Research Article

Evolutionary History of LTR Retrotransposon Chromodomains in Plants

Anton Novikov,¹ Georgiy Smyshlyaev,² and Olga Novikova^{3,4}

¹ Laboratory of Molecular Genetic Systems, Institute of Cytology and Genetics, Novosibirsk, 630090, Russia

² Department of Natural Sciences, Novosibirsk State University, Novosibirsk, 630090, Russia

³ Department of Plant Pathology, University of Kentucky, Lexington, KY 40546, USA

⁴ Department of Biological Sciences, University at Albany, Life Sciences Building 2061, 1400 Washington Avenue, Albany, NY 12222, USA

Correspondence should be addressed to Olga Novikova, novikova.olga.uky@gmail.com

Received 15 September 2011; Revised 27 January 2012; Accepted 12 February 2012

Academic Editor: Jim Leebens-Mack

Copyright © 2012 Anton Novikov et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Chromodomain-containing LTR retrotransposons are one of the most successful groups of mobile elements in plant genomes. Previously, we demonstrated that two types of chromodomains (CHDs) are carried by plant LTR retrotransposons. Chromodomains from group I (CHD.I) were detected only in Tcn1-like LTR retrotransposons from nonseed plants such as mosses (including the model moss species *Physcomitrella*) and lycophytes (the *Selaginella* species). LTR retrotransposon chromodomains from group II (CHD.II) have been described from a wide range of higher plants. In the present study, we performed computer-based mining of plant LTR retrotransposon CHDs from diverse plants with an emphasis on spike-moss *Selaginella*. Our extended comparative and phylogenetic analysis demonstrated that two types of CHDs are present only in the *Selaginella* genome, which puts this species in a unique position among plants. It appears that a transition from CHD.I to CHD.II and further diversification occurred in the evolutionary history of plant LTR retrotransposons at approximately 400 MYA and most probably was associated with the evolution of chromatin organization.

1. Introduction

A chromodomain (CHD) is a protein domain involved in chromatin remodeling and the regulation of gene expression in eukaryotes (e.g., [1–3]). CHDs perform a wide range of functions, including chromatin targeting and interactions between different proteins, RNA and DNA [3]. There are two major groups of CHDs that are found in eukaryotic chromodomain-containing proteins. The so-called “classical” CHDs carry the characteristic chromo-box motif (Y/f)-(L/F/Y)-(L/I/V)-K-(W/y)-(k/r)-g (single-letter code, capital letters standing for the most prominent aminoacid) [4]. The “classical” CHDs are highly conserved among eukaryotes and are represented in a large number of proteins in many genomes. They are believed to have a similar three-dimensional structure, which consists of an N-terminal three-stranded β -barrel capped by a C-terminal helix [5].

Three conserved residues, Y24, W45, and Y48, are essential for aromatic pocket formation [6, 7].

The second group of CHDs, “shadow” chromodomains, is more variable and includes chromo-related domains, which are well conserved in their central region, but they deviate significantly in other regions. The majority of the “shadow” CHDs contain the conserved residue W45 and lack Y24 and Y48 [3, 4]. In comparison with the “classical” CHDs, the shadow chromodomains contain one helix at the N-terminus and another inserted before the C-terminal helix [8]. The best-known protein with both types of chromodomains is heterochromatin protein-1 (HP1). HP1 presence is a hallmark of constitutive heterochromatin in *Drosophila*, a condensed and highly repressive type of chromatin that organizes the repetitive pericentromeric DNA. This protein contains an N-terminal “classical” CHD and a C-terminal shadow chromo-related domain. The CHD of HP1 binds to histone

H3 dimethyl-K9 (H3K9me2) and histone H3 trimethyl-K9 (H3K9me3) to help establish transcriptionally silent heterochromatin [9–11].

The chromodomain has been found not only in eukaryotic functional proteins but also in diverse LTR retrotransposons, which are called chromodomain-containing Gypsy LTR retrotransposons or chromoviruses [12, 13]. Chromoviruses are the most widespread lineage of Gypsy LTR retrotransposons and are present in the genomes of fungi, plants, and vertebrates [13, 14]. Two distinct groups of retrotransposon CHDs have been described. Group I CHDs from retrotransposons (CHDs_I) are similar to “classical” CHDs of chromodomain-containing proteins [15]. This group of CHDs was found in diverse eukaryotic LTR retrotransposons, including fungal and vertebrate Gypsy elements, as well as in LTR retrotransposons from moss *Physcomitrella patens* and spike-moss *Selaginella moellendorffii*, which belong to the Tcn1 clade [13, 16–18]. The information on the role of CHDs_I in the retrotransposition of LTR retrotransposons is limited. The transposition activity of the MAGGY retrotransposon of the rice blast fungus *Magnaporthe oryzae* dramatically decreased with the loss or alteration of the chromodomain [19]. On the other hand, the chromointegrase of the Tfl LTR retrotransposon from *Schizosaccharomyces pombe* that lacks the chromodomain demonstrated a significantly higher activity and a substantially reduced substrate specificity [20]. As was demonstrated recently, the MAGGY chromodomain interacts with H3K9me2 and H3K9me3 in a similar way compared to the HP1 “classical” chromodomain. It was proposed that chromodomains can target the integration of chromoviruses into heterochromatic regions [15].

Representatives of group II CHDs from retrotransposons (CHDs_II) lack the first conserved aromatic residue (Y24) and usually the third (Y48). Group II has only been identified in plant Gypsy LTR retrotransposons. Little is known about the activity and role of CHDs_II from plant retrotransposons. The mostly heterochromatic distribution of plant chromoviruses along with data describing the localization of chromodomain-YFP fused protein in heterochromatin can be used as indirect evidence for recognizing heterochromatin and directing the integration role of CHDs_II [15]. Nevertheless, the actual mechanisms with which these chromodomains act are still unknown.

Previously, we demonstrated that our knowledge of plant chromodomain-containing LTR retrotransposons is mostly limited to knowledge about seed plants (mostly angiosperms) [16–18]. An investigation of retrotransposons from nonseed plants could shed light on the evolutionary history of retrotransposons and their impact on the evolutionary history of plant genomes. For example, it is still not clear whether CHDs_I and CHDs_II were acquired independently by distinct lineages of LTR retrotransposons or whether they evolved from a common ancestor. The present survey of chromodomains from diverse plants with an emphasis on the spike-moss *Selaginella* demonstrated that a transition from CHDs_I to CHDs_II occurred in the evolutionary history of plants approximately 500–400 MYA. Moreover, several types of clade-specific CHDs_II were found in plants; sequence dissimilarities among these clade-specific CHDs

hypothesized to indicate functional differences. We examined the evolutionary constraints that shaped the diversity of CHDs_II in plants and demonstrated that positive selection contributed to the diversification of clade-specific LTR retrotransposon CHDs. We propose that the presence of CHDs_I or CHDs_II is related to the distribution of heterochromatin/euchromatin marks and molecular differences in these marks between distinct lineages of eukaryotes, such as fungi/metazoa and plants. Both the transition from CHDs_I to CHDs_II and the diversification of clade-specific CHDs reflect evolutionary changes that occurred in plant chromatin organization.

2. Results

2.1. Novel LTR Retrotransposons from *Selaginella moellendorffii*. Previously, we described SM-Tcn1 CHD-containing Gypsy LTR retrotransposons, which presumably appeared as a result of a horizontal transfer from fungi during the early evolution of plants [18]. Several other families of LTR retroelements were identified by performing BLASTN and TBLASTN searches of the *S. moellendorffii* Whole Genome Shotgun (WGS) database (<http://genome.jgi-psf.org/Selmo1>) using the SM-Tcn1 retrotransposon as well as previously described retrotransposons from other species as queries (see Section 4). The newly identified retrotransposons were classified as representatives of the same or different families based on the levels of their similarities. More than 80% identity at the nucleotide level is believed to be sufficient for the classification of retrotransposons in the same family [21]. An additional criterion to designate retrotransposon families is a minimum of 50% nucleotide identity in LTRs [22]. The exemplar element was retrieved or reconstructed based on copies that were available for each family. These retrotransposons were used for further classification based on comparative and phylogenetic analysis, which included known LTR retrotransposons from other plants. The result of this analysis is shown in Figure 1. In total, five diverse families of CHD-containing LTR retrotransposons were found in addition to the previously described SM-Tcn1 [18]. The five families were named SM1-Galahad and SM2-Galahad, SM-Fogey, SM-Diluvium, and SM-Cranky (Table 1).

