *Tc1/IS630* consensus and transposon families identified in *Paramecium*) and of the *piggyBac* element from *Trichoplusia ni*. Flanking sequences are in black (the conserved TA found at each boundary is highlighted in bold) and internal nucleotides are in red. Note that for *piggyBac*, the target site duplication is made of 4 base pairs (TTAA in black). (b) The geometry of double-strand DNA cleavages introduced by *Tc1* (left) and *piggyBac* (right) transposases is shown on top, together with that of PiggyMac-dependent DSBs detected at *Paramecium* IES ends (middle). The conserved TAs are represented by black bold letters. Based on their transposition mechanism, *Tc1* and *piggyBac* transposons are delimited by their cleaved 3′ ends and are represented by red lines. By analogy with *Tc1*, IESs are drawn as red lines bounded by two flanking TAs (in black), although this does not reflect the actual position of DNA cleavages. At the bottom of each panel, the structure of chromosomal junctions formed after excision from the donor site is shown. For *Tc1* transposons and *Paramecium* IESs, the nucleotides that are neosynthesized during gap filling and repair are represented in blue.

at some point during evolution, a cellular gene (possibly a domesticated or preexisting *Tc1/mariner* transposase gene) took over the catalysis of excision of all these elements, allowing them to accumulate internal mutations and give rise to current IESs, while still being able to excise from the MAC. As already discussed [45], one caveat of this model is that *Tc1/mariner* transposition leaves a characteristic footprint at the excision (or donor) site [10], while IESs are excised precisely at the nucleotide level, leaving only one copy of the original duplicated TA at the excision junction (Figure 3(b)).

Molecular analyses of IES excision intermediates formed *in vivo* during sexual processes in *P. tetraurelia* provided important information about the mechanisms involved in this process. IES excision starts after a few rounds of endoduplication of the germline genome have taken place, so that at least 16 copies of each IES need to be excised in each developing MAC [56]. It is initiated by 4 bp staggered double-strand DNA cleavages at both ends of each IES, centered on the conserved TA dinucleotides [57]. As a result, transient double-strand breaks (DSBs) with characteristic 4-base 5′ overhangs can be detected by ligation-mediated PCR during MAC development, at the ends of linear excised IES molecules and at flanking MAC-destined DNA ends (Figure 3(b) and [57, 58]). Strikingly, these DSBs have the same geometry as those catalyzed *in vitro* by *piggyBac* transposases [59]. Indeed, *piggyBac* cut-and-paste transposons duplicate a 5′-TTAA-3′ target site upon integration, and when they transpose to a new locus, their transposase cleaves DNA on each side of each duplicated TSD to generate a 5′ TTAA overhang [59]. Thus, *piggyBac* excision is highly precise and reconstitutes the TTAA sequence at the donor site [60]. The discovery of *PiggyMac* (*PGM*), a domesticated *piggyBac* transposase gene in *P. tetraurelia*, represented a significant breakthrough towards the identification of protein partners involved in IES excision [24]. This gene is only expressed during sexual processes, with an induction peak during the development of new MACs, which corresponds to the time when IES excision starts. It encodes a large 1065 aa protein with a recognizable central domain homologous to the transposase of *piggyBac* transposons, including a potentially active DDD catalytic triad (Figure 4(a)). During sexual processes, Pgm-GFP fusion proteins were found to localize specifically in the developing new MACs, in which IES excision takes place (Figure 4(b) and [24], Dubois, unpublished). In cells silenced for expression of the *PGM* gene, IES excision is blocked as well as other known programmed genome rearrangements (chromosome fragmentation and heterogeneous elimination of *Sardine* transposons); as a result, strong lethality is observed in the sexual progeny. Nuclei of *PGM*-silenced cells (purified during the development of new MACs, before the cells die) provided the source of DNA that was used for whole-genome sequencing and identification of the set of 45,000 IESs described above [41]. Indicative of a catalytic function of Pgm, microinjection of a mutant transgene encoding a protein in which the DDD catalytic triad was switched to AAA induces a dominant negative effect on the survival of sexual progeny, while a normal phenotype is obtained with a wild-type transgene (Dubois, unpublished). Thus, even though *Paramecium* IESs are probably relics of *Tc1/mariner* transposons, their precise excision from the MAC genome appears to be carried out by a domesticated transposase related to a different family of transposable elements, the *piggyBac* family.

DNA transposons generally assemble a synaptic nucleoprotein complex called the "transpososome," which includes both transposon ends and oligomers of the transposase (see [61] for a review). Assembly of this complex activates the successive hydrolysis and transesterification steps that ultimately lead to transposon excision. Likewise, genetic evidence has indicated that IES excision in *Paramecium* involves an interaction between the two ends of each IES, before DNA cleavage [62]. For three IESs of different sizes (28, 66, and 370 bp), it was indeed shown that a mutation

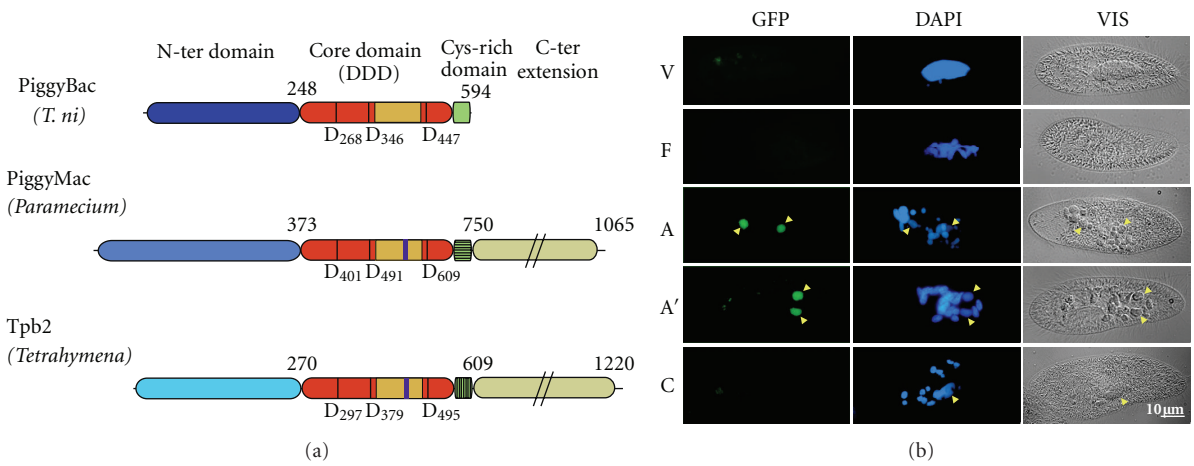


FIGURE 4: PiggyMac: a domesticated PiggyBac transposase in *Paramecium*. (a) Conserved domains in ciliate domesticated transposases. PiggyMac and the related Tpb2 protein from *Tetrahymena thermophila* were aligned with the transposase of the *piggyBac* transposon isolated from *Trichoplusia ni*, using the MUSCLE software (http://www.ebi.ac.uk/Tools/msa/muscle/). The conserved catalytic core domain is represented in orange, and the putative DDD catalytic residues are indicated for each protein (numbers refer to amino acid positions in the primary sequence). A β strand-rich module (yellow box) can be predicted between the second and third catalytic residues using the PSIPRED package (http://bioinf.cs.ucl.ac.uk/psipred/); in ciliate proteins, additional residues inserted within this module are represented by a purple bar. The cysteine-rich domain is drawn as a light green box at the C-terminus of PiggyBac, and corresponding variant domains in Pgm and Tpb2 as hatched boxes. The C-terminal coiled-coil extensions of ciliate domesticated proteins are not drawn to scale (light beige boxes). The divergent N-terminal domains of the three proteins are represented in different shades of blue. (b) A PiggyMac-GFP fusion localizes to developing new MACs. A transgene encoding a C-terminal GFP fusion expressed under the control of endogenous *PGM* transcription signals was microinjected into the MAC of vegetative cells (Dubois, unpublished). During autogamy, cells were fixed and nuclei were stained with DAPI and observed with a Zeiss epifluorescence microscope (magnification 630x). No GFP fluorescence was observed in vegetative cells (V) or at early stages during autogamy, when the old MAC starts its fragmentation (F). The GFP fusion protein was detected specifically in the two developing MACs of autogamous cells (arrowheads in A and A′) and GFP fluorescence disappeared from the new MAC after karyonidal division (C).

within the TA at one end not only inhibits cleavage of the mutant end but also strongly impairs DNA cleavage at the wild-type end of the same IES. Moreover, thorough analysis of the currently available set of 45,000 IESs of *P. tetraurelia* provided support to the hypothesis that IES excision, similar to transposition, involves the formation of an intramolecular DNA loop on a double-strand substrate [41]. Indeed, the size distribution of IESs exhibits a striking 10 bp periodicity, which coincides with the length covered by one turn of the DNA double helix. This suggests that interactions between Pgm molecules bound at each end of an IES depend critically on helical phasing, especially for very short sequences (93% of *Paramecium* IESs are shorter than 150 bp, the persistence length of double-strand DNA). As discussed for other systems (site-specific recombination, transposition, or repression), DNA looping between very closely spaced sites might also be favored by DNA bending factors and/or local melting of the double helix [63].

3.2. How May PiggyMac Recognize Paramecium IESs? The ends of cut-and-paste transposons are generally made of two parts: an internal sequence-specific binding site for their cognate transposase and a few nucleotides at their termini that constitute the DNA cleavage site *per se* (see, for example, [64]). For *piggyBac*, the site cleaved by the transposase is the TTAA duplicated target sequence on each side of the integrated copy of the element. A 13 bp terminal repeat (TR) and a 19 bp internal repeat (IR) separated by a spacer are present in inverted orientation at each end of the element and may be binding sites for the transposase [65, 66]. Analysis of the nucleotide sequence of 45,000 IESs from *P. tetraurelia* showed that the TTAA tetranucleotide is actually largely underrepresented at IES ends (Marmignon, unpublished), even though the sequence cleaved by PiggyMac at the termini of *Paramecium* IESs bears some similarity with the site cleaved by the PiggyBac transposase (i.e., a 4 bp sequence with a central TA). Furthermore, neither TR nor IR repeats are found at the ends of *Paramecium* IESs. The situation is even more striking for *Tetrahymena* IESs, which depend on a close PiggyMac homolog, the PiggyBac-like transposase called Tpb2, for their elimination [25]; indeed, the sequence cleaved at their ends (5′ANNNNT 3′) does not even carry a conserved central TA [67]. This suggests that ciliate domesticated *piggyBac* transposases do not recognize a specific nucleotide motif at IES ends and raises the question of how germline sequences are targeted for elimination.

Part of the answer may lie in the epigenetic mechanisms that control programmed genome rearrangements in *Paramecium* and *Tetrahymena* (reviewed in [26, 27, 69, 70]). It was proposed for both ciliates that a comparison between the DNA content of parental MIC and MAC genomes takes place during MIC meiosis through the annealing of two kinds of noncoding RNA molecules. Short RNAs, also called

scnRNAs (25 nt in *Paramecium*, 28 nt in *Tetrahymena*), are generated by a specialized RNA interference pathway from noncoding RNA precursors transcribed specifically from the MIC during meiosis. According to the "scanning" model, these scnRNAs would pair to larger transcripts that are produced constitutively by generalized transcription of the parental MAC genome, which was rearranged during the previous sexual cycle. Those scnRNAs that do not find homologous MAC sequences, and therefore represent the fraction of the germline genome that was absent from the parental MAC, are then imported into the new developing MAC, in which they are thought to target the deletion of homologous sequences. In *Tetrahymena*, the methylation of IES-associated histones is clearly one of the scnRNA-dependent epigenetic modifications that trigger the elimination of heterochromatin regions [71, 72]. In contrast, the putative epigenetic marks that are deposited by scnRNAs on *Paramecium* germline eliminated sequences have not been identified yet, especially for IESs, the vast majority of which are much shorter than the length of DNA wrapped around a nucleosome (~150 bp). Whatever the exact mechanism may be, a strong implication of the scanning model is that ciliates tend to reproduce their pattern of developmentally programmed genome rearrangements from one sexual generation to the next. Thus, epigenetic control may have contributed to loosen the requirement for a specific nucleotide sequence to direct ciliate domesticated PiggyBac transposases towards regions that have to be eliminated from the developing MAC. Quite interestingly, PiggyMac and Tpb2 present variant domains relative to PiggyBac transposases (within their catalytic site and a downstream cysteine-rich region) and have acquired long C-terminal extensions (Figure 4(a)); the role of these domains in IES recognition still has to be elucidated.

3.3. The Cellular Nonhomologous End Joining Double-Strand Break Repair Pathway Closes IES Excision Sites. Thanks to the particular cleavage properties of their transposase, the transposition of *piggyBac* transposons leaves no footprint at the donor site (Figure 3(b)), and excision junctions can be closed through direct annealing of the fully complementary 5′-TTAA-3′ overhangs generated on flanking DNA ends, with no need for any additional processing step [59]. For *Paramecium* IESs, however, the situation is quite different, since only the central TA is conserved on the 4-base overhangs created by Pgm-dependent cleavage. It was proposed that the closure of IES excision junctions on MAC chromosomes involves partial pairing of the flanking ends through annealing of their conserved TAs and limited additional processing (removal of the unpaired 5′-terminal nucleotides and gap-filling by addition of one nucleotide at each 3′ recessive end), before the final ligation step (Figure 5 and [57]). IESs are assumed to be excised as linear molecules and, at least for those larger than 200 bp, to be circularized in a second step using the same pathway (partial pairing of overhangs, 5′ and 3′ processing, ligation). The enzymes that carry out the additional processing steps have not been identified. However, recent work uncovered the essential role played by the ligase IV and its partner Xrcc4 in the closure of IES excision sites and the circularization of excised IESs [58]. Ligase IV and Xrcc4 are core actors of the nonhomologous end joining (NHEJ) pathway, which repairs DSBs through the direct joining of broken ends, without requiring sequence homology [73]. Two very closely related *LIG4* genes originating from the most recent WGD and a single *XRCC4* gene were identified in the genome of *P. tetraurelia*. Their expression reaches a peak during MIC meiosis, even before new MACs have differentiated from mitotic copies of the zygotic nucleus; this implies that induction of DSB repair genes is part of a developmental program in *Paramecium* rather than a response to DNA damage. In cells depleted either for ligase IV or Xrcc4, Pgm-dependent cleavages are introduced normally, but no detectable chromosomal—nor circular—junctions are formed; unrepaired DSBs accumulate at IES excision sites as well as linear forms of excised IESs. Noteworthy, DSBs are processed normally at their 5′ end in ligase IV-depleted cells (removal of the 5′terminal nucleotide), but no nucleotide addition is observed at their 3′ recessive end [58]. As already inferred from *in vitro* studies of reconstituted eukaryotic NHEJ systems [74], this indicates that the ligase IV participates in the recruitment or activation of a gap-filling DNA polymerase, prior to end joining (Figure 5).

The participation of actors of the NHEJ pathway in the final step of IES excision raises the question of how accurate end joining is achieved following the massive introduction of programmed DSBs throughout the genome. Indeed, given the number of IESs per haploid genome, thousands of DSBs are introduced all along chromosomes within a restricted time window during MAC development. The formation of a Pgm-containing synaptic excision complex prior to IES end cleavage might contribute to hold together adjacent fragments of MAC-destined DNA, ensuring, therefore, that somatic chromosomes are assembled in the right order during DSB repair (Figure 6). Furthermore, the human Xrcc4 protein was recently shown to form filaments with another NHEJ factor, Cernunnos (or XLF), independently of ligase IV [75–77]. These filaments are thought to promote the bridging between broken DNA ends [78]. Two Cernunnos homologs are encoded by paralogs of the recent WGD in *P. tetraurelia*, and their expression is induced during sexual processes [58]. During IES excision, such filaments may provide an alignment scaffold and favor the correct assembly of broken MAC ends. At the nucleotide level, an additional requirement for assembly of a functional MAC genome is the highly precise joining of each IES excision site to reconstitute open reading frames. Increasing evidence that the "classical" NHEJ pathway is inherently precise ([79], reviewed in [80]) has pointed to the key role played by the Ku70/Ku80 heterodimer in protecting broken DNA ends against resection and inhibiting other DSB repair pathways, such as alternative end-joining (which would create imprecise deletions) or homologous recombination (which would restore the non-rearranged molecule). Ku proteins have been conserved through evolution, from bacteria to humans [81], and several genes encoding putative Ku70 and Ku80 homologs were found in the genome of *P. tetraurelia* [58].

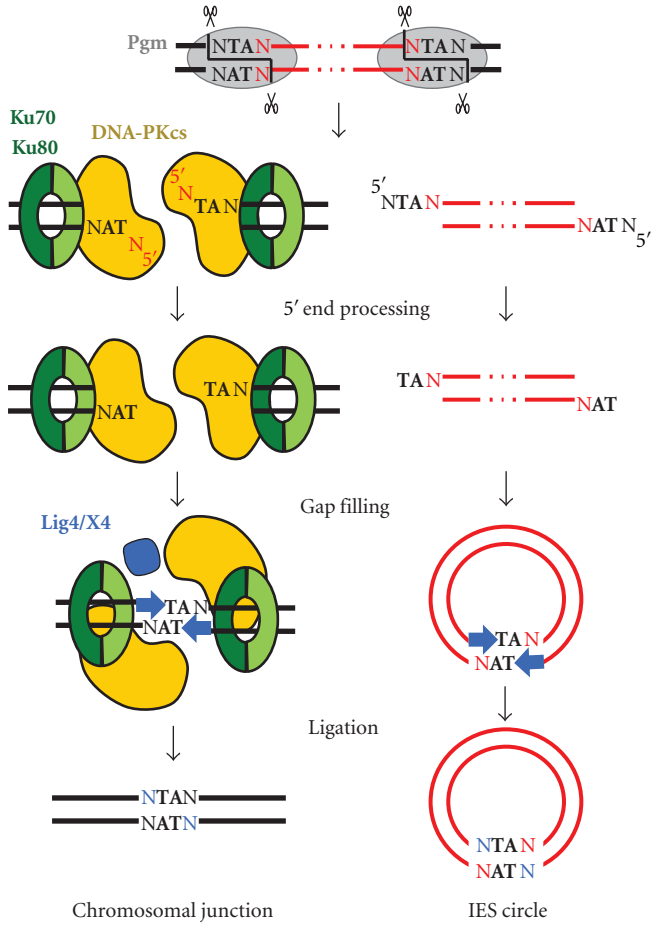


Figure 5: Molecular mechanism of IES excision in *Paramecium*. The successive DNA intermediates that are formed during IES excision are displayed, with IESs in red and flanking MAC-destined DNA in black. The proteins that were shown to be required for proper IES excision are also represented. The first step is the introduction of 4-base staggered double-strand breaks at each IES end and depends on the PiggyMac domesticated transposase (Pgm). According to available knowledge of the classical NHEJ pathway in other organisms, a Ku70/Ku80 dimer is proposed to bind to each broken flanking DNA end and recruits the DNA-PKcs catalytic subunit. The last steps of the reaction were proposed to take place within a paired-end intermediate guided by annealing of the central TA present on each 5′ overhang [57]. The proteins involved in the removal of the 5′ terminal nucleotide have not been identified. For the 3′ processing step, the ligase IV is required for recruiting or activating a gap-filling DNA polymerase, which adds one nucleotide to the recessive end, prior to final ligation. A similar mechanism is proposed for the circularization of excised linear IES molecules (right part of the figure), providing that they are long enough. IES circles do not replicate and are actively degraded.

Efficient recruitment of Ku proteins at IES excision sites probably plays a determinant role in the precision of DNA rearrangements. *P. tetraurelia* also harbors a unique gene encoding a homolog of the DNA-PKcs, a DNA-dependent protein kinase (Malinsky et al., in preparation) that interacts with the Ku dimer, facilitates the synapsis of broken DNA ends, and, after autophosphorylation, activates downstream NHEJ proteins [82]. The conservation of DNA-PKcs in *Paramecium*, even though this protein has been lost from other model organisms such as budding yeast or *Drosophila*, suggests that this protein was present in the ancestral eukaryotic NHEJ core machinery. Functional inactivation of the *KU* and *DNA-PKcs* genes by RNA interference indicates that Ku70/Ku80 and the DNA-PKcs homolog are required for IES excision (Marmignon, unpublished; Malinsky et al., in preparation). Strikingly, among the three *KU80* genes identified in the genome, only one is specifically expressed during MAC development and appears to have acquired a specialized function in genome rearrangements (Marmignon, unpublished).

3.4. Revisiting the IBAF Model. The availability of a genome-wide set of ~45,000 IESs in *Paramecium tetraurelia* has broadened our current view of the evolutionary history of the germline genome of ciliates [41]. All IESs that have been identified so far in *Paramecium* belong to the TA-IES family. Consistent with the IBAF model proposed by Klobutcher and Herrick, some of them at least seem to have evolved from ancestral *Tc1/mariner*-related transposons, still recognizable in the fraction of the MIC genome that is eliminated in an imprecise manner. This putative evolutionary link between IESs and TEs is similar to that proposed for *Euplotes*, a distant

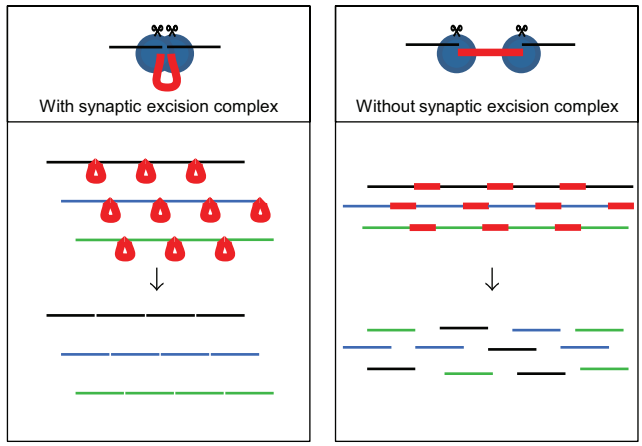


Figure 6: PiggyMac and the NHEJ pathway orchestrate accurate assembly of MAC chromosomes. On the diagrams shown on top of each panel, the IES is drawn in red and its flanking DNA in black. The excisase complex, which includes PiggyMac and putative additional partners, is represented in blue. In bottom panels, different germline chromosomes are displayed in different colors, with IESs in red. (a) If a synaptic complex is formed prior to cleavage, adjacent MAC-destined chromosome fragments are brought together, which might favor their alignment during the repair step, therefore limiting the risk of translocation. (b) In the absence of a synaptic complex, IES excision could lead to catastrophic chromosome breakage and translocation, as described in [68].

spirotrichous ciliate in which Tec transposons and related TA-IESs are excised precisely from the MAC genome, although the details of the mechanism may be somewhat different from IES excision in *Paramecium*. In particular, the enzyme responsible for IES excision in *Euplotes* has not been identified yet. In another stichotrichous ciliate, *Oxytricha*, at least three families of TBE transposons, also related to the *Tc1* family and initially designated as telomere-bearing elements, have been identified in the eliminated fraction of the MIC genome [83]. Along the lines of the IBAF model, RNA interference experiments have suggested that the TBE transposase itself mediates the elimination of TBE transposons from the somatic genome [84]. It appears to be also involved in other genome rearrangements reported in *Oxytricha*, such as IES excision and the unscrambling of a subset of genes, for which macronuclear-destined sequences are not collinear in the MIC and MAC genomes. This situation has provided a nice example of mutualism, rather than domestication, between resident transposons and their host (discussed in [85]).

In *Paramecium*, the discovery that elimination of IESs and *Tc1/mariner*-like transposons depends on a domesticated transposase related to the *piggyBac* family has provided an unexpected extension of the IBAF hypothesis [24]. As discussed previously [24, 41], the existence of a catalytically active PiggyMac homolog, Tpb2, also required for programmed genome rearrangements in *Tetrahymena thermophila* [25], indicates that domestication of a PiggyBac transposase occurred early during ciliate evolution, before the divergence between *Paramecium* and *Tetrahymena* (Figure 7(a)). The initial role of this ancestral PiggyBac transposase might have been to cope with a first invasion of *piggyBac* elements, by removing them from ciliate genomes. It may then have been recruited to carry out the elimination of other unrelated germline sequences from the MAC genome. Intriguingly, except for a few TTAA-IESs that may originate from *piggyBac* transposons (some of which add 3′ exons to genes that would be expressed transiently during MAC development), *Tetrahymena* IESs are generally not flanked by TA dinucleotides and differ significantly from those of *Paramecium*; they are larger and are usually multicopy elements, their excision generates microheterogeneity at chromosomal junctions, and they are very rarely found within coding sequences [26, 86]. This suggests, therefore, that invasion of the *Paramecium* germline genome by *Tc1/mariner* transposons took place after the separation of the two ciliate lineages (Figure 7(a)). This idea has been supported by an analysis of IES evolution in *Paramecium*, which led to the conclusion that the majority of TA-IESs appeared between the intermediate and recent WGDs [41], that is after divergence of *Paramecium* and *Tetrahymena* [39]. In *Paramecium*, the ability of PiggyMac and the NHEJ pathway to carry out the precise excision of *Tc1/mariner*-related elements from the MAC may have allowed these transposons to spread throughout the germline genome without harmful consequences on gene expression (Figure 7(b)). Thus, thanks to nuclear dimorphism and to the existence of a precise mechanism for transposon elimination from the somatic genome, *Paramecium*, in contrast to other organisms, may have tolerated insertions within genes. This raises the question of whether currently known IESs are the relatively harmless remnants of ancient *Tc1/mariner* invaders or whether they have acquired some useful function for the cell. As suggested earlier [87], some of them may contribute to the structuration of MIC chromosomes, for example, by providing centromere-related functions or ensuring the condensation of chromosomes. IESs may also have a regulatory role, if they carry sequences

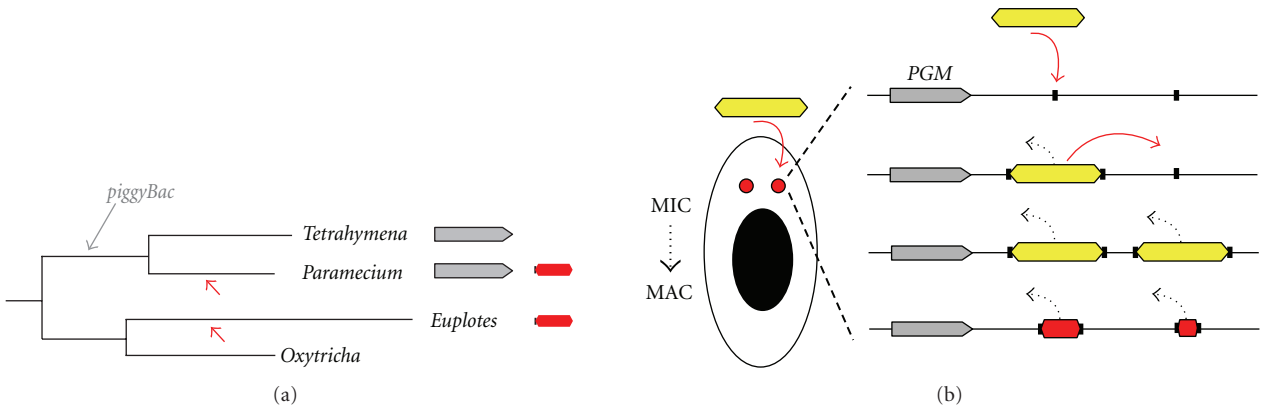


Figure 7: Evolutionary scenario for the origin of *Paramecium* IESs. (a) Putative timing for transposon invasion in the ciliate phylum (ciliate tree adapted from [28]). Identification of closely related domesticated *piggyBac* transposases in *Tetrahymena* and *Paramecium* (grey boxes) led to the hypothesis that a *piggyBac* transposon invaded the germline genome of one of their common ancestors (grey arrow), prior to the divergence between these two ciliate lineages. Because of the absence of TA-IESs in *Tetrahymena*, only the germline genome of *Paramecium* is thought to have undergone subsequent invasion by *Tc1/mariner* transposons (red arrowhead). TA-IESs (red box flanked by two black squares) and related transposons were found in the more distant ciliate *Euplotes*, but the protein(s) required for their developmentally programmed excision have not been identified. (b) In the revisited version of the IBAF model in *Paramecium*, the ancestor of the *PGM* gene (in grey) was already present when the first *Tc1/mariner* transposon (yellow box) started to invade the MIC. During the blooming step, Pgm may have been recruited to rid the genome from deleterious transposon insertions within genes. Thanks to the preexistence of the Pgm domesticated transposase and to the NHEJ repair pathway, *Tc1*-related transposons could be excised precisely from the somatic genome of the next sexual generation (programmed elimination from the MAC is represented by black dotted arrows), between the two duplicated copies of their TA target site (black squares). This has allowed invading *Tc1/mariners* to spread throughout the germline genome as a consequence of transposition catalyzed by their own transposase (mobility inside the MIC is symbolized by red arrows). During evolution, most copies of *Tc1/mariner* transposons have lost their coding capacity and have shortened in size, while being kept under selection pressure for their Pgm-dependent precise excision from the MAC, to give the currently known IESs (red boxes).

that can control transient gene expression specifically before they are removed from genes during genome rearrangements.

4. Conclusion

A recent study of ~10 million genes annotated in sequenced genomes from individual bacteria, archaea, eukaryotes, and viruses as well as in metagenomes, has pointed to the remarkable evolutionary success of transposase genes, which appear to be "the most abundant, the most ubiquitous genes in nature" [88]. This brought further support to the idea that transposons should not simply be considered as selfish or parasitic elements, but also as a source of novel and sometimes essential functions for their host. In mammalian genomes for instance, numerous transposase genes seem to have been domesticated, but, for the most part, their function has remained elusive [8]. The best documented example is the RAG1 nuclease involved in V(D)J recombination, a process that generates the highly diverse repertoire of immunoglobulin genes in differentiating B and T lymphocytes (reviewed in [19]). RAG1 is clearly a domesticated transposase from the *Transib* family, and its target sites within immunoglobulin genes, also called the recombination signal sequences, present significant sequence similarities with the TIRs of *Transib* transposable elements [18]. As in *Paramecium*, the NHEJ double-strand break repair pathway has been recruited in this system to join the coding (and signal) ends and assemble functional immunoglobulin genes. In V(D)J recombination, however, additional factors (such as nucleases and a template-free DNA polymerase) contribute to the observed variability of the coding junctions.

In addition to being yet another example of a catalytically active domesticated transposase involved in programmed DNA elimination during differentiation, PiggyMac in *Paramecium* represents a novel variation on the theme of how a genome can cope with invasion by transposable elements. Here, a domesticated *piggyBac* transposase, the NHEJ pathway and epigenetic control by noncoding RNAs orchestrate a highly precise and accurate system for the programmed elimination of transposon-related sequences from somatic chromosomes. As discussed in [89], recent observations have indicated that some IESs in *Paramecium* may carry promoters or parts of coding sequences, the excision of which would be regulated during the development of a new MAC and could also be submitted to homology-dependent epigenetic control of the old MAC. Whether IES excision may have provided an additional layer of variability for the control of gene expression at the genome-wide scale is an attractive hypothesis that will need to be investigated.

Acknowledgments

The authors would like to thank all members of the Sperling, Cohen, Meyer, and Duharcourt labs for the very stimulating

discussions. They also thank Sophie Malinsky for sharing unpublished data and, together with Linda Sperling, for critical reading of the authors' paper. Work in the authors' lab has been supported by the Agence Nationale pour la Recherche (ANR BLAN08-3_310945 "ParaDice" and ANR 2010 BLAN 1603 "GENOMAC") and the CNRS (ATIP Plus grant to M. Bétermier). Much of the work reviewed here was carried out in the context of the CNRS-supported European consortium "*Paramecium* Genome Dynamics and Evolution." A. Marmignon is the recipient of a Ph.D. fellowship from the Ministère de l'Enseignement Supérieur et de la Recherche.

References

[1] B. McClintock, "The origin and behavior of mutable loci in maize," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 36, no. 6, pp. 344–355, 1950.

[2] M. J. Curcio and K. M. Derbyshire, "The outs and ins of transposition: from MU to kangaroo," *Nature Reviews Molecular Cell Biology*, vol. 4, no. 11, pp. 865–877, 2003.

[3] T. Wicker, F. Sabot, A. Hua-Van et al., "A unified classification system for eukaryotic transposable elements," *Nature Reviews Genetics*, vol. 8, no. 12, pp. 973–982, 2007.

[4] P. Siguier, J. Filée, and M. Chandler, "Insertion sequences in prokaryotic genomes," *Current Opinion in Microbiology*, vol. 9, no. 5, pp. 526–531, 2006.

[5] E. S. Lander, L. M. Linton, B. Birren et al., "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, pp. 860–921, 2001.

[6] A. Goffeau, G. Barrell, H. Bussey et al., "Life with 6000 genes," *Science*, vol. 274, no. 5287, pp. 546–567, 1996.

[7] Y. W. Yuan and S. R. Wessler, "The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 19, pp. 7884–7889, 2011.

[8] C. Feschotte and E. J. Pritham, "DNA transposons and the evolution of eukaryotic genomes," *Annual Review of Genetics*, vol. 41, pp. 331–368, 2007.

[9] B. E. Staveley, T. R. Heslip, R. B. Hodgetts, and J. B. Bell, "Protected P-element termini suggest a role for inverted-repeat-binding protein in transposase-induced gap repair in *Drosophila melanogaster*," *Genetics*, vol. 139, no. 3, pp. 1321–1329, 1995.

[10] R. H. A. Plasterk and H. G. A. M. van Luenen, "The *Tc1/mariner* family of transposable elements," in *Mobile DNA II*, N. Craig, R. Craigie, M. Gellert, and A. Lambowitz, Eds., pp. 519–532, American Society for Microbiology, Washington, DC, USA, 2002.

[11] V. J. Robert, M. W. Davis, E. M. Jorgensen, and J. L. Bessereau, "Gene conversion and end-joining-repair double-strand breaks in the *Caenorhabditis elegans* germline," *Genetics*, vol. 180, no. 1, pp. 673–679, 2008.

[12] J. Bender, J. Kuo, and N. Kleckner, "Genetic evidence against intramolecular rejoining of the donor DNA molecule following IS*10* transposition," *Genetics*, vol. 128, no. 4, pp. 687–694, 1991.

[13] W. R. Engels, D. M. Johnson-Schlitz, W. B. Eggleston, and J. Sved, "High-frequency P element loss in *Drosophila* is homolog dependent," *Cell*, vol. 62, no. 3, pp. 515–525, 1990.

[14] C. D. Malone and G. J. Hannon, "Small RNAs as Guardians of the Genome," *Cell*, vol. 136, no. 4, pp. 656–668, 2009.

[15] J. E. Galagan and E. U. Selker, "RIP: the evolutionary cost of genome defense," *Trends in Genetics*, vol. 20, no. 9, pp. 417–423, 2004.

[16] L. Sinzelle, Z. Izsvák, and Z. Ivics, "Molecular domestication of transposable elements: from detrimental parasites to useful host genes," *Cellular and Molecular Life Sciences*, vol. 66, no. 6, pp. 1073–1093, 2009.

[17] J. N. Volff, "Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes," *BioEssays*, vol. 28, no. 9, pp. 913–922, 2006.

[18] V. V. Kapitonov and J. Jurka, "RAG1 core and V(D)J recombination signal sequences were derived from *Transib* transposons," *PLoS Biology*, vol. 3, no. 6, article e181, 2005.

[19] D. G. Schatz and P. C. Swanson, "V(D)J recombination: mechanisms of initiation," *Annual Review of Genetics*, vol. 45, pp. 167–202, 2011.

[20] E. Barsoum, P. Martinez, and S. U. Åström, "*α*3, a transposable element that promotes host sexual reproduction," *Genes and Development*, vol. 24, no. 1, pp. 33–44, 2010.

[21] D. Liu, J. Bischerour, A. Siddique, N. Buisine, Y. Bigot, and R. Chalmers, "The human SETMAR protein preserves most of the activities of the ancestral *Hsmar1* transposase," *Molecular and Cellular Biology*, vol. 27, no. 3, pp. 1125–1132, 2007.

[22] S. H. Lee, M. Oshige, S. T. Durant et al., "The SET domain protein Metnase mediates foreign DNA integration and links integration to nonhomologous end-joining repair," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 50, pp. 18075–18080, 2005.

[23] M. Shaheen, E. Williamson, J. Nickoloff, S. H. Lee, and R. Hromas, "Metnase/SETMAR: a domesticated primate transposase that enhances DNA repair, replication, and decatenation," *Genetica*, vol. 138, no. 5, pp. 559–566, 2010.

[24] C. Baudry, S. Malinsky, M. Restituito et al., "PiggyMac, a domesticated *piggyBac* transposase involved in programmed genome rearrangements in the ciliate *Paramecium tetraurelia*," *Genes and Development*, vol. 23, no. 21, pp. 2478–2483, 2009.

[25] C. Y. Cheng, A. Vogt, K. Mochizuki, and M. C. Yao, "A domesticated piggyBac transposase plays key roles in heterochromatin dynamics and DNA cleavage during programmed DNA deletion in *Tetrahymena thermophila*," *Molecular Biology of the Cell*, vol. 21, no. 10, pp. 1753–1762, 2010.

[26] D. L. Chalker and M. C. Yao, "DNA elimination in ciliates: transposon domestication and genome surveillance," *Annual Review of Genetics*, vol. 45, pp. 227–246, 2011.

[27] U. E. Schoeberl and K. Mochizuki, "Keeping the soma free of transposons: programmed DNA elimination in ciliates," *The Journal of Biological Chemistry*, vol. 286, pp. 37045–37052, 2011.

[28] S. L. Baldauf, A. J. Roger, I. Wenk-Siefert, and W. F. Doolittle, "A kingdom-level phylogeny of eukaryotes based on combined protein data," *Science*, vol. 290, no. 5493, pp. 972–977, 2000.

[29] J. D. Berger, "Nuclear differentiation and nucleic acid synthesis in well fed exconjugants of *Paramecium aurelia*," *Chromosoma*, vol. 42, no. 3, pp. 247–268, 1973.

[30] D. M. Prescott, "The DNA of ciliated protozoa," *Microbiological Reviews*, vol. 58, no. 2, pp. 233–267, 1994.

[31] C. L. Jahn and L. A. Klobutcher, "Genome remodeling in ciliated protozoa," *Annual Review of Microbiology*, vol. 56, pp. 489–520, 2002.

[32] M. C. Yao, S. Duharcourt, and D. L. Chalker, "Genome-wide rearrangements of DNA in ciliates," in *Mobile DNA II*, N. L. Craig, R. Craigie, M. Gellert, and A. M. Lambowitz, Eds., pp. 730–758, ASM Press, Washington, DC, USA, 2002.

[33] F. Caron and E. Meyer, "Molecular basis of surface antigen variation in paramecia," *Annual Review of Microbiology*, vol. 43, pp. 23–42, 1989.

[34] A. Le Mouël, A. Butler, F. Caron, and E. Meyer, "Developmentally regulated chromosome fragmentation linked to imprecise elimination of repeated sequences in *Paramecium*," *Eukaryotic Cell*, vol. 2, no. 5, pp. 1076–1090, 2003.

[35] J. D. Forney and E. H. Blackburn, "Developmentally controlled telomere addition in wild-type and mutant paramecia," *Molecular and Cellular Biology*, vol. 8, no. 1, pp. 251–258, 1988.

[36] L. Amar and K. Dubrana, "Epigenetic control of chromosome breakage at the 5′ end of *Paramecium tetraurelia* gene A," *Eukaryotic Cell*, vol. 3, no. 5, pp. 1136–1146, 2004.

[37] M. McCormick-Graham and D. P. Romero, "A single telomerase RNA is sufficient for the synthesis of variable telomeric DNA repeats in ciliates of the genus *Paramecium*," *Molecular and Cellular Biology*, vol. 16, no. 4, pp. 1871–1879, 1996.

[38] M. McCormick-Graham, W. J. Haynes, and D. P. Romero, "Variable telomeric repeat synthesis in *Paramecium tetraurelia* is consistent with misincorporation by telomerase," *EMBO Journal*, vol. 16, no. 11, pp. 3233–3242, 1997.

[39] J. M. Aury, O. Jaillon, L. Duret et al., "Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*," *Nature*, vol. 444, no. 7116, pp. 171–178, 2006.

[40] K. W. Jones, *Nuclear differentiation in Paramecium [Ph.D. thesis]*, Aberystwyth, University of Wales, 1956.

[41] O. Arnaiz, N. Mathy, C. Baudry et al., "The *Paramecium* germline genome provides a niche for intragenic parasitic DNA: evolutionary dynamics of internal eliminated sequences," under revision.

[42] M. Nowacki, W. Zagorski-Ostoja, and E. Meyer, "Nowa1p and Nowa2p: novel putative RNA binding proteins involved in trans-nuclear crosstalk in *Paramecium tetraurelia*," *Current Biology*, vol. 15, no. 18, pp. 1616–1628, 2005.

[43] L. B. Preer, G. Hamilton, and J. R. Preer, "Micronuclear DNA from *Paramecium tetraurelia*: serotype 51 A gene has internally eliminated sequences," *Journal of Protozoology*, vol. 39, no. 6, pp. 678–682, 1992.

[44] C. J. Steele, G. A. Barkocy-Gallagher, L. B. Preer, and J. R. Preer, "Developmentally excised sequences in micronuclear DNA of *Paramecium*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 91, no. 6, pp. 2255–2259, 1994.

[45] A. Gratias and M. Bétermier, "Developmentally programmed excision of internal DNA sequences in *Paramecium aurelia*," *Biochimie*, vol. 83, no. 11-12, pp. 1009–1022, 2001.

[46] M. Bétermier, "Large-scale genome remodelling by the developmentally programmed elimination of germ line sequences in the ciliate *Paramecium*," *Research in Microbiology*, vol. 155, no. 5, pp. 399–408, 2004.

[47] M. E. Jacobs and L. A. Klobutcher, "The long and the short of developmental DNA deletion in *Euplotes crassus*," *Journal of Eukaryotic Microbiology*, vol. 43, no. 6, pp. 442–452, 1996.

[48] L. A. Klobutcher and G. Herrick, "Consensus inverted terminal repeat sequence of *Paramecium* IESs: resemblance to termini of *Tc1*-related and *Euplotes* Tec transposons," *Nucleic Acids Research*, vol. 23, no. 11, pp. 2006–2013, 1995.

[49] L. Duret, J. Cohen, C. Jubin et al., "Analysis of sequence variability in the macronuclear DNA of *Paramecium tetraurelia*: a somatic view of the germline," *Genome Research*, vol. 18, no. 4, pp. 585–596, 2008.

[50] L. A. Klobutcher and G. Herrick, "Developmental genome reorganization in ciliated protozoa: the transposon link," *Progress in Nucleic Acid Research and Molecular Biology*, vol. 56, pp. 1–62, 1997.

[51] W. John Haynes, K. Y. Ling, R. R. Preston, Y. Saimi, and C. Kung, "The cloning and molecular analysis of pawn-B in *Paramecium tetraurelia*," *Genetics*, vol. 155, no. 3, pp. 1105–1117, 2000.

[52] A. Matsuda, K. M. Mayer, and J. D. Forney, "Identification of single nucleotide mutations that prevent developmentally programmed DNA elimination in *Paramecium tetraurelia*," *Journal of Eukaryotic Microbiology*, vol. 51, no. 6, pp. 664–669, 2004.

[53] K. M. Mayer and J. D. Forney, "A mutation in the flanking 5′-TA-3′ dinucleotide prevents excision of an internal eliminated sequence from the *Paramecium tetraurelia* genome," *Genetics*, vol. 151, no. 2, pp. 597–694, 1999.

[54] K. M. Mayer, K. Mikami, and J. D. Forney, "A mutation in *Paramecium tetraurelia* reveals functional and structural features of developmentally excised DNA elements," *Genetics*, vol. 148, no. 1, pp. 139–149, 1998.

[55] F. Ruiz, A. Krzywicka, C. Klotz et al., "The *SM19* gene, required for duplication of basal bodies in *Paramecium*, encodes a novel tubulin, η-tubulin," *Current Biology*, vol. 10, no. 22, pp. 1451–1454, 2000.

[56] M. Bétermier, S. Duharcourt, H. Seitz, and E. Meyer, "Timing of developmentally programmed excision and circularization of *Paramecium* internal eliminated sequences," *Molecular and Cellular Biology*, vol. 20, no. 5, pp. 1553–1561, 2000.

[57] A. Gratias and M. Bétermier, "Processing of double-strand breaks is involved in the precise excision of *Paramecium* internal eliminated sequences," *Molecular and Cellular Biology*, vol. 23, no. 20, pp. 7152–7162, 2003.

[58] A. Kapusta, A. Matsuda, A. Marmignon et al., "Highly precise and developmentally programmed genome assembly in *Paramecium* requires ligase IV-dependent end joining," *PLoS Genetics*, vol. 7, no. 4, Article ID e1002049, 2011.

[59] R. Mitra, J. Fain-Thornton, and N. L. Craig, "*piggyBac* can bypass DNA synthesis during cut and paste transposition," *EMBO Journal*, vol. 27, no. 7, pp. 1097–1109, 2008.

[60] T. A. Elick, C. A. Bauser, and M. J. Fraser, "Excision of the *piggyBac* transposable element in vitro is a precise event that is enhanced by the expression of its encoded transposase," *Genetica*, vol. 98, no. 1, pp. 33–41, 1996.

[61] S. P. Montaño and P. A. Rice, "Moving DNA around: DNA transposition and retroviral integration," *Current Opinion in Structural Biology*, vol. 21, no. 3, pp. 370–378, 2011.

[62] A. Gratias, G. Lepère, O. Garnier et al., "Developmentally programmed DNA splicing in *Paramecium* reveals short-distance crosstalk between DNA cleavage sites," *Nucleic Acids Research*, vol. 36, no. 10, pp. 3244–3251, 2008.

[63] L. Saiz and J. M. Vilar, "DNA looping: the consequences and its control," *Current Opinion in Structural Biology*, vol. 16, no. 3, pp. 344–350, 2006.

[64] M. Chandler and J. Mahillon, "Insertion sequences revisited," in *Mobile DNA II*, N. L. Craig, R. Craigie, M. Gellert, and A. M. Lambovitz, Eds., pp. 305–366, ASM Press, Washington, DC, 2002.

[65] L. C. Cary, M. Goebel, B. G. Corsaro, H. G. Wang, E. Rosen, and M. J. Fraser, "Transposon mutagenesis of baculoviruses: analysis of *Trichoplusia ni* transposon IFP2 insertions within the FP-locus of nuclear polyhedrosis viruses," *Virology*, vol. 172, no. 1, pp. 156–169, 1989.

[66] X. Li, N. Lobo, C. A. Bauser, and M. J. Fraser Jr., "The minimum internal and external sequence requirements for transposition of the eukaryotic transformation vector *piggyBac*,"

Molecular Genetics and Genomics, vol. 266, no. 2, pp. 190–198, 2001.

[67] S. V. Saveliev and M. M. Cox, "Developmentally programmed DNA deletion in *Tetrahymena thermophila* by a transposition-like reaction pathway," *EMBO Journal*, vol. 15, no. 11, pp. 2858–2869, 1996.

[68] P. J. Stephens, C. D. Greenman, B. Fu et al., "Massive genomic rearrangement acquired in a single catastrophic event during cancer development," *Cell*, vol. 144, no. 1, pp. 27–40, 2011.

[69] S. Duharcourt, G. Lepère, and E. Meyer, "Developmental genome rearrangements in ciliates: a natural genomic subtraction mediated by non-coding transcripts," *Trends in Genetics*, vol. 25, no. 8, pp. 344–350, 2009.

[70] R. S. Coyne, M. Lhuillier-Akakpo, and S. Duharcourt, "RNA-guided DNA rearrangements in ciliates: is the best genome defense a good offense?" *Biology of the Cell*, vol. 104, pp. 1–17, 2012.

[71] Y. Liu, K. Mochizuki, and M. A. Gorovsky, "Histone H3 lysine 9 methylation is required for DNA elimination in developing macronuclei in *Tetrahymena*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 6, pp. 1679–1684, 2004.

[72] Y. Liu, S. D. Taverna, T. L. Muratore, J. Shabanowitz, D. F. Hunt, and C. D. Allis, "RNAi-dependent H3K27 methylation is required for heterochromatin formation and DNA elimination in *Tetrahymena*," *Genes and Development*, vol. 21, no. 12, pp. 1530–1545, 2007.

[73] M. R. Lieber, "The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway," *Annual Review of Biochemistry*, vol. 79, pp. 181–211, 2010.

[74] J. Budman, S. A. Kim, and G. Chu, "Processing of DNA for nonhomologous end-joining is controlled by kinase activity and XRCC4/ligase IV," *Journal of Biological Chemistry*, vol. 282, no. 16, pp. 11950–11959, 2007.

[75] M. Hammel, M. Rey, Y. Yu et al., "XRCC4 protein interactions with XRCC4-like factor (XLF) create an extended grooved scaffold for DNA ligation and double strand break repair," *The Journal of Biological Chemistry*, vol. 286, pp. 32638–32650, 2011.

[76] V. Ropars, P. Drevet, P. Legrand et al., "Structural characterization of filaments formed by human Xrcc4-Cernunnos/XLF complex involved in nonhomologous DNA end-joining," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 31, pp. 12663–12668, 2011.

[77] S. N. Andres, A. Vergnes, D. Ristic, C. Wyman, M. Modesti, and M. Junop, "A human XRCC4-XLF complex bridges DNA," *Nucleic Acids Research*, vol. 40, pp. 1868–1878, 2012.

[78] S. Roy, S. N. Andres, A. Vergnes et al., "XRCC4's interaction with XLF is required for coding (but not signal) end joining," *Nucleic Acids Res*, vol. 40, pp. 1684–1694, 2012.

[79] J. Guirouilh-Barbat, S. Huck, P. Bertrand et al., "Impact of the KU80 pathway on NHEJ-induced genome rearrangements in mammalian cells," *Molecular Cell*, vol. 14, no. 5, pp. 611–623, 2004.

[80] J. E. Haber, "Alternative endings," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, pp. 405–406, 2008.

[81] S. Shuman and M. S. Glickman, "Bacterial DNA repair by non-homologous end joining," *Nature Reviews Microbiology*, vol. 5, no. 11, pp. 852–861, 2007.

[82] E. Weterings and D. J. Chen, "The endless tale of non-homologous end-joining," *Cell Research*, vol. 18, no. 1, pp. 114–124, 2008.

[83] K. Williams, T. G. Doak, and G. Herrick, "Developmental precise excision of *Oxytricha trifallax* telomere-bearing elements and formation of circles closed by a copy of the flanking target duplication," *EMBO Journal*, vol. 12, no. 12, pp. 4593–4601, 1993.

[84] M. Nowacki, B. P. Higgins, G. M. Maquilan, E. C. Swart, T. G. Doak, and L. F. Landweber, "A functional role for transposases in a large eukaryotic genome," *Science*, vol. 324, no. 5929, pp. 935–938, 2009.

[85] D. L. Chalker, "Transposons that clean up after themselves," *Genome Biology*, vol. 10, no. 6, p. 224, 2009.

[86] J. N. Fass, N. A. Joshi, M. T. Couvillion et al., "Genome-scale analysis of programmed DNA elimination sites in *Tetrahymena thermophila*," *G3*, vol. 1, pp. 515–522, 2011.

[87] R. S. Coyne, D. L. Chalker, and M. C. Yao, "Genome downsizing during ciliate development: nuclear division of labor through chromosome restructuring," *Annual Review of Genetics*, vol. 30, pp. 557–578, 1996.

[88] R. K. Aziz, M. Breitbart, and R. A. Edwards, "Transposases are the most abundant, most ubiquitous genes in nature," *Nucleic Acids Research*, vol. 38, no. 13, pp. 4207–4217, 2010.

[89] L. Sperling, "Remembrance of things past retrieved from the *Paramecium* genome," *Research in Microbiology*, vol. 162, no. 6, pp. 587–597, 2011.

Hindawi Publishing Corporation
International Journal of Evolutionary Biology
Volume 2012, Article ID 207958, 14 pages
doi:10.1155/2012/207958

Review Article

Why Chromosome Palindromes?

Esther Betrán, Jeffery P. Demuth, and Anna Williford

Department of Biology, University of Texas at Arlington, Box 19498, Arlington, TX 76019, USA

Correspondence should be addressed to Esther Betrán, betran@uta.edu

Received 31 March 2012; Accepted 9 May 2012

Academic Editor: Hideki Innan

We look at sex-limited chromosome (Y or W) evolution with particular emphasis on the importance of palindromes. Y chromosome palindromes consist of inverted duplicates that allow for local recombination in an otherwise nonrecombining chromosome. Since palindromes enable intrachromosomal gene conversion that can help eliminate deleterious mutations, they are often highlighted as mechanisms to protect against Y degeneration. However, the adaptive significance of recombination resides in its ability to decouple the evolutionary fates of linked mutations, leading to *both* a decrease in degeneration rate *and* an increase in adaptation rate. Our paper emphasizes the latter, that palindromes may exist to accelerate adaptation by increasing the potential targets and fixation rates of incoming beneficial mutations. This hypothesis helps reconcile two enigmatic features of the "palindromes as protectors" view: (1) genes that are not located in palindromes have been retained under purifying selection for tens of millions of years, and (2) under models that only consider deleterious mutations, gene conversion benefits duplicate gene maintenance but not initial fixation. We conclude by looking at ways to test the hypothesis that palindromes enhance the rate of adaptive evolution of Y-linked genes and whether this effect can be extended to palindromes on other chromosomes.

1. Evolution of Sex-Limited Chromosomes

1.1. Evolution of Sex-Limited Chromosomes-Theory. Sex-limited chromosomes are unique in that they often have a small, peculiar gene content [1–5]. Classically, sex chromosomes are thought to originate from a pair of autosomes in three phases: (1) one homolog acquires a sex determining factor; (2) selection favors linkage between sexually antagonistic variants and the sex determination factor, thereby reducing or eliminating regional recombination; (3) the forces of mutation, drift, and selection in regions of low recombination lead to rapid gene loss (Figure 1; [6–8]). To the extent that this model is true, *positive selection* for reduced recombination (e.g., selection to fix chromosomal inversions and/or other modifiers of recombination [9, 10]) is responsible for providing the spark that ignites proto-Y chromosome morphological differentiation from the proto-X chromosome (Figure 1).

