Genetics in Genomic Era

Guest Editors: Eugenia Poliakov, David N. Cooper, Elena I. Stepchenkova, and Igor B. Rogozin
Genetics in Genomic Era
Genetics in Genomic Era

Guest Editors: Eugenia Poliakov, David N. Cooper, Elena I. Stepchenkova, and Igor B. Rogozin
Editorial Board

Urs Albrecht, Switzerland
Marc Billaud, France
Claudio Bravi, Argentina
Leanne Dibbens, Australia
Patrizio Dimitri, Italy
Norman A. Doggett, USA
Francine Durocher, Canada
Shinichi Fukushige, Japan
Jonathan A. Harton, USA
M. Hattori, Japan
Ziaril H. Hawi, Australia
Akira Horii, Japan
M. Horikoshi, Japan
Martin Hlskamp, Germany
Maj Hulten, UK
Y. Kashi, Israel
M. Kreitman, USA
Ulf Kristoffersson, Sweden
Jerzy Kulski, Australia
Martin Kupiec, Israel
Paul J. Lockhart, Australia
Arne Ludwig, Germany
Fabio M. Macciardi, USA
Alexander MacKenzie, Canada
E. Martínez-Romero, Mexico
Melvin G. McInnis, USA
Wieland Meyer, Australia
Giuseppe Novelli, Italy
Hilmi Ozcelik, Canada
Svetlana D. Pack, USA
Dorien Peters, Netherlands
Wendy Robinson, Canada
Sevtap Savas, Canada
Edgar Serfling, Germany
Marcel Tilanus, Netherlands
V. W. van Beusechem, Netherlands
Jim Wainscoat, UK
Bernard Weissman, USA
Haim Werner, Israel
Meredith Yeager, USA
Jia L. Zhuo, USA
Drazen B. Zimonjic, USA
Contents

**Genetics in Genomic Era**, Eugenia Poliakov, David N. Cooper, Elena I. Stepchenkova, and Igor B. Rogozin
Volume 2015, Article ID 364960, 2 pages

**Importance of Genetic Diversity Assessment in Crop Plants and Its Recent Advances: An Overview of Its Analytical Perspectives**, M. Govindaraj, M. Vetriventhan, and M. Srinivasan
Volume 2015, Article ID 431487, 14 pages

**Unlimited Thirst for Genome Sequencing, Data Interpretation, and Database Usage in Genomic Era: The Road towards Fast-Track Crop Plant Improvement**, Arun Prabhu Dhanapal and Mahalingam Govindaraj
Volume 2015, Article ID 684321, 15 pages

Volume 2014, Article ID 210418, 5 pages

**Generalized Portrait of Cancer Metabolic Pathways Inferred from a List of Genes Overexpressed in Cancer**, Eugenia Poliakov, David Managadze, and Igor B. Rogozin
Volume 2014, Article ID 646193, 8 pages

**Complexity of Gene Expression Evolution after Duplication: Protein Dosage Rebalancing**, Igor B. Rogozin
Volume 2014, Article ID 516508, 8 pages
Genetics is a relatively young science compared to many other fields in biology, for example, evolutionary biology or physiology. However, genetics is a central theme in many fields, especially the medical sciences, because genetic studies can provide explanations and may even allow predictions to be made, in the context of a range of biological problems including the field of inherited human disorders. This has become especially important during the last decade as we have entered the “genomic era.” With the dramatic improvement of sequencing technologies and the enormous reduction in the cost of sequencing, biologists are faced with a “data avalanche.” Next generation whole exome or genome sequencing has provided us with an arsenal of tools to study human, animal, and plant genetics. The underlying genetic lesions responsible for Mendelian diseases may now be found and mapped with even a limited number (as few as one) of affected individuals (rare and neglected diseases), with the help of large quantities of “control” genome data such as those emanating from the 1000 Genomes Project. Genomic data analysis may also be useful for the dissection of the genetic mechanisms underlying complex polygenic diseases or in exploring the role of modifiers genes in influencing the age of onset or clinical severity of a given Mendelian disease entity. Although we have acquired new research capabilities, we are also encountering new problems with the analysis of genomic data and the reanalysis of already published (deposited in databases) data. Such problems of genomic data presentation, format, sharing, and reanalysis are now starting to be addressed.

This special issue is dedicated to problems of genetics in the genomic era and comprises five articles: three review articles and two research articles. In the first review article, entitled “Unlimited Thirst for Genome Sequencing, Data Interpretation, and Database Usage in Genomic Era: The Road towards Fast-Track Crop Plant Improvement,” A. P. Dhanapal and M. Govindaraj discuss the use of crop plant databases in advancing research in the genomic era. The major focus of this review is to provide knowledge on platforms for comparative genomics of agriculturally important crop plants with industrial and environmental significance. Recent advances in sequencing and resequencing of plant genomes have potentiated new analyses of genomic variation and gene function. Genetic variation databases could facilitate research that helps to improve the efficiency of plant breeding programs. This review should aid researchers in the plant science research community by providing information on available databases and platforms for genome-based analyses that help to link model systems with other plants in the genomics context. This is a timely review which highlights the recent finding of frequent gene and whole genome duplication events. Many unresolved questions however remain regarding the number and timing of such events in plant evolution.

Gene duplication is a key mechanism of genomic change in evolution. I. B. Rogozin discusses recent analyses of the “ortholog conjecture” (OC) hypothesis. Under the OC hypothesis, which is central to the functional annotation of genomes, orthologous genes are functionally more similar...
than paralogous genes at the same level of sequence divergence. A recent study found a greater functional similarity, in terms of gene ontology (GO) annotations and expression profiles, among within-species paralogs as compared to orthologs. These findings have suggested that the functional similarity of homologous genes is primarily determined by the cellular context of those genes, rather than their evolutionary history. Subsequent studies have suggested that the OC hypothesis appears to be generally valid but that a comprehensive picture of the evolution of gene expression requires the incorporation of lineage-specific aspects of paralogy. The observed complexity of gene expression evolution after duplication may be most parsimoniously explained by the duplication-degeneration-complementation model combined with selection for gene dosage.

E. Poliakov et al.’s research paper reinforces the central importance of gene expression. Some ninety years on from the first articulation of the Warburg theory of cancer cell origin, the question of altered metabolism in cancer is again assuming a central role. Analyses of signaling pathways and oncogenes in different types of cancer have been the focus of research for several decades. Now, empowered by a wealth of knowledge about tumor suppressor genes, oncogenes, and signaling pathways, the reprogramming of cellular metabolism (e.g., increased glycolysis to respiration ratio in cancer cells) has reemerged as a key step in cancer progression. To obtain a general picture of cancer metabolism and to analyze the level of expression of various genes encoding proteins including metabolic enzymes across various cancers, E. Poliakov et al. employed dbEST and Unigene data. With the appearance of abundant RNA-seq cancer data, it is interesting to ascertain if dbEST-based conclusions will hold. The authors delineated a list of genes that are overexpressed in most types of cancer. They also grouped overexpressed enzymes into KEGG pathways and analyzed adjacent pathways to describe the enzymatic reactions that take place in cancer cells thereby identifying major players in the cancer protein machinery. Glycolysis/gluconeogenesis, oxidative phosphorylation, and pyruvate metabolism appear to be the most abundant pathways although several other pathways are enriched in genes from the list. Ubiquitously overexpressed genes could be marked as nonspecific cancer-associated genes when analyzing genes that are overexpressed in certain types of cancer. Thus, the list of overexpressed genes is likely to be a useful tool for systems biology approaches to cancer research.

D. Hmida-Ben Brahim et al. discuss Huntington disease (HD) (an autosomal dominant neurodegenerative disorder). The causative mutation is an expansion of more than 36 CAG repeats in the first exon of the HTT (IT15) gene. Many studies have shown that the HTT gene interacts with several modifier genes to regulate the age at onset of HD. Their study aims to investigate the involvement of CAG expansion and 9 modifiers in the age at onset variance of 15 HD Tunisian patients. The authors establish a correlation between these modifier genes and the age of onset of this disease. Their results demonstrate a specific effect of modifier genes in each population. Despite the small number of studied patients, this report constitutes the first North African study of Huntington disease patients.

This special issue is concluded by an article by M. Srinivasan et al. that discusses the importance of genetic diversity assessment in crop plants and its recent advances. The importance of plant genetic diversity is now being widely recognized by agricultural scientists. This paper comprehensively reviews four important areas: (i) the significance of plant genetic diversity (PGD); (ii) the risk associated with the narrowing down of the genetic base of current commercial cultivars and climate change; (iii) analysis of existing PGD analytical methods in both the pregenomic and genomic eras; and (iv) the tools now available for PGD analysis in the postgenomic era. This review describes the new methods and technology for the improved and rapid assessment of the genetic diversity of crops and for the utilization of germplasm from gene-banks in their applied breeding programs. Since plant breeding research and cultivar development are integral components of improving food production, the availability of and access to diverse genetic stocks will help make the global food production network become more sustainable. The pros and cons of the use of basic and advanced statistical tools available for measuring genetic diversity are discussed.

We have no doubt that there are many topics that remain uncovered in this special issue. However, we hope that the approaches described here will become widely used by the scientific community.
Review Article

Importance of Genetic Diversity Assessment in Crop Plants and Its Recent Advances: An Overview of Its Analytical Perspectives

M. Govindaraj,1,2 M. Vetriventhan,1,2 and M. Srinivasan3

1Centre for Plant Breeding and Genetics, Tamil Nadu Agricultural University, Coimbatore 641 003, India
2International Crops Research Institute for the Semi-Arid Tropics, Patancheru, Telangana 502324, India
3School of Life Science, Bharathidasan University, Tiruchirappalli 620 024, India

Correspondence should be addressed to M. Srinivasan; mahasiva2@gmail.com

Received 17 July 2014; Revised 24 November 2014; Accepted 27 November 2014

Academic Editor: Igor B. Rogozin

Copyright © 2015 M. Govindaraj 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.

The importance of plant genetic diversity (PGD) is now being recognized as a specific area since exploding population with urbanization and decreasing cultivable lands are the critical factors contributing to food insecurity in developing world. Agricultural scientists realized that PGD can be captured and stored in the form of plant genetic resources (PGR) such as gene bank, DNA library, and so forth, in the biorepository which preserve genetic material for long period. However, conserved PGR must be utilized for crop improvement in order to meet future global challenges in relation to food and nutritional security. This paper comprehensively reviews four important areas; (i) the significance of plant genetic diversity (PGD) and PGR especially on agriculturally important crops (mostly field crops); (ii) risk associated with narrowing the genetic base of current commercial cultivars and climate change; (iii) analysis of existing PGD analytical methods in pregenomic and genomic era; and (iv) modern tools available for PGD analysis in postgenomic era. This discussion benefits the plant scientist community in order to use the new methods and technology for better and rapid assessment, for utilization of germplasm from gene banks to their applied breeding programs. With the advent of new biotechnological techniques, this process of genetic manipulation is now being accelerated and carried out with more precision (neglecting environmental effects) and fast-track manner than the classical breeding techniques. It is also to note that gene banks look into several issues in order to improve levels of germplasm distribution and its utilization, duplication of plant identity, and access to database, for prebreeding activities. Since plant breeding research and cultivar development are integral components of improving food production, therefore, availability of and access to diverse genetic sources will ensure that the global food production network becomes more sustainable. The pros and cons of the basic and advanced statistical tools available for measuring genetic diversity are briefly discussed and their source links (mostly) were provided to get easy access; thus, it improves the understanding of tools and its practical applicability to the researchers.

1. Introduction

Diversity in plant genetic resources (PGR) provides opportunity for plant breeders to develop new and improved cultivars with desirable characteristics, which include both farmer-preferred traits (yield potential and large seed, etc.) and breeders preferred traits (pest and disease resistance and photosensitivity, etc.). From the very beginning of agriculture, natural genetic variability has been exploited within crop species to meet subsistence food requirement, and now it is being focused to surplus food for growing populations. In the middle of 1960s developing countries like India experienced the green revolution by meeting food demand with help of high-yielding and fertilizer responsive dwarf hybrids/varieties especially in wheat and rice (Figure 1). These prolonged activities that lead to the huge coverage of single genetic cultivars (boom) made situation again worse in other forms such as genetic erosion (loss of genetic diversity) and extinction of primitive and adaptive genes (loss of landraces). Today with an advancement of agricultural and allied science and technology, we still ask ourselves whether we can feed the world in 2050; this
question was recently sensitized at the world food prize event in 2014 and remains unanswered in every one hands since global population will exceed 9 billion in 2050. The per capita availability of food and water will become worse year after year coping with the undesirable climate change. Therefore, it becomes more important to look at the agriculture not only as a food-producing machine, but also as an important source of livelihood generation both in the farm and nonfarm sectors. Keeping the reservoir for cultivated and cultivable crops species is a principle for future agriculture, just like keeping a museum of cultural and spiritual specialty of diverse civilized humans in various geography for their historical evidence for future. The former can play a very important role in providing adaptive and productive genes, thus leading to long-term increases in food productivity which is further associated with environmental detriment. This paper will indicate the significance of genetic conservation and its analytical tools and techniques that are made widely available for utilization in postgenomic era. Plant and animal breeders introduced desirable genes and eliminated undesirable ones slowly, altering in the process of underlying heredity principle for several decades [1]. With the advent of new biotechnological tools and techniques, this process of genetic manipulation is being accelerated and it shortened the breeding cycles, and it can be carried out with more precision (neglecting environmental effects) and fast-track manner than the classical breeding techniques.

2. Significance of Genetic Conservation of Crop Plants

The growing population pressure and urbanization of agricultural lands and rapid modernization in every field of our day-to-day activities that create biodiversity are getting too eroded in direct and indirect way. For instance, land degradation, deforestation, urbanization, coastal development, and environmental stress are collectively leading to large-scale extinction of plant species especially agriculturally important food crops. On the other hand, system driven famine such as, Irish potato famine and Southern corn leaf blight epidemic in USA are the two instances of food crises caused by large-scale cultivation of genetically homogenous varieties of potato and corn, respectively. Even after these historical events, the importance of PGR had only got popular recognition when the spread of green revolution across cultivated crops threatened the conservation of land races [2]. Green revolution technologies introduced improved crop varieties that have higher yields, and it was hoped that they would increase farmers’ income. Consequently, the Consultative Group of International Agricultural Researches (CIGAR) initiated gene banks and research centers of domestication for conserving PGR in most of the stable food crops around the world. Center for domestication: maize (Mexico), wheat and barley (middle/near East and North Africa), rice (North China), and potatoes (Peru); for further information see http://www.cigar.org/center/index.html.) The Food and Agriculture Organization (FAO) supported the International Treaty on Plant Genetic Resources (ITPGR) and UN supported the Convention on Biological Diversity (CBD) which are the international agreements that recognize the important role of genetic diversity conservation. Such treaty still plays in current and future food production as one of the major supremo [3].

Genetic diversity is the key pillar of biodiversity and diversity within species, between species, and of ecosystems (CBD, Article 2), which was defined at the Rio de Janeiro Earth Summit. However, the problem is that modern crop varieties, especially, have been developed primarily for high yielding potential under well endowed production conditions. Such varieties are often not suitable for low income farmers in marginal production environments as they are facing highly variable stress conditions [4]. Land races or traditional varieties have been found to have higher stability (adaptation over time) in low-input agriculture under marginal environments, thus, their cultivation may contribute farm level resilience in face of food production shocks [5, 6]. This is especially true in some part of Ethiopia where agroclimatic conditions are challenging, technological progress is slow, and market institutions are poorly developed and have no appropriate infrastructure [7, 8].

Why is genetic diversity important? The goal of conservation genetics is to maintain genetic diversity at many levels and to provide tools for population monitoring and assessment that can be used for conservation planning. Every individual is genetically unique by nature. Conservation efforts and related research are rarely directed towards individuals but genetic variation is always measured in individuals and this can only be estimated for collections of individuals in a population/species. It is possible to identify the genetic variation from phenotypic variation either by quantitative traits (traits that vary continuous and are governed by many genes, e.g., plant height) or discrete traits traits that fall into
discrete categories and are governed by one or few major
genes (e.g., white, pink, or red petal color in certain flowers)
which are referred to as qualitative traits. Genetic variation
can also be identified by examining variation at the level of
enzymes using the process of protein electrophoresis. Fur-
ther, genetic variations can also be examined by the order of
nucleotides in the DNA sequence.

3. Erosion of Genetic Diversity due to
Population Size: A Bottleneck Concept

It is well known that inbreeding is the most common phe-
omena in cross-pollinated crops, and in small outcross pop-
ulations it has resulted in deleterious effects and loss of fitness
of the population due to recombination between undesirable
genes (recessive identical alleles). In natural population too,
severe reductions in population size, the so-called genetic
bottleneck, leads to loss of genetic diversity and increased
susceptibility to infectious pests and diseases that supervene
increased chances of extinction of an individual crop in ques-
tion. Genetic models that predict the proportion of initial
heterozygosity retained per generation is $1 - (1/2N_e)$ where
$N_e$ is the effective population size, usually less than $N$, the
actual population size. Thus a population of $N_e = 10$ indi-
viduals loses 5% of its heterozygosity per generation. This
indicates that severe bottlenecks degrade heterozygosity and
genetic diversity [9]. Therefore, plant breeders have been
advised to maintain the optimum population size for any trait
conservation for specific purpose and its utilization for crop
improvement. Thus, before quantifying the genetic diversity,
it is essential to know the optimum population size and its
representatives to ensure no biasness in diversity assessment
that leads to wrong prediction of its value.

4. Climate Change and Its Impact on
Plant Genetic Resources

The most profound and direct impacts of climate change over
previous decade and the next few decades will surely be on
agriculture and food security. The effects of climate change
will also depend on current production conditions. The area
where already being obstructed by other stresses, such as pol-
lution and will likely to have more adverse impact by chang-
ing climate. Food production systems rely on highly selected
cultivars under better endowed environments but it might
be increasingly vulnerable to climate change impacts such as
pest and disease spread. If food production levels decreases
over the year, there will be huge pressure to cultivate the crops
under marginal lands or implement unsustainable practices
that, over the long-term, degrade lands and resources and
adversely impact biodiversity on and near agricultural areas.
In fact, such situations have already been experienced by most
of the developing countries. These changes have been seen
to cause a decrease in the variability of those genetic loci (alle-
les of a gene) controlling physical and phenotypic responses
to changing climate [10]. Therefore, genetic variation holds
the key to the ability of populations and species to persist over
evolutionary period of time through changing environments
[II]. If this persists, neither any organism can predict its
future (and evolutionary theory does not require them to)
nor can any of those organisms be optimally adapted for all
environmental conditions. Nonetheless, the current genetic
composition of a crop species influences how well its mem-
ers will adapt to future physical and biotic environments.
The population can also migrate across the landscape
over generations. By contrast, populations that have a narrow
range of genotypes and are more phenotypically uniform may
merely fail to survive and reproduce at all as the conditions
become less locally favorable. Such populations are more
likely to become extirpated (locally extinct), and in extreme
cases the entire plant species may end up at risk of extinction.
For example, the Florida Yew (Torreya taxifolia) is currently
one of the rarest conifer species in North America. But in
the early Holocene (10,000 years ago), when conditions in
southeastern North America were cooler and wetter than
today, the species was probably widespread. The reasons
for that are not completely understood, but T. taxifolia failed
to migrate towards the northward as climate changed during
the Holocene. Today, it is restricted to a few locations in the
Apalachicola River Basin in southern Georgia and the Florida
panhandle. As the T. taxifolia story illustrates, once plant
species are pushed into marginal habitat at the limitations
of their physiological tolerance, they may enter an extinction
vortex, a downward cycle of small populations, and so on [12,
13]. Reduced genetic variability is a key step in the extinction
vortex. Gene banks must be better to respond to novel and
increased demands on germplasm for adapting agriculture to
climate change. Gene banks need to include different char-
acteristics in their screening processes and their collections
need to be comprehensive, including what are now consid-
ered minor crops, and that may come with huge impact on
food baskets.

5. Assessment of Genetic Diversity in
Crop Plants

The assessment of genetic diversity within and between plant
populations is routinely performed using various techniques
such as (i) morphological, (ii) biochemical characterization/evaluation (allozyme), in the pregenomic era, and
(iii) DNA (or molecular) marker analysis especially single
nucleotide polymorphism (SNPs) in postgenomic era. Mark-
ers can exhibit similar modes of inheritance, as we observe for
any other traits, that is, dominant/recessive or codominant.
If the genetic pattern of homozygotes can be distinguished from
that of heterozygotes, then a marker is said to be codominant.
Generally codominant markers are more informative than
the dominant markers.

Morphological markers are based on visually accessible
traits such as flower color, seed shape, growth habits, and
pigmentation, and it does not require expensive technology
but large tracts of land area are often required for these field
experiments, making it possibly more expensive than molec-
ular assessment in western (developed) countries and equally
expensive in Asian and Middle East (developing) countries
considering the labour cost and availability. These marker
traits are often susceptible to phenotypic plasticity; conversely, this allows assessment of diversity in the presence of environmental variation which cannot be neglected from the genotypic variation. These types of markers are still having advantage and they are mandatory for distinguishing the adult plants from their genetic contamination in the field, for example, spiny seeds, bristled panicle, and flower/leaf color variants.