SM1-Galahad and SM2-Galahad are closely related to each other and form a common clade with Galadriel-like retrotransposons from monocots and dicots, suggesting that this clade originated before nonvascular and vascular plants separated from a common ancestor, roughly 400 MYA [23]. Among the sequences that are available in the NCBI protein database based on BLASTP analysis, the retroelement that is most closely related to SM1-Galahad and SM2-Galahad is the Galadriel LTR retrotransposon from *Lycopersicon esculentum*, which was identified in the Cf-9 disease resistance gene cluster [24].

We did not identify a putative intact copy of SM1-Galahad in *S. moellendorffii* WGS; thus, we used a number of copies to obtain a consensus sequence. SM1-Galahad is highly repeated and represented by hundreds of copies per genome, which share 95% nucleotide similarity on average. Many copies contain large deletions and/or insertions.

FIGURE 1: A neighbor-joining (NJ) phylogenetic tree based on multiple alignment of reverse transcriptase aminoacid sequences of LTR retrotransposons, including newly identified chromodomain-containing LTR retrotransposons from *Selaginella moellendorffii* (a) and the structural organization of novel retroelements from *Selaginella* (b). Statistical support was evaluated by bootstrapping (1000 replications); nodes with bootstrap values over 50% are shown. The plant-specific clades: Reina, CRM, Galadriel, and Tekay, as well as the Tcn1 clade, are indicated. The name of the host species and the accession number is indicated for the LTR elements that were taken from GenBank. Abbreviations: ORF: open reading frame, PR: aspartyl protease, RT: reverse transcriptase, RNH: ribonuclease H, INT: integrase, CHD.I and CHD.II: chromodomain group I and group II, CCHC: Zn-finger motif, TSD: target site duplications, and 5' and 3'LTRs: 5' and 3' long terminal repeats. The positions of stop-codons (SM-Cranky) are marked by asterisks.

TABLE 1: The characteristics of chromodomain-containing Gypsy LTR retrotransposons from *Selaginella moellendorffii*.

Element name	Element size [bp]	Intact copies	LTRs size [bp]	LTRs identity [%]	TSD	ITR
SM1-Galahad*	~7500	ND	1076/1146	90.2	ND	TG...CA
SM2-Galahad	7353	+	919	99.7	CCTAT...CCTAT	TGT...ACA
SM-Fogey	4916	+	233	98.7	TGCCC...TGCCC	TG...CA
SM-Diluvium	5038	+	112	97.3	GCGTA...GCGTA	TG...CA
SM-Cranky	5594	—	217	98.6	GTTTCT...GTTTCT	TG...CA

* SM1-Galahad was reconstructed based on a number of sequences; TSD: terminal site duplications; ITR: dinucleotide inverted repeat; ND: not detected.

The reconstructed SM1-Galahad is 7.5 kbp in length and carries LTRs that are more than 1 kbp in length and which possess a short inverted terminal repeat (TG...CA), typical for LTR retrotransposons and retroviruses [25]. They also contain polyguanine tracks that vary in length (from 11 to 25 bp) between different copies. A 13 bp primer binding site-like sequence (PBS) that is complementary to the 3' end region of tRNA^{Met} is present downstream of the 5' LTR, and a polypurine tract (PPT) was detected immediately upstream of the 3' LTR (Figure 1(b)). The putative single open reading frame (ORF) of SM1-Galahad is 3885 bp in length. The hypothetical protein product of the ORF of SM1-Galahad (1295 aa) exhibits significant similarity to both Gag and Pol gene products of known retroelements, especially with those of Gypsy retrotransposons. A search for characteristic motifs within the polyprotein sequence identified several functional domains, such as the cysteine or Zn-finger motif with Cys-X₂-Cys-X₄-His-X₄-Cys (CCHC) composition, proteinase (PR), reverse transcriptase (RT), ribonuclease H (RNH), and core integrase (core Int), in the order indicated, confirming that SM1-Galahad belongs to the Ty3/Gypsy LTR retrotransposons. No chromodomain (CHD) was found at the C-terminal end of the Gag-Pol protein from the reconstructed SM1-Galahad. However, analysis of the DNA sequence downstream of the putative ORF revealed two additional short ORFs (316 bp and 309 bp in length). The 309-bp ORF3 encodes a protein (113 aa) containing a CHD group I (CHD_I) domain.

The full-length putatively intact SM2-Galahad was found in scaffold_24 (between position 1030224 and 1037576). This LTR retrotransposon is represented by hundreds of copies per genome. SM2-Galahad is a 7.4 kbp LTR retrotransposon with a single ORF (Figure 1(b)). LTRs are 919 bp in length and share 99.7% similarity. A 5 bp target site duplication (TSD) was also detected (CCTAT...CCTAT). An 11 bp PBS complementary to the tRNA^{Met} is located just downstream of the 5' LTR, and 13 bp PPT is located upstream of 3' LTR. The putative ORF (4731 bp) encodes a fused Gag-Pol protein (1577 aa), which carries several functional domains, including the CCHC Zn-finger motif, PR, RT, RNH, core Int, and CHD_I.

The complete retrotransposons SM-Fogey and SM-Diluvium are 4.9 kbp and 5 kbp long, respectively (Figure 1(b)). The full-length intact copies of SM-Fogey and SM-Diluvium can be found in scaffold_20 (position from 2127930 to 2132846) and scaffold_73 (position from 936919 to 941957). Target site duplications of five base pairs mark the integration of the intact SM-Fogey and SM-Diluvium in the genomic sequence. Both LTR retrotransposons contain four catalytic regions, PR, RT, RNH, and Int of the retroviral genes *gag*

and *pol*, in a continuous single open reading frame. SM-Diluvium also has a CHD group II (CHD_II), whereas SM-Fogey does not carry a chromodomain. SM-Fogey is characterized by 233 bp terminal repeats; in contrast, the LTRs of SM-Diluvium are only 112 bp in length. The LTRs of SM-Fogey and SM-Diluvium contain consensus short inverted terminal repeats (TG...CA), which are important for the integration of retroviral sequences [26]. The primer binding site (PBS), which is complementary to the 3' end of tRNA^{Met} and has a one-nucleotide spacer next to the adjacent 5' LTR, was identified in both retrotransposons. The PPT with the stretch of 14 purines for SM-Fogey and SM-Diluvium is located immediately before the 3' LTR. Several full-length highly similar copies for both SM-Fogey and SM-Diluvium can be found in *S. moellendorffii* WGS. For example, nine putatively intact copies of SM-Fogey presented in WGS share 98.5% average similarity at the DNA level. All of them are flanked by 5-bp TSD. Despite the fact that only one full-length copy was detected for SM-Diluvium, seven full-length copies presented in the genome showed an average 99.1% DNA similarity. Additionally, the 5' and 3' LTRs of elements have high similarity, 98.7% on average for SM-Fogey and 97.3% on average for SM-Diluvium. Altogether, these features indicate that SM-Fogey and SM-Diluvium recently retrotransposed and may still be active.

The putative Gag-Pol polyproteins of SM-Fogey and SM-Diluvium were compared to those reported for other plant LTR retrotransposons. The aminoacid domains show the highest similarity to the corresponding regions of the Tekay-like retrotransposon SHMIDT that is adjacent to the disease resistance-priming gene NPR1 in *Beta vulgaris* (EF101866; [27]) and retrotransposons from the diverse *Oryza* species, including LTR retrotransposons from *Oryza sativa* and *Oryza australiensis* (e.g., DQ365821, DP000086; [28]). It appears that SM-Fogey and SM-Diluvium are the most closely related to the Tekay clade of CHD-containing Gypsy LTR retrotransposons from plants. They formed a common branch on the phylogenetic tree, but neither SM-Fogey nor SM-Diluvium can be assigned to this clade because of the low bootstrap support for this cluster (bootstrap 77%; Figure 1(a)).

One more CHD-containing Gypsy LTR retrotransposon, SM-Cranky, was found to be present with a few copies per genome. We detected only one full-length SM-Cranky located in scaffold_1 (between position 5651434 and 5657027) and six additional truncated copies. The retrotransposon is 5.6 kbp long with LTRs of 217 bp, terminating in the two-nucleotide inverted repeat (TG...CA). The LTRs are 98.6% identical to each other. A stretch of 12 bp located

with two-nucleotide spacers downstream of the 5'LTR is complementary to the 3' end of tRNA^{Met}, probably providing a primer site for reverse transcription. The 14-bp PPT was found upstream of the 3'LTR. The putative pseudo-ORF is interrupted by three stop codons, but there are no frameshifts. This sequence can be translated to the protein that bears a resemblance to the characteristic PR, RT, Int, and CHD_II motifs (Figure 1(b)).