In the third phase of sex chromosome differentiation, three different processes—Muller's ratchet, background selection, and genetic hitchhiking—may contribute to degeneration of the Y (or W) chromosome once recombination is reduced in all or part of the nascent sex-specific chromosome [11–13]. These three mechanisms are instances of the general Hill-Robertson effect that describes the reduction in the efficiency of selection in the presence of segregating mutations under selection when recombination is either absent or reduced [14–16]. Muller's ratchet will operate when deleterious mutations occur, and the class of Y chromosomes with the least deleterious mutations is lost from the population by drift and cannot be recovered because of the lack of recombination. Background selection will lead to the fixation of weakly deleterious mutations due to the reduction in effective population size brought about by the selection against strongly deleterious mutations in regions with reduced recombination. Genetic hitchhiking will occur when a beneficial mutation drags along the fixation of deleterious mutations in the nonrecombining region of the Y chromosome. The long-term consequences for Y chromosome fitness are very different for each of these processes (Figure 2). The first two processes make the fitness of Y chromosomes worse on average as time goes by while genetic hitchhiking improves the Y on average. Interestingly, these processes have different likelihood of operating at different times in the process of Y chromosome differentiation. Muller's ratchet and background selection are predicted to

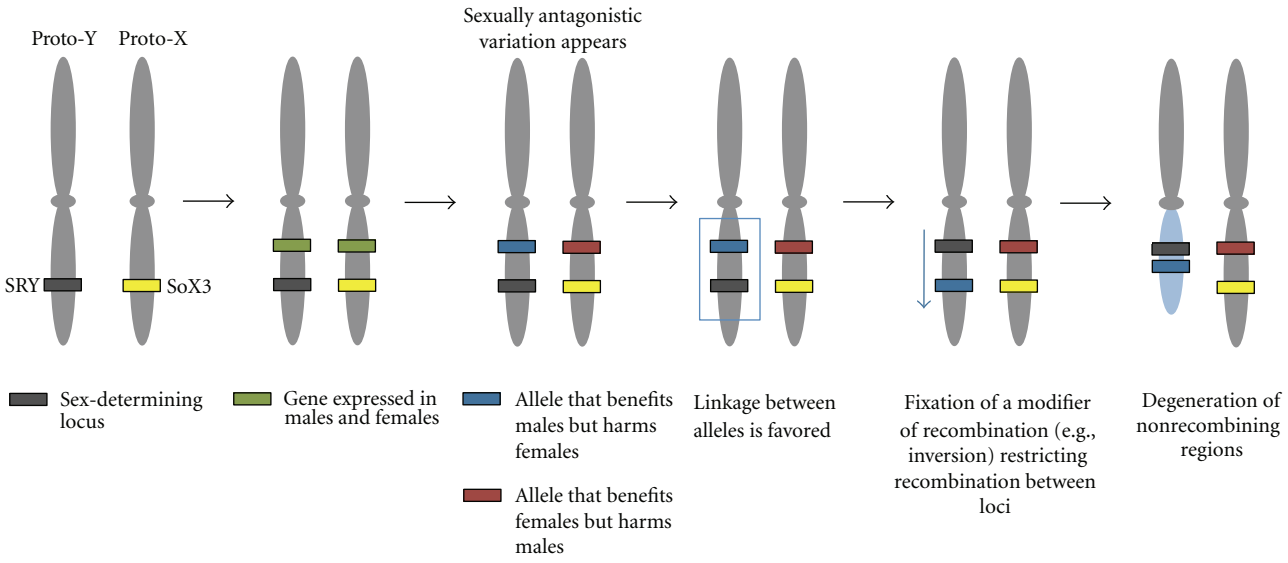


Figure 1: The model of sex chromosome evolution. Close linkage between sexually antagonistic variation and the sex-determining gene has been proposed to start Y chromosome morphological differentiation from the X chromosome.

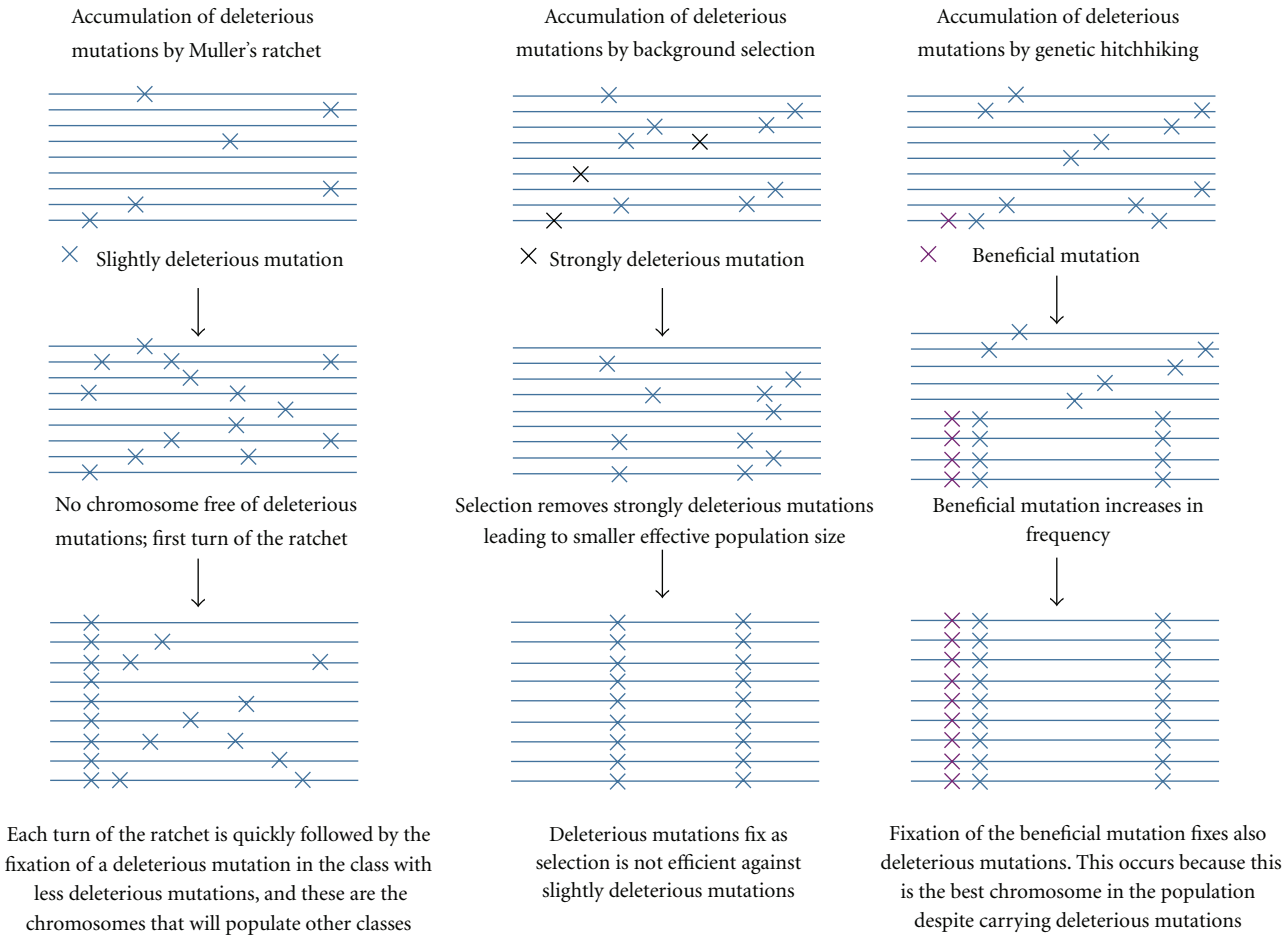


Figure 2: (a) The three processes that lead to degeneration of the Y chromosome: Muller's ratchet (see [109] for details of how every turn of the ratchet is followed by fixation of a deleterious allele), background selection, and genetic hitchhiking. Only in the case of genetic hitchhiking, the fitness of the Y chromosome increases through time.

be strong degenerating forces when there are many genes on the nonrecombining region of the Y, while genetic hitchhiking will dominate the nonrecombining region of the Y when the genic content is smaller [17, 18]. Thus, genetic hitchhiking is predicted to be the dominant process on older Y chromosomes that have small gene content. In addition, since the cessation of recombination often occurs in strata [10, 19, 20] and consequently a limited number of genes are involved in each bout of degeneration, Bachtrog [17] proposed that genetic hitchhiking might actually play an important role throughout the chromosome's degeneration.

The relative contribution of the above mechanisms to the evolution of the Y chromosome is difficult to assess as neither the rate of mutations under selection nor the distribution of their fitness effects are well known. In addition, the fitness effects of Y-linked mutations might differ depending on how and when Y inactivation and dosage-compensation evolve. X- and Y-linked loci are expected to differ in fitness either because more beneficial mutations can fix on the X than on the Y [21] or because more deleterious mutations fix on the Y than on the X [13]. This will generate selection pressure for transcriptional downregulation of the Y-linked loci and upregulation of the X-linked loci. If such Y inactivation and dosage compensation occur, subsequent mutations in dosage compensated regions may not be deleterious anymore but rather neutral or sometimes beneficial (i.e., if they facilitate the silencing of the maladapted Y-linked genes). The fraction of Y-linked genes that are now neutral or whose inactivation can be beneficial could be large if dosage compensation occurs "block by block" [22]. Recent studies of systems with young sex chromosomes suggest that gene silencing might be an early step in Y chromosome degeneration [23] and dosage compensation may evolve concomitantly with Y chromosome degeneration [24]. That is, much of the Y degeneration might be a neutral or even adaptive process [18, 21, 22].

1.2. Evolution of Sex-Limited Chromosomes—Data. In humans, most of the approximately 1,000 genes present on the X chromosome are absent from the Y chromosome [25]. Apart from the 29 genes that are present within the recombining regions of the X and Y chromosomes (i.e., the pseudoautosomal regions), the male specific region of the Y chromosome (MSY) contains only 19 genes that can be traced to ancestral autosomes (Table 1; [25, 26]). Similar extensive gene loss from the Y is seen in other mammalian lineages, where the number of extant genes that originated on the protosex chromosomes does not exceed ~20 genes [27, 28]. Independently evolved sex chromosomes in other taxa, including birds, snakes, plants, and insects, followed similar patterns of gene loss after recombination ceased on the sex-limited chromosome (Y or W; [19, 20, 29–31]).

Interestingly, the pattern of gene loss in humans and other lineages (including species where females are the heterogametic sex) suggests that phases 2 and 3 of Y (or W) chromosome evolution recur multiple times, generating a series of strata with different levels of degeneration relative to the X (or Z) [10, 19, 20, 31]. The pattern of rapid gene loss following stratum formation and subsequent stabilization of gene content [26] is consistent with temporal dynamics of the evolutionary forces implicated in the degeneration of the Y chromosome [30, 32]. Furthermore, among the major lineages of birds, the same pair of ancestral autosomes independently proceeded through phases 2 and 3 [33]. These two patterns support the idea that, once the process of sex chromosome differentiation initiates, the presence of strong sexually antagonistic variation will drive the chromosomes through similar steps and to convergent ends in independent lineages.

Recent sequencing of primate Y chromosomes has uncovered what might be called the 4th phase of sex chromosome evolution characterized by gene preservation and Y chromosome specialization through acquisition and amplification of genes with testis expression [26, 34, 35]. Inter- and intraspecific sequence comparisons suggest that purifying selection on the Y chromosome is strong enough to prevent the full decay of genes that originated on the protosex chromosomes. Analyses of gene loss in three primates—human, chimpanzee, and rhesus macaque—indicates that lineage-specific gene losses in the human and rhesus MSY are restricted to the stratum that most recently ceased to recombine with the X, while the few genes in older strata (1–4) have been conserved by purifying selection for more than 25 million years [26, 36]. While no lineage-specific gene losses were detected in gorilla, the chimpanzee MSY has lost 5 ancestral genes since splitting from the human lineage ~6 million years ago being the only lineage that shows instability among primates thus far (Table 1; [26, 36, 37]). Conservation of gene content is also found outside primates. At least 6 Y-linked X-degenerate genes specific to marsupial lineage have been preserved for ~50 million years [38]. Polymorphism data also supports the efficient retention of some genes by purifying selection. Within human populations, analysis of sequence variation in 16 Y-linked single-copy X-degenerate genes indicates efficient purifying selection, finding little difference in the protein sequence among males [39].

In addition to the preservation of the X-degenerate genes, Y chromosomes show clear signs of differentiation through lineage-specific gene gain. In humans, 80% of genes on the MSY (60 out of 78) are members of 9 gene families (Table 1). Some of these families originated by duplication of X-degenerate genes (TSPY, RBMY, and HSFY), but other families arose through gene duplication and subsequent amplification of autosomal genes (DAZ and CDY) while others possibly originated de novo on the Y (PRY and BPY2) as no X-linked or autosomal homologues have been identified [26, 34, 40–43]. Two single-copy genes (TGIF2LY and PCDH11Y) were also recently acquired by the human Y via translocation of 3.4 Mb from the X chromosome [34]. New genes are added to the Y chromosome in other mammals as well. For example, studies of MSY in horse identified 17 novel and acquired genes that are also present on the donkey Y but are absent in other mammalian Y chromosomes [28]. New gene families have been independently gained on the bovine and carnivore MSY through translocation of autosomal gene blocks followed by amplification [44–46]. In *Drosophila melanogaster*, gene acquisition plays

Table 1: Copy number and expression profiles of MSY genes in primates.

Origin	Gene	Rhesus	Human	Chimp	X-homolog
Stratum			Ancestral genes		
1	SRY	1 (?)	1 (pT)	1 (T + S)	Yes
	RBMY	1 (T)	6 (T)	6 (pT)	Yes
	RPS4Y1	1 (T+P)	1 (B)	1 (B)	Yes
	RPS4Y2	1 (T)	1 (B)	1 (B)	Yes
	HSFY	3 (T)	2 (T)	—	Yes
2	KDM5D	1 (B)	1 (B)	1 (B)	Yes
	TSPY	5 (T)	35 (T)	6 (pT + L + Li)	Yes
3	ZFY	1 (B)	1 (B)	1 (B)	Yes
	DDX3Y	1 (B)	1 (B)	1 (B)	Yes
	UTY	1 (B)	1 (B)	1 (B)	Yes
	EIF1AY	1 (B)	1 (B)	1 (B)	Yes
	CYorf15A	1 (B)	1 (B)	1 (B)	Yes
	CYorf15B	1 (B)	1 (B)	ps (B)	Yes
	USP9Y	1 (B)	1 (B)	ps (B)	Yes
	TMSB4Y	1 (B)	1 (B)	ps	Yes
4	AMELY	1 (?)	1 (B)	1 (?)	Yes
	NLGN4Y	1 (B)	1 (B)	1 (B)	Yes
	TBL1Y	1 (B)	1 (B)	ps (B)	Yes
5	PRKY	1 (B)	1 (B)	1 (B)	Yes
	MXRA5Y	1 (B)	ps	ps	Yes
	Total:	**26**	**59**	**24**	
			Added genes		
A-transposed	DAZ	2 (T)	4 (T)	4 (T)	No
A-retroposed	CDY	2 (T)	4 (T)	5 (pT)	No
	XKRY	1 (B)	2 (T)	ps (pT)	No
	BPY2	—	3 (T)	2 (T)	No
	PRY	—	2 (T)	—	No
	VCY	—	2 (T)	2 (?)	Yes
X-transposed	PCDH11Y	—	1 (Br)	—	Yes
X-transposed	TGIF2LY	—	1 (T)	—	Yes
	Total:	**5**	**19**	**13**	
	Total AG[a]	**12**	**60**	**25**	
	Total AF[b]	**4**	**9**	**6**	
	Grand Total:	**31**	**78**	**37**	

Modified from [26]. Expression data from [26, 34, 36, 49]. T: testis, pT: predominantly testis, B: broad, Br: brain, P: prostate, S: spleen, L: lung, Li: liver, ?: not known. Absent gene (—), pseudogene (ps). [a]Ampliconic genes; [b]Ampliconic families.

a major role in the evolution of the Y chromosome as all protein-coding genes (<20) result from duplication of autosomal genes [47, 48]. These findings favor the view that Y chromosome gene content is not merely characterized by degeneration. Rather, it is much more dynamic than previously recognized, having evolutionary stages that vary dramatically in gene birth and death rates.

It is notable that the vast majority of the genes that have been amplified or acquired on the Y chromosome in different lineages are expressed predominantly or exclusively in testis and have spermatogenesis-related functions [26, 28, 34–36, 44, 46, 49–55]. In mammals, the testis-specific expression of amplified and acquired genes contrasts the much broader expression profile of single-copy X-degenerate genes (see Table 1 for primate examples). Such acquisition and retention of different testis-specific genes in different lineages suggest that specialization for male-fertility functions is a driver of Y chromosome evolution.

1.3. Models for the 4th Phase of Y Evolution. There are several models invoking positive selection to explain Y chromosome gene content that are consistent with Y-linked genes being a lasting and important determinant of male fitness. First, phase 2 of the classical model introduced above suggests that positive selection favoring tight linkage between sex

determining loci and those with sexually antagonistic variation starts the differentiation and degeneration of the Y (Figure 1; [7]). While the classical model explains the emergence of a sex-specific gene (i.e., a sexually antagonistic gene that becomes Y linked) linked to sex determination factors on the Y, sexual antagonism also provides a framework to explain recruitment of new Y-linked genes. For instance, intralocus sexual conflict on autosomes can be resolved by duplicating the allele benefiting males onto the Y chromosome [56, 57]. A likely example of this model includes sexually selected loci in guppies. Selection in male guppies to make them more attractive to females has been proposed to be so strong that it leads to the Y-linkage of traits that are likely costly in females [58] although it is unclear whether these genes moved to the Y chromosome or they evolved *in situ*. An alternative resolution of intralocus sexual antagonism that more drastically reshuffles sex chromosome gene content is achieved by invasion of a new male determining gene linked to the male benefiting allele as proposed for a cichlid fish [59]. This particular example follows a previously proposed model related to the resolution of sexual antagonism that involves the turnover of sex determination genes [60].

In addition to the role of sexual antagonism, strong epistasis between Y-linked and X- or autosomal genes could also impact Y chromosome gene content [61]. Beneficial Y-X or Y-autosome combinations will experience positive selection for genomic rearrangements that result in tight linkage. In the case of Y-X epistasis, this could favor the spread of nonrecombining regions as observed in Y chromosomes with multiple strata of differentiation. Y-autosome epistasis would favor duplications of the autosomal genes to the nonrecombining portion of the Y where their linkage would no longer be disrupted. Consistent with this model, the Y chromosome of *D. melanogaster* shows strong epistatic variance for fitness [61]. The Y chromosome that is the best in one genetic background is worst in another contributing nothing to additive genetic variation for fitness in males [61]. Furthermore, introgressions of *Drosophila* Y chromosomes into different conspecific [62] or heterospecific [63] genomic backgrounds result in misexpression of more than 100 X- and autosomal genes.

Additional models that may govern Y chromosome gene content likely include efficient sex-limited selection, selfish genetic elements, and subfunctionalization. The fact that most Y-linked genes have rapidly evolving sex-specific functions (e.g., only expressed in testes) is a clear indication that sex-limited selection on a haploid chromosome is a large determinant of what remains and/or is duplicated to the Y [64]. The evolution of selfish elements could also explain the Y linkage of some genes. One kind of selfish elements is segregation distorters (i.e., selfish systems that increase in frequency because they bias their transmission to the next generation). In *Drosophila*, one RNA gene family on the Y chromosome (*suppressor of stellate*) has been proposed to be a gene that acquired Y-linkage under positive selection as it acts to suppress *Stellate* expression that has been proposed to be a X-Y selfish segregation distorter [65]. Finally, Koerich and colleagues (2008) also considered that neutral duplication of a testis gene followed by chance loss of the parent copy or the neutral duplication of a broadly expressed gene followed by subfunctionalization could explain some of the gene gains observed for the *Drosophila* Y chromosome [47].

Several studies of DNA sequence divergence have suggested the action of positive selection in some Y(W) genes. Gerrard and Filatov [66] studied three genes in 12 mammalian species and concluded that two of them (*USP9Y* and *UTY*) evolved under positive selection. The basis for the selection of these genes is not clear as both of them are broadly transcribed among tissues (Table 1; [66, 67]). Another example, *DAZ*, might also have evolved under positive selection in humans and in this case is easier to explain owing to testes specific expression of the gene [68]. Signatures of positive selection have also been found in female specific W genes. For instance, the *HINTW* gene is under positive selection on the W chromosome of birds [69] and has sex-specific functions in the developing female urogenital tract and ovaries. In plants, Marais et al. [70] analyzed seven Y-linked genes in *Silene latifolia* and revealed patterns of divergence in two of these genes (*SlssY* and *DD44Y*) that are consistent with positive selection.

In addition to analyses of substitution patterns across taxa, polymorphism data has also been analyzed for a few Y chromosome systems. A signature consistent with ongoing positive selection was found on the neo-Y chromosome of *D. miranda* [71]. However, it is likely that adaptation is, at most, restricted to a few loci and that the faster accumulation of amino acid substitutions and unpreferred codons on the neo-Y compared to neo-X chromosome is the result of reduced efficiency of purifying selection on the nonrecombining neo-Y [32, 72]. Particular models of background selection that include interference between strongly negatively selected sites are also compatible with this polymorphism data [73]. So a population analysis of the fitness effects of these chromosomes is needed to distinguish among the models. In *Silene,* polymorphism is reduced on the Y but it is unclear whether background selection or genetic hitchhiking with beneficial mutations (or both) contribute to the observed reduction [74, 75]. Human Y polymorphism data reveals very low levels of polymorphism on the Y and have been taken as evidence of the small effective population size that accompanies nonrecombining Y chromosome degeneration [76]. However, a more detailed look reveals that reduced variation is mainly due to gene conversion in ampliconic regions [76].

In sum, positive selection may not only be the spark that ignites Y (or W) chromosome differentiation in phase 1 but also continues to influence Y (or W) chromosome evolution in phases 2 through 4, leading to degeneration of some genes due to genetic hitchhiking and possibly the addition of others by duplications and translocations. In the following, we propose that palindromes and amplicons that seem to originate on the Y and W chromosomes late in the process of sex chromosome differentiation might be important chromosomal mutations whose primary role could be to increase the rate of incoming beneficial mutations and accelerate adaptation in old sex chromosomes.

1.4. Y(W) Palindromes. The assembly of the Y chromosome of humans, chimpanzee, and rhesus macaque revealed surprising sequence heterogeneity of the Y chromosome with a substantial portion of these chromosomes occupied by large repeat units, referred to as amplicons [26, 34, 35]. Ampliconic sequences can be organized as tandem arrays as well as palindromes (inverted repeats). The amplicons are extensive, comprising 45% (10.2 Mb) of the euchromatic portion of the MSY in humans, 57% (14.7 Mb) in chimpanzee, and 5% (0.5 Mb) in rhesus macaque [26, 34, 35]. Compared to X-degenerate sequences (i.e., orthologous single-copy X-Y sequences), ampliconic sequences have a higher density of genes and pseudogenes but markedly lower density of retrotransposable elements [34, 35].

Palindromes are the most impressive feature of the Y chromosome. These structures are made up of inverted repeats (palindrome arms) separated by a nonduplicated spacer. The length of each arm varies among the three primate species, ranging between 73 kb in rhesus and 344 kb in humans on average [26]. Rhesus macaque has 3 palindromes that occupy about 87% (437 kb) of the ampliconic region, while chimpanzee and humans have 19 and 8 palindromes that make up about 50% (7.5 Mb) and 54% (5.5 Mb) of the ampliconic region, respectively. Twelve of 19 palindromes are specific to chimpanzee lineage [35]. Two of 3 palindromes in rhesus macaque are also found in humans [26] revealing that some palindromes have endured for at least 25 millions of years.

A striking feature of ampliconic MSY regions is the high intrachromosomal sequence identity. In humans, 60% (6.1 Mb) of ampliconic sequences (including all 8 palindromes) show 99.9% or greater intrachromosomal sequence identity. The sequence comparison of the 4 palindromes between humans and chimpanzee revealed that such high sequence identity is maintained by ongoing gene conversion between the arms of the palindromes. Sequence divergence between orthologous palindrome arms was found to be 1.44%, while arm-to-arm divergence within each species is much lower, 0.021% and 0.028% for human and chimpanzee palindromes, respectively (Figure 3; [77]). The rate of gene conversion required to maintain the observed level of sequence identity is estimated to be 2.2×10^{-4} per site per generation which means that ~600 duplicated nucleotides have undergone gene conversion between palindrome arms every generation [77].

Gene conversion is a standard type of recombination but, unlike crossing over (Figure 4(a) (a1) and Figure 4(b) (b1)), it involves the nonreciprocal transfer of information. This is shown in the central panels of Figure 4 (see Figure 4(a) (a2) and Figure 4(b) (b2); [78]). Gene conversion was first observed as an outcome of allelic recombination (between orthologous sequences of homologous chromosomes) but is now widely recognized as a mechanism of genetic transfer between paralogous sequences ([79] and references therein). The extent of gene conversion is influenced by a number of factors, including sequence identity, physical proximity, and the length of the identical regions [80]. Discovery of ampliconic regions on the primate Y arranged as palindromes and tandem arrays that are expected to promote gene conversion largely changed the view of the Y chromosome from a vestigial part of the genome to a vital chromosome that is capable of escaping the debilitating consequences of the absence of recombination [34, 77].

So far, only a few cases of gene conversion on sex-limited chromosomes have been documented outside primates. In the European rabbit, gene conversion occurs between the 23 kb long palindrome arms that house the SRY genes and are 99.94% identical [81]. In galliform birds, multiple tandem copies of W-linked *HINTW* genes undergo gene conversion maintaining high sequence identity between copies within each of the four species studied [82]. A W-linked palindrome in white-throated sparrow shows signs of conversion in a region containing a portion of *CHD1W* intron [83]. Ampliconic regions with large tandem and palindrome-like repeats containing active genes and pseudogenes have been found on the bovine MSY where sequence identity within repeat families ranges from 99.4% to 99.7% [46]. Preliminary analysis of the mouse Y chromosome sequence also identifies multiple palindromes and large repeat units [84]. Whether or not gene conversion is operating in these species awaits further analyses. Despite the current scarcity of information about the detailed organization of most Y chromosomes, data are rapidly accumulating and it is becoming increasingly clear that gene duplication is a common feature of differentiated/old Y chromosomes [28, 46, 51, 84, 85], and we anticipate that more cases of gene conversion will be discovered.

In addition to the possibility of gene conversion, ampliconic structures create an opportunity for ectopic crossing over. For genes located in palindromes, crossing over can occur between gene copies on the same chromatid or between different copies on different sister chromatids. It has been observed that crossover events that involve paralogs from different sister chromatids (Figure 4(a) (a3) and Figure 4(b) (b3)) lead to isodicentric and acentric chromosomes and can result in gene loss and gain [78]. This process may underlie several disease phenotypes in humans including spermatogenesis failure, sex reversal, and Turner syndrome which are associated with inheritance of a rearranged Y and gene loss [78]. Although ectopic recombination does not always lead to reduced male fertility [53, 78, 86, 87], fitness-reducing consequences of ampliconic structure are likely to be frequent enough to impose an upper limit on the number of duplicates that can be maintained in a Y chromosome as a higher number is expected to lead to more ectopic crossovers [88]. Given that palindromes and tandem arrays are fixed in a population and are maintained for long periods of time, the benefits associated with gene duplications must be large enough to offset their deleterious effects.

2. Why Chromosome Palindromes?

The available data suggest that the most important consequence of ampliconic structure relevant to the evolution of Y chromosomes is the opportunity for gene conversion. Some palindromes are very complex in structure and gene content

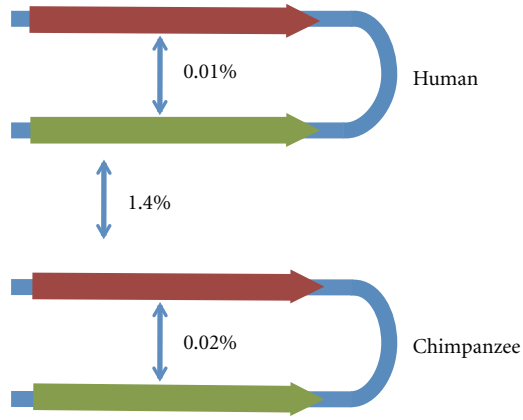


Figure 3: Concerted evolution by gene conversion in primate palindrome 6 [77]: low divergence between paralogs within a lineage but "normal" divergence between orthologs between lineages.

[35, 77, 89] and are maintained for long periods of time (i.e., >25 My in some instances). Palindromic regions also seem to be under purifying selection as it has been observed that transposable elements and retroviruses are removed from the palindromic regions [26, 34]. High levels of intrachromosomal sequence identity are consistent with high rates of ongoing gene conversion within Y(W) chromosome palindromes [26, 35, 77, 81, 82]. In fact, the rate of gene conversion in the palindromes of the human Y chromosome is three orders of magnitude higher than that of human paralogs that are similarly arranged but located elsewhere in the genome [90]. This observation supports the view that the evolution of high levels of gene conversion on the Y chromosome has been favored by selection [76]. Thus, the reason for palindrome emergence and maintenance should be sought in understanding the benefits of gene conversion for the evolution of gene families on the Y chromosome.

The consequences of gene conversion for the evolution of Y-linked duplicates have been recently investigated using analytical and simulation approaches [76, 88]. Both works considered the evolution of gene duplicates in the presence of deleterious mutations and examined how gene conversion affects the probability of fixation of new duplicates and preservation of duplicated genes once they are fixed. Both studies find that gene conversion does not enhance the probability of duplicate fixation, and, unless there are direct fitness benefits of having a duplicate (e.g., increase in dose), the fixation of Y-linked duplicates is expected to occur by drift [76, 88]. However, once duplicates are fixed, gene conversion can effectively counteract the degeneration of the Y chromosome. Gene conversion exerts its effect through regeneration of the least-mutated haplotype allowing for more efficient removal of deleterious mutations and reducing the chance that the least-mutated class will be lost by drift. These benefits of gene conversion are higher when the rate of gene conversion and the total mutation rate are high and the fitness effects of deleterious mutations are small [88]. The advantage of gene conversion can be further extended to cases where the deleterious effect of a mutation in one copy is masked by another functional copy. In this situation, selection is inefficient in removing these mutations (effectively recessive deleterious mutations). Gene conversion can expose such mutations to selection, preventing accumulation of deleterious mutations that would otherwise eventually lead to the loss of the functional copy [88]. High rates of gene conversion observed on the human Y palindromes that maintain nearly identical copies [77] might have been favored to allow efficient selection against recessive deleterious mutations.

The results of the above studies highlight the beneficial effect of gene conversion on the removal of deleterious mutations (i.e., protection against further degeneration). But gene conversion between members of a gene family can also have the complementary effect of increasing the fixation rate of beneficial mutations. The effect of gene conversion on the rate of adaptive evolution in gene families has been investigated by Mano and Innan [91]. Using analytical and simulation approaches the authors studied the dynamics of a beneficial mutation that initially occurs in one member of the gene family and eventually spreads to all members through gene conversion reaching fixation. They show that gene conversion increases the effective population size by a factor that is equal to the size of the gene family. This leads to a higher fixation rate of beneficial mutations and a lower fixation rate of deleterious mutations in multigene families [91]. This result holds in cases with or without crossing over and should be applicable to gene families on the Y chromosome [92] although the effects are expected to be smaller due to reduced population size and the haploid nature of the Y.

Mano and Innan's model [91] might provide a better fit to Y chromosome data than models that consider the effect of gene conversion in the presence of deleterious mutations only. The common feature of the Y chromosome across different species is the peculiar composition of its gene content with respect to function and expression. With few exceptions, genes can be divided into two broad categories: there are single-copy genes that are expressed broadly and multicopy genes that are expressed predominantly in testis and have functions related to male fertility. Furthermore, testis-expressed gene copies within gene families share high sequence identity as a result of intrachromosomal gene

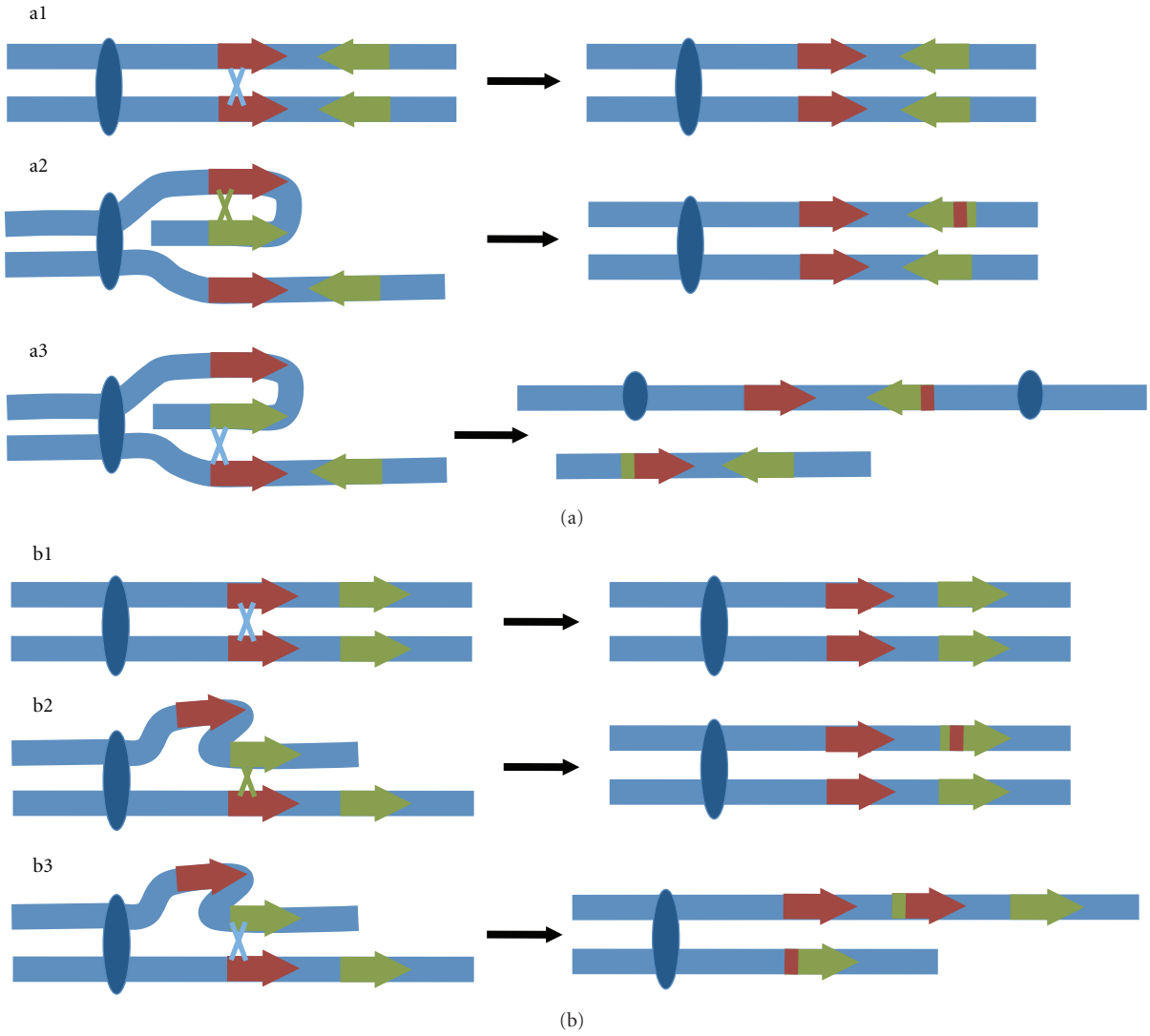


Figure 4: Effects of crossovers (blue lines) and gene conversion (green lines) in Y(W) palindromes (a) and tandem arrays (b). No effect of a crossover is observed if it occurs within the same gene between sister chromatids (a1 and b2). Gene conversion (nonreciprocal transfer of information) is observed if it occurs between different genes within the same palindrome or between tandem duplicates (a2 and b2). Acentric and dicentric chromosomes are produced from a crossover between different genes in palindromes located in sister chromatids (a3). Acentric chromosomes will not segregate properly, and dicentric chromosomes will likely break and lose information when they are pulled to opposite cell poles [78]. Gene gains and losses are produced from a crossover between different duplicates within array located in sister chromatids (b3).

conversion that occurs within ampliconic regions. Given all the data that has accumulated over the years demonstrating adaptive evolution of genes with reproduction-related functions [93–98], it would not be surprising if the same pattern was found for Y-linked spermatogenesis genes. What differentiates a nontestis gene from a testis gene is the fraction of sites that can receive beneficial mutations. When in a single-copy state, the adaptive evolution of a testis gene is compromised by linkage to deleterious mutations [99]. When duplicated, gene conversion allows for more efficient removal of deleterious mutations and the beneficial mutation can now occur on a chromosome with fewer deleterious mutations [76, 88, 91]. While the fixation rate of beneficial mutations that occur anywhere on the Y is expected to increase in the presence of gene conversion, the adaptive evolution of duplicated testis genes is further accelerated by gene conversion that facilitates the spread of beneficial mutations between paralogs as described by Mano and Innan [91]. In this scenario, ampliconic/palindromic structure is maintained because it allows rapid adaptive evolution of testis genes.

In the absence of beneficial mutations fixation of duplicates occurs by drift unless the duplicate has an immediate fitness benefit associated with the increased dosage of gene

product. The effect of gene conversion on the fixation of Y-linked duplicates in the presence of both beneficial and deleterious mutations has not been modeled, but it is interesting to note that gene conversion can slow down the loss of redundant duplicates [88], thereby increasing the time period during which functional duplicates are segregating in a population. This effect of gene conversion is expected to increase the chance of duplicate fixation where the direct fitness benefit is supplied by the beneficial mutations that improve gene function. The differences in the target size for beneficial mutations between nontestis and testis genes may help explain the fixation of duplications containing testis genes. Let us consider first the case of a testis gene. Duplication of a testis gene would immediately double the rate of incoming beneficial mutations. If a beneficial mutation occurs while a duplicate is segregating, gene conversion is expected to enhance the fixation of the duplicate by spreading the beneficial mutation among paralogs and by freeing beneficial mutation from its association with deleterious mutations within the ampliconic region thereby increasing the fitness of the Y chromosome that carries the duplicated genes. Duplication here can be viewed as a modifier of recombination that is under direct positive selection when a beneficial mutation occurs in one of the copies. While a duplication event will also immediately double the rate of deleterious mutations, efficient selection on a haploid chromosome and a high rate of gene conversion are expected to efficiently remove them [39, 76, 88, 91]. In the case of X-degenerate nontestis genes, mutations are less likely to have a beneficial effect as they are broadly expressed and gene conversion would only bring the potential benefit of a reduced rate of fixation of deleterious mutations. This beneficial effect might not be enough to offset the deleterious effects of ectopic crossing over between gene duplications [78, 88].

It has been also proposed that ampliconic regions have evolved gradually as the fixation of large duplications is extremely unlikely when the benefits of gene conversion associated only with the removal of deleterious mutations are considered [76]. However, the analyses of ampliconic sequence in primates suggest that some of the steps in the evolution of palindromes may involve duplication of large regions [89, 100]. Furthermore, new genes are not always acquired gene by gene; in bovine MSY, a new testis gene family has been acquired by "gene block" transposition from an autosome [46]. The proposed-above dependence of the duplicate fixation on the presence of gene conversion and adaptive mutations suggested for the testis-specific genes might also allow for fixation of large-scale duplications.

A prediction of the model of Mano and Innan [91] is that the rate of evolution of multicopy genes located in regions undergoing gene conversion (palindromes) should be higher than the rate of evolution of single-copy genes if they are evolving under positive selection [91, 92]. Alternatively, if adaptive evolution in testis genes is rare, the main consequence of gene conversion (and consequently, palindrome presence) would be increased efficiency of purifying selection, leading to reduced rate of evolution in multicopy genes compared to single-copy genes. This comparison is analogous to that between genes in regions of high and low recombination [97]. Comparing human and rhesus macaque Y-linked genes (data from [26]), genes in ampliconic regions show accelerated rate of evolution, with higher ratio of nonsynonymous to synonymous substitution rates compared to single-copy X-degenerate genes (Figure 5). This result might be interpreted as indicative of adaptive evolution in testis genes. However, given differences in expression profiles between the two classes of genes and the fact that rates of protein evolution correlate negatively with expression levels and not only with expression breadth [101, 102], further analyses are needed to remove the effect of gene expression on the rates of evolution. A more adequate way to test the model of Mano and Innan [91] is to look for an acceleration or deceleration of the rate of evolution in genes with the same function and expression by comparing lineages where gene is present in many copies to the lineages where the gene still remains a single-copy gene (Figure 6). Among Y-linked genes in primates (Table 1), there is one gene (RBMY) that at first glance satisfies these conditions. RBMY is a single-copy gene in rhesus macaque but has 6 copies in humans and chimps. However, the single-copy status of RBMY is a derived state as RBMY is present in multiple copies in nonprimate species [103]. We should therefore wait until data from more species becomes available to directly test the effects of gene conversion on the adaptive evolution of Y-linked gene families.

More generally, it would be expected that fast-evolving genes should be members of gene families that undergo high levels of gene conversion because ampliconic structures accelerate the rate of adaptive evolution by permitting high levels of gene conversion. A whole genome analysis of palindromes in the human genome revealed that palindromes are not only overrepresented on the Y chromosome but also overrepresented on the X chromosome, and among those palindromes with >99% arm-to-arm identity, most contain genes with testis expression [104]. The mouse X chromosome also contains many genes showing postmeiotic expression in testis that are part of amplicons including some palindromes [105]. It has been suggested that the role of these palindromes on the sex chromosomes might be to prevent meiotic sex chromosome inactivation allowing the expression of spermatogenesis genes that reside in palindromes [104]. However, recent discovery of Z-linked amplicons with testis genes in chicken [106] argues against the role of palindromes in escaping gene silencing since it is typically the heterogametic sex that undergoes meiotic sex chromosome inactivation [107], but male chickens are ZZ. The rate of evolution or gene conversion of these testis genes has not been studied, but it is notable that amplicons are enriched for the kinds of genes that frequently evolve under positive selection [93–98]. In other instances, the genes in palindromes have functions that might be under positive selection in both sexes. These are patterns that were observed in palindromes in worms for genes that were speculated to act as antimicrobial peptides [108]. There is therefore a need for systematic studies of genes in amplicons in association with the rates of evolution in the regions undergoing gene

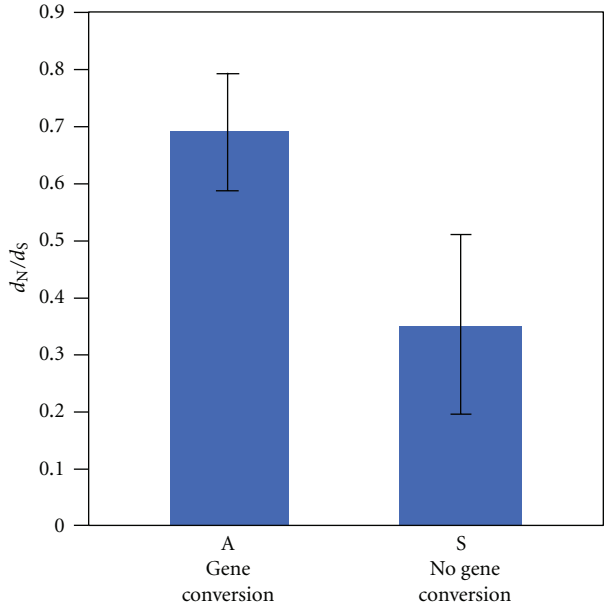


Figure 5: d_N/d_S comparison between ampliconic (A) and single-copy (S) genes in the human-rhesus Y chromosome ([26]; Mann-Whitney test, $Z = 3.75$, $P = 0.0002$). If a gene is ampliconic in one species and not in another, it was counted as ampliconic in this comparison. Error bars indicate 95% confidence interval.

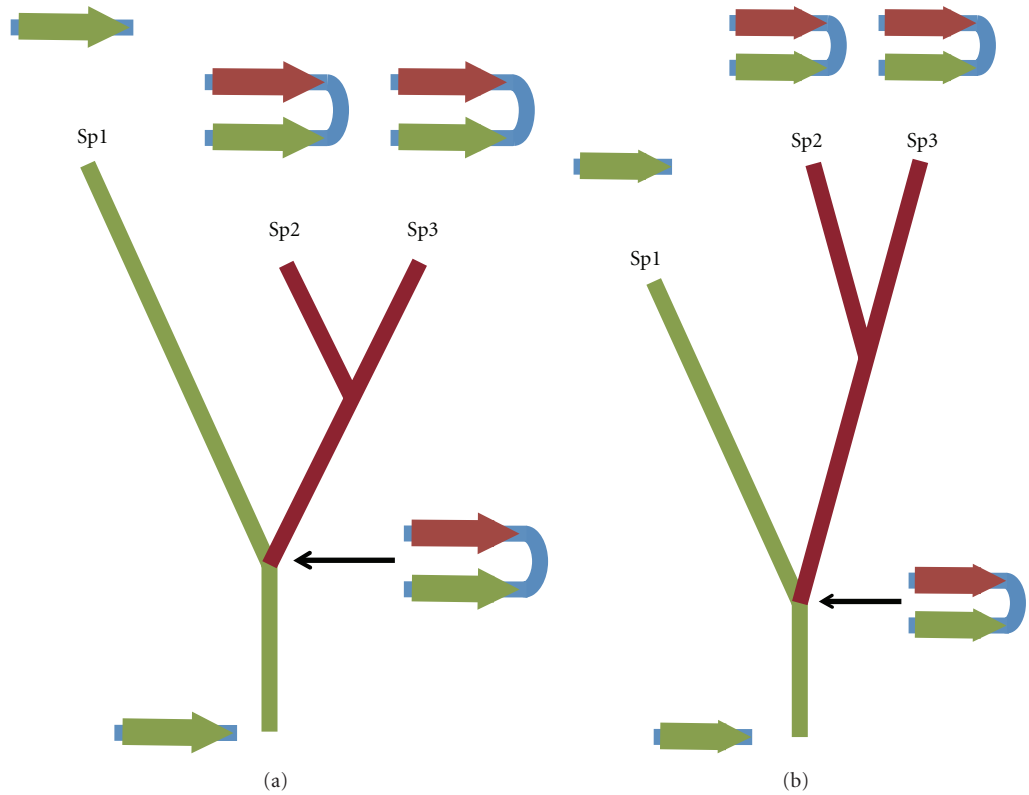


Figure 6: Two models of the evolution of gene families on the Y chromosome under concerted evolution. The length of the branches shown is proportional to d_N/d_S ratio. (a) When gene conversion does not increase the fixation rate of beneficial mutations in multigene families, the rate of evolution is reduced compared to that of single-copy gene because gene conversion is expected to reduce the fixation rate of deleterious mutations. (b) When gene conversion increases the fixation rate of beneficial mutations in multigene families, the rate of evolution is higher compared to single-copy genes.

conversion in order to evaluate the relative contribution of gene conversion to patterns of gene preservation and adaptation.

Acknowledgments

This research was supported by Grants no. R01GM071813 and R01GM065414 from National Institutes of Health, USA, to E. Betrán and J. P. Demuth, respectively. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] D. Charlesworth, "Plant sex determination and sex chromosomes," *Heredity*, vol. 88, no. 2, pp. 94–101, 2002.

[2] D. Charlesworth, "Plant sex chromosomes," *Genome Dynamics*, vol. 4, pp. 83–94, 2008.

[3] M. Schartl, "Sex chromosome evolution in non-mammalian vertebrates," *Current Opinion in Genetics and Development*, vol. 14, no. 6, pp. 634–641, 2004.

[4] J. A. Marshall Graves, "Weird animal genomes and the evolution of vertebrate sex and sex chromosomes," *Annual Review of Genetics*, vol. 42, pp. 565–586, 2008.

[5] E. J. Vallender and B. T. Lahn, "How mammalian sex chromosomes acquired their peculiar gene content," *BioEssays*, vol. 26, no. 2, pp. 159–169, 2004.

[6] I. Marn, M. L. Siegal, and B. S. Baker, "The evolution of dosage-compensation mechanisms," *BioEssays*, vol. 22, no. 12, pp. 1106–1114, 2000.

[7] W. R. Rice, "Evolution of the Y sex chromosome in animals," *BioScience*, vol. 46, no. 5, pp. 331–343, 1996.

[8] S. Ohno, *Sex Chromosome and Sex-Linked Genes*, Springer, Berlin, Germany, 1967.

[9] C. Lemaitre, M. D. Braga, C. Gautier et al., "Footprints of inversions at present and past pseudoautosomal boundaries in human sex chromosomes," *Genome Biology and Evolution*, vol. 1, pp. 56–66, 2009.

[10] B. T. Lahn and D. C. Page, "Four evolutionary strata on the human X chromosome," *Science*, vol. 286, no. 5441, pp. 964–967, 1999.

[11] W. R. Rice, "Genetic hitchhiking and the evolution of reduced genetic activity of the Y sex chromosome," *Genetics*, vol. 116, no. 1, pp. 161–167, 1987.

[12] B. Charlesworth, "Model for evolution of Y chromosomes and dosage compensation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 75, no. 11, pp. 5618–5622, 1978.

[13] B. Charlesworth, "Background selection and patterns of genetic diversity in *Drosophila melanogaster*," *Genetical Research*, vol. 68, no. 2, pp. 131–149, 1996.

[14] W. G. Hill and A. Robertson, "The effect of linkage on limits to artificial selection," *Genetical Research*, vol. 8, no. 3, pp. 269–294, 1966.

[15] J. Felsenstein, "The evolution advantage of recombination," *Genetics*, vol. 78, no. 2, pp. 737–756, 1974.

[16] B. Charlesworth and D. Charlesworth, "The degeneration of Y chromosomes," *Philosophical Transactions of the Royal Society B*, vol. 355, no. 1403, pp. 1563–1572, 2000.

[17] D. Bachtrog, "The temporal dynamics of processes underlying Y chromosome degeneration," *Genetics*, vol. 179, no. 3, pp. 1513–1525, 2008.

[18] J. Engelstädter, "Muller's ratchet and the degeneration of Y chromosomes: a simulation study," *Genetics*, vol. 180, no. 2, pp. 957–967, 2008.

[19] K. Nam and H. Ellegren, "The chicken (*Gallus gallus*) Z chromosome contains at least three nonlinear evolutionary strata," *Genetics*, vol. 180, no. 2, pp. 1131–1136, 2008.

[20] R. Bergero, A. Forrest, E. Kamau, and D. Charlesworth, "Evolutionary strata on the X chromosomes of the dioecious plant *Silene latifolia* evidence from new sex-linked genes," *Genetics*, vol. 175, no. 4, pp. 1945–1954, 2007.

[21] H. A. Orr and Y. Kim, "An adaptive hypothesis for the evolution of the Y chromosome," *Genetics*, vol. 150, no. 4, pp. 1693–1698, 1998.

[22] B. Vicoso and D. Bachtrog, "Progress and prospects toward our understanding of the evolution of dosage compensation," *Chromosome Research*, vol. 17, no. 5, pp. 585–602, 2009.

[23] Q. Zhou and D. Bachtrog, "Chromosome-wide gene silencing initiates Y degeneration in *Drosophila*," *Current Biology*, vol. 22, no. 6, pp. 522–525, 2012.

[24] A. Muyle, N. Zemp, C. Deschamps, S. Mousset, A. Widmer, and G. A. B. Marais, "Rapid de novo evolution of X chromosome dosage compensation in *Silene latifolia*, a plant with young sex chromosomes," *PLoS Biology*, vol. 10, no. 4, Article ID e1001308, 2012.

[25] M. T. Ross, D. V. Grafham, A. J. Coffey et al., "The DNA sequence of the human X chromosome," *Nature*, vol. 434, no. 7031, pp. 325–337, 2005.

[26] J. F. Hughes, H. Skaletsky, L. G. Brown et al., "Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes," *Nature*, vol. 483, no. 7387, pp. 82–86, 2012.

[27] P. D. Waters, M. C. Wallis, and J. A. M. Graves, "Mammalian sex-Origin and evolution of the Y chromosome and SRY," *Seminars in Cell and Developmental Biology*, vol. 18, no. 3, pp. 389–400, 2007.

[28] N. Paria, T. Raudsepp, A. J. Wilkerson et al., "A gene catalogue of the euchromatic male-specific region of the hors chromosome: comparison with human and other mammals," *PLoS ONE*, vol. 6, no. 7, Article ID e21374, 2011.

[29] A. K. Fridolfsson, H. Cheng, N. G. Copeland et al., "Evolution of the avian sex chromosomes from an ancestral pair of autosomes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 14, pp. 8147–8152, 1998.

[30] V. B. Kaiser and B. Charlesworth, "Muller's ratchet and the degeneration of the *Drosophila* miranda neo-Y chromosome," *Genetics*, vol. 185, no. 1, pp. 339–348, 2010.

[31] K. Matsubara, H. Tarui, M. Toriba et al., "Evidence for different origin of sex chromosomes in snakes, birds, and mammals and step-wise differentiation of snake sex chromosomes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 48, pp. 18190–18195, 2006.

[32] D. Bachtrog, E. Hom, K. M. Wong, X. Maside, and P. de Jong, "Genomic degradation of a young Y chromosome in *Drosophila miranda*," *Genome Biology*, vol. 9, no. 2, article R30, 2008.

[33] H. Ellegren and A. Carmichael, "Multiple and independent cessation of recombination between avian sex chromosomes," *Genetics*, vol. 158, no. 1, pp. 325–331, 2001.

[34] H. Skaletsky, T. Kuroda-Kawaguchl, P. J. Minx et al., "The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes," *Nature*, vol. 423, no. 6942, pp. 825–837, 2003.

[35] J. F. Hughes, H. Skaletsky, T. Pyntikova et al., "Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content," *Nature*, vol. 463, no. 7280, pp. 536–539, 2010.

[36] J. F. Hughes, H. Skaletsky, T. Pyntikova et al., "Conservation of Y-linked genes during human evolution revealed by comparative sequencing in chimpanzee," *Nature*, vol. 437, no. 7055, pp. 100–103, 2005.