Second type of genetic marker is called biochemical markers, allelic variants of enzymes called isozymes that are detected by electrophoresis and specific staining. Isozyme markers are codominant in nature. They detect diversity at functional gene level and have simple inheritance. It requires only small amounts of plant material for its detection. However, only a limited number of enzymes markers are available and these enzymes are not alone but it has complex structural and special problems; thus, the resolution of genetic diversity is limited to explore.

The third and most widely used genetic marker type is molecular markers, comprising a large variety of DNA molecular markers, which can be employed for analysis of genetic and molecular variation. These markers can detect the variation that arises from deletion, duplication, inversion, and/or insertion in the chromosomes. Such markers themselves do not affect the phenotype of the traits of interest because they are located only near or linked to genes controlling the traits. These markers are inherited both in dominant and codominant patterns. Different markers have different genetic qualities (they can be dominant or codominant, can amplify anonymous or characterized loci, can contain expressed or nonexpressed sequences, etc.). A molecular marker can be defined as a genomic locus, detected through probe or specific starter (primer) which, in virtue of its presence, distinguishes unequivocally the chromosomal trait which it represents as well as the flanking regions at the 3' and 5' extremity [14]. Molecular markers may or may not correlate with phenotypic expression of a genomic trait. They offer numerous advantages over conventional, phenotype-based alternatives as they are stable and detectable in all tissues regardless of growth, differentiation, development, or defense status of the cell. Additionally, they are not confounded by environmental, pleiotropic, and epistatic effects. We are not describing much about the pregenomic era tools, since our paper deals with genomic advances and its assistance in crop genetic diversity assessment.

6. Analyses of Genetic Diversity in Genomic Era

A comprehensive study of the molecular genetic variation present in germplasm would be useful for determining whether morphologically based taxonomic classifications reveal patterns of genomic differentiation. This can also provide information on the population structure, allelic richness, and diversity parameters of germplasm to help breeders to use genetic resources with less prebreeding activities for cultivar development more effectively. Now germplasm characterization based on molecular markers has gained importance due to the speedy and quality of data generated. For the readers benefit, the availability of different DNA markers acronym is given in Abbreviations section.

6.1. Molecular Markers. DNA (or molecular) markers are the most widely used type of marker predominantly due to their abundance. They arise from different classes of DNA mutations such as substitution mutations (point mutations), rearrangements (insertions or deletions), or errors in replication of tandemly repeated DNA [15]. These markers are selectively neutral because they are usually located in noncoding regions of DNA in a chromosome. Unlike other markers, DNA markers are unlimited in number and are not affected by environmental factors and/or the developmental stage of the plant [16]. DNA markers have numerous applications in plant breeding such as (i) marker assisted evaluation of breeding materials like assessing the level of genetic diversity, parental selection, cultivar identity and assessment of cultivar purity [16–26], study of heterosis, and identification of genomic regions under selection, (ii) marker assisted backcrossing, and (iii) marker assisted pyramiding [27].

Molecular markers may be broadly divided into three classes based on the method of their detection: hybridization-based, polymerase chain reaction- (PCR-) based, and DNA sequence-based. Restriction fragment length polymorphisms (RFLPs) are hybridization-based markers developed first in human-based genetic study during 1980s [28, 29] and later they were used in plant research [30]. RFLP is based on the variation(s) in the length of DNA fragments produced by a digestion of genomic DNAs and hybridization to specific markers of two or more individuals of a species is compared. RFLPs have been used extensively to compare genomes in the major cereal families such as rye, wheat, maize, sorghum, barley, and rice [31–33]. The advantages of RFLPs include detecting unlimited number of loci and being codominant, robust, and reliable and results are transferable across populations. However, RFLPs are highly expensive, time consuming, labour intensive, larger amounts of DNA required, limited polymorphism especially in closely related lines [34]. At present polymerase chain reaction- (PCR-) based marker systems are more rapid and require less plant material for DNA extraction. Rapid amplified polymorphic DNAs (RAPDs) were the first of PCR-based markers and are produced by PCR machines using genomic DNA and arbitrary (random) primers which act as both forward and backward primers in creation of multiple copies of DNA strands [35, 36]. The advantages of RAPDs include being quick and simple and inexpensive and the facts that multiple loci from a single primer are possible and a small amount of DNA is required. However, the results from RAPDs may not be reproduced in different laboratories and only can detect the dominant traits of interest [34]. Amplified fragment length polymorphisms (AFLPs) combine both PCR and RFLP [37]. AFLP is generated by digestion of PCR amplified fragments using specific restriction enzymes that cut DNA at or near specific recognition site in nucleotide sequence. AFLPs are highly reproducible and this enables rapid generation and high frequency of identifiable AFLPs, making it an attractive technique for identifying polymorphisms and for determining linkages by
analyzing individuals from a segregating population [37]. Another class of molecular markers which depends on the availability of short oligonucleotide repeats sequences in the genome of plants such as SSR, STS, SCAR, EST-SSR, and SNP. Many authors reviewed in detail different markers techniques [38, 39]. In this paper we are presenting the most widely used molecular markers and next generation sequencing technologies in detail in the following section.

6.2. Simple Sequence Repeat or Microsatellite. Microsatellites [40] are also known as simple sequence repeats (SSRs), short tandem repeats (STRs), or simple sequence length polymorphisms (SSLPs) which are short tandem repeats, their length being 1 to 10 bp. Some of the literature define microsatellites as 2–8 bp [41], 1–6 bp [42], or even 1–5 pb repeats [43]. SSRs are highly variable and evenly distributed throughout the genome and common in eukaryotes, their number of repeated units varying widely among crop species. The repeated sequence is often simple, consisting of two, three, or four nucleotides (di-, tri-, and tetrancleotide repeats, resp.). One common example of a microsatellite is a dinucleotide repeat (CA)n, where n refers to the total number of repeats that ranges between 10 and 100. These markers often present high levels of inter- and intraspecific polymorphism, particularly when tandem repeats number is 10 or greater [44]. PCR reactions for SSRs are performed in the presence of forward and reverse primers that anneal at the 5’ and 3’ ends of the template DNA, respectively. These polymorphisms are identified by constructing PCR primers for the DNA flanking the microsatellite region. The flanking regions tend to be conserved within the species, although sometimes they may also be conserved in higher taxonomic levels.

PCR fragments are usually separated on polyacrylamide gels in combination with AgNO3 staining, autoradiography, or fluorescent detection systems. Agarose gels (usually 3%) with ethidium bromide (EBr) can also be used when differences in allele size among samples are larger than 10 bp. However, the establishment of microsatellite primers from scratch for a new species presents a considerable technical challenge. Several protocols have been developed [43, 45–47] and details of the methodologies are reviewed by many authors [48–50]. The loci identified are usually multiallelic and codominant. Bands can be scored either in a codominant or as present or absent. The microsatellite-derived primers can often be used with many varieties and even other species because the flanking DNA is more likely to be conserved. These markers are evenly distributed throughout the genome, easily automated, and highly polymorphic and have good analytic resolution and high reproducibility making them a preferred choice of markers [51], most widely used for individual genotyping, germplasm evaluation, genetic diversity studies, genome mapping, and phylogenetic and evolutionary studies. However, the development of microsatellites requires extensive knowledge of DNA sequences, and sometimes they underestimate genetic structure measurements; hence they have been developed primarily for agricultural species, rather than wild species [39].

6.3. EST-SSRs. An alternative source of SSRs development is development of expressed sequence tag (EST-) based SSRs using EST databases has been utilized [52–58]. With the availability of large numbers of ESTs and other DNA sequence data, development of EST-based SSR markers through data mining has become fast, efficient, and relatively inexpensive compared with the development of genomic SSRs [59]. This is due to the fact that the time-consuming and expensive processes of generating genomic libraries and sequencing of large numbers of clones for finding the SSR containing DNA regions are not needed in this approach [60]. However, the development of EST-SSRs is limited to species for which this type of database exists as well as being reported to have lower rate of polymorphism compared to the SSR markers derived from genomic libraries [61–64].

6.4. Single Nucleotide Polymorphisms (SNPs). Single nucleotide polymorphisms (SNPs) are DNA sequence variations that occur when a single nucleotide (A, T, C, or G) in the genome sequence is changed, that is, single nucleotide variations in genome sequence of individuals of a population. These polymorphisms are single-base substitutions between sequences. SNPs occur more frequently than any other type of markers and are very near to or even within the gene of interest. SNPs are the most abundant in the genomes of the majority of organisms, including plants, and are widely dispersed throughout genomes with a variable distribution among species. SNPs can be identified by using either microarrays or DHPLC (denaturing high-performance liquid chromatography) machines. They are used for a wide range of purposes, including rapid identification of crop cultivars and construction of ultrahigh-density genetic maps. They provide valuable markers for the study of agronomic or adaptive traits in plant species, using strategies based on genetic mapping or association genetics studies.

6.5. Diversity Arrays Technology (DArT). A DArT marker is a segment of genomic DNA, the presence of which is polymorphic in a defined genomic representation. A DArT was developed to provide a practical and cost-effective whole genome fingerprinting tool. This method provides high throughput and low cost data production. It is independent from DNA sequence; that is, the discovery of polymorphic DArT markers and their scoring in subsequent analysis does not require any DNA sequence data. The detail of methodology for DArT is described by Jaccoud et al. [65] and Semagn et al. [38] as well as in website http://www.diversityarrays.com/.

To identify the polymorphic markers, a complexity reduct method is applied on the metagenome, a pool of genomes representing the germplasm of interest. The genomic representation obtained from this pool is then cloned and individual inserts are arrayed on a microarray resulting in a “discovery array.” Labelled genomic representations prepared from the individual genomes included in the pool are hybridized to the discovery array. Polymorphic clones (DArT markers) show variable hybridization signal intensities for different individuals. These clones are subsequently assembled into a “genotyping array” for routine genotyping. DArT is one of
the recently developed molecular techniques and it has been used in rice [66], wheat [38, 67, 68], barley [69], eucalyptus [70], Arabidopsis [71], cassava [72], pigeon-pea [73], and so forth.

DArT markers can be used as any other genetic marker. With DArT, comprehensive genome profiles are becoming affordable regardless of the molecular information available for the crop. DArT genome profiles are very useful for characterization of germplasm collections, QTL mapping, reliable and precise phenotyping, and so forth. However, DArT technique involves several steps, including preparation of genomic representation for the target species, cloning, data management, and analysis, requiring dedicated software such as DArTsoft and DArTdb. DArT markers are primarily dominant (present or absent) or differences in intensity, such as DArTsoft and DArTdb. DArT markers are primarily dominant (present or absent) or differences in intensity, which limits its value in some application [38].

7. Next Generation Sequencing

DNA sequencing is the determination of the order of the nucleotide bases, A (adenine), G (guanine), C (cytosine), and T (thymine), present in a target molecule of DNA. DNA sequencing technology has played a pivotal role in the advancement of molecular biology [74]. Next generation sequencing (NGS) or second generation sequencing technologies are revolutionizing the study of variation among individuals in a population. Most NGS technologies reduce the cost and time required for sequencing than Sanger-style sequencing machines (first generation sequencing). The following is the list of NGS technologies available at present, namely, the Roche/454 FLX, the Illumina/Solexa Genome Analyzer, the Applied Biosystems SOLiD System, the Helicos single-molecule sequencing, and pacific Biosciences SMRT instruments. These techniques have made it possible to conduct robust population-genetic studies based on complete genomes rather than just short sequences of a single gene.

The Roche/454 FLX, based on sequencing-by-synthesis with pyrophosphate chemistry, was developed by 454 Life Sciences and was the first next generation sequencing platform available on the market [75]. The Solexa sequencing platform was commercialized in 2006. The working principle is sequencing-by-synthesis chemistry. The Life Technologies SOLiD system is based on a sequencing-by-ligation technology. This platform has its origins in the system described by Shendure et al. [76] and in work by McKernan et al. [77] at Agencourt Personal Genomics (acquired by Applied Biosystems in 2006). Helicos single molecule sequencing (tSMS) technology is an entirely novel approach to DNA sequencing and genetic analysis and offers significant advantages over both traditional and “next generation” sequencing technologies. Helicos offers the first universal genetic analysis platform that does not require amplification. Pursuing a single molecule sequencing strategy simplifies the DNA sample preparation process, avoids PCR-induced bias and errors, simplifies data analysis, and tolerates degraded samples. Helicos single-molecule sequencing is often referred to as third generation sequencing. The detailed methodology, advantages, and disadvantages of each NGS technology were reviewed by many authors [78–81].

8. Analysis of Genetic Diversity from Molecular Data

It is essential to know the different ways that the data generated by molecular techniques can be analyzed before their application to diversity studies. Two main types of analysis are generally followed: (i) analysis of genetic relationships among samples and (ii) calculation of population genetics parameters (in particular diversity and its partitioning at different levels). The analysis of genetic relationships among samples starts with the construction of a matrix, sample × sample pair-wise genetic distance (or similarities). The advent and explorations of molecular genetics led to a better definition of Euclidean distance to mean a quantitative measure of genetic difference calculated between individuals, populations, or species at DNA sequence level or allele frequency level. Genetic distance and/or similarity between two genotypes, populations, or individuals may be calculated by various statistical measures depending on the data set. The commonly used measures of genetic distance (GD) or genetic similarity (GS) are (i) Nei and Li’s [82] coefficient (GDNL), (ii) Jaccard’s [83] coefficient (GDJ), (iii) simple matching coefficient (GDSM) [84], and (iv) modified Rogers’ distance (GDMR). Genetic distance determined by the above measures can be estimated as follows:

\[
GD_{NL} = 1 - \frac{2N_{11}}{(2N_{11} + N_{10} + N_{01})},
\]

\[
GD_{J} = 1 - \frac{N_{11}}{(N_{11} + N_{10} + N_{01})},
\]

\[
GD_{SM} = 1 - \frac{(N_{11} + N_{00})}{(N_{11} + N_{10} + N_{01} + N_{00})},
\]

\[
GD_{MR} = \left(\frac{N_{10} + N_{01}}{2N}\right)^{0.5},
\]

where \(N_{11}\) is the number of bands/alleles present in both individuals; \(N_{00}\) is number of bands/alleles absent in both individuals; \(N_{10}\) is the number of bands/alleles present only in the individual \(i\); \(N_{01}\) is the number of bands/alleles present only in the individual \(j\); and \(N\) represents the total number of bands/alleles. Readers are requested to read Mohammadi and Prasanna [85] review paper for more details about different GD measures.

There are two main ways of analyzing the resulting distance (or similarity) matrix, namely, principal coordinate analysis (PCA) and dendrogram (or clustering, tree diagram). PCA is used to produce a 2 or 3 dimensional scatter plot of the samples such that the distances among the samples in the plot reflect the genetic distances among them with a minimum of distortion. Another approach is to produce a dendrogram (or tree diagram), that is, grouping of samples together in clusters that are more genetically similar to each other than to samples in other clusters. Different algorithms were used for clustering, but some of the more widely used ones include unweighted pair group method with arithmetic
<table>
<thead>
<tr>
<th>Concept terms</th>
<th>Description/Features</th>
<th>Formulae/pros/cons</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Band-based approaches</strong></td>
<td>Easiest way to analyze and measure diversity by focusing on presence or absence of banding pattern.</td>
<td>Routinely use individual level. Totally relay on marker type and polymorphism</td>
</tr>
<tr>
<td><strong>(1) Measuring polymorphism</strong></td>
<td>Observing the total number of polymorphic bands (PB) and then calculating the percentage of polymorphic bands.</td>
<td>This “band informativeness” (I_b) can be represented on a scale ranging from 0 to 1 according to the formula: ( I_b = 1 - (2 \times</td>
</tr>
<tr>
<td><strong>(2) Shannon’s information index (I)</strong></td>
<td>It is called the Shannon index of phenotypic diversity and is widely applied.</td>
<td>( I = -\sum \pi_i \log_2 \pi_i ). These methods depend on the extraction of allelic frequencies.</td>
</tr>
<tr>
<td><strong>(3) Similarity coefficients</strong></td>
<td>Utilize similarity or dissimilarity (the inverse of the previous one) coefficients. The Jaccard coefficient ( J ) only takes into account the bands present in at least one of the two individuals. It is therefore unaffected by homoplastic absent bands (where the absence of the same band is due to different mutations). The simple-matching index (SM) maximizes the amount of information provided by the banding patterns considering all scored loci. The Nei and Li index (SD) doubles the weight for bands present in both individuals, thus giving more attention to similarity than dissimilarity.</td>
<td>( (i) ) Jaccard similarity coefficient or Jaccard index ( J = a/(a + b + c) ). ( (ii) ) Simple matching coefficient or index SM = ( (n - b - c)/n ). ( (iii) ) Sorensen-Dice index or Nei and Li index SD = ( 2a/2a + b + c ) where ( a ) is the number of bands (1s) shared by both individuals; ( b ) is the number of positions where individual ( i ) has a band, but ( j ) does not; ( c ) is the number of positions where individual ( j ) has a band, but ( i ) does not; and ( n ) is the total number of bands (0s and 1s).</td>
</tr>
<tr>
<td><strong>(4) Allele frequency based approaches</strong></td>
<td>Measure variability by describing changes in allele frequencies for a particular trait over time, more population oriented than band-based approaches.</td>
<td>These methods depend on the extraction of allelic frequencies from the data. The accurate estimates of frequencies essentially influence the results of different indices calculated for further measurements of genetic diversity.</td>
</tr>
<tr>
<td><strong>(5) Allelic diversity (A)</strong></td>
<td>Easiest ways to measure genetic diversity is to quantify the number of alleles present. Allelic diversity (A) is the average number of alleles per locus and is used to describe genetic diversity.</td>
<td>( A = \bar{n}_l/n_l ) where ( n_l ) is the total number of alleles over all loci; ( n_l ) is the number of loci. It is less sensitive to sample size and rare alleles and is calculated as ( n_l = 1/\sum p_i^2 ) ( p_i^2 ) ability; it provides information about the dispersal ability of the organism and the degree of isolation among populations.</td>
</tr>
<tr>
<td><strong>(6) Effective population size (N_e)</strong></td>
<td>It provides a measure of the rate of genetic drift, the rate of genetic diversity loss, and increase of inbreeding within a population.</td>
<td>Effective size of a population is an idealized number, since many calculations depend on the genetic parameters used and on the reference generation. Thus, a single population may have many different effective sizes which are biologically meaningful but distinct from each other.</td>
</tr>
<tr>
<td><strong>(7) Heterozygosity (H)</strong></td>
<td>There are two types of heterozygosity observed ( (H_o) ) and expected ( (H_e) ). The ( H_o ) is the portion of genes that are heterozygous in a population and ( H_e ) is estimated fraction of all individuals that would be heterozygous for any randomly chosen locus. Typically values for ( H_e ) and ( H_o ) range from 0 (no heterozygosity) to nearly 1 (a large number of equally frequent alleles). If ( H_o ) and ( H_e ) are similar (they do not differ significantly), mating in the populations is random. If ( H_o &lt; H_e ), the population is inbreeding; if ( H_o &gt; H_e ), the population has a mating system avoiding inbreeding.</td>
<td>Expected ( H_e ) is calculated based on the square root of the frequency of the null (recessive) allele as follows: ( H_e = 1 - \sum p_i^2 ) where ( p_i ) is the frequency of the ( i )th allele. ( H_o ) is calculated for each locus as the total number of heterozygotes divided by sample size.</td>
</tr>
</tbody>
</table>
averages (UPGMA), neighbour-joining method, and Ward’s method [86].

The molecular data can be scored in presence/absence matrices manually or with the aid of specific software. However, because these techniques are based on the incorporation of genomic elements in the primer sets or else target specific regions in the genome, biases affecting the evaluation process can occur. Although many recently developed targeting methods detect large numbers of polymorphisms, not many studies to date have utilized them, largely due to their unfamiliarity. In many cases the drawbacks are unknown. These mainly affect the analysis of the banding patterns produced, largely depending on the nature of the methods and whether they generate dominant or codominant markers. We presented a brief description of common/basic statistical approaches and its principle with the pros and cons of each method for measuring genetic diversity and it is given in Table 1. These are self-explanatory; therefore, the features and method of calculations were not much discussed separately in our text.

9. Assessment of Genetic Diversity in Postgenomic Era

Many software programs are available for assessing genetic diversity; however, most of them are freely available through source link to internet and corresponding institute web links are given in Table 2. In this section, we described some of the programs available which are mostly used in molecular diversity analyses in the postgenomic era (Table 2). Many of these perform similar tasks, with the main differences being in the user interface, type of data input and output, and platform. Thus, choosing which to use depends heavily on individual preferences.