2.2. Several Types of Retrotransposon Chromodomains in Plants. The most intriguing finding that concerns the CHD-containing Gypsy LTR retrotransposons from the spike-moss *S. moellendorffii* is the presence of both types of LTR retrotransposon chromodomains, CHD_I and CHD_II, in the same plant genome. The SM-Tcn1 [18], SM1-Galahad, and SM2-Galahad LTR retrotransposons from *Selaginella* carry CHD_I, while SM-Diluvium and SM-Cranky contain CHD_II. The distribution of both retrotransposon CHD types among different taxa is considered to be well known. CHD_I is typical for fungal and animal LTR retrotransposons as well as for green algae (Chlamyvir clade), whereas only CHD_II was found in LTR retrotransposons from seed plants (Tekay, Galadriel, and Reina clades [13]). However, it appears that phylogenetic distribution of CHD_I-containing elements extends to most if not all extant green plant lineages. Previously, we reported that CHD_I-containing LTR retrotransposons belonging to the Tcn1 clade can be found in both moss *Physcomitrella* and spike-moss *Selaginella* [16, 18].

To expand our understanding of the evolution and diversity of retrotransposon chromodomains in green plants, we implemented a search throughout sequence databases including but not limited to PlantGDB (<http://www.plantgdb.org/>) and Phytozome (<http://www.phytozome.net/>). A plant retrotransposon CHD search was performed with (TBLASTN) using aminoacid sequences of known CHDs as queries (see Section 4). Altogether 114 plant species were investigated and results for some of them are presented in Table 2. The full list and primary information can be found in Supporting Information Table 1S and Table 2S. It should be noted that the majority of sequences available were derived from EST databases. In fact, CHD-containing retrotransposons are barely expressed, and as a rule, they are underrepresented in ESTs [13, 16]. Another factor that has an effect on the final result is the presence of a strong bias with respect to the species diversity that is represented in databases. More than 83% of the analyzed species (95 out of a total of 114) were angiosperms, and only a few representatives from other green plant groups are currently available. Among the analyzed species, only 26 did not produce any significant hits. For 80 species, retrotransposon CHDs were detected in ESTs; CHDs were also found in GenomeSurvey Sequences (GSS) and Whole Genomic Sequences (WGS) databases for 46 investigated species. Among those species for which the WGS database is available, a few did not show the presence of retrotransposon CHDs, including red algae (*Porphyra* *zeoensis*, *Galdieria* *sulphuraria*, and *Cyanidioschyzon* *merolae*) as well as one green algae, *Micromonas pusilla* CCMP1545. This result can arise from either the limited

number of sources of sequences available or the loss of this type of retrotransposon.

The majority of the CHDs detected belonged to group II. Only the representatives of the Chlamyvir clade showed the presence of CHD_I which was found in green algae *Chlorella vulgaris* C-169 and *Chlamydomonas reinhardtii* (Table 2; [13]). The source of ESTs as well as genomic sequences of gymnosperms is very limited, especially in comparison to those for angiosperms. Nevertheless, we were able to identify a few LTR retrotransposon CHDs in all of the species investigated (Table 2; Supporting Information Table 1S).

Comparative analysis indicated that almost all of the chromodomains of type II can be easily classified as Reina-, Tekay-, or Galadriel-like CHDs based on their sequence similarity. Almost all of the plant species that produced hits in our search contain Reina- and Tekay-like CHDs. Galadriel-like CHDs were underrepresented in the analyzed databases. The phylogenetic analysis based on the aminoacid sequences of newly identified CHDs and CHDs from known plant LTR retrotransposons support these findings with several exceptions: (i) Galadriel-like LTR retrotransposons from *Selaginella* have CHD_I and are grouped with the Tcn1 clade; (ii) CHDs from conifers, which were previously believed to belong to Reina-like LTR retrotransposons from angiosperms, actually form their own branch; (iii) three CHDs identified in green algae grouped together with the *Selaginella* SM-Cranky LTR retrotransposon and not with CHDs from the Chlamyvir clade (Figure 2; [13]). Additionally, Tekay-like LTR retrotransposon CHDs from gymnosperms formed a common cluster that appears to be distinct from other Tekay-like CHDs. It is worthwhile to note that Tekay-like CHDs retrieved from representatives of the family Poaceae formed their own branch, with fairly high bootstrap support (bootstrap value 77%; Figure 2). Tekay-like LTR retrotransposons are known to be highly repeated in grass genomes. Moreover, retrotransposon activity has been implicated as playing a major role in genome size evolution in angiosperm lineage. This has been especially well-characterized in the Poaceae (e.g., [29, 30]).

2.3. Clade-Characteristic Protein Motifs Can Be Found in Plant CHDs. Comparative analysis of primary sequences and tertiary structures of retrotransposon CHDs and the CHDs from functional proteins with known function can provide important insights into the possible roles of some of the specific aminoacids and the retrotransposon CHDs as a whole [3, 31–33]. Based on multiple alignments of selected CHDs from different clades as well as “classical” and shadow CHDs from cellular functional proteins, we identified changes in plant CHDs starting with the CHD_I found in LTR retrotransposons from green algae, moss *Physcomitrella*, and spike moss *Selaginella* and culminating in the CHD_II that is isolated from genomes of gymnosperms and angiosperms (Figure 3). The CHDs of LTR retrotransposons obtained from the *Selaginella* genome represent transitional stages between CHD_I and CHD_II. For example, SM2-Galahad CHD is close to the “classical” CHDs, whereas SM1-Galahad CHD contains a few substitutions in the chromo box motif (Y/f)-(L/F/Y)-(L/I/V)-K-(W/y)-(k/r)-g. The chromo-box is one of

TABLE 2: List of some plant species used in this study, their taxonomy (according to NCBI Taxonomy: <http://www.ncbi.nlm.nih.gov/taxonomy>) and the results of *in silico* mining of chromodomains by LTR retrotransposon clades. The full list is available in Supporting Information Table S1 in Supplementary Material available online at doi: 10.1155/2012/874743.

Class	Order	Family	Species	EST			GSS/WGS		
				Chlamyvir	Other		Chlamyvir	other	
Bangiophyceae	Bangiales	Bangiaceae	<i>Porphyra yezoensis</i>	—	—		NA	NA	
	Cyanidiales	Cyanidiaceae	<i>Galdieria sulphuraria</i>	—	—		NA	NA	
Trebouxiophyceae	Chlorellales	Chlorellaceae	<i>Chlorella vulgaris</i>	—	—		7	—	
Chlorophyceae	Chlamydomonadales	Chlamydomonadaceae	<i>Chlamydomonas reinhardtii</i>	—	—		13	3	
		Volvocaceae	<i>Volvox carteri f. nagariensis</i>	3	—		241	20	
				Reina	Tekay	Galadriel	Reina	Tekay	Galadriel
Lycopodiopsida	Lycopodiales	Lycopodiaceae	<i>Huperzia serrata</i>	—	—	—	NA	NA	NA
Polypodiopsida	Polypodiales	Pteridaceae	<i>Adiantum capillus-veneris</i>	—	1	—	NA	NA	NA
Coniferopsida	Coniferales	Pinaceae	<i>Picea glauca</i>	6	2	—	—	—	—
			<i>Pinus taeda</i>	4	5	—	18	2	—
			<i>Pinus banksiana</i>	9	3	—	—	—	—
Monocotyledons	Poales	Poaceae	<i>Triticum aestivum</i>	37	5	—	10	91	—
			<i>Brachypodium distachyon</i>	11	—	—	46	371	—
			<i>Oryza sativa Indica Group</i>	5	1	—	64	239	—
			<i>Oryza sativa Japonica Group</i>	55	6	—	23	78	—
			<i>Sorghum bicolor</i>	14	—	—	18	>1000	—
			<i>Zea mays</i>	53	105	—	600	>10000	—
Eudicotyledons	Solanales	Solanaceae	<i>Solanum lycopersicum</i>	—	11	—	17	>1000	80
			<i>Solanum tuberosum</i>	2	8	1	11	>1000	12
			<i>Nicotiana tabacum</i>	23	39	2	185	2973	297
	Lamiales	Phrymaceae	<i>Mimulus guttatus</i>	2	—	1	16	370	13
	Brassicales	Brassicaceae	<i>Arabidopsis thaliana</i>	2	6	—	14	23	—
			<i>Brassica napus</i>	2	3	—	5	99	2
	Sapindales	Rutaceae	<i>Citrus clementina</i>	—	—	1	11	8	35
	Fabales	Fabaceae	<i>Glycine max</i>	1	2	—	22	252	—
			<i>Lotus japonicus</i>	2	3	—	8	62	—
	Vitales	Vitaceae	<i>Vitis vinifera</i>	—	4	4	18	59	327

NA: no data available.

the motifs that is essential for hydrophobic core formation [31, 32]. This motif is characteristic for the “classical” CHDs of chromodomain-containing proteins [3]. SM-Diluvium, lacks Y24 (corresponding to position 1 in the multiple alignment represented in Figure 3), but it still has both aromatic

aminoacids, W45 and Y48 (positions 28 and 31), which form methyl binding cages [3, 32].

