Send Orders for Reprints to reprints@benthamscience.ae

Current Genomics, 2016, *17*, 368-378

Bioinformatics Approach in Plant Genomic Research

Quang Ong[a,b], Phuc Nguyen[c], Nguyen Phuong Thao[c,*] and Ly Le[c,*]

[a]*Plant Abiotic Stress Research Group, Ton Duc Thang University, Ho Chi Minh City, Vietnam;* [b]*Faculty of Applied Sciences, Ton Duc Thang University, Ho Chi Minh City, Vietnam;* [c]*School of Biotechnology, International University, Vietnam National University, Ho Chi Minh City, Vietnam*

L. Le

Abstract: The advance in genomics technology leads to the dramatic change in plant biology research. Plant biologists now easily access to enormous genomic data to deeply study plant high-density genetic variation at molecular level. Therefore, fully understanding and well manipulating bioinformatics tools to manage and analyze these data are essential in current plant genome research. Many plant genome databases have been established and continued expanding recently. Meanwhile, analytical methods based on bioinformatics are also well developed in many aspects of plant genomic research including comparative genomic analysis, phylogenomics and evolutionary analysis, and genome-wide association study. However, constantly upgrading in computational infrastructures, such as high capacity data storage and high performing analysis software, is the real challenge for plant genome research. This review paper focuses on challenges and opportunities which knowledge and skills in bioinformatics can bring to plant scientists in present plant genomics era as well as future aspects in critical need for effective tools to facilitate the translation of knowledge from new sequencing data to enhancement of plant productivity.

Keywords: Bioinformatics, Comparative genomics, GWAS, Next-generation sequencing, Plant genomics, Phylogenomics.

Received: June 22, 2015 **Revised: September 11, 2015** **Accepted: September 18, 2015**

1. INTRODUCTION

The Plant kingdom is very important not only for human but also for other living organisms. One of the crucial role of plants is to provide a huge amount of food [1]. Plants are also used in making many human medicines [2] and have been selected as model organisms to study transposable elements in heterochromatin and epigenetic control [3]. Study of plant biology has, therefore, been conducted broadly since the early stage of human life because of its vital role.

Modern technologies have pushed the study of plant biology to a higher level than before [4]. The innovation of high-throughput sequencing methods gives scientists the ability to exploit the structure of the genetic material at the molecular level which is known as "genomics". Plant genomics study has exploded recently and becomes the main theme in plant research due to the rapid increase of sequenced genomes of many plant species [5]. It is easy to see the huge impact of plant genome research on the improvement of economically important plants and the knowledge of plant biology [6]. Open-access and constant updates to this plant genomic information create a fertile environment for plant research to grow. This requires strong connection and cooperation among global biological community [7].

In this paper, we review firstly the development of genomic sequencing technologies and their applications in plant genomic research. Then, we introduce recent approaches of bioinformatics in managing and analyzing plant genomic databases. Particularly, we summarize most popular plant genomic resources. In addition, we also provide fundamental knowledge of key methods for integration and analysis of these genomic data such as comparative genomic analysis, phylogenomics, evolutionary analysis and genome-wide association study (GWAS) in plant.

2. NEXT GENERATION SEQUENCING TECHNOLOGY IN PLANT GENOMIC RESEARCH

The development of DNA sequencing technology has been a great and memorial journey filled with many historical events. In the last decade, nearly all of DNA sequence production has restrictively been executed with capillary-based, semi-automated applications of the Sanger biochemistry and its variations [8-10]. Over the years, the field of DNA sequencing has been revived and prospered due to various scientific breakthroughs. These technological advancements eventually lead to the encouragement for developing novel experimental designs for this field due to various reasons [11]. Ultimately, next-generation sequencing (NGS) technologies were released in 2005 [12]. They are known as "high throughput sequencing technologies that parallelize the sequencing process, producing millions of sequences at once at a much lower per-base cost than conventional Sanger sequencing" [13].

Based on NGS technologies, big companies like Roche, Illumina, Applied Biosystems and so forth have recently developed many autonomous and ultrahigh-throughput platforms. All of them are all well-fitted for the current and even future large sequence needs. Generally, Sanger's dideoxy chain termination sequencing technology is no longer utilized in these NGS platforms. Instead, more advanced methods are applied such as pyrosequencing, sequencing-by-

*Address correspondence to these authors at the School of Biotechnology, International University, Vietnam National University HCMC; Quarter 6, Linh Trung Ward, Thu Duc District, Ho Chi Minh city, Vietnam; Tel/Fax: +84-8-3724-4270, +84-8-3724-4271;
E-mails: npthao@hcmiu.edu.vn; ly.le@hcmiu.edu.vn

1875-5488/16 $58.00+.00 ©2016 Bentham Science Publishers

Current Genomics

synthesis, sequencing-by-ligation, ion semiconductor-based non-optical sequencing, single molecule sequencing and nanopore sequencing [14].

Sequencing-by-synthesis platform utilizes DNA polymerase to extend many DNA strands in parallel [15]. This method uses modified deoxynucleoside triphosphates (dNTPs) containing a terminator which prevents further polymerization, thus, only one single base can be added by DNA polymerase to each growing DNA copy strand. Therefore, the newly incorporated nucleotide or oligonucleotide can be determined as extension proceeds. The pyrosequencing platform is based on the principle of sequencing-by-synthesis (SBS) [16]. It relies on the detection of pyrophosphate released on nucleotide incorporation by DNA polymerase to facilitate a following series of enzymatic reactions that finally produces light signal from the cleavage of oxyluciferin by luciferase. Sequencing-by-ligation platform uses DNA ligase to create sequential ligation of dye-labeled oligonucleotides. This process enables massively parallel sequencing of clonally amplified DNA fragments [17]. The discrepancy sensitivity of these clonally amplified DNA fragments is then used to determine the hidden sequence of the target DNA molecule. Ion semiconductor-based non-optical sequencing platform detects the hydrogen ions which are released during DNA polymerization. Single molecule sequencing is based on "the successive enzymatic degradation of fluorescently labeled single DNA molecules, and the detection and identification of the released monomer molecules according to their sequential order in a microstructured channel" [18]. Single molecule sequencer does not require any amplification of DNA fragments prior to sequencing [19]. Nanopore sequencing identifies individual nucleotide sequences as the DNA strand is passed through a membrane-inserted protein nanopore, one base at a time, by alterations in the ion current [20].

Some examples for well-known NGS platforms commercially available are Genome Sequencer from Roche/454 (Pyrosequencing); Genome Analyzer from Illumina/Solexa (Sequencing-by-synthesis); SOLiD from Applied Biosystems (Sequencing-by-ligation) and Polonator from Dover SystemsP (Sequencing-by-ligation), Ion Torrent from Life Science, Inc. (Ion semiconductor-based non-optical sequencing); Heliscope sequencers from Helicos Bioscience Corporation (True single molecule sequencing); PacBio RS sequencers from Pacific Biosciences (Single molecule, real-time sequencing); GridION and miniaturized MinION sequencers from Oxford Nanopore Technologies (Nanopore sequencing) [4, 14].

The main differences among these systems are the length of a sequence read, the unique error model that they applied and the operation cost [21-24]. These differences may affect how the reads are utilized in bioinformatics analyzes, depending upon the application [19]. However, most of the results finally showed that the data produced are similar among these methods [21-24]. Therefore, it mainly depends on the ultimate goal of a particular research that one may choose the appropriate sequencing methods.