10. Conclusion

Agriculturist has been realized that diverse plant genetic resources are priceless assets for humankind which cannot be lost. Such materials increasingly required to accessible for feeding a burgeoning world population in future (>9 billion in 2050). Presence of genetic variability in crops is essential for its further improvement by providing options for the breeders to develop new varieties and hybrids. This can be achieved through phenotypic and molecular characterization of PGR. Sometimes, large size of germplasm may limit their use in breeding. This may be overcome by developing and using subsets like core and minicore collection representing the diversity of the entire collection of the species. Molecular markers are indispensable tools for measuring the diversity of plant species. Low assay cost, affordable hardware, throughput, convenience, and ease of assay development and automation are important factors when choosing a technology. Now with the high throughput molecular marker technologies ensuring speed and quality of data generated, it is possible to characterize the larger number of germplasm with limited time and resources. Next generation sequencing reduced the cost and time required for sequencing the whole genome. Many software packages are available for assessing phenotypic and molecular diversity parameters that increased the efficiency of germplasm curators and, plant breeders to speed up the crop improvement. Therefore, we believe that this paper provides useful and contemporary information at one place; thus, it improves the understanding of tools for graduate students and also practical applicability to the researchers.

Table 1: Continued.

<table>
<thead>
<tr>
<th>Concept terms</th>
<th>Description/features</th>
<th>Formulae/pros/cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>(8) F-statistics</td>
<td>In population genetics the most widely applied measurements besides heterozygosity are $F$-statistics, or fixation indices, to measure the amount of allelic fixation by genetic drift. The $F$-statistics are related to heterozygosity and genetic drift. Since inbreeding increases the frequency of homozygotes, as a consequence, it decreases the frequency of heterozygotes and genetic diversity.</td>
<td>Three indexes can be calculated as follows: $F_{IT} = 1 - (H_I/H_T)$, $F_{IS} = 1 - (H_I/H_S)$, $F_{ST} = 1 - (H_S/H_T)$, where $H_I$ is the average $H_O$ within each population, $H_S$ is the average $H_E$ of subpopulations assuming random mating within each population, and $H_T$ is the $H_E$ of the total population assuming random mating within subpopulations and no divergence of allele frequencies among subpopulations.</td>
</tr>
</tbody>
</table>

Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFLP</td>
<td>Amplified fragment length polymorphism</td>
</tr>
<tr>
<td>AP-PCR</td>
<td>Arbitrarily primed PCR</td>
</tr>
<tr>
<td>ARMS</td>
<td>Amplification refractory mutation system</td>
</tr>
<tr>
<td>ASAP</td>
<td>Arbitrary signatures from amplification</td>
</tr>
<tr>
<td>ASH</td>
<td>Allele-specific hybridization</td>
</tr>
<tr>
<td>ASLP</td>
<td>Amplified sequence length polymorphism</td>
</tr>
<tr>
<td>ASO</td>
<td>Allele specific oligonucleotide</td>
</tr>
<tr>
<td>CAPS</td>
<td>Cleaved amplification polymorphic sequence</td>
</tr>
<tr>
<td>CAS</td>
<td>Coupled amplification and sequencing</td>
</tr>
<tr>
<td>DAF</td>
<td>DNA amplification fingerprint</td>
</tr>
<tr>
<td>DGGE</td>
<td>Denaturing gradient gel electrophoresis</td>
</tr>
<tr>
<td>GBA</td>
<td>Genetic bit analysis</td>
</tr>
<tr>
<td>IRAO</td>
<td>Interretrotransposon amplified polymorphism</td>
</tr>
<tr>
<td>ISSR</td>
<td>Intersimple sequence repeats</td>
</tr>
<tr>
<td>ISTR</td>
<td>Inverse sequence-tagged repeats</td>
</tr>
<tr>
<td>MP-PCR</td>
<td>Microsatellite-primed PCR</td>
</tr>
<tr>
<td>OLA</td>
<td>Oligonucleotide ligation assay</td>
</tr>
<tr>
<td>RAHM</td>
<td>Randomly amplified hybridizing microsatellites</td>
</tr>
<tr>
<td>RAMPs</td>
<td>Randomly amplified microsatellite polymorphisms</td>
</tr>
<tr>
<td>Analytical tools</td>
<td>Data type</td>
</tr>
<tr>
<td>------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>Arlequin</td>
<td>RFLPs, DNA sequences, SSR data, allele frequencies, or standard multilocus genotypes.</td>
</tr>
<tr>
<td>DnaSP</td>
<td>DNA sequence data</td>
</tr>
<tr>
<td>PowerMarker</td>
<td>SSR, SNP, and RFLP data</td>
</tr>
<tr>
<td>DARwin</td>
<td>Single data (for haploids, homozygote diploids, and dominant markers), allelic data, and sequence data</td>
</tr>
<tr>
<td>NTSYSpc</td>
<td>Single data (for haploids, homozygote diploids, and dominant markers), allelic data, and sequence data</td>
</tr>
<tr>
<td>MEGA</td>
<td>DNA sequence, protein sequence, evolutionary distance, or phylogenetic tree data</td>
</tr>
<tr>
<td>PAUP</td>
<td>Molecular sequences, morphological data, and other data types</td>
</tr>
<tr>
<td>Analytical tools</td>
<td>Data type</td>
</tr>
<tr>
<td>-----------------</td>
<td>-----------</td>
</tr>
<tr>
<td>STRUCTURE</td>
<td>All types of markers including mostly used markers like SSRs, SNPs, RFLPs, dArT, and so forth.</td>
</tr>
<tr>
<td>fastSTRUCTURE</td>
<td>SNP</td>
</tr>
<tr>
<td>ADMIXTURE</td>
<td>SNP</td>
</tr>
<tr>
<td>fineSTRUCTURE</td>
<td>Sequencing data</td>
</tr>
<tr>
<td>POPGENE</td>
<td>Use the dominant, codominant, and quantitative data for population genetic analysis</td>
</tr>
<tr>
<td>GENEPOP</td>
<td>Haploid or diploid data</td>
</tr>
<tr>
<td>GenAlEx</td>
<td>Codominant, haploid, and binary genetic data. It accommodates the full range of genetic markers available, including allozyms, SSRs, SNPs, AFLP, and other multifocus markers, as well as DNA sequences</td>
</tr>
</tbody>
</table>
RAPD: Randomly amplified polymorphic DNA
RBIP: Retrotransposon-based insertion polymorphism
REF: Restriction endonuclease fingerprinting
REMAP: Retrotransposon-microsatellite amplified polymorphism
RFLP: Restriction fragment length polymorphism
SAMPL: Selective amplification of polymorphic loci
SCAR: Sequence characterised amplification regions
SNP: Single nucleotide polymorphism
SPAR: Single primer amplification reaction
SPLIT: Single polymorphic amplification test
S-SAP: Sequence-specific amplification polymorphisms
SSCP: Single strand conformation polymorphism
SSL: Single sequence length polymorphism
SSR: Simple sequence repeats
STMS: Sequence-tagged microsatellite site
STS: Sequence-tagged site
TGGE: Thermal gradient gel electrophoresis
VNTR: Variable number tandem repeats
RAMS: Randomly amplified microsatellites.

Conflict of Interests
The authors declare that there is no conflict of interests regarding the publication of this paper.

References


Review Article

Unlimited Thirst for Genome Sequencing, Data Interpretation, and Database Usage in Genomic Era: The Road towards Fast-Track Crop Plant Improvement

Arun Prabhu Dhanapal1 and Mahalingam Govindaraj2

1Division of Plant Sciences, University of Missouri, Columbia, MO 65211, USA
2International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru 502324, India

Correspondence should be addressed to Arun Prabhu Dhanapal; dhanapala@missouri.edu

Received 17 July 2014; Revised 14 October 2014; Accepted 3 November 2014

Copyright © 2015 A. P. Dhanapal and M. Govindaraj. 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.

The number of sequenced crop genomes and associated genomic resources is growing rapidly with the advent of inexpensive next generation sequencing methods. Databases have become an integral part of all aspects of science research, including basic and applied plant and animal sciences. The importance of databases keeps increasing as the volume of datasets from direct and indirect genomics, as well as other omics approaches, keeps expanding in recent years. The databases and associated web portals provide at a minimum a uniform set of tools and automated analysis across a wide range of crop plant genomes. This paper reviews some basic terms and considerations in dealing with crop plant databases utilization in advancing genomic era. The utilization of databases for variation analysis with other comparative genomics tools, and data interpretation platforms are well described. The major focus of this review is to provide knowledge on platforms and databases for genome-based investigations of agriculturally important crop plants. The utilization of these databases in applied crop improvement program is still being achieved widely; otherwise, the end for sequencing is not far away.

1. Introduction

Most recent development of high-throughput methods for analyzing the structure and function of genes is collectively referred to as “genomics.” The comprehensive information of this kind is currently available for only a few plants and is rapidly being available for most of the higher plants and several underutilized crop plant species. Public access to this information will exploit biological selections and have direct impact on application of genomics to the improvement of economically important plants. Getting sequences of major plants on the one hand and access to all sequenced information for further applications on the other hand are most important. Therefore, global biological community should have open-access database for all plant genome sequenced so far.

Plant databases are facilities or long-lived record that are systematically updated with massive amount of data which has been generated as research outcomes in the context of the whole field of plant biology to ensure maximal accessibility and visibility to use by researchers in different fields of interest. These databases assist in drawing conclusion to make some new hypotheses to address basic questions of researchers. Internet-accessible information has become an integral part of most scientific enterprise, including the plant sciences. It now seems that it is impossible to conceive of future significant progress being made without the internet and the databases and many other similar resources the internet makes openly available. This is particularly true as the information flows from genomics and other high-throughput technologies to all aspects of crop plant sciences. The ultimate goal of plant genomics is to improve our ability
to identify the genotypes with optimal agronomic traits in order to improve yield, a must with the increasing world population [1].

2. Omics Research on Crop Plants: Present Status

"Omnics" refers to the collective technologies that are made available in recent years which are used to explore the roles, relationships, and actions of the various types of molecules that make up the cells of a living organism. The "omics" technology includes genomics (the study of genes and their function), proteomics (the study of proteins), metabolomics (the study of molecules involved in cellular metabolism), transcriptomics (the study of the mRNA), glycomics (the study of cellular carbohydrates), and lipomics (the study of cellular lipids). These omics technologies provide the tools needed to look at the differences in DNA, RNA, proteins, and other cellular molecules between species and among individuals of the same or different species. A combinatorial approach using multiple omics platforms and integration of their outcomes is now an effective strategy for clarifying molecular systems integral to improving crop plant productivity (Figure 1). Recent progress in plant genomics and utilization of genetic resources has allowed us to discover and isolate important genes and analyze their functions that regulate yields as well as stress tolerance [2].

A technological advance in omics research integrating animal and plant science has become essential resources for the investigation of gene function in association with phenotypic changes. Some of these advances include the development of high-throughput methods for profiling expressions of thousands of genes, for identifying modification events and interactions in the plant proteome and for measuring the abundance of many metabolites simultaneously. In addition, large-scale collections of bioresources, such as mass-produced mutant lines and clones of full-length cDNAs and their integrative relevant databases, are now made available [3, 4]. The importance of crop plant genetic resources and insights that have been emerged in recent years through genomics are well reviewed [5, 6]. The recent high-throughput technological advances have provided opportunities to develop collections of sequence-based resources and other related resource platforms for specific organisms. Various bioinformatics platforms have become essential tools for accessing omics dataset for the efficient mining and integration of biologically significant knowledge to deposit in databases for public access (Figure 1).

3. Crop Plant Genome Sequence Resources

In recent years, many crop plant genomes have been sequenced and data is available to public (Table 1). On the other hand, collected sequence data provide essential genomic resources for accelerating molecular understanding of biological properties and for promoting the application of such knowledge to the benefit of humans. The recent accumulation of nucleotide sequences of model plants and other crop species has provided fundamental information for the design of sequence-based research applications in functional genomics. Species-specific nucleotide sequence collections also provide opportunities to identify the genomic aspects of phenotypic characters based on genome-wide comparative analyses and knowledge of model organisms [46].

3.1. Rationale of Genome Sequencing Projects. Recent revolution in DNA sequencing technology has brought down the cost of DNA sequencing of several crop plant species and made the sequencing of an increased number of genomes both feasible and cost effective [46]. The first plant genome Arabidopsis was completely sequenced in December 2000, and it was the third complete genome of a higher eukaryote and further studies were carried out in recent years on Arabidopsis thaliana and Arabidopsis lyrata [30, 31]. Subsequently, after Arabidopsis, several other crop plants have been sequenced (Table 1). These genomes reveal numerous species-specific details, including genome size, gene number, patterns of sequence duplication, a catalog of transposable elements, and syntenic relationships. To understand
<table>
<thead>
<tr>
<th>Name of crop plant</th>
<th>Consortium/initiative</th>
<th>URL</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barley (<em>Hordeum vulgare</em>)</td>
<td>International Barley Genome Sequencing Consortium</td>
<td><a href="http://www.public.iastate.edu/~imagefpc/IBSC%20Webpage/IBSC%20Template-home.html">http://www.public.iastate.edu/~imagefpc/IBSC%20Webpage/IBSC%20Template-home.html</a></td>
<td>The International Barley Genome Sequencing Consortium [10]</td>
</tr>
<tr>
<td>Cannabis (<em>Cannabis sativa</em>)</td>
<td>Consortium</td>
<td><a href="http://genome.ccb.iuttoronto.ca/index.html?org=C.+sativa&amp;db=canSat3&amp;hgsid=I252">http://genome.ccb.iuttoronto.ca/index.html?org=C.+sativa&amp;db=canSat3&amp;hgsid=I252</a></td>
<td>van Bakel et al., 2011 [12]</td>
</tr>
<tr>
<td>Castor bean (<em>Ricinus communis</em>)</td>
<td>TIGR</td>
<td><a href="http://castorbean.jcvi.org/">http://castorbean.jcvi.org/</a></td>
<td>Chan et al., 2010 [13]</td>
</tr>
<tr>
<td>Chickpea (<em>Cicer arietinum</em>)</td>
<td>Consortium (ICRISAT-BGI)</td>
<td><a href="http://www.icrisat.org/gt-bt/ICGGC/GenomeManuscript.htm">http://www.icrisat.org/gt-bt/ICGGC/GenomeManuscript.htm</a></td>
<td>Varshney et al., 2013 [14]</td>
</tr>
<tr>
<td>Crucifer (<em>Thellungiella parvula</em>)</td>
<td>Consortium</td>
<td><a href="http://www.brassica.info/info/events.php">http://www.brassica.info/info/events.php</a></td>
<td>Dassanayake et al., 2011 [17]</td>
</tr>
<tr>
<td>Cucumber (<em>Cucumis sativus</em>)</td>
<td>International Cucurbit Genomics Initiative (ICuGI)</td>
<td><a href="http://www.icugi.org/">http://www.icugi.org/</a></td>
<td>Huang et al., 2009 [18]</td>
</tr>
<tr>
<td>Date palm (<em>Phoenix dactylifera</em>)</td>
<td>Consortium</td>
<td><a href="http://qatar-weill.cornell.edu/research/datepalmGenome/download.html">http://qatar-weill.cornell.edu/research/datepalmGenome/download.html</a></td>
<td>Al-Dous et al., 2011 [19]</td>
</tr>
<tr>
<td>Grape (<em>Vitis vinifera</em>)</td>
<td>Consortium</td>
<td><a href="http://www.genoscope.cns.fr/externe/GenomeBrowser/Vitis/">http://www.genoscope.cns.fr/externe/GenomeBrowser/Vitis/</a></td>
<td>Jaillon et al., 2007 [23]</td>
</tr>
<tr>
<td>Maize (<em>Zea mays</em>)</td>
<td>Consortium</td>
<td><a href="http://www.maizegdb.org/">http://www.maizegdb.org/</a></td>
<td>Schnable et al., 2009 [26]</td>
</tr>
<tr>
<td>Mosses (<em>Physcomitrella patens</em>)</td>
<td>JGI</td>
<td><a href="http://genome.jgi-psf.org/PhyspaL1/PhyspaL1.home.html">http://genome.jgi-psf.org/PhyspaL1/PhyspaL1.home.html</a></td>
<td>Rensing et al., 2008 [28]</td>
</tr>
<tr>
<td>Papaya (<em>Carica papaya</em>)</td>
<td>Consortium</td>
<td><a href="http://www.plantgdb.org/CpGDB/">http://www.plantgdb.org/CpGDB/</a></td>
<td>Ming et al., 2008 [32]</td>
</tr>
<tr>
<td>Peach (<em>Prunus persica</em>)</td>
<td>International Peach Genome Initiative</td>
<td><a href="http://www.rosaceae.org/peach/genome">http://www.rosaceae.org/peach/genome</a></td>
<td>International Peach genome initiative 2013 [33]</td>
</tr>
<tr>
<td>Name of crop plant</td>
<td>Consortium/initiative</td>
<td>URL</td>
<td>References</td>
</tr>
<tr>
<td>--------------------</td>
<td>-----------------------</td>
<td>-----</td>
<td>------------</td>
</tr>
<tr>
<td>Pigeon pea (Cajanus cajan)</td>
<td>International Initiative for Pigeonpea Genomics (IIPG)</td>
<td><a href="http://www.icrisat.org/gt-bt/iipg/Home.html">http://www.icrisat.org/gt-bt/iipg/Home.html</a></td>
<td>Varshney et al., 2011 [34]</td>
</tr>
<tr>
<td>Poplar (Populus trichocarpa)</td>
<td>JGI</td>
<td><a href="http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html">http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html</a></td>
<td>Tuskan et al., 2006 [35]</td>
</tr>
<tr>
<td>Rape seed (Brassica napus)</td>
<td>Consortium (MGBP)</td>
<td><a href="http://www.plantgdb.org/BrGDB/">http://www.plantgdb.org/BrGDB/</a></td>
<td>Wang et al., 2011 [37]</td>
</tr>
<tr>
<td>Rice (Oryza sativa ssp. indica and japonica)</td>
<td>Consortium (IRGSP)</td>
<td><a href="http://rgp.dna.affrc.go.jp/E/IRGSP/index.html">http://rgp.dna.affrc.go.jp/E/IRGSP/index.html</a></td>
<td>Yu et al., 2002 [38]; Goff et al., 2002 [39]</td>
</tr>
<tr>
<td>Sorghum (Sorghum bicolor)</td>
<td>JGI</td>
<td><a href="http://www.plantgdb.org/SbGDB/">http://www.plantgdb.org/SbGDB/</a></td>
<td>Paterson et al., 2009 [40]</td>
</tr>
<tr>
<td>Soybeans (Glycine max)</td>
<td>JGI</td>
<td><a href="http://www.phytozone.net/soybean_er.php">http://www.phytozone.net/soybean_er.php</a></td>
<td>Schmutz et al., 2010 [41]</td>
</tr>
<tr>
<td>Strawberry (Fragaria vesca)</td>
<td>Consortium</td>
<td><a href="http://www.strawberrygenome.org/">http://www.strawberrygenome.org/</a></td>
<td>Shulaev et al., 2011 [42]</td>
</tr>
<tr>
<td>Tomato (Solanum lycopersicum)</td>
<td>Consortium</td>
<td><a href="http://solgenomics.net/organism/Solanum_lycopersicum/genome">http://solgenomics.net/organism/Solanum_lycopersicum/genome</a></td>
<td>The Tomato Genome Consortium 2012 [43]</td>
</tr>
<tr>
<td>Watermelon (Citrullus lanatus)</td>
<td>International Watermelon Genomics Initiative</td>
<td><a href="http://www.iwgi.org/">http://www.iwgi.org/</a></td>
<td>Guo et al., 2013 [44]</td>
</tr>
</tbody>
</table>
the complex instructions contained in all these raw sequence information of the plant genome, large-scale functional genomics projects are required. Progress towards a complete understanding of gene regulatory networks shared among many crop plants is important for improving cultivated species and for complete understanding of crop plant evolution.

3.2. Contribution of Whole-Genome Resequencing. Advancement in next-generation sequencing (NGS) technology coupled with many reference genomes sequence data allows us to discover variations among many crop plants. A whole-genome resequencing project to discover whole-genome sequence variations in 1,001 strains (accessions) of Arabidopsis resulted in dataset that became a fundamental resource for promoting future genetics studies to identify alleles in association with phenotypic diversity across the entire genome and across the entire crop plant species (http://1001genomes.org/) [47, 48]. In rice, a high-throughput method for genotyping recombinant populations that used whole-genome resequencing data generated by the Illumina Genome Analyzer was performed [18] and recently resequencing of 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes has been completed [49].

3.3. Analyzing Crop Plant Genome Sequences. Galaxy (http://galaxyproject.org) is a software system that provides knowledge and support through a framework that provides researchers with simple interfaces to powerful data interpretation tools. Galaxy is web-based framework designed for use of experimental and computational biologists in all fields of biological science. With Galaxy, one can easily use analysis tools through a web-based interface [50]. Another tool made available from the Sanger institute (http://www.sanger.ac.uk/) is Artemis, a free genome browser and annotation tool that allows visualization of sequence features, next generation data, and the results of analyses [51]. The Broad’s Genome Sequencing and Analysis Program (GSAP) plays a major role in providing several analyses tools for genome sequences coming out of the NGS platforms in all biological fields (http://www.broadinstitute.org/).

4. Crop Plant Genome Resources and Variation Analysis

Genome-wide study of both structural and gene content variation are hypothesized to drive important phenotypic variation within a crop plant species. Previous studies have shown that both structural and gene content variations were assessed in several crops using array hybridization and targeted resequencing. Genetic variation within and between species is most commonly quantified by single nucleotide polymorphisms (SNPs). There has been increased interest in recent years to resolve genetic differences in terms of structural variation (SV), which includes copy number variation (CNV) caused by large insertions and deletions, and other types of rearrangements such as inversions and translocations. CNV together with SV is thought to be an important factor in determining phenotypic variation for a wide range of traits reviewed [52] in both crop plant and animal species.