With respect to different clades of LTR retrotransposons, Galadriel-like CHDs lost essential aminoacids (Y24 and Y48) and diverged significantly from CHD.I. Nevertheless,

FIGURE 2: Maximum likelihood (ML) phylogenetic tree based on LTR retrotransposon chromodomain (CHD) amino acid sequences including CHDs from newly identified chromodomain-containing LTR retrotransposons from *Selaginella moellendorffii*. Statistical support was evaluated by using aLTR; nodes with aLTR statistics over 50% are shown. The plant-specific clades Chlamyvir, Reina, CRM, Galadriel, and Tekay, as well as the Tcn1 clade, are indicated. The names of the host species and the accession numbers for the LTR elements are available in Supporting Information Figure 1S. The taxonomic range of the host species is indicated by colored boxes and includes angiosperms, gymnosperms, green algae, and nonseed plants (*Selaginella* and *Physcomitrella*).

Galadriel-like CHDs have the conservative motif (Y/f)-(L/Y)-(I/V)-k-W-k-g (single-letter code, capital letters represent the most prominent amino acid), which is very close to the chromo box. Reina-like LTR retrotransposons have traces of this motif, but Tekay-like CHDs have lost the motif

with the exception of V43 (position 26 in the multiple alignment) and the highly conservative residue W45 (Figure 3). At the same time, Tekay-like CHDs have a number of conservative characteristic motifs. For example, the protein motif (K/R)-X-(L/T)-R-X-(k/r) is present in all of the investigated

FIGURE 3: Multiple alignment of chromodomains. The following were used in the alignment: the “classical” and “shadow” chromo-like motifs of functional proteins from diverse animal species and *Arabidopsis* (At_CMT1); group I and group II chromodomains from diverse plant LTR retrotransposons belonged to Tcn1, Galadriel, Reina, and Tekay clades as well as unclassified (Other); CHD-I from Tcn1 LTR retrotransposons of fungi *Cryptococcus neoformans*. The GenBank accession numbers are indicated for each sequence. The most conservative domains for each of the groups are shown on the top. The chromo-box is indicated on the very top, which also shows the secondary structure elements (arrows indicate β -strands; rectangle α -helix) and conserved residues that form the complementary surface that is responsible for H3 peptide recognition (green boxes) [3–7]. Three aminoacids that have been shown to be under positive selection are highlighted in yellow.

Tekay-like retrotransposon CHDs but not in Reina- or Galadriel-like CHDs. One more Tekay-characteristic motif, EEXTWEXE, is highly conserved in Tekay CHDs. A similar motif can be found in diverse LTR retrotransposon CHDs; however, this motif is not as conserved in other clades as it is in Tekay. The aminoacid motif TWE is extremely conserved among all of the retrotransposon CHDs, with a few exceptions in green algae and is believed to have important functions. Interestingly, this particular motif is absent in the majority of shadow CHDs and in a few “classical” CHDs of chromodomain-containing proteins. This motif corresponds to the $\beta 3$ strand in the tertiary structure [3, 31].

While a majority of plant retrotransposon CHDs lack conserved aromatic residues Y24 and Y48, they still retain high sequence similarity with known “classical” CHDs of functional proteins. This similarity is much higher than the similarity among “classical” and shadow CHDs. Moreover, the analysis of tertiary structure shows the presence of all of the structural features that are characteristic of “classical” CHDs (see below).

2.4. Positive Selection for Retrotransposon CHDs in Plants.

For further understanding of processes that lead to the current diversity of plant retrotransposon CHDs, we examined evolutionary constraints that shape the essential domains of plant LTR retrotransposon polyproteins. The presence of conservative motifs in CHD sequences among diverse LTR retrotransposons suggests that evolution has been strongly constrained. At the same time, the presence of conserved clade-specific motifs, as well as the transition from CHD_I to CHD_II can indicate that some aminoacid changes had a selective advantage during diversification between clades and could be accumulated at a rate higher than expected under natural evolution (positive selection). To distinguish between these possibilities and to indicate the positions that evolved under positive selection, we analyzed the nonsynonymous to synonymous substitutions rate ratio (ω , see Section 4 for details).

Only retrotransposons that maintained intact domains were chosen for further analysis. CHD, core integrase (Int), and reverse transcriptase (RT) domains of 39 LTR retrotransposons were used as datasets. Phylogenetic trees based on multiple alignments of nucleotide sequences were reconstructed for each domain separately. As expected, the major difference in tree topologies among datasets included the appearance of common clusters for Galadriel-like LTR retrotransposons from *Selaginella* and the Tcn1 clade on tree reconstructions based on CHD sequences. Overall, RT- and CHD-based phylogenies are similar, whereas the Int-based tree has a different position for the Reina clade (Figure 4).

First, we estimated the ω ratio averaged over all of the sites and all of the lineages using M0 model. This model yields estimates that are close to 0: $\hat{\omega} = 0.022$ for RT, $\hat{\omega} = 0.039$ for Int, and $\hat{\omega} = 0.068$ for the CHD domain. This low rate indicated that there is a dominating role of purifying selection in the evolution of all of the domains. The clade-site test demonstrated that no events of positive selection were inferred for RT, which was expected for a highly conserved

protein domain evolving under strict constraints (e.g., [34–36]). Unexpectedly, strong positive selection was detected by clade-site test on the branch subtending the Reina clade on the Int phylogenetic tree (R branch; Figure 4(b)). This result was confirmed by a branch-site test of positive selection (modified model A—modified model A with $\omega_2 = 1$ fixed comparison) that detected positive selection on the R branch with a significance level of 0.01 (Table 3), taking into consideration multiple testing corrections (see Section 4 for details). Inference of positive selection can be an artifact when the synonymous substitutions reaching saturation. However, it is highly unlikely, taken in consideration the nature of the analyzed sequences and the pattern of substitutions (see Figure 3).

The branch-site tests also revealed several events of positive selection on the CHD phylogenetic tree. The evolutionary changes of CHDs in the Galadriel-Tekay-Reina group most probably occurred under positive selection at the 0.05 significance level based on the mixture distribution by Hommel correction procedure (GTR branch on Figure 4(c)). Nonsynonymous substitutions were also significantly elevated above background on branches subtending the Tekay-Reina group and the Galadriel clade (branch site significance level of 0.1; TR and G branches on Figure 4(c), resp.). We found evidence for the positive selection of CHDs for the GTR and TR branches, with positive signals coming from a few codons (significance level 0.1; Figure 5). Specifically, three codons appeared to be under positive selection for the TR branch at the cutoff posterior probability 95% (corresponding to aminoacid residues in positions 3, 46, and 48 of the multiple alignments presented in Figure 5(a)). Proline residues that are located in positions 3 and 48 of CHDs from Tekay and Reina LTR retrotransposons are highly conserved in these clades (see also Figure 3). Such a high degree of conservation may indicate that these residues are functionally important for both the Reina- and Tekay-like CHDs. P3 is located in a position that corresponds to the residue V3 in “classical” CHDs and is believed to participate in the formation of a complementary surface that is responsible for histone 3 peptide (H3) recognition (based on a study of CHDs from histone protein 1, dmHP1, from *Drosophila melanogaster*; [3, 31–33]).

The P48 and E46 residues are located in the area that corresponds to the helical structure in dmHP1 CHD and other CHDs. The presence of P48 in the region, which is expected to form an α -helix, should have significant effects on secondary structure. Prolines are rarely found in α and β structures, because the structure’s side chain α -N can form only one hydrogen bond [37], which would reduce the stability of such structures. At the same time, prolines are easily accommodated in a variety of turns; for example, as a Pro-X corner (where X is a variable aminoacid residue) [38]. We reconstructed the tertiary structure of some representatives of plant retrotransposon CHDs using I-TASSER [39]. All of the representatives clearly exhibited the presence of tertiary structure similar to that of dmHP1 (Figure 5(b)). However, as expected due to the presence of P48, CHDs from LTR retrotransposons that belong to the Reina and Tekay clades bear additional helix structures in comparison to dmHP1.

FIGURE 4: Maximum likelihood (ML) phylogenetic trees of sampled reverse transcriptase (a), integrase (b), and chromodomains (c) from 39 LTR retrotransposons. The plant-specific clades: Reina, Galadriel, and Tekay, as well as the Tcn1 clade, are indicated. Statistical support was evaluated by bootstrapping (100 replications); bootstrap values within clades are not shown. The changing in position of SM1-Galahad and SM2-Galahad LTR retrotransposons from *Selaginella* are shown by a gray box for each tree. The results of selection tests are reported for the tested branches, as is the proportion of sites under particular selective regimes for clades/groups. The red color indicates branches where sites under positive selection at the cutoff posterior probability 90% were identified.