With its rapid innovation, NGSs have been well applied to many aspects in plant genomic research, such as exome sequencing and studying genetic transmission of alleles/quantitative trait loci (QTLs) through whole genome sequencing [14]. Exome sequencing can effectively help in exploring biodiversity, studying host–pathogen interactions, investigating the natural evolution of crops, testing for the inheritance of genetic markers, providing large-scale genetic resources for the crop improvement, identifying the genes and establishing the presence of functional gene sets that are involved in symbiotic or other co-existential systems [14]. In addition, NGS methods with single-base resolution can provide epigenomic information. For instance, a study in *A. thaliana* epigenome revealed that the location and abundance of small RNA targets were significantly related to cytosine methylation [25]. Another application of plant genome sequencing is genotyping by sequencing (GBS), which is emerging as high through-put and inexpensive method for optimizing genotype populations. GBS has many approaches for enhancing genomic map construction, especially single nucleotide polymorphisms (SNPs) identification [26]. 681,257 SNP markers of 2,815 maize inbred accessions were found to be positively associated with trait related genes by performing GBS [27].

The successful application of NGSs in plant genomic research is undoubtable. However, there are challenges in developing computational tools for analyzing genome sequences. Galaxy (http://galaxyproject.org) is one of the software systems in which researchers can easily use analysis tools through web-based interfaces comprised of enormous free-accessed biological data [28]. Another software is Artemis, which is freely available from Sanger institute (http://www.sanger.ac.uk/). It provides genome browser and annotation tool [29]. There are several other genome sequence analysis tools given by The Broad's Genome Sequencing and Analysis Program (GSAP) (http://www.broadinstitute.org/). Additionally, the rapid decrease in cost of genome sequencing leads to the urgent requirement of a development of huge database storage and management. In fact, there are more and more plant genomic databases have been generated to confront with that demand.

3. PLANT GENOMIC RESOURCES

The history of plant genomics has been changed dramatically by the creation of expressed sequence tag (EST) sequencing, a high-throughput gene discovery method [30], and the release of the complete *Arabidopsis thaliana* genomic sequence in 2000 [31]. Following that success, the complete genomic sequence of rice became available only 2 years later [32]. These events have created powerful waves on both plant biotechnology and crop bioinformatics. For the advancement of learning, more sequencing projects on vital plant species have been carried out by combining novel *in silico* technologies from genomic research with traditional breeding schemes for further enhancing the quality of crops.

With the advent of NGS technology in 2005 [33], the number of plant genomes sequenced have dramatically increased to more than 100 species in 2014 according to CoGepedia, a platform that aims to record all plant genomes with published or in-processed sequences [12]. Throughout the years, these genomes have contributed many valuable materials for plant research in modern molecular genomics era. Based on that foundations, genetical/biological activities

of many critical genes and pathways have been revealed [34]. For instance, plant species such as *Arabidopsis* [31], *Brachypodium distachyon* (grass) [35], *Physcomitrella patens* (moss) [36] and *Setaria italic* (millet) [37, 38] can be used as scientific model for genomic studies in drought tolerance [39]. Others like *Oryza sativa* (rice) [40, 41], *Populus trichocarpa* (poplar) [42], *Zea mays* (maize) [43], *Glycine max* (soybean) [44], *Solanum lycopersicum* (tomato) [45], and *Pinus taeda* (loblolly pine) [46] can serve as both crops and functional models [34].

Non-model and non-crop plant genomes can also tell a story about genome construction and flowering plant evolution [34]. For examples, *Utricularia gibba* (bladderwort) and *Genlisea aurea* (corkscrew) genomes can provide significant understanding about genome size variation [47, 48]. Furthermore, *Spirodela polyrhiza* (greater duckweed) genome which share the similarity in size with that of *Arabidopsis* but only needs 28% fewer genes to function normally [49]. In another case, the genomes of *Selaginella moellendorffii* (spikemoss) and *Amborella trichopoda* present the bridge between the evolution of vascular plants and angiosperms respectively, revealing fundamental understandings about the trajectory of plant specific gene families and the radiance of flowering plants, thus, shedding more light in the evolution of flowering plant [34].

The gene knowledge drawn from genomics can be utilized to recognize, classify, exploit and tag individual alleles as well as to promote and manipulate molecular markers to track the desired alleles in breeding programs [50]. For those reasons, many genome sequencing projects in the field of horticultural crops were carried out such as Tomato genome sequencing project (www.sgn.cornell.edu/about/tomato) [45], Potato genome sequencing consortium, (www.potato genome.net) [51], Papaya genome sequencing project (www.asgpb.mhpcc.hawaii.edu/papaya/) [52], Grape genome sequencing project (www.vitaceae.org) [53], Floral genome sequencing project (www.fgp.bio.psu.edu/) [54] and hopefully many more will be available in public domain for scientific usages in near future. Combining with traditional methods, these projects were armed with advanced sequencing technologies, to fully certify generation of high-quality sequences and budget-efficient design [55]. Therefore, these whole-genome sequencing projects may have great significant impact in global food insurance and bio-energy advancement by providing invaluable resources for comparative and functional genomic studies [55]. If current research keeps moving forward, noticeable impact on global human well-being may be seen through applications of genomic science resources to horticulture plant species.

The availability of complete genome sequences, as well as the explosion of sequence data, is leading to an urgent need for well-catalogued and annotated DNA sequence databases. The largest and most well-known of these sequence databases are GenBank, EMBL and DNA Data Bank of Japan [32]. These databases are acknowledged as the standard figure for public annotated DNA sequence collection worldwide and contain millions of plant DNA sequences. Take NCBI as an example, up to 2015, NCBI Genome database have been increased to a total of 5,132,285 plant accession entries according to RefSeq Growth Statistics (http://www.ncbi.nlm.nih.gov/refseq/statistics/). Back to 2004, there were only 88,972 entries, thus, the growth rate is approximately 458,483 entries per year over ten years, which means more than 38,000 sequences are updated monthly.

There are other public databases which may provide extra information on plant genome such as Phytozome [56], PlantGDB [57], EnsembPlants, ChloroplastDB [58], KEGG [59], Genomes On-Line Database (GOLD) [12] and the wiki of CoGepedia web page (Table **1**). Recently, in addition to these general sequence data banks, other databases that focus on specific plant species have been available. Some examples for species-specific sequence databases are The *Arabidopsis* Initiative Resource (TAIR) [60], The Salk Institute Genomics Analysis Laboratory (SIGnAL), The RIKEN *Arabidopsi*s Genome Encyclopedia (RARGE) [61], The Rice Genome Annotation Project (RGAP) [62],The Rice Annotation Project (RAP-DB) [63], The *Solanaceae* (SOL) Genomics Network (SGN) [64], *Gramene* [65], GrainGenes [66], SoyBase [67], MaizeGDB [68], CyanoBase [69], the Genome Database for *Rosaceae* (GDR) [70], *Brassica* Genome Gateway and Cucurbit Genomics Database (Table **1**) [71]. Commonly, these databases and associated web portals incorporate a set of analytical, visualization and interrogation tools to study the genomic sequences they process such as BLAST for identifying sequence similarity in large datasets.

4. PLANT COMPARATIVE GENOMIC ANALYSIS

Once whole genomes have been sequenced, defining and describing the gene and non-coding content in these sequences is an important process [72]. For that reason, plant comparative genomic analysis has arisen as a new field of modern biotechnology since its main function is to predict functions for many unknown genes by studying the significant differences and similarities among species. These genes, however, are required to appear in the available datasets of orthologs evolved from the same ancestor [73]. As can be seen, developing new tools, strategies to manage and analyze these tremendous data has been urgently needed. Recent approaches in bioinformatics and systematic biology have reached those demands but still faced further challenges.

4.1. Tools and Databases for Plant Comparative Genomic Analysis

Using comparative genomic approach, more and more genes in plant species have been annotated. For instance, several known stress-responsive transcription factors (TFs) in *Arabidopsis* and rice were used to correctly predict stress-responsive TFs in many other plant species, such as soybean, maize, sorghum, barley, and wheat [74-76]. Moreover, not only comparing within plant species, comparative genomics between plants and distantly related prokaryotes can be greatly presumed the genes functionally associated. The function of NiaP protein family in plants was determined from knowing the role of those proteins in bacteria [77]. Similar strategies to identify functional genes among different plants using comparative analysis also help researchers study genes annotation in newly sequenced plant species [78].