4.1. Molecular Breeding Tools

4.1.1. Role of Molecular Markers. Among various DNA markers available to research community, single sequence repeats (SSRs) and single nucleotide polymorphisms (SNPs) are most widely used today. SSRs are demonstrated to be of high degree of transferability between species and could easily be transferred to related species to amplify the same corresponding locus. SNPs represent the most frequent type of genetic polymorphism and may therefore provide a high density of markers near a locus of interest compared to SSRs. The high density of SNPs makes them valuable for genome mapping, and in particular they allow the generation of ultra-high density genetic maps and haplotyping systems for genes or regions of interest and map-based positional cloning in crop plants. SNPs are used routinely in crop breeding programs, for genetic diversity analysis, cultivar identification, phylogenetic analysis, characterization of genetic resources, and association with agronomic and physiological traits in both cereals and legumes [53, 54]. Application of SNP markers for genetic dissection of complex traits like delta 13C and delta 15N in legume like soybean with high density SNP chips has also increased and been made available [55–57].

4.1.2. Biparental QTL Mapping. The quantitative traits loci (QTL) identified for a trait of interest that contribute to higher phenotypic variation are considered major QTL. These identified QTLs, after validation in desired germplasm, can be used for introgression of the trait from the donor genotypes (generally used for identification of the QTL for the trait) into elite cultivars to traits of less phenotypic variation cultivars or breeding lines (recipient parents) without transfer of undesirable genes from the donors (linkage drag). The process is commonly referred to as marker-assisted backcrossing (MABC) most commonly employed by plant breeders. Superior lines or cultivars are developed which contain only the major QTL from the donor parent while retaining the whole-genome of the recurrent parent [58]. MABC has been used extensively for introgression of resistance to abiotic stresses and abiotic stress in crop plants. To overcome the limitations of MABC, particularly when multiple QTLs control the expression of a complex trait, the MARS approach, which involves intermating selected individuals in each selection cycle, has been recommended [59, 60]. It generally involves the use of an F2 base population and can be used in self-pollinated crops like wheat, barley, and chickpea for developing pure lines with superior per se performance (for more details, see [60]). MARS has the additional advantage of overcoming the limitation of inadequate improvement in the frequency of superior alleles in F2 enrichment since MAS is practiced in each cycle following intermating to improve the frequency of favourable alleles [59].

4.1.3. Genome-Wide Association Analysis. Genome-wide association analysis (GWAS) is a powerful approach to
identify the causal genetic polymorphisms underlying both simple and complex traits in crop plants. Advancement in genomics has provided alternative tools to improve breeding efficiency in plant breeding programs. Molecular markers linked to the causal genes and/or QTLs can be used for marker-assisted selection (MAS) [61]. Recent advances in genome sequencing and single nucleotide polymorphism (SNP) genotyping have increased the applicability of association analysis for QTL mapping in crop plants [62, 63]. Genome-wide association analyses with SNP markers have been conducted for several important traits in many plant species, including Arabidopsis thaliana [64], maize [65], rice [66], and soybean [67–69], and also in tree crops like peach [70].

4.1.4. Genomic Selection. Genomic selection (GS) is more reliable and relatively simple and most powerful approaches used in crop plant species where breeding values of the genotype/cultivar lines are predicted using their marker genotypes and phenotypes [71]. GS captures the small QTL effect that governs the variation including epistatic interaction effects. GS has been successfully used in wheat, maize, and soybean [71–73]. The accuracy of GS depends on genetic × environmental (G × E) interaction and major challenge of GS is to arrive with the accurate genomic estimated breeding values (GEBVs) with respect to the G × E interaction. Application of GS has been extended to other crops plants like Arabidopsis, sugarcane, and sugar beet in recent years.

4.2. Application of Molecular Platforms for Variation Analysis. High-throughput polymorphism analysis is an essential tool for facilitating any genetic map-based approach, and the number of platforms has been developed and applied to genetic map construction, marker-assisted selection, and QTL cloning using multiple segregation populations in major crop plants. These types of genotyping systems have been successfully used in postgenome sequencing era with extending of their projects on genotyping of genetic resources, identifying their population structure, and association of their phenotypic values to identify their genomic regions. This recent expansion of analysis platforms provides an essential resource in the “variome” study of crop plants. The increasing demand for high-throughput and cost-effective platforms for comprehensive variation analysis (also called variome analysis) has rapidly increased. Whole-genome resequencing approaches are already being realized as a direct solution for variome analysis in species whose reference genome sequence data are available [74, 75].

 Diversity Array Technology (DArT) is a high-throughput genotyping system developed based on a microarray platform (http://www.diversityarrays.com/index.html) [76]. In various crop species such as wheat, barley, and sorghum, DArT markers have been used together with conventional molecular markers to construct denser genetic maps and perform association studies [77–79]. The Illumina GoldenGate assay allows the simultaneous analysis of up to 1,536 SNPs in 96 samples and has been used to analyze genotypes of segregation populations in order to construct genetic maps allocating SNP markers in crops such as barley, wheat, soybean [80–82], and peach [70, 83]. Recently 3K to 700K Infinium i Select HD and HTS custom genotyping bead chips are made available for the high-throughput genotyping of SNPs, indels, and CNVs.

4.3. Databases for Variation Analysis. Characterizing the genetic basis of variation in crop plants and linking to observable traits will provide an important framework for understanding evolutionary patterns and population structure and could specially increase the efficiency of selection made in the crop plant breeding programmes.

GRAMENE. The Genetic Diversity Database in GRAMENE specializes in storage of genotypes, phenotypes and their environments, germplasm, and association data. Genomic Diversity and Phenotype Data Model (GDPM) database schema which efficiently stores anything from small-scale SSR diversity studies to large-scale SNP/indel-based genotype-phenotype studies with billions of allele calls [84].

The Plant Variation Mart Database. It holds a catalogue of DNA variants for single nucleotide polymorphisms (SNPs) and insertions/deletions (indels) for Arabidopsis, rice, and grapes.

5. Crop Plant Comparative Genomics Resources

The number of sequenced crop plant genomes and their associated genomic resources is growing rapidly with the advent of increased focus on crop plant genomics from funding agencies and other NGS technologies. Among several comparative genomics platform available today, Phytozome, a comparative hub to plant genome and gene family data and analysis, provides a view of the evolutionary history of every plant gene at the level of sequence, gene structure, gene family, and genome organization. Through their comprehensive plant genome database and web portal, these data are available to the broader plant science research community, providing powerful comparative genomics tools that help link model systems with other plants of economic and ecological importance. A number of information resources to plant genomics accessible on the web have appeared, along with appropriate analytical tools. The integrative databases promoting plant comparative genomics and URLs of each integrative database in plant genomics are shown (Table 2).

5.1. Crop Plant Comparative Genomics Databases. Several plant traits, namely, anatomical, morphological, biochemical, and physiological features of individuals or their component organs or tissues, serve as the key to understanding and predicting the adaptation of ecosystems in the face of biodiversity loss and global change. The reduced genome sequencing cost is opening up significant opportunities for crop improvement through plant breeding and increased understanding of plant biology. Many crop plant genomes are large and have complex evolutionary histories, making
Table 2: List of databases and their information of application to crop plant research community.

<table>
<thead>
<tr>
<th>Name of databases</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>AgBase—a unified resource for functional analysis in agriculture</td>
<td>Search and analyze functional genomics datasets in agricultural species</td>
</tr>
<tr>
<td>AutoSNPdb—an annotated single nucleotide polymorphism database for crop plants</td>
<td>Identify SNPs from assembled EST sequences for the crops rice, barley, and Brassica</td>
</tr>
<tr>
<td>BarleyBase—an expression profiling database for plant genomics</td>
<td>Analyze and visualize plant microarray data</td>
</tr>
<tr>
<td>BBGD—an online database for blueberry genomic data.</td>
<td>It stores both EST and microarray data and allows scientists to correlate expression profiles with gene function</td>
</tr>
<tr>
<td>BIOGEN BASE—CASSAVA</td>
<td>A web accessible resource for investigating cassava phenomics and genomics information</td>
</tr>
<tr>
<td>CastorDB—a comprehensive knowledge base for Ricinus communis.</td>
<td>CastorDB provides a user friendly comprehensive resource on castor with particular emphasis on its genome, transcriptome, and proteome and on protein domains, pathways, protein localization, presence of sumoylation sites, expression data, and protein interacting partners</td>
</tr>
<tr>
<td>ChromDB—The Chromatin Database</td>
<td>Locate chromatin-associated proteins, including RNAi-associated proteins, for a broad range of organisms</td>
</tr>
<tr>
<td>CR-EST—a resource for crop ESTs</td>
<td>Search for sequence, classification, clustering, and annotation data of crop EST projects</td>
</tr>
<tr>
<td>CSRDB—a small RNA integrated database and browser resource for cereals</td>
<td>Search for sequence information on rice, maize, and other cereal crops small RNAs</td>
</tr>
<tr>
<td>DEBDOM—Database Exploring Banana Diversity of Manipur</td>
<td>The database DEBDOM provides a sophisticated web base access to the details of the taxonomy, morphological characteristics, utility, and sites of collection of Musa genotypes</td>
</tr>
<tr>
<td>DRASTIC—Database Resource for the Analysis of Signal Transduction in Cells</td>
<td>Search for information of plant gene expression in response to pathogens and environmental changes</td>
</tr>
<tr>
<td>FLAGdb++—A Database for the Functional Analysis of the Arabidopsis Genome</td>
<td>Search and visualize data for high-throughput functional analysis of Arabidopsis, rice, and other plant genomes</td>
</tr>
<tr>
<td>GabiPD—a plant integrative “omics” database</td>
<td>Search for comprehensive and extensive information on various plant genomes generated by a German collaborative network of plant genomics research</td>
</tr>
<tr>
<td>GCP—The Generation Challenge Programme</td>
<td>An online resource documenting stress-responsive genes comparatively across plant species</td>
</tr>
<tr>
<td>GDR—Genome Database for Rosaceae</td>
<td>A central repository of curated and integrated genetics and genomics data of Rosaceae, which includes apple, cherry, peach, pear, raspberry, rose, and strawberry</td>
</tr>
<tr>
<td>GeneCAT—gene co-expression analysis toolbox</td>
<td>Novel web tools that combine BLAST and coexpression analyses</td>
</tr>
<tr>
<td>GeneSeqer@PlantGDB—gene structure prediction in plant genomes</td>
<td>Predict gene structures of plant genomes</td>
</tr>
<tr>
<td>GERMINATE</td>
<td>A generic database for integrating genotypic and phenotypic information for plant genetic resource collections</td>
</tr>
<tr>
<td>GGT—Graphical GenoTypes</td>
<td>Software for visualization and analysis of genetic data</td>
</tr>
<tr>
<td>GrainGenes—The genome database for small-grain crops</td>
<td>Search for molecular and phenotypic information on wheat, barley, rye, triticale, and oats</td>
</tr>
<tr>
<td>Gramene—a resource for comparative grass genomics</td>
<td>Curated resource for genetic, genomic, and comparative genomics data for the major crop species, including rice, maize, wheat, and many other plant (mainly grass) species</td>
</tr>
<tr>
<td>MaizeGDB—the Community Database for Maize Genetics and Genomics</td>
<td>Search genetic and genomic information about maize</td>
</tr>
</tbody>
</table>
### Table 2: Continued.

<table>
<thead>
<tr>
<th>Name of databases</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>MANET—The Molecular Ancestry Network</td>
<td>Tracing evolution of protein architecture in metabolic networks</td>
</tr>
<tr>
<td>Medicago—a database for personalized data mining of the model legume Medicago truncatula transcriptome</td>
<td>Search for integrated genomic, genetic, and biological information on cool season legume <em>Medicago truncatula</em> (<em>Mt</em>)</td>
</tr>
<tr>
<td>MetaCrop—a detailed database of crop plant metabolism</td>
<td>A database that summarizes diverse information about metabolic pathways in crop plants and allows automatic export of information for the creation of detailed metabolic models</td>
</tr>
<tr>
<td>MetaCrop 2.0—managing and exploring information about crop plant metabolism.</td>
<td>It contains information about seven major crop plants with high agronomical importance and two model plants; MetaCrop is intended to support research aimed at the improvement of crops for both nutrition and industrial use</td>
</tr>
<tr>
<td>Narcisse—a mirror view of conserved syntenies</td>
<td>A database dedicated to the study of genome conservation</td>
</tr>
<tr>
<td>NIASGDbdb—National Institute of Agrobiological Sciences Gene Bank Data Base</td>
<td>Find information about agricultural plant genetics and diseases</td>
</tr>
<tr>
<td>P3DB—Plant Protein Phosphorylation Database</td>
<td>Find information about protein phosphorylation in plants</td>
</tr>
<tr>
<td>Panzea—a database and resource for molecular and functional diversity in the maize genome</td>
<td>Search for information on relationship between genotype and functional phenotype variations</td>
</tr>
<tr>
<td>Pepper EST database—in silico exploitation of EST data to extensivley score genes of Capsicum annuum</td>
<td>Comprehensive in silico tool for analyzing the chili pepper (<em>Capsicum annuum</em>) transcriptome</td>
</tr>
<tr>
<td>PIP—a database of potential intron polymorphism markers</td>
<td>A database of potential intron polymorphism markers in plants</td>
</tr>
<tr>
<td>PLACE—plant cis-acting regulatory DNA elements</td>
<td>Search for documented motifs found in plant cis-acting regulatory DNA elements</td>
</tr>
<tr>
<td>Plant snoRNA database</td>
<td>Search for comprehensive information on small nucleolar RNAs in plants</td>
</tr>
<tr>
<td>PlantCARE—a database of plant cis-acting elements</td>
<td>Search for information on plant cis-acting regulatory elements, transcription sites, enhancers, and repressors</td>
</tr>
<tr>
<td>PlantTFDB—Plant Transcription Factor Databases</td>
<td>A comprehensive plant transcription factor database</td>
</tr>
<tr>
<td>PlantTribes—a gene and gene family resource for comparative genomics in plants</td>
<td>A plant gene family database based on the inferred proteomes of five sequenced plant species: <em>Arabidopsis thaliana</em>, <em>Carica papaya</em>, <em>Medicago truncatula</em>, <em>Oryza sativa</em>, and <em>Populus trichocarpa</em></td>
</tr>
<tr>
<td>PLecDom—Plant Lectin Domains server</td>
<td>Find information about plant lectin domains</td>
</tr>
<tr>
<td>PhtTFDB—Plant Transcription Factor Database</td>
<td>Find information about transcription factors in plants</td>
</tr>
<tr>
<td>PmiRKB—Plant MicroRNA Knowledge Base</td>
<td>Find information about plant microRNAs</td>
</tr>
<tr>
<td>PMRD—Plant MicroRNA Database</td>
<td>Find information about microRNA sequences and targets in plants</td>
</tr>
<tr>
<td>PODB—the Plant Organelles Database</td>
<td>Search a collection of visualized plant organelles and protocols for plant organelle research</td>
</tr>
<tr>
<td>POGs/PlantRBP—a resource for comparative genomics in plants</td>
<td>Search for information on putative orthologous proteins among rice, maize, and <em>Arabidopsis</em> with emphasis on RNA-binding proteins</td>
</tr>
<tr>
<td>PoMaMo—a comprehensive database for potato genome data</td>
<td>Search for comprehensive genomic information on potato</td>
</tr>
<tr>
<td>PREP Suite—Predictive RNA Editor for Plants</td>
<td>Use to predict sites of RNA editing in plants</td>
</tr>
<tr>
<td>PRGBD—Plant Resistance Genes DataBase</td>
<td>Find information about genes involved in plant defense mechanisms</td>
</tr>
<tr>
<td>psRNAMiner—a plant short small RNA regulatory cascade analysis server</td>
<td>Identify both the clusters of phased small RNAs as well as the potential phase-initiator</td>
</tr>
<tr>
<td>RadishBase—a database for genomics and genetics of radish.</td>
<td>A database containing radish pathways predicted from unigene sequences is also included in RadishBase</td>
</tr>
<tr>
<td>RoBuST—an integrated genomics resource for the root and bulb crop families Apiaceae and Alliaceae.</td>
<td>The RoBuST database has been developed to initiate a platform for collecting and organizing genomic information useful for RBV (root and bulb vegetables) researchers</td>
</tr>
<tr>
<td>Name of databases</td>
<td>Application</td>
</tr>
<tr>
<td>-------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>SALAD—Surveyed contained motif Alignment diagram and the Associating Dendrogram</td>
<td>Perform systematic comparison of proteome data among species</td>
</tr>
<tr>
<td>SGN—SOL Genomics Network</td>
<td>A comparative map viewer dedicated to the biology of the Solanaceae family</td>
</tr>
<tr>
<td>Shanghai RAPESEED Database—a resource for functional genomics studies of seed development and fatty acid metabolism of <em>Brassica</em></td>
<td>Find information on EST, gene expression profiles, and bioresources for the promotion of functional genomics studies and quality breeding of <em>Brassica</em> crops</td>
</tr>
<tr>
<td>SolRgene—an online database to explore disease resistance genes in tuber-bearing Solanum species</td>
<td>The SolRgene database contains data on resistance to <em>P. infestans</em> and presence of R genes and R gene homologues in Solanum section Petota</td>
</tr>
<tr>
<td>Shanghai RAPESEED Database—a resource for functional genomics studies of seed development and fatty acid metabolism of <em>Brassica</em></td>
<td>Find information on EST, gene expression profiles, and bioresources for the promotion of functional genomics studies and quality breeding of <em>Brassica</em> crops</td>
</tr>
<tr>
<td>SoyBase—USDA-ARS soybean genetics and genomics database</td>
<td>Find genetic information about soybeans</td>
</tr>
<tr>
<td>SoyTEdb—a comprehensive database of transposable elements in the soybean genome.</td>
<td>SoyTEdb provides resources and information related to transposable elements in the soybean genome, representing the most comprehensive and the largest manually curated transposable element database for any individual plant genome completely sequenced to date</td>
</tr>
<tr>
<td>SoyXpress—a database for exploring the soybean transcriptome</td>
<td>A soybean gene expression and transcription database</td>
</tr>
<tr>
<td>Sputnik—a database platform for comparative plant genomics</td>
<td>Search for ESTs from over 20 different plant species</td>
</tr>
<tr>
<td>TFGD—Tomato Functional Genomics Database</td>
<td>Find information about tomato genes</td>
</tr>
<tr>
<td>The Adaptive Evolution Database (TAED)—a phylogeny based tool for comparative genomics</td>
<td>Search for information on adaptive evolution in gene families of higher plants and chordate</td>
</tr>
<tr>
<td>The Legume Information System (LIS)—an integrated information resource for comparative legume biology</td>
<td>Search for integrated genetic and molecular data from multiple legume species</td>
</tr>
<tr>
<td>The Plant DNA C-values Database</td>
<td>Search for information on plant DNA C-values and genome sizes</td>
</tr>
<tr>
<td>The Plant Ontology Database—a resource for plant structure and developmental stages</td>
<td>View, search, and query plant ontology terms</td>
</tr>
<tr>
<td>The PlantsP Functional Genomics Database</td>
<td>Search for information on plant kinases and phosphatases</td>
</tr>
<tr>
<td>The TIGR Maize Database</td>
<td>Search for annotated genomic sequences of maize</td>
</tr>
<tr>
<td>The TIGR Plant Repeat Databases—A Collective Resource for the Identification of Repetitive Sequences in Plants</td>
<td>Identify, classify, and analyze repetitive sequences in plant genomes</td>
</tr>
<tr>
<td>TomatEST database—in silico exploitation of EST data to explore expression patterns in tomato species</td>
<td>Find expressed sequence tag (EST)/cDNA sequence information from different libraries of multiple tomato species</td>
</tr>
<tr>
<td>TriMEDB—a database to integrate transcribed markers and facilitate genetic studies of the tribe Triticeae.</td>
<td>The Triticeae mapped expressed sequence tag (EST) database</td>
</tr>
<tr>
<td>TropGENE-DB—A Multi-tropical Crop Information System</td>
<td>Search for genetic, molecular, and phenotypic data of tropical crop species</td>
</tr>
<tr>
<td>TropGENE-DB—A Multi-tropical Crop Information System</td>
<td>Search for genetic, molecular, and phenotypic data of tropical crop species</td>
</tr>
<tr>
<td>UK CropNet—a collection of databases and bioinformatics resources for crop plant genomics</td>
<td>Search sequences and genomic information on crop plants</td>
</tr>
<tr>
<td>WhETS—Wheat Estimated Transcript Server</td>
<td>A tool to provide the best estimate of hexaploid wheat transcript sequence</td>
</tr>
</tbody>
</table>
their analysis theoretically challenging and highly demanding of computational resources. Issues also include genome size, polyploidy, and the quantity, diversity, and dispersed nature of data in need of integration.