TABLE 3: Maximum likelihood estimates and LRT statistics for the chromodomain.

Foreground branch	$2\Delta\ell$	$p_{1/2\chi_0^2+1/2\chi_1^2}$	$\hat{p}_0$	$\hat{p}_1$	$\hat{w}_0$	$\hat{w}_2$	Positively selected sites (90%)
base T	1.634722	0.1005255	0.71854	0.02675	0.07183	∞	
base R	1.11799	0.145176	0.69307	0.02571	0.07074	19.50879	
base G*	4.766098	0.014513	0.84189	0.03127	0.07093	∞	none
base TR*	3.720944	0.026867	0.70023	0.02601	0.07199	∞	3/9/23/46/48
base GTR*	5.652052	0.0087175	0.71684	0.02660	0.07097	∞	1/30/44/45/46/47/49

* Branches Galadriel (G), Tekay-Reina (TR), Galadriel-Tekay-Reina (GTR) are detected to be under positive selection by Hommel test procedure.

The P48 is located between the two helices and appears to be crucial for the helix-helix structure formation that is specific to plant LTR retrotransposon CHDs.

3. Discussion

The evolutionary history of retrotransposons includes the gain (and loss) of functional enzymatic domains, which allows them to adapt to a constantly changing genomic environment [12, 43–45]. The chromodomain (CHD) is believed to be a comparatively recent acquisition of LTR retrotransposons [12, 45]. The role of the chromodomains most likely is in targeting the insertion of new LTR retrotransposon copies into heterochromatic regions by recognizing specific heterochromatic histone marks and/or other factors [15]. As a consequence, LTR retrotransposons can easily avoid subsequent inactivation and elimination (purifying selection) because the chance to interfere with any coding sequence is small in heterochromatic regions [46].

Comparative and phylogenetic analysis demonstrates that plant LTR retrotransposon CHDs represent heterogeneous groups of enzymatic domains with a complex evolutionary history. First, it appears that plant retrotransposon CHD.II evolved from CHD.I, which can still be found in genomic sequences of green algae (Chlamyvir clade) and nonseed plants such as mosses and lycophytes (Tcn1 clade; Figure 6; and [13, 16, 18]). In addition, the SM1-Galahad and SM2-Galahad found in lycophyte *Selaginella* carried CHD.I, while belonged to the Galadriel clade of plant LTR retrotransposons. All of the other known representatives of this clade possess typical plant CHD.II domain. Lycophyte *Selaginella* appears to be a unique model species for the investigation of chromodomains among plants (Figure 6). This species is the only plant species known to have both types of retrotransposon CHDs in its genome. Moreover, LTR retrotransposon CHDs found in *Selaginella* represent transitional stages between CHD.I (found in fungi and animals) and CHD.II (described from angiosperms). Interestingly, the few CHDs that were found in gymnosperms are also distant from “typical” angiosperm CHD.II domain. We believe that the evolutionary history of plant LTR retrotransposon CHDs and their diversity among different plant groups reflect changes that occurred in chromatin organization (e.g., the distribution of heterochromatin/euchromatin mark; or molecular differences in specific heterochromatin/euchromatin marks) from green algae to higher plants. It was proposed earlier that although the histone methylation

marks are conserved among eukaryotes, the distribution of the individual marks and their functional meaning may have diverged as different phyla evolved [47].

The occurrence of heterochromatic marks in plants differs from that of fungi and mammals. For example, histone H3 trimethyl-K9 (H3K9me3) is a heterochromatin-specific mark in *Schizosaccharomyces pombe* [48] and in mammals [49, 50] but has never been found to be associated with heterochromatin in plants (reviewed in [51]). Moreover, heterochromatin-specific marks have an uneven distribution among plants: histone H3 dimethyl-K27 (H3K27me2) has been shown to be a typical modification in the heterochromatic regions of *Arabidopsis thaliana*, *Vicia faba*, *Zea mays*, and *Secale cereale*, but it was not detected in species such as *Glycine max*, *Plantago ovate*, and *Hordeum vulgare* [52–58]; only two species so far, *S. cereale* and in *V. faba*, showed labeling of heterochromatin with histone H3 trimethyl-K27 (H3K27me3) [53, 55]. Very little is known about chromatin organization in mosses, lycophytes, and ferns [59]. The limited information that is available for gymnosperms indicates that their heterochromatic marks seem to be quite different from those of angiosperms [47]. For example, histone H3 monomethyl-K9 (H3K9me1), H3K9me2, and histone H3 monomethyl-K27 (H3K27me1) modifications, which are believed to be associated with silencing and heterochromatin formation in *A. thaliana*, are underrepresented in *Picea abies* and *Pinus sylvestris*. At the same time, H3K9me3 and H3K27me3 are typical heterochromatic marks in two gymnosperm species but are not present in *Arabidopsis* [47].

One would expect mobile elements to be very sensitive to any changes in host genome function and organization; they must adapt or be eliminated from the genome. Divergence in the distribution of individual heterochromatin-associated histone methylation marks could trigger the evolutionary changes of LTR retrotransposon CHDs in plants, which would result in a shift from the original CHD.I (still present in green algae, mosses, and lycophytes) to CHD.II (all higher plants) and subsequent subdivision into clade-specific CHDs (Tekay-, Reina-, and Galadriel-like CHDs). The initial plasticity of CHDs provided a wide range of possibilities for evolution. Chromodomains carry diverse functions in cells, from the recognition of specific H3 histone modifications to protein dimerisation as well as DNA and RNA binding (for review [3]). It is believed that considerable diversity of recognition by CHDs is generated within the CHD family through relatively few aminoacid substitutions at the aromatic cage or the peptide-binding sites [31]. While the function of retrotransposon CHDs is generally unknown, it was

FIGURE 5: Multiple alignment including the “classical” chromodomain (dmHP1 from *Drosophila melanogaster*) and sampled representatives of group I and group II chromodomains from LTR retrotransposons (a); estimated tertiary structure for dmHP1 CHD (pdb: 1q3l; [7]) and predicted tertiary structures for some LTR retrotransposon CHDs (b). The GenBank accession numbers are indicated for each sequence used. The secondary structure elements are shown on the top (arrows indicate β -strands; rectangle α -helix) [3–7]. Three aminoacids that were shown to be under positive selection in the Tekay-Reina cluster are highlighted in yellow. The aminoacids that are potentially under positive selection in the Galadriel-Reina-Tekay cluster are highlighted in red. An additional helix structure is indicated by an arrow for representatives of the Reina and Tekay clades.

demonstrated that CHD.I of MAGGY LTR retrotransposon from rice-blast fungus *Magnaporthe oryzae* targets the integration of new copies to heterochromatin by recognizing H3K9me2 and H3K9me3 modifications [15]. Although a colocalization of TFL2, a dmHP1-like homolog in *Arabidopsis*, and CHD.II of the Tma LTR retrotransposon from *A. thaliana* were shown in the same study, the actual interacting factor(s) for plant CHD.II was not found.

The sequence divergence between CHD.I and CHD.II is a key to the functional differences, and positive selection appears to be involved in the diversification of CHDs during

the evolutionary history of LTR retrotransposons. The presence of positive selection is uncommon among LTR retrotransposon domains [34–36]. It was proposed earlier that LTR retrotransposons rarely undergo substitution events that are driven by positive selection, which allows elements to remain unrecognized by the host genome and to escape silencing [36]. However, the CHD itself provides the possibility of escaping silencing by the specific targeting of heterochromatic regions when LTR retrotransposons integrate [15], and their rapid evolution could be advantageous. Most of the genes for which positive selection has been

FIGURE 6: Distribution of different clades of chromodomain-containing LTR retrotransposons in plants as well as CHD.I and CHD.II. The evolutionary tree is represented according to Bowman et al., 2007 [40] and Berbee and Taylor, 2001 [41] with minor modifications. Divergence times (Mya: million years ago) are indicated according to Hedges, 2002 [42]. Other: SM-Diluvium and SM-Cranky.

documented are involved in interactions between the organism and the environment (e.g., [60]) and/or are subjected to genetic conflict (e.g., [61–63]). What could be driving the evolution of chromodomains? One of the most attractive explanations is coevolutionary pressures; for example, plant LTR retrotransposon CHDs might evolve after heterochromatin-associated histone methylation marks the host species. This scenario could explain the shift from CHD.I to CHD.II after the divergence of the plant and fungi/metazoa groups as well as the divergence of CHDs between angiosperms and gymnosperms. It is possible that rapid adaptation coupled with subsequent strong selective pressure not only led to the adaptation of LTR retrotransposons to a changing “chromatin” environment in plants in general, but it also may have contributed to a functional diversification of clade-specific LTR retrotransposon CHDs. In other words, while Galadriel-, Reina- and Tekay-like CHDs were still involved in targeted-integration of new LTR retrotransposon copies, they could possibly recognize different chromatin marks or factors.