[37] H. Goto, L. Peng, and K. D. Makova, "Evolution of X-degenerate Y chromosome genes in greater apes: conservation of gene content in human and gorilla, but not chimpanzee," *Journal of Molecular Evolution*, vol. 68, no. 2, pp. 134–144, 2009.

[38] V. J. Murtagh, D. O'Meally, N. Sankovic et al., "Evolutionary history of novel genes on the tammar wallaby Y chromosome: implications for sex chromosome evolution," *Genome Research*, vol. 22, no. 3, pp. 498–507, 2012.

[39] S. Rozen, J. D. Marszalek, R. K. Alagappan, H. Skaletsky, and D. C. Page, "Remarkably little variation in proteins encoded by the Y chromosome's single-copy genes, implying effective purifying selection," *American Journal of Human Genetics*, vol. 85, no. 6, pp. 923–928, 2009.

[40] R. Saxena, L. G. Brown, T. Hawkins et al., "The DAZ gene cluster on the human Y chromosome arose from an autosomal gene that was transposed, repeatedly amplified and pruned," *Nature Genetics*, vol. 14, no. 3, pp. 292–299, 1996.

[41] B. T. Lahn and D. C. Page, "Retroposition of autosomal mRNA yielded testis-specific gene family on human Y chromosome," *Nature Genetics*, vol. 21, no. 4, pp. 429–433, 1999.

[42] M. L. Delbridge, G. Longepied, D. Depetris et al., "TSPY, the candidate gonadoblastoma gene on the human Y chromosome, has a widely expressed homologue on the X—implications for Y chromosome evolution," *Chromosome Research*, vol. 12, no. 4, pp. 345–356, 2004.

[43] M. L. Delbridge, P. A. Lingenfelter, C. M. Disteche, and J. A. Marshall Graves, "The candidate spermatogenesis gene RBMY has a homologue on the human X chromosome," *Nature Genetics*, vol. 22, no. 3, pp. 223–224, 1999.

[44] W. J. Murphy, A. J. Pearks Wilkerson, T. Raudsepp et al., "Novel gene acquisition on carnivore Y chromosomes," *PLoS genetics*, vol. 2, no. 3, p. e43, 2006.

[45] T. C. Chang, Y. Yang, H. Yasue, A. K. Bharti, E. F. Retzel, and W. S. Liu, "The expansion of the PRAME gene family in Eutheria," *PLoS ONE*, vol. 6, no. 2, Article ID e16867, 2011.

[46] Y. Yang, T. C. Chang, H. Yasue, A. K. Bharti, E. F. Retzel, and W. S. Liu, "*ZNF280BY* and *ZNF280AY*: autosome derived Y-chromosome gene families in *Bovidae*," *BMC Genomics*, vol. 12, article 13, 2011.

[47] L. B. Koerich, X. Wang, A. G. Clark, and A. B. Carvalho, "Low conservation of gene content in the *Drosophila* Y chromosome," *Nature*, vol. 456, no. 7224, pp. 949–951, 2008.

[48] A. B. Carvalho, L. B. Koerich, and A. G. Clark, "Origin and evolution of Y chromosomes: *Drosophila* tales," *Trends in Genetics*, vol. 25, no. 6, pp. 270–277, 2009.

[49] M. A. Wilson and K. D. Makova, "Genomic analyses of sex chromosome evolution," *Annual Review of Genomics and Human Genetics*, vol. 10, pp. 333–354, 2009.

[50] B. T. Lahn and D. C. Page, "Functional coherence of the human Y chromosome," *Science*, vol. 278, no. 5338, pp. 675–680, 1997.

[51] F. J. Krsticevic, H. L. Santos, S. Januário, C. G. Schrago, and A. B. Carvalho, "Functional copies of the Mst77F gene on the Y chromosome of *Drosophila melanogaster*," *Genetics*, vol. 184, no. 1, pp. 295–307, 2010.

[52] A. Touré, E. J. Clemente, P. Ellis et al., "Identification of novel Y chromosome encoded transcripts by testis transcriptome analysis of mice with deletions of the Y chromosome long arm," *Genome biology*, vol. 6, no. 12, p. R102, 2005.

[53] V. S. Vineeth and S. S. Malini, "A journey on Y chromosomal genes and male infertility," *International Journal of Human Genetics*, vol. 11, no. 4, pp. 203–215, 2011.

[54] T. Kuroda-Kawaguchi, H. Skaletsky, L. G. Brown et al., "The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men," *Nature Genetics*, vol. 29, no. 3, pp. 279–286, 2001.

[55] L. Ferguson, P. J. I. Ellis, and N. A. Affara, "Two novel mouse genes mapped to chromosome Yp are expressed specifically in spermatids," *Mammalian Genome*, vol. 20, no. 4, pp. 193–206, 2009.

[56] R. A. Fisher, "The evolution of dominance," *Biological Reviews*, vol. 6, pp. 345–368, 1931.

[57] M. Gallach, S. Domingues, and E. Betran, "Gene duplication and the genome distribution of sex-biased genes," *International Journal of Evolutionary Biology*, vol. 2011, Article ID 989438, 20 pages, 2011.

[58] E. Postma, N. Spyrou, L. A. Rollins, and R. C. Brooks, "Sex-dependent selection differentially shapes genetic variation on and off the guppy Y chromosome," *Evolution*, vol. 65, no. 8, pp. 2145–2156, 2011.

[59] R. B. Roberts, J. R. Ser, and T. D. Kocher, "Sexual conflict resolved by invasion of a novel sex determiner in lake malawi cichlid fishes," *Science*, vol. 326, no. 5955, pp. 998–1001, 2009.

[60] G. S. van Doorn and M. Kirkpatrick, "Turnover of sex chromosomes induced by sexual conflict," *Nature*, vol. 449, no. 7164, pp. 909–912, 2007.

[61] A. K. Chippindale and W. R. Rice, "Y chromosome polymorphism is a strong determinant of male fitness in *Drosophila melanogaster*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 10, pp. 5677–5682, 2001.

[62] B. Lemos, L. O. Araripe, and D. L. Hartl, "Polymorphic Y chromosomes harbor cryptic variation with manifold functional consequences," *Science*, vol. 319, no. 5859, pp. 91–93, 2008.

[63] T. B. Sackton, H. Montenegro, D. L. Hartl, and B. Lemos, "Interspecific Y chromosome introgressions disrupt testis-specific gene expression and male reproductive phenotypes in *Drosophila*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 41, pp. 17046–17051, 2011.

[64] B. T. Lahn, N. M. Pearson, and K. Jegalian, "The human Y chromosome, in the light of evolution," *Nature Reviews Genetics*, vol. 2, no. 3, pp. 207–216, 2001.

[65] L. D. Hurst, "Is *Stellate* a relict meiotic driver?" *Genetics*, vol. 130, no. 1, pp. 229–230, 1992.

[66] D. T. Gerrard and D. A. Filatov, "Positive and negative selection on mammalian Y chromosomes," *Molecular Biology and Evolution*, vol. 22, no. 6, pp. 1423–1432, 2005.

[67] A. Luddi, M. Margollicci, L. Gambera et al., "Spermatogenesis in a man with complete deletion of USP9Y," *The New England Journal of Medicine*, vol. 360, no. 9, pp. 881–885, 2009.

[68] J. P. Bielawski and Z. Yang, "Positive and negative selection in the DAZ gene family," *Molecular Biology and Evolution*, vol. 18, no. 4, pp. 523–529, 2001.

[69] H. Ceplitis and H. Ellegren, "Adaptive molecular evolution of HINTW, a female-specific gene in birds," *Molecular Biology and Evolution*, vol. 21, no. 2, pp. 249–254, 2004.

[70] G. A. B. Marais, M. Nicolas, R. Bergero et al., "Evidence for degeneration of the Y chromosome in the dioecious plant *Silene latifolia*," *Current Biology*, vol. 18, no. 7, pp. 545–549, 2008.

[71] D. Bachtrog, "Evidence that positive selection drives Y-chromosome degeneration in *Drosophila miranda*," *Nature Genetics*, vol. 36, no. 5, pp. 518–522, 2004.

[72] C. Bartolomé and B. Charlesworth, "Evolution of amino-acid sequences and codon usage on the *Drosophila miranda* neo-sex chromosomes," *Genetics*, vol. 174, no. 4, pp. 2033–2044, 2006.

[73] V. B. Kaiser and B. Charlesworth, "The effects of deleterious mutations on evolution in non-recombining genomes," *Trends in Genetics*, vol. 25, no. 1, pp. 9–12, 2009.

[74] D. A. Filatov, V. Laporte, C. Vitte, and D. Charlesworth, "Dna diversity in sex-linked and autosomal genes of the plant species *Silene latifolia* and *Silene dioica*," *Molecular Biology and Evolution*, vol. 18, no. 8, pp. 1442–1454, 2001.

[75] S. Qiu, R. Bergero, A. Forrest, V. B. Kaiser, and D. Charlesworth, "Nucleotide diversity in *Silene latifolia* autosomal and sex-linked genes," *Proceedings of the Royal Society B*, vol. 277, no. 1698, pp. 3283–3290, 2010.

[76] G. A. B. Marais, P. R. A. Campos, and I. Gordo, "Can intra-Y gene conversion oppose the degeneration of the human Y chromosome? A simulation study," *Genome Biology and Evolution*, vol. 2, no. 1, pp. 347–357, 2010.

[77] S. Rozen, H. Skaletsky, J. D. Marszalek et al., "Abundant gene conversion between arms of palindromes in human and ape Y chromosomes," *Nature*, vol. 423, no. 6942, pp. 873–876, 2003.

[78] J. Lange, H. Skaletsky, S. K. M. van Daalen et al., "Isodicentric Y chromosomes and sex disorders as byproducts of homologous recombination that maintains palindromes," *Cell*, vol. 138, no. 5, pp. 855–869, 2009.

[79] H. Innan and F. Kondrashov, "The evolution of gene duplications: Classifying and distinguishing between models," *Nature Reviews Genetics*, vol. 11, no. 2, pp. 97–108, 2010.

[80] S. P. Mansai, T. Kado, and H. Innan, "The rate and tract length of gene conversion between duplicated genes," *Genes*, vol. 2, no. 2, pp. 313–331, 2011.

[81] A. Geraldes, T. Rambo, R. A. Wing, N. Ferrand, and M. W. Nachman, "Extensive gene conversion drives the concerted evolution of paralogous copies of the SRY gene in European rabbits," *Molecular Biology and Evolution*, vol. 27, no. 11, pp. 2437–2440, 2010.

[82] N. Backström, H. Ceplitis, S. Berlin, and H. Ellegren, "Gene conversion drives the evolution of HINTW, an ampliconic gene on the female-specific avian W chromosome," *Molecular Biology and Evolution*, vol. 22, no. 10, pp. 1992–1999, 2005.

[83] J. K. Davis, P. J. Thomas, and J. W. Thomas, "AW-linked palindrome and gene conversion in new world sparrows and blackbirds," *Chromosome Research*, vol. 18, no. 5, pp. 543–553, 2010.

[84] J. Alfoldi, *Sequence of the Mouse Y Chromosome*, Massachusetts Institute of Technology, Cambridge, Mass, USA, 2008.

[85] E. L. C. Verkaar, C. Zijlstra, E. M. van 't Veld, K. Boutaga, D. C. J. van Boxtel, and J. A. Lenstra, "Organization and concerted evolution of the ampliconic Y-chromosomal TSPY genes from cattle," *Genomics*, vol. 84, no. 3, pp. 468–474, 2004.

[86] M. J. Noordam, S. K. M. van Daalen, S. E. Hovingh, C. M. Korver, F. van der Veen, and S. Repping, "A novel partial deletion of the Y chromosome azoospermia factor c region is caused by non-homologous recombination between palindromes and may be associated with increased sperm counts," *Human Reproduction*, vol. 26, no. 3, pp. 713–723, 2011.

[87] E. Kichine, V. Rozé, J. Di Cristofaro et al., "HSFY genes and the P4 palindrome in the AZFb interval of the human Y chromosome are not required for spermatocyte maturation," *Human Reproduction*, vol. 27, no. 2, pp. 615–624, 2012.

[88] T. Connallon and A. G. Clark, "Gene duplication, gene conversion and the evolution of the Y chromosome," *Genetics*, vol. 186, no. 1, pp. 277–286, 2010.

[89] Y. H. Yu, Y. W. Lin, J. F. Yu, W. Schempp, and P. H. Yen, "Evolution of the DAZ gene and the AZFc region on primate Y chromosomes," *BMC Evolutionary Biology*, vol. 8, no. 1, article 96, 2008.

[90] E. Bosch, M. E. Hurles, A. Navarro, and M. A. Jobling, "Dynamics of a human interparalog gene conversion hotspot," *Genome Research*, vol. 14, no. 5, pp. 835–844, 2004.

[91] S. Mano and H. Innan, "The evolutionary rate of duplicated genes under concerted evolution," *Genetics*, vol. 180, no. 1, pp. 493–505, 2008.

[92] J. A. Fawcett and H. Innan, "Neutral and non-neutral evolution of duplicated genes with gene conversion," *Genes*, vol. 2, no. 1, pp. 191–209, 2011.

[93] L. M. Turner, E. B. Chuong, and H. E. Hoekstra, "Comparative analysis of testis protein evolution in rodents," *Genetics*, vol. 179, no. 4, pp. 2075–2089, 2008.

[94] V. C. Li, J. C. Davis, K. Lenkov, B. Bolival, M. T. Fuller, and D. A. Petrov, "Molecular evolution of the testis TAFs of *Drosophila*," *Molecular Biology and Evolution*, vol. 26, no. 5, pp. 1103–1116, 2009.

[95] S. Dorus, E. R. Wasbrough, J. Busby, E. C. Wilkin, and T. L. Karr, "Sperm proteomics reveals intensified selection on mouse sperm membrane and acrosome genes," *Molecular Biology and Evolution*, vol. 27, no. 6, pp. 1235–1246, 2010.

[96] M. Pröschel, Z. Zhang, and J. Parsch, "Widespread adaptive evolution of *Drosophila* genes with sex-biased expression," *Genetics*, vol. 174, no. 2, pp. 893–900, 2006.

[97] Z. Zhang and J. Parsch, "Positive correlation between evolutionary rate and recombination rate in *Drosophila* genes with male-biased expression," *Molecular Biology and Evolution*, vol. 22, no. 10, pp. 1945–1947, 2005.

[98] G. J. Wyckoff, W. Wang, and C. I. Wu, "Rapid evolution of male reproductive genes inn the descent of man," *Nature*, vol. 403, no. 6767, pp. 304–309, 2000.

[99] J. R. Peck, "A ruby in the rubbish: beneficial mutations, deleterious mutations and the evolution of sex," *Genetics*, vol. 137, no. 2, pp. 597–606, 1994.

[100] B. K. Bhowmick, Y. Satta, and N. Takahata, "The origin and evolution of human ampliconic gene families and ampliconic structure," *Genome Research*, vol. 17, no. 4, pp. 441–450, 2007.

[101] S. Subramanian and S. Kumar, "Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome," *Genetics*, vol. 168, no. 1, pp. 373–381, 2004.

[102] D. A. Drummond and C. O. Wilke, "Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution," *Cell*, vol. 134, no. 2, pp. 341–352, 2008.

[103] K. Ma, J. D. Inglis, A. Sharkey et al., "A Y chromosome gene family with RNA-binding protein homology: candidates for the azoospermia factor AZF controlling human spermatogenesis," *Cell*, vol. 75, no. 7, pp. 1287–1295, 1993.

[104] P. E. Warburton, J. Giordano, F. Cheung, Y. Gelfand, and G. Benson, "Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeated that contain testes genes," *Genome Research*, vol. 14, no. 10A, pp. 1861–1869, 2004.

[105] J. L. Mueller, S. K. Mahadevaiah, P. J. Park, P. E. Warburton, D. C. Page, and J. M. A. Turner, "The mouse X chromosome is enriched for multicopy testis genes showing postmeiotic expression," *Nature Genetics*, vol. 40, no. 6, pp. 794–799, 2008.

[106] D. W. Bellott, H. Skaletsky, T. Pyntikova et al., "Convergent evolution of chicken Z and human X chromosomes by expansion and gene acquisition," *Nature*, vol. 466, no. 7306, pp. 612–616, 2010.

[107] J. M. A. Turner, "Meiotic sex chromosome inactivation," *Development*, vol. 134, no. 10, pp. 1823–1831, 2007.

[108] J. H. Thomas, "Concerted evolution of two novel protein families in caenorhabditis species," *Genetics*, vol. 172, no. 4, pp. 2269–2281, 2006.

[109] B. Charlesworth and D. Charlesworth, "Rapid fixation of deleterious alleles can be caused by Muller's ratchet," *Genetical Research*, vol. 70, no. 1, pp. 63–73, 1997.

Hindawi Publishing Corporation
International Journal of Evolutionary Biology
Volume 2012, Article ID 418964, 10 pages
doi:10.1155/2012/418964

Review Article

The Role of Reticulate Evolution in Creating Innovation and Complexity

Kristen S. Swithers, Shannon M. Soucy, and J. Peter Gogarten

Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06269-3125, USA

Correspondence should be addressed to J. Peter Gogarten, gogarten@uconn.edu

Received 3 February 2012; Revised 8 May 2012; Accepted 10 May 2012

Academic Editor: Wen Wang

Reticulate evolution encompasses processes that conflict with traditional Tree of Life efforts. These processes, horizontal gene transfer (HGT), gene and whole-genome duplications through allopolyploidization, are some of the main driving forces for generating innovation and complexity. HGT has a profound impact on prokaryotic and eukaryotic evolution. HGTs can lead to the invention of new metabolic pathways and the expansion and enhancement of previously existing pathways. It allows for organismal adaptation into new ecological niches and new host ranges. Although many HGTs appear to be selected for because they provide some benefit to their recipient lineage, other HGTs may be maintained by chance through random genetic drift. Moreover, some HGTs that may initially seem parasitic in nature can cause complexity to arise through pathways of neutral evolution. Another mechanism for generating innovation and complexity, occurring more frequently in eukaryotes than in prokaryotes, is gene and genome duplications, which often occur through allopolyploidizations. We discuss how these different evolutionary processes contribute to generating innovation and complexity.

1. Introduction

Reconstruction of the Tree of Life attempts to represent the organismal histories of all of life on earth on a single bifurcating tree. Since the dawn of the molecular age, and, more so recently, with the numerous whole-genome sequences that are now available, it has become apparent that reticulate evolutionary processes such as horizontal gene transfer (HGT), genome fusion, and incomplete lineage sorting have a profound impact on microbial and eukaryotic evolution. These processes dissolve or embed the lines of vertical descent that are a hallmark of the tree of life into net-like relationships between genomes and organisms. To more accurately describe the complexity of organismal histories many groups have proposed net-like reconstructions of life's history [1] to account for the lines of vertical descent and lateral lines created from reticulate processes; the "rooted net of life" [2], the "forest of life" [3, 4], and the "rhizome of life" [5, 6] are a few examples.

HGT is the nonvertical transmission of genetic material, that is, the exchange of genetic information between organisms not in an ancestor descendant relationship. HGT causes individual genes in a genome to have vastly different evolutionary histories. Studies show HGT occurs more frequently between closely related organisms than in divergent organisms [7, 8]. Closely related organisms tend to have similar sequences and intracellular environments. These similarities allow for more opportunity for homologous recombination and for an easier integration of the transferred gene into the metabolic and regulatory networks of the recipient. However, there are increasing examples of HGTs between divergent species, even across domain boundaries, revealing that barriers to HGT can occasionally be overcome. Examples include the highways of HGTs [9] that exist between divergent organisms: members of the Thermotogae phylum share about half of their genes with both the Firmicutes and the Archaea [10], and the Aquificae share many genes with the Epsilonproteobacteria [11]. Many of these successful HGTs allow for innovations in metabolism and body plan that provide a selective advantage to the organisms involved and allow expansion into new ecological niches.

Table 1: Categories of HGTs leading to innovation and complexity.

Type	"Beneficial" HGTs	"Neutral" HGTs	"Parasitic" HGTs
Definition	HGTs that provide an initial selective advantage to the recipient	HGTs are maintained by random genetic drift	HGTs do not provide an initial selective advantage to the recipient but over time may adapt to have a beneficial function or be maintained via pathways to neutral complexity in the recipient
Examples	(i) Metabolic pathway expansion and invention (ii) Adaptation to new ecological niches	(i) Many ORFan genes and genes of limited distribution and with unknown function may be in this category [14, 15]	(i) Inteins (ii) Group I Introns (iii) Group II Introns

Transferred genes can be distinguished based on their long- and short-term impact on the fitness of the recipient (Table 1). Genes that provide an adaptation create a selective advantage for the recipient and have a higher chance to persist over longer periods of time. As their frequency in the population increases over time these genes will become fixed. Examples of these "beneficial" HGTs are those that allow the recipient to expand into a previously empty ecological niche. These provide a huge increase in fitness to the recipient, even if the transferred gene has not yet adapted perfectly to the genomic and regulatory environment of the recipient [12]. Many of the genes that extend, enhance, or create new metabolic pathways fall into this category. These genes may be selfish in Dawkins' [13] original definition, but they cooperate with the other genes in the organism's genome and provide a selective advantage for the organism.

Many other, and possibly most, transferred genes that can be identified in the pan-genome [16] of bacterial or archaeal populations may be selectively neutral or nearly neutral to their carriers [14]. Many of these genes will be lost after a few generations; however, a few may be fixed through random genetic drift. It could be argued that most of the endosymbiant *Wolbachia* to host transfers are selectively neutral or nearly neutral. Almost all of the *Wolbachia* genes are found in the host genome and their transcript levels are very low [17]. This low transcript level may indicate that these genes do not provide a function to the host and supports the notion that many genes transferred from the symbiont are only transiently present in the host nuclear genome. Although the majority of these transferred genes are transcribed at very low level, two hypothetical proteins in the *Aedyes aegypti* originating from *Wolbachia* have been maintained in the nuclear genome for a long period of time and are transcribed at higher levels than background suggesting these genes were fixed in the population [18].

Some transferred genes initially are like infections in that their survival and spread is through a mechanism that decouples the genes propagation from host replication and host fitness. Although the propagation of these selfish genetic elements is decoupled from the host's genetic machinery, the element does utilize the host's resources to propagate through a population. In this sense these genetic elements can be considered parasitic. To more clearly distinguish them from the selfish gene concept in Dawkins' gene-centered view of evolution, which considers all genes as selfish, we term these elements as parasitic genetic elements and their transfers "parasitic HGTs"; examples include inteins and self-splicing introns. Initially, a self-splicing molecular parasite may provide little or no advantage to the host but may later adapt a function to benefit the host. Many inteins and group I introns contain a homing endonuclease (HE) that provides mobility to the element and allows them to follow a life cycle known as the homing cycle [19]. Briefly, the homing cycle begins when an allele with an HE is horizontally transferred to a recipient in a new population or species that before the invasion harbored only alleles without HE [20]. Through faster than Mendalian inheritance the HE containing parasite spreads through the population, leaving little or no detrimental effects on the host. However, once all the members of the population have the HE containing element the HE containing genetic element starts to degrade. To escape this cycle, over time the parasites may adapt to provide a beneficial function for the host [7] or are maintained through neutral pathways to complexity as discussed below for the case of the *dnaE* intein [21, 22].

Transferred genes can be integrated into the recipient genome by homologous recombination or through illegitimate recombination [23]. The former process requires stretches of similar sequences; however, the stringency of this requirement depends on the activity of the mismatch repair system [24]. The similarities necessary for homologous recombination can be due to the presence of a homolog in the recipient genome or can be created through transposable elements present in the recipient that jump into the transferred extrachromosomal genetic material [25]. Transferred DNA also can be integrated independent of sequence similarity through double-strand break repair pathways, such as nonhomologous end joining, allowing for the integration of DNA from divergent organisms [7]. Transposable elements can also facilitate transfer and integration into recipient DNA. One such example is the integrative and conjugative elements (ICEs). ICEs have been implicated in transfer of genes involved in antibiotic and heavy metal resistance, nitrogen fixation, virulence, biofilm formation, and the degradation of aromatic compounds (for reviews see [26, 27] and references therein), providing another example for multiple levels of selection, in this case benefiting both the transferred genes and the recipient.

Although HGT appears to be more prevalent in prokaryotes, more and more examples of HGT are being documented in single-celled and even multicellular eukaryotes (see [28] and below for examples of transfer from bacteria to eukaryotes). Related driving forces in creating innovation and complexity in eukaryotic lineages are gene and whole genome duplications. Genome fusion resulting from hybridization between members of related species, a frequent pathway towards polyploidization, is akin to HGT in that it results in mosaic genomes and that the resulting gene family expansion is due to reticulate evolution. Observed in plants [29], animals [30, 31], and fungi [32, 33] whole-genome duplication followed by neofunctionalization and/or subfunctionalizations have been implicated in providing the building blocks for more complex developmental and metabolic pathways.

Gene, genome duplication, and HGT, regardless of the type of selection, beneficial, neutral, or parasitic, are all reticulate processes that affect evolution across all domains of life. Here we explore how the process of HGTs can expand metabolic pathways, allow for microorganisms to adapt to new host ranges, expand environmental niches, and even influence multicellular eukaryotes. We also explore how "parasitic HGTs" can ultimately lead to innovation and increased complexity. Additionally, we discuss how gene and whole-genome duplications can give rise to novel pathways that are important for development.

2. HGT and Expansion Metabolic Pathways

HGTs can lead to the enhancement, expansion, and construction of more complex metabolic pathways. About two-thirds of the annual biogenic methane is produced from the acetoclastic methanogenesis pathway, which is exclusively carried out by the methanogenic eurarchaeal order Methanosarcinales [34]. Most members of this group carry out the conversion of acetate to acetyl-coenzyme A using the acetyl-CoA synthesis pathway. However, members of the more widely distributed *Methanosarcina* use a variation on this pathway, which uses the enzymes acetate kinase (AckA) and phosphoacetyl treasferase (Pta) [34]. Both the *ackA* and *pta* genes were shown through multiple phylogenetic methods to be transferred in one event from the cellulolytic clostridia, where the encoded enzymes are used to produce acetate as a product of fermentation, to *Methanosarcina* [35], where the same enzymes are used to produce acetyl-CoA.

Another example of an expanded pathway created by HGT is found in the Thermotogae phylum. Some of the lower-temperature lineages are able to produce vitamin B_{12} using the cobinamide salvage pathway [36] (Figure 2). In this pathway a partial B_{12} molecule is scavenged from the environment and subsequently modified to produce an active B_{12} molecule. This method of B_{12} production was shown to be the ancestral pathway for the Thermotogae lineage by presence and absence of the genes in the phylum (Figure 2). A later HGT allowed the *Thermosipho* genus to synthesize B_{12} *de novo* from glutamate, through transfer of twenty-one genes from the Firmicutes.

An enhancement of a pathway is observed in HGT events between eukaryotic species of grasses. Some members of the *Alloteropsis* grasses have acquired highly functional genes for C_4 photosynthesis from the Cenchrinae and Melinidinae: phosphoenolpyruvate carboxylases (ppc) were likely transferred from both the Cenchrinae and Melinidinae, and phosphoenolpyruvate carboxykinase (pck) was transferred from the Cenchrinae. Christin et al. hypothesize that before the arrival of these genes the *Alloteropsis* may have had a subfunctional C_4 CO_2-fixation pathway, as in the case of the extant *A. semialata* subsp. *semialata* grass, which did not receive these HGTs. This enhancement of the C_4 pathways allows for adaptation of the grass to warm and arid climates [37].

The metabolic pathways expanded and enhanced through HGT allow for an occupation of a new ecological niche. The *Thermosipho* can now produce B_{12} and thrive in an environment where no partial B_{12} derivatives are present, while members of the genus *Methanosarcina* are able to produce most of the world's methane from acetate and the *Alloteropsis* grasses can thrive in warm and arid climates.

3. HGT and Metabolic Innovations

Members of at least six different bacterial phyla use chlorophyll-based photosynthesis to gain energy from light [38, 39]. Comparative phylogenetic analysis revealed that horizontal gene transfer played an important role in evolution and distribution of bacterial photosynthesis [40, 41]. The assembly of the electron transport chain that allows the use of water as electron donor likely represents the gene transfer event that most changed Earth's biosphere [42, 43]. Chloroflexi (green filamentous bacteria) and purple bacteria possess a photosynthetic reaction center similar to photosystem II of the cyanobacteria; whereas the reaction centers in Chlorobi (green sulfur bacteria) and Heliobacteria (Firmicutes) are similar to the photosynthetic reaction center I in cyanobacteria [39, 44]. However, in the cyanobacteria photosystem I and photosystem II are present, and only when the two divergent types of reaction centers work in series do the harvested photons provide sufficient energy to lift electrons over the electrochemical potential difference between water and NADP. It is theoretically possible that photosystems I and II arose through a within-lineage gene duplication, diverged within the cyanobacteria, and subsequently individual photosystems were transferred to other bacteria. A more likely scenario is that the two photosystems diverged from an ancestral photosystem in diverging lineages (Figure 1(b)), which each used a single photosystem, and that the two distinct photosystems were brought together in the cyanobacterial ancestor through HGT.

The recently described methylaspartate cycle in Haloarchaea [45] provides another example for the creative power of HGT. This cycle provides an alternative to the glyoxylate cycle and the ethylmalonyl-CoA pathway for acetyl CoA to enter central carbon metabolism to synthesize cellular building blocks. According to analyses reported in [45] the key enzymes of the methylaspartate cycle were acquired by

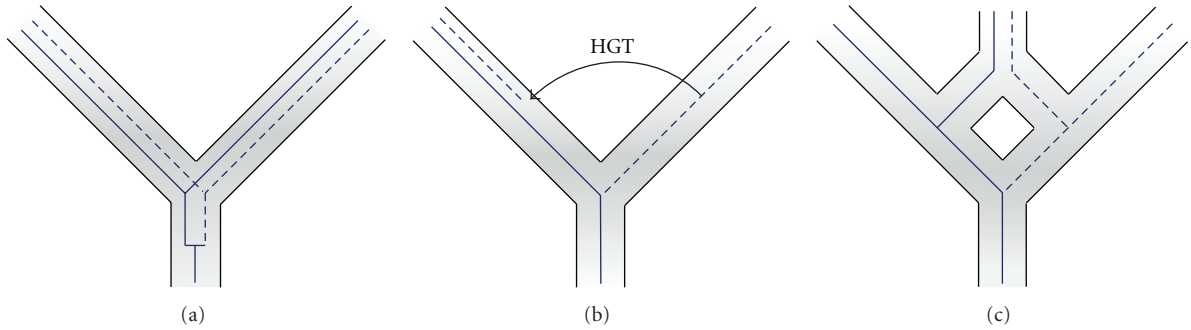


Figure 1: Types of genetic duplications. (a) Shows an autochthonous duplication, which can happen either through tandem duplication, segmental duplication, chromosomal duplication, genome duplications, or retro-transposition. (b) Shows gene family expansion through HGT. Following the divergence of two lineages orthologous genes diverge in sequence and possibly in function. These orthologs can be brought together in a single genome through HGT or allopolyploidization (c). The scenarios depicted in (c) and (b) explain an apparent duplication through reticulated evolution.

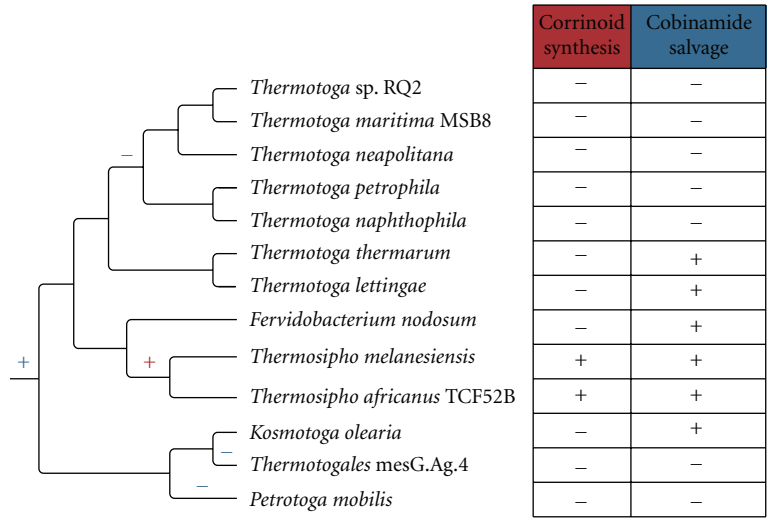


Figure 2: Distribution of the two gene clusters involved in vitamin B_{12} biosynthesis among the Thermotogae phylum. The corrinoid synthesis gene cluster contains genes for the first part of the *de novo* B_{12} synthesis pathway and the cobinamide salvage gene cluster contains genes that synthesize vitamin B_{12} from cobinamides, incomplete B_{12} molecules. Together these two gene clusters complete the *de novo* B_{12} biosynthesis pathway. Presence of a gene cluster is denoted by (+) and absence is denoted by (–). The most parsimonious explanation for the extant presence/absence patterning for the cobinamide salvage gene cluster is one gain at the root of the phylum and three losses marked by blue and (+) and (–) and for the corrinoid synthesis gene cluster one gain marked by a red (+). This suggests the cobinamide salvage pathway was present in the ancestor of the Thermotogae phylum and the genes for complete *de novo* synthesis were gained in a later event by the *Thermosipho* lineage.

the Haloarchaea through gene transfer from different bacteria. Furthermore, before the transfer, these enzymes were part of different pathways in the donor organisms, such as propionate assimilation or glutamate fermentation. The methylaspartate cycle thus represents a metabolic patchwork of enzymes acquired from different donors and combining fragments of different pathways into a novel enzymatic cycle.

4. HGT and Innovations in Communities

The human microbiome provides an opportunity to understand a complex community of microorganisms and how HGT has facilitated innovation within a large community of microorganisms. Many traits, such as antibiotic resistance, and xenobiotic metabolism observed in the human gut microbiota are a consequence of HGT. One study showed that antibiotic resistance genes can be transferred to the gastrointestinal microbiome from food sources [46]. Volunteers were fed chicken, which had a strain of vancomycin-resistant *Enterococcus faecium*, and vancomycin resistance was transferred to *E. faecium* in the human gut. Other studies in Japanese individuals showed that genes for porphyranases, agarases, and alginases, which facilitate the breakdown of red and brown algae (seaweed) in the human gut, were likely transferred from marine bacteria to Japanese gut symbiont

Bacteroidetes [47, 48]. These HGTs not only allow the gut bacteria to utilize seaweeds as a novel carbon source, but confer secondary benefits to the human host, which can now utilize seaweed as a nutrient source. The act of introducing foreign material to the gut microbiota (consuming a food source) facilitates interactions between the microbiome and the microorganisms on that food source. This interaction encourages possible HGTs from microorganisms outside the gut and allows for constant innovation and evolution of our microbiome to cope with the frequent changes in the gut environment, reinforcing the "you are what you eat" saying. These findings also confirm that the holobiont (host plus symbiont) can evolve and gain new adaptations without changes in the host's genome, simply by acquiring new symbionts with novel metabolic capabilities [49].

5. "Parasitic HGTs" Can Lead to Innovation and Complexity

"Parasitic HGT" involving molecular parasites, such as inteins and group I introns, are HGTs that confer no immediate selective advantage to the host but over time adapt to benefit the host. These inteins and group I introns are self-splicing genetic elements that are made mobile by homing endonucleases, an endonuclease that recognizes target sequences of 12–40 bps [50]. They can evade purifying selection on the organismal level as they cause little or no harm to their host [51]. These HE containing parasites have their own life cycle described by the homing cycle [20, 50, 52]. A possible escape route from this cycle presents itself, if the HE or the intein/intron evolves a beneficial function in the host. One such example of this is found in the mating type switching HO endonuclease in yeast [53]. This endonuclease is left over from what once was a close relative to the large intein in the yeast vacuolar ATPase catalytic subunit, but now facilitates genetic recombination from one mating type to another. This innovation is beneficial to the organism in that it expands the reproductive capabilities of the yeast cell. Another example where an intein may have been retained and adapted to benefit its host is found in bacterial intein-like (BIL) domains. These are degenerated remnants of the HINT domain intein family, which are now thought to function to facilitate rearrangements in hypervariable surface proteins [54, 55]. Over time the HEs of some group I introns are maintained as functional maturases to aid in the folding and splicing of the intron they reside in or other introns that may have lost their self-splicing ability [21, 56]. In these cases parasitic HGTs have facilitated beneficial innovations; however, most of these innovations evolved after a long period of neutral or nearly neutral association between the parasite and host.

Although many "parasitic HGTs" eventually provide some benefit for the host, there are other cases where they are maintained via selectively neutral pathways, which also can lead to higher complexity. The *dnaE* gene, of some cyanobacterial species, is split on two parts (*dnaE1* and *dnaE2*), and each portion has part of an N-terminal or C-terminal intein [57]. An autocatalytic mechanism allows the split inteins to find each other after translation and splice the split protein together, resulting in a functional DNA polymerase III. Deletion or mutation of the intein portions of the split gene results in a nonfunctional DNA polymerase III, a major selective disadvantage for the organism and even possibly detrimental. This intein likely never supplied a selective advantage for the host. Through a series of intermediate steps, each of them neutral or nearly neutral to the organism, a complex processing system emerged that places the intein under strong purifying selection, because the self-splicing reaction of the intein now is necessary to synthesize a functioning DNA polymerase III [22]. The wide distribution of the split intein in *dnaE* in cyanobacteria [58] suggests that this rather complex gene structure is an evolutionarily stable arrangement.

Another mobile genetic element that is frequently transferred and creates novelties and complexity is group II introns. They are thought to be the predecessors of both the eukaryotic spliceosomal introns and non-LTR retrotransposon [59–61]. These self-splicing elements are found in all domains of life; they are made mobile either via retrohoming, using an endonuclease [62], or retrotransposition mechanisms, using a reverse transcriptase [63]. Evidence for group II introns being the ancestors of the spliceosomal intron in eukaryotes includes similar splicing mechanisms, comparable boundary sequences, and secondary structure similarities [64–66]. One hypothesis suggests the group II intron originated in the bacteria and were horizontally transferred from the alphaproteobacterial endosymbiont ancestor of the mitochondria to the genome of the ancestor of the eukaryotic nucleocytoplasm. The presence of introns in most transcripts might have necessitated a separation between transcription and translation, facilitating the emergence of a nucleus [67]. Some of the original introns may have lost their self-splicing activity and relied on other introns and their associated proteins to catalyze the splicing reaction in trans, evolving over time into the spliceosomal machinery. In this scenario, the introns initially proliferated as molecular parasites; however, on the long run they allowed for exon shuffling, alternative splicing, and the nonsense mediated decay pathway to evolve. Interestingly, extant bacterial group II introns maintain self-splicing and mobility, while most mitochondrial and chloroplast group II introns are not mobile and have lost the ability to self-splice. For example, about 20 group II introns present in the organelles of plants have lost their ability to self-splice [68, 69]. However, to maintain functional genes, they must be spliced out thus their maintenance is dependent on the complex interactions with nuclear and plastid splicing factors. Group II introns have also been implicated in genome rearrangements and gene conversion events [70], both of which can cause innovations in gene function and structure.

6. Interdomain HGT and Innovation

One of the benefits of HGT is that it can provide a selective advantage for organisms to occupy new niches and expand host ranges. Many interdomain transfers from bacteria to

single-celled eukaryotes provided for innovations and adaptation to new environments [28, 71]. In many instances these genes were subsequently transferred between divergent single-cell eukaryotes [28]. One example is the parasitic protozoan *Blastocystis*, which is found in many different animal gut environments and causes gastrointestinal diseases, and has acquired genes for energy metabolism, adhesion, and osmotrophy from various bacterial donors. These transfers have allowed the successful adaptation of Blastocystis to the gut environment [72].

Surprisingly many genes were transferred from bacteria into multicellular eukaryotes. The ancient bacterivorous nematodes acquired cell wall degrading enzymes from several bacterial lineages via HGT [73–75]. The cell wall degrading genes are required for the initial stages in plant pathogenesis, without them plants would be an unavailable niche for the nematode [76]. Therefore, the transfer of those genes allowed the transition of the nematode from a free living state to a plant parasite [77]. Other examples of innovative interdomain HGTs can be found in the tunicates. A cellulose synthase gene (*cesA*) is proposed to have been transferred to the ancestor of the tunicates from a bacterial lineage [78]. Following a gene duplication, CesA1 produces cellulose for the larval tail and CesA2 synthesizes cellulose for the complex filter-feeding house of the ascidians and larvaceans [78]. This HGT played a role in body plan development in tunicates.

Examples of bacteria to animal transfers also reveal the adaptive benefits. The HhMAN1 gene in the coffee berry borer, *Hypothenemus hampei*, was likely transferred from a bacterial lineage [79]. The gene encodes a secreted mannanase that allows the coffee berry borer access the primary seed storage polysaccharide in the coffee plant and ultimately confers an adaptive advantage because *H. hampei* uses the coffee berry as a specific host [79]. The spider mite *Tetranychus urticae* has several genes likely transferred from bacterial lineages; those are genes that encode a secreted fructosidase and a cyanate lyase-encoding gene that may be involved in feeding on cyanogenic plants [80]. These acquisitions have allowed the spider mite to utilize different plants for feeding thereby expanding its host range [80].

The aphid genome, *Acyrthosiphon pisum*, encodes for multiple carotenoids transferred from fungal lineages. These genes allow the aphid to synthesize its own carotenoids rather than to acquire them from food sources as many other animals do [81]. These are only a few of the current examples of interdomain HGTs. As more and more genomes from multicellular organisms become available more interdomain transfers are likely to be revealed.

7. Gene Duplication and Gene Transfer

The emergence of new genes from previously noncoding DNA is a rare event (e.g., [82, 83]). Most new genes are believed to originate through gene duplication [84]. In Eukaryotes gene duplications frequently occur in an autochthonous fashion within a single lineage (Figure 1(a)). Mechanisms include tandem, segmental, and chromosomal duplication, retrotransposition, and genome duplications [85]. Of the two genes created, most frequently one accumulates mutations and is no longer maintained under purifying selection and decays [86]. There are two mechanisms by which the duplicated gene can be maintained, subfunctionalization or neofunctionalization. In subfunctionalization, functions of the parent gene are divided among the duplicated genes; in neofunctionalization, after duplication one copy diverges to create a new function. The creation of new functions from duplicated genes appears to be a rare event [87].

Ancient genome duplications have played an important role in vertebrate, plant, and fungi evolution (see [88] for review). In these ancient duplications it is difficult to decide if the whole genome duplication resulted from an autochthonous autopolyploidization or an allopolyploidization following a between-species hybridization (Figure 1(c)). The latter process is particularly important in plant evolution and breeding [89]. Many of these whole-gene duplications are followed by neofunctionalization and subfunctionalizations of various genes throughout the genome. However the above example of the cellulose synthatase genes in the larvacean lineage of tunicates is an example of a gene duplication leading to neofunctionalization in a eukaryote.

The whole-genome duplication of the fungus *Saccharomyces cerevisiae* followed by neofunctionalization of various genes led to the emergence of viral defense mechanisms from translation elongation and the emergence of gene silencing from origin of replication binding proteins [33]. Subfunctionalization events after gene or genome duplications can also arise and create novel regulatory pathways. For example, the maize genome arose from an allotetraploidization between two grass species [90–93]. In the extant maize lineage the *ZAG1* and *ZMM2* genes are necessary for the development of stamens and carpals in the plant. The *ZAG1* gene is expressed throughout carpal development, and the *ZMM2* gene is expressed in maize stamen but not in the immature carpal [94]. It is thought that these genes were expressed in both developing stamens and carpals in the allotetraploid ancestor shortly after the polyploidization event [95]. Over time mutations affecting the regulation of *ZAG1* decrease expression of ZAG1 in stamens but not carpals and mutations affecting the regulation of *ZMM2* eliminated expression in the early carpal but not in stamens [95].

In Bacteria and Archaea autochthonous gene duplications appear to be rare [42, 96]. The typical pathway for gene family extension is through HGT followed by nonhomologous recombination in the recipient. Following the divergence of two lineages, orthologous genes experience substitutions. These might be associated with altered properties of the encoded protein; for example, mutations in an ion translocating subunit of an ATP synthase/ATPase might increase its specificity for protons, thereby changing the specificity for the transported ion from Na^+ to H^+ [97], allowing the organism to use the proton motive force for ATP synthesis. When subsequently the two genes end up in the same cell following horizontal gene transfer, they have diverged so much that homologous recombination between the divergent forms is no longer possible (Figure 1(b)).

As both genes have different functions, both can be maintained in the recipient through purifying selection. For example, one ATPase might function as ATP synthase driven by a Na^+ gradient, and the homolog might function in controlling the cellular pH.

8. Conclusions

The processes of reticulate evolution lead to innovations and complexity. Horizontal gene transfer whether beneficial or parasitic in nature can lead to innovations and increased complexity. "Beneficial" HGTs provide an immediate selective advantage to the recipient, which increases fitness and guarantees that the transferred gene will be fixed in the recipient's population. Such benefits include but are not limited to innovations in metabolic pathways, expansion of niche adaptations, and in the case of the human gut microbiome can have important secondary implications for the human. "Parasitic" HGTs can also provide innovation, although innovation is more likely to be formed through neutral or nearly neutral pathways to complexity. Gene and genome duplications are another way to spawn innovation and complexity, more so in Eukaryotes than in prokaryotic lineages. In both cases, the horizontal transfer of genetic material and gene and genome duplications are a driving factor in organismal evolution.

Acknowledgment

This work was supported through the National Science Foundation (DEB 0830024) and the NASA Exobiology program (NNX08AQ10G).

References

[1] T. Dagan, Y. Artzy-Randrup, and W. Martin, "Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 29, pp. 10039–10044, 2008.

[2] D. Williams, G. P. Fournier, P. Lapierre et al., "A rooted net of life," *Biology Direct*, vol. 6, article 45, 2011.

[3] P. Lopez and E. Bapteste, "Molecular phylogeny: reconstructing the forest," *Comptes Rendus Biologies*, vol. 332, no. 2-3, pp. 171–182, 2009.

[4] P. Puigb, Y. I. Wolf, and E. V. Koonin, "Search for a "Tree of Life" in the thicket of the phylogenetic forest," *Journal of Biology*, vol. 8, no. 6, article 59, 2009.

[5] V. Merhej, C. Notredame, M. Royer-Carenzi, P. Pontarotti, and D. Raoult, "The rhizome of life: the sympatric *Rickettsia felis* paradigm demonstrates the random transfer of DNA sequences," *Molecular Biology and Evolution*, vol. 28, no. 11, pp. 3213–3223, 2011.

[6] D. Raoult, "The post-Darwinist rhizome of life," *The Lancet*, vol. 375, no. 9709, pp. 104–105, 2010.

[7] O. Popa, E. Hazkani-Covo, G. Landan, W. Martin, and T. Dagan, "Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes," *Genome Research*, vol. 21, no. 4, pp. 599–609, 2011.

[8] C. P. Andam and J. P. Gogarten, "Biased gene transfer in microbial evolution," *Nature Reviews Microbiology*, vol. 9, no. 7, pp. 543–555, 2011.

[9] R. G. Beiko, T. J. Harlow, and M. A. Ragan, "Highways of gene sharing in prokaryotes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 40, pp. 14332–14337, 2005.

[10] O. Zhaxybayeva, K. S. Swithers, P. Lapierre et al., "On the chimeric nature, thermophilic origin, and phylogenetic placement of the Thermotogales," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 14, pp. 5865–5870, 2009.

[11] B. Boussau, L. Guéguen, and M. Gouy, "Accounting for horizontal gene transfers explains conflicting hypotheses regarding the position of aquificales in the phylogeny of Bacteria," *BMC Evolutionary Biology*, vol. 8, no. 1, article 272, 2008.

[12] J. P. Gogarten, W. F. Doolittle, and J. G. Lawrence, "Prokaryotic evolution in light of gene transfer," *Molecular Biology and Evolution*, vol. 19, no. 12, pp. 2226–2238, 2002.

[13] R. Dawkins, *The Selfish Gene*, Oxford University Press, 1976.

[14] J. P. Gogarten and J. P. Townsend, "Horizontal gene transfer, genome innovation and evolution," *Nature Reviews Microbiology*, vol. 3, no. 9, pp. 679–687, 2005.

[15] P. Lapierre and J. P. Gogarten, "Estimating the size of the bacterial pan-genome," *Trends in Genetics*, vol. 25, no. 3, pp. 107–110, 2009.

[16] H. Tettelin, V. Masignani, M. J. Cieslewicz et al., "Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial 'pan-genome'," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 39, pp. 13950–13955, 2005.

[17] J. C. Dunning Hotopp, M. E. Clark, D. C. S. G. Oliveira et al., "Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes," *Science*, vol. 317, no. 5845, pp. 1753–1756, 2007.

[18] L. Klasson, Z. Kambris, P. E. Cook, T. Walker, and S. P. Sinkins, "Horizontal gene transfer between *Wolbachia* and the mosquito Aedes aegypti," *BMC Genomics*, vol. 10, article 33, 2009.

[19] M. R. Goddard and A. Burt, "Recurrent invasion and extinction of a selfish gene," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 24, pp. 13880–13885, 1999.

[20] J. P. Gogarten and E. Hilario, "Inteins, introns, and homing endonucleases: recent revelations about the life cycle of parasitic genetic elements," *BMC Evolutionary Biology*, vol. 6, article 94, 2006.

[21] D. Mo, L. Wu, Y. Xu et al., "A maturase that specifically stabilizes and activates its cognate group I intron at high temperatures," *Biochimie*, vol. 93, no. 3, pp. 533–541, 2011.

[22] K. S. Swithers and J. P. Gogarten, "Introns and Inteins," in *Bacterial Integrative Mobile Genetic Elements*, chapter 4, Landes Bioscience, Austin, Tex, USA, 2012.

[23] J. De Vries and W. Wackernagel, "Integration of foreign DNA during natural transformation of *Acinetobacter* sp. by homology-facilitated illegitimate recombination," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 4, pp. 2094–2099, 2002.

[24] M. Vulić, R. E. Lenski, and M. Radman, "Mutation, recombination, and incipient speciation of bacteria in the laboratory," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 13, pp. 7348–7351, 1999.

[25] H. Ochman, J. G. Lawrence, and E. A. Grolsman, "Lateral gene transfer and the nature of bacterial innovation," *Nature*, vol. 405, no. 6784, pp. 299–304, 2000.

[26] A. P. Roberts and P. Mullany, "A modular master on the move: the Tn916 family of mobile genetic elements," *Trends in Microbiology*, vol. 17, no. 6, pp. 251–258, 2009.

[27] R. A. F. Wozniak and M. K. Waldor, "Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow," *Nature Reviews Microbiology*, vol. 8, no. 8, pp. 552–563, 2010.

[28] J. O. Andersson, "Gene transfer and diversification of microbial eukaryotes," *Annual Review of Microbiology*, vol. 63, pp. 177–193, 2009.

[29] A. H. Paterson, M. Freeling, H. Tang, and X. Wang, "Insights from the comparison of plant genome sequences," *Annual Review of Plant Biology*, vol. 61, pp. 349–372, 2010.

[30] O. Jatllon, J. M. Aury, F. Brunet et al., "Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype," *Nature*, vol. 431, no. 7011, pp. 946–957, 2004.

[31] J. M. Aury, O. Jaillon, L. Duret et al., "Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*," *Nature*, vol. 444, no. 7116, pp. 171–178, 2006.

[32] K. H. Wolfe and D. C. Shields, "Molecular evidence for an ancient duplication of the entire yeast genome," *Nature*, vol. 387, no. 6634, pp. 708–713, 1997.

[33] M. Kellis, B. W. Birren, and E. S. Lander, "Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*," *Nature*, vol. 428, no. 6983, pp. 617–624, 2004.

[34] J. G. Ferry, "Methane from acetate," *Journal of Bacteriology*, vol. 174, no. 17, pp. 5489–5495, 1992.

[35] G. P. Fournier and J. P. Gogarten, "Evolution of acetoclastic methanogenesis in Methanosarcina via horizontal gene transfer from cellulolytic Clostridia," *Journal of Bacteriology*, vol. 190, no. 3, pp. 1124–1127, 2008.

[36] J. D. Woodson, C. L. Zayas, and J. C. Escalante-Semerena, "A new pathway for salvaging the coenzyme B12 precursor cobinamide in archaea requires cobinamide-phosphate synthase (CbiB) enzyme activity," *Journal of Bacteriology*, vol. 185, no. 24, pp. 7193–7201, 2003.

[37] P.-A. Christin, E. J. Edwards, G. Besnard et al., "Adaptive evolution of C_4 photosynthesis through recurrent lateral gene transfer," *Current Biology*, vol. 22, no. 5, pp. 445–449, 2012.

[38] R. E. Blankenship, "Molecular evidence for the evolution of photosynthesis," *Trends in Plant Science*, vol. 6, no. 1, pp. 4–6, 2001.

[39] J. Raymond, "Coloring in the tree of life," *Trends in Microbiology*, vol. 16, no. 2, pp. 41–43, 2008.

[40] J. Raymond, O. Zhaxybayeva, J. P. Gogarten, S. Y. Gerdes, and R. E. Blankenship, "Whole-genome analysis of photosynthetic prokaryotes," *Science*, vol. 298, no. 5598, pp. 1616–1620, 2002.

[41] J. Xiong and C. E. Bauer, "Complex evolution of photosynthesis," *Annual Review of Plant Biology*, vol. 53, pp. 503–521, 2002.

[42] D. Williams, C. P. Andam, and J. P. Gogarten, "Horizontal gene transfer and the formation of groups of microorganisms," in *Molecular Phylogeny of Microorganisms. Hethersett*, A. Oren and R. T. Papke, Eds., Caister Academic Press, Norwich, UK, 2010.

[43] J. Raymond, "The role of horizontal gene transfer in photosynthesis, oxygen production, and oxygen tolerance," *Methods in Molecular Biology*, vol. 532, pp. 323–338, 2009.

[44] S. Sadekar, J. Raymond, and R. E. Blankenship, "Conservation of distantly related membrane proteins: photosynthetic reaction centers share a common structural core," *Molecular Biology and Evolution*, vol. 23, no. 11, pp. 2001–2007, 2006.

[45] M. Khomyakova, O. Bukmez, L. K. Thomas, T. J. Erb, and I. A. Berg, "A methylaspartate cycle in haloarchaea," *Science*, vol. 331, no. 6015, pp. 334–337, 2011.