**Plant Trait Database.** The main focus of TRY (https://www.try-db.org/TryWeb/Home.php) database is to bring together the different plant trait databases worldwide into a comprehensive web-archive of the functional biodiversity of plants at the global scale by assembling, harmonizing, and distributing published and unpublished data on functional plant traits as well as a wide range of ancillary methodological and environmental information. It contains 3 million trait records for 750 traits of 1 million individual plants, representing 69,000 plant species [85, 86].

**TransPLANT.** Recently II European partners gathered to address growing database challenges and to develop a transnational database called “transplant” (http://www.transplantdb.eu/about) to help increasing database needs. Bringing together groups with strengths in data analysis, plant science, and computer science and from the academic and commercial sectors, transPLANT has developed integrated standards and services and undertaken new research and development needed to capitalize on the sequencing revolution, across the spectrum of agricultural and model plant species.

**PlantsDB.** This is another most commonly used database by various degree of researchers, and it comprises database instances for tomato, Medicago, Arabidopsis, Brachypodium, Sorghum, maize, rice, barley, and wheat. Building up on that, state-of-the-art comparative genomics tools such as CrowsNest are integrated to visualize and investigate syntenic relationships between monocot genomes. Results from novel genome analysis strategies targeting the complex and repetitive genomes of Triticeae species (wheat and barley) were provided and cross-linked with model species [87, 88].

5.2. Application of Comparative Genomics Platforms. Advancing genomic tools have provided higher boost for researchers in plant science community to understand the functional roles of genes and their evolutionary histories. Recently, resequencing additional genomes of a reference species has been made available [89], improving the understanding of genomic variation. Comparison of genomes gives insights into the evolution and adaptation of species to specific environments when compared to the information of genes provided by a single genome. To do comparative genomics studies there is a need of additional cost and as the number of available genomes increases, large-scale analyses become increasingly difficult for nonexperts, where need for computational biologist becomes essential [17]. Furthermore, biological variation between species and differences in sequence quality enhance the complexity of evolutionary analyses. Therefore, platforms for comparative genomics that take care of some of these challenges are valuable resources for experimental biologists [90, 91]. Comparative genomics has proven to be a valuable approach to understanding biology, not only for dissecting patterns and processes of genome evolution but also in revealing aspects of different gene function. The rapid advancement in comparative genomics technology, both for sequencing and for determining expression and interaction patterns, will continue to propel comparative genomics area of research in near future.

5.3. Emerging Databases for Comparative Genomics Analysis. To cope up and interact with increased data due to higher number of plant genome sequencing and inexpensive NGS technologies, recently developed and improved Phytozome database (http://www.phytozome.net) has provided a comparative hub for crop plant genome and gene family data analysis. The number of sequencing crop plant genomes is rapidly increasing and, at the same time, comparative sequence analysis has significantly changed our vision on the complexity of gene function, genome organization, and regulatory pathways. To explore all this genome information, a centralized infrastructure is required where all data generated by different sequencing initiatives is integrated and combined with advanced methods for data mining.

**PLAZA.** It is an online platform of plant comparative genomics (http://bioinformatics.psb.ugent.be/plaza/) that integrates functional and structural annotation of published crop plant genomes together with a large set of interactive tools to study gene and genome evolution along with their gene function. Precomputed datasets cover, intraspecies dot plots, whole-genome multiple sequence alignments, homologous gene families, phylogenetic trees, and genomic colinearity between species are provided by PLAZA. In conclusion, PLAZA provides the most comprehensible and up-to-date research environment to aid researchers in the exploration of genome information [92].

**GreenPhylDB.** GreenPhylDB is a component of the South Green Bioinformatics Platform (http://southgreen.cirad.fr/) and is open to public access (http://greenphyll.cirad.fr). GreenPhylDB is a database designed for functional and comparative genomics-based study on complete genomes. GreenPhylDB contains sixteen full genomes of members of the plantae kingdom, ranging from algae to angiosperms, automatically clustered into gene families. The database offers various lists of gene families including plant, phylum, and species specific gene families. Gene families are manually annotated and then analyzed phylogenetically in order to elucidate orthologous and paralogous relationships. It enables comparative genomics in a broad taxonomy context to enhance the understanding of evolutionary processes and thus tends to speed up gene discovery [91].

**iPlant Collaborative.** It enables transformative research through the use of a unified cyberinfrastructure funded by National Science Foundation (NSF) Plant Science Cyberinfrastructure Collaborative (PSCIC). iPlant (http://www.iplantcollaborative.org/) is a community of educators, researchers, and students working to enrich all plant sciences through the development of cyberinfrastructure,
the physical computing resources, virtual machine resources, collaborative environment and interoperable analysis software and data services that are essential components of modern biology.

**KBase.** It (http://kbase.us/) provides an open, extensible framework for secure sharing of data, tools, and scientific conclusions in predictive and systems biology. The Department of Energy Systems Biology Knowledgebase (KBase) is an emerging software and data environment designed to enable researchers to collaboratively generate, test, and share new hypotheses about gene and protein functions and also to perform large-scale analyses on a scalable computing infrastructure and model interactions in microbes, plants, and their communities.

### 6. Cross-Talk between Different Databases

Although several databases are available to public, still there is a lack of information needed for researchers exactly for what they are looking for. The update should not only take place in individual plant databases but also in all comparative genomic databases holding the genome. Updating the new version of genome for crop plant species should be uniform with several databases holding the genomes. The crop/plant specific databases should be updated periodically with new variety/germplasm lines whenever it becomes available including the ploidy level of the genome information for the easy access to researchers. Integration of data types and sources will continue to be a struggle in the future. In addition to the technical problems with integration, there is a need for vision at all community levels as to the role of integrating databases in the crop plant sciences for better usage. Several species focused databases like Graingenes (http://wheat.pw.usda.gov/) for *triticca*, oats, and sugarcane; *Brachypodium* database (http://www.brachypodium.org/) for *B. distachyon*; MaizeGDB (http://www.maizegdb.org/) for maize; Oryzabase (http://www.shigen.nig.ac.jp/rice/oryzabase/) for rice; BRAD (http://brassicadb.org/brad/) for Brassica crops; Legume information system (http://www.comparative-legumes.org/) for legumes; and SOL Genomics Network (SGN) (http://solgenomics.net/) for Solanaceae crop species should come forward for an integrated platform for researchers in field of crop plant science. The integrated breeding platform (IBP) of iPlant collaborative (http://www.integratedbreeding.net/) is playing big role to help plant breeders accelerate the creation and delivery of new crop varieties in the context of an increasing global demand for food.

### 7. Tools Needed for Data Interpretation and Utilization for Crop Improvement

All crop plant databases should be updated with basic statistical to advanced sequence analysis tools. As the sequence information has been made available to public for several crop plant genomes. Data interpretation tools should be developed within the databases for easy access of researchers. Reality is that many potential users will not use available resources for a number of reasons including lack of basic training in the use of bioinformatics, resources too difficult to learn and extract data, and simple inertia at learning new tools. Training of scientists for the current and future bioinformatics landscape is essentially important. Part of the solution is time since younger researchers are more attuned to the importance of bioinformatics than many established researchers. But more formal training in all aspects of bioinformatics tools, including database essentials and use, should be done for all future biological scientists. Having inbuilt tools for QTL linkage mapping, association mapping, genomic selection, and many more tools will aid the plant researchers to use the tool of interest and speed up the process of crop improvement.

### 8. Need for More Applied Research in Crop Plants

Alike quantitative trait loci (QTL), the genome sequencing project has provided much of the raw data for most of model as well as cultivated crops, which has shaped our view on genetics insights and evolution over the past two decades. Since it is a well teaching stuff to understand the complete architecture of organism, however, no applied researches have been undertaken so far in many of the sequenced crops that are already available to public (i.e., research impact is as same as the presequencing era) and now such work is just pleasure to read with beautiful chromosome maps and dizzying Venn diagrams. For instance, cereal genome sequencing (rice, wheat, sorghum, etc.,) was completed, but yet no demonstrated work on the cultivar development had been published or undertaken for wider applied research. Genome papers have been the bread and butter of evolution- ary biologists and geneticists for decades [93]. Everyone is jumping from one genome sequence to the next and looking to score a major publication aiming long-run project funding as some donors encourage them [93]. Everyone would like to see the genome sequencing projects in an optimistic way (any innovation takes its own time to influence the community) that can help us break some of the genetic bottle-neck for crop improvement in the early phase of 21st century. One and all, we should agree that every genome sequence project should have been deliberately designed to study the function of the gene in addition to the structural architecture for applied research since applied research is badly required for ongoing multi-sector crisis including agricultural food production under marginal lands. Product oriented research will have more impact than basic research alone. For instance, if more applied research is not undertaken then “genome-based research” could soon be dead which would affect the applied breeding for new cultivar development with respect to food crops as food security has still been a critical challenge for coming decades; populations blowing up unexpectedly in most of the developing countries and the novel agricultural research system should be in place to feed more than 9 billion people around the world in 2050.
9. Major Limitation of the Databases

As new sequencing technologies come online and the costs continue their downward trend, there will always be “more” worthy sequencing projects. Already we see multiple sequencing from the same genera with both the *Oryza japonica* and *Oryza indica* genomes sequenced and additional *Arabidopsis* genome projects following that of *Arabidopsis thaliana*. Making the crop plant databases and related bioinformatics tools easily accessible to research community is going to be a continual problem. As the volume of data power of computers increases, what is not possible is the software to fully use the potentials and the expertise of users in accessing those potentials. The amount of sequence data generated in crop plant research has dramatically increased over the last few years and will continue to accelerate in near future.

Researchers would want the complete genome sequence of every line of every organism under study; thus, an effectively unlimited thirst for sequence information will happen in near future. There will be whole-genomes of additional plants, the already mentioned sequence of additional versions of plant genomes, and intense resequencing of specific regions over tens, hundreds, and thousands of genomes. Custom microarrays are already made to resequence hundreds of thousands of dispersed DNA sequences. Resequencing to discover SNPs allows rapid genotyping through various array technologies. Currently, the planning is based more towards a minimal number necessary for a given program, but as cost declines and higher resolutions are within range of breeding programs, the density of desired SNPs may approach the entire genome level. There will also be more integration of data as knowledge, database, and analysis tools interlink. Functional genomics data on mRNA transcription and expression will tie to proteomic analyses and metabolomics of entire plants.

10. Conclusions

The implications of genomics on crop production can be envisioned on many fronts since fundamental advances in genomics would greatly accelerate the acquisition of knowledge and in turn will directly impact many aspects of the processes associated with crop plant trait improvement thereby considering productivity in a given environment. However, the complexity of possible higher orders of interactions can only be speculated with much more information, but the reasonable assumption is that it will dwarf our current limited views. A consequence of more voluminous and complex data is essential for better visualization and final validations. Better graphic tools to consolidate and summarize, and integration of data in a flexible manners to customize each researchers requirement. There will be more adoption of simultaneous data presentations and near future will involve ever more powerful computers, computational capability, sophisticated displays and interpretation tools, and greater practical expertise in the capabilities and exploitation of databases. Unless all these datasets are utilized in applied/product-oriented breeding program, the sequence data’s just to stay with its obituary notes in database network. Hence, scientist needs critical attention and discussion within and among disciplinary on the applied platforms of outcomes for better recognition of their novel research for betterment of humankind.

Conflict of Interests

Arun Prabhu Dhanapal and Mahalingam Govindaraj approve this paper and declare that they do not have any conflict for interests.

References


Research Article

Modulation at Age of Onset in Tunisian Huntington Disease Patients: Implication of New Modifier Genes

Dorra Hmida-Ben Brahim,1 Marwa Chourabi,1 Sana Ben Amor,2 Imed Harrabi,3 Saoussen Trabelsi,1 Marwa Haddaji-Mastouri,1 Moez Gribaa,1 Sihem Sassi,1 Fatma Ezzahra Gahbiche,1 Turkia Lamouchi,2 Soumaya Mougou-Zereli,1 Sofiane Ben Ammou,2 and Ali Saad1

1 Department of Cytogenetics and Reproductive Biology, Farhat HACHED Hospital, Ibn El Jazzar Road, 4000 Sousse, Tunisia
2 Department of Neurology, Sahloul Hospital, Sahloul Road, 4054 Sousse, Tunisia
3 Department of Epidemiology and Medical Statistics, Farhat HACHED Hospital, Ibn El Jazzar Road, 4000 Sousse, Tunisia

Correspondence should be addressed to Dorra Hmida-Ben Brahim; dorrahmida@yahoo.fr

Received 7 May 2014; Revised 26 July 2014; Accepted 13 August 2014; Published 1 September 2014

Academic Editor: Eugenia Poliakov

Copyright © 2014 Dorra Hmida-Ben Brahim 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.

Huntington’s disease (HD) is an autosomal dominant neurodegenerative disorder. The causative mutation is an expansion of more than 36 CAG repeats in the first exon of IT15 gene. Many studies have shown that the IT15 interacts with several modifier genes to regulate the age at onset (AO) of HD. Our study aims to investigate the implication of CAG expansion and 9 modifiers in the age at onset variance of 15 HD Tunisian patients and to establish the correlation between these modifiers genes and the AO of this disease. Despite the small number of studied patients, this report consists of the first North African study in Huntington disease patients. Our results approve a specific effect of modifiers genes in each population.

1. Background

Huntington disease (HD) is an autosomal dominant disorder caused by the expansion of polymorphic CAG repeat in the coding sequence of IT15 (OMIM no. 143100) also called huntingtin (htt) gene that leads to a progressive loss of neurons preferentially in the striatum and cortex. The symptoms, which usually appear between 40 and 50 years of age, are cognitive defects, psychiatric disorders, and motor dysfunction [1]. More than 36 repeats can cause the disease with the age at onset (AO) being inversely related to CAG repeat number on expanded chromosome. However, variation in CAG repeat number alone explains only around 60% of the variability in AO [2]. Evidence has been provided for genetic modifiers as well as for environmental factors that affect the AO. Several well-powered studies clearly implicate environmental modifiers in the AO and progression of HD [3]. Many genomic variations have also been tested for their influence on AO. Possible candidates are genes encoding products interacting with wild type or mutant huntingtin. Polymorphisms in these genes, exerting no effects in unaffected individuals, could modify the course of disease. Several studies have shown an effect of the unexpanded CAG repeat of the IT15 gene [4]. Other studies suggested a contribution of the gene variations of the N-methyl-D-aspartate receptor (NMDAR) subtypes NR2A (GRIN2A) (OMIM no. 6139971) and NR2B (GRIN2B) (OMIM no. 613970) to critically influence the variability in AO [2]. Many other modifier genes have also been tested for their influence on AO among them the polymorphic (Gln-Ala) repeat in the transcriptional coactivator CAI50 (or TCERG1 gene) (OMIM no. 605409) [5] and the CAG repeat expansion at the TATA-Box binding protein (TBP) (OMIM no. 607136), which is a factor of transcription [6].

In the present study we analyzed 4 polymorphisms that have not yet been described in modulation of AO in HD: CAG repeat of the DRPLA gene (OMIM no. 125370), CTG repeat of the DMPK (OMIM no. 160900), CAG repeat of the ATXN1 (OMIM no. 164400) gene, and CTG repeat of...
the JPH3 (OMIM no. 606438) implicated in Huntington’s disease-like-2 (HDL2), a phenocopy of Huntington’s disease. We also investigated 6 polymorphisms that are already known to be associated with HD: rs1969060 in GRIN2A, rs890 in GRIN2B, CAG repeat at TBP gene, (Gln-Ala) repeat in JPH3, and unexpanded and expanded CAG repeat of the IT15 gene.

2. Material and Methods

2.1. HD Patients. Our study population consisted of 15 Tunisian unrelated patients with clinical diagnosis of HD. All patients gave their consent for all information to be published. This research paper consists of a retrospective study that does not require ethics committee approval at our institution.

For all patients, the age-at-onset was estimated as the age when motor or cognitive symptoms were first noticed. The age-at-onset ranged from 24 to 72 years with a mean age-at-onset of 44 years. Informed consent was obtained from all patients. HD CAG repeat sizes were determined by polymerase chain reaction using an assay counting the perfectly repeated (CAG) units. Repeated numbers of patients derived from European institutions were randomly checked in our laboratory with a reference control. The expanded CAG repeats ranged from 40 to 49 trinucleotide repeats. The median repeat number was 44.

2.2. DNA Samples. Genomic DNA was isolated from 5 mL of peripheral blood with the flexiGen DNA kit from Qiagen according to manufactured protocol and then diluted to a final concentration of 200 ng/μL.

2.3. Genotyping of Polymorphic Repeats. Polymorphic repeats in htt, DMPK, DRPLA, ATXN1, JPH3, TBP, and TCERG1 genes were determined using a modification of the PCR amplification assay reported by Warner et al., with fluorescent oligonucleotide primers flanking the repeats detailed in Table 1 [7]. Determination of the number of repeats was carried out by fragment analysis with the ABI 310 Genetic Analyzer System (Applied Biosystems) according to the manufacturer’s instructions. PCR conditions are available upon request.

2.4. Genotyping of GRIN2A and GRIN2B. Genotyping for SNP rs1969060 (GRIN2B) and SNP rs890 (GRIN2A) was performed by direct sequencing (see Table 1). PCR conditions are available upon request.

2.5. Statistical Analysis. Statistical analysis was performed using SPSS 17.0. A possible modifying effect on the HD age-at-onset of the respective polymorphisms was investigated by applying a model of analysis regression with multiple variables. The goodness of fit was evaluated by the proportion of variation in the age-at-onset, explained by the coefficient of determination ($R^2$). For analysis, variance in the age-at-onset for the CAG repeats in htt was determined alone as well as in addition to different polymorphisms. A change of $R^2$ indicated a relative improvement of the model of multiple regressions when the respective factors were added to the effect of the expanded huntingtin allele ($\Delta R^2$). This identified the percentage of the variance that was attributable to the candidate modifier genes, when there was a significant $P$ value. A $P$ value of less than 0.05 was considered significant.

### Table 1: Studied polymorphisms and corresponding oligonucleotide primers.

<table>
<thead>
<tr>
<th>Genes</th>
<th>Polymorphisms</th>
<th>Primers</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRIN2A</td>
<td>SNP rs890</td>
<td>Forward 5'TGT ACC CAG ATA TAT ACA GAC AC 3'</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reverse 5'TGA GAT GGT AGA GTA CGA GTC A 3'</td>
</tr>
<tr>
<td>GRIN2B</td>
<td>SNP rs1969060</td>
<td>Forward 5'GGG GGG TAG AGC GGA GAA AG 3'</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reverse 5'TCC TTT CAC AAG CAG TGT GCG 3'</td>
</tr>
<tr>
<td>IT15</td>
<td>CAG</td>
<td>Forward 5'ATG AAG GCC TTC GAG TCC CTC AAG TCC TTC 3'</td>
</tr>
<tr>
<td>TCERG1</td>
<td>Gln-Ala</td>
<td>Reverse 5'TCA GAA TCG CAG TTT TGT GAG ACG TGC TGC TGC TGC 3'</td>
</tr>
<tr>
<td>TBP</td>
<td>CAG</td>
<td>Forward 5'CAC CCC ACA GCC TAT TCA GA 3'</td>
</tr>
<tr>
<td>JPH3</td>
<td>CTG</td>
<td>Reverse 5'GGG GAA ACA GAC AGT TGC GTC TGC TGC TGC TGC 3'</td>
</tr>
<tr>
<td>ATXN1</td>
<td>CAG</td>
<td>Forward 5'GAG GCT CTC CTC ATC TAG ATC TAG ATC TAG ATC 3'</td>
</tr>
<tr>
<td>DMPK</td>
<td>CTG</td>
<td>Reverse 5'TTC GCG GGG TGG CGG AGC TGC TGC TGC TGC TGC TGC TGC TGC TGC TGC TGC TGC TGC TGC TGC TGC TGC TGC 3'</td>
</tr>
<tr>
<td>DRPLA</td>
<td>CAG</td>
<td>Forward 5'CAC CCA CCAGTC TCA ACA CA 3'</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reverse 5'TCA GCA TCC CAG TTT TGT GAG ACG TGC TGC TGC TGC TGC TGC TGC TGC TGC TGC TGC TGC TGC TGC TGC TGC TGC TGC TGC TGC TGC TGC TGC 3'</td>
</tr>
</tbody>
</table>

*The fluorescent primer is identical to 5' half (underlined) of reverse primers of all triplet repeat polymorphisms. It was added to the PCR mixture in combination with each pair of triplet repeat nucleotide primers.*
Correlation between respective polymorphic repeats lengths and the AO was performed using the Pearson correlation coefficient (r). As adopted for the coefficient of determination \(R^2\), a P value of less than 0.05 was considered as significant.

3. Results

Both negative correlation and significant effect of the expanded CAG repeat number in the huntingtin gene on the AO of HD patients have been shown in numerous studies [3, 8]. In our study, we could also confirm these observations by applying a statistical multiple regression model of an analysis of variance (Table 2). CTG repeat in the JPH3 gene was also found to be inversely correlated to the AO (see Table 2).

To investigate the possible modifiers that can influence the AO, we followed a multiple regression model. The value of \(R^2\) was determined for genotypes at each individual locus together with the CAG repeat number on expanded chromosomes. This value was then compared with the value of \(R^2\) obtained by considering expanded CAG repeat alone. A change of \(R^2\) (\(\Delta R^2\)) is the measure of the influence of putative modifying factors on the variation of AO (Table 3).