In addition, comparative genomics can discover missing biosynthetic genes by co-expression analysis [79]. This

Table 1. List of plant genomic databases.

Type of Database	URL
General Plant Genome Database	
NCBI Genome	http://www.ncbi.nlm.nih.gov/genome/
Phytozome v10.2	http://phytozome.jgi.doe.gov/pz/portal.html
PLAZA	http://plaza.psb.ugent.be/
PlantGDB	http://www.plantgdb.org/
Ensembl Plants	http://plants.ensembl.org/index.html
ChloroplastDB	http://chloroplast.cbio.psu.edu/
KEGG	http://www.genome.jp/kegg/
GOLD v.5	https://gold.jgi-psf.org/
CoGepedia	https://genomevolution.org/wiki/index.php/Main_Page
Species-specific sequence databases	
TAIR (*Arabidopsis*)	http://www.arabidopsis.org/
SIGnAL (*Arabidopsis*)	http://signal.salk.edu/
RARGE (*Arabidopsis*)	http://rarge.psc.riken.jp/
RGAP v.7 (Rice)	http://rice.plantbiology.msu.edu/
RAP-DB (Rice)	http://rapdb.dna.affrc.go.jp/
SGN (*Solanaceae*)	http://solgenomics.net/solanaceae-project/index.pl
Gramene (*Gramineae*)	http://www.gramene.org/
GrainGenes (*Triticeae* and *Avena*)	http://wheat.pw.usda.gov/GG3/
SoyBase (Soybean)	http://soybase.org/
MaizeGDB (Maize)	http://www.maizegdb.org/
CyanoBase (Cyanobacteria)	http://genome.microbedb.jp/cyanobase/
GDR (*Rosaceae*)	https://www.rosaceae.org/
Brassica Genome Gateway (*Brassica*)	http://brassica.nbi.ac.uk/
Cucurbit Genomics Database (*Cucurbitacea*e)	http://www.icugi.org/cgi-bin/ICuGI/index.cgi
Comparative genomics analysis databases	
Golm transcriptome db	http://csbdb.mpimp-golm.mpg.de/csbdb/dbxp/ath/ath_xpmgq.html
ATTED-II	http://atted.jp/
Other database and tools resources	
Galaxy	http://galaxyproject.org
Sanger institute	http://www.sanger.ac.uk/
GSAP	http://www.broadinstitute.org/

method performs by considering an unknown gene that is co-expressed with various genes from a metabolic pathway which is expected to have a function in that particular pathway [80, 81]. GolmTranscriptome DB [82] and ATTED-II [83] are two popular tools for such type of analysis in plants. One case for this analysis is the discovery of trans-prenyldiphosphate synthase responsible for making the solanesyl moiety of ubiquinone-9. *Arabidopsis* gene At2g34630 was identified as an alternative candidate using the co-expression and under-expression analysis in *Arabidopsis* and by functional complementation in yeast [84].

Besides tools and strategies for analysis, powerful computational resources are essential to store and manage massive genomic data. Many online platforms have been developed, published and available to perform comparative genomic study among different plant species. For instance, several plant genomic data platforms described below have been the most representative and widely used recently.

Phytozome. One of the largest comparative databases for plant species (http://phytozome.jgi.doe.gov/pz/portal.html). It contains plant genome, gene family data, and evolutionary history information. From the beginning, only 25 plant genomes were sequenced and annotated. This number has increased up to more than 50 species at the current state. Phytozome also provides impressive tools for comparative analysis in level of sequence, gene structure, gene family, and genome organization. With those tools and comprehensive web portal, Phytozome makes it accessible for scientist worldwide conducting plant research intensively [56].

PLAZA. Being known as the most comprehensible plant comparative genomics online platform, PLAZA integrates functional and structure annotation of all currently published crop plant genomes (http://plaza.psb.ugent.be/). Together with that huge set of data, PLAZA provides many interactive tools to study gene, genome evolution, and gene function. Those tools include pre-computed datasets cover, intraspecies dot plots, whole-genome multiple sequence alignments, homologous gene families, phylogenetic trees, and genomic colinearity between species [85].

GreenPhylDB. A web resource belongs to South Green Bioinformatics Platform (http://southgreen.cirad.fr/) and is open to public access. GreenPhylDB is designed for comparative and functional genomics in plants. This database contains 37 full genomes of members of the Plant kingdom at the current release version 4. Catalogue of gene families from GreenPhylDB is provided by gene predictions of genomes, covering a broad taxonomy of green plants. Its web interfaces have been continually developed to improve the navigation through information related to each gene or gene family, such as gene composition, protein domains, publications, orthologous gene predictions, and also external links. The latest version of this database is now possible to browse the full Gene Oncology, which supports gene discovery [86].

PlantsDB. This is one of the most commonly used plant database resources for integrative and comparative plant genome research (http://mips.helmholtz-muenchen.de/plant/genomes.jsp). PlantsDB comprises database instances for tomato, *Medicago*, *Arabidopsis*, *Brachypodium*, *Sorghum*, maize, rice, barley and wheat. This platform stores and provides individual plant genomes. Moreover, it is also equipped with up-to-date bioinformatics tools to visualize synteny, transfer data from model systems to crops and explore similarities and peculiarities of different plant species. Further important analysis strategies developed from PlantsDB are repeat catalogs and classification systems for all plant species [87].

4.2. Remaining Challenges

The enormous amount of genomic data for plants rapidly increases. Thousands of Gb of plant sequences are deposited in NCBI and other public databases monthly. However, reference genome sequence with basic annotation provided by current comparative genomic databases is simply a foundation. It still needs to be integrated with specific biological data such as plant epigenetic decorations and gene expression under vary conditions of environment, development stages and tissue types in order to get better detailed genome maps [34].

Moreover, since plant genomes have been constantly sequenced and re-sequenced, there is rising problem in updating databases. The update process should occur in all comparative genomic databases, not just solely in that individual genome database. This technical problem requires efforts to synchronize update data resources among different plant genomic platforms. Developing a strong community network of plant researchers might be one solution for this issue [88].

Several databases have been developed, published and available to compare plant genomes and tentatively identify orthologs (Table **1**). Having powerful application in gene prediction, comparative genomics recently has played an important role in contributing the functional annotation infrastructure on which future plant biotechnology researchers rely on.

5. PHYLOGENOMICS AND EVOLUTIONARY ANALYSIS IN PLANT

Phylogenomics is known as molecular phylogenetic analysis, in which using sets of genomic database for gene function prediction and exploration of the evolutionary relationships among species. This definition of phylogenomics was formed from the early studies in the late 1990s when a scientific hypothesis about protein function via evolutionary analysis of a gene and its homologs was published [89]. Phylogenomics was also defined as the new era of phylogenetic analysis when there are more complete genomes sequenced [90]. Plant phylogenomics has an advantage over other species, which is the ability to identify hundreds of low copy number nuclear genes, hence easily to study the molecular systematic and evolutionary biology [91]. Current approaches of NGS also provide plant phylogenomics research useful information about plant genome diversity, such as the nature and frequency of genome duplication among a diversity of plant lineages [92-94].

There are two important goals in phylogenomic research aims to accomplish. First is to discover the evolutionary patterns among plant species using nuclear genomic information. Second is to derive new hypothesis for the unknown function of plant genes associated to major divergence events in the evolution of plant species [95]. Genomic data give more advantages in the evolutionary study than morphological data which are easily misleading or fossil data which are usually fragmented. Phylogenomics also uses a set of orthologs from genomic sequence via a phylogenetic context to detect hypotheses for the genes and biological processes [96]. The main difference between functional phylogenomics compare to classical phylogenetic analysis methods and current functional genomic methods is that in phylogenomics research, genomic information is mined without incorporating a phylogenetic context during the search for

orthologs or candidate genes of functional importance [97]. However, it remains a debating issue in constructing the tree of life (phylogeny of all organisms), which inferred evolutionary relationship using phylogenomics as the advance method. Some studies continuously revalidated the positions of certain plant species in biological taxonomy [98-100] to get the most accurate tree as possible. Therefore, how to draw a scientifically significant topology is still problematic due to some limitations, such as the confliction among methodologies and character sets [101] and systematic errors from merely adding more sequences [102].