[46] C. H. Lester, N. Frimodt-Møller, T. L. Sørensen, D. L. Monnet, and A. M. Hammerum, "In vivo transfer of the vanA resistance gene from an *Enterococcus faecium* isolate of animal origin to an E. faecium isolate of human origin in the intestines of human volunteers," *Antimicrobial Agents and Chemotherapy*, vol. 50, no. 2, pp. 596–599, 2006.

[47] J. H. Hehemann, G. Correc, T. Barbeyron, W. Helbert, M. Czjzek, and G. Michel, "Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota," *Nature*, vol. 464, no. 7290, pp. 908–912, 2010.

[48] F. Thomas, T. Barbeyron, T. Tonon, S. Genicot, M. Czjzek, and G. Michel, "Characterization of the first alginolytic operons in a marine bacterium: from their emergence in marine Flavobacteriia to their independent transfers to marine Proteobacteria and human gut Bacteroides," *Environmental Microbiology*. In press.

[49] E. Rosenberg, G. Sharon, and I. Zilber-Rosenberg, "The hologenome theory of evolution contains Lamarckian aspects within a Darwinian framework," *Environmental Microbiology*, vol. 11, no. 12, pp. 2959–2962, 2009.

[50] B. S. Chevalier and B. L. Stoddard, "Homing endonucleases: structural and functional insight into the catalysts of intron/intein mobility," *Nucleic Acids Research*, vol. 29, no. 18, pp. 3757–3774, 2001.

[51] L. Olendzenski and J. P. Gogarten, "Evolution of genes and organisms: the tree/web of life in light of horizontal gene transfer," *Annals of the New York Academy of Sciences*, vol. 1178, pp. 137–145, 2009.

[52] B. Dujon, "Group I introns as mobile genetic elements: facts and mechanistic speculations—a review," *Gene*, vol. 82, no. 1, pp. 91–114, 1989.

[53] V. Koufopanou and A. Burt, "Degeneration and domestication of a selfish gene in yeast: molecular evolution versus site-directed mutagenesis," *Molecular Biology and Evolution*, vol. 22, no. 7, pp. 1535–1538, 2005.

[54] M. Dori-Bachash, B. Dassa, O. Peleg, S. A. Pineiro, E. Jurkevitch, and S. Pietrokovski, "Bacterial intein-like domains of predatory bacteria: a new domain type characterized in *Bdellovibrio bacteriovorus*," *Functional and Integrative Genomics*, vol. 9, no. 2, pp. 153–166, 2009.

[55] G. Amitai, O. Belenkiy, B. Dassa, A. Shainskaya, and S. Pietrokovski, "Distribution and function of new bacterial intein-like protein domains," *Molecular Microbiology*, vol. 47, no. 1, pp. 61–73, 2003.

[56] O. G. Wikmark, C. Einvik, J. F. De Jonckheere, and S. D. Johansen, "Short-term sequence evolution and vertical inheritance of the *Naegleria* twin-ribozyme group I intron," *BMC Evolutionary Biology*, vol. 6, article 39, 2006.

[57] B. Dassa, N. London, B. L. Stoddard, O. Schueler-Furman, and S. Pietrokovski, "Fractured genes: a novel genomic arrangement involving new split inteins and a new homing endonuclease family," *Nucleic Acids Research*, vol. 37, no. 8, pp. 2560–2573, 2009.

[58] J. Caspi, G. Amitai, O. Belenkiy, and S. Pietrokovski, "Distribution of split DnaE inteins in cyanobacteria," *Molecular Microbiology*, vol. 50, no. 5, pp. 1569–1577, 2003.

[59] T. R. Cech, "The generality of self-splicing RNA: relationship to nuclear mRNA splicing," *Cell*, vol. 44, no. 2, pp. 207–210, 1986.

[60] P. A. Sharp, "On the origin of RNA splicing and introns," *Cell*, vol. 42, no. 2, pp. 397–400, 1985.

[61] S. Zimmerly, H. Guo, P. S. Perlman, and A. M. Lambowitz, "Group II intron mobility occurs by target DNA-primed reverse transcription," *Cell*, vol. 82, no. 4, pp. 545–554, 1995.

[62] B. Cousineau, D. Smith, S. Lawrence-Cavanagh et al., "Retrohoming of a bacterial group II intron: mobility via complete reverse splicing, independent of homologous DNA recombination," *Cell*, vol. 94, no. 4, pp. 451–462, 1998.

[63] N. Toro, J. I. Jiménez-Zurdo, and F. M. García-Rodríguez, "Bacterial group II introns: not just splicing," *FEMS Microbiology Reviews*, vol. 31, no. 3, pp. 342–358, 2007.

[64] G. C. Shukla and R. A. Padgett, "A catalytically active group II intron domain 5 can function in the U12-dependent spliceosome," *Molecular Cell*, vol. 9, no. 5, pp. 1145–1150, 2002.

[65] H. D. Madhani and C. Guthrie, "A novel base-pairing interaction between U2 and U6 snRNAs suggests a mechanism for the catalytic activation of the spliceosome," *Cell*, vol. 71, no. 5, pp. 803–817, 1992.

[66] K. S. Keating, N. Toor, P. S. Perlman, and A. M. Pyle, "A structural analysis of the group II intron active site and implications for the spliceosome," *RNA*, vol. 16, no. 1, pp. 1–9, 2010.

[67] E. V. Koonin, "The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate?" *Biology Direct*, vol. 1, article 22, 2006.

[68] T. S. Kroeger, K. P. Watkins, G. Friso, K. J. Van Wijk, and A. Barkan, "A plant-specific RNA-binding domain revealed through analysis of chloroplast group II intron splicing," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 11, pp. 4537–4542, 2009.

[69] K. P. Watkins, M. Rojas, G. Friso, K. J. van Wijk, J. Meurer, and A. Barkan, "APO1 promotes the splicing of chloroplast group II introns and harbors a plant-specific zinc-dependent RNA binding domain," *Plant Cell*, vol. 23, no. 3, pp. 1082–1092, 2011.

[70] S. Leclercq, I. Giraud, and R. Cordaux, "Remarkable abundance and evolution of mobile group II introns in *Wolbachia* bacterial endosymbionts," *Molecular Biology and Evolution*, vol. 28, no. 1, pp. 685–697, 2011.

[71] J. Huang and J. P. Gogarten, "Ancient horizontal gene transfer can benefit phylogenetic reconstruction," *Trends in Genetics*, vol. 22, no. 7, pp. 361–366, 2006.

[72] F. Denoeud, M. Roussel, B. Noel et al., "Genome sequence of the stramenopile *Blastocystis*, a human anaerobic parasite," *Genome Biology*, vol. 12, no. 3, article R29, 2011.

[73] S. E. Kalla, D. C. Queller, A. Lasagni, and J. E. Strassmann, "Kin discrimination and possible cryptic species in the social amoeba *Polysphondylium violaceum*," *BMC Evolutionary Biology*, vol. 11, no. 1, article 31, 2011.

[74] J. P. McCarter, "Nematology: terra incognita no more," *Nature Biotechnology*, vol. 26, no. 8, pp. 882–884, 2008.

[75] M. Mitreva, G. Smant, and J. Helder, "Role of horizontal gene transfer in the evolution of plant parasitism among nematodes," *Methods in Molecular Biology*, vol. 532, pp. 517–535, 2009.

[76] G. Smant, J. P. W. G. Stokkermans, Y. Yan et al., "Endogenous cellulases in animals: isolation of β-1,4-endoglucanase genes from two species of plant-parasitic cyst nematodes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 9, pp. 4906–4911, 1998.

[77] C. Dieterich and R. J. Sommer, "How to become a parasite—lessons from the genomes of nematodes," *Trends in Genetics*, vol. 25, no. 5, pp. 203–209, 2009.

[78] Y. Sagane, K. Zech, J. M. Bouquet, M. Schmid, U. Bal, and E. M. Thompson, "Functional specialization of cellulose synthase genes of prokaryotic origin in chordate larvaceans," *Development*, vol. 137, no. 9, pp. 1483–1492, 2010.

[79] R. Acuña, B. E. Padilla, C. P. Flórez-Ramos et al., "Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 11, pp. 4197–4202, 2012.

[80] M. Grbić, T. Van Leeuwen, R. M. Clark et al., "The genome of *Tetranychus urticae* reveals herbivorous pest adaptations," *Nature*, vol. 479, no. 7374, pp. 487–492, 2011.

[81] N. A. Moran and T. Jarvik, "Lateral transfer of genes from fungi underlies carotenoid production in aphids," *Science*, vol. 328, no. 5978, pp. 624–627, 2010.

[82] M. T. Levine, C. D. Jones, A. D. Kern, H. A. Lindfors, and D. J. Begun, "Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 26, pp. 9935–9939, 2006.

[83] D. G. Knowles and A. McLysaght, "Recent de novo origin of human protein-coding genes," *Genome Research*, vol. 19, no. 10, pp. 1752–1759, 2009.

[84] S. Ohno, *Evolution by Gene Duplication*, Springer, Berlin, Germany, 1970.

[85] M. Long, E. Betrán, K. Thornton, and W. Wang, "The origin of new genes: glimpses from the young and old," *Nature Reviews Genetics*, vol. 4, no. 11, pp. 865–875, 2003.

[86] M. Lynch and J. S. Conery, "The evolutionary fate and consequences of duplicate genes," *Science*, vol. 290, no. 5494, pp. 1151–1155, 2000.

[87] M. W. Hahn, "Distinguishing among evolutionary models for the maintenance of gene duplicates," *Journal of Heredity*, vol. 100, no. 5, pp. 605–617, 2009.

[88] Y. Van De Peer, S. Maere, and A. Meyer, "The evolutionary significance of ancient genome duplications," *Nature Reviews Genetics*, vol. 10, no. 10, pp. 725–732, 2009.

[89] J. P. Gogarten and L. Olendzenski, "Orthologs, paralogs and genome comparisons," *Current Opinion in Genetics and Development*, vol. 9, no. 6, pp. 630–636, 1999.

[90] M. M. Goodman, C. W. Stuber, K. Newton, and H. H. Weissinger, "Linkage relationships of 19 enzyme Ioci in maize," *Genetics*, vol. 96, pp. 697–710, 1980.

[91] T. Helentjaris, D. Weber, and S Wright, "Identification of the genomic locations of duplicate nucleotide sequences in maize by analysis of restriction fragment length polymorphisms," *Genetics*, vol. 118, pp. 353–363, 1988.

[92] B. S. Gaut and J. F. Doebley, "DNA sequence evidence for the segmental allotetraploid origin of maize," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 13, pp. 6809–6814, 1997.

[93] S. White and J. Doebley, "Of genes and genomes and the origin of maize," *Trends in Genetics*, vol. 14, no. 8, pp. 327–332, 1998.

[94] M. Mena, B. A. Ambrose, R. B. Meeley, S. P. Briggs, M. F. Yanofsky, and R. J. Schmidt, "Diversification of C-function activity in maize flower development," *Science*, vol. 274, no. 5292, pp. 1537–1540, 1996.

[95] A. Force, M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, and J. Postlethwait, "Preservation of duplicate genes by complementary, degenerative mutations," *Genetics*, vol. 151, no. 4, pp. 1531–1545, 1999.

[96] T. J. Treangen and E. P. C. Rocha, "Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes," *PLoS Genetics*, vol. 7, no. 1, Article ID e1001284, 2011.

[97] J. Dzioba, C. C. Häse, K. Gosink, M. Y. Galperin, and P. Dibrov, "Experimental verification of a sequence-based prediction: F_1F_0-type ATPase of *Vibrio cholerae* transports protons, not Na^+ ions," *Journal of Bacteriology*, vol. 185, no. 2, pp. 674–678, 2003.

Hindawi Publishing Corporation
International Journal of Evolutionary Biology
Volume 2012, Article ID 708980, 11 pages
doi:10.1155/2012/708980

Review Article

Repeated Evolution of Testis-Specific New Genes: The Case of Telomere-Capping Genes in *Drosophila*

Raphaëlle Dubruille,[1] Gabriel A. B. Marais,[2] and Benjamin Loppin[1]

[1] *UMR 5534, Centre de Génétique et de Physiologie Moléculaire et Cellulaire, Centre National de la Recherche Scientifique, Université Claude Bernard Lyon 1, Université de Lyon, 69622 Villeurbanne, France*
[2] *UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Centre National de la Recherche Scientifique, Université Claude Bernard Lyon 1, 69622 Villeurbanne, France*

Correspondence should be addressed to Benjamin Loppin, benjamin.loppin@univ-lyon1.fr

Received 16 February 2012; Accepted 9 May 2012

Academic Editor: Hideki Innan

Comparative genome analysis has allowed the identification of various mechanisms involved in gene birth. However, understanding the evolutionary forces driving new gene origination still represents a major challenge. In particular, an intriguing and not yet fully understood trend has emerged from the study of new genes: many of them show a testis-specific expression pattern, which has remained poorly understood. Here we review the case of such a new gene, which involves a telomere-capping gene family in *Drosophila. hiphop* and its testis-specific paralog *K81* are critical for the protection of chromosome ends in somatic cells and male gametes, respectively. Two independent functional studies recently proposed that these genes evolved under a reproductive-subfunctionalization regime. The 2011 release of new *Drosophila* genome sequences from the *melanogaster* group of species allowed us to deepen our phylogenetic analysis of the *hiphop/K81* family. This work reveals an unsuspected dynamic of gene birth and death within the group, with recurrent duplication events through retroposition mechanisms. Finally, we discuss the plausibility of different evolutionary scenarios that could explain the diversification of this gene family.

1. Introduction

In the past decade, rapid progress has been made on the origin and evolution of new genes thanks to the genomics revolution [1]. Many cases of gene birth are now documented, and they have revealed that the mechanisms for new gene formation are surprisingly diverse. They include DNA-based duplication, RNA-based duplication (retroposition or retroduplication), gene fusion (chimerization), *de novo* gene origination, domestication of transposable elements, and horizontal gene transfer [1, 2]. Remarkably, many new genes show a male-biased expression and a majority of these are actually specifically expressed in the testis. Indeed, this organ seems to have a critical role in gene birth and evolution [1]. Two of the first documented cases of gene origination, *Pgk2* in mammals and *Jingwei* in *Drosophila,* are both testis-specific [3, 4]. More recent work on retroduplication showed an overall propensity of young retrogenes to be testis specific [5, 6]. Other types of new genes also tend to show testis-specificity or testis-biased transcription (e.g, [7–12]).

Several explanations have been proposed for this tendency of new genes to be testis specific [1, 13, 14]. The first explanation relies on a peculiarity of sex chromosome biology called MSCI (male sex chromosomes inactivation). In mammals and *C. elegans,* the sex chromosomes are inactivated during male meiosis, probably as a consequence of a general mechanism to avoid recombination between nonhomologous sequences [15, 16]. MSCI is expected to drive genes expressed during male meiosis out of the sex chromosomes. This was observed by looking at retrogenes in mammals [17–19]. In mice, in particular, it has been shown that the X parental genes are ubiquitously expressed except in testis, and this is complemented by a testis-specific expression of their daughter autosomal retrocopies

in agreement with the "escape from MSCI" hypothesis [19]. In *Drosophila*, the "exodus" of testis-specific genes out of X affects RNA-based [5] and possibly DNA-based duplicates [20]. However, the actual contribution of MSCI to this phenomenon and even the very existence of MSCI in *Drosophila* are actively debated issues [21–26].

Another hypothesis has been proposed to explain this above-mentioned pattern, especially in *Drosophila* where MSCI is controversial. It involves the interaction between dosage compensation (DC) and sex-biased expression [24, 27, 28]. The massive Y gene loss or silencing generates an imbalance of expression for X-linked genes compared to autosomal genes in males. DC mechanisms have evolved to even X and autosomal gene expression [29]. In mammals, where one X is inactivated in females, the X is hypertranscribed in both sexes, while in *Drosophila* the X is hypertranscribed in males [29, 30]. In *Drosophila*, male-biased genes have been shown to evolve mostly by hyperexpression in males only [28, 31]. However, such evolution of male-biased expression is difficult on the X chromosome because it is already hypertranscribed due to DC [24]. In agreement with this model, it has been shown that highly expressed male-biased genes are underrepresented on the X chromosome [28] and that dosage-compensated X genes tend to have autosomal retrocopies with male-biased expression [27].

However, the "escape from MSCI" and "escape from DC" hypotheses can only explain the evolution of new testis-specific genes involving the relocation out of the sex chromosomes but not those involving autosomes only. Another more general explanation has been recently proposed [13, 14]. In species with two sexes, mutations with sex antagonism (beneficial for one sex, deleterious for the other) can arise [32]. The presence of two sex-antagonistic alleles of a gene can cause an intralocus sexual antagonism [33]. Evolving sex-biased expression is a way to solve the conflict. However, this cannot work for housekeeping genes that need to be expressed in both sexes. In this case, duplication can resolve the intralocus sexual conflict, with the parental copy remaining expressed in both sexes and the new one being expressed only in one sex. Data in *Drosophila* suggests that testis is the tissue where sex antagonism is by far the strongest, and most male-biased genes are indeed expressed in testis [34]. In practice, solving intralocus sex conflict for housekeeping genes will imply getting a new copy expressed in testis [13, 14]. Gallach et al. [35] reported that 83% of the relocated copies of the mitochondrial genes found in the nuclear genome exhibit testis-specific expression. Importantly, about half of these relocation events involved autosomes only and could not be explained by the "escape from MSCI" and "escape from DC" hypotheses. dN/dS analysis of these genes suggested that the testis-specific copies tend to evolve under positive selection. Other examples of housekeeping genes show similar patterns, which fits well with the idea of resolving sexual conflict by duplication [13, 14, 36].

Finally, testis-specific new genes may be more common just because new genes arise more easily when expressed in the testis (the "out of testis" hypothesis, see [1]). In mammals, the chromatin in male germ cells is characterized by the presence of histone variants and histone marks favoring open chromatin, widespread demethylation of CpG-enriched promoters, and elevated levels of the transcription machinery components [37, 38]. Similarly, in *Drosophila* primary spermatocytes, very high level of transcriptional activity ensures the production of most mRNAs required for the postmeiotic differentiation program of male germ cells [2, 39]. This highly permissive state of chromatin as well as other peculiar features of male germ cells may have facilitated the expression of newly arisen genes in testis during their early evolution [40].

Many papers call for more functional studies of new genes. Here we review the case of the *hiphop*/*K81* telomere capping genes in *Drosophila*, for which detailed functional studies are available. We also present new results on the evolution of the *hiphop*/*K81* genes and discuss functional and evolutionary data with respect to the hypotheses presented above.

2. K81 as a Case of Reproductive Specialization of an Essential Telomere Protein

2.1. Drosophila Telomeres and Capping Proteins. Telomeres are essential structures at the end of eukaryotic chromosomes that are generally composed of highly repetitive DNA associated with specific proteins. The elongation of repetitive telomeric DNA counteracts the slow erosion of chromosome arms caused by the incomplete replication of DNA extremities at each S-phase. Telomere elongation is mediated in most eukaryotes by the conserved enzyme telomerase, a reverse transcriptase that adds small G-rich repeats, such as (TTAGGG)n, at the end of chromosomes. In addition, telomeres function as protective caps that prevent the recognition of chromosome ends as DNA double-strand breaks by the DNA repair machinery and their irreversible and deleterious ligation [41–43]. In most eukaryotes, this capping function is largely dependent on several DNA binding proteins that specifically recognize the small repeats added by the telomerase complex. *Drosophila* represents an exception in telomere biology as this model organism lacks telomerase. In this species, the "end replication problem" is solved in an original manner, by the controlled insertion of specialized telomeric retrotransposons at chromosome extremities [44]. Although repetitive by nature, *Drosophila* telomeric DNA thus lacks large arrays of small repeat motifs and associated binding proteins. Instead, the capping function of *Drosophila* telomeres is ensured by proteins that possess the remarkable ability to bind chromosome ends in a sequence-independent manner [45–47].

Well-characterized *Drosophila* capping proteins include HOAP, HP1a, Modigliani (Moi), Verrocchio (Ver), and HipHop [48–53]. Mutations affecting capping genes are all zygotic lethal and induce chromosome end-to-end fusions that are detectable in rapidly dividing cells. Telomere fusions form dicentric chromosomes that break in mitotic anaphase result in genomic instability. Despite their critical role for the maintenance of genome integrity, *Drosophila* capping proteins are rapidly evolving. With the exception of heterochromatin protein 1a (HP1a), which has additional functions in

the nucleus, and possibly the OB-fold containing protein Ver, other capping proteins do not seem to have any ortholog in yeasts, mammals, or plants [46].

2.2. K81, a Male Germline Paralog of the HipHop Capping Protein. The *Drosophila ms(3)K81* (*K81*) gene was originally identified through a unique male sterile mutation found in a Japanese population of *D. melanogaster* [54]. *K81* mutant males produce apparently normal sperm that are capable of fertilizing eggs. However, the resulting embryos invariably die before hatching, a phenotype which actually makes *K81* one of the very rare paternal effect, embryonic lethal mutations. Furthermore, eggs fertilized by *K81* mutant sperm develop as nonviable, aneuploid, or haploid embryos, after the loss of paternal chromosomes during the first zygotic nuclear division [54–56]. Despite the critical requirement of K81 for the integration of paternal chromosomes into the diploid zygote, its molecular identification unexpectedly revealed a small, intronless gene, encoding a nonconserved protein [55]. In fact, the *K81* gene appeared restricted to the nine species comprising the *melanogaster* subgroup. Loppin et al. [55] also identified another gene paralogous to *K81*, now known as *Hiphop*, which was present in species of the *melanogaster* subgroup as well as in the more distantly related *D. pseudoobscura* genome. The conserved synteny around the *hiphop* locus in *D. melanogaster* and *D. pseudoobscura* strongly indicated that *hiphop* was the ancestor gene, while *K81* appeared after the duplication of *hiphop* at the root of the *melanogaster* subgroup.

The *hiphop* gene is located in chromosome arm 3L and has a unique predicted intron immediately upstream its coding sequence. *hiphop* is expressed in most tissues at low to moderate levels, but it is also strongly transcribed in adult ovaries, suggesting that the HipHop protein is required during early embryo development. *hiphop* mutants are zygotic lethal and die in larval stages. In contrast to *hiphop*, *K81* expression is essentially restricted to the male germline [55], and adult flies homozygous for a *K81* null allele are viable. The *K81* gene (chromosome arm 3R) has no intron and presumably shares its 5′ regulatory sequences with its neighbor gene *Rb97D*, which is also strongly expressed in the testis. Taken together, these features fit well with a retroposition event at the origin of *K81* [55]. More recently, the independent findings that *hiphop* and *K81* encoded telomere capping proteins [57, 58] eventually provided the functional frame that was required to revisit the molecular evolution of these paralogs.

3. Evolution of K81: Functional Innovation or Reproductive Specialization?

HipHop and K81 are small proteins (221 and 184 residues, resp.) that do not display any known domain or motif [55]. HipHop was originally implicated in telomere biology through its physical interaction with the HOAP and HP1a capping proteins [50]. Furthermore, knocking down *hiphop* in somatic cells induces telomere fusions at high frequency indicating that HipHop is critical for the capping of chromosome ends. Finally, the HipHop protein is specifically enriched at telomeres and this localization occurs independently of the DNA sequence [50]. Similarly, K81 was demonstrated to associate with telomeres in the male germline, in a way similar to HipHop in somatic cells. Indeed, functional GFP::K81 fusion protein was observed at telomeres in spermatocytes as well as in postmeiotic spermatids and in mature gametes [57–59]. In spermatids, GFP::K81 accumulates into a small number of foci (that presumably correspond to clustered telomeres) that also contain HOAP and HP1a, but not HipHop. During the condensation of spermatid nuclei in *Drosophila* as in many animals, histones are massively replaced by sperm-specific, nonhistone chromosomal proteins such as protamines [60]. Interestingly, in the absence of K81, HOAP and HP1a are not maintained at telomeres after the histone-to-protamine transition, suggesting that K81 is required for the stability of the capping complex in the peculiar chromatin environment of condensing spermatid nuclei [57]. Using the GFP::K81 transgene allowed to demonstrate that the K81 capping protein remains associated with paternal telomeres until zygote formation, where it is required for the protection of paternally-transmitted telomeres [57]. Accordingly, in eggs fertilized by *K81* sperm, paternal chromatin bridges resulting from telomere fusions are observed during the first mitosis [55, 57, 58]. After fertilization, maternally provided HipHop progressively replaces K81 at paternal telomeres, which is no longer detectable after two or three nuclear cycles.

Why does *Drosophila melanogaster* require a second HipHop-related protein to protect telomeres in postmeiotic germ cells when other species outside the *melanogaster* subgroup only have a single *hiphop* gene? First, experimental evidence clearly indicates that HipHop is not capable to functionally replace K81 in the male germline. Indeed, a transgene expressing *hiphop* in male germ cells using the *K81* regulatory sequences cannot restore the fertility of *K81* mutant males [57, 58]. Interestingly however, *hiphop* is nevertheless capable of restoring HOAP and HP1a foci at telomeres in early spermatid nuclei, but all three capping proteins eventually disappear when histones are replaced with protamines [57]. Thus, these observations support the idea that K81 has become specialized in protecting telomeres in the highly peculiar chromatin environment of condensing spermatid nuclei. Second, and quite remarkably, the single HipHop-related protein of *D. virilis* (which lacks K81) was found associated with telomeres throughout spermiogenesis, strongly indicating that the ancestral *hiphop* gene in the *melanogaster* lineage was required to protect telomere in all cells, including male germ cells [58]. Taken together, these studies suggest a reproductive subfunctionalization by duplication-degeneration-complementation (DDC) (see [61]), in which the ancestral HipHop lost its ability to protect telomeres in postmeiotic germ cells after the gene duplication event. Meanwhile, the duplicated copy acquired male germline specific expression and specialized in the capping of telomeres in the peculiar sperm chromatin environment. This scenario is actually supported by the analysis of nonsynonymous and synonymous nucleotide

substitutions of *hiphop* and *K81* sequences, which indicated that these genes evolved under purifying selection as in the typical DDC model [57].

4. Diversification of the HipHop Protein Family: The Rule or the Exception?

Based on the first available twelve *Drosophila* sequenced genomes, the *hiphop*/*K81* duplication appeared specific to the *melanogaster* subgroup of species. Notably, the *hiphop* gene was found at the same genomic position in all species of the *Sophophora* subgenus while *K81* was restricted to the *melanogaster* subgroup [55]. This view, however, was biased by the absence of sequenced genomes belonging to the other subgroups comprising the *melanogaster* group. Indeed, this large and complex group includes at least ten subgroups with many species that can be partitioned into three main phylogenetic clades [62]. Recently, the genome sequences of eight additional species representative of several other subgroups in all three clades were released by the modENCODE consortium (modencode.org) and were made available by Flybase (flybase.org).

Interestingly, our combined BLAST analyses and microsynteny comparisons revealed an unsuspected diversification of the *hiphop* family in the *melanogaster* group (Figure 1). First, the *K81* gene is also present in species from four other subgroups belonging to clade III: *ficusphila*, *eugracilis*, *takahashi*, and *suzukii* (represented by *D. biarmipes*). Conversely, within the *melanogaster* group, *K81* is absent from the two available species from clade I (*D. ananassae* and *D. bipectinata*) and from the single representative species of clade II (*D. kikkawai*). Thus, *K81* appears to have a broader phylogenetic distribution than initially thought, and the gene duplication event at the origin of this gene probably occurred at the base of clade III (Figure 1).

We have also noticed the absence of *K81* in two species of clade III (*D. elegans* and *D. rhopaloa*) where it is presumably replaced by a paralog at another genomic position (*K81-like*, in orange in Figure 1). Interestingly, synteny block comparisons indicates that these two *K81-like* genes are apparently located on the X chromosome (Table 1), in contrast to the general tendency of testis-specific retrogenes to avoid the X [5, 20].

Finally, we observed that the original *hiphop* gene was independently lost or relocated at least at three occasions. In *D. bipectinata* and *D. ananassae* (clade I), *hiphop* is apparently replaced by a single *hiphop-like* paralog. In *D. ficusphila* and *D. takahashi* (clade III), *hiphop* is absent but one or two additional paralogs are present, in addition to the original *K81* gene. Interestingly, one of the new paralogs found in species of clade III (represented in light gray in Figure 1) is conserved between *D. ficusphila*, *D. elegans*, and *D. rhopaloa*, but not in the *D. eugracilis* lineage. Thus, the repertoire of *hiphop/K81* related genes in the *Drosophila* group of species is extremely dynamic, with multiple gene gains and losses observed at several levels of this radiation. Some species have three members of this gene family in their genomes, while all other species have either one or two. The fact that at least one *hiphop*-related capping gene is present in all *Drosophila* genomes sampled so far underlines the essential role of these genes for telomere protection. Importantly, the tendency of *hiphop* to duplicate is not restricted to the *melanogaster* group since an independent duplication of this gene occurred in the lineage leading to *D. willistoni* (*willistoni* group) ([57]; Figure 1).

Based on a combination of K81/HipHop protein alignment and complementation tests with mutant proteins, Gao and colleagues [58] proposed that a small QFVH motif near the C-terminus of K81 is critical for the protection of telomeres in mature gametes. Interestingly, in HipHop proteins from the *melanogaster* subgroup, this motif is replaced with a PTV tripeptide which functions in somatic cells but not in mature male gametes. However, non-*melanogaster* species that harbor a single *hiphop*-related gene display a "male germline-like" motif which also begins with a glutamine residue as in QFVH [58]. The presence of such a motif is probably important for these proteins to fulfill their role in all cells, including postmeiotic male germ cells. We have extended this analysis to the new available members of the family and found that this tendency is generally confirmed for the additional proteins. For instance, the K81-like proteins from *D. elegans* and *D. rhopaloa* (Figure 1 and Table 1). In addition, the single HipHop-like proteins from *D. bipectinata* (clade I) and *D. kikkawai* (clade II) have also a motif of the male germline type (QFLV). The only exception is *D. ananassae* where the single HipHop protein is apparently of the somatic type (PTII).

The highly dynamic repertoire of *K81*/*hiphop* genes reported here is remarkable and suggests that a constant evolutionary pressure is forcing this gene diversification (see below). One can wonder whether other telomere capping genes display a comparable level of evolutionary instability and, notably, those that are known to functionally interact with HipHop and K81. A great diversity of HP1 paralogs has already been documented in *Drosophila* [63], but the situation is complicated by the fact that HP1a is associated with several other important functions not related to telomere capping. In contrast, the other K81/HipHop partner HOAP is only required for telomere protection [48]. In *D. melanogaster*, the HOAP protein is encoded by a unique and essential gene named *caravaggio* (*cav*) [48]. HOAP is a fast evolving protein, which belongs to the *Drosophila* terminin complex of telomere proteins [46, 64]. This complex also contains two other proteins, Ver and Moi, which are also rapidly evolving as demonstrated by dN/dS analyses of their respective genes [46]. Interestingly, a recent study has reported the existence of three independent duplication of the *cav* gene outside the *melanogaster* group, in the *D. willistoni*, *D. virilis*, and *D. pseudoobscura/D. persimilis* lineages [65]. The presence of introns in these *cav* duplicates strongly suggests that these duplications occurred through a DNA-based mechanism. We found two additional independent duplication events in the recently released *melanogaster* group genomes (*D. ficusphila* and *D. elegans/D. rhopaloa* lineages) (Figure 2 and Table 2). Thus, although the presence of a syntenic *cav* gene in all *Drosophila* genomes sequenced so far indicates that this gene is probably

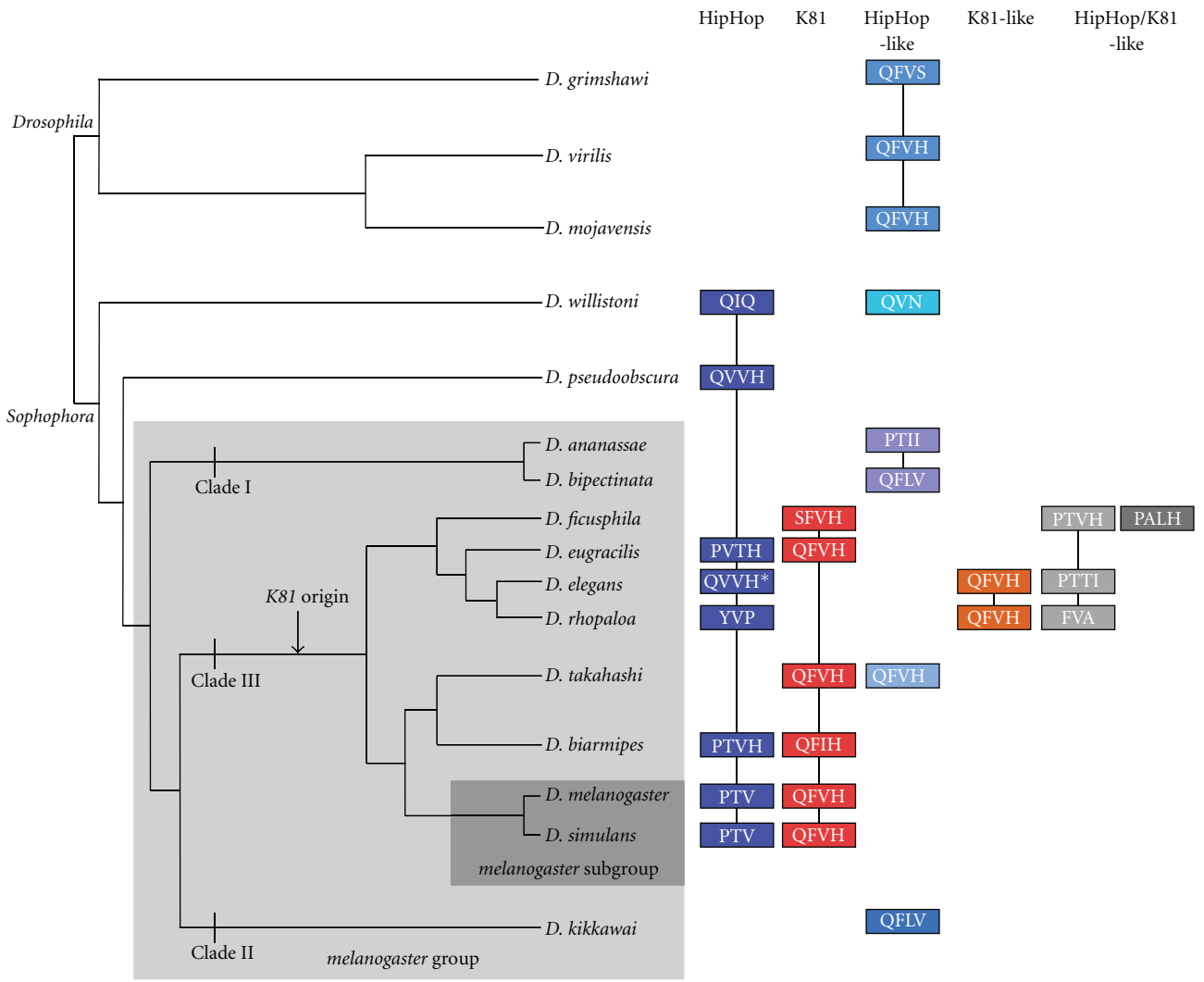


Figure 1: The HipHop/K81 protein family. A tree representing the schematic phylogeny of *Drosophila* species as described by Yang et al. [62]. For each species, the HipHop/K81-like proteins are represented as rectangles. These proteins were identified by tBLASTn search in Flybase (http://flybase.org/blast/). For the 8 new sequenced *Drosophila* genomes (*biarmipes, elegans, eugracilis, ficusphila, bipectinata, rhopaloa, takahashii,* and *kikkawai*), which are not yet annotated, an ORF corresponding to the protein was identified and used to determine the putative whole protein sequence (see also Table 1) except for *D. elegans*-HipHop due to poor sequence quality (∗). HipHop-like or K81-like proteins are proteins more closely related to HipHop or K81, respectively, whereas proteins whose phylogenetic origin was ambiguous HipHop or K81, are referred to as HipHop/K81-like proteins. For each protein, we identified the PTV or QFVH motif (see text) in the C-terminal domain that was described by Gao et al. [58] as responsible for the functional divergence between HipHop and K81 in sperm telomere protection. This PTV/ QFVH motif is indicated for each protein in the corresponding rectangle. BLAST analysis using 5 kbp upstream and downstream the *hiphop/K81*-like genes allowed to identify the orthologous region in the *melanogaster* genome. A same color code and a line connecting proteins indicate that the synteny block is conserved between the corresponding genes.

more ancient than *hiphop*, it is also subjected to recurrent duplication events.

5. What Evolutionary Forces Drive the Diversification of Telomere Genes?

If the functional partitioning of these paralogs is well established by experimental and phylogenetic analyses, we now face the challenge of understanding the nature of the evolutionary force responsible for the birth of *K81*. Escape from MSCI and escape from DC cannot explain the case of *hiphop*/*K81* since both parental and daughter copies are autosomal, at least in the *melanogaster* subgroup.

In the light of the duplication-degeneration-complementation classical model [61], the specialization of K81 in the capping of sperm telomeres as well as its restricted expression in the male germline are interpreted as the result of differential loss of function (i.e, subfunctionalization) of the duplicated copies [57–59]. In agreement with the DDC model, HipHop performs both somatic tissues and sperm-telomere capping in species without duplicates while in *D. melanogaster*, HipHop has lost its ability to protect

Table 1: hiphop and K81-like genes in *Drosophila*.

Species\gene	ID# or GI#	Orthologous region in *D. mel*	PTV/QFVH motif	Position of putative start codon	Position of stop codon
D. mel\hiphop	CG6874	*D.mel hiphop*	RR**PTV**-LDKQSMD		
D. mel\K81	CG14251	*D.mel K81*	RR**QFVH**LNREAMA		
D. sim\hiphop	GD14769	*D.mel hiphop*	RR**PTV**-LDKPSMD		
D. sim\K81	GD21311	*D.mel K81*	RR**QFVH**LNHQAMA		
D. bia\hiphop	358392949	*D.mel hiphop*	RR**PTVH**LNKEAMD	690387	689698
D. bia\K81	358402098	*D.mel K81*	RR**QFIH**LNKEAMD	2964671	2965297
D. tak\K81	343975433	*D.mel K81*	RR**QFVH**LNKEAMD	141804	141217
D. tak\hiphop-like	343974900	chro2L in fred gene	RR**QFVH**LNKEAMD	211517	212122
D. rho\hiphop	358405427	*D.mel hiphop*	RR**YVP**-LNKVAMD	33547	32867
D. rho\K81-like	358404732	chroX in Sh gene	RR**QFVH**LNKEAMD	683852	683265
D. rho\hiphop-K81-like-1	358405183	chroX Roc1a/CG13367	RR**FVA**-PNKEVMD	799350	800057
D. ele\hiphop	343972741	*D.mel hiphop*	RR**QVVH**PNKKAMD	ND	1725959
D. ele\K81-like	343972552	chroX in Sh gene	RR**QFVH**LNKNAMD	34447	35022
D. ele\hiphop-K81-like-1	343972719	chroX Roc1a/CG13367	RR**PTTI**LNKESMD	1005656	1006243
D. eug\hiphop	358409234	*D.mel hiphop*	RR**PVTH**LNKEAMD	677060	676191
D. eug\K81	358409002	*D.mel K81*	RR**QFVH**LNKEAME	154852	155409
D. fic\K81	343464569	*D.mel K81*	RR**SFVH**LNKEAMD	2599414	2600109
D. fic\hiphop-K81-like-1	343464682	chroX Roc1a/CG13367	RR**PTVH**LNKEAMD	461020	461505
D. fic\hiphop-K81-like-2	343464675	chro2L CG34163/zuc	RR**PALH**LNKEAMD	185420	184971
D. kik\hiphop-like	343973849	chro2L bsf/CR43344	RR**QFLV**PNKKVMD	92534	92040
D. ana\hiphop	GF10272	chro3L YT521-B/Drs	RR**PTII**LNKAVMD		
D. bip\hiphop	358403122	chro3L YT521-B/Drs	RR**QTVI**LNKAAMD	1284107	1283427
D. pse\hiphop	GA19922	*D.mel hiphop*	RR**QVVH**LNKTAMD		
D. wil\hiphop	GK12110	*D.mel hiphop*	RR**QIQ**-LTGPHLD		
D. wil\hiphop-like	GK15167	chro2L Or33c/Cry	RR**QVN**-RSGIDLD		
D. moj\hiphop-like	GI17239	chro2L CG13398	RR**QFVH**LNKDVMD		
D. vir\hiphop-like	GJ17998	chro2L CG13398	RR**QFVH**LNKDVMD		
D. gri\hiphop-like	GH13489	chro2L CG13398	RR**QFVS**LNKDVMD		

The *hiphop* and *K81*-like genes were identified by tBLASTn search in Flybase (http://flybase.org/blast/).

For each gene, the ID number, when available, is indicated. For species whose genome is not yet annotated, a GI number corresponding to the scaffold DNA sequence is indicated with the position in the scaffold of the putative start codon (first methionine in phase with the homolog protein identified) and the stop codon. ND: not determined.

The orthologous region in the *D. melanogaster* genome surrounding the *hiphop-K81*-like gene is indicated as follows: chromosome and neighbor genes. When two genes are indicated, the *hiphop-K81*-like gene is placed in between. D. mel hiphop and D. mel K81 means that the synteny block is conserved between the gene of interest and *hiphop* or *K81* from *D. melanogaster*, respectively.

The PTV or QFVH motifs of the HipHop/K81 proteins as defined by Gao et al. [58] are highlighted in bold.

D. mel: Drosophila melanogaster; D. sim: D. simulans; D. bia: D. biarmipes; D. tak: D. takahashi; D. rho: D. rhopaloa; D. ele: D. elegans; D. eug: D. eugracilis; D. fic: D. ficusphila; D. kik: D. kikkawai; D. ana: D. ananassae; D. bip: D. bipectinata; D. pse: D. pseudoobscura; D. wil: D. willistoni; D. moj: D. mojavensis; D. vir: D. virilis; D. gri: D. grimshawi.

chromosome ends in spermatids. Indeed, HipHop cannot replace K81 in complementation experiments. However, a simple subfunctionalization scenario does not predict the observed recurrent duplications of these capping genes that we have found here. A possibility is that the expression of a gene in testis increases the chance to get a testis-specific duplicate for mechanistic reasons (see the "out of testis" hypothesis in the introduction).

The high gene turnover observed within the *hiphop*/*K81* gene family is more consistent with ongoing sexual conflicts, as recently proposed by Gallach and Betrán [13]. Their model states that a preexisting sexual conflict between different

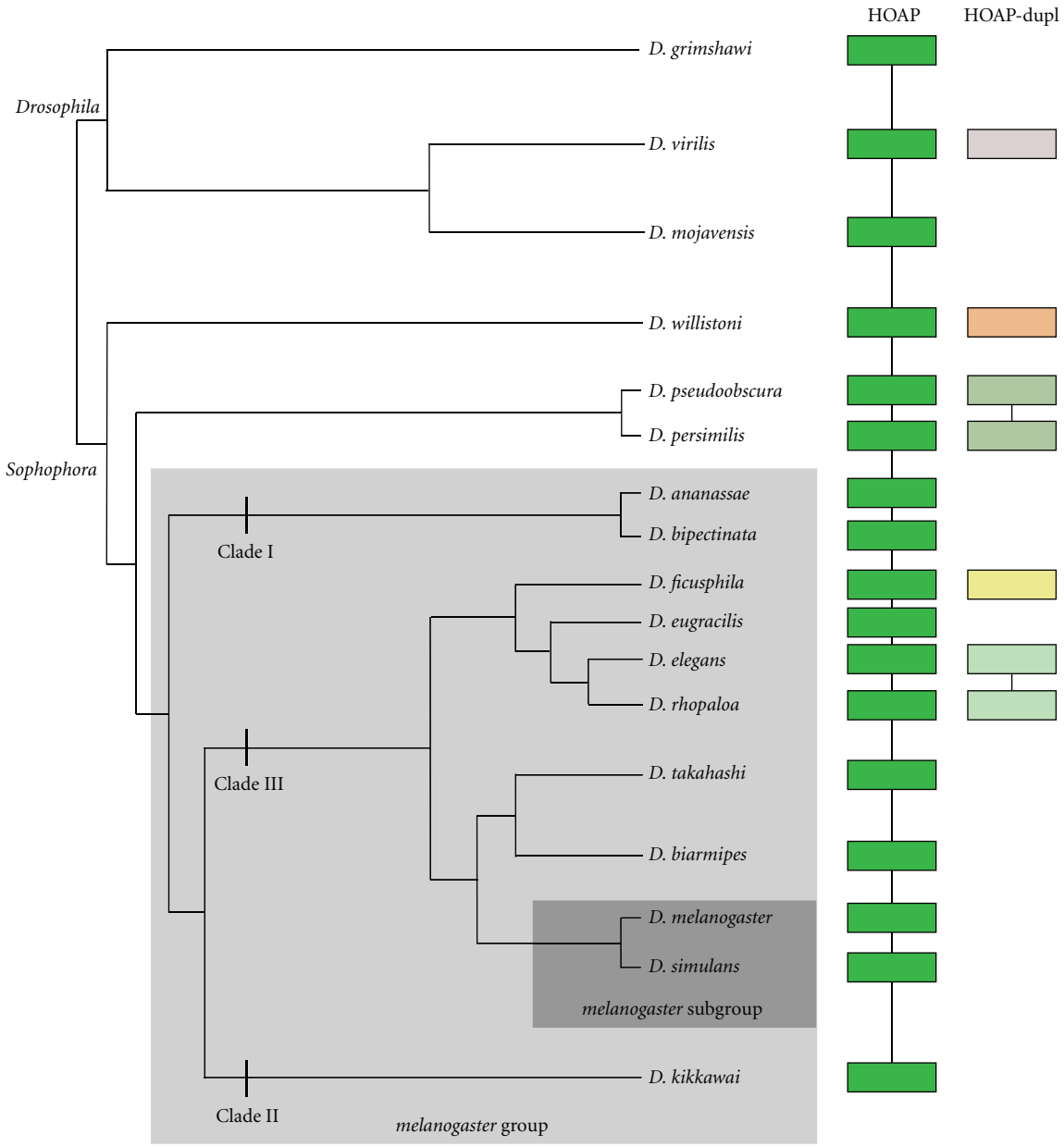


Figure 2: The HOAP protein family. The HOAP proteins were identified by tBLASTn search in Flybase (http://flybase.org/blast/) and are represented as rectangles. For the unannotated new sequenced genomes, BLAST analysis using 5 kbp upstream and downstream the *cav* and *cav-dupl* genes allowed to identify the orthologous region in the *melanogaster* genome. A same color code and a line connecting proteins indicate that the synteny block is conserved between the corresponding genes.

alleles of a gene could be solved by a gene duplication event and the acquisition of testis-specific expression of the duplicate. They propose that many testis-specific gene duplicates could have emerged under this scenario, including *K81*. This model implies that the ancestor gene plays an essential function in all cells (housekeeping gene), which is indeed the case of *hiphop*, as demonstrated by its associated lethal mutant phenotype and its critical protection of chromosome ends. In contrast, the duplicate gene *K81* is specifically expressed in the male germline and is first detected in spermatocytes. The K81 protein then decorates telomeres throughout spermiogenesis (postmeiotic stages of spermatogenesis) and after fertilization on paternal chromosomes [57, 58]. *hiphop* is actually also expressed in the male germline, but only in premeiotic cells [58]. Moreover, as previously mentioned, complementation analyses have demonstrated that HipHop and K81 have functionally diverged. Although this divergence could reflect an initial sexual conflict between different allelic variants of the ancestor protein, the Gallach and Betrán model states that acquisition of a testis-specific duplicate could also solve conflicting constraints on the expression of a ubiquitous parental gene. Indeed, genes that

Table 2: *cav* and *cav*-like genes in *Drosophila*.

Species\gene	ID# or GI#	Orthologous region in *D. melanogaster*
D. mel\cav	CG6219	*D.mel cav*
D. sim\cav	GD21077	*D.mel cav*
D. bia\cav	358402078	*D.mel cav*
D. tak\cav	343975000	*D.mel cav*
D. rho\cav	358405209	*D.mel cav*
D. rho\cav-dupl	358407419	chro2R CG1441/CG1513
D. ele\cav	343972724	*D.mel cav*
D. ele\cav-dupl	343972624	chro2R CG1441/CG1513
D. eug\cav	358408974	*D.mel cav*
D. fic\cav	343464694	*D.mel cav*
D. fic\cav-dupl	343464518	chro3L Eip74EF/CG7510
D. kik\cav	343973540	*D.mel cav*
D. ana\cav	GF16116	*D.mel cav*
D. bip\hiphop	358402982	*D.mel cav*
D. pse\cav	GA27250	*D.mel cav*
D. pse\cav-dupl	GA26940	chro3R CG2218/CG15536
D. per\cav	GL23417	*D.mel cav*
D. per\cav-dupl	GL14051	chro3R CG2218/CG15536
D. wil\cav	GK11387	*D.mel cav*
D. wil\cav-dupl	GK24325	chro2L jhamt
D. moj\cav	GI24179	*D.mel cav*
D. vir\cav	GJ14215	*D.mel cav*
D. vir\cav-dupl	GJ17001	chroX Upf2
D. gri\cav	GH18668	*D.mel cav*

The *cav* genes and duplications in the 8 new sequenced *Drosophila* genomes were identified by tBLASTn search in Flybase (http://flybase.org/blast/). For these genes, a GI number corresponding to the scaffold DNA sequence is indicated. *cav* homologs and duplication in other species are identified with their ID number.

For each gene, the orthologous region in the *D. melanogaster* genome surrounding the identified *cav* homologous gene is indicated as follows: chromosome and neighbor genes. D. mel cav means that the synteny block is conserved between the gene of interest and *cav* from *Drosophila melanogaster*.

D. mel: D. melanogaster; D. sim: D. simulans; D. bia: D. biarmipes; D. tak: D. takahashi; D. rho: D. rhopaloa; D. ele: D. elegans; D. eug: D. eugracilis; D. fic: D. ficusphila; D. kik: D. kikkawai; D. ana: D. ananassae; D. bip: D. bipectinata; D. pse: D. pseudoobscura; D. wil: D. willistoni; D. moj: D. mojavensis; D. vir: D. virilis; D. gri: D. grimshawi.

are specifically expressed in male germ cells are characterized by peculiar 5' regulatory elements and are often clustered in genome regions, suggesting the existence of higher order chromatin structure that favors transcription in spermatocytes or even in postmeiotic spermatids (reviewed in [2]). In this context, the existence of a duplicated copy could provide a more robust expression in the male germline than the ubiquitously expressed parental gene. This prediction could be experimentally tested by comparing the expression of *hiphop/K81*-like genes in species where a duplication has occurred or not.

These features fit with the possible existence of an initial sexual antagonism at the ancestor locus, which has been resolved by duplication followed by the specialization of the new copy. The fact that *hiphop* is actually expressed in the male germline is in apparent contradiction with this hypothesis. However, the critical difference between *hiphop* and *K81* is their differential expression in postmeiotic germ cells. Indeed, *K81* regulatory sequences drive robust and specific expression of K81 in spermatids, while the ubiquitously expressed HipHop is essentially excluded from these differentiating cells.

Thus, the birth of *K81* may have removed this possible source of conflict at the ancestor locus. In this model, telomere capping genes that do not function in postmeiotic male germ cells are not expected to give rise to testis-specific duplicates. It would thus be interesting to investigate the distribution and function of other essential telomere capping genes in the male germline, such as *Ver* and *Moi,* that do not show any duplicate in the species analyzed in the present study (not shown). Interestingly, our phylogenetic analysis of the *cav* (HOAP) gene revealed a rather different diversification pattern. In contrast to *hiphop*, *cav* duplication events seem to occur only through a DNA-based mechanism, and we did not observe any obvious correlation between the *hiphop*/*K81* and the *cav* respective diversification patterns. *cav* is notably characterized by the presence of a fixed parental gene throughout the analyzed genomes, which is not the case for *hiphop*. Reis et al. [65] observed that the *D. willistoni cav-dup* is specifically (albeit weakly) expressed in males,

but the other *cav* duplicates are expressed in both sexes. Thus, despite their apparent close functional relationship, these telomere genes are probably not subjected to the same evolutionary constraints. In addition, the functional status of *cav* in spermatids and sperm remains to be established.

6. Concluding Remarks

The molecular identification of the *K81* paternal effect gene about a decade ago was soon followed by the surprising observation that this essential male fertility gene in *D. melanogaster* was absent in the only other sequenced *Drosophila* genome available at that time (*D. pseudoobscura*) [55]. We now know that the acquisition of essential functions by recently evolved genes is not exceptional. A large-scale functional analysis of recently arisen genes in *Drosophila* revealed that most of them rapidly acquire essential developmental functions [66]. The functional characterization of new genes is invaluable to approach the intimacy of the evolutionary forces responsible for their origination and selection. Our phylogenetic analysis of the *hiphop*/*K81* gene family over twenty *Drosophila* species has revealed a highly dynamic pattern of gene gains and losses. Instead of our initial vision of a sporadic event specifically affecting the *melanogaster* subgroup, the *hiphop/K81* family is apparently subjected to a constant diversification. Future work should aim at determining if this diversification is compatible with the resolution of a sexual antagonism or with the "out of testis" hypothesis.

Acknowledgments

The authors thank Béatrice Horard and Pierre Couble for helpful discussions. The authors are grateful to modENCODE and Flybase for having made publicly available the new *Drosophila* genome sequences before publication. This work was supported by the Agence Nationale de la Recherche (ANR-08-BLAN-0139-01) and by the Fondation ARC pour la Recherche sur le Cancer (PDF20110603152).

References

[1] H. Kaessmann, "Origins, evolution, and phenotypic impact of new genes," *Genome Research*, vol. 20, no. 10, pp. 1313–1326, 2010.

[2] H. White-Cooper and N. Bausek, "Evolution and spermatogenesis," *Philosophical Transactions of the Royal Society B*, vol. 365, no. 1546, pp. 1465–1480, 2010.

[3] M. Long and C. H. Langley, "Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila," *Science*, vol. 260, no. 5104, pp. 91–95, 1993.

[4] J. R. McCarrey and K. Thomas, "Human testis-specific PGK gene lacks introns and possesses characteristics of a processed gene," *Nature*, vol. 326, no. 6112, pp. 501–505, 1987.