Using this model in HD, the value of \(R^2\) was 0.745 (\(P = 0.000\)). Thus, for HD 74.5% of the variation of AO can be explained by expanded CAG repeats.

We also established the implication of other polymorphisms studied in DRPLA, DMPK, TCERG1, TBP, ATXN1, JPH3, GRIN2B (rs890), and GRIN2A (rs1969060) genes in the unexplained variance of AO (see Table 3).

4. Discussion

The evaluation of the age-at-onset in HD presents a challenge that has to be solved as precise as possible. In this pilot study, we characterized a large number of polymorphisms in genes that are suggested to act as possible modifiers for the AO of HD. Environmental modifiers, already known to have a considerable effect on AO and progression of HD, were not investigated in the present study. It would be of great interest to also look for environmental modifiers in African population, where environmental factors may differ between countries and thus be highly informative. Furthermore, via gene-environment interactions, environmental modifiers may affect the detectability of genetic modifiers in specific cohorts. Wexler [3] reported that 60% of the variance remaining in AO is attributable to environmental modifiers in Venezuelan population.

Genetic modifier factors have been indicated in HD as the length of the disease causing expanded polyglutamine tract in huntingtin explains only 65–70% of the variance in the age-at-onset [2]. In our report, the CAG repeat accounts for 74.5% of the variance in AO which is in accord with other previous studies [9]. Previous studies have established that the unexpanded HD allele has also been shown to have some association with OA in 754 patients from North America, Europe, and Australia [10]. Later AO was being associated with longer repeats lengths in normal HD gene [10], whereas a study of 138 patients from Wales found negative correlation between the normal CAG repeat and the AO (\(P = 0.014\)) [11]. Other studies did not confirm any correlation [2, 8, 12]. We have found a negative correlation between unexpanded CAG and AO but, that is, not significant. This result can be explained by the low number of patients. The combination with the expanded allele, using the model of analysis regression, indicated elevation of value of \(R^2\) from 0.745 to 0.789 (\(P = 0.000\)); this elevation indicated that 4.4% of the total variability in the AO was attributable to the variations in CAG normal repeats in the huntingtin gene. Thus, we could interpret 17.18% of the unexplained variation in AO by variation at this locus.

Several data have suggested that the NMDA subset of glutamate receptor can contribute to neurodegeneration in HD. These receptors permit influx of calcium and, when activated, can generate neuronal death. NMDARs are multimeric complexes composed of NR1 subunits together with NRA2, NR2B, NR2C, and/or NR2D. In the cortex and striatum, NR2A and NR2B predominate; moreover the NR2B/NR2A ratio is higher in the striatum than other brain regions [13]. Both NR2A and NR2B subunits are encoded, respectively, by GRIN2A and GRIN2B genes. A study with 167 patients from Germany interested to investigate the association between the AO and two SNPs: a C/T SNP (rs1969060) in GRIN2A and a T/G SNP (rs890) in GRIN2B, showed that 4.5% of the unexplained variation in AO could be interpreted by a C/T SNP and 2.31% of the unexplained variation in OA could be contributed to the T/G SNP [13]. A Venezuelan study was conducted on the same number of samples and was interested in these two polymorphisms and indicates that 4.5% of the unexplained variation in AO could be attributed to the GRIN2A gene, although the GRIN2B gene had no implication in the variation of the AO [2].

Here, we found that the polymorphism studied in the GRIN2A added to the CAG expansion shows a low variation from 0.745 to 0.748; we might explain that 1.17% of the unexplained variation in AO can be associated to this polymorphism. When the GRIN2B polymorphism added the regression model, the value of \(R^2\) increases (from 0.745 to 0.765). This data can explain that 7.81% of the unexplained

<table>
<thead>
<tr>
<th>Gene (polymorphism)</th>
<th>Pearson correlation coefficient r</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>IT15 (expanded CAG)</td>
<td>−0.863</td>
<td>0.000*</td>
</tr>
<tr>
<td>IT15 (unexpanded CAG)</td>
<td>−0.383</td>
<td>0.176</td>
</tr>
<tr>
<td>DRPLA (CAG)</td>
<td>−0.225</td>
<td>0.440</td>
</tr>
<tr>
<td>DMPK (CTG)</td>
<td>0.295</td>
<td>0.306</td>
</tr>
<tr>
<td>ATXN1 (CAG)</td>
<td>−0.290</td>
<td>0.315</td>
</tr>
<tr>
<td>TBP (CAG)</td>
<td>−0.299</td>
<td>0.299</td>
</tr>
<tr>
<td>JPH3 (CTG)</td>
<td>−0.547</td>
<td>0.043*</td>
</tr>
<tr>
<td>TCERG1 (Gln-Ala)</td>
<td>−0.151</td>
<td>0.606</td>
</tr>
<tr>
<td>GRIN2A (rs1969060)</td>
<td>0.287</td>
<td>0.320</td>
</tr>
<tr>
<td>GRIN2B (rs890)</td>
<td>−0.129</td>
<td>0.661</td>
</tr>
</tbody>
</table>

*P < 0.5.
Variance in AO for the CAG repeats is indicated as well as in combination with the different examined polymorphisms. In regression analysis, the TCERG1 gene when we added the ATXN1 locus, respectively, at 1.07% and 2.19% [2,15]. In this present report, patients demonstrated that this locus can influence the AO, of 432 American patients and another of 427 Venezuelan with n-terminal fragments of huntingtin [2]. The study homologue of a Caenorhabditis elegans protein that interacts encoded protein forms insoluble aggregates in the nucleus of neuronal cells in HD patients [4]. Similar to huntingtin, gene with HD. This gene is a good candidate as the encoded protein forms insoluble aggregates in the nucleus (in addition to the contribution of the expanded CAG repeat (HD CAG)).

<table>
<thead>
<tr>
<th>Gene (polymorphism)</th>
<th>$R^2$</th>
<th>$\Delta R^2$</th>
<th>% unexplained variability</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>HD CAG</td>
<td>0.745</td>
<td>—</td>
<td>—</td>
<td>0.000*</td>
</tr>
<tr>
<td>HD CAG + normal CAG</td>
<td>0.789</td>
<td>0.044</td>
<td>17.18</td>
<td>0.000*</td>
</tr>
<tr>
<td>HD CAG + DRPLA (CAG)</td>
<td>0.793</td>
<td>0.048</td>
<td>18.75</td>
<td>0.000*</td>
</tr>
<tr>
<td>HD CAG + DMPK (CTG)</td>
<td>0.750</td>
<td>0.005</td>
<td>1.95</td>
<td>0.000*</td>
</tr>
<tr>
<td>HD CAG + ATXN1 (CAG)</td>
<td>0.782</td>
<td>0.037</td>
<td>14.45</td>
<td>0.000*</td>
</tr>
<tr>
<td>HD CAG + TBP (CAG)</td>
<td>0.747</td>
<td>0.002</td>
<td>0.78</td>
<td>0.001*</td>
</tr>
<tr>
<td>HD CAG + JPH3 (CTG)</td>
<td>0.774</td>
<td>0.029</td>
<td>11.32</td>
<td>0.000*</td>
</tr>
<tr>
<td>HD CAG + TCERG1 (Gln-Ala)</td>
<td>0.746</td>
<td>0.001</td>
<td>0.39</td>
<td>0.001*</td>
</tr>
<tr>
<td>HD CAG + GRIN2A (rs1969060)</td>
<td>0.748</td>
<td>0.003</td>
<td>1.17</td>
<td>0.000*</td>
</tr>
<tr>
<td>HD CAG + GRIN2B (rs890)</td>
<td>0.765</td>
<td>0.020</td>
<td>7.81</td>
<td>0.000*</td>
</tr>
</tbody>
</table>

$^*P < 0.05.$

Variance in AO for the CAG repeats is indicated as well as in combination with the different examined polymorphisms. $R^2$ illustrates the relative improvement of the regression model when the various genotypes are considered in addition to the HD CAG repeats. $\Delta R^2$ values quantify these differences. P value refers to $R^2$.

Variation in AO was attributable to the T/G SNP in the GRIN2B gene.

Some previous studies have indicated an association of the TBP gene with HD. This gene is a good candidate as the encoded protein forms insoluble aggregates in the nucleus of neuronal cells in HD patients [4]. Similar to huntingtin, TBP contains a polymorphic CAG repeat; mutant huntingtin interacts with TBP and impairs the functional conformation of the transcription factor [14]. In this report, the $R^2$ statistic increases slightly from 0.745 to 0.747 with the TBP genotypes included. This low variation might explain that 0.78% of the unexplained variation in AO can be contributed to this locus.

Age at onset in HD might be influenced by the length of the polymorphic (Gln-Ala) repeat in TCERG1 gene. This gene encodes the transcriptional coactivator CA150, a human homologue of a Caenorhabditis elegans protein that interacts with n-terminal fragments of huntingtin [2]. The study of 432 American patients and another of 427 Venezuelan patients demonstrated that this locus can influence the AO, respectively, at 1.07% and 2.19% [2,15]. In this present report, when we added the TCERG1 genotypes in the model of regression analysis, the $R^2$ statistic rose from 0.744 to 0.746. This increase indicates that 0.78% of the unexplained variance of the AO in HD could be explained by variation of imperfect (Gln-Ala) repeat at TCERG1 locus. This strong variability of results between different populations may suggest a specific effect of modifiers genes in each population.

We are also interested in investigating the implication of four novel genes: CAG polymorphic repeat of the DRPLA and ATXN1 genes and CTG polymorphic repeat of the JPH3 and DMPK genes. DRPLA, ATXN1, DMPK, and JPH3 are responsible, respectively, for the dentatorubral-pallidolysian atrophy disorder (DRPLA), spinocerebellar ataxia type 1 (SCA1), myotonic dystrophy 1 (DM1), and Huntington’s disease-like-2 (HDL2). As it is the case with HD, DRPLA, SCA1, and DM1 are all caused by a gain of function mechanism [16, 17]. Similar to HD, the HDL2 major symptom is chorea. Based on those arguments we conducted the analysis of the 4 novel polymorphisms to look for their involvement in the variation in AO.

Here, we found that JPH3 CTG repeat was inversely correlated to the AO (see Table 2); this correlation was significant at $P = 0.043$. Given that HD is uncommon in our country, JPH3 finding needs to be improved by studies in larger international cohorts.

Using the model of regression analysis, we could confirm the implication of DRPLA, DMPK, ATXN1, and JPH3 in the variability of AO in HD. For the DRPLA gene, we noticed an increase in the $R^2$ value from 0.745 to 0.793. Thus, we indicated that 18.75% of the unexplained variation in AO might be influenced by the variation in this locus. When we added the DMPK gene to the HD expanded allele we observed that $R^2$ rose slightly from 0.745 to 0.750. This increase allows interpreting 1.95% of the unexplained variation in AO by this gene. In our sample, we can interpret 14.45% of the unexplained variation in AO by the variation of the CAG repeat in the ATXN1 gene ($R^2$ rose from 0.745 to 0.782). Finally, we have established the involvement of the polymorphic CTG repeat in the JPH3 gene at 11.32% of the unexplained variation in AO ($R^2$ rose modestly from 0.745 to 0.774).

5. Conclusion

In conclusion, the objective of all studies is to obtain a complete understanding of the variance in AO, that is, attributable to genetic factors other than the length of the unexpanded repeat in the IT15 gene. We report here the first North African study on Huntington disease. Despite the poor knowledge of HD epidemiological distribution in North African population, we succeeded to prove the implication of 9 polymorphisms encompassing 4 novel ones located at DRPLA, DMPK, ATXN1, and JPH3 genes. Our results suggest a specific Tunisian age of onset prediction model that should be tested in larger Tunisian HD population.

Conflict of Interests

The authors declare that they have no conflict of interests.
Authors’ Contribution

Dorra Hmida-Ben Brahim is responsible for the overall content as a guarantor. Dorra Hmida-Ben Brahim and Marwa Chourabi have contributed to the paper through involvement in study design, statistical analysis and interpretation, and paper preparation. Imed Harrabi contributed to the paper through involvement in study design, data collection, and final paper revision. Soumaya Mastouri have contributed to the paper through involvement in study design, paper preparation, and revision. Soufiane Ben Ammouh have contributed to the paper through involvement in statistical study and analysis. Sana Ben Amor, Turkia Lamouchi, Fatma Ezzahra Gahbiche, and Soufien Ben Ammouh have contributed to the paper through involvement in study design, data collection, and final paper revision. Saoussen Trabelsi, Sihem Sassi, and Marwa Haddaji-Mastouri have contributed to the paper through involvement in study design, paper preparation, and revision. Soumaya Mougou-Zereli, Moez Gribaa, and Ali Saad have contributed to the paper through involvement in study design, paper editing and revision, and final review.

Acknowledgments

Authors are grateful to the patients and their families. They thank Miss Ahlem Msakni and Miss Safa Bouker for their technical help. They are also grateful to Miss Wiem Manoubi for her help in paper revision.

References


More than half a century from postulated Warburg theory of cancer cells origin [1], question of changed metabolism in cancer is again taking central place. For several decades a generalized picture of cancer metabolism was replaced by analysis of signaling and oncogenes in each type of cancer for several decades. However, now empowered with wealth of knowledge about tumor suppressors, oncogenes, and signaling pathways, reprogramming of cellular metabolism described as Warburg effect (increased glycolysis to respiration ratio in cancer cells) [2] reemerged as an important element of cancer progression [3–5]. The metabolism of one molecule of glucose to two molecules of pyruvate in glycolysis has a net yield of two molecules of ATP. Glycolysis does not require or consume oxygen. Eukaryotic aerobic respiration (oxidative phosphorylation) produces approximately 34 additional molecules of ATP for each glucose molecule. The lower-energy production, per glucose, of anaerobic respiration relative to aerobic respiration, results in greater flux through the pathway under hypoxic (low-oxygen) conditions. It has been hypothesized that, in these cells, glycolytic enzymes associate into a large complex, which results in an increased efficiency of glycolytic flux [6]. Another explanation is based on alternative glycolytic pathway that bypasses ATP production but produces pyruvate [7]. Recent analysis of microarray data across major cancer types showed activation of certain metabolic pathways in cancer cells [8, 9]. These data confirmed that cancer cells upregulate biosynthesis and metabolism of certain nutrients like glycine and glutamine along with upregulated glycolysis [10, 11].

Analysis of cancer molecular signatures deduced from genomics data recently appeared in the literature [12–14]. For example, overexpressed membrane receptors as suggested by an analysis of ESTs (expressed sequences tags) could be used as hallmark of cancer cells [15]. EST datasets were also analyzed by Aouacheria and coworkers to distinguish between normal and tumor tissues [16]. Analysis was done
Table 1: Examples of genes overexpressed in cancer: ten genes with the largest number of EST libraries.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Number of cancer ESTs</th>
<th>Number of cancer EST libraries</th>
<th>Number of normal ESTs</th>
<th>Number of normal EST libraries</th>
<th>Gene description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EEF1A1</td>
<td>26747</td>
<td>380</td>
<td>54872</td>
<td>1436</td>
<td>Translation elongation factor</td>
</tr>
<tr>
<td>GAPDH</td>
<td>11388</td>
<td>300</td>
<td>9446</td>
<td>858</td>
<td>Glyceroldehyde-3-phosphate dehydrogenase</td>
</tr>
<tr>
<td>ACTG1</td>
<td>9831</td>
<td>295</td>
<td>8826</td>
<td>879</td>
<td>Actin gamma</td>
</tr>
<tr>
<td>FTH1</td>
<td>5264</td>
<td>260</td>
<td>6627</td>
<td>758</td>
<td>Ferritin, heavy polypeptide 1</td>
</tr>
<tr>
<td>EEF1G</td>
<td>6575</td>
<td>255</td>
<td>7185</td>
<td>783</td>
<td>Eukaryotic translation elongation factor 1 gamma</td>
</tr>
<tr>
<td>RPLP0</td>
<td>6583</td>
<td>253</td>
<td>6400</td>
<td>683</td>
<td>Ribosomal protein, large subunit, P0</td>
</tr>
<tr>
<td>HSP90AB1</td>
<td>4035</td>
<td>253</td>
<td>7178</td>
<td>801</td>
<td>Heat shock protein 90 kDa alpha (cytosolic), class B member 1</td>
</tr>
<tr>
<td>PKM2</td>
<td>4794</td>
<td>248</td>
<td>7593</td>
<td>760</td>
<td>Pyruvate kinase, muscle</td>
</tr>
<tr>
<td>FTL</td>
<td>4495</td>
<td>246</td>
<td>7784</td>
<td>599</td>
<td>Ferritin, light polypeptide</td>
</tr>
<tr>
<td>RPL3</td>
<td>6278</td>
<td>242</td>
<td>7463</td>
<td>753</td>
<td>Ribosomal protein L3</td>
</tr>
</tbody>
</table>

The complete list of genes overexpressed in cancer is shown in Supplementary Table S1 (also available at ftp://ftp.ncbi.nlm.nih.gov/pub/managdav/paper_suppl/est_cancer_pathways/).

in a tissue-specific manner but demonstrated similar patterns of enzymes enrichment in various cancers, for example, glycolytic enzymes, alpha enolase ENO1 (EC 4.2.1.11), pyruvate kinase, muscle PKM2 (EC 2.7.1.40), and glyceraldehyde-3-phosphate dehydrogenase GAPDH (EC 1.2.1.12), were found in three different types of cancer [16]. This result demonstrates that some genes are ubiquitously highly expressed in cancer cells and such genes may be important hallmarks of cancer cells [12].

To analyze level of expression of various proteins including metabolic enzymes across various cancers we used EST data. We also used the NCBI Unigene data across the panel of cancers. We grouped overexpressed enzymes into KEGG pathways and manually analyzed adjacent pathways to analyze enzymatic reactions that take place in cancer cells and to identify major players that are abundant in cancer protein machinery.

2. Materials and Methods

The 23,586 nonredundant coding sequences (CDS) of human genes from the human genome draft build 35, the April 2012 freeze, obtained at the NCBI ftp server were used as reference sequences to be compared with the EST sequences. The EST sequences were from the dbEST release of June 2012. Each EST library was assigned to either cancer-related (1,437 libraries; 1,848,538 ESTs (26%)) or normal tissue (7,309 libraries; 5,276,385 ESTs (74%)) bins (the file EST_libraries.xls is available at ftp://ftp.ncbi.nlm.nih.gov/pub/managdav/paper_suppl/est_cancer_pathways/). The CDS set was searched against dbEST using the BLASTN program with the default parameters. The number of ESTs with at least 97% identity and alignment length of at least 200 nucleotides or longer was counted. The Unigene database was used to verify results of BLASTN-dbEST searches.

We used two statistical filters to delineate a set of genes overexpressed in cancer (10 genes with the largest number of cancer EST libraries are shown in the Table 1).

(1) We included a gene in a list of preliminarily candidates if the number of cancer ESTs is greater in comparison with normal ESTs using Fisher $2 \times 2$ exact test ($X_1 - X_2, X_2, X_3 - X_4, X_4$) (an arbitrary threshold $P$ value = 0.01). $X_1$ is the total number of cancer ESTs, $X_2$ is the number of cancer ESTs for a given gene, $X_3$ is the total number of normal ESTs, and $X_4$ is the number of normal ESTs for a given gene.

(2) We accepted a gene in the final list (Supplementary Table S1; Supplementary Material available online at http://dx.doi.org/10.1155/2014/646193) from the preliminarily candidates if the number of cancer ESTs per EST library is greater in comparison with normal ESTs per EST library using Fisher $2 \times 2$ exact test ($X_1, X_2, X_3, X_4$) (an arbitrary threshold $P$ value = 0.01). $X_1$ is the total number of cancer ESTs for a given gene, $X_2$ is the number of cancer EST libraries with at least one BLASTN hit for a given gene, $X_3$ is the total number of normal ESTs for a given gene, and $X_4$ is the number of normal EST libraries with at least one BLASTN hit for a given gene. In other words, if a gene has an EST-detectable expression level in an EST library, it should be expressed in cancer EST libraries higher than in normal EST libraries.

A KEGG pathway enrichment in genes overexpressed in cancer libraries (Table 2) was estimated using the Fisher $2 \times 2$ exact test ($X_1-X_2, X_2, X_3-X_4, X_4$) and the KEGG database [17]. $X_1$ is the total number of proteins in the initial set, $X_2$ is the number of proteins in a given pathway, $X_3$ is the total number of genes that are overexpressed in cancer, $X_4$
Table 2: KEGG metabolic pathways enriched in genes that are overexpressed in cancer.