As shown above, the main problem of phylogenomics comes from how to handle the large scale of genomic data in a proper way to avoid systematic misleading (bias) assumptions. Statistical confidence (*P value*) which is normally used in such phylogenetic issue manner, however, was reported as unreliable. The authors then suggested that the magnitudes of differences (effect sizes) and biological relevance are those should be more focus on to get trustworthy results [103]. Another solution is the improvement of existing phylogenetic algorithms so that phylogenomic relationships can be inferred with minimal technical biases and greater computer efficiency [104].

New methods and tools have been developed to gradually overcome these limitations of plant phylogenomics. For instance, *de novo* assembly of short read RNA-seq data dramatically improves gene coverage by non-redundant and non-chimeric transcripts that are optimized for downstream phylogenomic analysis [105]. Another protocol is called Hyb-Seq, which combines target enrichment of low-copy nuclear exons and flanking regions, as well as genome skimming of high-copy repeats and organelle genomes, to efficiently produce genome-scale data sets for plant phylogenomics [106]. More recently, ExaML (Exascale Maximum Likelihood), which is usually known as new code for large-scale phylogenetic analyzes on Intel MIC (Many Integrated Core) hardware platform, has been updated its version 3. This coding program represents the achievement of developing better phylogenetic analysis algorithms, it is now possible to analyze datasets with 10-20 genes and up to 55,000 taxa [107]. However, even though it is just released few months ago, ExaML still has its limit since it can only run on supercomputer with Linux/Mac system. Obviously, new plant phylogenomic tools similar to ExaML is desperately needed with high quality performance and easy to operate in any computational system in the future.

6. GENOME-WIDE ASSOCIATION STUDIES IN PLANT

Basic knowledge of phenotypic variation, such as those agronomically important traits used for plant breeding resources has been the main trend of plant genetic studies. In classical crop breeding, biparental cross-mapping is still a major method for genetics dissections of the traits although its limitation is giving the QTLs mapping with low resolution (typically with several megabases in distance) [108]. To overcome that disadvantage, GWAS is currently a favorable tool to explore the allelic variation in a broader scope for extensive phenotypic diversity and higher resolution of QTL mapping thanks to the advent of NGS. Using GWAS, many research projects have been done to investigate the association between genetic variation and valuable plant traits. GWAS has been successfully applied to study *Arabidopsis thaliana*, a typical model plant organism, in which more than 1,300 distinct accessions have been genotyped for 250,000 SNPs [109] and 107 phenotypes have been studies [110]. Following this initial foundation, there were numerous achievements in conducting GWAS on other traits of interests in *Arabidopsis*, such as glucosinolate levels [111], shade avoidance [112], heavy metal [113], salt tolerance [114] and flowering time [115], etc. Beside *Arabidopsis*, rice, one of the most important crop species in the world, also has been focus of intense efforts to map the ancestral genetic variation that underlines agronomic traits such as heading date, grain size, and starch quality [116]. A few rice genes having large effects in controlling traits are involved in determining yield, morphology, stress tolerance, and nutritional quality were also identified [117]. GWAS has been widely used to dissect complex traits in some other major crops, e.g., maize and soybean [118-122].

It is undeniable that GWAS has the powerful application to plant species for identifying phenotypic diversity in trait-associated loci, as well as allelic variation in candidate genes addressing quantitative and complex traits [123, 124]. However, to accelerate genetic mapping and gene discovery in plant using GWAS, besides massive DNA variation data from NGS, it requires having a high-through put phenotyping facility that is capable to capture in details specific traits to enhance GWAS results and gain more significant gene identification information [125]. It is a challenging and promising road for future plant genomic mapping research. Hence, there are efforts on making high quality phenotyping data [126-129]. Furthermore, having computational tools to assist GWAS is also concerning issue. There are three main factors required for a GWAS tool to well perform including computing speed, memory requirements, and statistical power [130]. At the current stage, several bioinformatics approaches have been introduced as GWAS acceleration tools. Following are some examples:

Heap. Heap is a SNPs detection tool for NGS data with special reference to GWAS and genomic. Heap detects larger number of variants taking advantage of the information whether the samples are inbred (homozygosity assumption) or not. For data portability to GWAS/GP, Heap outputs variant information in vcf, beagle and PED/MAP format files that are compatible with existing GWAS/GP tools [131].

GnpIS-Asso. *GnpIS-Asso* is a generic database for managing and exploiting plant genetic association studies. This database provides tools that allow plant scientists or breeders to get associations values between traits and markers obtained in several association studies. It is also easy to view graphically the results with dedicated plots (QQPlot, Manhattan Plot), generated dynamically and to extract data in files to continue the analysis with external tools. After selecting the best markers associated to trait of interest, one specific tool automatically jumps on the genome to find where those markers are located on chromosomes and to identify which genes or other markers or features of interest are nearby. This database is already currently used for dealing GWAS for two species: tomato and maize [132].

BioGPU. As a high performance computing tool for GWAS, BioGPU effectively controls false positives caused by population structure and unequal relatedness among individuals and improves statistical power when compared to mixed linear model methods. The BioGPU method requires much less complex computing time. BioGPU was developed with parallel computational capacity to increase computing speed, so that computing time decreases linearly with the number of central processing units. To solve the memory footprint bottleneck, BioGPU allows users to directly control memory usage when big data are analyzed on computers with limited memory, which means users have the option to trade computing time for less memory usage. Based on these features, BioGPU makes analyzes of large and complex datasets feasible without supercomputers [130].

BHIT. Bayesian high-order interaction toolkit (BHIT) first builds a Bayesian model on both continuous data and discrete data, which is capable of detecting high-order interactions in SNPs related to case-control or quantitative phenotypes. Using both simulation data and soybean nutritional seed composition studies on oil content and protein content, BHIT effectively detects the high-order interactions associated with phenotypes, and it outperformed a number of other currently available tools. BHIT are also used on Soybean 50K SNP array analysis by diversity computational strategies. Then a series of SNP interactions in multiple-orders are detected associated with oil and protein phenotypes. BHIT is freely available at http://digbio.missouri.edu/BHIT/ for academic users [133].

While it was time-consuming in the past to perform QTL analysis a small data, recent bioinformatics approach helps running GWAS with a simple marker scan of few hundred thousand SNPs on PC or web-based software within few minutes [123]. However, future GWAS assisted tools still need to be improved in speed and increased memory capacity in order to integrate with rapidly growing plant genomic data. Moreover, to ensure the accuracy of GWAS results, statistical test is very important factor and must be applied intensively, in which mixed models are set as the error-making factor of genetic background [134, 135]. One example for this is a GWAS online tool is the one for *Arabidopsis*, which was developed based on R and Python programming languages [136]. This web-based server comprises of common accessions with their genotyping information and several statistical options as well as integrates correlation analysis among published traits [136].

In combination with high resolution phenotyping technologies, performing GWAS is a novel strategy for conducting research on plant genetics, genomics, gene characterization and breeding [137]. Nevertheless, GWAS analysis still has another limitation, which is failure in detecting epistatic and gene-environment interactions in most studies [138]. Due to the fact that living organisms express their phenotypes as the result of not only one but several factors including epistatic effects and their interactions with environment; hence it is important to estimate those gene-gene and gene-environment interactions for better breeding system [139, 140]. Focusing on one main SNP that correlates with a specific phenotype as normal GWAS output may miss the key genetic variants with particular environment response in the context of complex traits [141]. For this issue, bioinformatics approach is again a current solution. Generalize multifactor dimensionality reduction (GMDR) algorithm on a computing system with graphics processing units (GPUs) is one in some available methods at the moment that can screen potential candidate variants and then use the mixed liner model to detect the epistatic and gene-environment interactions [142]. This new GWAS strategy was applied and showed its success in identifying four significant SNPs associated with additive, epistatic, and gene-environment interaction effects in rice [138]. Similar GWAS method using epistatic association mapping (EAM) also successfully detected three epistatic QTLs in soybean [143]. Those presented methods are just the groundwork, future bioinformatics tools have to be more powerful in statistical methodology and overcome the heavy burden of current computation [144-146].