[5] E. Betrán, K. Thornton, and M. Long, "Retroposed new genes out of the X in Drosophila," *Genome Research*, vol. 12, no. 12, pp. 1854–1859, 2002.

[6] A. C. Marques, I. Dupanloup, N. Vinckenbosch, A. Reymond, and H. Kaessmann, "Emergence of young human genes after a burst of retroposition in primates," *PLoS Biology*, vol. 3, no. 11, p. e357, 2005.

[7] D. J. Begun, H. A. Lindfors, A. D. Kern, and C. D. Jones, "Evidence for de novo evolution of testis-expressed genes in the Drosophila yakuba/Drosophila erecta clade," *Genetics*, vol. 176, no. 2, pp. 1131–1137, 2007.

[8] S. T. Chen, H. C. Cheng, D. A. Barbash, and H. P. Yang, "Evolution of hydra, a recently evolved testis-expressed gene with nine alternative first exons in Drosophila melanogaster," *PLoS Genetics*, vol. 3, no. 7, p. e107, 2007.

[9] T. J. Heinen, F. Staubach, D. Häming, and D. Tautz, "Emergence of a new gene from an intergenic region," *Current Biology*, vol. 19, no. 18, pp. 1527–1531, 2009.

[10] M. T. Levine, C. D. Jones, A. D. Kern, H. A. Lindfors, and D. J. Begun, "Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 26, pp. 9935–9939, 2006.

[11] C. A. Paulding, M. Ruvolo, and D. A. Haber, "The Tre2 (USP6) oncogene is a hominoid-specific gene," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 5, pp. 2507–2511, 2003.

[12] X. She, J. E. Horvath, Z. Jiang et al., "The structure and evolution of centromeric transition regions within the human genome," *Nature*, vol. 430, no. 7002, pp. 857–864, 2004.

[13] M. Gallach and E. Betrán, "Intralocus sexual conflict resolved through gene duplication," *Trends in Ecology & Evolution*, vol. 26, no. 5, pp. 222–228, 2011.

[14] M. Gallach, S. Domingues, and E. Betran, "Gene duplication and the genome distribution of sex-biased genes," *International Journal of Evolutionary Biology*, vol. 2011, Article ID 989438, 20 pages, 2011.

[15] W. G. Kelly, C. E. Schaner, A. F. Dernburg et al., "X-chromosome silencing in the germline of C. elegans," *Development*, vol. 129, no. 2, pp. 479–492, 2002.

[16] J. M. Turner, "Meiotic sex chromosome inactivation," *Development*, vol. 134, no. 10, pp. 1823–1831, 2007.

[17] J. J. Emerson, H. Kaessmann, E. Betrán, and M. Long, "Extensive gene traffic on the mammalian X chromosome," *Science*, vol. 303, no. 5657, pp. 537–540, 2004.

[18] H. Kaessmann, N. Vinckenbosch, and M. Long, "RNA-based gene duplication: mechanistic and evolutionary insights," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 19–31, 2009.

[19] L. Potrzebowski, N. Vinckenbosch, A. C. Marques, F. Chalmel, B. Jégou, and H. Kaessmann, "Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes," *PLoS Biology*, vol. 6, no. 4, p. e80, 2008.

[20] M. D. Vibranovski, Y. Zhang, and M. Long, "General gene movement off the X chromosome in the Drosophila genus," *Genome Research*, vol. 19, no. 5, pp. 897–903, 2009.

[21] W. Hense, J. F. Baines, and J. Parsch, "X chromosome inactivation during Drosophila spermatogenesis," *PLoS Biology*, vol. 5, no. 10, p. e273, 2007.

[22] C. D. Meiklejohn, E. L. Landeen, J. M. Cook, S. B. Kingan, and D. C. Presgraves, "Sex chromosome-specific regulation in the Drosophila male germline but little evidence for chromosomal dosage compensation or meiotic inactivation," *PLoS Biology*, vol. 9, no. 8, Article ID e1001126, 2011.

[23] R. P. Meisel, M. V. Han, and M. W. Hahn, "A complex suite of forces drives gene traffic from Drosophila X chromosomes," *Genome Biology & Evolution*, vol. 1, pp. 176–188, 2009.

[24] L. M. Mikhaylova and D. I. Nurminsky, "Lack of global meiotic sex chromosome inactivation, and paucity of tissue-specific gene expression on the Drosophila X chromosome," *BMC Biology*, vol. 9, article 29, 2011.

[25] M. D. Vibranovski, H. F. Lopes, T. L. Karr, and M. Long, "Stage-specific expression profiling of Drosophila spermatogenesis suggests that meiotic sex chromosome inactivation drives genomic relocation of testis-expressed genes," *PLoS Genetics*, vol. 5, no. 11, Article ID e1000731, 2009.

[26] Y. E. Zhang, M. D. Vibranovski, B. H. Krinsky, and M. Long, "Age-dependent chromosomal distribution of male-biased genes in Drosophila," *Genome Research*, vol. 20, no. 11, pp. 1526–1533, 2010.

[27] D. Bachtrog, N. R. Toda, and S. Lockton, "Dosage compensation and demasculinization of X chromosomes in Drosophila," *Current Biology*, vol. 20, no. 16, pp. 1476–1481, 2010.

[28] B. Vicoso and B. Charlesworth, "The deficit of male-biased genes on the D. melanogaster X chromosome is expression-dependent: a consequence of dosage compensation?" *Journal of Molecular Evolution*, vol. 68, no. 5, pp. 576–583, 2009.

[29] T. Straub and P. B. Becker, "Dosage compensation: the beginning and end of generalization," *Nature Reviews Genetics*, vol. 8, no. 1, pp. 47–57, 2007.

[30] X. Deng, J. B. Hiatt, D. K. Nguyen et al., "Evidence for compensatory upregulation of expressed X-linked genes in mammals, Caenorhabditis elegans and Drosophila melanogaster," *Nature Genetics*, vol. 43, no. 12, pp. 1179–1185, 2011.

[31] T. Connallon and L. L. Knowles, "Intergenomic conflict revealed by patterns of sex-biased gene expression," *Trends in Genetics*, vol. 21, no. 9, pp. 495–499, 2005.

[32] W. R. Rice, "Sexually antagonistic genes: experimental evidence," *Science*, vol. 256, no. 5062, pp. 1436–1439, 1992.

[33] R. Bonduriansky and S. F. Chenoweth, "Intralocus sexual conflict," *Trends in Ecology & Evolution*, vol. 24, no. 5, pp. 280–288, 2009.

[34] Y. Zhang, D. Sturgill, M. Parisi, S. Kumar, and B. Oliver, "Constraint and turnover in sex-biased gene expression in the genus Drosophila," *Nature*, vol. 450, no. 7167, pp. 233–237, 2007.

[35] M. Gallach, C. Chandrasekaran, and E. Betrán, "Analyses of nuclearly encoded mitochondrial genes suggest gene duplication as a mechanism for resolving intralocus sexually antagonistic conflict in Drosophila," *Genome Biology & Evolution*, vol. 2, pp. 835–850, 2010.

[36] N. Phadnis, E. Hsieh, and H. S. Malik, "Birth, death and replacement of karyopherins in Drosophila," *Molecular Biology & Evolution*, vol. 29, no. 5, pp. 1429–1440, 2012.

[37] S. Kimmins and P. Sassone-Corsi, "Chromatin remodelling and epigenetic features of germ cells," *Nature*, vol. 434, no. 7033, pp. 583–589, 2005.

[38] K. C. Kleene, "A possible meiotic function of the peculiar patterns of gene expression in mammalian spermatogenic cells," *Mechanisms of Development*, vol. 106, no. 1-2, pp. 3–23, 2001.

[39] M. T. Fuller, "Spermatogenesis," in *The Development of Drosophila Melanogaster*, M. Bate and A. M. Arias, Eds., pp. 71–147, Cold Spring Harbor Laboratory Press, 1993.

[40] K. C. Kleene, "Sexual selection, genetic conflict, selfish genes, and the atypical patterns of gene expression in spermatogenic cells," *Developmental Biology*, vol. 277, no. 1, pp. 16–26, 2005.

[41] J. P. Murnane, "Telomere dysfunction and chromosome instability," *Mutation Research*, vol. 730, no. 1-2, pp. 28–36, 2012.

[42] R. J. O'Sullivan and J. Karlseder, "Telomeres: protecting chromosomes against genome instability," *Nature Reviews Molecular Cell Biology*, vol. 11, no. 3, pp. 171–181, 2010.

[43] W. Palm and T. de Lange, "How shelterin protects mammalian telomeres," *Annual Review of Genetics*, vol. 42, pp. 301–334, 2008.

[44] M. L. Pardue and P. G. DeBaryshe, "Retrotransposons that maintain chromosome ends," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 51, pp. 20317–20324, 2011.

[45] S. Pimpinelli, "Drosophila telomeres," in *Telomeres*, T. L. De Lange and V. Blackburn, Eds., pp. 433–459, Cold Spring Harbor Laboratory Press, 2nd edition, 2006.

[46] G. D. Raffa, L. Ciapponi, G. Cenci, and M. Gatti, "Terminin: a protein complex that mediates epigenetic maintenance of Drosophila telomeres," *Nucleus*, vol. 2, no. 5, pp. 383–391, 2011.

[47] Y. S. Rong, "Telomere capping in Drosophila: dealing with chromosome ends that most resemble DNA breaks," *Chromosoma*, vol. 117, no. 3, pp. 235–242, 2008.

[48] G. Cenci, G. Siriaco, G. D. Raffa, R. Kellum, and M. Gatti, "The drosophila HOAP protein in required for telomere capping," *Nature Cell Biology*, vol. 5, no. 1, pp. 82–84, 2003.

[49] L. Fanti, G. Giovinazzo, M. Berloco, and S. Pimpinelli, "The heterochromatin protein 1 prevents telomere fusions in Drosophila," *Molecular Cell*, vol. 2, no. 5, pp. 527–538, 1998.

[50] G. Gao, J. C. Walser, M. L. Beaucher et al., "HipHop interacts with HOAP and HP1 to protect Drosophila telomeres in a sequence-independent manner," *The EMBO Journal*, vol. 29, no. 4, pp. 819–829, 2010.

[51] B. Perrini, L. Piacentini, L. Fanti et al., "HP1 controls telomere capping, telomere elongation, and telomere silencing by two different mechanisms in Drosophila," *Molecular Cell*, vol. 15, no. 3, pp. 467–476, 2004.

[52] G. D. Raffa, D. Raimondo, C. Sorino et al., "Verrocchio, a Drosophila OB fold-containing protein, is a component of the terminin telomere-capping complex," *Genes & Development*, vol. 24, no. 15, pp. 1596–1601, 2010.

[53] G. D. Raffa, G. Siriaco, S. Cugusi et al., "The Drosophila modigliani (moi) gene encodes a HOAP-interacting protein required for telomere protection," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 7, pp. 2271–2276, 2009.

[54] Y. Fuyama, "Gynogenesis in Drosophila," *The Japanese Journal of Genetics*, vol. 59, no. 1, pp. 91–96, 1984.

[55] B. Loppin, D. Lepetit, S. Dorus, P. Couble, and T. L. Karr, "Origin and neofunctionalization of a Drosophila paternal effect gene essential for zygote viability," *Current Biology*, vol. 15, no. 2, pp. 87–93, 2005.

[56] G. K. Yasuda, G. Schubiger, and B. T. Wakimoto, "Genetic characterization of ms(3)K81, a paternal effect gene of Drosophila melanogaster," *Genetics*, vol. 140, no. 1, pp. 219–229, 1995.

[57] R. Dubruille, G. A. Orsi, L. Delabaere et al., "Specialization of a drosophila capping protein essential for the protection of sperm telomeres," *Current Biology*, vol. 20, no. 23, pp. 2090–2099, 2010.

[58] G. Gao, Y. Cheng, N. Wesolowska, and Y. S. Rong, "Paternal imprint essential for the inheritance of telomere identity in Drosophila," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 12, pp. 4932–4937, 2011.

[59] R. Dubruille and B. Loppin, "Epigenetic maintenance of telomere identity in Drosophila: buckle up for the sperm ride," *Cell Cycle*, vol. 10, no. 7, pp. 1037–1042, 2011.

[60] C. Rathke, W. M. Baarends, S. Jayaramaiah-Raja, M. Bartkuhn, R. Renkawitz, and R. Renkawitz-Pohl, "Transition from a nucleosome-based to a protamine-based chromatin configuration during spermiogenesis in Drosophila," *Journal of Cell Science*, vol. 120, part 9, pp. 1689–1700, 2007.

[61] A. Force, M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, and J. Postlethwait, "Preservation of duplicate genes by complementary, degenerative mutations," *Genetics*, vol. 151, no. 4, pp. 1531–1545, 1999.

[62] Y. Yang, Z. C. Hou, Y. H. Qian, H. Kang, and Q. T. Zeng, "Increasing the data size to accurately reconstruct the phylogenetic relationships between nine subgroups of the Drosophila melanogaster species group (Drosophilidae, Diptera)," *Molecular Phylogenetics & Evolution*, vol. 62, no. 1, pp. 214–223, 2012.

[63] D. Vermaak and H. S. Malik, "Multiple roles for heterochromatin protein 1 genes in drosophila," *Annual Review of Genetics*, vol. 43, pp. 467–492, 2009.

[64] K. J. Schmid and D. Tautz, "A screen for fast evolving genes from Drosophila," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 18, pp. 9746–9750, 1997.

[65] M. Reis, S. Sousa-Guimarães, C. P. Vieira, C. E. Sunkel, and J. Vieira, "Drosophila genes that affect meiosis duration are among the meiosis related genes that are more often found duplicated," *PLoS ONE*, vol. 6, no. 3, Article ID e17512, 2011.

[66] S. Chen, Y. E. Zhang, and M. Long, "New genes in Drosophila quickly become essential," *Science*, vol. 330, no. 6011, pp. 1682–1685, 2010.

Hindawi Publishing Corporation
International Journal of Evolutionary Biology
Volume 2012, Article ID 821645, 9 pages
doi:10.1155/2012/821645

Review Article

Novel Genes from Formation to Function

Rita Ponce,[1] Lene Martinsen,[2,3] Luís M. Vicente,[4] and Daniel L. Hartl[3]

[1] *Centro de Biologia Ambiental (CBA), Faculdade de Ciências da Universidade de Lisboa, 1749 Lisboa, Portugal*
[2] *Centre for Ecological and Evolutionary Synthesis (CEES), Department of Biology, University of Oslo, Blindern, 0316 Oslo, Norway*
[3] *Department of Organismic and Evolutionary Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA*
[4] *Centre for Environmental and Marine Studies (CESAM), Faculdade de Ciências da Universidade de Lisboa, 1749 Lisboa, Portugal*

Correspondence should be addressed to Rita Ponce, arponce@fc.ul.pt

Received 7 February 2012; Accepted 26 April 2012

Academic Editor: Frédéric Brunet

The study of the evolution of novel genes generally focuses on the formation of new coding sequences. However, equally important in the evolution of novel functional genes are the formation of regulatory regions that allow the expression of the genes and the effects of the new genes in the organism as well. Herein, we discuss the current knowledge on the evolution of novel functional genes, and we examine in more detail the youngest genes discovered. We examine the existing data on a very recent and rapidly evolving cluster of duplicated genes, the *Sdic* gene cluster. This cluster of genes is an excellent model for the evolution of novel genes, as it is very recent and may still be in the process of evolving.

1. Introduction

The availability of complete genome sequences allows the comparison of genomes, thus revealing the differences in gene complement and demonstrating the nature of the changes that occur in the evolution of genomes. In particular, genomic analysis showed that more than one-third of the eukaryotic genome is composed of gene duplications and gene families (e.g., [1–4]), highlighting the prominence of duplications in the evolution of genomes. The study of whole genomes also allows an analysis of the rates and dynamics of duplications as well as the divergence and silencing of duplicates (e.g., [5–7]). Moreover, evidence indicates that the rates of duplication are extremely high and of the same order of magnitude as the rate of mutation per nucleotide; for example, the frequency of duplication for coding genes was found to be 0.01 per gene per million years [8] and for internal duplications of gene segments, 0.001–0.013 duplications/gene per million years [9].

The fates of gene duplicates can be very different according to the mutations they undergo and the selective pressures they are under. After a gene is duplicated, the most common fate seems to be the loss of function of one copy by the acquisition of degenerative mutations, while the other copy retains the original function. The originally identical copies can also both be maintained in the genome, allowing a higher production of the corresponding RNA or protein. Subsequent mutation by retrotransposon insertion into one of the copies can affect adaptive evolution because the effect on the phenotype can in some cases be beneficial to the organism [10]. Partial gene duplications can also occur, and if the duplicated part involves a structural or functional domain the new gene can increase the functional complexity of the encoded protein. It has also been proposed that if two copies of a gene acquire mutations in distinct subfunctions, the copies may undergo complementary loss of function in such a way that both copies are required to perform the original function; both copies would be maintained by selection [11]. Probably the most infrequent fate for a gene duplicate, and the one with most evolutionary impact, is for a duplicate copy to gain a new function by acquisition of mutations and to be maintained subsequently by selection while the old copy retains the original function. This last phenomenon leads to novel cellular functions and is the basis of the formation of divergent gene families. For a recent review of models of the evolution of gene duplications and the predictions for each model, see Innan and Kondrashov [12].

Gene duplication has long been thought to be a primary source of material for the evolution of new genes and new functions. The potential of gene duplication was first

recognized by Haldane [13], Bridges [14] and Muller [15]. In 1970, these ideas resurfaced with Ohno [16] who maintained that new genes and novel functions arise only by duplication and who presented evidence for the role of genome duplications in the evolution of multicellular organisms. Since then, the importance of duplications and of the divergence of duplicated copies has been widely supported (for a review see, e.g., [17–20]).

However, other evidence has accumulated showing that there are also other mechanisms responsable for the evolution of novel functions, such as exon shuffling, retroposition, gene fusion, and gene fission, and that several mechanisms can act together ([21–23] and recently reviewed in Kaessmann [24]). Another issue that has received attention is the formation of new regulatory regions, as these can be responsible for the development of novel expression patterns (for review see, e.g., [25–28]). In this paper, we will focus on the origin of very recent genes, which are of special interest since they may have not yet lost the signature of the events that took place during their formation, therefore conveying valuable information about those events.

2. Mechanisms of Origin of Novel Genes

The origin of novel genes has been extensively studied. In Table 1 is summarized the information on a sample of novel genes that reveals a picture of the events involved in their origin and their function. The sample is limited to genes that originated less than 50 million years ago to avoid confounding events of origination with effects of later evolution; this cutoff is also based on the dynamics of formation, preservation, and decay of duplicates and chimeras found by Rogers et al. [29]. Many of the novel genes described and characterized are from *Drosophila* species. The extensive work carried out on the evolution of new genes in *Drosophila* (see e.g., [30]) reflects the fact that *D. melanogaster* is a well-studied model organism at the molecular level that also has a long history of evolutionary studies.

Overall, there is no single mechanism of the molecular events involved in the formation and maintenance of novel genes, but rather several mechanisms. While gene duplication is conceded to have an important role in creating genetic novelties, an alternative mechanism for the origin of new genes is the shuffling of existing genes or functions to form chimeric genes. However, many chimeric genes derive from previously duplicated material, as shown by the presence of the parental genes in the genome and the organization of the chimeras [29, 31]. A chimeric gene would hardly be recognized as such were the parental genes not present. A second type of chimeric gene arises not from tandem duplication, but from duplication through retrotransposition. In addition, it has been shown that both nonhomologous recombination and nonallelic homologous recombination can lead to the formation of chimeric genes which both can be facilitated by repetitive elements (transposable elements or satellite DNA sequences) [32]. Yang et al. [32] identified 17 chimeric genes formed by ectopic recombination within the last 12 Mys in the *Drosophila melanogaster*.

The contribution of chimeric genes to the evolution of genetic novelties has been revealed by recent studies. Rogers and colleagues [29, 31] compare the fate of simple genetic duplicates versus genetic duplicates that underwent fusion to form chimeras in *D. melanogaster*. The results are very interesting: gene duplications are formed at a rate of about 80.4 duplicates per million years, but only 4.1% are preserved; chimeras are formed at a rate of about 11.4 per million years and show a similar rate of decay (with 1.4% preservation) [29]. Some chimeric genes were also implicated in selective sweeps [31], revealing the impact of this kind of molecular event. This analysis of the *D. melanogaster* genome is interesting, but raises the question whether the same pattern would be observed in species that do not have such a high rate of DNA loss as *D. melanogaster* as found by Petrov and colleagues [33–35].

Equally relevant, and conceptually on the opposite side of the spectrum, is the origination of new genes "from scratch" or *de novo*, the case when coding sequences are derived from noncoding DNA, for example, introns and other untranslated regions [36–39]. This reutilization of noncoding sequences may have important consequences in the fate of such genes.

3. Tissues Where Novel Genes Are Expressed

Among the examples of new genes with very recent origin, many are involved in male reproduction, in particular in spermatogenesis. It is the case of *Jingwei*, which appears to be expressed in the testis [43]; *Odysseus*, which contains rapidly evolving homeodomains involved in sperm function [56]; *Dntf-2r*, which has male-specific expression [50]; *K81*, which is expressed in primary spermatocytes [58]; *Sdic*, whose product is incorporated into the sperm tail [47]. The prominence of expression in male reproductive tissue has led to the hypothesis that spermatogenesis is more prone to the cooption of novel genes and functions, while other developmental processes may be under a higher level of selective constraint [63]. As debated in Nielsen et al. [64], the testes undergo intense selection pressures because the cells are subject to genetic conflict, sperm competition, reproductive isolation and are exposed to germline pathogens and mutations that cause segregation distortion. An interesting observation in this regard is that newly retroposed duplicated genes in *Drosophila* are dominated by genes going from the X chromosome to the autosomes, not the opposite direction, and most of these have evolved a testis-specific expression pattern [65]. A possible explanation is that this predominantly one-way pattern is driven by positive selection because the retroposed genes escape X inactivation during spermatogenesis and are therefore free to develop expression patterns in the testis [65].

Kaessmann [24] emphasizes the importance of transcription during spermatogenesis, during which a permissive state of the chromatin may allow a widespread transcription of genes that would not otherwise be expressed.

Additionally, during spermatogenesis, there are numerous cell divisons, affording an opportunity for new

Table 1: Novel genes described (less than 50 million years (mys) old): Their formation, expression and function. The genes are ordered by age.

Gene	Mechanisms	Functions	Expression tissue	Species, genome location	Age (mys)	References
Monkey-king (mkg)	Duplication by retroposition and gene fission	Not determined	*D. simulans/D. sechellia* only males, *D. mauritania* males and females	*D. simulans, D. sechellia, D. mauritania, X* chromosome (chr.)	1-2	[23]
Quetzalcoatl	Duplication, gene fusion	Not determined	Early pupae, adult females, and male testes	*D. melanogaster*	<2	[40]
Ifc-2h	Duplication by retroposition	Not determined	Whole body	*D. simulans, D. sechellia, D. mauritiana*	2	[41]
Hun	Exon shuffling by illegitimate recombination	Not determined	Testis	*D. sechellia, D. mauritania, D. simulans*	2-3	[42]
Jingwei	Exon shuffling, retroposition	Not determined	Testis	*D. teissieri, D. yakuba, D. santomea*	2-3	[43]
Sphinx	Exon shuffling, retroposition	Male courtship behaviour	Embryo, pupae, adult of males and females	*D. melanogaster, chr. 4*	2-3	[44, 45]
Poldi	*De novo* emergence from intergenic DNA	Involved in spermatogenesis, perhaps through chromatin modification pathways	Testis	*Mus mus domesticus* and four other species in the *Mus* genus.	2.5-3.5	[38]
Adh-Twain	Duplication, gene fusion	Not determined	Larvae, adult males and females	*D. subobscura, D. guanche, D. madeirensis*	3	[22, 46]
CG32582, CG32712, CG15323, CG32690, CG31909	From non-coding sequences	Not determined	Testis	*D. melanogaster, D. simulans*	2–5	[37]
Sdic	Duplication, gene fusion	Sperm competition	Testis	*D. melanogaster*, X chr.	<5.4	[47–49]
Dntf-2r	Duplication by retroposition	Not determined	Testis	*D. melanogaster, D. simulans, D. sechellia, D. mauritiana, 2nd chr.*	3–12	[50]
Nsr (*kep1* gene family)	Duplication	Spermatogenesis	Testis (enriched in primary spermatocytes)	*D. melanogaster* subgroup, 2nd chr.	5.4–12	[51]
CG3927 (kep1 gene family)	Duplication	Spermatogenesis	Testis (enriched in primary spermatocytes)	*D. melanogaster* subgroup, 2nd chr.	5.4–12	[51]
CG4021 (kep1 gene family)	Duplication	Spermatogenesis	Testis (enriched in primary spermatocytes)	*D. melanogaster* subgroup, 2nd chr.	5.4–12	[51]
Sflc	Duplication	Male and female fertility, life expectancy	Not specified	*D. melanogaster* subgroup	6–11	[52]
Hydra	*De novo* emergence from intergenic DNA	Not determined	Testis	*D. melanogaster* subgroup	<13	[53]
PIPSL	Exon shuffling, retrotransposition	Ubiquitin binding	Testis	*Humans Chimpanzees*	15–19	[54]
Adh-Finnegan	Duplication, gene fusion	Not determined	Pupae and adults	*D. buzzatii, D. hydei, D. mettleri, D. mojavensis, D. mulleri* (repleta group)	20–30	[45, 46, 55]

Table 1: Continued.

Gene	Mechanisms	Functions	Expression tissue	Species, genome location	Age (mys)	References
Odysseus (OdsH)	Duplication	Involved in sperm function	Mainly in male reproductive tissue, but also detected in larvae and embryo	*D. melanogaster, D. simulans, D. sechellia, D. mauritiana, D. yakuba*	<40 (after split of subgenus *Sophophora* and *Drosophila*	[56, 57]
K81	Duplication by retroposition	Telomere capping in post-meiotic male germ cells	Male specific expression (testis: primary spermatocytes), but weak expression also in wild-type females	*D. melanogaster* subgroup	<30	[58, 59]
Obp57d and Obp57e	Duplication	Taste sensation of octanoic acid	Taste sensilla on the legs	*D. melanogaster* subgroup	<30	[60, 61]
PGAM3	Exon shuffling, retroposition	Possibly phosphoglycerate mutase activity	White blood cells	Humans, chimpanzees, macaques, *X chr.*	>25	[45, 62]

mutations to arise, and those may affect sperm performance. Hence, spermatogenesis offers a large arena of competition in which cells and their descendants are under intensive selective pressure.

An important pattern observed in speciation, often considered the first step in the evolution of complete sterility and inviability, is Haldane's rule, which states that, in an interspecific cross, if one sex is sterile or inviable, this sex is the heterogametic sex [66]. If novel genes appear more often in male reproductive tissue than in female reproductive tissue—as observed in the summary of novel genes in Table 1—this could be an explanation for Haldane's rule when the heterogametic sterile F_1 hybrid is the male, but not in the cases when it is the female. In *Xenopus*, in which males are the homogametic sex, males are always sterile in interspecific crosses [67]. However, while *Xenopus* seems an exception to Haldane's rule, it may be a consequence of the fast evolution of male-specific genes. Strong sexual selection in males is reflected by rapid evolution of genes in male reproductive tissue, which then may cause hybrid sterility in recently diverged sister species even when males are homogametic. From this angle, tissue specificity of novel genes seems more important to hybrid sterility than heterogamy. It would be interesting to examine the tissue specificity of novel genes in species in which the female is the heterogametic sex, but this is beyond the scope of our paper.

4. The *Sdic* Gene Cluster

The *Sdic* gene is a recently evolved chimeric gene in *D. melanogaster*, discovered and described by Nurminsky and colleagues in 1998 [47, 68]. This gene possesses several unique features that provide an exceptional opportunity for the study of new gene functions, the fate of gene duplications, and the evolution of male reproductive traits.

Sequence analysis of the *Sdic* gene revealed that *Sdic* is a chimera of two genes that exist intact in the genome. *Sdic* is composed of parts of *AnnX*, which encodes an annexin protein, and *Cdic* (also referred to in the literature and FlyBase as *sw*), which encodes an intermediate polypeptide chain for the cytoplasmic dyneins [47]. The structure of *Sdic* along with the fact that *AnnX* and *Cdic* exist intact in the genome indicates that *Sdic* originated as a duplication and fusion of *AnnX* and *Cdic*, followed by small deletions and rearrangements. Its formation involved the creation of novel promoter elements (which provided testis-specific expression) from the fusion of portions of an *AnnX* exon and a *Cdic* intron. Its coding region, however, derived solely from *Cdic*. The comparison of the coding region of *Sdic* with *Cdic* shows that *Sdic* lacks the 3′ region of *Cdic* (which corresponds to 100 amino acids residues at the C-terminal part of the *Cdic* protein) and at its 5′ end underwent extensive refashioning by the occurrence of multiple mutations, deletions (including frameshift deletions), and insertions, culminating in a new 5′ exon that encodes a totally novel N-terminus for the protein [47, 69] (Figure 1).

There are several copies of *Sdic* located in tandem at the base of the X chromosome, in region 19 of the larval salivary gland polytene chromosomes, forming a gene cluster. This repeated region is flanked by the parental genes, on the 5′ side by *Cdic* and on the 3′ side by *AnnX*. According to the available genomic sequence of *D. melanogaster, the Sdic* gene is repeated four times in tandem between the genes *Cdic* and *AnnX* genes. Within this cluster there are also four dead-on-arrival retrotransposable elements of the *RT1C* family, one *RT1C* copy located upstream of each *Sdic* gene copy [70] (Figure 2).

Sdic is present in all wild-type strains of *D. melanogaster* [47], but it has not been found in any other species of the *D. melanogaster* subgroup. In the species closest to *D. melanogaster* (*D. simulans*, *D. mauritiana*, *D. yakuba*, *D. teissieri*, and *D. erecta*) *AnnX* and *Cdic* are adjacent to each other, without signs of an ancestral *Sdic* gene or *RT1C* element between them [71]. Furthermore, in these species, *AnnX* and *Cdic* do not show any signs of duplication and are reasonably conserved across species [71]. Consequently, the formation of the *Sdic* gene and the subsequent duplication that formed the entire cluster happened only in the lineage that gave rise to *D. melanogaster*; the original *Sdic* and the *Sdic* cluster were formed within the last 2 million years, after the split of the lineage that formed *D. melanogaster* and its sibling species *D. simulans*.

A model for the evolution of this cluster was developed based on the available genomic data [70], and the divergence of the *RT1C* copies, which are expected to be evolving neutrally, was used to date the duplications [48]. In the ancestral situation *AnnX* and *Cdic* must have been adjacent to each other. An initial event duplicated *AnnX* and *Cdic*, and then one or more deletions fused one copy of *AnnX* and one copy of *Cdic* giving rise to an ancestral *Sdic*. During the early steps of the formation of the ancestral *Sdic*, a dead-on-arrival retrotransposable element from the *RT1C* family was inserted upstream of the ancestral *Sdic*. The ancestral *Sdic* and its upstream region were duplicated in the last 232–463 thousand years, giving rise to two *Sdic* genes (and two *RT1C*, one upstream of each *Sdic*). Another duplication in the last 100–180 thousand years gave rise to four *Sdic* genes (and four *RT1C*, one upstream of each *Sdic* copy).

While the *Sdic* cluster has been isolated in the extremities of BAC clones, it is difficult to determine the overlap of the BAC sequences; moreover, there are nonassembled pieces of BAC clones containing *Sdic* portions that do not seem to match any of the assembled copies. The recent *in situ* hybridization work by Yeh et al. [49] indicates that all *Sdic* copies should be in region 19 of chromosome X. Given the young age of the cluster, the duplicates are predicted to be very recent and to have few or no differences in sequence, making it extremely difficult to confirm the exact number of genes by present sequencing and assembly techniques. The available sequences of assembled and nonassembled BAC clones could be explained either by four copies in tandem or eight copies that resulted from a duplication of preexisting four, yielding eight with the same pattern of similarities. The hypothesis of a cluster formed of four copies is the most parsimonious; however eight copies in tandem would be closer to the original Southern blot experimental results by Nurminsky et al. [47], which suggested as many as 10 duplicates.

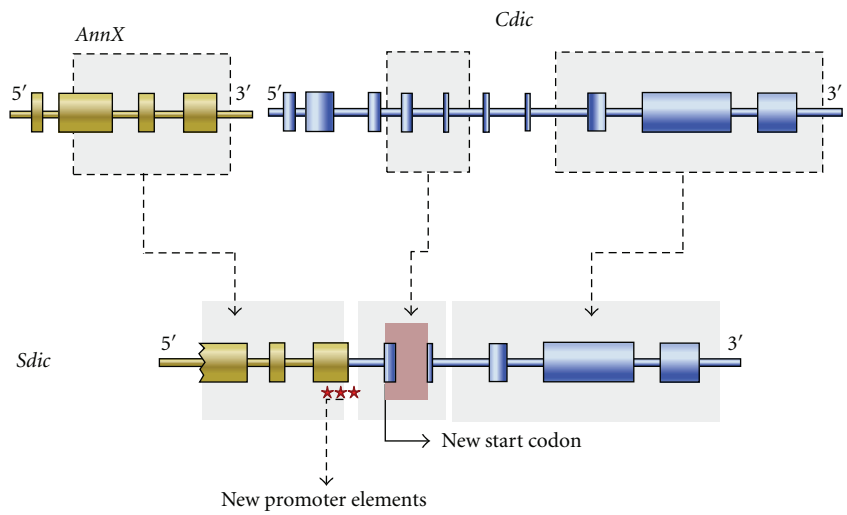


Figure 1: Formation of the *Sdic* gene from parts of the genes *AnnX* and *Cdic*. Introns are represented as thin cylinders and exons as thick cylinders. The stars represent *Sdic* promoter elements.

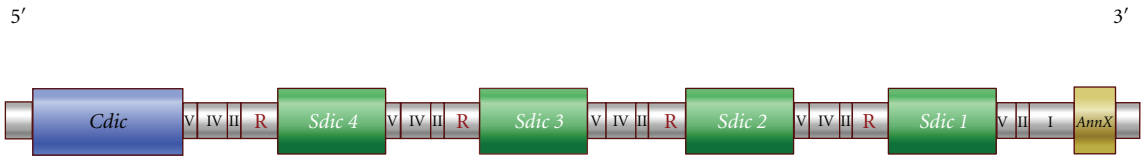


Figure 2: The *Sdic* gene cluster. The cluster is composed by four *Sdic* genes, with an *RT1C* retrotransposable element upstream of each Sdic gene. The cluster is located between the parental genes *Cdic* and *AnnX*. *Cdic* is represented in blue, *Sdic* genes in green and *AnnX* in yellow; intergenic regions are grey; R represents *RT1C* elements. These genes are located in the minus strand, so the order of genes in this figure is the opposite order of these genes in Flybase.

One question still unanswered is that of the expression of the genes that belong to this cluster. Although the cluster is composed of four almost identical *Sdic* genes, there is no evidence for expression of all copies. The RNA sequences described in Nurminsky et al. [47] and Kulathinal et al. [63] are similar to what would be the expected transcription of *Sdic1*. Later work found evidence of expression for both *Sdic1* and *Sdic3* [49, 72]. *Sdic1* is the oldest gene in the array and *Sdic3* was originated from a duplication of *Sdic1*, being the second oldest. The promoter sequences of all copies are identical, except for the promoter of *Sdic1*, which differs from the others by two nucleotides. Although these changes at the promoter level might not be relevant, there is experimental evidence that the promoter of *Sdic1*, at least, is a functional promoter: Nurminsky et al. [47] tested the *Sdic1* promoter *in vivo* and *in vitro* and found it to be functional.

The presence of extra genetic material in the *Sdic* region when one compares *D. melanogaster* with the other species of the *D. melanogaster* subgroup, extra material that contains a cluster of nearby identical genes and dead-on-arrival transposable elements is remarkable. *D. melanogaster* has a high rate of gene loss of unessential genetic material caused by deletions [33–35]. Additionally, according to existing models of the fate of gene duplicates and in light of the high rate of nonessential DNA loss in *D. melanogaster*, if not all copies of *Sdic* are functional then one would predict that the extra copies of *Sdic* would be in the process of degeneration showing many deletions and mutations in the coding sequences as well as the *RT1C* elements. However, the *Sdic* copies and *RT1C* elements appear very similar to each other. It is possible that, given the young age of the cluster, degeneration is in such early stages that such mutations have not occurred or that those that may have occurred have been repaired by gene conversion from nonmutated copies. It is also possible that the *Sdic* cluster has important cellular functions yet to be discovered. Novel gene functions are frequently associated with rapid changes and show signs of positive selection [44, 50, 73], supporting the idea that novel genetic functions allow adaptive changes. There is limited evidence for positive selection of *Sdic* and at least one selective sweep in this region [47, 63, 74].

A key question concerning any novel gene and the functional role it fulfills is whether the function itself is novel or whether the function is redundant or overlapping with an already existing functional gene or genes. The function of *Sdic* protein was deduced from its sequence to be a sperm-specific dynein intermediate chain, and this deduction was supported by the finding of the *Sdic1* gene product in the sperm tail [47]. Is the Sdic protein redundant with another sperm dynein intermediate chain or has it supplanted a pre-existing gene? Has *Sdic* played a role in the speciation events in the split of the lineage of *D. melanogaster* from its sibling

species and has *Sdic* had an effect in adaptation or sexual selection in *D. melanogaster*?

These questions on the role of *Sdic* have been addressed very recently in experiments in which Yeh and colleagues [49] knocked out the *Sdic* region in *D. melanogaster* and looked for phenotypic effects on male reproduction. While no effects were detected in progeny size or sex ratio, males without *Sdic* had sperm that were less competitive in the female reproductive tract in being more easily displaced by sperm from subsequent males and, to a lesser extent, in being less able to displace sperm from previous males. These results support the role of the *Sdic* protein in sperm competition, but most importantly they answer the central question for new genes of "how did the sibling species manage without this function?" the work by Yeh et al. [49] shows that *Sdic* plays an important role in reproduction, and although it is not an essential gene, it gives an advantage in sperm competition, which could in turn be related to the positive selection.

It remains to be determined whether positive selection has acted upon all copies of *Sdic* and if the copies are subjected to the same selective pressures, but the importance of this gene cluster in reproduction has at least been demonstrated, and the results may help to shed light on the general issue of the function of novel genes

5. Concluding Remarks

Ultimately, differences observed between species are due to differences at the genome level. Genomic studies are revealing the extent of these differences—in gene number, in encoded functions, in expression—and are also revealing the mechanisms involved in the evolution of genomes. The analysis of particular newly evolved genes provides information in finer detail, which hopefully can be generalized and help to understand the evolution of new genes and new functions. Equally as important as the formation of new coding sequences is the formation of regulatory regions responsible for new patterns of expression as well as the processes leading to spread and maintenance of the novel gene in the population.

Bacterial genome studies have made very clear that, at least in bacterial species, a great part of the genes are not shared by all individuals of a species [75]. Different strains of the same species share a core genome containing genes present in all strains; however there is also a pan-genome consisting of genes present in only a subset of strains. As more complete genome sequences become available, we will be able to determine if similar patterns are observed in eukaryotes.

Author's Contribution

R. Ponce and L. Martinsen contributed equally to this work.

Acknowledgments

R. Ponce is supported by SFRH/BPD/42801/2008 postdoctoral fellowship from Fundação para a Ciência e Tecnologia, Portugal. L. Martinsen is supported by The Norwegian Research Council project 204693/F20. This work was also supported by NIH Grants GM084236 and GM065169 to D. Hartl.

References

[1] S. Kaul, H. L. Koo, J. Jenkins et al., "Analysis of the genome sequence of the flowering plant Arabidopsis thaliana," *Nature*, vol. 408, no. 6814, pp. 796–815, 2000.

[2] G. M. Rubin, M. D. Yandell, J. R. Wortman et al., "Comparative genomics of the eukaryotes," *Science*, vol. 287, no. 5461, pp. 2204–2215, 2000.

[3] J. C. . Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton et al., "The sequence of the human genome," *Science*, vol. 291, pp. 1304–1351, 2001.

[4] R. D. Morin, E. Chang, A. Petrescu et al., "Sequencing and analysis of 10,967 full-length cDNA clones from Xenopus laevis and Xenopus tropicalis reveals post-tetraploidization transcriptome remodeling," *Genome Research*, vol. 16, no. 7, pp. 947–803, 2006.

[5] Z. Gu, D. Nicolae, H. H. S. Lu, and W. H. Li, "Rapid divergence in expression between duplicate genes inferred from microarray data," *Trends in Genetics*, vol. 18, no. 12, pp. 609–613, 2002.

[6] P. Zhang, Z. Gu, and W. H. Li, "Different evolutionary patterns between young duplicate genes in the human genome," *Genome Biology*, vol. 4, no. 9, article R56, 2003.

[7] L. Bu, U. Bergthorsson, and V. Katju, "Local synteny and codon usage contribute to asymmetric sequence divergence of Saccharomyces cerevisiae gene duplicates," *BMC Evolutionary Biology*, vol. 11, article 279, 2011.

[8] M. Lynch and J. S. Conery, "The evolutionary fate and consequences of duplicate genes," *Science*, vol. 290, no. 5494, pp. 1151–1155, 2000.

[9] X. Gao and M. Lynch, "Ubiquitous internal gene duplication and intron creation in eukaryotes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 49, pp. 20818–20823, 2009.

[10] A. Kanazawa, B. Liu, F. Kong, S. Arase, and J. Abe, "Adaptive evolution involving gene duplication and insertion of a novel Ty1/copia-like retrotransposon in soybean," *Journal of Molecular Evolution*, vol. 69, no. 2, pp. 164–175, 2009.

[11] A. Force, M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, and J. Postlethwait, "Preservation of duplicate genes by complementary, degenerative mutations," *Genetics*, vol. 151, no. 4, pp. 1531–1545, 1999.

[12] H. Innan and F. Kondrashov, "The evolution of gene duplications: classifying and distinguishing between models," *Nature Reviews Genetics*, vol. 11, no. 2, pp. 97–108, 2010.

[13] J. B. S. Haldane, *The Causes of Evolution*, Princeton University Press, 1990.

[14] C. B. Bridges, "The Bar "gene" a duplication," *Science*, vol. 83, no. 2148, pp. 210–211, 1936.

[15] H. J. Muller, "Bar duplication," *Science*, vol. 83, no. 2161, pp. 528–530, 1936.

[16] S. Ohno, *Evolution by Gene Duplication*, Springer, 1970.

[17] V. F. Irish and A. Litt, "Flower development and evolution: gene duplication, diversification and redeployment," *Current Opinion in Genetics and Development*, vol. 15, no. 4, pp. 454–460, 2005.

[18] T. Ohta, "Role of gene duplication in evolution," *Genome*, vol. 31, no. 1, pp. 304–310, 1989.

[19] J. S. Taylor and J. Raes, "Duplication and divergence: the evolution of new genes and old ideas," *Annual Review of Genetics*, vol. 38, pp. 615–643, 2004.

[20] J. Zhang, "Evolution by gene duplication: an update," *Trends in Ecology and Evolution*, vol. 18, no. 6, pp. 292–298, 2003.

[21] M. Long, E. Betrán, K. Thornton, and W. Wang, "The origin of new genes: glimpses from the young and old," *Nature Reviews Genetics*, vol. 4, no. 11, pp. 865–875, 2003.

[22] C. D. Jones, A. W. Custer, and D. J. Begun, "Origin and evolution of a chimeric fusion gene in Drosophila subobscura D. madeirensis and D. guanche," *Genetics*, vol. 170, no. 1, pp. 207–219, 2005.

[23] W. Wang, H. Yu, and M. Long, "Duplication-degeneration as a mechanism of gene fission and the origin of new genes in Drosophila species," *Nature Genetics*, vol. 36, no. 5, pp. 523–527, 2004.

[24] H. Kaessmann, "Origins, evolution, and phenotypic impact of new genes," *Genome Research*, vol. 20, no. 10, pp. 1313–1326, 2010.

[25] G. A. Wray, "The evolutionary significance of cis-regulatory mutations," *Nature Reviews Genetics*, vol. 8, no. 3, pp. 206–216, 2007.

[26] P. J. Wittkopp, "Evolution of cis-regulatory sequence and function in Diptera," *Heredity*, vol. 97, no. 3, pp. 139–147, 2006.

[27] S. B. Carroll, "Evolution at two levels: on genes and form," *PLoS Bology*, vol. 3, no. 7, p. e245, 2005.

[28] F. Rodriguez-Trelles, R. Tarríol, and F. J. Ayala, "Evolution of cis-regulatory regions versus codifying regions," *International Journal of Developmental Biology*, vol. 47, no. 7-8, pp. 665–673, 2003.

[29] R. L. Rogers, T. Bedford, and D. L. Hart, "Formation and longevity of chimeric and duplicate genes in Drosophila melanogaster," *Genetics*, vol. 181, no. 1, pp. 313–322, 2009.

[30] Q. Zhou, G. Zhang, Y. Zhang et al., "On the origin of new genes in Drosophila," *Genome Research*, vol. 18, no. 9, pp. 1446–1455, 2008.

[31] R. L. Rogers and D. L. Hartl, "Chimeric genes as a source of rapid evolution in Drosophila melanogaster," *Molecular Biology and Evolution*, vol. 29, no. 2, pp. 517–529, 2012.

[32] S. Yang, J. R. Arguello, X. Li et al., "Repetitive element-mediated recombination as a mechanism for new gene origination in Drosophila," *PLoS Genetics*, vol. 4, no. 1, article e3, 2008.

[33] D. A. Petrov and D. L. Hartl, "High rate of DNA loss in the Drosophila melanogaster and Drosophila virilis species groups," *Molecular Biology and Evolution*, vol. 15, no. 3, pp. 293–302, 1998.

[34] D. A. Petrov, E. R. Lozovskaya, and D. L. Harti, "High intrinsic rate of DNA loss in Drosophila," *Nature*, vol. 384, no. 6607, pp. 346–349, 1996.

[35] D. A. Petrov and D. L. Hartl, "Trash DNA is what gets thrown away: high rate of DNA loss in Drosophila," *Gene*, vol. 205, no. 1-2, pp. 279–289, 1997.

[36] M. E. Johnson, L. Viggiano, J. A. Bailey et al., "Positive selection of a gene family during the emergence of humans and African apes," *Nature*, vol. 413, no. 6855, pp. 514–519, 2001.

[37] M. T. Levine, C. D. Jones, A. D. Kern, H. A. Lindfors, and D. J. Begun, "Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 26, pp. 9935–9939, 2006.

[38] T. J. A. J. Heinen, F. Staubach, D. Häming, and D. Tautz, "Emergence of a New Gene from an Intergenic Region," *Current Biology*, vol. 19, no. 18, pp. 1527–1531, 2009.

[39] M. Toll-Riera, N. Bosch, N. Bellora et al., "Origin of primate orphan genes: a comparative genomics approach," *Molecular Biology and Evolution*, vol. 26, no. 3, pp. 603–612, 2009.

[40] R. L. Rogers, T. Bedford, A. M. Lyons, and D. L. Hartl, "Adaptive impact of the chimeric gene Quetzalcoatl in Drosophila melanogaster," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 24, pp. 10943–10948, 2010.

[41] C. Fan and M. Long, "A new retroposed gene in Drosophila heterochromatin detected by microarray-based comparative genomic hybridization," *Journal of Molecular Evolution*, vol. 64, no. 2, pp. 272–283, 2007.

[42] J. R. Arguello, Y. Chen, S. Yang, W. Wang, and M. Long, "Origination of an X-linked testes chimeric gene by illegitimate recombination in Drosophila," *PLoS Genetics*, vol. 2, no. 5, article e77, 2006.

[43] W. Wang, J. Zhang, C. Alvarez, A. Llopart, and M. Long, "The origin of the jingwei gene and the complex modular structure of its parental gene, yellow emperor, in Drosophila melanogaster," *Molecular Biology and Evolution*, vol. 17, no. 9, pp. 1294–1301, 2000.

[44] W. Wang, F. G. Brunet, E. Nevo, and M. Long, "Origin of sphinx, a young chimeric RNA gene in Drosophila melanogaster," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 7, pp. 4448–4453, 2002.

[45] M. Long, M. Deutsch, W. Wang, E. Betrán, F. G. Brunet, and J. Zhang, "Origin of new genes: evidence from experimental and computational analyses," *Genetica*, vol. 118, no. 2-3, pp. 171–182, 2003.

[46] C. D. Jones and D. J. Begun, "Parallel evolution of chimeric fusion genes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 32, pp. 11373–11378, 2005.

[47] D. I. Nurminsky, M. V. Nurminskaya, D. De Aguiar, and D. L. Hartl, "Selective sweep of a newly evolved sperm-specific gene in Drosophila," *Nature*, vol. 396, no. 6711, pp. 572–575, 1998.

[48] R. Ponce, "The use of a non-LTR element to date the formation of the Sdic gene cluster," *Genetica*, vol. 131, no. 3, pp. 315–324, 2007.

[49] S.-D. Yeh, T. Do, C. Chan et al., "Functional evidence that a recently evolved Drosophila sperm-specific gene boosts sperm competition," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 6, pp. 2043–2048, 2012.

[50] E. Betrán and M. Long, "Dntf-2r, a young Drosophila retroposed gene with specific male expression under positive Darwinian selection," *Genetics*, vol. 164, no. 3, pp. 977–988, 2003.

[51] Y. Ding, L. Zhao, S. Yang et al., "A young Drosophila duplicate gene plays essential roles in spermatogenesis by regulating several Y-linked male fertility genes," *PLoS Genetics*, vol. 6, no. 12, article e1001255, 2010.

[52] S. Chen, H. Yang, B. H. Krinsky, A. Zhang, and M. Long, "Roles of young serine-endopeptidase genes in survival and reproduction revealed rapid evolution of phenotypic effects at adult stages," *Fly*, vol. 5, pp. 345–351, 2011.

[53] S. T. Chen, H. C. Cheng, D. A. Barbash, and H. P. Yang, "Evolution of hydra, a recently evolved testis-expressed gene with nine alternative first exons in Drosophila melanogaster," *PLoS Genetics*, vol. 3, no. 7, article e107, 2007.

[54] D. V. Babushok, K. Ohshima, E. M. Ostertag et al., "A novel testis ubiquitin-binding protein gene arose by exon shuffling in hominoids," *Genome Research*, vol. 17, no. 8, pp. 1129–1138, 2007.

[55] D. T. Sullivan, W. T. Starmer, S. W. Curtiss, M. Menotti-Raymond, and J. Yum, "Unusual molecular evolution of an Adh pseudogene in Drosophila," *Molecular Biology and Evolution*, vol. 11, no. 3, pp. 443–458, 1994.

[56] C. T. Ting, S. C. Tsaur, M. L. Wu, and C. I. Wu, "A rapidly evolving homeobox at the site of a hybrid sterility gene," *Science*, vol. 282, no. 5393, pp. 1501–1504, 1998.

[57] C. T. Ting, S. C. Tsaur, S. Sun et al., "Gene duplication and speciation in Drosophila: evidence from the Odysseus locus," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 33, pp. 12232–12235, 2004.

[58] B. Loppin, D. Lepetit, S. Dorus, P. Couble, and T. L. Karr, "Origin and neofunctionalization of a Drosophila paternal effect gene essential for zygote viability," *Current Biology*, vol. 15, no. 2, pp. 87–93, 2005.

[59] R. Dubruille, G. A. Orsi, L. Delabaere et al., "Specialization of a drosophila capping protein essential for the protection of sperm telomeres," *Current Biology*, vol. 20, no. 23, pp. 2090–2099, 2010.

[60] T. Matsuo, "Rapid evolution of two odorant-binding protein genes, Obp57d and Obp57e, in the Drosophila melanogaster species group," *Genetics*, vol. 178, no. 2, pp. 1061–1072, 2008.

[61] E. Harada, J. Nakagawa, T. Asano et al., "Functional evolution of duplicated odorant-binding protein genes, Obp57d and Obp57e, in Drosophila," *PLoS ONE*, vol. 7, no. 1, article e29710, 2012.

[62] E. Betrán, W. Wang, L. Jin, and M. Long, "Evolution of the Phosphoglycerate mutase processed gene in human and chimpanzee revealing the origin of a new primate gene," *Molecular Biology and Evolution*, vol. 19, no. 5, pp. 654–663, 2002.

[63] R. J. Kulathinal, S. A. Sawyer, C. D. Bustamante, D. Nurminsky, R. Ponce, and J. M. Ranz, "Selective sweep in the evolution of a new sperm-specific gene in Drosophila," in *Selective Sweep*, D. Nurminsky, Ed., pp. 22–33, Springer, Boston, Mass, USA, 2005.

[64] R. Nielsen, C. Bustamante, A. G. Clark et al., "A scan for positively selected genes in the genomes of humans and chimpanzees," *PLoS Bology*, vol. 3, no. 6, article e170, 2005.

[65] E. Betrán, K. Thornton, and M. Long, "Retroposed new genes out of the X in Drosophila," *Genome Research*, vol. 12, no. 12, pp. 1854–1859, 2002.

[66] J. B. S. Haldane, *The Causes of Evolution*, Princeton University Press, 1990.

[67] M. J. Madison-Villar and P. Michalak, "Misexpression of testicular microRNA in sterile Xenopus hybrids points to tetrapod-specific microRNAs associated with male fertility," *Journal of Molecular Evolution*, vol. 73, pp. 316–324, 2011.

[68] D. I. Nurminsky, M. V. Nurminskaya, E. V. Benevolenskaya, Y. Y. Shevelyov, D. L. Hartl, and V. A. Gvozdev, "Cytoplasmic dynein intermediate-chain isoforms with different targeting properties created by tissue-specific alternative splicing," *Molecular and Cellular Biology*, vol. 18, no. 11, pp. 6816–6825, 1998.

[69] J. M. Ranz, A. R. Ponce, D. L. Hartl, and D. Nurminsky, "Origin and evolution of a new gene expressed in the Drosophila sperm axoneme," *Genetica*, vol. 118, no. 2-3, pp. 233–244, 2003.

[70] R. Ponce and D. L. Hartl, "The evolution of the novel Sdic gene cluster in Drosophila melanogaster," *Gene*, vol. 376, no. 2, pp. 174–183, 2006.

[71] R. Ponce, "The recent origin of the Sdic gene cluster in the melanogaster subgroup," *Genetica*, vol. 135, no. 3, pp. 415–418, 2009.