Mobile elements in general, and LTR retrotransposons in particular, are important parts of plant genomes. While it is still believed that mobile elements are selected against and silenced in host genomes to prevent their harmful effects, increasing numbers of studies indicate that mobile elements have been positively selected as major components of heterochromatin (see [64]). Chromodomain-containing LTR retrotransposons are the most remarkable example of mobile elements that developed the mechanism of targeted integration into heterochromatic regions. In the present study, we inferred that interactions between chromodomain-containing LTR retrotransposons and the host genome resulted in the present diversity of plant LTR retrotransposon CHDs and, most likely, led to the retrotransposon-enriched genome organization in plants. It is necessary to note that chromodomain-containing LTR retrotransposons in plant genomes

represent a large pool of diverse chromatin remodeling domains, which also possess high evolutionary plasticity (for a review, see [3]). The potential roles of LTR retrotransposon chromodomains in genome and chromatin organization are poorly understood but should not be underestimated.

4. Materials and Methods

4.1. Genomic Sequence Screening and Sequence and Phylogenetic Analysis. *Selaginella moellendorffii* genomic sequence is available at the DOE Joint Genome Institute ([65]; <http://genome.jgi-psf.org/Selmo1/Selmo1.info.html>). We performed BLASTN and TBLASTN searches of the *S. moellendorffii* database (<http://genome.jgi-psf.org/Selmo1/Selmo1.info.html>; with default parameters) using the SM-Tcn1 retrotransposon [18] and previously described retrotransposons from other species as queries: Os35—AC068924; rn377-208—AK068625; Reina—U69258; RIRE3—AC119148; Tekay—AF050455; Retrosor2—AF061282; Tma—AF147263; Galadriel—AF119040. The full-length copies of newly identified Gypsy LTR retrotransposons were discovered in genomic sequences and analyzed by UniPro uGENE software (<http://ugene.unipro.ru/>). The LTR retrotransposon sequences obtained during BLASTN and TBLASTN searches were localized using UniPro uGENE “Find pattern” option with default parameters (both strands, match percent: 100%, whole sequence range). Open reading frames were detected by UniPro uGENE “Find ORFs” option with default parameters. Pseudo-ORFs were manually reconstructed. Putative consensus sequences were reconstructed based on multiple alignments of copies. All DNA alignments were performed by ClustalW [66] with default parameters and were edited manually in UniPro uGENE.

The LTR retrotransposon chromodomain search was carried out using BLAST (BLASTN, TBLASTN, and BLASTP).

BLAST analysis was performed using sequence databases that were accessible from the National Center for Biotechnology Information (NCBI) server (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>; BLASTP and MegaBLAST with default parameters), the U.S. Department of Energy Joint Genome Institute (<http://genome.jgi-psf.org/>), PlantGDB (<http://www.plantgdb.org/>; BLASTN, TBLASTN and BLASTP with default parameters), and Phytozome, a tool for green plant comparative genomics (<http://www.phytozome.net/>; BLASTN with default parameters). The full list of species investigated and primary information can be found in Supporting Information Table 1S and Table 2S. Amino acid sequences of the known CHDs were used as queries: MAGGY—L35053; Tcn1—XM_571377; Retrosor2—AF061282; Tma—AF147263; Galadriel—AF119040. To discriminate between functional proteins and retrotransposon CHDs, the next round of BLASTP was performed using newly identified CHDs as queries.

All multiple alignments were performed by ClustalW [66] and were edited manually in UniPro uGENE (<http://ugene.unipro.ru/>). Phylogenetic analyses were performed using the Maximum Likelihood (ML) method in the PhyML 3.0 program [67]. Neighbor-Joining analysis was performed using the MEGA5 software [68]. Statistical support for the NJ tree was evaluated by bootstrapping (number of replications, 1000) [69]. Statistical support for the ML tree was evaluated by approximate likelihood ratio test (aLRT; Figure 2) and by bootstrapping (number of replications, 100; Figure 4) [69, 70]. The tertiary structures of investigated CHD peptides were predicted using I-TASSER server with default parameters (<http://zhanglab.ccmb.med.umich.edu/I-TASSER/>) [39]. The tertiary structure of dmHP1 from *Drosophila melanogaster* is available in Protein Data Bank (<http://www.rcsb.org/pdb/home/home.do>) under ID—1q3l [7].

4.2. Test for Selection. The multiple alignment of CHD sequences was performed using ClustalW [66] available at the RevTrans 1.4 Server ([71]; <http://www.cbs.dtu.dk/services/RevTrans>). RevTrans takes a set of DNA sequences, virtually translates them, aligns the peptide sequences, and uses this as a scaffold for constructing the corresponding DNA multiple alignment. The phylogenetic trees of domains (Figure 4) were obtained with the maximum likelihood algorithm implemented in the PhyML 3.0 program [67]. From nucleotide sequence alignments for each domain, we reconstructed the phylogenetic trees under HKY85+G model. The PhyML tree searching algorithm was chosen as the best of subtree pruning and regrafting (SPR) and nearest neighbor interchange (NNI) for more thorough explorations of the space of topologies. To assess the reliability of the reconstructed phylogenies, we performed 100 bootstrap reconstructions for each domain [69].

The nonsynonymous to synonymous substitution rate ratio ($\omega = d_N/d_S$) provides a measure of natural selection at the protein level, with $\omega = 1$, >1 , and <1 , indicating neutral evolution, purifying selection, and positive selection, respectively. We use codeml program (the PAML package) to perform lineage and clade-specific analyses of d_N/d_S ratios (ω) [72–74]. The program codeml implements large collec-

tion of codon substitution models. Positive selection is tested using a likelihood ratio test (LRT) comparing a null model that does not allow $\omega > 1$ with an alternative model that does. When two models are nested, twice the log-likelihood difference between the two models can be compared with the χ^2 distribution, with the difference in the number of parameters between the two models as the degrees of freedom (df).

At first, to evaluate whether any of three chosen domains from four diverse clades are undergoing positive selection affecting all sites over prolonged time, the simplest one-ratio site model (M0) was used (codeml parameters used were as follows: model = 0, NSsites = 0). One should keep in mind, since our null hypothesis was always the absence of positive selection failing to reject a null hypothesis and/or providing a null hypothesis were interpreted as an absence of positive selection in either case. At the second stage, we tested each of the clades/groups for a signature of positive selection using clade-site test, this compares the modified clade model C (model = 3, NSsites = 2) with the neutral M1a model (model = 0, NSsites = 0). For this comparison df was set to 3. We conducted one specific test for each domain using extension of clade model C which allows for more than two branch types. Branches leading to appropriate clade/group were labeled: Tekay: T; Reina: R; Galadriel: G; Tekay-Reina: TR; Galadriel-Tekay-Reina: GTR; Reina-Galadriel: RG; Tekay-Reina-Galadriel: TRG, other branches were used as “background”. The main purpose of the test was to identify whether there is at least one branch potentially under positive selection. All maximum likelihood estimates for RT branches of the site class 2 ω ratio were close to 0, thus taking in the consideration previous results we conclude that there are not any events of positive selection. For CHD and Int domains there was at least one estimate greater than one. Nevertheless, this clade-based test does not directly examine whether any ω ratio is significantly greater than one.

Finally, we applied the LRT based on two branch-site models with one, where ω was to be estimated and other with ω fixed to 1, to test every of five aforementioned branch of the CHD and Int trees for evidence of positive selection, this test is also known as the branch-site test of positive selection. Branch-site models of codon substitution allow ω to vary both among sites in the protein and across branches on the tree and provide a means to detect short episodes of molecular adaptation affecting just a few sites [75]. In these models it is assumed that the branches are *a priori* divided into foreground and background. Only foreground lineages may have experienced positive selection. One of two branch-site models presented in codeml—modified model A (model = 2, NSsites = 2) was used for comparison with null hypothesis [75]. The model assumes four classes of sites. Site class 0 includes codons that are conserved throughout the tree, with $0 < \omega_0 < 1$ estimated. Site class 1 includes codons that are evolving neutrally throughout the tree with $\omega_1 = 1$. Site classes 2a and 2b include codons that are conserved or neutral on the background branches, but become under positive selection on the foreground branches with $\omega_2 > 1$ estimated. The model involves four parameters in the ω distribution: p_0 , p_1 , ω_0 , and ω_2 . The null hypothesis was also modified model A but with $\omega_2 = 1$

fixed (codeml options were switched from `fix.omega = 0` to `fix.omega = 1` and `omega = 1`). A likelihood ratio test (LRT) based on models was found to have satisfactory accuracy and reasonable power [75–79]. Branch-site models allow only two types of branches thus the most common approach to test several branches on the tree is to treat every branch as foreground in turn. The probability of rejecting falsely at least one of null hypotheses in such tests can be high. The correction for multiple testing becomes necessary. The Hommel procedure that controls family-wise error rate (FWER) was used as correction method. The results of Bayes empirical Bayes (BEB) approach which accommodates uncertainties in the maximum likelihood estimates were used to identify sites under positive selection if the likelihood ratio test after correction procedure was significant. Table 3 summarizes maximum likelihood estimates and test statistics for LRTs corresponding to the every of 5 branches leading to appropriate clades/groups of the CHD tree (see Supporting Information Protocol S1 for details).