7. BIOINFORMATIC ADVANCES BEYOND PLANT GENOMIC RESEARCH

The world is now at the post genomic era since DNA sequencing technology continues reaching unprecedented innovations in sequencing scale and throughput. In particular, the term "genomics" by itself is only just a small part in the whole picture named "Omics". With the development of modern technology, several new omics layers have been emerged to deepen the knowledge of plant molecular system [147]. The most recent added omics layers include interactomics, epigenomics, hormonomics, and metabolomics. While NGS provides feature for whole-genome sequencing/re-sequencing for various genomic analysis, such as those are discussed across this paper, RNA sequencing (RNA-seq) is established for transcriptome and non-coding RNAome analysis, quantitative detection of epigenomic dynamics, and Chip-seq analysis for DNA–protein interactions [148]. In addition, approaches in transcriptional regulatory networks research based on omics data have been published such as interactome analysis for networks formed by protein–protein interactions [149], hormonome analysis for phytohormone-mediated cellular signaling [150], and metabolome analysis for metabolic systems [151]. Apparently, these rapidly growing omics databases widen the large-scale of genomic resources. Therefore, bioinformatics has become more essential than ever for every aspect of omic-based research to be well managed and effectively analyzed.

8. CONCLUSIONS

Recent advances in bioinformatics application for plant genomes not only provide huge potential for large-scale genomic research among plant species but also many technical challenges. NGS technologies and platforms will make plant genetic data become abundant in the next few years. With these accessible genomic data, development of effective tools for these data management and analysis become increasingly important. Indeed, there are more and more genome databases of plant species continuously established merging with different analysis methods. Comparative genomic analysis gives a specific insight of functional genes within the same and among plant species. Phylogenomic results show more accurate evidences for evolution studies and hypothesized function of genes in plant. GWAS, which has been currently used in plant research, successfully point

out loci and allelic variation related to valuable traits. On the contrary, one of the main challenges facing plant genomic researchers is the high demand of knowledge and skills in bioinformatics as well as computer sciences in order to well manage and intensively manipulate the results from the increasing of large-scale plant genomic data. Moreover, since high density genotype information rapidly exploited, high-throughput phenotyping is urgently needed to provide plant genomic analysis results at high resolution.

In brief, the recent wealth of plant genomic resources, along with advances in bioinformatics, have enabled plant researchers to achieve fundamental and systematic understanding of economically important plants and plant processes, critical for advancing crop improvement. Despite these exciting achievements, there remains a critical need for effective tools and methodologies to advance plant biotechnology, to tackle questions that are hardly solved using current approaches, and to facilitate the translation of this newly discovered knowledge to improve plant productivity.

CONFLICT OF INTEREST

The author(s) confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

"This work was supported by Vietnam National University, HCM under grant number C2014-28-07 to N.P.T. This work was also funded by the Asian Office of Aerospace Research & Development of The United Statesunder grant number FA2386-15-1-419 to L.L."

REFERENCES

[1] Millstone, E.; Lang, T. *The atlas of food: who eats what, where and why*. Earthscan Publications Ltd., **2003**.

[2] Mann, J. Natural products in cancer chemotherapy: past, present and future. *Nat. Rev. Cancer*, **2002**, *2* (2), 143-148.

[3] Lippman, Z.; Gendrel, A.-V.; Black, M.; Vaughn, M.W.; Dedhia, N.; McCombie, W.R.; Lavine, K.; Mittal, V.; May, B.; Kasschau, K.D. Role of transposable elements in heterochromatin and epigenetic control. *Nature*, **2004**, *430* (6998), 471-476.

[4] Schuster, S.C. Next-generation sequencing transforms today's biology. *Nature*, **2007**, *200* (8), 16-18.

[5] Govindaraj, M.; Vetriventhan, M.; Srinivasan, M. Importance of genetic diversity assessment in crop plants and its recent advances: an overview of its analytical perspectives. *Genet. Res. Int.*, **2015**, *2015*, 431487.

[6] Feuillet, C.; Leach, J.E.; Rogers, J.; Schnable, P.S.; Eversole, K. Crop genome sequencing: lessons and rationales. *Trends Plant Sci.*, **2011**, *16* (2), 77-88.

[7] Raes, J.; Bork, P. Molecular eco-systems biology: towards an understanding of community function. *Nat. Rev. Microbiol.*, **2008**, *6* (9), 693-699.

[8] Langeveld, S.; van Mansfeld, A.; Baas, P.; Jansz, H.; Van Arkel, G.; Weisbeek, P. Nucleotide sequence of the origin of replication in bacteriophage phiX174 RF DNA. *Nature*, **1978**, *271* (5644), 417-420.

[9] Swerdlow, H.; Wu, S.; Harke, H.; Dovichi, N.J. Capillary gel electrophoresis for DNA sequencing: laser-induced fluorescence detection with the sheath flow cuvette. *J. Chromatogr.*, **1990**, *516* (1), 61-67.

[10] Hunkapiller, T.; Kaiser, R.; Koop, B.; Hood, L. Large-scale and automated DNA sequence determination. *Science*, **1991**, *254* (5028), 59-67.

[11] Shendure, J.; Mitra, R.D.; Varma, C.; Church, G.M. Church. Advanced sequencing technologies: methods and goals. *Nat. Rev. Genet.*, **2004**, *5* (5), 335-344.

[12] Pagani, I.; Liolios, K.; Jansson, J.; Chen, I.-M.A.; Smirnova, T.; Nosrat, B.; Markowitz, V.M.; Kyrpides, N.C. Kyrpides. The Genomes On-Line Database (GOLD) v. 4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.*, **2012**, *40* (D1), D571-D579.

[13] Siegel, J. A.; Saukko, P.J. *Encyclopedia of forensic sciences*. Academic Press, **2012**.

[14] Singh, D.; Singh, P.K.; Chaudhary, S.; Mehla, K.; Kumar, S. Exome sequencing and advances in crop improvement. *Adv. Genet.*, **2012**, 79, 87-121.

[15] Fuller, C.W.; Middendorf, L.R.; Benner, S.A.; Church, G.M.; Harris, T.; Huang, X.; Jovanovich, S.B.; Nelson, J.R.; Schloss, J.A.; Schwartz, D.C. The challenges of sequencing by synthesis. *Nat. Biotechnol.*, **2009**, *27* (11), 1013-1023.

[16] Ahmadian, A.; Gharizadeh, B.; Gustafsson, A.C.; Sterky, F.; Nyrén, P.; Uhlén, M.; Lundeberg, J. Single-nucleotide polymorphism analysis by pyrosequencing. *Anal. Biochem.*, **2000**, *280* (1), 103-110.

[17] Myllykangas, S.; Buenrostro, J.; Ji, H.P. Overview of sequencing technology platforms. In *Bioinformatics for high throughput sequencing*, Springer, **2012**; pp 11-25.

[18] Dörre, K.; Brakmann, S.; Brinkmeier, M.; Han, K.T.; Riebeseel, K.; Schwille, P.; Stephan, J.; Wetzel, T.; Lapczyna, M.; Stuke, M. Techniques for single molecule sequencing. *Bioimaging*, **1997**, *5* (3), 139-152.

[19] Mardis, E.R. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, **2008**, *9*, 387-402.

[20] Branton, D.; Deamer, D.W.; Marziali, A.; Bayley, H.; Benner, S.A.; Butler, T.; Di Ventra, M.; Garaj, S.; Hibbs, A.; Huang, X. The potential and challenges of nanopore sequencing. *Nat. Biotechnol.*, **2008**, *26* (10), 1146-1153.