[72] B. R. Graveley, A. N. Brooks, J. W. Carlson et al., "The developmental transcriptome of Drosophila melanogaster," *Nature*, vol. 471, no. 7339, pp. 473–479, 2011.

[73] M. Long and C. H. Langley, "Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila," *Science*, vol. 259, no. 5104, pp. 91–95, 1993.

[74] D. Nurminsky, D. De Aguiar, C. D. Bustamante, and D. L. Hartl, "Chromosomal effects of rapid gene evolution in drosophila melanogaster," *Science*, vol. 291, no. 5501, pp. 128–130, 2001.

[75] M. Touchon, C. Hoede, O. Tenaillon et al., "Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths," *PLoS Genetics*, vol. 5, no. 1, article e1000344, 2009.

Hindawi Publishing Corporation
International Journal of Evolutionary Biology
Volume 2012, Article ID 596274, 10 pages
doi:10.1155/2012/596274

Review Article

Alternative Splicing: A Potential Source of Functional Innovation in the Eukaryotic Genome

Lu Chen, Jaime M. Tovar-Corona, and Araxi O. Urrutia

Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, UK

Correspondence should be addressed to Araxi O. Urrutia, a.urrutia@bath.ac.uk

Received 11 February 2012; Revised 19 April 2012; Accepted 7 May 2012

Academic Editor: Ben-Yang Liao

Alternative splicing (AS) is a common posttranscriptional process in eukaryotic organisms, by which multiple distinct functional transcripts are produced from a single gene. The release of the human genome draft revealed a much smaller number of genes than anticipated. Because of its potential role in expanding protein diversity, interest in alternative splicing has been increasing over the last decade. Although recent studies have shown that 94% human multiexon genes undergo AS, evolution of AS and thus its potential role in functional innovation in eukaryotic genomes remain largely unexplored. Here we review available evidence regarding the evolution of AS prevalence and functional role. In addition we stress the need to correct for the strong effect of transcript coverage in AS detection and set out a strategy to ultimately elucidate the extent of the role of AS in functional innovation on a genomic scale.

1. Introduction

The first draft of the human genome sequence [1, 2] was unveiled in February 2001 and surprisingly it was shown to contain ~23000 genes, only a fraction of the numbers of genes originally predicted [3]. To put this into perspective, there are ~20,000 genes in the genome of the nematode *C. elegans*. The lack of an association between gene number and organismal complexity has resulted in an increased interest in alternative splicing (AS) given it has been proposed to be a major factor in expanding the regulatory and functional complexity, protein diversity, and organismal complexity of higher eukaryotes [4–6]. However, despite the best efforts of many research groups we still understand very little about the actual role played by AS in the evolution of functional innovation—here understood as the appearance of novel functional transcripts—underpinning the increased organismal complexity observed.

Alternative splicing is a posttranscriptional process in eukaryotic organisms by which multiple distinct transcripts are produced from a single gene [4]. Previous studies using high-throughput sequencing technology have reported that up to 92%~94% of human multiexon genes undergo AS [7, 8], often in a tissue/developmental stage-specific manner [7, 9]. With the development and constant improvement of whole genome transcription profiling and bioinformatics algorithms, the ubiquity of AS in the mammalian genome began to become clear. The concept of one gene-one protein gave way as evidence mounted for the high percentage of AS incidence in nonhuman species [7, 8], such as fruit fly [10], *Arabidopsis* [11] and other eukaryotes [5]. Despite the advances in our understanding and characterisation of AS several questions remain unanswered. First, the large difference in transcript coverage between species has hampered direct comparisons of the prevalence of alternative splicing in different species [6]. Secondly, even if comparable AS estimates between species could be obtained, it is unclear to what extent any changes in AS prevalence along evolution have contributed to overall protein diversity or rather reflect splicing noise. Finally, we understand very little about how AS has evolved through time and how this is related to functional parameters of genes. Here we review how alternative is regulated and recent progress in our understanding of the evolution of alternative splicing.

2. Alternative Splicing and Its Regulation

In 1977, Chow et al. [12–15] reported that 5′ and 3′ terminal sequences of several adenovirus 2 (Ad2) mRNAs varied, implying a new mechanism for the generation of several distinct mRNAs. Following this study, alternative splicing was also found in the gene encoding thyroid hormone calcitonin in mammalian cells. Subsequent studies revealed that many other genes were also able to generate more than one transcript by cuttingout different sections from its coding regions (reviewed in [4, 16]).

Depending on the location of the exonic segments cut-out-or if introns are left in, splicing events can be classified into four basic types (Figure 1). These four major modes of splicing are (1) exon skipping (2) intron retention (3) alternative 5′ splicing site (5′ss), and (4) alternative 3′ splicing site (3′ss) [22, 23]. In addition, mutually exclusive exons, alternative initiation, and alternative polyadenylation provide two other mechanisms for generating various transcript isoforms. Moreover, different types of alternative splicing can occur in a combinatorial manner and one exon may be subject to more than one AS mode, for example, 5′ss and 3′ss at the same time (Figure 1). Prevalence of each type of AS has been found to vary between different taxa. Several studies have shown that exon skipping is common in metazoan genomes [24] whereas intron retention is the most common type of AS among plants [25] and fungi [26].

Alternative splicing is tightly regulated by cis elements as well as transacting factors that bind to these cis elements. Transacting factors, mainly RNA-binding proteins, modulate the activity of the spliceosome and cis elements such as exonic splicing enhancers (ESEs), exonic splicing silencers (ESSs), intronic splicing enhancers (ISEs), and intronic splicing silencers (ISSs). Canonical mechanism of AS suggests that serine/arginine-rich (SR) proteins typically bind to ESEs, whereas heterogeneous nuclear ribonucleoproteins (hnRNP) tend to bind to ESSs or ISSs [27]. Given the crucial roles of these regulators in the splicing machinery, the cis and transacting mutations, which disrupt the splicing code, are known to cause disease (reviewed in [28–30]) . It has been estimated that 15–60% of mutations cause disease by affecting the splicing pattern of genes ([31] and reviewed in [30]). Moreover, AS has also been shown to be regulated without the involvement of auxiliary splicing factors [32] and AS may be also combined with other posttranscriptional events such as the use of multiple internal translation initiation sites, RNA editing, mRNA decay, and microRNA binding and other noncoding RNAs [33, 34], suggesting the existence of additional noncanonical mechanism of AS that are yet to be identified [35].

Recently, a direct role of histone modifications in alternative splicing has been reported, in which histone modification (H3-K27m3) affects the splicing outcome by influencing the recruitment of splicing regulators via a chromatin-binding protein in a number of human genes such as *FGFR2,TPM2,TPM1* and *PKM2* [36]. Moreover, it has been reported that CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing, providing the first evidence of developmental regulation of splicing outcome through heritable epigenetic marks [37]. Additionally, non-coding RNAs also have emerged as key determinants of alternative splicing patterns [34]. Therefore these findings reveal an additional epigenetic layer in the regulation of transcription and alternative splicing [38]. Genomewide genetic and epigenetic studies, therefore, have been proposed in at least 100 specific blood cell types [39], which will provide high quality reference epigenomes (using DNA methylation and histone marks assays) with detailed genetic and transcriptome data (whole genome sequencing, RNA-Seq, and miRNA-Seq), providing us with an opportunity to assess the genomewide influence of epigenetic factors in the regulation of AS in specific blood cell types. We are expecting the rise of comparative epigenetics will provide different perspective of the evolution of transcriptome.

3. Identification of Alternative Splicing Events

Alternative splicing is difficult to estimate from genomic parameters alone [40]. A number of regulatory motifs for AS have been uncovered but the presence of known alternative splicing motifs does not guarantee that a gene is actually alternatively spliced [40]. Thus, alternative splicing patterns are generally assessed from examining transcript data. For any gene of interest, alternative splicing events can be identified by using reverse transcription polymerase chain reaction (RT-PCR) conducted on a complementary DNA (cDNA) library. Over the last decade, as high-throughput transcriptome technologies have improved, it has become possible to assess alternative splicing patterns on a genomewide scale. Three main sources of transcriptome data have been used to assess splicing patterns: expressed sequence tags (ESTs), splice-junction microarrays, and RNA sequencing (RNA-Seq).

The first wave of genomewide transcriptome analysis consisted in direct sequencing cDNA and ESTs carried out at large scale [41], which allowed alternative splicing events to be identified by aligning cDNA/EST sequences to the reference genome. ESTs are 200–800 nucleotide bases in length, unedited, randomly selected single-pass sequence reads derived from cDNA libraries [42]. Currently, there are eight million ESTs for human, including about one million sequences from cancer tissues, and about 71 million ESTs for around 2000 species in dbEST [43]. However, ESTs are based on low-throughput Sanger sequencing and are aggregated over a wide range of tissues, developmental states, and diseases using widely different levels of sensitivity.

More recently, splice-junction microarrays and RNA-Seq have been increasingly used to quantitatively analyse alternative splicing events. Splicing microarrays target specific exons or exon-exon junctions with oligonucleotide probes. The fluorescent intensities of individual probes reflect the relative usage of alternatively splicing exons in different tissues and cell lines [44]. High-density splice-junction microarrays are a cost-effective way to assay previously known exons and AS events with low false positive rate. The disadvantage is that it requires prior knowledge of existing AS variants

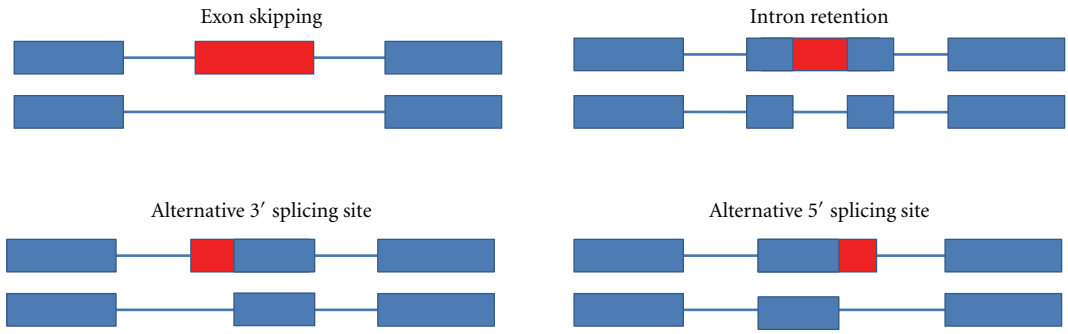


Figure 1: Different types of alternative splicing. The blue boxes are constitutive exons and alternatively spliced regions in red. Introns are represented by straight lines between boxes. Four types of common splicing events were identified: (1) exon skipping (2) intron retention (3) alternative 5′ splicing site (5′ss), and (4) alternative 3′ splicing site (3′ss).

and gene structures. More importantly unlike RNA-Seq and EST, microarrays do not provide additional sequence information.

RNA-Seq has emerged as a powerful technology for transcriptome analysis due to its ability to produce millions of short sequence reads [45–47]. RNA-Seq experiments provide in-depth information on the transcriptional landscape [45]. The ever-increasing accumulation of high-throughput data will continue to provide ever richer opportunities to investigate further aspects of AS such as low-frequency AS events as well as tissue-specific and/or development-specific AS events [7, 8, 47–49]. Earlier datasets consist of RNA read sequences of 50 bp or less, limiting the information about combinations of AS events in a single transcript but it is likely that the length of short reads will continue to increase over the next decade. With the increasing capacity of next-generation sequencing (RNA-Seq) the study of alternative spicing is likely to undergo a revolution [50]. The higher depth of sequencing of transcriptomes in human and other species has increased our understanding of the occurrence of AS event and AS expression patterns in different tissues [7, 51], developmental stages [10].

Transcript assembly of sequence-based technologies, such as ESTs and RNA-Seq, can use either align-then-assemble or assemble-then-align, depending on the quality of reference genome and sequence data [47]. An algorithm can be employed to detect AS event by comparing different transcripts. However, detecting AS isoforms, as opposed to single AS event, is still challenging because short sequences provide little information in terms of the combination of exons. Several applications have been developed for transcript assembly and AS isoform detection, different strategies and comparison of these applications have been reviewed previously [47].

4. Prevalence of Alternative Splicing across Eukaryotic Genomes

Initial whole genome analyses suggested that 5%–30% of human genes were alternatively spliced (reviewed in [6, 16]). EST-based AS databases identify AS events in 40–60% of human genes [5, 52, 53]; however, recently this number has been revised over and over with the latest estimates showing that up to 94% of human multiexon genes produce more than one transcript through alternative splicing [7, 8, 16]. Understanding how alternative splicing has changed over time could provide insights as to how alternative splicing has impacted on transcript and protein diversity and phenotype evolution [6]. In fungi, AS is thought to be rare due to the low number of exons in yeast [23]. In plants it has been estimated that around 20% of genes undergo AS based on EST data [25], a recent study using RNA-Seq, however, suggests that at least approximately 42% of intron-containing genes in *Arabidopsis* are alternatively spliced [11]. We are expecting significantly higher percentages of AS occurrence will be discovered from various eukaryotes given the in-depth studies of transcriptome using next-generation sequencing such as RNA-Seq are ongoing. A few studies have attempted to compare AS prevalence among different taxa with animals generally reported to have higher AS incidence than plants [16] and vertebrates having a higher AS incidence than invertebrates [24]. However, these studies are either based on limited data or failed to correct for differences in transcript coverage [6].

There are a number of databases that provide AS data for multiple species [5, 52–54]. However, these existing resources are primarily focused on animal species and have poor coverage for protist, fungal, and plant genomes thus making it difficult to compare divergent taxa. Most importantly, none of these resources take into account the well-documented effects of differential transcript coverage across genes within and between species which greatly influences AS detection rates [6, 24, 55, 56]. Random sampling has been used [24] and shown to minimize the bias of transcript coverage (Figure 2). We expect that similar strategies will be employed in future comparative AS data resources.

5. Is Alternative Splicing Functional or Mostly Just Noise?

If an increase in AS levels in vertebrate species compared to invertebrates is confirmed, given the limitations of current proteomics resources, it is hard to assess the extent to which alternatively spliced transcripts are translated into

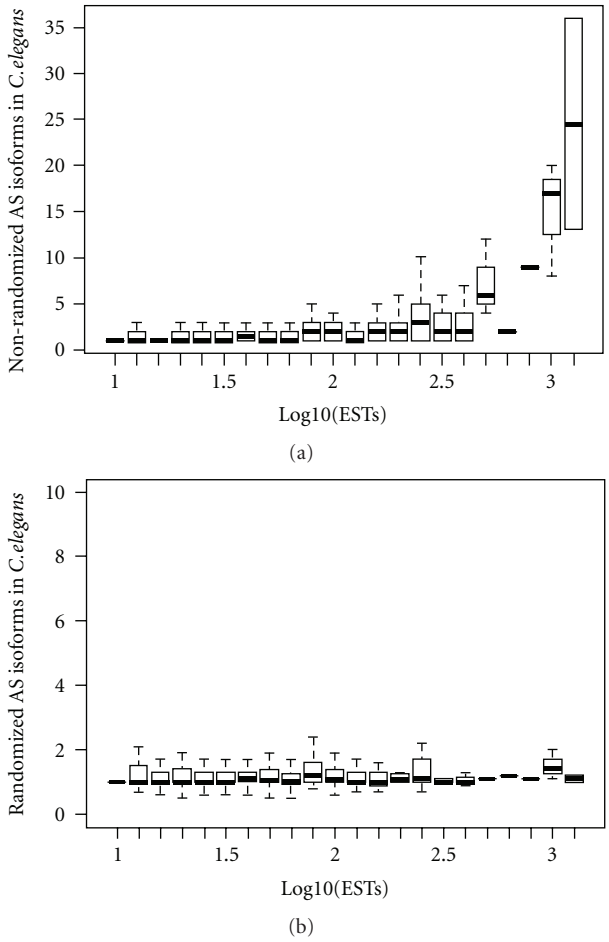


Figure 2: Total transcript number influences AS detection but bias can be corrected by using a sampling method. AS detection in genes divided by transcript coverage for the nematode (a and b) using the full transcript dataset (a) or a random sampling method (b).

an expanded proteome. The evolution of many phenotypes that we most associate with human being such as longer lifespan, encephalization, or even increased complexity have been accompanied by sharp reductions in effective population size, possibly explaining the proliferation of a variety of genomic features in more complex organisms ([57] but see [58]). Therefore, it is possible that increased AS through evolution results from aberrant splicing and therefore it does not play any functional role [59–61]. If alternative splicing has increased along the phylogenetic tree and it is indeed functional, we can expect the following.

(A) Transcripts should have a low incidence of premature stop codons which would render them vulnerable to nonsense mediated decay. Between 4% and 35% of AS human transcripts have been found to contain a premature termination codon in human and mouse transcripts [62, 63]. These transcripts have been found to be enriched in nonconserved exons likely to cause frame shifts [64]. It is unknown whether the proportion of premature stop codon containing AS transcripts has changed along the phylogenetic tree.

(B) It has been proposed that most low copy number alternative isoforms produced in human cells are likely to be nonfunctional [65, 66]. A recent study has shown that although cancer-specific alternative-splicing variants can be found, these events are mostly found as single-copy events and thus unlikely to contribute to the core cancer transcriptome [67].

(C) Conservation of alternative-splicing events along evolution can be taken as an indicator of their functional role. Conservation levels of AS have been studied in many species. The estimation ranges from 11% to 67% between human and mouse [68–70]. Notably, major AS forms tend to be have higher conservation levels compared to minor forms. On the other hand, the conserved AS forms vary among different AS; for example, exon skipping between *C. elegans* and *C. briggsae* has shown more than 81% conservation level, compared to 28% for intron retention [71, 72].

(D) Presence of identifiable functional domains in AS areas may also be an indicator of functional relevance for AS transcripts [67]. To our best knowledge there are no reports of the prevalence of functional domains in AS areas in model species. To examine the presence of functional domains in AS transcripts, we compiled a set of 267,996 AS events obtained from the analysis of 8,315,254 ESTs from normal human tissues. We found that about 50% of AS areas in human contain known functional components using InterProScan [17] which contains 14 applications for the prediction of protein domains (Figure 3, see methods in [67]), suggesting a possible functional role for AS. The extent of the variations in the prevalence of functional domains among AS areas between species remains to be explored but would provide additional insights on the evolution of AS.

Taken together above observations suggest that although alternative splicing-events are indeed conserved throughout evolution a significant proportion are not and some may result from noisy transcript splicing not contributing to the protein pool. However, until further studies using comparable AS indexes it will be impossible to estimate the extent to which increases in AS levels along the phylogenetic tree have impacted on the pool of functional transcripts.

6. Alternative Splicing and Gene Duplication

Gene duplication (GD) is considered a prime source of functional innovation in the genome. Newly duplicated genes can evolve functional divergence [73], and it is thought to be key in driving the evolution of developmental and morphological complexity in vertebrates [74]. Alternative splicing, as a prevalent mechanism that also increases protein diversity, has been proposed as a potential player in the evolution of eukaryotes [4, 6]. By examining the relationship

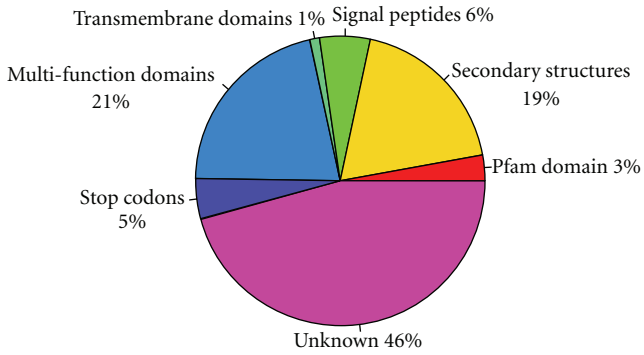


Figure 3: Percentage of AS areas containing identifiable functional domains, secondary structures, and stop codons in human. Functional components were identified using InterProScan which contains 14 applications for the prediction of protein domains [17], including Pfam for the prediction of protein domains [18], SignalP 3.0 for signal peptide predictions [19], and TMHMM [20] for the predictions of transmembrane domains. PSORT II [21] was used to identify the likely subcellular localization of protein products. Secondary protein structures were predicted by CLC Main Workbench 5.7, which is based on extracted protein sequences from the protein databank (http://www.rcsb.org/pdb/).

between gene duplication and alternative splicing we can better understand the extent to which both mechanisms are equivalent means for protein diversification. Several studies have reported a negative correlation between AS and gene family size in human and mouse [6, 65, 75, 76] and worm [71, 77] (Table 1). It is easy to lead to a conclusion that AS and GD are interchangeable and there is a universal negative correlation from worm to human. However, the relationship between the two variables is marginal at best and it is not consistent when including singleton genes which have a lower AS level compared to multigene families [76, 78, 79]. Jin et al. [76] suggested that singletons have more evolutionary constriction than duplicates which hampers their AS isoform gain Consistent with this hypothesis, Lin et al. [78] found that singletons differ from multigene families in several aspects suggesting that they have differing evolutionary paths. Even if we focus on multigene families only, a negative correlation between AS and gene family size may be explained or byproduct of AS and gene family size covariance with other factors. For example, gene age and biased duplication have been proposed to be the explanation [79]. This study has cast doubt over the relationship between AS and GD and it may indeed provide support to the suggestion that AS and GD have little or no equivalence concerning effects on protein sequence, structure, and function [80]. As most studies have examined a small number of model species it is difficult to assess the extent of the link between AS and GD. In addition, the snapshot approach of comparing GFS and AS in a single genome might hide the true relationship between AS and GFS.

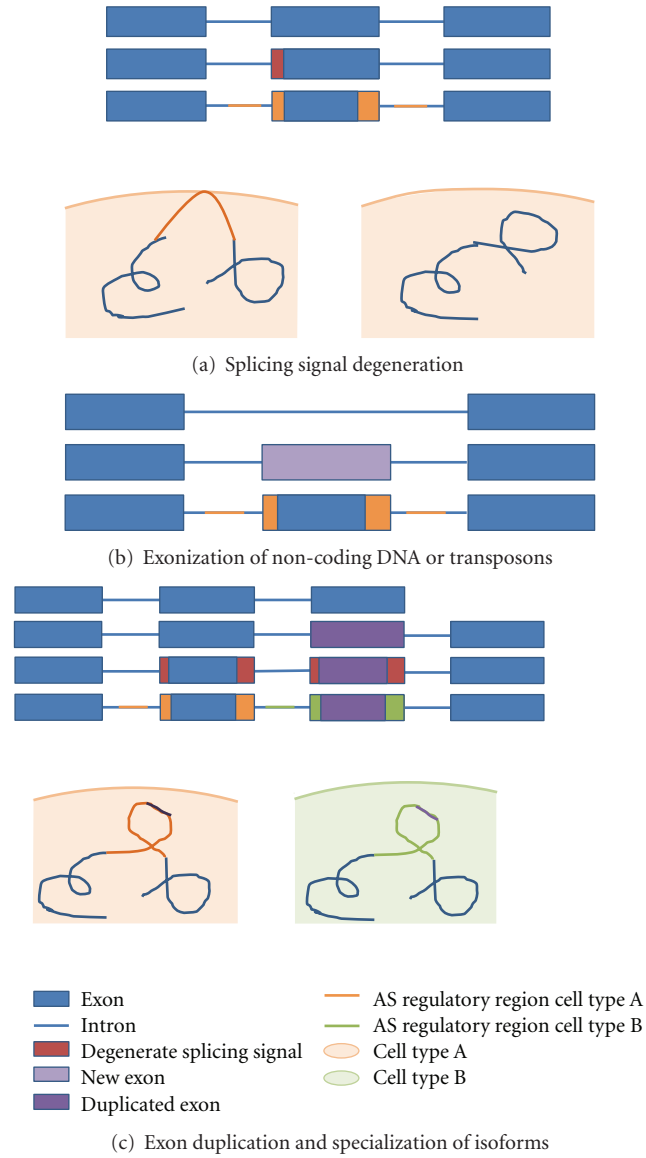


Figure 4: Novel AS variants can take on specialised or novel roles. Novel splicing variants can arise from (a) mutations in the exon recognition site of a constitutive exon and subsequent acquisition of AS regulatory elements. (b) Exonization of introns or intron regions or transposable elements with subsequent acquisition of AS regulatory regions. Novel proteins may interact with different proteins or localise in different subcellular regions. (c) Exon duplication and subsequent specialization functional domains and AS regulatory regions. Resulting specialised proteins may take on partial roles relevant in different cell types or developmental stages or result in novel interactions and functions.

7. Alternative Splicing's Contribution to Functional Innovation

Alternative splicing has been hailed as the missing source of information in the genome accounting for the evolution of higher complexity despite the near static gene number in metazoans over the last 800 million years. Wegmann et al. [81] found that width of gene expression is positively

Table 1: Summary for the relationship between AS and GFS.

Species	Data	Alternative splicing	Orthology	Bias control	Correlation	Reference
Human	Ensembl	ASD's AltSplice database	BLSATP	Exons, EST coverage, gene family size, isoform count	Negative correlation	[75]
	NCBI, UCSC	GeneSplicer program	EnsMart	Remove garbage EST, EST coverage,	Negative correlation	[65]
	H-InvDB 5.0	H-InvDB 5.0	BLAST		Positive correlation when includes all gene families. Negative correlation within multigene families	[76]
Mouse	Ensembl	ASD's AltSplice database	BLSATP	Exons, EST coverage, gene family size, isoform count	Negative correlation	[75]
	NCBI, UCSC	GeneSplicer program	EnsMart	Remove garbage EST, EST coverage,	Negative correlation	[65]
	Riken's FANTOM3	Riken's FANTOM3	BLAST		Positive correlation when includes all gene families. Negative correlation within multigene families	[76]
C. elegans	WormPep	WormPep	BLAST		Lower AS occurrence in multigene families	[77]
Rice	TIGR 4.0	PASA program	BLASTP	Remove genes that lack transcript evidence	Multigene families have significantly higher AS incidence than singletons	[78]
Arabidopsis	TAIR7	TAIR7	TAIR7		Multigene families have significantly higher AS incidence than singletons	[78]

correlated to the number of new transcript isoforms and proposed that the increase of gene expression breadth is essential for acquiring new transcript isoforms, which could be maintained by a new form of balancing selection. Moreover, experimental and bioinformatics analyses have shown that AS can generate a variety of functional mRNAs and protein products, displaying distinct stability properties, subcellular localization, and function [9] as well as in specific stages in cell differentiation [82], sex differentiation [83, 84], and development [9].

Single-gene studies have provided examples where alternative splicing can lead to functional innovation before any events of gene duplication have taken place. One such example is that of Troponin I (TnI), which plays a key role in muscle contraction. In the vertebrate genome, TnI exists in three copies each expressed in a different muscle type (skeletal, fast and slow, and cardiac). In *Ciona*, one of the closest relatives of vertebrates TnI is present as a single gene. Interestingly, however, the *Ciona* gene produces three distinct alternatively spliced isoforms, each found to resemble the expression profile of one of the vertebrate genes suggesting that the specialisation of the TnI proteins to function in each muscle type preceded gene duplication events [85]. This pattern of alternative splice variants in ancestrally single genes resembling expression profiles of genes later duplicated has also been found in synapsin-2 genes in tetrapods [86] and *MITF* genes in teleost fish species [87, 88]. These examples suggest that alternative splicing can be a mechanism for functional innovation preceding events of gene duplication through one of the three possible paths (Figure 4).

Genes may also further gain alternative splicing and regulation after duplication along with the complexity of the organ systems after the divergence of protochordates and vertebrates. Comparison between transcriptional factors *Pax* genes in vertebrates and amphioxus has shown that at least 52 reported alternative-splicing events in vertebrates compared to 23 events in amphioxus [89]. Furthermore, vertebrate *Pax* genes have maintained most of their ancestral functions and also expanded their expression [90]. Novel alternative splicing of *Pax* genes has been shown to modify the functional domain content (e.g., DNA binding) and transactivation capacities of the resulting protein products [89]. For example, a novel alternative transcript of *Pax3* can transactivate a cMET reporter construct in mouse [91]. These additional isoforms of *Pax3* have been proposed to play a functional role in the acquisition of new roles at neural plate in vertebrates [89]. Similarly, vertebrate-specific AS events of exon 5a in *Pax4* and *Pax6* have been linked to functional roles in the development of vertebrate eye [89, 92]. Therefore, it is reasonable to propose the hypothesis that, besides gene duplication, alternative splicing plays important roles in acquiring novel functions contributing to the complexity of the organ systems after the divergence of protochordates and vertebrates [93]. The potential roles of the increasing prevalence of AS in vertebrates in functional innovation will be largely explored in more gene families or genomewide level in the future, which will further

our understanding of how AS contributes to functional innovation.

8. Conclusion

Here we have reviewed evidence from genomewide studies as well as possible avenues for future comparatives studies for the potential of alternative splicing as a source of functional innovation during the evolution of the eukaryotic genome. While it is now clear that AS is prevalent in the human genome, obstacles still remain in the assessment how alternative splicing has evolved through time. The main obstacle lies in that while most other genomic features can be directly measured or estimated from genomic sequences alone, no accurate estimates of alternative splicing can be obtained from genomic sequence analysis. The reliance in transcript sequences availability to measure AS together with the strong bias brought by unequal transcript coverage has hampered the genomewide assessment of AS in all but a few model species and makes difficult any direct comparison between species. This has slowed down the study of how alternative splicing has evolved over time, how AS is regulated, and how it may relate to other genomic features and most crucially to phenotype. The ever-increasing transcript profiling for many more species combined with the use of comparable index estimates will allow addressing a number of evolutionary questions regarding the evolution of AS and its implications for the evolution of transcript diversity and functional innovation.

Conflict of Interests

The authors declare no conflict of interests.

Acknowledgments

The authors wish to thank Humberto Gutierrez for comments on earlier versions of this paper. This work was funded by UK-China Scholarship for Excellence and University of Bath Research Studentship to L. Chen, a CONACyT Scholarship to J. M. Tovar-Corona, and a Royal Society Dorothy Hodgkin Research Fellowship, Royal Society Research Grant, and a Royal Society Research Grant for Fellows to A. O. Urrutia.

References

[1] E. S. Lander, L. M. Linton, B. Birren et al., "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, 2001.

[2] J. C. Venter, M. D. Adams, E. W. Myers et al. et al., "The sequence of the human genome," *Science*, vol. 291, no. 5507, pp. 1304–1351, 2001.

[3] H. R. Crollius, O. Jaillon, A. Bernot et al., "Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence," *Nature Genetics*, vol. 25, no. 2, pp. 235–238, 2000.

[4] B. R. Graveley, "Alternative splicing: increasing diversity in the proteomic world," *Trends in Genetics*, vol. 17, no. 2, pp. 100–107, 2001.

[5] N. Kim, A. V. Alekseyenko, M. Roy, and C. Lee, "The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species," *Nucleic Acids Research*, vol. 35, no. 1, pp. D93–D98, 2007.

[6] T. W. Nilsen and B. R. Graveley, "Expansion of the eukaryotic proteome by alternative splicing," *Nature*, vol. 463, no. 7280, pp. 457–463, 2010.

[7] E. T. Wang, R. Sandberg, S. J. Luo et al., "Alternative isoform regulation in human tissue transcriptomes," *Nature*, vol. 456, no. 7221, pp. 470–476, 2008.

[8] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe, "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing," *Nature Genetics*, vol. 40, pp. 1413–1415, 2008.

[9] S. Stamm, S. Ben-Ari, I. Rafalska et al., "Function of alternative splicing," *Gene*, vol. 344, pp. 1–20, 2005.

[10] B. R. Graveley, A. N. Brooks, J. W. Carlson et al., "The developmental transcriptome of *Drosophila melanogaster*," *Nature*, vol. 471, no. 7339, pp. 473–479, 2011.

[11] S. A. Filichkin, H. D. Priest, S. A. Givan et al., "Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*," *Genome Research*, vol. 20, no. 1, pp. 45–58, 2010.

[12] L. T. Chow, R. E. Gelinas, T. R. Broker, and R. J. Roberts, "An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA," *Cell*, vol. 12, no. 1, pp. 1–8, 1977.

[13] S. M. Berget, C. Moore, and P. A. Sharp, "Spliced segments at the 5′ terminus of adenovirus 2 late mRNA," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, no. 8, pp. 3171–3175, 1977.

[14] F. W. Alt, A. L. M. Bothwell, M. Knapp et al., "Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mRNAs that differ at their 3′ ends," *Cell*, vol. 20, no. 2, pp. 293–301, 1980.

[15] P. Early, J. Rogers, M. Davis et al., "Two mRNAs can be produced from a single immunoglobulin μ gene by alternative RNA processing pathways," *Cell*, vol. 20, no. 2, pp. 313–319, 1980.

[16] I. I. Artamonova and M. S. Gelfand, "Comparative genomics and evolution of alternative splicing: the pessimists' sciene," *Chemical Reviews*, vol. 107, no. 8, pp. 3407–3430, 2007.

[17] E. M. Zdobnov and R. Apweiler, "InterProScan—an integration platform for the signature-recognition methods in InterPro," *Bioinformatics*, vol. 17, no. 9, pp. 847–848, 2001.

[18] A. Bateman, L. Coin, R. Durbin et al., "The Pfam protein families database," *Nucleic Acids Research*, vol. 32, pp. D138–D141, 2004.

[19] J. D. Bendtsen, H. Nielsen, G. von Heijne, and S. Brunak, "Improved prediction of signal peptides: SignalP 3.0," *Journal of Molecular Biology*, vol. 340, no. 4, pp. 783–795, 2004.

[20] A. Krogh, B. Larsson, G. von Heijne, and E. L. L. Sonnhammer, "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes," *Journal of Molecular Biology*, vol. 305, no. 3, pp. 567–580, 2001.

[21] K. Nakai and P. Horton, "PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization," *Trends in Biochemical Sciences*, vol. 24, no. 1, pp. 34–35, 1999.

[22] D. B. Malko, V. J. Makeev, A. A. Mironov, and M. S. Gelfand, "Evolution of exon-intron structure and alternative splicing in fruit flies and malarial mosquito genomes," *Genome Research*, vol. 16, no. 4, pp. 505–509, 2006.

[23] G. Ast, "How did alternative splicing evolve?" *Nature Reviews Genetics*, vol. 5, no. 10, pp. 773–782, 2004.

[24] E. Kim, A. Magen, and G. Ast, "Different levels of alternative splicing among eukaryotes," *Nucleic Acids Research*, vol. 35, no. 1, pp. 125–131, 2007.

[25] B. B. Wang and V. Brendel, "Genomewide comparative analysis of alternative splicing in plants," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 18, pp. 7175–7180, 2006.

[26] E. Kim, A. Goren, and G. Ast, "Alternative splicing: current perspectives," *BioEssays*, vol. 30, no. 1, pp. 38–47, 2008.

[27] M. Chen and J. L. Manley, "Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches," *Nature Reviews Molecular Cell Biology*, vol. 10, no. 11, pp. 741–754, 2009.

[28] B. M. N. Brinkman, "Splice variants as cancer biomarkers," *Clinical Biochemistry*, vol. 37, no. 7, pp. 584–594, 2004.

[29] J. P. Venables, "Unbalanced alternative splicing and its significance in cancer," *BioEssays*, vol. 28, no. 4, pp. 378–386, 2006.

[30] G. S. Wang and T. A. Cooper, "Splicing in disease: disruption of the splicing code and the decoding machinery," *Nature Reviews Genetics*, vol. 8, no. 10, pp. 749–761, 2007.

[31] N. López-Bigas, B. Audit, C. Ouzounis, G. Parra, and R. Guigó, "Are splicing mutations the most frequent cause of hereditary disease?" *FEBS Letters*, vol. 579, no. 9, pp. 1900–1903, 2005.

[32] Y. Yu, P. A. Maroney, J. A. Denker et al., "Dynamic regulation of alternative splicing by silencers that modulate 5′ splice site competition," *Cell*, vol. 135, no. 7, pp. 1224–1236, 2008.

[33] T. A. Hughes, "Regulation of gene expression by alternative untranslated regions," *Trends in Genetics*, vol. 22, no. 3, pp. 119–122, 2006.

[34] R. F. Luco and T. Misteli, "More than a splicing code: integrating the role of RNA, chromatin and non-coding RNA in alternative splicing regulation," *Current Opinion in Genetics & Development*, vol. 21, no. 4, pp. 366–372, 2011.

[35] B. R. Graveley, "Alternative splicing: regulation without regulators," *Nature Structural & Molecular Biology*, vol. 16, no. 1, pp. 13–15, 2009.

[36] R. F. Luco, Q. Pan, K. Tominaga, B. J. Blencowe, O. M. Pereira-Smith, and T. Misteli, "Regulation of alternative splicing by histone modifications," *Science*, vol. 327, no. 5968, pp. 996–1000, 2010.

[37] S. Shukla, E. Kavak, M. Gregory et al., "CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing," *Nature*, vol. 479, no. 7371, pp. 74–79, 2011.

[38] R. F. Luco, M. Allo, I. E. Schor, A. R. Kornblihtt, and T. Misteli, "Epigenetics in alternative pre-mRNA splicing," *Cell*, vol. 144, no. 1, pp. 16–26, 2011.

[39] D. Adams, L. Altucci, S. E. Antonarakis et al., "BLUEPRINT to decode the epigenetic signature written in blood," *Nature Biotechnology*, vol. 30, no. 3, pp. 224–226, 2012.

[40] Y. Barash, J. A. Calarco, W. Gao et al., "Deciphering the splicing code," *Nature*, vol. 465, no. 7294, pp. 53–59, 2010.

[41] E. W. Sayers, T. Barrett, D. A. Benson et al., "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Research*, vol. 37, no. 1, pp. D5–D15, 2009.

[42] S. H. Nagaraj, R. B. Gasser, and S. Ranganathan, "A hitchhiker's guide to expressed sequence tag (EST) analysis," *Briefings in Bioinformatics*, vol. 8, no. 1, pp. 6–21, 2007.

[43] M. S. Boguski, T. M. J. Lowe, and C. M. Tolstoshev, "dbEST—database for 'expressed sequence tags'," *Nature Genetics*, vol. 4, no. 4, pp. 332–333, 1993.

[44] J. M. Johnson, J. Castle, P. Garrett-Engele et al., "Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays," *Science*, vol. 302, no. 5653, pp. 2141–2144, 2003.

[45] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[46] G. Robertson, J. Schein, R. Chiu et al., "De novo assembly and analysis of RNA-seq data," *Nature Methods*, vol. 7, no. 11, pp. 909–912, 2010.

[47] J. A. Martin and Z. Wang, "Next-generation transcriptome assembly," *Nature Reviews Genetics*, vol. 12, no. 10, pp. 671–682, 2011.

[48] R. D. Hawkins, G. C. Hon, and B. Ren, "Next-generation genomics: an integrative approach," *Nature Reviews Genetics*, vol. 11, no. 7, pp. 476–486, 2010.

[49] F. Ozsolak and P. M. Milos, "RNA sequencing: advances, challenges and opportunities," *Nature Reviews Genetics*, vol. 12, no. 2, pp. 87–98, 2011.

[50] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nature Methods*, vol. 5, no. 7, pp. 621–628, 2008.

[51] H. J. Kang, Y. I. Kawasawa, F. Cheng et al., "Spatio-temporal transcriptome of the human brain," *Nature*, vol. 478, no. 7370, pp. 483–489, 2011.

[52] A. Bhasi, P. Philip, V. T. Sreedharan, and P. Senapathy, "AspAlt: a tool for inter-database, inter-genomic and user-specific comparative analysis of alternative transcription and alternative splicing in 46 eukaryotes," *Genomics*, vol. 94, no. 1, pp. 48–54, 2009.

[53] Y. Lee, B. Kim, Y. Shin et al., "ECgene: an alternative splicing database update," *Nucleic Acids Research*, vol. 35, no. 1, pp. D99–D103, 2007.

[54] G. Koscielny, V. Le Texier, C. Gopalakrishnan et al., "ASTD: the alternative splicing and transcript diversity database," *Genomics*, vol. 93, no. 3, pp. 213–220, 2009.

[55] D. Brett, H. Pospisil, J. Valcárcel, J. Reich, and P. Bork, "Alternative splicing and genome complexity," *Nature Genetics*, vol. 30, no. 1, pp. 29–30, 2002.

[56] Z. Y. Kan, D. States, and W. Gish, "Selecting for functional alternative splices in ESTs," *Genome Research*, vol. 12, no. 12, pp. 1837–1845, 2002.

[57] M. Lynch and J. S. Conery, "The origins of genome complexity," *Science*, vol. 302, no. 5649, pp. 1401–1404, 2003.

[58] K. D. Whitney and T. Garland, "Did genetic drift drive increases in genome complexity?" *PLoS Genetics*, vol. 6, no. 8, 2010.

[59] Q. Xu and C. Lee, "Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences," *Nucleic Acids Research*, vol. 31, no. 19, pp. 5635–5643, 2003.

[60] R. I. Skotheim and M. Nees, "Alternative splicing in cancer: noise, functional, or systematic?" *The International Journal of Biochemistry & Cell Biology*, vol. 39, no. 7-8, pp. 1432–1449, 2007.

[61] E. Kim, A. Goren, and G. Ast, "Insights into the connection between cancer and alternative splicing," *Trends in Genetics*, vol. 24, no. 1, pp. 7–10, 2008.

[62] R. E. Green, B. P. Lewis, R. T. Hillman et al., "Widespread predicted nonsense-mediated mRNA decay of alternatively-spliced transcripts of human normal and disease genes," *Bioinformatics*, vol. 19, no. 1, pp. i118–i121, 2003.

[63] B. P. Lewis, R. E. Green, and S. E. Brenner, "Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 1, pp. 189–192, 2003.

[64] Z. Zhang, D. Xin, P. Wang et al., "Noisy splicing, more than expression regulation, explains why some exons are subject to nonsense-mediated mRNA decay," *BMC Biology*, vol. 7, article 23, 2009.

[65] Z. Su, J. Wang, J. Yu, X. Huang, and X. Gu, "Evolution of alternative splicing after gene duplication," *Genome Research*, vol. 16, no. 2, pp. 182–189, 2006.

[66] J. K. Pickrell, A. A. Pai, Y. Gilad, and J. K. Pritchard, "Noisy splicing drives mRNA isoform diversity in human cells," *PLoS Genetics*, vol. 6, no. 12, Article ID e1001236, 2010.

[67] L. Chen, J. M. Tovar-Corona, and A. O. Urrutia, "Increased levels of noisy splicing in cancers, but not for oncogene-derived transcripts," *Human Molecular Genetics*, vol. 20, no. 22, pp. 4422–4429, 2011.

[68] T. A. Thanaraj, F. Clark, and J. Muilu, "Conservation of human alternative splice events in mouse," *Nucleic Acids Research*, vol. 31, no. 10, pp. 2544–2552, 2003.

[69] Q. Pan, M. A. Bakowski, Q. Morris et al., "Alternative splicing of conserved exons is frequently species-specific in human and mouse," *Trends in Genetics*, vol. 21, no. 2, pp. 73–77, 2005.

[70] J. M. Mudge, A. Frankish, J. Fernandez-Banet et al., "The origins, evolution and functional potential of alternative splicing in vertebrates," *Molecular Biology and Evolution*, vol. 28, no. 10, pp. 2949–2959, 2011.

[71] M. Irimia, J. L. Rukov, D. Penny, J. Garcia-Fernandez, J. Vinther, and S. W. Roy, "Widespread evolutionary conservation of alternatively spliced exons in Caenorhabditis," *Molecular Biology and Evolution*, vol. 25, no. 2, pp. 375–382, 2008.

[72] M. Irimia, J. L. Rukov, S. W. Roy, J. Vinther, and J. Garcia-Fernandez, "Quantitative regulation of alternative splicing in evolution and development," *BioEssays*, vol. 31, no. 1, pp. 40–50, 2009.

[73] M. Long, E. Betrán, K. Thornton, and W. Wang, "The origin of new genes: glimpses from the young and old," *Nature Reviews Genetics*, vol. 4, no. 11, pp. 865–875, 2003.

[74] P. Dehal and J. L. Boore, "Two rounds of whole genome duplication in the ancestral vertebrate," *PLoS Biology*, vol. 3, no. 10, pp. 1700–1708, 2005.

[75] N. M. Kopelman, D. Lancet, and I. Yanai, "Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms," *Nature Genetics*, vol. 37, no. 6, pp. 588–589, 2005.

[76] L. Jin, K. Kryukov, J. C. Clemente et al., "The evolutionary relationship between gene duplication and alternative splicing," *Gene*, vol. 427, no. 1-2, pp. 19–31, 2008.

[77] A. L. Hughes and R. Friedman, "Alternative splicing, gene duplication and connectivity in the genetic interaction network of the nematode worm *Caenorhabditis elegans*," *Genetica*, vol. 134, no. 2, pp. 181–186, 2008.

[78] H. Lin, S. Ouyang, A. Egan et al., "Characterization of paralogous protein families in rice," *BMC Plant Biology*, vol. 8, article 18, 2008.

[79] J. Roux and M. Robinson-Rechavi, "Age-dependent gain of alternative splice forms and biased duplication explain the relation between splicing and duplication," *Genome Research*, vol. 21, no. 3, pp. 357–363, 2011.

[80] D. Talavera, C. Vogel, M. Orozco, S. A. Teichmann, and X. de la Cruz, "The (In)dependence of alternative splicing and gene duplication," *PLoS Computational Biology*, vol. 3, no. 3, pp. 375–388, 2007.

[81] D. Wegmann, I. Dupanloup, and L. Excoffier, "Width of gene expression profile drives alternative splicing," *PLoS ONE*, vol. 3, no. 10, article e3587, 2008.

[82] E. L. Heinzen, D. Ge, K. D. Cronin et al., "Tissue-specific genetic control of splicing: implications for the study of complex traits," *PLoS Biology*, vol. 6, no. 12, Article ID e1000001, pp. 2869–2879, 2008.

[83] B. Hartmann, R. Castelo, B. Miñana et al., "Distinct regulatory programs establish widespread sex-specific alternative splicing in *Drosophila melanogaster*," *RNA*, vol. 17, no. 3, pp. 453–468, 2011.

[84] R. Blekhman, J. C. Marioni, P. Zumbo, M. Stephens, and Y. Gilad, "Sex-specific and lineage-specific alternative splicing in primates," *Genome Research*, vol. 20, no. 2, pp. 180–189, 2010.

[85] D. W. MacLean, T. H. Meedel, and K. E. M. Hastings, "Tissue-specific alternative splicing of ascidian troponin I isoforms: redesign of a protein isoform-generating mechanism during chordate evolution," *The Journal of Biological Chemistry*, vol. 272, no. 51, pp. 32115–32120, 1997.

[86] W. P. Yu, S. Brenner, and B. Venkatesh, "Duplication, degeneration and subfunctionalization of the nested synapsin-Timp genes in Fugu," *Trends in Genetics*, vol. 19, no. 4, pp. 180–183, 2003.

[87] J. A. Lister, J. Close, and D. W. Raible, "Duplicate mitf genes in zebrafish: complementary expression and conservation of melanogenic potential," *Developmental Biology*, vol. 237, no. 2, pp. 333–344, 2001.

[88] J. Altschmied, J. Delfgaauw, B. Wilde et al., "Subfunctionalization of duplicate mitf genes associated with differential degeneration of alternative exons in fish," *Genetics*, vol. 161, no. 1, pp. 259–267, 2002.

[89] S. Short and L. Z. Holland, "The evolution of alternative splicing in the *Pax* family: the view from the basal chordate amphioxus," *Journal of Molecular Evolution*, vol. 66, no. 6, pp. 605–620, 2008.

[90] L. Chen, Q. J. Zhang, W. Wang, and Y. Q. Wang, "Spatiotemporal expression of *Pax* genes in amphioxus: insights into *Pax*-related organogenesis and evolution," *Science China Life Sciences*, vol. 53, no. 8, pp. 1031–1040, 2010.

[91] T. D. Barber, M. C. Barber, T. E. Cloutier, and T. B. Friedman, "*PAX3* gene structure, alternative splicing and evolution," *Gene*, vol. 237, no. 2, pp. 311–319, 1999.

[92] S. Singh, R. Mishra, N. A. Arango, J. M. Deng, R. R. Behringer, and G. F. Saunders, "Iris hypoplasia in mice that lack the alternatively spliced *Pax6(5a)* isoform," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 10, pp. 6812–6815, 2002.

[93] L. Z. Holland and S. Short, "Alternative splicing in development and function of chordate endocrine systems: a focus on *Pax* genes," *Integrative and Comparative Biology*, vol. 50, no. 1, pp. 22–34, 2010.

Hindawi Publishing Corporation
International Journal of Evolutionary Biology
Volume 2012, Article ID 843167, 8 pages
doi:10.1155/2012/843167

Research Article

Where Do Phosphosites Come from and Where Do They Go after Gene Duplication?

Guillaume Diss, Luca Freschi, and Christian R. Landry

Département de Biologie, PROTEO and Institut de Biologie Intégrative et des Systèmes, Université Laval, Pavillon Charles-Eugène-Marchand, 1030, Avenue de la Médecine, Québec, QC, Canada G1V 0A6

Correspondence should be addressed to Christian R. Landry, christian.landry@bio.ulaval.ca

Received 21 March 2012; Accepted 3 May 2012

Academic Editor: Frédéric Brunet

Gene duplication followed by divergence is an important mechanism that leads to molecular innovation. Divergence of paralogous genes can be achieved at functional and regulatory levels. Whereas regulatory divergence at the transcriptional level is well documented, little is known about divergence of posttranslational modifications (PTMs). Protein phosphorylation, one of the most important PTMs, has recently been shown to be an important determinant of the retention of paralogous genes. Here we test whether gains and losses of phosphorylated amino acids after gene duplication may specifically modify the regulation of these duplicated proteins. We show that when phosphosites are lost in one paralog, transitions from phosphorylated serines and threonines are significantly biased toward negatively charged amino acids, which can mimic their phosphorylated status in a constitutive manner. Our analyses support the hypothesis that divergence between paralogs can be generated by a loss of the posttranslational regulatory control on a function rather than by the complete loss of the function itself. Surprisingly, these favoured transitions cannot be reached by single mutational steps, which suggests that the function of a phosphosite needs to be completely abolished before it is restored through substitution by these phosphomimetic residues. We conclude by discussing how gene duplication could facilitate the transitions between phosphorylated and phosphomimetic amino acids.

1. Introduction

Gene duplication is one of the most prominent mechanisms by which organisms acquire new functions [1]. Spectacular examples of such gains of function resulting from gene duplications are the evolution of trichromatic vision in primates [2], the evolution of human beta-globin genes that are involved in the oxygen transport at different developmental stages [3] as well as the expansion of the family of immunoglobulins and other immunity-related genes that shaped the vertebrate immune system [4, 5]. Because of the central role of gene duplication in evolution, there has been a profound interest for a better understanding of how these new functions evolve at the molecular level [6], for determining at what rate gene duplication occurs [7–9] and for testing whether the retention of paralogous genes necessarily requires the evolution of new functions [6, 10, 11]. One of the most important challenges has been to determine mechanistically how specific mutations translate into new functions, as establishing sequence-function relationships remains a difficult task [12].

After a gene duplication event, the two sister paralogs are identical copies of their ancestor and encode two identical functions, thus relaxing the selective constraints on each paralog [8]. Under most evolutionary models, both paralogs have to diverge to be retained on evolutionary time scales, otherwise one paralog would be lost and the system would return to its ancestral state (nonfunctionalization) [6]. There are two ways for paralogs to diverge in function. The first one is the acquisition of new functions by one or both of the two paralogs, a mechanism called neofunctionalization [1, 8, 10]. The second mechanism, called subfunctionalization, implies the complementary partitioning of the ancestral function between the two paralogs by losses of functions [8, 10, 13]. These two mechanisms are not mutually exclusive

because the ancestral function can be partitioned by subfunctionalization and then one or both paralogs may acquire new functions by neofunctionalization, a mechanism called neosubfunctionalization [14]. An increase in the dosage of a gene product by the addition of a second identical copy of the ancestral gene can also contribute to the retention of paralogous pairs, without the need for the gain or loss of functions [15, 16].

Divergence between paralogs does not necessarily imply a divergence in a specific function but can also involve a change in the regulation of that function. For instance, the regulatory control of a protein function can be modified at the transcriptional or at the posttranslational level. Divergence in expression pattern of duplicated transcript is well documented [1, 10, 17, 18]. For example, Gu et al. showed that a large fraction of ancient duplicated gene pairs in yeast shows divergent gene expression patterns [18]. A more recent study showed that nearly half of the genes that duplicated after a whole genome duplication event (WGD) in a forest tree species have diverged in expression by a random degeneration process [19]. However, little is known about the divergence of regulation by posttranslational modifications (PTMs), which take place after transcription and translation and directly affect protein activities [20].

PTMs are covalent modifications of one or more amino acids that affect the activity of a protein, its localization in the cell, its turnover rate, and its interactions with other molecules [21]. Cells use a wide range of different PTMs to exert distinct regulations on proteins. Although only 20 amino acids are encoded by the genetic code, more than 200 amino acid variants or their derivatives are found in proteins after PTMs [22]. Phosphorylation, the addition of a phosphate moiety from an ATP donor to a serine (Ser), threonine (Thr), or tyrosine (Tyr) residue by a protein kinase, is by far the best-known PTM, as it is the most common and is involved in the regulation of key biological processes of fundamental and medical interest, such as signal transduction and cell-cycle regulation [23]. Phosphorylation of these amino acids modifies their biochemical properties in several manners. Of particular interest for this study is the addition of a phosphate group that brings two new negative charges that allow the formation of a salt bridge or that contribute to the local charge of the protein [24]. Given that a phosphate group is a relatively large molecule, phosphorylation can also have sterical effects. Such properties can notably induce conformational changes of the protein, modify its catalytic activity, or block the access to its catalytic site, which result in the activation or inhibition of the activity of the target protein by direct or allosteric effects [24].