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Number of genes in a pathway (X2)</th>
<th>Number of genes overexpressed in cancer (X4)</th>
<th>Pathway enrichment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycolysis/gluconeogenesis</td>
<td>62</td>
<td>10</td>
<td>$2.8 \times 10^{-7}$</td>
</tr>
<tr>
<td>Oxidative phosphorylation</td>
<td>135</td>
<td>10</td>
<td>0.00015</td>
</tr>
<tr>
<td>Pyruvate metabolism</td>
<td>40</td>
<td>5</td>
<td>0.0083</td>
</tr>
<tr>
<td>Metabolism of xenobiotics by cytochrome P450</td>
<td>70</td>
<td>6</td>
<td>0.0016</td>
</tr>
<tr>
<td>Fructose and mannose metabolism</td>
<td>34</td>
<td>4</td>
<td>0.0034</td>
</tr>
<tr>
<td>One carbon pool by folate</td>
<td>17</td>
<td>3</td>
<td>0.0041</td>
</tr>
<tr>
<td>TCA cycle (citrate cycle)</td>
<td>32</td>
<td>3</td>
<td>0.02</td>
</tr>
<tr>
<td>Glycerolipid metabolism</td>
<td>49</td>
<td>3</td>
<td>0.054</td>
</tr>
<tr>
<td>Steroid hormone biosynthesis</td>
<td>55</td>
<td>3</td>
<td>0.07</td>
</tr>
<tr>
<td>Pentose phosphate pathway</td>
<td>27</td>
<td>2</td>
<td>0.082</td>
</tr>
<tr>
<td>Purine metabolism/purine de novo biosynthesis</td>
<td>159</td>
<td>4</td>
<td>0.35</td>
</tr>
<tr>
<td>Pyrimidine metabolism/pyrimidine de novo biosynthesis</td>
<td>98</td>
<td>3</td>
<td>0.23</td>
</tr>
<tr>
<td>Cysteine and methionine metabolism</td>
<td>34</td>
<td>2</td>
<td>0.12</td>
</tr>
<tr>
<td>Aminoacyl tRNA biosynthesis</td>
<td>41</td>
<td>2</td>
<td>0.16</td>
</tr>
<tr>
<td>Fatty acid metabolism</td>
<td>43</td>
<td>2</td>
<td>0.17</td>
</tr>
<tr>
<td>N-Glycan biosynthesis</td>
<td>46</td>
<td>2</td>
<td>0.19</td>
</tr>
<tr>
<td>Phenylalanine/tyrosine/histidine metabolism</td>
<td>18/42/29</td>
<td>3</td>
<td>—</td>
</tr>
<tr>
<td>Enzymes introducing posttranslational</td>
<td>?</td>
<td>6</td>
<td>—</td>
</tr>
<tr>
<td>modifications</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Additional antioxidant systems</td>
<td>?</td>
<td>3</td>
<td>—</td>
</tr>
</tbody>
</table>

Pathway enrichment is estimated using the Fisher $2 \times 2$ exact test ($X_1 \times X_2$, $X_2$, $X_3 \times X_4$, $X_4$). $X_1$ is the total number of genes in the initial set (23,586 genes), $X_2$ is the number of proteins in a given pathway, $X_3$ is the total number of genes that are overexpressed in cancer (394 genes), and $X_4$ is the total number of genes in a given pathway that are overexpressed in cancer.

is the total number of genes in a given pathway that are overexpressed in cancer. It should be noted that all three statistical tests are deliberately simplified because EST datasets can be used as semiquantitative estimators of gene expression and are considered to be an approximate measure of expression in this study (see Section 4).

3. Results

3.1. Analysis of Genes/Pathways Overexpressed in Cancer. We analyzed large collections of EST libraries associated with cancer and compared them to EST libraries from normal tissues. We tried to select genes that are highly expressed in cancer and the level of expression is substantially different between cancer and normal samples taking into account that EST data is inherently semiquantitative. We used two statistical filters to delineate a list of genes that are highly expressed in many cancer-associated EST libraries (see Section 2). In this way we attempted to remove genes that are associated with specific types of cancer in order to reconstruct a generalized portrait of a cancer cell. We delineated 394 genes that are overexpressed in cancer cells according to statistical filters (see Section 2, Supplementary Table S1; the file Supp_table_S1.xls is also available at ftp://ftp.ncbi.nlm.nih.gov/pub/managdav/paper_suppl/est_cancer_pathways/). A wide variety of functional themes were found in this list and many products of these genes are proteins involved in translation. For example, among 10 genes with the largest number of cancer-related EST libraries, two ribosomal proteins and two translation elongation factors were found (Table 1). This is consistent with many previous observations of highly significant elevation of protein synthesis rates and the expression of several translation components in various cancer cells indicating an importance of ribosome function and translational control in tumor progression [18–21].

We used the KEGG database [17] to assign proteins overexpressed in cancer to various metabolic pathways. We found several KEGG pathways that are significantly enriched in proteins overexpressed in cancer whereas many other pathways do not show a significant enrichment in genes overexpressed in cancer although they contain two or more genes overexpressed in cancer (Table 2). We discuss all these pathways below in more detail.

3.2. Glycolysis/Gluconeogenesis. Ten enzymes (TPI1, PGAM1, ENO1, PKM2, ALDH3A1, GAPDH, LDHB, ALDH3B1,
ALDH3B2, and ALDOA) from the glycolysis pathway (62 genes) are overexpressed in cancer (Table 2 and Supplementary Table S1). The PKM2 embryonic isoform of pyruvate kinase is important for cancer metabolism and tumor growth [22, 23]. PGAM1 and PKM2 are involved in the alternative glycolytic pathway producing pyruvate [7]. Our data confirm that glycolysis is the central metabolic process for cancer cells.

3.3. Oxidative Phosphorylation. Ten enzymes (ATPSFI, SDHD, ATP5B, UQRC1, NDUF7, ATP6V0E1, NDUF9, NDUF2, NDUF5, and CYC1) out of 135 proteins participating in this pathway were overexpressed in cancer (Table 2 and Supplementary Table S1). It was demonstrated that hypoxic cancer cells maintain active, though diminished, oxidative phosphorylation even at 1% oxygen. ATP production in these cells is around 40% of ATP production under normal oxygen conditions and their results suggest that, under hypoxia, the autophagy is required to support ATP production [24]. Our data demonstrate that the oxidative phosphorylation is important for cancer survival and growth. Overexpression of LDHB according to cancer EST analysis (Supplementary Table S1) is also pointing to a possible utilization of lactate for the oxidative metabolism [25, 26].

3.4. Pyruvate Metabolism. Pyruvate is produced in glycolysis and is used for lactate production. High ratio of lactate/pyruvate is used for metabolic imaging of prostate cancer [27]. Lactate is a prominent substrate that fuels the oxidative metabolism of oxygenated tumor cells. There is a symbiosis in which glycolytic and oxidative tumor cells mutually regulate their access to energy metabolites [25]. Preferential utilization of lactate for oxidative metabolism spares glucose that may in turn reach hypoxic tumor cells [26]. Five enzymes (LDHB, MDH2, PKM2, AKRIB1, and AKRIB10) out of 40 enzymes in pyruvate metabolism pathway are overexpressed in cancer (Table 2 and Supplementary Table S1). Surprisingly, we found an overexpression of the lactate dehydrogenase B enzyme in certain cancers (Supplementary Table S1) and no significant upregulation of the lactate dehydrogenase A enzyme that has been suggested to have a ubiquitous role in tumor metabolism and growth [28].

3.5. Metabolism of Aldehydes (Xenobiotics, Drugs) by Cytochrome P450. We found that our EST dataset is enriched in various aldehyde dehydrogenases (ALDH3A1, ALDH3B1, and ALDH3B2) (Table 2 and Supplementary Table S1). Aldehyde dehydrogenase superfamily plays an important role in the enzymatic detoxification of endogenous and exogenous aldehydes and in the formation of molecules that are important in cellular processes. Additionally, ALDH3B1 expression is upregulated in many human tumors and this enzyme is catalytically active toward aldehydes derived from lipid peroxidation, suggesting a potential role against oxidative stress [29]. Moreover, three members of aldo-keto reductases (AKRs), AKR1C1, AKR1C2, and AKR1C3, are overexpressed in cancer (Table 2 and Supplementary Table S1). The resistance towards the chemotherapeutic drug cisplatin in colon cancers is believed to be a result of decreased sensitivity toward cellular damages evoked by oxidative stress-derived aldehydes, 4-hydroxy-2-nonenal and 4-oxo-2-nonenal, that are detoxified by AKR1C1 and AKR1C3 [30]. Metabolism of xenobiotics by cytochrome P450 pathway has six overexpressed enzymes from total 70 enzymes. Furthermore, one member of the superfamily of short-chain dehydrogenases/reductases (SDR) was found to be overexpressed in cancer (DHRS2 or Hep27, Supplementary Table S1), SDR catalyzes the NADPH-dependent reduction of dicarboxyl compounds [31].

3.6. Fructose and Mannose Metabolism. Fructose provides an alternative carbon source for glycolysis, entering downstream of glucose and bypassing two key rate-limiting steps. Whereas glucose favors overall growth kinetics, fructose enhances protein and nucleotide synthesis and appears to promote a more aggressive cancer phenotype [32, 33]. Four enzymes (TPP1, AKRIB10, ALDOA, and AKRIB1) out of 34 enzymes in this pathway are overexpressed in cancer (Table 2 and Supplementary Table S1). All these enzymes participate in a variety of metabolic pathways.

3.7. One Carbon Pool by Folate. Three enzymes (MTHFD2, ATIC, and SHMT2) in this pathway were found to be overexpressed in many cancers (Table 2). This result suggests that most of tetrahydrofolate (THF) forms (5,6,7,8-THF; 10-formyl-THF; 5,10-methyl-THF; 5,10-methylene-THF) are synthesized in cancer cells. Byproduct of serine hydroxymethyltransferase 2 (SHMT2) is glycine and its biosynthesis and metabolism is upregulated in proliferating cancer cells [11]. The 5-methylthreohydrofolate-homocysteine methyltransferase (MTR), which synthesizes L-methionine and THF from L-homocysteine and 5-methyltetrahydrofolate, was not included in our list (Supplementary Table S1). However, the MTR cofactor vitamin B12 (cobalamin) ATP-transporter MMADHC is overexpressed in cancer (Supplementary Table S1).

3.8. TCA Cycle (Citrate Cycle). Three enzymes (FH, SDHD, and MDH2) were found to be overexpressed in the TCA cycle (32 enzymes, Table 2 and Supplementary Table S1). Even though mutations in FH and SDHD lead to development of tumors [34], overexpression of these three enzymes is likely to lead to accumulation of oxaloacetate. TCA cycle is subject to metabolic reprogramming in cancer cells [35, 36]. In the transformed cells, the tricarboxylic acid (TCA) cycle was active but was characterized by an efflux of substrates for use in biosynthetic pathways, particularly fatty acid synthesis. Glutamine metabolism in these cells supports restoration of oxaloacetate for continued TCA cycle function as well as NADPH production [37]. We found in our list of genes (Supplementary Table S1) the glutamine transporter SLC1A5 that is important for survival of lung cancer cells [38].

3.9. Glycerolipid Metabolism. We observed overexpression of monoglyceride lipase (MGLL) that catalyzes the conversion of monoacylglycerides to free fatty acids and glycerol. Expression of this gene may play a role in tumorigenesis and metastasis. Also aldo-keto reductases AKR1B1 and
AKR1B10 enriched in cancer ESTs catalyze reduction of D-glyceraldehyde to glycerol and 2-hydroxypropanal to propane-1,2-diol in glycerolipid metabolism. A total of three enzymes out of 49 enzymes in this pathway are overexpressed in cancer although the pathway enrichment is not significant (Table 2 and Supplementary Table S1).

3.10. Steroid Hormone Biosynthesis. The same three members of AKRs (AKR1C1, AKR1C2, and AKR1C3) that are involved in metabolism of aldehydes (xenobiotics, drugs) by cytochrome P450 and are overexpressed in cancer (see above) participate in the steroid biosynthesis pathway and could play more than detoxifying role in cancer cells. Only these three enzymes are present in our list among 55 enzymes in this pathway (Table 2).

3.11. Pentose Phosphate Pathway. Two enzymes (ALDOA, PGD) from pentose phosphate pathway (27 genes) are overexpressed in cancer (Table 2 and Supplementary Table S1). This pathway is adjacent to glycolysis pathway and feeds into purine and pyrimidine metabolism producing 5-phospho-α-D-ribose 1-pyrophosphate (PRPP).

3.12. Purine Metabolism/Purine De Novo Biosynthesis. The de novo synthesis of the purine ring is mostly required in cells when DNA replication occurs and the activity of the metabolic pathway in most of tissues is relatively low [39]. Differentiated cells largely employ the salvage pathway, which recycle nucleotides by retrieving the purine ring after nucleic acid or coenzyme breakdown [39].

Four enzymes in the purine metabolism/purine de novo biosynthesis KEGG pathway are overexpressed in cancer (Table 2 and Supplementary Table S1). ATIC encodes a bifunctional protein that catalyzes the last two steps of the de novo purine biosynthetic pathway. ATIC inhibitors are being developed as anticancer therapy [40, 41]. IMPDH2 encodes the rate-limiting enzyme in the de novo guanine nucleotide biosynthesis. These two enzymes are key enzymes in the de novo synthesis of purine nucleotides. The other two enzymes involved in this pathway are not exclusive for this pathway. PKM2 catalyzes the transfer of a phosphoryl group from phosphoenolpyruvate to ADP (GDP), generating ATP (GTP) and pyruvate. The DNA polymerase delta subunit 2 (POLD2) is involved in DNA synthesis and repair [42].

3.13. Pyrimidine Metabolism/Pyrimidine De Novo Biosynthesis. Carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase (CAD) enzyme catalyzing the first three steps in the 6-step pathway of pyrimidine de novo biosynthesis are overexpressed in cancer (Supplementary Table S1). Additionally, the uridine phosphorylase 1 (UPPI) participates in degradation and salvage of pyrimidine ribonucleosides. The DNA polymerase delta subunit 2 (POLD2) incorporates pyrimidine and purine nucleotides in DNA. Three enzymes in this KEGG pathway (98 genes) are overexpressed in cancer (Table 2 and Supplementary Table S1).

3.14. Cysteine and Methionine Metabolism. Adenosylhomocysteinase (AHCY) enzyme that produces L-homocysteine and adenosine by hydrolysis of S-adenosyl-L-homocysteine was found to be overexpressed in cancer (Table 2). This protein is in cysteine and methionine metabolism pathway and also may be used by the MTR enzyme in the one carbon pool by folate pathway (see above). LDHB is another enzyme in the pathway that is overexpressed in cancer (Supplementary Table S1).

3.15. Aminoacyl tRNA Biosynthesis. The cytoplasmic methionyl-tRNA synthetase (MARS) and the cytoplasmic alanyl-tRNA synthetase (AARS) enzymes that charge tRNAs with their cognate amino acids were found to be overexpressed in cancer (Table 2 and Supplementary Table S1).

3.16. Fatty Acid Metabolism. Two enzymes that participate in the fatty acid metabolism are overexpressed in cancer. The mitochondrial enoyl CoA hydratase, short chain 1 (ECHS1) catalyzes the second step of the mitochondrial fatty acid beta-oxidation pathway. Second enzyme is the stearoyl-CoA desaturase (delta-9-desaturase) (SCD), which is involved in synthesis of monounsaturated fatty acids, mostly the oleic acid. Recently, it has been shown that cancer survival is dependent on unsaturated fatty acids and is implicated SCD in this process [43]. Additionally, SCD inhibition causes cancer cell death by depleting monounsaturated fatty acids [44]. We also found that the fatty acid transporter SCP2, which protects fatty acids from oxidation, is overexpressed in cancer. SCD expression is upregulated by retinoic acid in various untransformed cell lines [45]. In our EST database we see an overexpression of cellular retinoic acid binding protein 2 (CRABP2) together with MYCN proteins that have been shown to be upregulated and correlated in variety of cancers [46]. Aldo-keto reductases overexpressed in our cancer dataset (AKR1C1, AKR1B10, and AKR1C3) are involved in the reduction of retinal to retinol (Table 2 and Supplementary Table S1).

3.17. N-Glycan Biosynthesis. Glycosylation is one of the most common posttranslational modification reactions and changes in oligosaccharide structures are associated with many physiological and pathological events, including cell growth, migration, differentiation, and tumor invasion [47]. N-Glycans are involved in cancer progression and MGAT4 mainly participate in branching of N-glycans [48]. Number of Mga4b transcripts increased considerably in diethylnitrosamine-induced hepatocellular carcinoma mice [49]. DPM1 is enzyme that forms dolichol phosphate mannose (Dol-P-Man), which is the mannose donor in pathways leading to N-glycosylation and O-mannosylation. We found these two enzymes to be overexpressed in cancer (2 out of 46 enzymes in this pathway). One enzyme (farnesyl diphosphate synthase (FPDS) from the terpenoid biosynthesis pathway (1 out of 15 enzymes) that feeds in N-glycan biosynthesis pathway was also overexpressed in cancer (Table 2 and Supplementary Table S1).

3.18. Phenylalanine/Tyrosine/Histidine Metabolism. Three enzymes (ALDH3B1, ALDH3B2, and ALDH3A1) that
participate in the oxidation of acetaldehyde to acetate in the glycolysis/gluconeogenesis pathway and are involved in xenobiotics and drug oxidation by cytochrome P450 also participate in metabolism of amino acids phenylalanine (18 enzymes), tyrosine (42 enzymes), and histidine (29 enzymes) (Table 2 and Supplementary Table S1). We did not attempt to analyze enrichment of phenylalanine/tyrosine/histidine metabolism due to the obvious overlap with the glycolysis/gluconeogenesis pathway that is highly enriched in cancer-associated genes (Table 2).

3.19. Enzymes Introducing Posttranslational Modifications. Introduction of disulfide bonds by Cys mutations has been shown to improve the physical stability of some proteins [50]. We observed that three disulfide isomerases, AGR2, AGR3, and TXNDC5, are overexpressed in cancer (Table 2). AGR2 is the prionocogenic protein that could be used as a tumor biomarker [51]. AGR3 is overexpressed by a hormone-(estrogen-receptor α-) independent mechanism and identifies a novel protein-folding associated pathway that could mediate resistance to DNA-damaging agents in human cancers [52].

We also observed that one atypical excreted disulfide oxidase quiescin sulphydryl oxidase 1 (QSOX1), which supports cell-matrix adhesion and cell migration [53], is markedly overexpressed in cancer (Supplementary Table S1).

Third class of protein modifying enzymes we observed in our dataset (Supplementary Table S1) is peptidyl-prolyl isomerases (PPIA) that are known to accelerate protein folding. PPIA is one member we found in our list (Supplementary Table S1). Another member is the peptidyl-prolyl isomerase FKBP4 that is a cochaperone which activates RNA interference-mediated silencing in mammalian cells [54, 55].

3.20. Additional Antioxidant Systems. Our list of genes overexpressed in cancer is also enriched in certain antioxidant enzymes (peroxiredoxins 1 and 3 (PRXD1 and PRXD3) and cytochrome b561 (CYB561)) (Supplementary Table S1). Peroxiredoxins reduce hydrogen peroxide and alkyl hydroperoxides and CYB561 is involved in a reduction of ascorbate radicals.

4. Discussion

The use of EST to measure gene expression requires a lot of caution because many libraries may have insufficient coverage of low and moderately expressed genes. It should be noted that EST-based expression analyses were used in several studies of cancer cells [15, 16]. We assume that cancer ESTs are a good tool to study general properties of gene highly expressed in cancer cells because the statistics were collected over a large number of EST libraries that compensate to some extent a semiquantitative nature of any EST-based expression measure. We have no doubt that when RNAseq data will be available for a range of libraries/tissues similar to dbEST, this will substantially improve the generalized portrait of cancer metabolic pathways.

There are many computational approaches for analysis of metabolic and signaling pathway enrichment [56]. We used the simplest approach in this paper: the Fisher exact test (Table 2). The DAVID system [57] uses a more conservative implementation of this test. We analyzed the dataset using the DAVID system and found a very similar list of metabolic pathways although only three pathways were significantly enriched in genes that were found to be overexpressed in cancer (Supplementary Table S1): glycolysis/gluconeogenesis (EASE Score = 1.2 × 10^{-4}), oxidative phosphorylation (EASE Score = 2.6 × 10^{-5}), and metabolism of xenobiotics by cytochrome P450 (EASE Score = 4.5 × 10^{-2}). We do not see this as a contradiction with our results (Table 2) and differences in P values are likely to reflect known methodological problems with the analysis of pathway enrichment [56]. We also observed obvious problems with some KEGG pathways: for example, for the purine metabolism, the de novo purine biosynthesis pathway and the salvage pathway are merged into one KEGG pathway (Table 2). However it was shown that only de novo purine biosynthesis is overexpressed in cancer cells [40, 41]. The same problem was found for the pyrimidine metabolism (Table 2). Therefore the insignificant P values for the pathway enrichment (Table 2) do not necessarily indicate that such pathways are not important for cancer initiation and progression.