[21] Luo, C.; Tsementzi, D.; Kyrpides, N.; Read, T.; Konstantinidis, K.T. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PloS one*, **2012**, *7* (2), e30087.

[22] Metzker, M.L. Sequencing technologies—the next generation. *Nat. Rev. Genet.*, **2010**, *11* (1), 31-46.

[23] Quail, M.A.; Smith, M.; Coupland, P.; Otto, T.D.; Harris, S.R.; Connor, T.R.; Bertoni, A.; Swerdlow, H.P.; Gu, Y. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, **2012**, *13* (1), 341.

[24] Suzuki, S.; Ono, N.; Furusawa, C.; Ying, B.-W.; Yomo, T. Comparison of sequence reads obtained from three next-generation sequencing platforms. *PloS one*, **2011**, *6* (5), e19534.

[25] Lister, R.; O'Malley, R.C.; Tonti-Filippini, J.; Gregory, B.D.; Berry, C.C.; Millar, A.H.; Ecker, J.R.. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, **2008**, *133* (3), 523-536.

[26] Beissinger, T.M.; Hirsch, C.N.; Sekhon, R.S.; Foerster, J.M.; Johnson, J.M.; Muttoni, G.; Vaillancourt, B.; Buell, C.R.; Kaeppler, S.M.; de Leon, N. Marker density and read depth for genotyping populations using genotyping-by-sequencing. *Genetics*, **2013**, *193* (4), 1073-1081.

[27] Romay, M.C.; Millard, M.J.; Glaubitz, J.C.; Peiffer, J.A.; Swarts, K.L.; Casstevens, T.M., Elshire, R.J.; Acharya, C.B.; Mitchell, S.E.; Flint-Garcia, S.A. Comprehensive genotyping of the USA national maize inbred seed bank. *Genome Biol.*, **2013**, *14* (6), R55.

[28] Blankenberg, D.; Gordon, A.; Von Kuster, G.; Coraor, N.; Taylor, J.; Nekrutenko, A. Manipulation of FASTQ data with Galaxy. *Bioinformatics*, **2010**, *26* (14), 1783-1785.

[29] Rutherford, K.; Parkhill, J.; Crook, J.; Horsnell, T.; Rice, P.; Rajandream, M.-A.; Barrell, B. Artemis: sequence visualization and annotation. *Bioinformatics*, **2000**, *16* (10), 944-945.

[30] Adams, M.D.; Kelley, J.M.; Gocayne, J.D.; Dubnick, M.; Polymeropoulos, M.H.; Xiao, H.; Merril, C.R.; Wu, A.; Olde, B.; Moreno, R.F. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **1991**, *252* (5013), 1651-1656.

[31] Initiative, A. G. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, **2000**, *408* (6814), 796.

[32] Edwards, D.; Batley, J. Plant bioinformatics: from genome to phenome. *Trends Biotechnol.*, **2004**, *22* (5), 232-237.

[33] Metzker, M. L. Emerging technologies in DNA sequencing. *Genome Res.*, **2005**, *15* (12), 1767-1776.

[34] Michael, T.P.; VanBuren, R. Progress, challenges and the future of crop genomes. *Curr. Opin. Plant Biol.*, **2015**, *24*, 71-81.

[35] Vogel, J.P.; Garvin, D.F.; Mockler, T.C.; Schmutz, J.; Rokhsar, D.; Bevan, M.W.; Barry, K.; Lucas, S.; Harmon-Smith, M.; Lail, K. Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature,* **2010**, *463* (7282), 763-768.

[36] Rensing, S.A.; Lang, D.; Zimmer, A.D.; Terry, A.; Salamov, A.; Shapiro, H.; Nishiyama, T.; Perroud, P.-F.; E. Lindquist, A.; Kamisugi, Y. The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science*, **2008**, *319* (5859), 64-69.

[37] Bennetzen, J.L.; Schmutz, J.; Wang, H.; Percifield, R.; Hawkins, J.; Pontaroli, A.C.; Estep, M.; Feng, L.; Vaughn, J.N.; Grimwood, J. Reference genome sequence of the model plant Setaria. *Nat. Biotechnol.*, **2012**, *30* (6), 555-561.

[38] Zhang, G.; Liu, X.; Quan, Z.; Cheng, S.; Xu, X.; Pan, S.; Xie, M.; Zeng, P.; Yue, Z.; Wang, W. Genome sequence of foxtail millet (Setaria italica) provides insights into grass evolution and biofuel potential. *Nat. Biotechnol.*, **2012**, *30* (6), 549-554.

[39] Langridge, P; Reynolds, M.P. Genomic tools to assist breeding for drought tolerance. *Curr. Opin. Biotechnol.*, **2015**, *32*, 130-135.

[40] Goff, S.A.; Ricke, D.; Lan, T.-H.; Presting, G.; Wang, R.; Dunn, M.; Glazebrook, J.; Sessions, A.; Oeller, P.; Varma, H. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science*, **2002**, *296* (5565), 92-100.

[41] Yu, J.; Hu, S.; Wang, J.; Wong, G.K.-S.; Li, S.; Liu, B.; Deng, Y.; Dai, L.; Zhou, Y; Zhang, X. A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science*, **2002**, *296* (5565), 79-92.

[42] Tuskan, G.A.; Difazio, S.; Jansson, S.; Bohlmann, J.; Grigoriev, I.; Hellsten, U.; Putnam, N.; Ralph, S.; Rombauts, S.; Salamov, A. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science*, **2006**, *313* (5793), 1596-1604.

[43] Schnable, P.S.; Ware, D.; Fulton, R.S.; Stein, J.C.; Wei, F.; Pasternak, S.; Liang, C.; Zhang, J.; Fulton, L.; Graves, T.A. The B73 maize genome: complexity, diversity, and dynamics. *Science*, **2009**, *326* (5956), 1112-1115.

[44] Schmutz, J.; Cannon, S.B.; Schlueter, J.; Ma, J.; Mitros, T.; Nelson, W.; Hyten, D.L.; Song, Q.; Thelen, J.J.; Cheng, J. Genome sequence of the palaeopolyploid soybean. *Nature*, **2010**, *463* (7278), 178-183.

[45] Consortium, T.G. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, **2012**, *485* (7400), 635-641.

[46] Zimin, A.; Stevens; K.A.; Crepeau, M.W.; Holtz-Morris, A.; Koriabine, M.; Marçais, G.; Puiu, D.; Roberts, M.; Wegrzyn, J.L.; de Jong, P.J. Sequencing and assembly of the 22-Gb loblolly pine genome. *Genetics*, **2014**, *196* (3), 875-890.

[47] Leushkin, E.V.; Sutormin, R.A.; Nabieva, E.R.; Penin, A.A.; Kondrashov, A.S.; Logacheva, M.D. The miniature genome of a carnivorous plant Genlisea aurea contains a low number of genes and short non-coding sequences. *BMC Genomics*, **2013**, *14* (1), 476.

[48] Ibarra-Laclette, E.; Lyons, E.; Hernández-Guzmán, G.; Pérez-Torres, C.A.; Carretero-Paulet, L.; Chang, T.-H.; Lan, T.; Welch, A.J.; Juárez, M.J.A.; Simpson, J. Architecture and evolution of a minute plant genome. *Nature*, **2013**, *498* (7452), 94-98.

[49] Wang, W.; Haberer, G.; Gundlach, H.; Gläßer, C.; Nussbaumer, T.; Luo, M.; Lomsadze, A.; Borodovsky, M.; Kerstetter, R.; Shank-lin J. The Spirodela polyrhiza genome reveals insights into its neotenous reduction fast growth and aquatic lifestyle. *Nat. Commun.,* **2014**, *5*, ncomms 4311.

[50] Gupta, P.; Langridge, P.; Mir, R. Marker-assisted wheat breeding: present status and future possibilities. *Mol. Breed.*, **2010**, *26* (2), 145-161.

[51] Consortium, P.G.S. Genome sequence and analysis of the tuber crop potato. *Nature*, **2011**, *475* (7355), 189-195.