Several of the effects of protein phosphorylation can be mimicked by the negatively charged amino acids aspartic acid (Asp) and glutamic acid (Glu). Indeed, the biochemical properties of these amino acids are close to those of phosphorylated Ser or Thr residues [25]. In particular conditions, Asp and Glu are constitutive functional equivalents of phosphosites in a phosphorylated state. This functional resemblance has been exploited by biochemists by replacing Ser and Thr residues by Asp and Glu in proteins of interest in order to mimic their phosphorylated status. This molecular mimicry led them to call Asp and Glu phosphomimetic amino acids [25]. This trick appears to have been also used by nature to evolve new phosphosites. A striking example comes from the evolution of the Activation Induced cytidine Deaminase (AID) across vertebrates, an enzyme involved in the generation of antibody diversity. The interaction of this enzyme with the Replication Protein A (RPA) promotes AID access to transcribed double-stranded DNA during immunoglobulin class switch recombination. This interaction requires a negative charge on AID, which is provided by an Asp in bony fish. In these organisms, the enzyme is constitutively capable of interacting with RPA. In amphibians and mammals, the function of the Asp residue is carried out by a phosphorylatable Ser (pSer), which allows the regulation of the protein interaction by protein kinases in a condition-specific fashion [26]. It was recently suggested that this type of evolutionary transitions might be common. Globally, it was shown that pSer tends to evolve from or to phosphomimetic amino acids (Asp and Glu) when gained and lost, respectively, throughout the evolution of eukaryotes [27, 28].

Protein phosphoregulation has been suggested to play a role in the evolutionary fate of paralogous proteins. Most studies done so far focused on the paralogous genes of the budding yeast *Saccharomyces cerevisiae* because its phosphoproteome has been intensely studied [29–31]. Using the yeast paralogs that derive from the WGD event, Amoutzias et al. showed that the number of phosphosites on a phosphoprotein is an important determinant for the retention of its duplicated descendants [32]. In a following study, Freschi et al. studied the gains and losses of phosphosites in paralogous phosphoproteins and found that the great majority of them are present in one paralog and not in the other. This divergence was shown to be principally driven by losses rather than gains of phosphosites on one paralog [33]. Finally, Kaganovich and Snyder found that phosphosites tend to diverge more asymmetrically than nonphosphorylated amino acids, playing thus an important role in paralogous genes divergence and retention [34]. These observations raise the question of where do phosphosites come from and where do they go after a gene duplication. According to the observations on phosphomimetic amino acids described above, gains and losses of phosphosites could represent two distinct types of divergence. On the one hand, the gain or the loss of phosphosites from or to a nonphosphomimetic residue would represent a divergence in the function of the protein. On the other hand, a gain or a loss could occur from or to phosphomimetic residues, leading to a modification of the control of the charged residue by the cell rather than a modification of function per se. Here we test whether this second scenario could have contributed to the divergence of paralogous proteins using the yeast phosphoproteome as a model.

2. Methods

2.1. Dataset. All analyses were performed using the dataset we compiled in a previous study [33], and that is

available at http://www.bio.ulaval.ca/landrylab/download/ (Dataset 1). This dataset contains 20,342 phosphosites on 2688 proteins from eight large-scale studies [29–31, 35–39]. It also provides the alignments of all *S. cerevisiae* WGD paralogous genes with their ancestral sequence and with the orthologs of *Lachancea kluyveri* and *Zygosaccharomyces rouxii*. The alignments were performed using MUSCLE [40] while the ancestral sequence was inferred using the Codeml method implemented in PAML [41]. We chose to analyze only two species that diverged before the WGD event for the following reasons. The majority of phosphorylation sites are located in disordered regions [42], and these regions are fast evolving. Alignment of sequences from distantly related species leads to spurious alignments or to alignments that may contain several indels. Indels decrease the number of phosphorylation sites available for the analysis, as ancestral sequences cannot be computed at these positions. Further, in Freschi et al. [33], we performed the analyses including an additional species that diverged prior to the whole-genome duplication, and we found that this did not significantly affect our results. Finally, this dataset also provides information about the localization of each residue in ordered or disordered regions of the protein, according to predictions made with DISOPRED [43].

2.2. Approaches to Study Gains and Losses of Phosphosites. We applied different approaches to study gains and losses coming from or going to negatively charged amino acids. In the first approach, we used the ancestral sequence as a reference to assess the presence of a gain or a loss at a specific position. For the gains, we compared the proportion of phosphomimetic amino acids in the ancestral sequence (Asp or Glu) going to pSer or pThr to the proportion of phosphomimetic amino acids going to cSer and cThr. For the losses, we compared the proportion of phosphorylated residues (pSer and pThr) coming to Asp or Glu to the proportion of nonphosphorylated residues (cSer and cThr) coming to Asp or Glu, respectively. We required the ancestral sequence to have a phosphorylatable residue and one of the two paralogs to be phosphorylated at the homologous position. Comparisons of proportions were performed using Fisher's exact tests as implemented in R [44]. In our second approach, we used a parsimony method to calculate the same proportions. This time we used the sequences of *L. kluyveri* and *Z. rouxii* as reference. In the case of a gain of phosphosites, we required the presence of the same negatively charged residue (Asp or Glu) in the reference species as well as in one of the two paralogs and a phosphorylatable residue (Ser or Thr) in the other paralog. In the case of losses of phosphosites, we required the presence of the same phosphorylatable residue (Ser or Thr) in the reference species as well as in one of the two paralogs and a negatively charged residue (Asp or Glu) in the other paralog. All proportions were calculated by dividing the number of sites coming from or going to an Asp or a Glu by the number of sites that come from or go to any of the 17 nonphosphorylatable amino acids following the same criteria (Figure 1).

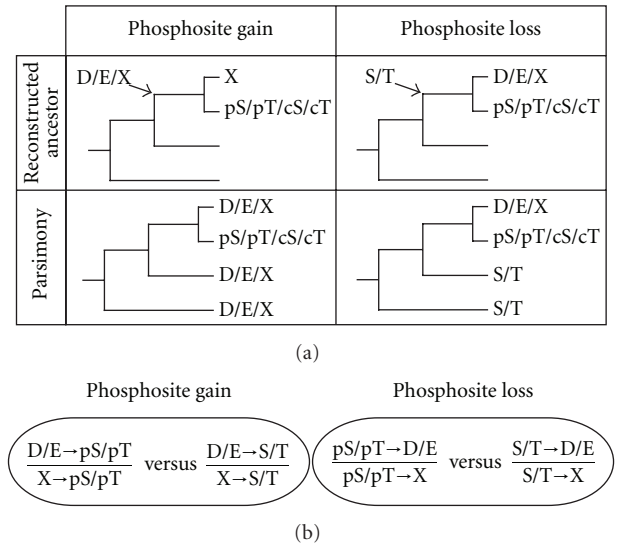


Figure 1: Algorithm used to calculate and compare the proportions of transitions between phosphorylated and phosphomimetic residues relative to control sites. (a) Phosphosite (pS, pT) gains from phosphomimetic amino acids were identified as cases where only one of the paralog has a phosphosite and the ancestral sequence has a phosphomimetic residue at the same position. Control sites (cS, cT) were identified in the same way but considering Ser and Thr that are not known to be phosphorylated. The ancestral sequence was inferred using likelihood or parsimony approaches. Phosphosites losses to phosphomimetic amino acids were identified as cases where one paralog has a phosphosite in a position that is occupied by a phosphomimetic amino acid in the other paralog and a phosphorylatable amino acid at the same position in the ancestral sequence. (b) The proportion of pS or pT that evolved from or to D or E was compared to the proportion of cS or cT that evolved from or to D or E. X represents any amino acid with the exception of Ser, Thr and Tyr.

3. Results

The phosphoproteome of *S. cerevisiae* is the best described among eukaryotes and has been mapped by mass spectrometry, leading to the identification of high-confidence phosphosites [29–31]. We assembled a data set [33] that consists of 2,726 phosphosites (Ser, 82%; Thr, 16%; Tyr, 2%) that belong to one or the other member of the 352 pairs of yeast WGD paralogs for which at least one of the two proteins is a phosphoprotein. We inferred the ancestral sequence for each pair of paralogs using alignments with orthologous sequences from *L. kluyveri* and *Z. rouxii*, two species that diverged from *S. cerevisiae* before the WGD event. For each pair, we aligned all five sequences, we mapped the phosphosites on the sequences of the paralogs and analysed phosphosites that diverged, that is, cases where a phosphorylatable residue was present in only one paralog.

Under a scenario where gains of phosphosites would result from selection for transitions from phosphomimetic amino acids to phosphorylated residues, we would expect phosphorylated Ser or Thr (pSer and pThr, resp.) to evolve more often from Asp or Glu than nonphosphorylated ones

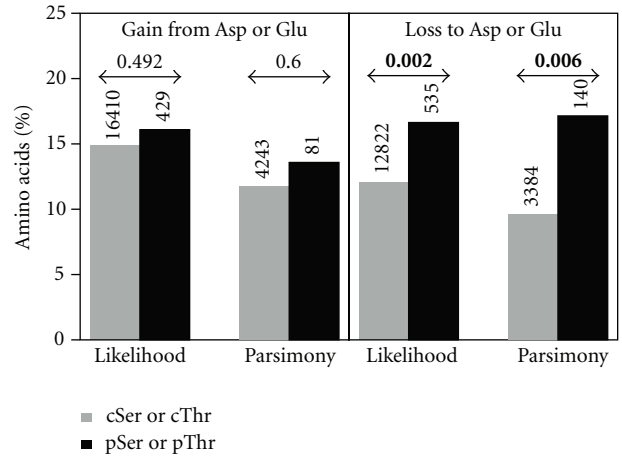


Figure 2: Phosphosites that are differentially lost in paralogous phosphoproteins evolve toward negatively charged residues. Each bar represents the percentage of sites (pSer and pThr, cSer and cThr) that evolved from or to Asp or Glu. Numbers above the bars represent the total number of pSer, cSer, pThr, or cThr sites that were gained or lost. Numbers above the arrows indicate *P*-values of the Fisher's exact tests, bold ones being below 0.05.

(cSer and cThr, resp.). Similarly, under a scenario where losses of phosphosites would result from transitions from phosphorylated residues to phosphomimetic amino acids, we would expect pSer and pThr to evolve more often to Asp and Glu than equivalent cSer and cThr. We tested these two hypotheses as described in Figure 1. In the first case, we compared the proportion of pSer and pThr that were gained from Asp and Glu with that of cSer and cThr, that is, all serines and threonines from the same set of proteins that were gained from Asp and Glu but that are not known to be phosphorylated. In the second case, we compared the ratio of sites that were lost and replaced by phosphomimetic residues in only one paralog with the ratios derived from cSer and cThr. We performed the analysis using paralogous ancestral sequences inferred with a likelihood method and also using a parsimonious approach, whereby the ancestral state of phosphosites was inferred based on the conservation of the site in one of the two paralogs and its two orthologs (Figure 1(a)). Global results are presented in Figure 2, and detailed analyses are presented in Figure 3.

A gobal analysis of pSer, pThr, Asp, and Glu shows that phosphosites tend to be lost to Asp and Glu more frequently than cSer and cThr, and this holds true for both likelihood (16.6% versus 12.1%, resp., $P = 0.002$) and parsimony (17.1% versus 9.6%, resp., $P = 0.006$) reconstruction methods (Figure 2). However, although there is a tendency towards the gains of phosphosites from Asp and Glu, the observed differences are not significant (Figure 2). When studied separately, phosphosites in ordered and disordered regions show the same global tendency to go toward phosphomimetic amino acids (likelihood: 17.5% versus 10.0% in ordered regions, $P = 0.058$; 16.5% versus 13.7% in disordered regions, $P = 0.086$, parsimony: 20.0% versus 8.1% in ordered regions, $P = 0.076$; 16.7% versus 11.7% in disordered regions, $P = 0.110$). Further, we found that phosphosites are not preferentially gained from phosphomimetic amino acids in disordered regions, while there is a nonsignificant tendency for this type of transition in ordered regions (likelihood: 16.0% versus 15.7% in disordered regions, $P = 0.943$; 18.8% versus 13.7% in ordered regions, $P = 0.294$, parsimony: 14.1% versus 14.2% in disordered regions, $P = 1.000$; 11.8% versus 10.2% in ordered regions, $P = 0.691$). This suggests that the effect might be more important in ordered regions of proteins, as would be expected if these residues were playing structural roles. Because the distinction between order and disorder reduces the number sites in each category and does not provide opposite results, we considered both regions simultaneously in the following analyses.

We also examined which class of substitution could be contributing to this overall result (Figure 3). We first found that pSer and pThr that were gained after gene duplication follow trends that are in the expected direction although some of the comparisons are not statistically significant and other results are in the opposite direction (Figure 3). However, this detailed analysis showed that pSer is significantly more likely to evolve to Glu than cSer (11.6% versus 5.3%, $P = 0.008$) while pThr evolves significantly more frequently to Asp than cThr (9.8% versus 4.3% resp., $P = 0.013$).

4. Discussion

Protein phosphorylation is known to have a key role in regulating protein activities [45]. Evolutionary events such as gains and losses of phosphosites can lead to changes in protein regulation, thus rewiring the protein regulatory network of the cell [33]. In the literature, there is evidence for gains of new phosphosites coming from negatively charged residues among orthologs [26, 27] as well as cases of losses of phosphosites to these amino acids [28]. The biochemical properties of Glu and Asp mimic the ones of pSer and pThr with the exception that their charge is not regulatable [25]. These observations led us to hypothesize that coding sequence divergence of paralogous genes by neo- and subfunctionalization does not strictly involve the apparition or the partitioning of protein function. Paralogous genes could also diverge in how these functions are regulated. Divergence in the regulatory control is well known at the transcriptional level [19, 46] but has not been specifically addressed at the posttranslational level. We tested this hypothesis on the complete set of WGD phosphoproteins of the budding yeast *S. cerevisiae*.

Using two different methods to infer the ancestral state of phosphorylated and nonphosphorylated Ser and Thr, we found that pSer and pThr globally have a tendency to evolve from negatively charged amino acids in paralogous phosphoproteins compared to their nonphosphorylated counterparts. The tendencies observed are in agreement with our hypothesis and with the observations made by Pearlman et al. across eukaryotes [27]. However, the observed differences are not significant, which could be explained

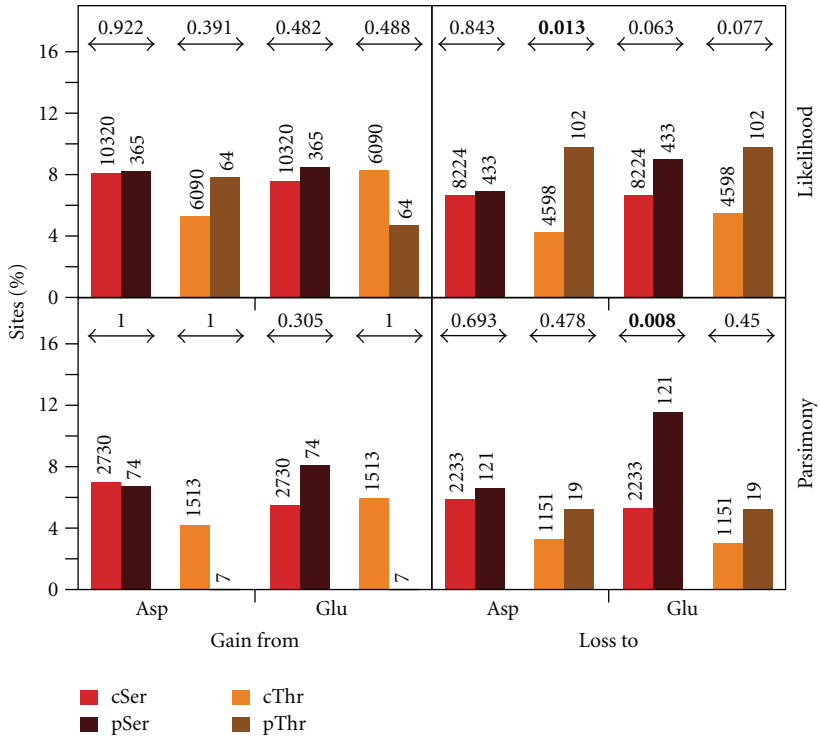


Figure 3: Detailed analysis of the patterns of evolution of pSer and pThr sites. Each bar represents the percentage of sites (pSer, cSer, pThr, or cThr) that evolved from or to Asp or Glu. Numbers above the bars represent the total number of pSer, cSer, pThr, or cThr sites that were gained or lost. Numbers above the arrows indicate *P*-values of the Fisher's exact tests, bold ones being below 0.05. The top panel shows results obtained by ancestral sequence reconstruction using a likelihood approach and the bottom panel using parsimony.

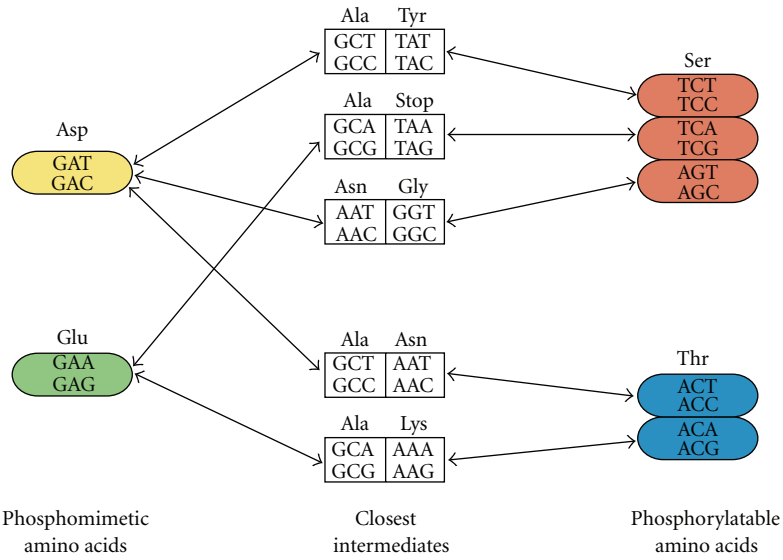


Figure 4: Transitions between phosphorylatable and phosphomimetic amino acids need to go through a nonnegatively charged intermediate.

by a few nonexclusive scenarios. First, we are looking at a narrow evolutionary window (100 My), which contrasts with the analysis conducted by Pearlman et al., who used aligned sequences from organisms spanning the entire tree of life [27]. Further, the mechanism proposed may apply primarily to few sites and in ordered regions of proteins. Only few phosphosites in these regions could be analysed here since the majority of them are found in disordered regions [42], which reduces the statistical power of our analysis. Our results regarding gains of phosphosites are in

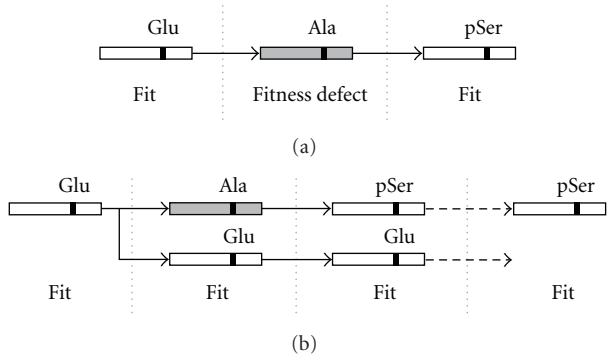


Figure 5: A duplication event could provide the conditions for the intermediate nonfunctional site to be neutral, which would allow a transition without affecting the fitness of the organism. (a) Without a duplication event, the loss of a negative charge could have deleterious effects if the charge is important for the function of the protein. (b) The redundant paralogous gene copy could serve as a backup and prevent deleterious effects created by the loss of the charge. The backup copy could then be retained or lost. In the latter case, the system would be different from its ancestor.

line with this hypothesis. Finally, a significant fraction of phosphosites are thought to be nonfunctional [42]. Because these nonfunctional sites are not under selective pressure, they may contribute to decrease the signal coming from functional sites. Nevertheless, from our results, we cannot rule out the possibility that gains of phosphosites are not more likely to derive from phosphomimetic residues after gene duplications. A larger sample size, the study of a time window of a different length and a better knowledge of the functional importance of phosphosites may be needed to provide a final answer.

Following the same approach, we examined whether phosphorylated residues, when lost, are more likely to be replaced by Asp and Glu than when nonphosphorylated equivalent residues are lost. We found that this is the case globally and also when considering individual cases for both pSer and pThr; pSer are more likely to be replaced by Glu residues while pThr by Asp residues. A similar trend was detectable for the transitions from pThr to Glu. These results are in agreement with those from Kurmangaliyev et al. [28] who also showed that pSer are more likely to evolve to phosphomimetic amino acids than cSer in the divergence of orthologs between species. Our results show that the evolutionary trajectories of pSer and pThr provide a mechanism for paralogous protein divergence. Our analyses support the hypothesis that divergence between paralogs can be generated by a loss of the posttranslational regulatory control on a function rather than by the complete loss of the function itself. Indeed, the substitution of a phosphosite for an Asp or a Glu residue may block one paralog into a single constitutive functional state whereas the other one remains regulatable by protein kinases and phosphatases.

Our results raise the question of how these transitions are made possible during evolution. The genetic code is organized in such a way that transitions between phosphorylatable and phosphomimetic amino acids involve a transition state with an amino acid that is not negatively charged, except for transitions between two Asp and two Ser codons that involve a Tyr residue (Figure 4). However, Tyr is only rarely phosphorylated in yeast, and Tyr residues are not phosphorylated by the serine/threonine kinases [47], which suggests that this path would not be favoured. Our results also suggest that this evolutionary route is uncommon. A nonnegatively charged intermediate could lead to a complete loss of the function that was performed by the negative charge and could thus be deleterious (Figure 5(a)). Here we propose that the relaxed constraints that follow a gene duplication event could provide the mean to reach this intermediate state and to go beyond (Figure 5(b)). After gene duplication, when one of the duplicated copies is lost, the system is assumed to go back to its ancestral state, a process called nonfunctionalization [8]. However, following our model, the duplicated copy could serve as a backup for a transition period, which would allow the other copy to reach a state that would have been unreachable otherwise [48–50]. After the loss of the backup copy, the system would remain different from its ancestral state since the phosphorylation profile and thus the phosphoregulation of this protein has changed. The term nonfunctionalization may thus not be suitable for such cases. In the case of a WGD event, where the vast majority of the duplicated genes are eventually lost and are thought to return back to their ancestral state, these 2-step transitions could potentially lead to a great burst in the evolution of phosphoregulation. Further studies at different time points following gene duplication would be needed to determine how important this mechanism could be for the evolution of phosphorylation networks.

Authors' Contribution

G. Diss and L. Freschi equally contributed to this work. G. Diss, L. Freschi, and C. R Landry planned the analyses, G. Diss and L. Freschi performed the analyses and wrote the paper. C. R Landry edited the manuscript.

Acknowledgments

This work was supported by a Canadian Institute of Health Research (CIHR) Grant GMX-191597 and Natural Sciences and Engineering Research Council of Canada discovery grant to C. R Landry. C. R Landry is a CIHR New Investigator. G. Diss, and L. Freschi were supported by fellowships from the Quebec Research Network on Protein Function, Structure and Engineering (PROTEO). The authors thank the members of the Landry laboratory, two anonymous referees, and N. Aubin-Horth for comments on the paper.

References

[1] S. Ohno, *Evolution by Gene Duplication*, Allen & Unwin/ Springer, London, UK, 1970.

[2] K. S. Dulai, M. von Dornum, J. D. Mollon, and D. M. Hunt, "The evolution of trichromatic color vision by opsin gene duplication in New World and Old World primates," *Genome Research*, vol. 9, no. 7, pp. 629–638, 1999.

[3] A. Efstratiadis, J. W. Posakony, T. Maniatis et al., "The structure and evolution of the human β-globin gene family," *Cell*, vol. 21, no. 3, pp. 653–668, 1980.

[4] J. Z. Zhang, "Evolution by gene duplication: an update," *Trends in Ecology & Evolution*, vol. 18, no. 6, pp. 292–298, 2003.

[5] J. Boulais, M. Trost, C. R. Landry et al., "Molecular characterization of the evolution of phagosomes," *Molecular Systems Biology*, vol. 6, article 423, 2010.

[6] M. Hurles, "Gene duplication: the genomic trade in spare parts," *PLoS Biology*, vol. 2, no. 7, article E206, 2004.

[7] A. Wagner, "Birth and death of duplicated genes in completely sequenced eukaryotes," *Trends in Genetics*, vol. 17, no. 5, pp. 237–239, 2001.

[8] M. Lynch and J. S. Conery, "The evolutionary fate and consequences of duplicate genes," *Science*, vol. 290, no. 5494, pp. 1151–1155, 2000.

[9] M. Lynch, W. Sung, K. Morris et al., "A genome-wide view of the spectrum of spontaneous mutations in yeast," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 27, pp. 9272–9277, 2008.

[10] A. Force, M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, and J. Postlethwait, "Preservation of duplicate genes by complementary, degenerative mutations," *Genetics*, vol. 151, no. 4, pp. 1531–1545, 1999.

[11] A. van Hoof, "Conserved functions of yeast genes support the duplication, degeneration and complementation model for gene duplication," *Genetics*, vol. 171, no. 4, pp. 1455–1461, 2005.

[12] A. M. Dean and J. W. Thornton, "Mechanistic approaches to the study of evolution: the functional synthesis," *Nature Reviews Genetics*, vol. 8, no. 9, pp. 675–688, 2007.

[13] M. Lynch and A. Force, "The probability of duplicate gene preservation by subfunctionalization," *Genetics*, vol. 154, no. 1, pp. 459–473, 2000.

[14] X. L. He and J. Z. Zhang, "Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution," *Genetics*, vol. 169, no. 2, pp. 1157–1164, 2005.

[15] F. A. Kondrashov, I. B. Rogozin, Y. I. Wolf, and E. V. Koonin, "Selection in the evolution of gene duplications," *Genome Biology*, vol. 3, no. 2, pp. 0008–0008.9, 2002.

[16] F. A. Kondrashov and E. V. Koonin, "A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications," *Trends in Genetics*, vol. 20, no. 7, pp. 287–290, 2004.

[17] S. D. Ferris and G. S. Whitt, "Evolution of the differential regulation of duplicate genes after polyploidization," *Journal of Molecular Evolution*, vol. 12, no. 4, pp. 267–317, 1979.

[18] Z. L. Gu, D. Nicolae, H. H. S. Lu, and W. H. Li, "Rapid divergence in expression between duplicate genes inferred from microarray data," *Trends in Genetics*, vol. 18, no. 12, pp. 609–613, 2002.

[19] E. Rodgers-Melnick, S. P. Mane, P. Dharmawardhana et al., "Contrasting patterns of evolution following whole genome versus tandem duplication events in Populus," *Genome Research*, vol. 22, no. 1, pp. 95–105, 2012.

[20] A. M. Moses and C. R. Landry, "Moving from transcriptional to phospho-evolution: generalizing regulatory evolution?" *Trends in Genetics*, vol. 26, no. 11, pp. 462–467, 2010.

[21] M. Mann and O. N. Jensen, "Proteomic analysis of post-translational modifications," *Nature Biotechnology*, vol. 21, no. 3, pp. 255–261, 2003.

[22] J. Seo and K. J. Lee, "Post-translational modifications and their biological functions: proteomic analysis and systematic approaches," *Journal of Biochemistry and Molecular Biology*, vol. 37, no. 1, pp. 35–44, 2004.

[23] T. Hunter, "Signaling—2000 and beyond," *Cell*, vol. 100, no. 1, pp. 113–127, 2000.

[24] Z. Serber and J. E. Ferrell, "Tuning bulk electrostatics to regulate protein function," *Cell*, vol. 128, no. 3, pp. 441–444, 2007.

[25] M. K. Tarrant and P. A. Cole, "The chemical biology of protein phosphorylation," *Annual Review of Biochemistry*, vol. 78, pp. 797–825, 2009.

[26] U. Basu, Y. B. Wang, and F. W. Alt, "Evolution of phosphorylation-dependent regulation of activation-induced cytidine deaminase," *Molecular Cell*, vol. 32, no. 2, pp. 285–291, 2008.

[27] S. M. Pearlman, Z. Serber, and J. E. Ferrell, "A mechanism for the evolution of phosphorylation sites," *Cell*, vol. 147, no. 4, pp. 934–946, 2011.

[28] Y. Z. Kurmangaliyev, A. Goland, and M. S. Gelfand, "Evolutionary patterns of phosphorylated serines," *Biology Direct*, vol. 6, article 8, 2011.

[29] P. Beltrao, J. C. Trinidad, D. Fiedler et al., "Evolution of phosphoregulation: comparison of phosphorylation patterns across yeast species," *PLoS Biology*, vol. 7, no. 6, Article ID e1000134, 2009.

[30] F. Gnad, L. M. F. de Godoy, J. Cox et al., "High-accuracy identification and bioinformatic analysis of in vivo protein phosphorylation sites in yeast," *Proteomics*, vol. 9, no. 20, pp. 4642–4652, 2009.

[31] C. P. Albuquerque, M. B. Smolka, S. H. Payne, V. Bafna, J. Eng, and H. L. Zhou, "A multidimensional chromatography technology for in-depth phosphoproteome analysis," *Molecular & Cellular Proteomics*, vol. 7, no. 7, pp. 1389–1396, 2008.

[32] G. D. Amoutzias, Y. He, J. Gordon, D. Mossialos, S. G. Oliver, and Y. van de Peer, "Posttranslational regulation impacts the fate of duplicated genes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 7, pp. 2967–2971, 2010.

[33] L. Freschi, M. Courcelles, P. Thibault, S. W. Michnick, and C. R. Landry, "Phosphorylation network rewiring by gene duplication," *Molecular Systems Biology*, vol. 7, article 504, 2011.

[34] M. Kaganovich and M. Snyder, "Phosphorylation of yeast transcription factors correlates with the evolution of novel sequence and function," *Journal of Proteome Research*, vol. 11, no. 1, pp. 261–268, 2012.

[35] A. Gruhler, J. V. Olsen, S. Mohammed et al., "Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway," *Molecular & Cellular Proteomics*, vol. 4, no. 3, pp. 310–327, 2005.

[36] B. Bodenmiller, L. N. Mueller, M. Mueller, B. Domon, and R. Aebersold, "Reproducible isolation of distinct, overlapping segments of the phosphoproteome," *Nature Methods*, vol. 4, no. 3, pp. 231–237, 2007.

[37] A. Chi, C. Huttenhower, L. Y. Geer et al., "Analysis of phosphorylation sites on proteins from *Saccharomyces cerevisiae* by electron transfer dissociation (ETD) mass spectrometry," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 7, pp. 2193–2198, 2007.

[38] X. Li, S. A. Gerber, A. D. Rudner et al., "Large-scale phosphorylation analysis of α-factor-arrested *Saccharomyces cerevisiae*," *Journal of Proteome Research*, vol. 6, no. 3, pp. 1190–1197, 2007.

[39] J. Reinders, K. Wagner, R. P. Zahedit et al., "Profiling phosphoproteins of yeast mitochondria reveals a role of phosphorylation in assembly of the ATP synthase," *Molecular & Cellular Proteomics*, vol. 6, no. 11, pp. 1896–1906, 2007.

[40] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Research*, vol. 32, no. 5, pp. 1792–1797, 2004.

[41] Z. H. Yang, "PAML 4: phylogenetic analysis by maximum likelihood," *Molecular Biology and Evolution*, vol. 24, no. 8, pp. 1586–1591, 2007.

[42] C. R. Landry, E. D. Levy, and S. W. Michnick, "Weak functional constraints on phosphoproteomes," *Trends in Genetics*, vol. 25, no. 5, pp. 193–197, 2009.

[43] J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton, and D. T. Jones, "Prediction and functional analysis of native disorder in proteins from the three kingdoms of life," *Journal of Molecular Biology*, vol. 337, no. 3, pp. 635–645, 2004.

[44] *The R Project for Statistical Computing*, http://www.r-project.org.

[45] P. Cohen, "The regulation of protein function by multisite phosphorylation—a 25 year update," *Trends in Biochemical Sciences*, vol. 25, no. 12, pp. 596–601, 2000.

[46] X. Gu, Z. Zhang, and W. Huang, "Rapid evolution of expression and regulatory divergences after yeast gene duplication," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 3, pp. 707–712, 2005.

[47] J. A. Ubersax and J. E. Ferrell Jr., "Mechanisms of specificity in protein phosphorylation," *Nature Reviews*, vol. 8, no. 7, pp. 530–541, 2007.

[48] R. Gordon, "Evolution escapes rugged fitness landscapes by gene or genome doubling: the blessing of higher dimensionality," *Computers & Chemistry*, vol. 18, no. 3, pp. 325–331, 1994.

[49] T. F. Hansen, A. J. R. Carter, and C. H. Chiu, "Gene conversion may aid adaptive peak shifts," *Journal of Theoretical Biology*, vol. 207, no. 4, pp. 495–511, 2000.

[50] D. R. Scannell and K. H. Wolfe, "A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast," *Genome Research*, vol. 18, no. 1, pp. 137–147, 2008.

Hindawi Publishing Corporation
International Journal of Evolutionary Biology
Volume 2012, Article ID 490894, 8 pages
doi:10.1155/2012/490894

Review Article

The Evolution of Novelty in Conserved Gene Families

Gabriel V. Markov and Ralf J. Sommer

Department for Evolutionary Biology, Max Planck Institute for Developmental Biology, Spemannstraße 37, 72076 Tübingen, Germany

Correspondence should be addressed to Ralf J. Sommer, ralf.sommer@tuebingen.mpg.de

Received 16 March 2012; Accepted 23 April 2012

Academic Editor: Frédéric Brunet

One of the major aims of contemporary evolutionary biology is the understanding of the current pattern of biological diversity. This involves, first, the description of character distribution at various nodes of the phylogenetic tree of life and, second, the functional explanation of such changes. The analysis of character distribution is a powerful tool at both the morphological and molecular levels. Recent high-throughput sequencing approaches provide new opportunities to study the genetic architecture of organisms at the genome-wide level. In eukaryotes, one overarching finding is the absence of simple correlations of gene count and biological complexity. Instead, the domain architecture of proteins is becoming a central focus for large-scale evolutionary innovations. Here, we review examples of the evolution of novelty in conserved gene families in insects and nematodes. We highlight how in the absence of whole-genome duplications molecular novelty can arise, how members of gene families have diversified at distinct mechanistic levels, and how gene expression can be maintained in the context of multiple innovations in regulatory mechanisms.

1. Introduction

To understand evolutionary novelty and its contribution to the generation of new species, biologists search for character differences between closely related species and try to determine the functional meaning of such changes. Characters range from morphological traits, like the trichome pattern on the cuticle of a fruit fly larva, to molecular characters such as nucleotide sequences. The genomics era is now providing an increasing number and also new kinds of molecular characters, such as gene numbers in multigenic families, gene position on chromosomes, microRNAs, or insertions of mobile elements at various places in a genome. In addition, next-generation sequencing approaches provide genome-wide single-nucleotide polymorphism (SNP) and copy number variation (CNV) data in a number of model organisms [1]. Nonetheless, it remains challenging to articulate knowledge concerning all these characters in a comprehensive functional framework. While there are some cases with a direct link between various levels of character changes, such as the small number of single-nucleotide mutations in a transcriptional enhancer that can modify the trichome pattern on the larval cuticle in *Drosophila sechellia* [2], there are many other cases where great variation at the molecular level has no simply interpretable effect at the level of the organism. Although this does not mean that such changes are necessarily neutral, there is a strong tendency to correlate any unexpected genomic finding, such as genome duplication or peculiar gene family expansion, with an adaptation to special environmental conditions, often without proper justification [3]. These attempts reflect an understandable quest for generalization. It was recently stressed that evolutionary relevant mutations are not distributed at random in the genome, and that a precise understanding of the contribution of genetic factors to evolution requires the consideration of the specific functional properties of genes [4]. One aspect that is often forgotten in these discussions is that cross-genome comparisons between species are mostly challenged by the inherent difficulty to infer homology between deeply rooted species [3–5].

Here, we review some cases where spectacular molecular changes do not correlate with any clear phenotypic novelty at

the organismal level, and we highlight the need to cope with different types of variation to understand their reciprocal interactions. We will review three examples dealing with (i) novelty by genome diversification in the absence of whole-genome duplication, (ii) novelty in large gene families, and (iii) novelty in promoter regions. For practical reasons, we focus our example choices on ecdysozoans, an animal group that contains two of the best genetic models, the fruit fly *Drosophila melanogaster* and the nematode worm *Caenorhabditis elegans*, each of which is complemented by satellite organisms, allowing us to make sophisticated comparisons by functional investigations [6].

2. Genome Diversification in the Absence of Whole-Genome Duplication

The spectacular examples of land plants and vertebrates highlight the importance of genome duplications for evolutionary success measured in a number of ways, such as number of species, morphological innovations, and ecological diversification [7]. In the animals, however, two other phyla outcompete the vertebrates in all these characteristics. Insects and nematodes are the largest animal phyla with respect to species number as well as morphological and ecological diversification [8, 9]. It is often forgotten that they managed to reach the highest levels of species diversity among animals without the involvement of genome duplication. At the same time, it has to be stressed that the absence of genome duplications in insects and nematodes does not mean that these two groups are lacking noticeable genomic innovations. For example, genome-wide comparisons of aminoacid substitution patterns lead to the estimate that the 39-million-year time interval between the separation of dipterans and coleopterans and the split of the two main dipteran lineages was characterized by an episodic threefold increase in evolutionary rate relative to the mean rate found for the coleopteran representative *Tribolium castaneum* [10]. It was then established that lepidopterans have branches of similar length than dipterans, whereas other holometabolous insects have shorter branches, indicating substitution rates comparable to those of coleopterans [11, 12]. Both dipterans and lepidopterans are, along with a three less diverse orders, members of an insect clade called "Mecopterida" (Table 1), and members of these three orders also experienced a strong acceleration of aminoacid substitution rate at least for the ecdysone receptor gene, a major regulator of molting [13]. Taken together, these data suggest that the acceleration of evolutionary rate took place at the stem of Mecopterida. Interestingly, the interaction between the two proteins that make the ecdysone receptor, USP and EcR, was conserved in spite of the acceleration, because it was compensated by the acquisition of a new dimerization surface that stabilized both partners [14]. The fact that the name "Mecopterida" is almost unknown outside entomology circles illustrates the absence of correlation between this strong molecular divergence and any major phenotypic change. In contrast, the two other species-rich insect orders, hymenopterans and coleopterans, have not experienced such a genome-wide acceleration, indicating that the understanding of this process will require more than a simplistic adaptationist scenario.

Table 1: Decoupling between species number and genome acceleration rates in holometabolous insects.

Order	Approximate number of described species
Diptera	150 000
Mecoptera	600
Siphonaptera	1750
Trichoptera	7000
Lepidoptera	120 000
Total mecopterida	279 350 species
Coleoptera	350 000
Strepsiptera	600
Hymenoptera	115 000
Neuroptera	6000
Raphidioptera	210
Megaloptera	300
Total nonmecopterida	472 110 species

In nematodes, similar findings can be made. For example, in the major nematode model species *C. elegans,* one of the most salient genomic features is the presence of some protein families with high numbers of duplications or coding genes. This is the case for the hedgehog-related sterol-binding secreted signaling proteins [15], which are mainly expressed in cuticular cells [16]. It is also the case for the guanylyl cyclases [17], tyrosine kinases [18], seven-transmembrane receptors [19], and nuclear receptors [20], as well as a number of other families that have not been studied specifically in nematodes. The analysis of members of multigene families that duplicated early in metazoan evolution or even before requires detailed phylogenetic investigations of each of these families, which is not always available. Therefore, it is still impossible to provide a comprehensive overview of genome diversification based on single-gene duplications. Moreover, the precise functional meaning of such amplifications remains quite obscure, even if these expansions show readily discernible patterns.

The functional gene categories most prone to lineage-specific expansions in eukaryotes seem to be structural proteins. This involves enzymes functioning in response to pathogens and environmental stress and includes various components of signaling pathways responsible for specificity, such as ubiquitin ligase subunits and transcription factors [21]. While the duplication pattern of nematode genes is roughly consistent with this notion, lineage-specific variations also exist, especially concerning the spatial distribution of duplicates in the genome. In *C. elegans*, for example, the number of duplicates varies greatly depending on the chromosomal location. The highest concentration of duplicates is found on chromosome V, which reflects tandem amplification in a specifically dynamic chromosomal context and indicates that purely structural factors can also drive the pattern of gene duplication pattern [22]. At least in *C. elegans* and its close relative *C. briggsae,* there is a strong

difference with regard to the position along the chromosome, as duplicated genes are more abundant in the chromosomal arms than in the centromeric part [23–25].

Apart from gene duplications, a major unexpected outcome of nematode genome sequencing is the importance of horizontal gene transfer (HGT), a process that was previously thought to be rare among eukaryotes with sexual reproduction. It turns out that many genes encoding for plant cell-wall modifying proteins were acquired in some nematode lineages many times independently and from various donor organisms [26–28]. Recipients of HGT-acquired cell-wall modifying proteins were plant-parasitic nematodes of the genera *Bursaphelenchus*, *Meloidogyne, Heterodera, Globodera, Pratylenchus,* and *Xiphinema* (for review, see [29]). Additionally, some nonplant parasitic nematodes such as the necromenic species *Pristionchus pacificus* have obtained cellulases from protist-type donors [30].

A major bottleneck for better understanding gene duplications and other processes such as HGT in their short-term and long-term evolutionary consequences is the lack of precise functional knowledge about the majority of these paralogous genes. This includes the well-studied genetic model system *C. elegans*. Compounded with the absence of functional genetic data for many paralogous genes is that little is known about the population structure and polymorphism rates in *C. elegans*. For example, when positive selection was detected among the *srz* family (the Z family of the serpentine receptor superfamily), where no protein had a precisely known physiological function, there were no additional data that would help to interpret the meaning of this observation [31]. This represents an important challenge for future studies because pure computational detection of candidate gene for positive selection is not sufficient to ascertain that an evolutionary event has a real functional meaning. Indeed, the frequency of adaptative substitutions can sometimes be overestimated due to the interplay of other processes that also influence the frequency of nucleotide substitutions, such as genetic hitchhiking or epistatic effects between nonindependent sites, processes of which vary greatly among lineages [32].

Biases in codon composition that are taken as molecular signatures for positive selection can also be produced by a specific bias in DNA turnover at that particular part of the genome. Such a mechanism was already suggested 30 years ago under the concept "molecular drive" [33] and got further support by whole-genome studies on the distribution of sites that are predicted to be under positive selection [34, 35]. Advances in population-genetic theory showed the emergence of certain kinds of aminoacid substitutions and protein-protein complexes restricted to taxa with relatively small effective population sizes [36]. Besides this structure-driven effect of gene family amplification, one should also take into account the possibility that the structure of some signaling networks necessitates the retention of duplicates, similar to what occurs in Mecopterida, where many members of the same ecdysone-signaling network have undergone a supplemental acceleration in nucleotide turnover when compared to the rest of the genome [37]. Furthermore, it should be noted that following an original gene duplication, "new" functions that arise subsequently are not really new but represent cases of subfunctionalization and cooption [38]. In addition, it has recently been proposed that there is no definitive proof that orthologous genes are functionally more similar than closely related paralogs [39]. Comparisons of aminoacid substitution patterns between the speciation event that separated insects and chordates and the duplication event at the basis of vertebrate show similar trends, suggesting that speciation is as important as duplication to promote novelty in gene families [40].

Taken together, genome-wide analyses of species-rich groups of insects and nematodes have identified mechanisms by which genomes can diversify to create novelty in the absence of complete genome duplications. Although both groups show an extraordinary level of genome data and have been studied by the scientific community for more than a century, the functional understanding of genes is often limited. Given space restrictions, we are unable to discuss fully the many hypotheses that have been proposed in association with the limitations of our current understanding, and so we refer the reader to cited literature for more in depth discussions.

3. A Case Study: Molecular Novelty in a Conserved Gene Family

Gene families have been identified as a major target for the generation of molecular novelty in eukaryotes. One of the gene families with a spectacular duplication rate is the nuclear receptor family. In nematodes, for example, the majority of duplicates arose from the amplification of a single member of the family, named HNF4 (NR2A), up to more than 250 duplicates in *C. elegans* [41]. In general, nuclear receptors are currently defined as ligand-activated transcription factors that can undergo a conformational change in response to the binding of a small molecule [42]. Some studies based on ancestral sequence reconstruction and *in vitro* analysis document up to the level of individual mutations how innovations in ligand-binding ability have arisen [43, 44]. This family is one of the most stable at the metazoan level, its members showing a conserved modular structure comprising a DNA-binding domain and a ligand-binding domain that are well characterized at the structural level [45]. However, even among nuclear receptors, there are some atypical members that can have important functional roles in spite of an altered functional structure (Figure 1). For example, the vertebrate DAX-1 (NR0B) is a receptor that has no DNA-binding domain. It is involved in X-linked adrenal hypoplasia congenita, a developmental disorder of the human adrenal gland, where it acts as a dominant-negative receptor that blocks the activation ability of other nuclear receptors [46]. In *Drosophila*, a similar situation appears during the molting cycle. One of the receptors involved in the regulation of molting, E75 (NR1D), has many isoforms that are expressed at various stages in the molting cycle. The isoform E75B lacks half of its DNA-binding domain, having only one zinc finger instead of two for a canonical DNA-binding domain. Being itself unable to bind DNA, it acts as a transcriptional repressor by blocking

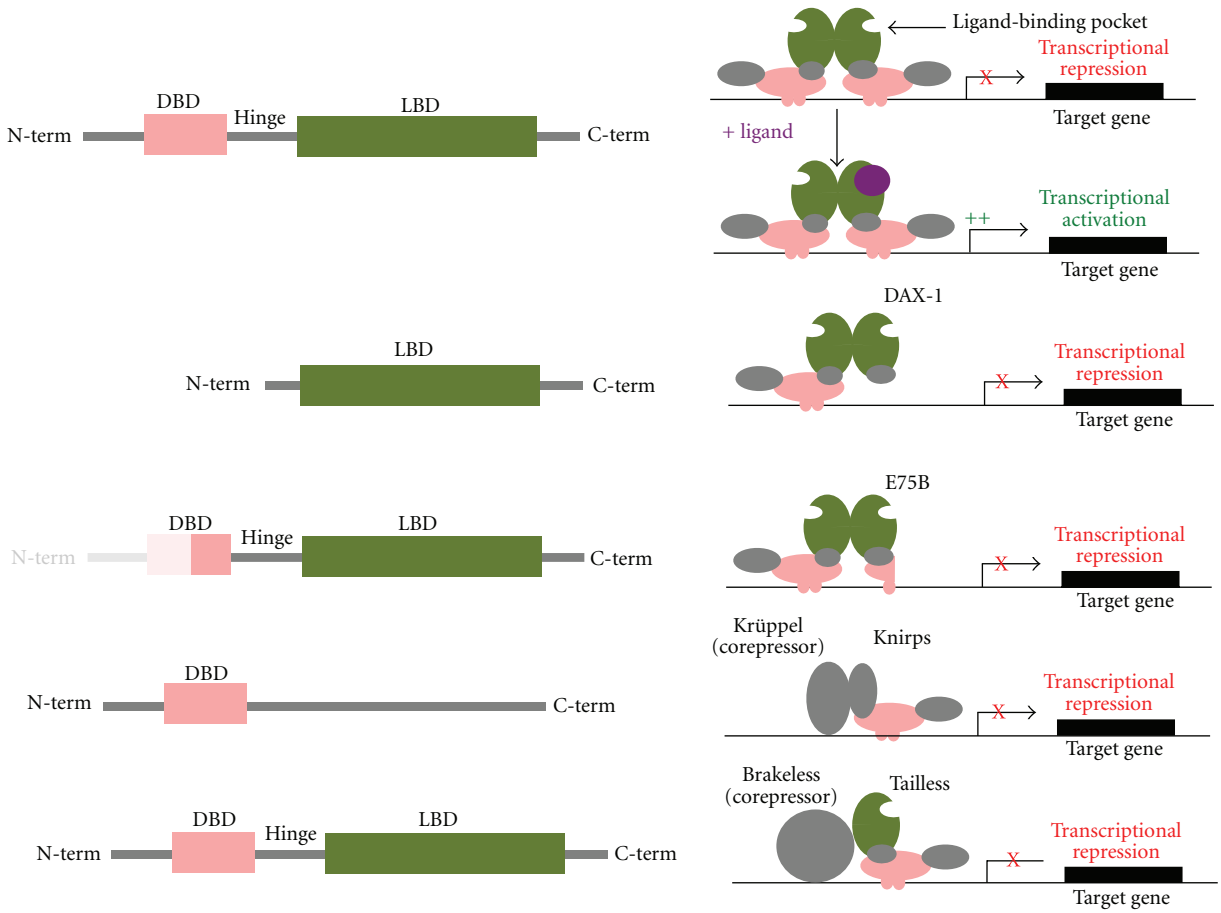


Figure 1: Various levels of functional diversity in a very conserved protein family. The first line shows the canonical structure of a nuclear receptor, comprising a DNA-binding domain (DBD) and a ligand-binding domain, that are structurally well conserved. A canonical receptor represses transcription in absence of a ligand (or it is even not in the nucleus) and activates transcription upon ligand binding. The second line shows a receptor that has lost its DNA-binding domain and that acts also as a transcriptional repressor. The third line shows a receptor that is complete at the gene level, but for which the expression of one isoform starts only at the half of the DNA-binding domain. It acts also as a transcriptional repressor. The fourth line shows a receptor having lost its ligand-binding domain. The last line shows an example of receptor that still has this canonical structure, but that has no known ligand and acts also as a constitutive transcriptional repressor. Whereas knirps and tailless bind to corepressors that are not nuclear receptors, DAX-1 and E75B act as dominant negatives, blocking the activation activity of another nuclear receptor with a canonical structure.

the transactivation abilities of its dimerization partner [47]. This example nicely demonstrates that even a single gene can give rise to proteins with different and even antagonistic functions, due to the variability generated by alternative splicing.

Other members of the nuclear receptor family are devoid of a ligand-binding domain. This is the case for the developmental control gene *knirps* (NR0A1), a *Drosophila* segmentation gene whose expression in the posterior part of the fly embryo is responsible for the presence of abdominal segments 1–7 [48]. It is also involved in head morphogenesis and in tracheal formation later in development. Interestingly, some of its functions in late development are redundant with those of its close paralog *knrl* [49], but the greater intron size of *knrl* relative to *knirps* prevents it from functional complementation duringsegmentation, where transcription time during short mitotic cycles provides a physiological barrier to transcript size [50]. While both genes arose from a duplication event in the cyclorrhaphan diptera, their nonduplicated ortholog in *Tribolium castaneum* also plays a role that is essential for head patterning [51].

Such a patterning role of nuclear receptors during insect segmentation is not restricted to receptors that have lost their ligand-binding domain but can also be observed in the orphan receptor encoded by the *tailless* gene (NR2E2). This protein has a recognizable ligand-binding domain but no known ligand. In *Drosophila*, this transcription factor belongs to the segmentation genes like *knirps*, and it functions in the segmentation gene hierarchy by providing an early subdivision into groups of segments of the embryo by acting as a transcriptional repressor [52]. This function seems to be conserved among holometabolous insects [53].

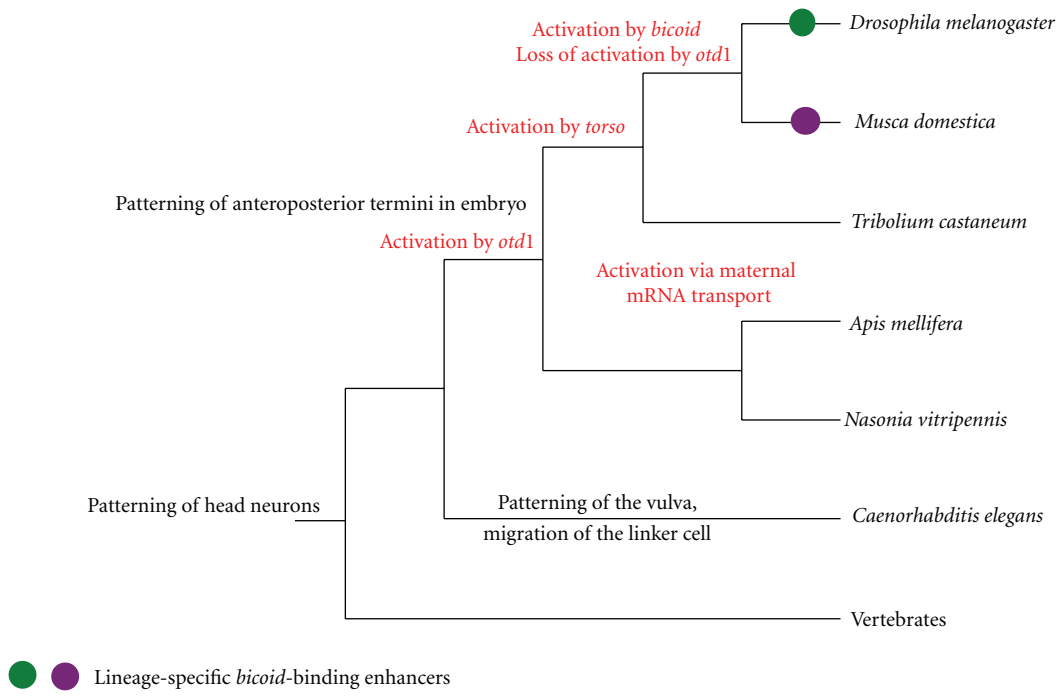


Figure 2: Examples of functional shifts at the level of a single protein. The transcriptional repressor *tailless* is considered to have a conserved function in bilaterians concerning the patterning of anterior neurons. But in addition to that, it has secondarily acquired a number of lineage-specific functions. In *Caenorhabditis elegans*, it contributes to the patterning of the vulva in hermaphrodites and in migration of the linker cell from the male gonad. In holometabolous insects, it participates in the patterning of the anterior and posterior tips of the embryo. Strikingly, even if the expression domain of *tailless* is conserved in holometabolous embryos, this is achieved through highly variable transcriptional pathways (in red on the figure).

However, the most conserved part of *tailless* function seems to be its role in the specification of the anterior nervous system (Figure 2). In *Drosophila*, *tailless* controls the formation of the protocerebral neuroblasts and acts in eye formation [54]. In mammals, it is involved in brain and visual system development, as well as in neural stem-cell renewal at the adult stage [55]. In *C. elegans*, the *tailless* ortholog, named *nhr-67*, is expressed in six head neurons, but is also involved in other processes, such as the patterning of the vulva [56] and the control of the migration of the linker cell in the male gonad [57].