Authors' Contribution

A. Novikov and G. Smyshlyaev contributed equally to this paper.

Acknowledgments

Authors thank Dr. Mark Farman (University of Kentucky, USA) for the helpful comments and stylistic suggestions. The *Selaginella* sequence data were produced by the US Department of Energy Joint Genome Institute (<http://www.jgi.doe.gov>) in collaboration with the user community.

References

- [1] R. Paro and D. S. Hogness, "The Polycomb protein shares a homologous domain with a heterochromatin-associated protein of *Drosophila*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 88, no. 1, pp. 263–267, 1991.
- [2] E. V. Koonin, S. Zhou, and J. C. Lucchesi, "The chromo superfamily: new members, duplication of the chrome domain and possible role in delivering transcription regulators to chromatin," *Nucleic Acids Research*, vol. 23, no. 21, pp. 4229–4233, 1995.
- [3] A. Brehm, K. R. Tufteland, R. Aasland, and P. B. Becker, "The many colours of chromodomains," *BioEssays*, vol. 26, no. 2, pp. 133–140, 2004.
- [4] F. Aasland and A. F. Stewart, "The chrome shadow domain, a second chrome domain in heterochromatin-binding protein 1, HP1," *Nucleic Acids Research*, vol. 23, no. 16, pp. 3168–3173, 1995.
- [5] L. J. Ball, N. V. Murzina, R. W. Broadhurst et al., "Structure of the chromatin binding (chromo) domain from mouse modifier protein 1," *EMBO Journal*, vol. 16, no. 9, pp. 2473–2481, 1997.
- [6] S. A. Jacobs, S. D. Taverna, Y. Zhang et al., "Specificity of the HP1 chromo domain for the methylated N-terminus of histone H3," *EMBO Journal*, vol. 20, no. 18, pp. 5232–5241, 2001.
- [7] S. A. Jacobs and S. Khorasanizadeh, "Structure of HP1 chromodomain bound to a lysine 9-methylated histone H3 tail," *Science*, vol. 295, no. 5562, pp. 2080–2083, 2002.
- [8] S. V. Brasher, B. O. Smith, R. H. Fogh et al., "The structure of mouse HP1 suggests a unique mode of single peptide recognition by the shadow chromo domain dimer," *EMBO Journal*, vol. 19, no. 7, pp. 1587–1597, 2000.
- [9] A. J. Bannister, P. Zegerman, J. F. Partridge et al., "Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain," *Nature*, vol. 410, no. 6824, pp. 120–124, 2001.
- [10] M. Lachner, D. O'Carroll, S. Rea, K. Mechtler, and T. Jenuwein, "Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins," *Nature*, vol. 410, no. 6824, pp. 116–120, 2001.
- [11] J. Nakayama, J. C. Rice, B. D. Strahl, C. D. Allis, and S. I. S. Grewal, "Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly," *Science*, vol. 292, no. 5514, pp. 110–113, 2001.
- [12] 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.
- [13] B. Gorinšek, F. Gubenšek, and D. Kordiš, "Evolutionary genomics of chromoviruses in eukaryotes," *Molecular Biology and Evolution*, vol. 21, no. 5, pp. 781–798, 2004.
- [14] I. Marín and C. Lloréns, "Ty3/Gypsy retrotransposons: description of new Arabidopsis thaliana elements and evolutionary perspectives derived from comparative genomic data," *Molecular Biology and Evolution*, vol. 17, no. 7, pp. 1040–1049, 2000.
- [15] X. Gao, Y. Hou, H. Ebina, H. L. Levin, and D. F. Voytas, "Chromodomains direct integration of retrotransposons to heterochromatin," *Genome Research*, vol. 18, no. 3, pp. 359–369, 2008.
- [16] O. Novikova, V. Mayorov, G. Smyshlyaev et al., "Novel clades of chromodomain-containing Gypsy LTR retrotransposons from mosses (Bryophyta)," *Plant Journal*, vol. 56, no. 4, pp. 562–574, 2008.
- [17] O. Novikova, "Chromodomains and LTR retrotransposons in plants," *Communitative and Integrative Biology*, vol. 2, no. 2, pp. 158–162, 2009.
- [18] O. Novikova, G. Smyshlyaev, and A. Blinov, "Evolutionary genomics revealed interkingdom distribution of Tcn1-like chromodomain-containing Gypsy LTR retrotransposons among fungi and plants," *BMC Genomics*, vol. 11, no. 1, article 231, 2010.
- [19] H. Nakayashiki, T. Awa, Y. Tosa, and S. Mayama, "The C-terminal chromodomain-like module in the integrase domain is crucial for high transposition efficiency of the retrotransposon MAGGY," *FEBS Letters*, vol. 579, no. 2, pp. 488–492, 2005.
- [20] A. Hizi and H. L. Levin, "The integrase of the long terminal repeat-retrotransposon Tf1 has a chromodomain that modulates integrase activities," *Journal of Biological Chemistry*, vol. 280, no. 47, pp. 39086–39094, 2005.
- [21] 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.
- [22] P. Sanmiguel and J. L. Bennetzen, "Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons," *Annals of Botany*, vol. 82, pp. 37–44, 1998.
- [23] W. N. Stewart and G. W. Rothwell, *Paleobotany and the Evolution of Plants*, Cambridge University Press, New York, NY, USA, 1993.
- [24] M. Parniske, B. B. H. Wulff, G. Bonnema, C. M. Thomas, D. A. Jones, and J. D. G. Jones, "Homologues of the Cf-9 disease