[52] Yu, Q.; Tong, E.; Skelton, R.L.; Bowers, J.E.; Jones, M.R.; Murray, J.E.; Hou, S.; Guan, P.; Acob, R.A.; Luo, M.-C. A physical map of the papaya genome with integrated genetic map and genome sequence. *BMC Genomics*, **2009**, *10* (1), 371.

[53] Welter, L.J.; Göktürk-Baydar, N.; Akkurt, M.; Maul, E.; Eibach, R.; Töpfer, R.; Zyprian, E.M. Genetic mapping and localization of quantitative trait loci affecting fungal disease resistance and leaf morphology in grapevine (Vitis vinifera L). *Mol. Breed.*, **2007**, *20* (4), 359-374.

[54] Soltis, D.E.; Ma, H.; Frohlich, M.W.; Soltis, P.S.; Albert, V.A.; Oppenheimer, D.G.; Altman, N.S.; Leebens-Mack, J. The floral genome: an evolutionary history of gene duplication and shifting patterns of gene expression. *Trends Plant Sci.*, **2007**, *12* (8), 358-367.

[55] Sonah, H.; Deshmukh, R.K.; Singh, V.P.; Gupta, D.K.; Singh, N.K.; Sharma, T.R. Genomic resources in horticultural crops: status, utility and challenges. *Biotechnol. Adv.*, **2011**, *29* (2), 199-209.

[56] Goodstein, D.M.; Shu, S.; Howson, R.; Neupane, R.; Hayes, R.D.; Fazo, J. ; Mitros, T.; Dirks, W.; Hellsten, U.; Putnam, N. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.*, **2012**, *40* (D1), D1178-D1186.

[57] Duvick, J.; Fu, A.; Muppirala, U.; Sabharwal, M.; Wilkerson, M.D.; Lawrence, C.J.; Lushbough, C.; Brendel, V. PlantGDB: a resource for comparative plant genomics. *Nucleic Acids Res.*, **2008**, *36* (suppl 1), D959-D965.

[58] Cui, L.; Veeraraghavan, N.; Richter, A.; Wall, K.; Jansen, R.K.; Leebens-Mack, J.; Makalowska, I. ChloroplastDB: the chloroplast genome database. *Nucleic Acids Res.*, **2006**, *34* (suppl 1), D692-D696.

[59] Tokimatsu, T.; Kotera, M.; Goto, S.; Kanehisa, M. KEGG and GenomeNet resources for predicting protein function from omics data including KEGG PLANT resource. In *Protein Function Prediction for Omics Era*, Springer, **2011**; pp 271-288.

[60] Swarbreck, D.; Wilks, C.; Lamesch, P.; Berardini, T.Z.; Garcia-Hernandez, M.; Foerster, H.; Li, D.; Meyer, T.; Muller, R.; Ploetz, L. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.*, **2008**, *36* (suppl 1), D1009- D1014.

[61] Sakurai, T.; Satou, M.; Akiyama, K.; Iida, K.; Seki, M.; Kuromori, T.; Ito, T.; Konagaya, A.; Toyoda, T.; Shinozaki, K. RARGE: a large-scale database of RIKEN Arabidopsis resources ranging from transcriptome to phenome. *Nucleic Acids Res.*, **2005**, *33* (suppl 1), D647-D650.

[62] Kawahara, Y.; de la Bastide, M.; Hamilton, J.P.; Kanamori, H.; McCombie, W.R.; Ouyang, S.; Schwartz, D.C., Tanaka, T.; Wu, J.; Zhou, S. Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. *Rice*, **2013**, *6* (1), 4.

[63] Sakai, H.; Lee, S. S.; Tanaka, T.; Numa, H.; Kim, J.; Kawahara, Y.; Wakimoto, H.; Yang, C.-c.; Iwamoto, M.; Abe, T. Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. *Plant Cell Physiol.*, **2013**, *54* (2), e6.

[64] Mueller, L.A.; Solow, T.H.; Taylor, N.; Skwarecki, B.; Buels, R.; Binns, J.; Lin, C.; Wright, M.H.; Ahrens, R.; Wang, Y. The SOL Genomics Network. A comparative resource for Solanaceae biology and beyond. *Plant Physiol.*, **2005**, *138* (3), 1310-1317.

[65] Ware, D.; Jaiswal, P.; Ni, J.; Pan, X.; Chang, K.; Clark, K.; Teytelman, L.; Schmidt, S.; Zhao, W.; Cartinhour, S. Gramene: a resource for comparative grass genomics. *Nucleic Acids Res.*, **2002**, *30* (1), 103-105.

[66] Matthews, D.E.; Carollo, V.L.; Lazo, G.R.; Anderson, O.D. GrainGenes, the genome database for small-grain crops. *Nucleic Acids Res.*, **2003**, *31* (1), 183-186.

[67] Grant, D.; Nelson, R.T.; Cannon, S.B.; Shoemaker; R.C. SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res.*, **2009**, doi: 10.1093/nar/gkp798

[68] Lawrence, C.J.; Dong, Q.; Polacco, M.L.; Seigfried, T.E.; Brendel, V. MaizeGDB, the community database for maize genetics and genomics. *Nucleic Acids Res.*, **2004**, *32* (suppl 1), D393-D397.

[69] Nakamura, Y.; Kaneko, T.; Hirosawa, M.; Miyajima, N.; Tabata, S. CyanoBase, a www database containing the complete nucleotide sequence of the genome of Synechocystis sp. strain PCC6803. *Nucleic Acids Res.*, **1998**, *26* (1), 63-67.

[70] Jung, S.; Staton, M.; Lee, T.; Blenda, A.; Svancara, R.; Abbott, A.; Main, D. GDR (Genome Database for Rosaceae): integrated web-database for Rosaceae genomics and genetics data. *Nucleic Acids Res.*, **2008**, *36* (suppl 1), D1034-D1040.

[71] Mochida, K.; Shinozaki, K. Genomics and bioinformatics resources for crop improvement. *Plant Cell Physiol.*, **2010**, *51* (4), 497-523.

[72] Collins, F.S.; Green, E.D.; Guttmacher, A.E.; Guyer, M.S. A vision for the future of genomics research. *Nature*, **2003**, *422* (6934), 835-847.

[73] Devos, K.M.; Gale. M.D. Genome relationships: the grass model in current research. *Plant Cell*, **2000**, *12* (5), 637-646.

[74] Mochida, K.; Yoshida, T.; Sakurai, T.; Yamaguchi-Shinozaki, K.; Shinozaki, K.; Tran, L.-S.P. *In silico* analysis of transcription factor repertoire and prediction of stress responsive transcription factors in soybean. *DNA Res.*, **2009**, doi: 10.1093/dnares/dsp023

[75] Tran, L.-S.P.; Mochida, K. Identification and prediction of abiotic stress responsive transcription factors involved in abiotic stress signaling in soybean. *Plant Signal. Behav.*, **2010**, *5* (3), 255-257.

[76] Mochida, K.; Yoshida, T.; Sakurai, T.; Yamaguchi-Shinozaki, K.; Shinozaki, K.; Tran, L.-S.P. *In silico* analysis of transcription factor repertoires and prediction of stress-responsive transcription factors from six major gramineae plants. *DNA Res.*, **2011**, doi: 10.1093/dnares/dsr019.

[77] Jeanguenin, L.; Lara-Núñez, A.; Rodionov, D.A.; Osterman, A.L.; Komarova, N.Y., Rentsch, D.; Gregory III, J.F.; Hanson, A.D.Comparative genomics and functional analysis of the NiaP family uncover nicotinate transporters from bacteria, plants, and mammals. *Funct. Integr. Genomics*, **2012**, *12* (1), 25-34.

[78] Bradbury, L.M.; Niehaus, T.D.; Hanson, A.D. Comparative genomics approaches to understanding and manipulating plant metabolism. *Curr. Opin. Biotechnol.*, **2013**, *24* (2), 278-284.