All these examples from nuclear receptors that do not act as ligand-activated transcription factors show that even in a family that is very well conserved at the structural level, there can be many functional variations and novelty. These unusual cases are involved in the control of important biological processes, providing a powerful first glance at the complexity of eukaryotic genomes. One should be aware, however, that such studies can lead only to partial conclusions, which need to be completed by more precise functional investigations. The genes described above, like any other gene in eukaryotic genomes, might have acquired novel but simple protein domains, that are not easily detectable by bioinformatic means. For example, the origin of four amino acid SH3-binding domains in an otherwise conserved LIN-18/Ryk/derailed receptor in WNT signaling has allowed new wiring in the signaling pathway leading to vulva formation in the nematode *Pristionchus pacificus* but which is not present in *C. elegans* [58]. Such domains would go unnoticed without functional studies by unbiased genetic approaches.

4. *cis*-Regulatory Novelties in the Promoter of a Gene with Conserved Expression Pattern

In the recent years, there has been an ongoing debate about the contribution of *cis*-regulatory elements versus protein-coding regions in evolutionary innovation [59, 60]. The arguments are extensively reviewed in detail elsewhere [61], so we will concentrate on the complementary side of the problem, the fact that high promoter turnover and changes in transcriptional regulation are compatible with a conserved gene expression pattern.

The evolution of the promoter of the *tailless* gene is especially well studied in holometabolous insects (Figure 2). The expression of the *tailless* gene in *Drosophila melanogaster* is regulated by a complex set of transcription factors, the most important being *bicoid* [62] and two other genes that are in the downstream *torso* signaling pathway [63]. The promoter of another dipteran fly, *Musca domestica*, contains binding sites for all these transcription factors in similar numbers, although the binding sites are organized in a different order. The expression of *tailless* is highly similar in the two flies at the blastodermal stage [64], the only subtle difference being the split of the expression pattern in

the anterior cap from the *Musca* embryo, which does not occur in *Drosophila*. Additionally, the promoter of *Musca* is able to drive a *Drosophila*-like expression pattern of a reporter gene when inserted in *Drosophila* embryos [65]. These observations completed by estimations about the mutation rates in *Drosophila* gene promoters suggest that the promoter region was fully renewed during the 100 million years following the divergence of *Drosophila* and *Musca* [65]. It has therefore been argued that the regulation by the same transcription factors was maintained by constant loss and *de novo* acquisitions of similar promoter elements [65].

In the flour beetle *Tribolium castaneum*, the expression pattern of *tailless* during embryogenesis is similar to that in flies, and the activation by the *torso* pathway is conserved as well. In contrast, transcriptional control by *bicoid* is not possible, because *bicoid* is specific to flies [66]. In contrast to flies and beetles, hymenopterans represent again a different case. In the parasitic wasp *Nasonia vitripennis*, in which components of the *torso* pathway are missing, the expression of *tailless* is activated by *orthodenticle-1* [67]. In the honeybee *Apis mellifera*, anterior expression of *tailless* also depends on *orthodenticle-1*, but its posterior expression is due to maternal RNA [67]. A contribution of *orthodenticle-1* to *tailless* expression also in *T. castaneum* is likely, given the fact that *orthodenticle-1* is a proposed substitute for *bicoid* in this insect [68], but direct binding from *orthodenticle-1* to the *tailless* promoter is yet to be reported. In spite of this remaining question, the comparison of the regulatory mechanisms for the *tailless* expression shows that there are already four slightly different ways that are known to maintain this pattern in holometabolous insects.

The *tailless* case is particularly well documented in terms of species sampling. Yet there are many other examples of high promoter turnover in genes whose expression is conserved (reviewed in [69]). This illustrates the notion of developmental system drift, describing the fact that many changes in developmental pathways occur during evolution without phenotypic effect and thus are more likely to be the result of contingent historical events than the response to selection pressure [70].

5. Conclusion

In his autobiography, Darwin [71] wrote the following about the reasons that pushed him to write two extensive monographies about cirripedes: "*When on the coast of Chile, I found a most curious form, which burrowed into the shells of Concholepas, and which differed so much from all other Cirripedes that I had to form a new sub-order for its sole reception. [...] To understand the structure of my new Cirripede, I had to examine and dissect many of the common forms: and this gradually led me on to take up the whole group.*" This illustrates perfectly well what lies on the agenda of today's evolutionary biologists. What has changed since Darwin's time is that now we have the tools required to describe natural variation from the molecular level to the ecological one. It follows that, for a given node of a phylogenetic tree, variation and repartition of characters can and need to be addressed at all these levels. What we have to understand is the connection between the various layers of biological complexity, combining and integrating the results of laboratory and fieldwork approaches [6].

We argue that the partial data on genomic variation are already sufficient to indicate that no specific attribute of a given molecular structure can indicate *a priori* more potentials than others to contribute to novelty at a higher phenotypic level. Uncoupling and buffering of natural variation at various integration scales has been clearly demonstrated, implying that the number of molecular events that can be directly correlated with a phenotypic change at the organismal level is probably very low, and that they are the results of exceptional contingency and structural constraints [72]. The possibility to detect *de novo* such interesting changes in nonmodel organisms is thus also probably very low, and it decreases quickly with an increase of the phylogenetic distance. Additionally, one should not forget that to really understand the link between genetic variation and phenotypic diversity, it is necessary to be able to explain cases where a molecular change triggers novelty, and those where phenotypic traits are maintained in spite of molecular innovations in the genes that specify them.

Acknowledgments

The authors thank members of the Sommer Lab for discussions and Dr. Erik J. Ragsdale for carefully reading the paper. This work was supported by funding from the Max Planck Society.

References

[1] M. Nordborg and D. Weigel, "Next-generation genetics in plants," *Nature*, vol. 456, no. 7223, pp. 720–723, 2008.

[2] N. S. Frankel, D. F. Erezyilmaz, A. P. McGregor, S. Wang, F. Payre, and D. L. Stern, "Morphological evolution caused by many subtle-effect substitutions in regulatory DNA," *Nature*, vol. 474, no. 7353, pp. 598–603, 2011.

[3] M. Lynch, "The frailty of adaptive hypotheses for the origins of organismal complexity," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 1, pp. 8597–8604, 2007.

[4] D. L. Stern and V. Orgogozo, "Is genetic evolution predictable?" *Science*, vol. 323, no. 5915, pp. 746–751, 2009.

[5] R. J. Sommer, "Homology and the hierarchy of biological systems," *BioEssays*, vol. 30, no. 7, pp. 653–658, 2008.

[6] R. J. Sommer, "The future of evo-devo: model systems and evolutionary theory," *Nature Reviews Genetics*, vol. 10, no. 6, pp. 416–422, 2009.

[7] Y. Van De Peer, S. Maere, and A. Meyer, "The evolutionary significance of ancient genome duplications," *Nature Reviews Genetics*, vol. 10, no. 10, pp. 725–732, 2009.

[8] D. Grimaldi and M. S. Engel, *Evolution of the Insects*, Cambridge University Press, Cambridge, UK, 2005.

[9] D. L. Lee, Ed., *The Biology of Nematodes*, Taylor & Francis, London, UK, 2002.

[10] J. Savard, D. Tautz, and M. J. Lercher, "Genome-wide acceleration of protein evolution in flies (Diptera)," *BMC Evolutionary Biology*, vol. 6, article 7, 2006.

[11] J. Savard, D. Tautz, S. Richards et al., "Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of Holometabolous insects," *Genome Research*, vol. 16, no. 11, pp. 1334–1338, 2006.

[12] E. M. Zdobnov and P. Bork, "Quantification of insect genome divergence," *Trends in Genetics*, vol. 23, no. 1, pp. 16–20, 2007.

[13] F. Bonneton, F. G. Brunet, J. Kathirithamby, and V. Laudet, "The rapid divergence of the ecdysone receptor is a synapomorphy for Mecopterida that clarifies the Strepsiptera problem," *Insect Molecular Biology*, vol. 15, no. 3, pp. 351–362, 2006.

[14] T. Iwema, A. Chaumot, R. A. Studer et al., "Structural and evolutionary innovation of the heterodimerization interface between USP and the ecdysone receptor ECR in insects," *Molecular Biology and Evolution*, vol. 26, no. 4, pp. 753–768, 2009.

[15] G. Aspöck, H. Kagoshima, G. Niklaus, and T. R. Bürglin, "Caenorhabditis *elegans* has scores of hedgehog-related genes: sequence and expression analysis," *Genome Research*, vol. 9, no. 10, pp. 909–923, 1999.

[16] L. Hao, R. Johnsen, G. Lauter, D. Baillie, and T. R. Bürglin, "Comprehensive analysis of gene expression patterns of hedgehog-related genes," *BMC Genomics*, vol. 7, article 280, 2006.

[17] D. A. Fitzpatrick, D. M. O'Halloran, and A. M. Burnell, "Multiple lineage specific expansions within the guanylyl cyclase gene family," *BMC Evolutionary Biology*, vol. 6, article 26, p. 18, 2006.

[18] C. Popovici, R. Roubin, F. Coulier, P. Pontarotti, and D. Birnbaum, "The family of *Caenorhabditis elegans* tyrosine kinase receptors: similarities and differences with mammalian receptors," *Genome Research*, vol. 9, no. 11, pp. 1026–1039, 1999.

[19] H. M. Robertson and J. H. Thomas, "The putative chemoreceptor families of *C. elegans*," in *WormBook*, The C. elegans Research Community, Ed., 2006.

[20] A. E. Sluder, S. W. Mathews, D. Hough, V. P. Yin, and C. V. Maina, "The nuclear receptor superfamily has undergone extensive proliferation and diversification in nematodes," *Genome Research*, vol. 9, no. 2, pp. 103–120, 1999.

[21] O. Lespinet, Y. I. Wolf, E. V. Koonin, and L. Aravind, "The role of lineage-specific gene family expansion in the evolution of eukaryotes," *Genome Research*, vol. 12, no. 7, pp. 1048–1059, 2002.

[22] A. R. O. Cavalcanti, R. Ferreira, Z. Gu, and W. H. Li, "Patterns of gene duplication in *Saccharomyces cerevisiae* and Caenorhabditis *elegans*," *Journal of Molecular Evolution*, vol. 56, no. 1, pp. 28–37, 2003.

[23] C. Elegans Sequencing Consortium, "Genome sequence of the nematode *C. elegans*: a platform for investigating biology," *Science*, vol. 282, no. 5396, pp. 2012–2018, 1998.

[24] H. M. Robertson, "Updating the *str* and *srj* (*stl*) families of chemoreceptors in *Caenorhabditis* nematodes reveals frequent gene movement within and between chromosomes," *Chemical Senses*, vol. 26, no. 2, pp. 151–159, 2001.

[25] L. D. Stein, Z. Bao, D. Blasiar et al., "The genome sequence of *Caenorhabditis* briggsae: a platform for comparative genomics," *PLoS Biology*, vol. 1, no. 2, article E45, 2003.

[26] G. Smant, J. P. W. G. Stokkermans, Y. Yan et al., "Endogenous cellulases in animals: isolation of β-1,4-endoglucanase genes from two species of plant-parasitic cyst nematodes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 9, pp. 4906–4911, 1998.

[27] T. Kikuchi, J. T. Jones, T. Aikawa, H. Kosaka, and N. Ogura, "A family of glycosyl hydrolase family 45 cellulases from the pine wood nematode *Bursaphelenchus xylophilus*," *FEBS Letters*, vol. 572, no. 1–3, pp. 201–205, 2004.

[28] E. G. J. Danchin, M. N. Rosso, P. Vieira et al., "Multiple lateral gene transfers and duplications have promoted plant parasitism ability in nematodes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 41, pp. 17651–17656, 2010.

[29] R. J. Sommer and A. Streit, "Comparative genetics and genomics of nematodes: genome structure, development, and lifestyle," *Annual Reviews of Genetics*, vol. 45, pp. 1–20, 2011.

[30] C. Dieterich, S. W. Clifton, L. N. Schuster et al., "The *Pristionchus pacificus* genome provides a unique perspective on nematode lifestyle and parasitism," *Nature Genetics*, vol. 40, no. 10, pp. 1193–1198, 2008.

[31] J. H. Thomas, J. L. Kelly, H. M. Robertson, K. Ly, and W. J. Swanson, "Adaptive evolution in the SRZ chemoreceptor families of *Caenorhabditis elegans* and *Caenorhabditis briggsae*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 12, pp. 4476–4481, 2005.

[32] J. C. Fay, "Weighing the evidence for adaptation at the molecular level," *Trends in Genetics*, vol. 27, no. 9, pp. 343–349, 2011.

[33] G. Dover, "Molecular drive: a cohesive mode of species evolution," *Nature*, vol. 299, no. 5879, pp. 111–117, 1982.

[34] G. Marais, D. Mouchiroud, and L. Duret, "Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 10, pp. 5688–5692, 2001.

[35] N. Galtier and L. Duret, "Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution," *Trends in Genetics*, vol. 23, no. 6, pp. 273–277, 2007.

[36] M. Lynch, L.-M. Bobay, F. Catania, J.-F. Gout, and M. Rho, "The repatterning of eukaryotic genomes by random genetic drift," *Annual Reviews of Genomics and Human Genetics*, vol. 12, pp. 347–366, 2011.

[37] F. Bonneton, A. Chaumot, and V. Laudet, "Annotation of *Tribolium* nuclear receptors reveals an increase in evolutionary rate of a network controlling the ecdysone cascade," *Insect Biochemistry and Molecular Biology*, vol. 38, no. 4, pp. 416–429, 2008.

[38] G. C. Conant and K. H. Wolfe, "Turning a hobby into a job: how duplicated genes find new functions," *Nature Reviews Genetics*, vol. 9, no. 12, pp. 938–950, 2008.

[39] R. A. Studer and M. Robinson-Rechavi, "How confident can we be that orthologs are similar, but paralogs differ?" *Trends in Genetics*, vol. 25, no. 5, pp. 210–216, 2009.

[40] R. A. Studer and M. Robinson-Rechavi, "Large-scale analysis of orthologs and paralogs under covarion-like and constant-but-different models of amino acid evolution," *Molecular Biology and Evolution*, vol. 27, no. 11, pp. 2618–2627, 2010.

[41] M. Robinson-Rechavi, C. V. Maina, C. R. Gissendanner, V. Laudet, and A. Sluder, "Explosive lineage-specific expansion of the orphan nuclear receptor HNF4 in nematodes," *Journal of Molecular Evolution*, vol. 60, no. 5, pp. 577–586, 2005.

[42] H. Gronemeyer, J. Å. Gustafsson, and V. Laudet, "Principles for modulation of the nuclear receptor superfamily," *Nature Reviews Drug Discovery*, vol. 3, no. 11, pp. 950–964, 2004.

[43] J. T. Bridgham, S. M. Carroll, and J. W. Thornton, "Evolution of hormone-receptor complexity by molecular exploitation," *Science*, vol. 312, no. 5770, pp. 97–101, 2006.

[44] S. M. Carroll, E. A. Ortlund, and J. W. Thornton, "Mechanisms for the evolution of a derived function in the ancestral glucocorticoid receptor," *PLoS Genetics*, vol. 7, no. 6, Article ID e1002117, 2011.

[45] P. Huang, V. Chandra, and F. Rastinejad, "Structural overview of the nuclear receptor superfamily: insights into physiology and therapeutics," *Annual Review of Physiology*, vol. 72, pp. 247–272, 2009.

[46] E. Zanaria, F. Muscatelli, B. Bardoni et al., "An unusual member of the nuclear hormone receptor superfamily responsible for X-linked adrenal hypoplasia congenita," *Nature*, vol. 372, no. 6507, pp. 635–641, 1994.

[47] C. S. Thummel, "Dueling orphans—interacting nuclear receptors coordinate *Drosophila* metamorphosis," *BioEssays*, vol. 19, no. 8, pp. 669–672, 1997.

[48] U. Nauber, M. J. Pankratz, A. Kienlin, E. Seifert, U. Klemm, and H. Jackle, "Abdominal segmentation of the *Drosophila* embryo requires a hormone receptor-like protein encoded by the gap gene knirps," *Nature*, vol. 336, no. 6198, pp. 489–492, 1988.

[49] M. Gonzalez-Gaitan, M. Rothe, E. A. Wimmer, H. Taubert, and H. Jackle, "Redundant functions of the genes knirps and knirps-related for the establishment of anterior *Drosophila* head structures," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 91, no. 18, pp. 8567–8571, 1994.

[50] M. Rothe, M. Pehl, H. Taubert, and H. Jackle, "Loss of gene function through rapid mitotic cycles in the *Drosophila* embryo," *Nature*, vol. 359, no. 6391, pp. 156–159, 1992.

[51] A. C. Cerny, D. Grossmann, G. Bucher, and M. Klingler, "The Tribolium ortholog of knirps and knirps-related is crucial for head segmentation but plays a minor role during abdominal patterning," *Developmental Biology*, vol. 321, no. 1, pp. 284–294, 2008.

[52] E. Morán and G. Jiménez, "The tailless nuclear receptor acts as a dedicated repressor in the early Drosophila embryo," *Molecular and Cellular Biology*, vol. 26, no. 9, pp. 3446–3454, 2006.

[53] M. J. Wilson and P. K. Dearden, "Tailless patterning functions are conserved in the honeybee even in the absence of *Torso* signaling," *Developmental Biology*, vol. 335, no. 1, pp. 276–287, 2009.

[54] K. M. Rudolph, G. J. Liaw, A. Daniel et al., "Complex regulatory region mediating tailless expression in early embryonic patterning and brain development," *Development*, vol. 124, no. 21, pp. 4297–4308, 1997.

[55] H. Gui, M. L. Li, and C. C. Tsai, "A tale of tailless," *Developmental Neuroscience*, vol. 33, no. 1, pp. 1–13, 2011.

[56] J. S. Fernandes and P. W. Sternberg, "The tailless ortholog nhr-67 regulates patterning of gene expression and morphogenesis in the *C. elegans* vulva," *PLoS Genetics*, vol. 3, no. 4, article e69, 2007.

[57] M. Kato and P. W. Sternberg, "The *C. elegans tailless/Tlx* homolog *nhr-67* regulates a stage-specific program of linker cell migration in male gonadogenesis," *Development*, vol. 136, no. 23, pp. 3907–3915, 2009.

[58] X. Wang and R. J. Sommer, "Antagonism of LIN-17/frizzled and LIN-18/RyK in nematode vulva induction reveals evolutionary alterations in core developmental pathways," *PLoS Biology*, vol. 9, no. 7, Article ID e1001110, 2011.

[59] H. E. Hoekstra and J. A. Coyne, "The locus of evolution: evo devo and the genetics of adaptation," *Evolution*, vol. 61, no. 5, pp. 995–1016, 2007.

[60] S. B. Carroll, "Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution," *Cell*, vol. 134, no. 1, pp. 25–36, 2008.

[61] D. L. Stern and V. Orgogozo, "The loci of evolution: how predictable is genetic evolution?" *Evolution*, vol. 62, no. 9, pp. 2155–2177, 2008.

[62] G. J. Liaw and J. A. Lengyel, "Control of tailless expression by bicoid, dorsal and synergistically interacting terminal system regulatory elements," *Mechanisms of Development*, vol. 40, no. 1-2, pp. 47–61, 1992.

[63] G. J. Liaw, K. M. Rudolph, J. D. Huang, T. Dubnicoff, A. J. Courey, and J. A. Lengyel, "The torso response element binds GAGA and NTF-1/Elf-1, and regulates *tailless* by relief of repression," *Genes and Development*, vol. 9, no. 24, pp. 3163–3176, 1995.

[64] R. Sommer and D. Tautz, "Segmentation gene expression in the housefly *Musca domestica*," *Development*, vol. 113, no. 2, pp. 419–430, 1991.

[65] N. S. Wratten, A. P. McGregor, P. J. Shaw, and G. A. Dover, "Evolutionary and functional analysis of the tailless enhancer in *Musca domestica* and *Drosophila melanogaster*," *Evolution and Development*, vol. 8, no. 1, pp. 6–15, 2006.

[66] M. Schoppmeier and R. Schröder, "Maternal torso signaling controls body axis elongation in a short germ insect," *Current Biology*, vol. 15, no. 23, pp. 2131–2136, 2005.

[67] J. A. Lynch, E. C. Olesnicky, and C. Desplan, "Regulation and function of tailless in the long germ wasp *Nasonia vitripennis*," *Development Genes and Evolution*, vol. 216, no. 7-8, pp. 493–498, 2006.

[68] R. Schröder, "The genes orthodenticle and hunchback substitute for bicoid in the beetle *Tribolium*," *Nature*, vol. 422, no. 6932, pp. 621–625, 2003.

[69] M. T. Weirauch and T. R. Hughes, "Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same," *Trends in Genetics*, vol. 26, no. 2, pp. 66–74, 2010.

[70] J. R. True and E. S. Haag, "Developmental system drift and flexibility in evolutionary trajectories," *Evolution and Development*, vol. 3, no. 2, pp. 109–119, 2001.

[71] C. Darwin, *The Autobiography of Charles Darwin 1809–1882*, With Original Omissions Restored, Edited With Appendix and Notes by Nora Barlow, Harcourt, Brace & Co., New York, NY, USA, 1959.

[72] G. P. Wagner, "The developmental genetics of homology," *Nature Reviews Genetics*, vol. 8, no. 6, pp. 473–479, 2007.

Hindawi Publishing Corporation
International Journal of Evolutionary Biology
Volume 2012, Article ID 278981, 7 pages
doi:10.1155/2012/278981

Review Article

What Can Domesticated Genes Tell Us about the Intron Gain in Mammals?

Dušan Kordiš and Janez Kokošar

Department of Molecular and Biomedical Sciences, Josef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia

Correspondence should be addressed to Dušan Kordiš, dusan.kordis@ijs.si

Received 26 January 2012; Accepted 6 April 2012

Academic Editor: Frédéric Brunet

Domesticated genes, originating from retroelements or from DNA-transposons, constitute an ideal system for testing the hypothesis on the absence of intron gain in mammals. Since single-copy domesticated genes originated from the intronless multicopy transposable elements, the ancestral intron state for domesticated genes is zero. A phylogenomic approach has been used to analyse all domesticated genes in mammals and chordates that originated from the coding parts of transposable elements. A significant amount of intron gain was found only in domesticated genes of placental mammals, where more than 70 cases were identified. *De novo* gained introns show clear positional bias, since they are distributed mainly in 5′ UTR and coding regions, while 3′ UTR introns are very rare. In the coding regions of some domesticated genes up to 8 *de novo* gained introns have been found. Surprisingly, the majority of intron gains have occurred in the ancestor of placental mammals. Domesticated genes could constitute an excellent system on which to analyse the mechanisms of intron gain. This paper summarizes the current understanding of intron gain in mammals.

1. Introduction

Transposable elements (TEs) constitute a major component of eukaryotic genomes [1]. Because they can transpose at high frequency they act as insertional mutagens and are powerful endogenous mutators [2, 3]. The mobility and amplification of TEs constitutes a major source of genomic variation either by virtue of their insertion or by triggering a variety of small- and large-scale chromosomal rearrangements. In consequence, they can have a major impact on the host phenotype [1–5]. Evidence is growing that TEs sometimes contribute positively to the function and evolution of genes and genomes [1–5]. Genome-scale analyses confirmed that domesticated or exapted TE-derived sequences have contributed diverse and abundant regulatory and protein coding sequences to host genomes [5–9].

Domesticated genes [6, 7, 9–11], originating from retroelements or from DNA transposons, constitute an ideal system for testing the hypothesis on the absence of intron gain in mammals. Since single-copy domesticated genes [7] originated from the intronless multicopy TEs [3], the ancestral intron state for domesticated genes is zero. Therefore, any intron present in these genes will constitute a *de novo* gained intron. The prerequisite for recognizing the origin, extent, and timing of *de novo* gained introns is reliable and wide taxon sampling [12]. In the past few years a quite large and dense collection of vertebrate, and especially mammalian, genomes has been accumulated. For some of these taxa a number of well-annotated genomes and genes exist, human and mouse genomes and transcriptomes being especially useful, with the full-length mRNAs that enabled reconstructions of the complete gene structures in these species [13, 14]. By using annotated human or mouse introns, we can trace their origin in mammals through genome-wide comparisons of orthologous genes in placentals, marsupials, and monotremes.

Spliceosomal introns are one of the major eukaryote-specific genome components, and the availability of numerous eukaryotic genomes has enabled genome-wide studies of the intron loss and gain dynamics [15–19]. The large-scale comparisons of the evolutionary dynamics of introns in eukaryotes has revealed a significant excess of losses

and a nonuniform distribution of gains and losses [15–19]. A substantial excess of intron gains has been detected only for those intervals of eukaryotic evolution that are associated with major evolutionary innovations, such as the origin of eukaryotes and animals [15–19]. The large-scale comparisons of the evolutionary dynamics of introns have demonstrated surprising evolutionary stasis in the intron dynamics over the last 100–200 My [15, 16]. Large-scale intron studies in orthologous mammalian genes have indicated that very little intron turnover has occurred, with convincing evidence only for loss of introns [17, 18]. Such absence of intron gain in "recent" evolutionary history might be real, but could also be artifactual, the consequence of inadequate taxon sampling or inadequate comparisons, since only the "old" orthologous genes have been compared. To test the claims on the absence of intron gain in some taxonomic groups such as mammals [17, 18] and in recent evolutionary history (in the last 100–200 Mya) [15, 16], we need a quite simple and robust "gene model" that is independent of the inference procedures about intron gain. If the ancestral intron state is definitely known, the intron gain can be easily recognized. Such an approach, coupled with the known ancestral state (intronless), has been used in Kordis study [19] for evaluation of the hypotheses on the absence of intron gain in the recent evolutionary past [15, 16], and especially in mammals [17, 18].

2. Gene Structures of Domesticated Genes in Chordates and Mammals

Vertebrate, and especially mammalian, genomes contain a number of genes that have originated from TEs or their remains [6, 7, 9–11]. Since vertebrate retroelements and DNA transposons do not contain introns [3], the ancestral state for TE-derived genes is intronless. During the transition process from a multicopy TE to the single-copy domesticated gene, intron gain can occur, meaning that any intron present in the domesticated gene will constitute a *de novo* intron gain. An extensive phylogenomic analysis of all domesticated genes in chordate and mammalian genomes has therefore been made [19]. The rich collection of numerous mammalian genomes, belonging to all three major extant mammalian lineages, Eutheria (placentals), Metatheria (marsupials), and Prototheria (monotremes), was a major advantage in studying the origin and evolution of domesticated genes. By the phylogenomic analysis of all available domesticated genes in mammalian, vertebrate, and chordate genomes, unequivocal data about their origins (when and in which taxonomic group they originated) and numerous gene-related data (exon/intron structure, genome location, chromosomal position, etc.) have been obtained. The most important part of Kordis study has been the finding of transition point where and when TEs were transformed into domesticated genes, allowing *de novo* gain of introns to be precisely pinpointed in these genes. The gene structures of domesticated genes has provided direct evidence for extensive intron gain in placental mammals [19].

3. The Majority of Domesticated Genes Contain *De Novo* Gained Introns

The analysis of all known domesticated genes in chordates and mammals has shown that both retroelement- and DNA transposon-derived genes contain introns [19]. In the case of retroelement-derived genes the exon/intron structures are simple, since in these cases the process of gene fusion or exon shuffling to the preexisting "normal" genes is almost always absent (one such exception is the SCAND3 gene). However, the situation in the case of DNA transposon-derived genes is more complicated, since these genes can originate by three different routes: (a) from the entire DNA transposon, (b) by a complete DNA transposon being fused to the "normal" gene in the form of a single long exon that is 3′ end located, and (c) the most prevalent case, by gene fusion or exon shuffling of DNA binding domains (DBD) of DNA transposons with "normal" genes (where the exonization is necessary before the gene fusion). Therefore, in the case of DNA transposon-derived genes, intron gain can be recognized easily only in the first case, while the second and third cases are much more difficult for inferring intron gain in these genes. In the majority of the cases of fused entire transposases or just the DBDs, the newly recruited exons remain intact as very long or relatively short exons, but they are mostly without any intron. Therefore, in the case of DNA transposon-derived genes, these fused transposases and DBDs have been excluded from the analysis of intron gain. The situation regarding intron gain in retroelement-derived genes is definitely much simpler and less problematic, since no fusion genes have originated from retroelements (except SCAND3) [19].

4. The Burst of Intron Gain in Domesticated Genes Was in the Ancestor of Placental Mammals (Eutheria)

The analysis of all domesticated genes in chordates and mammals has shown that by far the greatest amount of intron gain occurred in the ancestor of placentals [19] (Figure 1). Twenty intron-containing domesticated genes originated in the ancestor of placentals, 18 of them being retroelement-derived and only two DNA transposon-derived genes. Interestingly, a recent study reported that 11 retrogenes with newly gained 5′ UTR introns also originated in the ancestor of placentals [20]. In the case of retrogenes they found 18 intron gains, 17 into the 5′ UTR, and a single gain in the 3′ UTR. In the case of domesticated genes 49 to 57 cases of *de novo* gained introns were found in the ancestor of placentals (Figure 1). In retroelement-derived genes 42 to 50 cases of intron gain have been found, while in DNA transposon-derived genes only 7 cases were found. Collectively, the 20 domesticated genes and 11 retrogenes provide evidence for at least 50 to 70 cases of intron gain in the ancestor of placentals [19]. This finding contrasts strongly with previous studies [17, 18], in which no intron gain could be found in mammals.

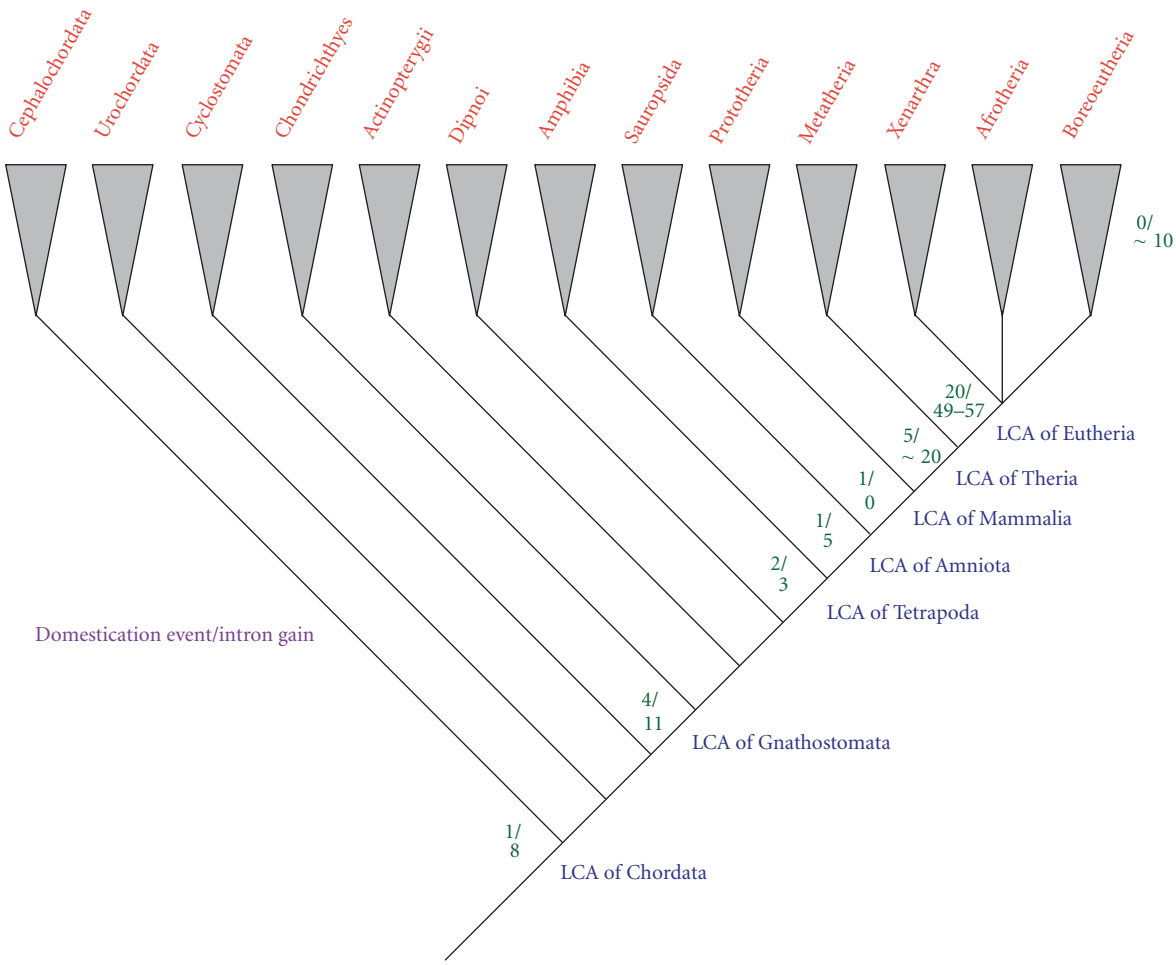


Figure 1: Numbers of transposable element-derived gene domestication events and intron gains mapped on the chordate phylogenetic tree. In the superorder Boreoeutheria some additional intron gains have occurred.

Although these genes represent a very small proportion of placental gene innovations, the observed extent of intron gain most probably represents just the tip of the iceberg. Regardless of the situation with the normal mammalian and vertebrate genes ("old genes"), there was a large-scale gene origination in the ancestor of placentals (evolutionarily young genes), at least in some classes of transcription factors (e.g., in C_2H_2 ZNFs). To test the extent of intron gain in some other placental-specific gene families the presence of intron gain has been analyzed in KRAB and SCAN ZNF genes [21, 22], especially in those orthologous genes that originated in the ancestor of placentals (>150 orthologous genes were analyzed). The analysis has shown that the amount of intron gain in these genes is not as high as in the case of TE-derived domesticated genes and retrogenes, but a number of cases with intron gain can, even so, be recognized [19].

The analysis of placental-specific domesticated genes, retrogenes and placental-specific transcription factors (~200 were analyzed) has shown that numerous intron gains occurred in the ancestor of placentals and that intron gain is still ongoing in mammals. At least 50 to 70 cases of intron gain have been documented from the analysis of >30 domesticated genes and retrogenes, and a few more cases have been documented also for placental-specific transcription factors (KRAB-ZNFs, SCAN-ZNFs and SCAN-KRAB-ZNFs). Up to 100 cases of intron gain have been recognized from the analysis of ~200 orthologous genes. These intron gains have occurred at different time points of placental evolution, the vast majority of them in the ancestor of placentals and the others in diverse lineages or species of placental mammals [19].

5. Numbers of Introns per Gene, Intron Densities, Sizes of Introns, and Preferred Locations of *De Novo* Gained Introns

An extensive phylogenomic analysis of all the domesticated genes in chordate and mammalian genomes has provided crucial information as to where and when TEs were

transformed into domesticated genes, allowing *de novo* gain of introns in these genes to be pinpointed precisely [19]. The number of gained introns in these genes varies greatly, from 1 to 8. Domesticated genes in placentals and chordates accumulated a large number of introns, such that their density in these genes has become close to that in "normal" genes [23]. The average intron density for Eutheria-specific domesticated genes is 4.01 intron per kb of 5′ UTR. Intron density for older domesticated genes that originated early in vertebrates (e.g., Gin-1 and PGBD5) is 4.09 intron per kb of CDS. This comparison indicates that intron densities are similar in domesticated genes, the major difference however being in their position. Intron densities in Eutheria-specific domesticated genes and in older domesticated genes that originated early in vertebrates are therefore lower than those for normal mammalian and vertebrate genes [15, 16]. The sizes of the gained introns in domesticated genes are highly variable, ranging from a few hundred to a few thousand base pairs. DNA transposon-derived genes contain longer introns than retroelement-derived genes, just as evolutionarily older domesticated genes contain much longer introns than evolutionarily younger domesticated genes. Surprisingly, the longest introns exist in the gag-derived ZCCHC16 gene, and the second intron in the mouse (~410 kb long) is also the longest intron in the chordate domesticated genes [19]. This gene resembles mammalian retrogenes with very long introns [20]. The preferred locations of *de novo* gained introns in domesticated genes are the 5′ UTRs and coding regions, while 3′ UTR locations are very rare [19]. These preferred intron locations are similar to those of the "normal" chordate genes [24]. However, the preferred locations of *de novo* gained introns in domesticated genes differ from the mammalian retrogenes, where newly gained introns are preferentially located in the 5′ UTRs [20].

6. Intron Positions and Nucleotide Sequences of *De Novo* Gained Introns Are Highly Conserved in Placental Mammals

The extensive information on intron position conservation collected from the genomic alignments for all placental-specific intron gains has shown that the great majority of intron positions in placental-specific domesticated genes are highly conserved. The great majority of gained introns in domesticated genes have been fixed in the eutherian ancestor, as demonstrated by their presence and sequence conservation in all eutherian superorders [19]. From the genomic alignments we can readily trace the genes and their exons and introns compared at the nucleotide level over quite large evolutionary distances (at least 80–100 Myr) and, what is surprising, see remarkable level of conservation at the nucleotide level.

The rate of sequence divergence in introns is very high, therefore many introns are less conserved in sequence between organisms than their associated exons [25, 26]. The sequence conservation of *de novo* gained introns has been analyzed, and a striking conservation of the intron sequences in their entire length was found in 9 out of 19 retroelement-derived domesticated genes, showing 70–75% nucleotide identity between humans and Afrotheria, Xenarthra, and Laurasiatheria. Comparison of the entire domesticated genes between human and the representatives of all placental superorders has also shown ~75% nucleotide identity between humans and Afrotheria, Xenarthra, and Laurasiatheria [19]. Conservation of intronic sequences has been observed in some other Boreoeutheria genes [26], however only several short regions were shown to be highly conserved. The unusually conserved introns are mostly located in the 5′ UTR regions. It is possible that some of these introns are so highly conserved because they may have some conserved regulatory role in enhancing expression, in mRNA localization, stability, or efficiency of translation [27, 28]. It has been demonstrated that some of the domesticated genes are evolving under negative selection [10], therefore the level of unusual conservation is not limited to the exons but may also include the introns. Some of these genes are located on the X chromosomes, which may cause unusual patterns of evolution, such as lower mutation rates than on the autosomes [29]. The mutation rate on human X chromosome is indeed low and X-linked genes evolving mainly under negative selection are therefore evolving slowly [29, 30]. The analysis of intron conservation in other randomly selected genes indicates that intron sequences in all placental superorders may be more highly conserved than is generally acknowledged [19]. Such a high level of conservation of intron sequences may reflect their functional significance for the expression and regulation of domesticated and some other genes [27, 28].

7. Eutheria-Specific Domesticated Genes Are Alternatively Spliced

The analysis of domesticated genes in ASTD database (Alternative Splicing and Transcript Diversity Database; http://www.ebi.ac.uk/asd/index.html) has shown the presence of alternative splicing. It is interesting that DNA transposon-derived genes possess a larger amount of alternative splicing than the retroelement-derived genes. Up to 14 alternative splicing events can be seen per domesticated gene. More alternative splicing events can be seen in human than in mouse orthologous genes [19]. Alternative splicing in domesticated genes may have originated by mutations in splicing sites (evolution of weaker splice sites), by sequence changes in the intronic and exonic splicing silencers or enhancers (generating lower or higher densities) or by accumulation of *Alu* SINEs that can change the mode of splicing of the flanking exons [27, 31]. Since most of the alternative splicing events in domesticated genes are limited to humans the involvement of *Alu* SINEs is among the most interesting possibilities. The presence of alternative splicing events in humans indicates that these events might be quite recent. Although the gained introns in domesticated genes have been fixed in the eutherian ancestor, the alternative splicing events can be found, in the majority of cases, only in humans and, possibly, in primates. Such a distribution

pattern may indicate very recent, and probably regulatory, adaptations in the human or primate lineages [19].

8. Lineage-Specific Enrichment of Intron Sequences with TEs in Diverse Placental Superorders

The majority of intron origination events in domesticated genes have occurred in the eutherian ancestor, but introns were later independently bombarded with lineage-specific TEs in all three eutherian sister groups Afrotheria, Xenarthra, and Boreoeutheria. Independent TE bombardment of introns occurred also inside Boreoeutheria, as evidenced by the large differences in TE repertoires in these introns between Laurasiatheria and Euarchontoglires, as well as between rodents and primates. Comparison of the orthologous introns in placental superorders has shown the presence of species- or lineage-specific enrichment of TEs and highly dynamic evolution of TE content in placental mammals [19]. These findings indicate that introns in each species are under constant bombardment with TEs [32]. By such accumulation of lineage-specific SINEs they may influence the alternative splicing of the flanking exons in some species [27, 31].

9. The Number of *De Novo* Gained Introns in Domesticated Genes is among the Highest in Eukaryotes

Genome-wide comparisons of closely related species in numerous intron-rich lineages have shown that recent intron gains are indeed very rare [15, 16, 33] and that intron losses outnumber intron gains in eukaryotic orthologous genes [15, 16]. Comparison of orthologous genes from mammalian genomes failed to reveal any intron gains at all, suggesting that all introns currently contained in mammalian genes were already present at the time of radiation of mammalian orders [17, 18]. However, in contrast to previous observations, Kordis study has demonstrated (based on the analysis of >200 orthologous genes) quite extensive intron gain, mainly in the ancestor of placental mammals. Therefore, the placental mammals can now be added to the list of taxonomic groups with significant amounts of intron gain arising in the relatively recent evolutionary past (100–200 Mya). Rates of intron gain in the past tens to hundreds of million years in diverse eukaryotes have been very low [15, 16, 25, 34]. Studies of closely related species have shown that diverse eukaryotic lineages experienced surprisingly few intron gains in this period (reviewed in [25, 34]). The highest rate of recent intron gain yet observed in genome-wide ortholog comparisons was in Oikopleura, where 4260 newly acquired introns have been detected [35]. As Kordis study has shown, the extent of intron gain in chordate, lower vertebrate, amniote, mammalian and therian ancestors has been much smaller. The domesticated genes have finally provided evidence for the numerous intron gains in the ancestor of placental mammals [19], more than 160 My ago [36]. At least 50–100 cases of intron gain have been observed in this ancestor. This extent of *de novo* gained introns is similar to that reported in diverse eukaryotic lineages [33, 34, 37, 38]. The comparative genomics of eutherian domesticated genes has shown differences in the numbers of introns, indicating that intron gain is still ongoing [19].

10. All Previous Claims for the Absence of Intron Gain in Mammals Were the Consequence of Inadequate Taxon Sampling and the Comparison of Only the "Old" Orthologous Genes

A substantial excess of intron gains has been detected only for those intervals of eukaryotic evolution that are associated with major evolutionary innovations, such as the origin of eukaryotes and animals [15, 16, 34]. The presence of ~100 intron gains in placental mammals is remarkable and clearly represents just the tip of the iceberg, the number of *de novo* gained introns in the ancestor of placental mammals probably being much higher. Kordis study pointed to the serious problems arising from comparison of orthologous introns in coding regions only and from sparse taxon sampling in the genome-wide analyses of intron gain [17, 18, 39]. None of the cases reported by Kordis were observed in the previous studies of closely related (human, mouse, rat and dog as an outgroup) [17] or distantly related (fish versus mammals) species [39]. In the closely related mammalian species analyzed [17, 18] intron gains occurred before those species originated. In comparisons of distantly related vertebrate species [39] only "old" orthologous genes have been compared, and evolutionary novelties were excluded from such analyses, however the neglected intron gains occurred after the analyzed species originated. Therefore, the overall extent of intron gain in eukaryotes could be much higher than reported in previous studies [19]. The solution to the above problems is to analyse the highly neglected evolutionary gene novelties at particular time points (like in the ancestor of placentals). Kordis study provides a further cautionary example in using only closely or distantly related species and sophisticated statistical methods in directionalizing intron loss/gain events, and underscores the importance of using appropriately selected taxa and evolutionary gene novelties for accurate inferences of genome evolution [19].

11. Intron Gain and Promoter Acquisition Are Intimately Linked in Domesticated Genes

The presence of numerous functional domesticated genes in mammals [6, 11] immediately raises the question of how they can obtain regulatory sequences that allow them to become transcribed—a precondition for gene functionality. To become expressed at a significant level and in the tissues where it can exert a selectively beneficial function, a new gene needs to acquire a core promoter and other structural

elements that regulate its expression. Various sources of promoters and regulatory sequences exist and provide general insights into how new genes can acquire promoters and evolve new expression patterns [20, 40, 41]. The expression of domesticated genes may benefit from preexisting regulatory machinery and expression capacities of genes in their vicinity. Transcribed domesticated genes are often located close to other genes, suggesting that their transcription might be facilitated by open chromatin and/or regulatory elements of nearby genes. This possibility is supported by the observations that domesticated genes may be transcribed from the bidirectional CpG-rich promoters of genes in their proximity [42]. Some domesticated genes might also recruit CpG dinucleotide-enriched proto-promoter sequences in their genomic vicinity not previously associated with other genes for their transcription. Sometimes the promoters of domesticated gene may have evolved *de novo* through small substitutional changes under the influence of natural selection.

The process of promoter acquisition often involved the evolution of new 5′ untranslated exon-intron structures, which may span substantial distances between the recruited promoters and domesticated genes and is very similar to the situation observed in retrogenes [20]. Through the acquisition of new 5′-UTR structures, domesticated genes might also become transcribed from distant CpG-enriched sequences, which often have inherent capacity to promote transcription, and were not previously associated with other genes. These distant CpG "proto-promoter" elements might have been optimized by natural selection after they became associated with a functional domesticated gene. The frequent inheritance of CpG promoters might also help to explain why a significant number of domesticated genes evolved paternally or maternally imprinted expression [6, 11]. Thus, the primary role and selective benefit of newly gained 5′ UTR introns has been to span the substantial distances to potent CpG promoters driving transcription of domesticated genes and to reduce the size of the UTR exons.

Abbreviations

DBD: DNA binding domain
LCA: Last common ancestor
Mya: Million years ago
My(r): Million year
TE: Transposable element
UTR: Untranslated region
ZNF: Zinc finger.

Acknowledgments

The authors thank Professor Roger H. Pain for critical reading of the paper. This study was supported by Grant P1-0207 from the Slovenian Research Agency.

References

[1] C. Biémont and C. Vieira, "Genetics: junk DNA as an evolutionary force," *Nature*, vol. 443, no. 7111, pp. 521–524, 2006.

[2] M. G. Kidwell and D. R. Lisch, "Perspective: transposable elements, parasitic DNA, and genome evolution," *Evolution*, vol. 55, no. 1, pp. 1–24, 2001.

[3] H. H. Kazazian Jr., "Mobile elements: drivers of genome evolution," *Science*, vol. 303, no. 5664, pp. 1626–1632, 2004.

[4] J. Jurka, V. V. Kapitonov, O. Kohany, and M. V. Jurka, "Repetitive sequences in complex genomes: structure and evolution," *Annual Review of Genomics and Human Genetics*, vol. 8, pp. 241–259, 2007.

[5] A. Böhne, F. Brunet, D. Galiana-Arnoux, C. Schultheis, and J. N. Volff, "Transposable elements as drivers of genomic and biological diversity in vertebrates," *Chromosome Research*, vol. 16, no. 1, pp. 203–215, 2008.

[6] J. N. Volff, "Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes," *BioEssays*, vol. 28, no. 9, pp. 913–922, 2006.

[7] C. Feschotte and E. J. Pritham, "DNA transposons and the evolution of eukaryotic genomes," *Annual Review of Genetics*, vol. 41, pp. 331–368, 2007.

[8] C. Feschotte, "Transposable elements and the evolution of regulatory networks," *Nature Reviews Genetics*, vol. 9, no. 5, pp. 397–405, 2008.

[9] L. Sinzelle, Z. Izsvák, and Z. Ivics, "Molecular domestication of transposable elements: from detrimental parasites to useful host genes," *Cellular and Molecular Life Sciences*, vol. 66, no. 6, pp. 1073–1093, 2009.

[10] J. Brandt, A. M. Veith, and J. N. Volff, "A family of neofunctionalized Ty3/gypsy retrotransposon genes in mammalian genomes," *Cytogenetic and Genome Research*, vol. 110, no. 1–4, pp. 307–317, 2005.

[11] M. Campillos, T. Doerks, P. K. Shah, and P. Bork, "Computational characterization of multiple Gag-like human proteins," *Trends in Genetics*, vol. 22, no. 11, pp. 585–589, 2006.

[12] T. A. Heath, S. M. Hedtke, and D. M. Hillis, "Taxon sampling and the accuracy of phylogenetic analyses," *Journal of Systematics and Evolution*, vol. 46, no. 3, pp. 239–257, 2008.

[13] K. D. Pruitt, T. Tatusova, W. Klimke, and D. R. Maglott, "NCBI reference sequences: current status, policy and new initiatives," *Nucleic Acids Research*, vol. 37, no. 1, pp. D32–D36, 2009.

[14] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova, "Entrez gene: gene-centered information at NCBI," *Nucleic Acids Research*, vol. 35, no. 1, pp. D26–D31, 2007.

[15] L. Carmel, Y. I. Wolf, I. B. Rogozin, and E. V. Koonin, "Three distinct modes of intron dynamics in the evolution of eukaryotes," *Genome Research*, vol. 17, no. 7, pp. 1034–1044, 2007.

[16] M. Csuros, I. B. Rogozin, and E. V. Koonin, "A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes," *PLoS Computational Biology*, vol. 7, no. 9, Article ID e1002150, 2011.

[17] J. Coulombe-Huntington and J. Majewski, "Characterization of intron loss events in mammals," *Genome Research*, vol. 17, no. 1, pp. 23–32, 2007.

[18] S. W. Roy, A. Fedorov, and W. Gilbert, "Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 12, pp. 7158–7162, 2003.

[19] D. Kordis, "Extensive intron gain in the ancestor of placental mammals," *Biology Direct*, vol. 6, article 59, 2011.

[20] M. Fablet, M. Bueno, L. Potrzebowski, and H. Kaessmann, "Evolutionary origin and functions of retrogene introns," *Molecular Biology and Evolution*, vol. 26, no. 9, pp. 2147–2156, 2009.

[21] H. D. Tadepally, G. Burger, and M. Aubry, "Evolution of C_2H_2-zinc finger genes and subfamilies in mammals: species-specific duplication and loss of clusters, genes and effector domains," *BMC Evolutionary Biology*, vol. 8, no. 1, article 176, 2008.

[22] R. O. Emerson and J. H. Thomas, "Gypsy and the birth of the SCAN domain," *Journal of Virology*, vol. 85, no. 22, pp. 12043–12052, 2011.

[23] M. Lynch, "The origins of eukaryotic gene structure," *Molecular Biology and Evolution*, vol. 23, no. 2, pp. 450–468, 2006.

[24] X. Hong, D. G. Scofield, and M. Lynch, "Intron size, abundance, and distribution within untranslated regions of genes," *Molecular Biology and Evolution*, vol. 23, no. 12, pp. 2392–2404, 2006.

[25] S. W. Roy and W. Gilbert, "The evolution of spliceosomal introns: patterns, puzzles and progress," *Nature Reviews Genetics*, vol. 7, no. 3, pp. 211–221, 2006.

[26] M. P. Hare and S. R. Palumbi, "High intron sequence conservation across three mammalian orders suggests functional constraints," *Molecular Biology and Evolution*, vol. 20, no. 6, pp. 969–978, 2003.

[27] E. Kim, A. Goren, and G. Ast, "Alternative splicing: current perspectives," *BioEssays*, vol. 30, no. 1, pp. 38–47, 2008.

[28] C. Cenik, A. Derti, J. C. Mellor, G. F. Berriz, and F. P. Roth, "Genome-wide functional analysis of human 5' untranslated region introns," *Genome Biology*, vol. 11, no. 3, article r29, 2010.

[29] B. Vicoso and B. Charlesworth, "Evolution on the X chromosome: unusual patterns and processes," *Nature Reviews Genetics*, vol. 7, no. 8, pp. 645–653, 2006.

[30] S. F. Schaffner, "The X chromosome in population genetics," *Nature Reviews Genetics*, vol. 5, no. 1, pp. 43–51, 2004.

[31] G. Lev-Maor, O. Ram, E. Kim et al., "Intronic Alus influence alternative splicing," *PLoS Genetics*, vol. 4, no. 9, Article ID e1000204, 2008.

[32] J. Brosius, "Genomes were forged by massive bombardments with retroelements and retrosequences," *Genetica*, vol. 107, no. 1–3, pp. 209–238, 1999.

[33] C. B. Nielsen, B. Friedman, B. Birren, C. B. Burge, and J. E. Galagan, "Patterns of intron gain and loss in fungi," *PLoS Biology*, vol. 2, no. 12, Article ID e422, 2004.

[34] S. W. Roy and M. Irimia, "Mystery of intron gain: new data and new models," *Trends in Genetics*, vol. 25, no. 2, pp. 67–73, 2009.

[35] F. Denoeud, S. Henriet, S. Mungpakdee et al., "Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate," *Science*, vol. 330, no. 6009, pp. 1381–1385, 2010.

[36] R. W. Meredith, J. E. Janečka, J. Gatesy et al., "Impacts of the cretaceous terrestrial revolution and KPg extinction on mammal diversification," *Science*, vol. 334, no. 6055, pp. 521–524, 2011.

[37] W. Li, A. E. Tucker, W. Sung, W. Kelley Thomas, and M. Lynch, "Extensive, recent lntron gains in Daphnia populations," *Science*, vol. 326, no. 5957, pp. 1260–1262, 2009.

[38] A. Farlow, E. Meduri, M. Dolezal, L. Hua, and C. Schlötterer, "Nonsense-mediated decay enables intron gain in Drosophila," *PLoS Genetics*, vol. 6, no. 1, Article ID e1000819, 2010.

[39] Y. H. Loh, S. Brenner, and B. Venkatesh, "Investigation of loss and gain of introns in the compact genomes of pufferfishes (Fugu and Tetraodon)," *Molecular Biology and Evolution*, vol. 25, no. 3, pp. 526–535, 2008.

[40] H. Kaessmann, N. Vinckenbosch, and M. Long, "RNA-based gene duplication: mechanistic and evolutionary insights," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 19–31, 2009.

[41] H. Kaessmann, "Origins, evolution, and phenotypic impact of new genes," *Genome Research*, vol. 20, no. 10, pp. 1313–1326, 2010.

[42] P. Kalitsis and R. Saffery, "Inherent promoter bidirectionality facilitates maintenance of sequence integrity and transcription of parasitic DNA in mammalian genomes," *BMC Genomics*, vol. 10, article 498, 2009.