- resistance gene (Hcr9s) are present at multiple loci on the short arm of tomato chromosome 1," *Molecular Plant-Microbe Interactions*, vol. 12, no. 2, pp. 93–102, 1999.
- [25] K. J. Dej, T. Gerasimova, V. G. Corces, and J. D. Boeke, "A hot-spot for the *Drosophila* gypsy retroelement in the ovo locus," *Nucleic Acids Research*, vol. 26, no. 17, pp. 4019–4024, 1998.
 - [26] H. M. Temin, "Origin of retroviruses from cellular moveable genetic elements," *Cell*, vol. 21, no. 3, pp. 599–600, 1980.
 - [27] D. Kuykendall, J. Shao, and K. Trimmer, "A nest of LTR retrotransposons adjacent the disease resistance-priming gene NPR1 in *Beta vulgaris* L. U.S. Hybrid H20," *International Journal of Plant Genomics*, vol. 2009, Article ID 576742, 2009.
 - [28] B. Piegue, R. Guyot, N. Picault et al., "Doubling genome size without polyploidization: dynamics of retrotransposon-driven genomic expansions in *Oryza australiensis*, a wild relative of rice," *Genome Research*, vol. 16, no. 10, pp. 1262–1269, 2006.
 - [29] G. Caetano-Anollés, "Evolution of genome size in the grasses," *Crop Science*, vol. 45, no. 5, pp. 1809–1816, 2005.
 - [30] C. Vitte and O. Panaud, "LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model," *Cytogenetic and Genome Research*, vol. 110, no. 1–4, pp. 91–107, 2005.
 - [31] K. Tajul-Arifin, R. Teasdale, T. Ravasi et al., "Identification and analysis of chromodomain-containing proteins encoded in the mouse transcriptome," *Genome Research*, vol. 13, no. 6 B, pp. 1416–1429, 2003.
 - [32] P. R. Nielsen, D. Nietlispach, H. R. Mott et al., "Structure of the HP1 chromodomain bound to histone H3 methylated at lysine 9," *Nature*, vol. 416, no. 6876, pp. 103–107, 2002.
 - [33] J. F. Flanagan, L. Z. Mi, M. Chruszcz et al., "Double chromodomains cooperate to recognize the methylated histone H3 tail," *Nature*, vol. 438, no. 7071, pp. 1181–1185, 2005.
 - [34] Y. Xiong and T. H. Eickbush, "Origin and evolution of retroelements based upon their reverse transcriptase sequences," *EMBO Journal*, vol. 9, no. 10, pp. 3353–3362, 1990.
 - [35] S. Boissinot and A. V. Furano, "Adaptive evolution in LINE-1 retrotransposons," *Molecular Biology and Evolution*, vol. 18, no. 12, pp. 2186–2194, 2001.
 - [36] R. S. Baucom, J. C. Estill, C. Chaparro et al., "Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome," *PLoS Genetics*, vol. 5, no. 11, Article ID e1000732, 2009.
 - [37] S. J. Wood, "Prolines and amyloidogenicity in fragments of the alzheimer's peptide β /A4," *Biochemistry*, vol. 34, no. 3, pp. 724–730, 1995.
 - [38] P. Y. Chou and G. D. Fasman, " β -Turns in proteins," *Journal of Molecular Biology*, vol. 115, no. 2, pp. 135–175, 1977.
 - [39] A. Roy, A. Kucukural, and Y. Zhang, "I-TASSER: a unified platform for automated protein structure and function prediction," *Nature protocols*, vol. 5, no. 4, pp. 725–738, 2010.
 - [40] J. L. Bowman, S. K. Floyd, and K. Sakakibara, "Green genes-comparative genomics of the green branch of life," *Cell*, vol. 129, no. 2, pp. 229–234, 2007.
 - [41] M. L. Berbee and J. W. Taylor, "Fungal molecular evolution: gene trees and geologic time," in *The Mycota: A Comprehensive Treatise on Fungi as Experimental Systems for Basic and Applied Research*, D. J. McLaughlin, E. G. McLaughlin, and P. A. Lemke, Eds., vol. 7 of *Systematics and Evolution, part B*, pp. 229–246, Springer, New York, NY, USA, 2001.
 - [42] S. B. Hedges, "The origin and evolution of model organisms," *Nature Reviews Genetics*, vol. 3, no. 11, pp. 838–849, 2002.
 - [43] H. S. Malik, "Ribonuclease H evolution in retrotransposable elements," *Cytogenetic and Genome Research*, vol. 110, no. 1–4, pp. 392–401, 2005.
 - [44] 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.
 - [45] C. Llorens, M. A. Fares, and A. Moya, "Relationships of gag-pol diversity between Ty3/Gypsy and Retroviridae LTR retroelements and the three kings hypothesis," *BMC Evolutionary Biology*, vol. 8, no. 1, article 276, 2008.
 - [46] P. Dimitri and N. Junakovic, "Revising the selfish DNA hypothesis: new evidence on accumulation of transposable elements in heterochromatin," *Trends in Genetics*, vol. 15, no. 4, pp. 123–124, 1999.
 - [47] J. Fuchs, G. Jovtchev, and I. Schubert, "The chromosomal distribution of histone methylation marks in gymnosperms differs from that of angiosperms," *Chromosome Research*, vol. 16, no. 6, pp. 891–898, 2008.
 - [48] T. Yamada, W. Fischle, T. Sugiyama, C. D. Allis, and S. I. S. Grewal, "The nucleation and maintenance of heterochromatin by a histone deacetylase in fission yeast," *Molecular Cell*, vol. 20, no. 2, pp. 173–185, 2005.
 - [49] A. H. F. M. Peters, S. Kubicek, K. Mechtler et al., "Partitioning and plasticity of repressive histone methylation states in mammalian chromatin," *Molecular Cell*, vol. 12, no. 6, pp. 1577–1589, 2003.
 - [50] J. C. Rice, S. D. Briggs, B. Ueberheide et al., "Histone methyltransferases direct different degrees of methylation to define distinct chromatin domains," *Molecular Cell*, vol. 12, no. 6, pp. 1591–1598, 2003.
 - [51] C. Liu, F. Lu, X. Cui, and X. Cao, "Histone methylation in higher plants," *Annual Review of Plant Biology*, vol. 61, pp. 395–420, 2010.
 - [52] A. M. Lindroth, D. Shultis, Z. Jasencakova et al., "Dual histone H3 methylation marks at lysines 9 and 27 required for interaction with CHROMOMETHYLASE3," *EMBO Journal*, vol. 23, no. 21, pp. 4286–4296, 2004.
 - [53] J. Fuchs, D. Demidov, A. Houben, and I. Schubert, "Chromosomal histone modification patterns—from conservation to diversity," *Trends in Plant Science*, vol. 11, no. 4, pp. 199–208, 2006.
 - [54] J. Shi and R. K. Dawe, "Partitioning of the maize epigenome by the number of methyl groups on histone H3 lysines 9 and 27," *Genetics*, vol. 173, no. 3, pp. 1571–1583, 2006.
 - [55] M. Carchilan, M. Delgado, T. Ribeiro et al., "Transcriptionally active heterochromatin in rye B chromosomes," *Plant Cell*, vol. 19, no. 6, pp. 1738–1749, 2007.
 - [56] S. Marschner, K. Kumke, and A. Houben, "B chromosomes of *B. dichromosomata* show a reduced level of euchromatic histone H3 methylation marks," *Chromosome Research*, vol. 15, no. 2, pp. 215–222, 2007.
 - [57] J. Fuchs and I. Schubert, "Chromosomal distribution and functional interpretation of epigenetic histone marks in plants," in *Plant Cytogenetics, Vol. 1: Genome Structure and Chromosome Function*, H. Bass and J. A. Birchler, Eds., Springer, New York, NY, USA, 2011.
 - [58] M. K. Dhar, J. Fuchs, and A. Houben, "Distribution of Eu- and heterochromatin in plantagoovata," *Cytogenetic and Genome Research*, vol. 125, no. 3, pp. 235–240, 2009.
 - [59] S. Spiker, "An evolutionary comparison of plant histones," *Biochimica et Biophysica Acta*, vol. 400, no. 2, pp. 461–467, 1975.
 - [60] Z. Yang and J. R. Bielawski, "Statistical methods for detecting molecular adaptation," *Trends in Ecology and Evolution*, vol. 15, no. 12, pp. 496–503, 2000.

- [61] C. F. Qi, F. Bonhomme, A. Buckler-White et al., "Molecular phylogeny of *Fv1*," *Mammalian Genome*, vol. 9, no. 12, pp. 1049–1055, 1998.
- [62] H. S. Malik and S. Henikoff, "Positive selection of iris, a retroviral envelope-derived host gene in *Drosophila melanogaster*," *PLoS Genetics*, vol. 1, no. 4, pp. 0429–0443, 2005.
- [63] D. Vermaak, S. Henikoff, and H. S. Malik, "Positive selection drives the evolution of rhino, a member of the heterochromatin protein 1 family in *Drosophila*," *PLoS Genetics*, vol. 1, no. 1, pp. 0096–0108, 2005.
- [64] C. Biémont, "Are transposable elements simply silenced or are they under house arrest?" *Trends in Genetics*, vol. 25, no. 8, pp. 333–334, 2009.
- [65] J. A. Banks, T. Nishiyama, M. Hasebe et al., "The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants," *Science*, vol. 332, no. 6032, pp. 960–963, 2011.
- [66] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Research*, vol. 22, no. 22, pp. 4673–4680, 1994.
- [67] 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.
- [68] 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.
- [69] M. Anisimova and O. Gascuel, "Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative," *Systematic Biology*, vol. 55, no. 4, pp. 539–552, 2006.
- [70] S. Guindon and O. Gascuel, "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood," *Systematic Biology*, vol. 52, no. 5, pp. 696–704, 2003.
- [71] R. Wernersson and A. G. Pedersen, "RevTrans: multiple alignment of coding DNA from aligned amino acid sequences," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3537–3539, 2003.
- [72] Z. Yang, "Maximum-likelihood models for combined analyses of multiple sequence data," *Journal of Molecular Evolution*, vol. 42, no. 5, pp. 587–596, 1996.
- [73] Z. Yang, "PAML: a program package for phylogenetic analysis by maximum likelihood," *Computer Applications in the Biosciences*, vol. 13, no. 5, pp. 555–556, 1997.
- [74] Z. Yang, "PAML 4: phylogenetic analysis by maximum likelihood," *Molecular Biology and Evolution*, vol. 24, no. 8, pp. 1586–1591, 2007.
- [75] 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.
- [76] Z. Yang, W. S. W. Wong, and R. Nielsen, "Bayes empirical Bayes inference of amino acid sites under positive selection," *Molecular Biology and Evolution*, vol. 22, no. 4, pp. 1107–1118, 2005.
- [77] J. P. Bielawski and Z. Yang, "A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution," *Journal of Molecular Evolution*, vol. 59, no. 1, pp. 121–132, 2004.
- [78] Z. Yang and R. Nielsen, "Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages," *Molecular Biology and Evolution*, vol. 19, no. 6, pp. 908–917, 2002.
- [79] M. Anisimova and Z. Yang, "Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites," *Molecular Biology and Evolution*, vol. 24, no. 5, pp. 1219–1228, 2007.