There are numerous attempts to build a census of human cancer genes [12–14, 58, 59]; for example, Futreal et al. [58] and Santarius et al. [12] identified ~400 candidate cancer-related genes. An important direction along this venue of research is development of predictive models for cancer-associated genes that could accelerate their identification. Such models have been developed for specific types of cancer. One example of such studies is a complex statistical model for the prediction of prostate cancer genes [60]. In our study we applied a complementary approach using simplified statistical filters for prediction of genes that are overexpressed in all available cancer EST libraries without classifying them into types and subtypes (Table 1 and Supplementary Table S1). In other words, we tried to delineate a list of broadly overexpressed genes and a generalized portrait of cancer metabolic pathways that are expected to be overrepresented in this list. Another important difference is that in this list we do not expect that genes involved in metabolic pathways have many somatic nonsynonymous mutations that are likely to cause inactivation or gain of a new function; an excess of such mutations has been used in previous attempts to build a census human cancer genes [12–14, 58, 59]. Genes that are found in many cancer EST libraries (Supplementary Table S1) should be marked (or even removed) as nonspecific cancer-associated genes when researchers analyze genes overexpressed in certain types of cancer. On the other hand, we know that cancer modifies normal metabolic processes to fulfill its high growth and energy demands and the above described metabolic pathways seem to be universally central for cancerous growth and progression. Enzymes that are highly expressed in cancer along these metabolic pathways could provide multiple targets for desirable inhibition of cancer progression.
Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors thank Mikhail Galperin, Youri Pavlov, Svetlana Shabalina, and Eugene Koonin for useful discussions. This work was supported by the National Institutes of Health Intramural Research Program of the National Eye Institute and the Intramural Research Program of the National Library of Medicine at the National Institutes of Health (US Department of Health and Human Services).

References


Review Article

Complexity of Gene Expression Evolution after Duplication: Protein Dosage Rebalancing

Igor B. Rogozin

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Correspondence should be addressed to Igor B. Rogozin; rogozin@ncbi.nlm.nih.gov

Received 26 March 2014; Accepted 3 August 2014; Published 17 August 2014

Academic Editor: Eugenia Poliakov

Copyright © 2014 Igor B. Rogozin. 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.

Ongoing debates about functional importance of gene duplications have been recently intensified by a heated discussion of the “ortholog conjecture” (OC). Under the OC, which is central to functional annotation of genomes, orthologous genes are functionally more similar than paralogous genes at the same level of sequence divergence. However, a recent study challenged the OC by reporting a greater functional similarity, in terms of gene ontology (GO) annotations and expression profiles, among within-species paralogs compared to orthologs. These findings were taken to indicate that functional similarity of homologous genes is primarily determined by the cellular context of the genes, rather than evolutionary history. Subsequent studies suggested that the OC appears to be generally valid when applied to mammalian evolution but the complete picture of evolution of gene expression also has to incorporate lineage-specific aspects of paralogy. The observed complexity of gene expression evolution after duplication can be explained through selection for gene dosage effect combined with the duplication-degeneration-complementation model. This paper discusses expression divergence of recent duplications occurring before functional divergence of proteins encoded by duplicate genes.

1. Models of Gene Duplications

With the increasing availability of genomic data, it became clear that numerous gene families have diverged in function through series of duplications, including many lineage-specific expansions (or gene copy-number variations (CNVs) at the population level) identified in each of the genomes sequenced [1–8]. This is not surprising taking into account that gene duplications are traditionally considered to be a major evolutionary source of new protein functions [1, 2, 6, 9]. The classic concept of the functional consequences of gene duplication, proposed by Susumu Ohno, holds that duplication produces two functionally redundant, paralogous genes and, thereby, frees one of them from selective constraints. This unconstrained paralog is then free to accumulate mutations that would have been deleterious in a unique gene but become neutral after the duplication [9]. Although the most likely outcome of such neutral evolution is for one of the paralogs to fix a null mutation and become a pseudogene, there is also the possibility of fixation of mutations that lead to a new function [10–15]. One of the predictions of this neofunctionalization model of evolution of duplicate genes is the rapid loss of paralogs due to null mutations [10, 14, 15]. However, this prediction was not supported by studies on isozyme spectra of polyploids in a number of organisms [16]. Furthermore, analysis of duplicate genes in the tetraploid frog Xenopus laevis has shown that both copies were subject to purifying selection [17], contrary to the prediction of the neutrality of one of the copies [9]. The failure of empirical research to support Ohno’s model has led to the proposal of alternative hypotheses, in particular, the general subfunctionalization model [2, 11], the more specific duplication-degeneration-complementation (DDC) model [2], and the dosage effect model [3].

The subfunctionalization hypothesis is based on the same assumptions as the Ohno’s model, namely, that newly duplicate genes are redundant in function and, accordingly, a duplication event is selectively neutral [2, 11–13, 18]. However, it was argued that, as natural selection does not “know” in advance which duplicate gene should be under selection...
and which remains free of selective constraint, both paralogs experience a period of relaxed selection and accelerated evolution. Measurements of the selection pressure affecting paralogs shortly after duplication appear to be compatible with this reasoning [12]. During this period, both genes are likely to accumulate mutations that impaire different functions of the ancestral gene, so that, after a certain point, none of the paralogs is capable of substituting for the ancestor [2, 12].

The duplication-degeneration-complementation (DDC) hypothesis is a special case of the subfunctionalization scenario. This hypothesis requires multiple losses of gene expression across tissues/cell types [2]. Under the DDC hypothesis, DNA regulatory elements are duplicated during gene duplication events. Subsequently, mutations increase specialization of gene function by degenerating modular regulatory elements in a complementary fashion in the duplicate genes, a process that is thought to contribute to the long-term preservation of duplicate genes [2]. The DDC model suggests that (1) degenerative mutations in regulatory elements and a divergence of expression patterns can increase rather than reduce the probability of duplicate gene preservation and (2) the usual mechanism of duplicate gene preservation is the partitioning of ancestral functions (e.g., expression profiles across tissues) rather than the evolution of new functions [2].

A major problem with the evolution of duplicate genes is that the creation of novel gene functions generally provides a long-term but not a short-term advantage for gene duplication. However, duplicate genes also appear to affect fitness immediately after duplication, providing a short-term advantage for duplication, conceivably through the gene dosage effect [3]. In the last decade, data have accumulated demonstrating that although a gene duplication does not necessarily double the dosage of the gene product, it nevertheless generally leads to the production of an increased amount of the product [19–21]. Some gene duplications (or gene copy-number variations (CNVs) at the population level) actually appear to be selected against due to the increase in gene dosage, including cases of gene duplications contributing to disease [5, 21–23]. Thus, the relative contributions of different factors to the evolution of paralogous genes after duplication remain a subject of intensive research and debate [7, 21, 24].

2. Ortholog Conjecture and Gene Duplications

The importance of appropriately designed studies to test various models of gene evolution between orthologs and paralogs was emphasized by Studer and Robinson-Rechavi [25]. It was suggested that functional changes between orthologs might be as common as between paralogs (the “uniform model”) and that more studies should be designed to test the impact of different models [25]. Robust identification of orthologs is of central importance for comparative and functional genomics due to a rarely stated but almost universally implied concept that recently has been denoted by ortholog conjecture (OC) [26]. The OC holds that orthologous genes perform equivalent functions in the respective organisms and, accordingly, experimentally determined functions of a gene can be transferred to its experimentally uncharacterized orthologs in other species (certainly, taking into account the biological differences between the organisms involved) [4, 26, 27]. Nehrt and coworkers argued that only rarely has it been noted that the OC is just a hypothesis although in most studies it is either assumed to be true or supported by evidence from a small number of genes. Therefore, Nehrt and coworkers decided to systematically test the OC hypothesis [26]. They used experiment-based annotations in the gene ontology (GO) database [28] and microarray gene expression data [29] to compare the functional and expression similarities of orthologs and paralogs in human and mouse [26]. They showed that at the same level of protein sequence divergence (i) orthologs are less similar than paralogs and (ii) between-species paralogs are less similar than within-species paralogs [26]. They further showed that (iii) functional and expression similarities between orthologs are independent of the protein sequence identity between the orthologs. These results are inconsistent with the OC hypothesis, prompting Nehrt and coworkers to propose that the primary determinant of the evolutionary rate of gene function and expression is a cellular context in which the genes act [26]. This “cellular context” hypothesis could explain why within-species paralogs were observed to be more similar in function and expression than between-species paralogs and orthologs [26].

Several consequent studies suggested that GO annotations should be used to test the OC hypothesis with a great caution [30, 31] or even should not be used for this purpose [32]. A general consensus is that GO annotations are compatible with the OC hypothesis [30, 32], although Altenhoff and coworkers suggested that GO annotations are better compatible with the “uniform” model [31]. In addition, Chen and Zhang [32] analyzed a large RNA-Seq dataset of multiple tissues and showed that the expression similarity between orthologs is significantly higher than that between within-species paralogs, supporting the OC hypothesis and refuting the “cellular context” hypothesis for gene expression [32].

Rogozin and coworkers reanalyzed these controversial results using approaches different as much as possible from those used before and reconciled them with the literature on gene duplications [34]. This analysis of a large RNA-Seq dataset of multiple tissues from human and mouse shows that rank/Z-score measures of the expression similarity between orthologs are significantly higher than that between within-species paralogs (Figures 1(a) and 1(b)), supporting the OC. This result is consistent with Chen and Zhang study [32]. The plots of expression similarity measured using linear or rank correlation coefficients were qualitatively similar to the analogous plots reported by Nehrt and colleagues [26] (Figures 1(c) and 1(d)) in that the strongest correlation was observed among within-species paralogs, followed by orthologs and then by between-species paralogs. For the between-species paralogs, significant expression similarity was observed only at low sequence divergence whereas at higher divergence, the correlation coefficient values were much lower (Figures 1(c) and 1(d)). Although the correlation among within-species paralogs was high for all values of sequence divergence, it also dropped with increasing divergence (Figures 1(c) and 1(d)). It was suggested that different measures of expression similarity...
could reflect different salient features of gene expression, namely, tissue-specificity in the case of the correlation coefficients and relative abundance of individual mRNAs in the case of Z-scores and ranking scores [34]. Further analysis in which expression profiles of orthologs and paralogs were compared separately for individual gene families provided a strong argument in support of the OC [34]. However, the OC, all its importance notwithstanding, reflects only one aspect of gene evolution. The complete picture must integrate vertical descent encapsulated in the OC with the lineage-specific aspects of the evolution of paralogs; it should be interpreted also in the context of various hypotheses on evolution of gene duplications [34].

3. Synthetic “Protein Dosage Rebalancing” Hypothesis

The major difference between the dosage effect model and the DDC model is the role of natural selection. The dosage effect model implies that paralogs are subject to purifying selection from the onset of evolution after the gene duplication [3, 7] whereas the DDC model assumes “constructive neutral evolution” [14] whereby the paralogs are maintained due to the partial, differential degeneration of their functions resulting in functional complementarities [2, 6, 35].

Results of previous studies of the “ortholog conjecture” hypothesis [26, 32, 34] are consistent with both models. (1) A significant positive correlation between gene expression and sequence divergence was found for within-species paralog measurements; this is best consistent with the dosage effect hypothesis (Figures 1(c) and 1(d)); (2) a significant difference between paralogs was found for all comparisons (Figures 1(a) and 1(b)); this is best consistent with the DDC hypothesis.

The neutrality of degenerative mutations assumed under the DDC model is amenable to a straightforward statistical test. Consider three genes, X, Y1, and Y2, in two species, among which Y1 and Y2 are lineage-specific paralogs in one species and X is the single ortholog of this pair of paralogs in the other species. Then, the neutral evolution under the strict DDC model predicts the following relationships between the expression profiles of the three genes: the profile of the gene X (E_x) is expected to show a greater similarity to the combination of the profiles of the genes Y1 and Y2 (E_y = E_{y1} + E_{y2}) than to either E_{y1} or E_{y2}, given the differential degeneration of the expression of the two

---

**Figure 1:** Expression and sequence similarity of orthologous and paralogous genes. (a) Z-score expression similarity averaged across 4 tissues. (b) Rank-based expression similarity averaged across 4 tissues. (c) Kendall’s τ rank correlation coefficient. (d) Pearson linear correlation coefficient. The raw data is taken from Rogozin and coworkers [34]; see Table 1 for more details about procedures used in this study.
paralogs. I identified all X-Y1-Y2 triplets within the human-mouse clusters of orthologs and paralogs (see Table 1 for details) [34] and compared the expression profiles of the respective genes. The results of this analysis revealed poor agreement with the neutral prediction: in a majority of the gene triplets, $E_x$ shows a greater similarity to $E_{y1}, E_{y2}$, or both than to the combined profile $E_y$, although this excess is not significantly different from the uniform distribution (Table 1).

This result is consistent with many previous observations. For example, Huminiecki and Wolfe examined how the gene expression profiles of orthologous gene sets in human and mouse are affected by the presence of recent species-specific paralogs [36]. Gene expression profiles were compared across 16 homologous tissues in human and mouse genomes using microarray data for 1,575 sets of orthologous genes including 250 with species-specific duplications. It was found that there is a general trend for paralogous genes to become more specialized in their expression patterns, with decreased breadth and increased specificity of expression as gene family size increases [36]. Often, the expression of both copies of a duplicate gene is likely to have changed relative to the predicted ancestral state [36].

An interesting example of a highly redundant genome is the microcrustacean Daphnia pulex genome which contains at least 30,907 genes [37]. This high gene count is a consequence of an elevated rate of gene duplication resulting in tandem gene clusters. More than a third of Daphnia's genes have no detectable homologs in any other available genomes, and the most amplified gene families are specific to the Daphnia lineage [37]. The coexpansion of gene families interacting within metabolic pathways suggested that the maintenance of duplicate genes is not random, and the analysis of gene expression under different environmental conditions revealed that numerous paralogs acquire divergent expression patterns soon after duplication events [37]. It was suggested that the persistence of some functionally divergent gene duplicates in Daphnia is likely to be due to preservation by entrainment (PBE) [37]. Entrainment was defined as the process of increasing the initial probability of preserving a duplicate gene through its functional interaction with existing or newly interacting genes sharing regulatory programs [37]. For example, genes with divergent expression patterns at the time of duplication, yet with regulation sufficiently similar to the expression patterns of a different interacting gene, may have combined products that are beneficial under a distinct environmental condition. In this scenario, the likelihood for preservation of these new gene duplicates is increased [37]. Thus, when genes are advantageous at the time of duplication, their coding regions are subject to purifying selection from the start and are entrained to a distinct regulatory pattern dictated by condition-specific gene-gene interactions [37].

Many studies have shown that gene duplicates in eukaryotes tend to have divergent expression patterns and that gene family expansions are associated with high levels of tissue specificity [37–44]. However, the timeframe in which these processes occur has rarely been investigated in detail, and most analyses do not include direct comparisons of orthologs as a baseline for the expected levels of tissue specificity in absence of duplications. To assess the contribution of duplications to expression divergence, Huerta-Cepas and coworkers combined phylogenetic analyses and expression data from human and mouse [42]. They analyzed differences in gene expression among human-mouse paralogs, specifically duplicated after the radiation of mammals, and compared them to pairs of orthologs in the same species. It was shown that gene duplication leads to increased levels of tissue specificity and that this tends to occur promptly after the duplication event [42].

Similar observations have been reported previously for paralogous genes in yeast [45] and fly [46]. Oakley and coworkers used a phylogenetic approach to demonstrate that the fast evolutionary rate of tissue-specific repression or loss of gene expression is significantly higher than the rate of activation or gain [46]. It was also found that DDC is consistent with only a portion of possible ancestral histories of gene expression [46]. Conceivably, the observed trend for paralogs to become more specialized in their expression patterns than expected from the strict DDC model (Table 1) as well as a significant positive correlation between gene expression and sequence divergence for within-species paralog measurements (Figures 1(c) and 1(d)) and the nonmonotonic dependency of the $Z$-scores and ranking scores on sequence divergence (Figures 1(a) and 1(b)) can be explained by selection for rebalancing of expression in different tissues and environmental conditions (Figure 2) [34]. This scenario, the “protein dosage rebalancing” [34], is consistent with several previous studies which suggest that rebalancing of expression after duplications, at least for some genes, could be beneficial [37–39, 41–44]. For example, Qian and colleagues have shown that yeast and mammalian genes often experienced a significant decrease in the level of expression after duplication. It was suggested that although the majority of the expression reduction is likely to be neutral, for some of duplicate genes, it could be beneficial through the rebalanced gene dosage [41].

4. Copy-Number Variations

Copy-number variations (CNVs) are alterations of a genome that results in individuals having an abnormal or, for certain genes, a normal variation in the number of copies of one or more sections of the genome. CNVs correspond to relatively large regions of the genome that have been deleted (fewer than the normal number) or duplicated (more than the normal number) on certain chromosomes. CNVs account for roughly 12% of the human genome and each variation may range from about one kilobase (1,000 nucleotides) to several megabases in size [47]. As any mutation, a duplication event by itself may have consequences on the organism's
fitness. However, two factors complicate studies of the short-term immediate fitness effects of gene duplication [7]. First, the conceptual appeal of gene duplications leading to novel functions was strong enough to overshadow potential short-term fitness effects of duplications [7]. Second, there are major technical difficulties in studying CNVs that persist to this day [48, 49]. One of the most obvious problems is analysis of expression levels for recently duplicated genes and CNVs. This problem is even worse for the Affymetrix microarray probes that have been designed to represent the unique portions of a gene. Each probe sequence is scanned against the available genomic sequence to minimize cross hybridization between duplicate genes. This process has a drawback of excluding many recently duplicated genes and CNVs from a microarray because unique probes cannot be designed for them [38].

CNVs were implicated in many human genetic diseases [50]; for example, it was suggested that rare CNV is an important source of risk for autism spectrum disorders (ASDs) [49, 51]. Pathogenic CNVs, often showing variable expressivity, included rare de novo and inherited events at over 30 gene loci, implicating several ASD-associated genes.
previously linked to other neurodevelopmental disorders [51]. It seems likely that the synergistic action of environmental hazards with genetic variations (including CNVs) that, in themselves, have limited or no deleterious effects but are potentiated by the environmental factors and result in dosage imbalance of neuron-specific proteins is a general principle that underlies the alarming increase in the ASD prevalence [52]. Genes affected by de novo CNVs converge on networks related to neuronal signaling and development, synapse function, and chromatin regulation [51]. These and many other observations of positive and negative fitness effects of CNVs [7] raised a question about validity of the so-called “backup” hypothesis (functionally redundant paralogs are used to backup important functions in the event of a severe mutation). It was suggested that the “backup” hypothesis is not supported by the analysis of expression data [41, 53]. This is consistent with the theoretical population genetic analysis by Clark [54]. It was concluded that the genetic robustness against mutations conferred by paralogous genes is a byproduct of other evolutionary processes [41]. Those processes may be extremely complicated; for example, in several cases, it appears that a gene duplication that is adaptive under a stressful condition comes at a fitness cost in a benign environment [7, 55].

5. Concluding Remarks

The concept of genetic balance traces back to the early days of genetics. Additions or subtractions of single chromosomes to the karyotype (aneuploidy) produced greater impacts on the phenotype than whole-genome changes (ploidy) (reviewed by [56]). Studies on changes in gene expression in aneuploid and ploidy series revealed a parallel relationship leading to the concept that many genes exhibited a stoichiometric balance, which, if upset, would modulate gene expression and protein dosage. Studies of retention of selected duplicate genes following diploidization of ancient polyploidization events have found that many duplicate genes have been preferentially maintained in a dosage-sensitive relationship [56]. Furthermore, it was hypothesized that stoichiometric alterations of macromolecular complexes or cellular networks are responsible for dominant phenotypes, because of the existing nonlinear relationships between the genotypic and phenotypic values with which they are associated [39, 57].

Many observations described in this paper are best consistent with the following possible scenario of gene duplications: many recent gene duplications (or rather gene copy-number variations (CNVs) at the population level) have a positive effect in some tissues and/or environmental conditions, whereas they also have a negative effect in some other tissues and/or environmental conditions (Figure 2) [3, 7, 21–23]. It seems likely that balancing of positive and negative dosage effects is an important factor which is causing diversification of expression patterns (rebalancing of expression) of duplicate genes in the course of fixation of gene duplications (Figure 2). This process is influenced by natural selection similar to the conventional dosage effect hypothesis [3]. After the gene duplication is fixed in a population, preservation of this gene duplication may be largely explained by the DDC model (maintenance of duplicate genes due to differential loss or reduction of expression in various tissues) that predicts that the usual mechanism of duplicate gene preservation is the partitioning of ancestral functions (expression profiles across tissues) rather than the evolution of new functions [2]. The suggested synthetic model, the “protein dosage rebalancing” model [34] (Figure 2), is a combination of the dosage effect [3] and DDC [2] models assuming importance of both natural selection and neutral evolution for maintenance of gene duplications. The “protein dosage rebalancing” model reverberates to some extent with the new mutation theory of phenotypic evolution which suggests that the driving force of phenotypic evolution is mutation, and natural selection is of secondary importance [58]. It is important to emphasize that the “ortholog conjecture,” all its importance notwithstanding, reflects only one aspect of gene evolution. The complete picture of eukaryotic evolution must integrate vertical descent encapsulated in the “ortholog conjecture” with the lineage-specific aspects of the evolution of paralogs [34, 59, 60]. This approach is embodied in a recently developed novel approach for computational annotation of gene function that incorporates information on both orthology and paralogy and yields significantly more annotations at the same average precision than a model that includes only orthologs [61].

Conflict of Interests

The author declares that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The author thanks David Manadadze, Yuri Wolf, Svetlana Shabalina, Pere Puigbo, and Eugene Koonin for useful discussions and Jean Thierry-Mieg for helpful advice on RNA-Seq data analysis. This work was supported by the Intramural Research Program of the National Library of Medicine at National Institutes of Health (US Department Health and Human Services).

References