[79] Usadel, B.; Obayashi, T.; Mutwil, M.; Giorgi, F.M.; Bassel, G.W.; Tanimoto, M.; Chow, A.; Steinhauser, D.; Persson, S.; Provart, N.J. Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. *Plant, Cell Environ.*, **2009**, *32* (12), 1633-1651.

[80] Aoki, K.; Ogata, Y.; Shibata, D. Shibata. Approaches for extracting practical information from gene co-expression networks in plant biology. *Plant Cell Physiol.*, **2007**, *48* (3), 381-390.

[81] Ehlting, J.; Provart, N.; Werck-Reichhart, D. Functional annotation of the Arabidopsis P450 superfamily based on large-scale co-expression analysis. *Biochem. Soc. Trans.*, **2006**, *34* (6), 1192-1198.

[82] Kopka, J.; Schauer, N.; Krueger, S.; Birkemeyer, C.; Usadel, B.; Bergmüller, E.; Dörmann, P.; Weckwerth, W.; Gibon, Y.; Stitt, M. GMD@ CSB. DB: the Golm metabolome database. *Bioinformatics*, **2005**, *21* (8), 1635-1638.

[83] Obayashi, T.; Kinoshita, K.; Nakai, K.; Shibaoka, M.; Hayashi, S.; Saeki, M.; Shibata, D.; Saito, K.; Ohta, H. ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in Arabidopsis. *Nucleic Acids Res.*, **2007**, *35* (suppl 1), D863-D869.

[84] Ducluzeau, A.L.; Wamboldt, Y.; Elowsky, C.G.; Mackenzie, S.A.; Schuurink, R.C.; Basset, G.J. Gene network reconstruction identifies the authentic trans-prenyl diphosphate synthase that makes the solanesyl moiety of ubiquinone-9 in Arabidopsis. *Plant J.*, **2012**, *69* (2), 366-375.

[85] Van Bel, M.; Proost, S.; Wischnitzki, E.; Movahedi, S.; Scheerlinck, C.; Van de Peer, Y.; Vandepoele, K. Dissecting plant genomes with the PLAZA comparative genomics platform. *Plant Physiol.*, **2011**, pp. 111.189514.

[86] Rouard, M.; Guignon, V.; Aluome, C.; Laporte, M.-A.; Droc, G.; Walde, C.; Zmasek, C.M.; Périn, C.; Conte, M.G. GreenPhylDB v2.0: comparative and functional genomics in plants. *Nucleic Acids Res.*, **2010**, doi: 10.1093/nar/gkq811

[87] Nussbaumer, T.; Martis, M.M.; Roessner, S.K.; Pfeifer, M.; Bader, K.C.; Sharma, S.; Gundlach, H.; Spannagl, M. MIPS PlantsDB: a database framework for comparative plant genome research. *Nucleic Acids Res.*, **2013**, *41* (D1), D1144-D1151.

[88] Dhanapal, A.P.; Govindaraj, M. Unlimited Thirst for Genome Sequencing, Data Interpretation, and Database Usage in Genomic Era: The Road towards Fast-Track Crop Plant Improvement. *Genet. Res. Int.*, **2015**, *2015*, 1-15.

[89] Eisen, J. A. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.*, **1998**, *8* (3), 163-167.

[90] Delsuc, F.; Brinkmann, H.; Philippe, H. Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.*, **2005**, *6* (5), 361-375.

[91] Zhang, N.; Zeng, L.; Shan, H.; Ma, H. Highly conserved low-copy nuclear genes as effective markers for phylogenetic analyses in angiosperms. *New Phytol.*, **2012**, *195* (4), 923-937.

[92] Grandbastien, M.-A.; Deloger, M.; Nichols, R.; Macas, J.; Novák, P.; Chase, M.W. Next generation sequencing reveals genome down-sizing in allotetraploid Nicotiana tabacum, predominantly through the elimination of paternally derived repetitive DNAs. *Mol. Biol. Evol.*, **2011**, doi: 10.1093/molbev/msr112

[93] Griffin, P.C.; Robin, C.; Hoffmann, A.A. A next-generation sequencing method for overcoming the multiple gene copy problem in polyploid phylogenetics, applied to Poa grasses. *BMC Biol.*, **2011**, *9* (1), 19.

[94] McKain, M.R.; Wickett, N.; Zhang, Y.; Ayyampalayam, S.; McCombie, W.R.; Chase, M.W.; Pires, J.C.; Leebens-Mack, J. Phylogenomic analysis of transcriptome data elucidates co-occurrence of a paleopolyploid event and the origin of bimodal karyotypes in Agavoideae (Asparagaceae). *Am. J. Bot.*, **2012**, *99* (2), 397-406.

[95] Cibrián-Jaramillo, A.; Jose, E.; Lee, E.K.; Katari, M.S.; Little, D.P.; Stevenson, D.W.; Martienssen, R.; Coruzzi, G.M.; DeSalle, R. Using phylogenomic patterns and gene ontology to identify proteins of importance in plant evolution. *Genome Biol. Evol.*, **2010**, *2*, 225-239.

[96] Conte, M.G.; Gaillard, S.; Droc, G.; Perin, C. Phylogenomics of plant genomes: a methodology for genome-wide searches for orthologs in plants. *BMC Genomics*, **2008**, *9* (1), 183.

[97] Lee, E.K.; Cibrian-Jaramillo, A.; Kolokotronis, S.-O.; Katari, M.S., Stamatakis, A.; Ott, M.; Chiu, J.C.; Little, D.P.; Stevenson, D.W.; McCombie, W.R. A functional phylogenomic view of the seed plants. *PLoS Genet.*, **2011**, *7* (12), e1002411.

[98] Wu, C.-S.; Chaw, S.-M.; Huang. Y.-Y. Chloroplast phylogenomics indicates that Ginkgo biloba is sister to cycads. *Genome Biol. Evol.*, **2013**, *5* (1), 243-254.

[99] Borhidi, A. Revalidation of the genus Tournefortiopsis Rusby, (Guettardeae, Rubiaceae) and a new Guettarda from Costa Rica. *Acta Bot. Hung.*, **2008**, *50* (1-2), 61-72.

[100] Medeiros, D.; de Senna Valle, L.; Valka Alves, R. J. Revalidation of the genera Bia and Zuckertia (Euphorbiaceae) with B. capivarensis sp. nov. from Serra da Capivara, Brazil. *Nord. J. Bot.*, **2013**, *31* (5), 595-602.

[101] Jeffroy, O.; Brinkmann, H.; Delsuc, F.; Philippe, H. Phylogenomics: the beginning of incongruence? *Trends Genet.*, **2006**, 22 (4), 225-231.

[102] Philippe, H.; Brinkmann, H.; Lavrov, D.V.; Timothy J; Littlewood, D.; Manuel, M.; Wörheide, G.; Baurain, D. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol.*, **2011**, *9* (3), 402.

[103] Kumar, S.; Filipski, A. J.; Battistuzzi, F.U.; Pond, S.L.K.; Tamura, K. Statistics and truth in phylogenomics. *Mol. Biol. Evol.*, **2011**, doi: 10.1093/molbev/msr202

[104] Chan, C.X.; Ragan, M.A. Next-generation phylogenomics. *Biol. Direct*, **2013**, *8* (3).

[105] Yang, Y.; Smith, S.A. Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. *BMC Genomics*, **2013**, *14* (1), 328.

[106] Weitemier, K.; Straub, S.C.; Cronn, R.C.; Fishbein, M.; Schmickl, R.; McDonnell, A.; Liston, A. Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Appl. Plant Sci.*, **2014**, *2* (9).

[107] Kozlov, A.M.; Aberer, A.J.; Stamatakis, A. ExaML version 3: a tool for phylogenomic analyses on supercomputers. *Bioinformatics*, **2015**, doi: 10.1093/bioinformatics/btv184.

[108] Myles, S.; Peiffer, J.; Brown, P.J.; Ersoz, E.S.; Zhang, Z.; Costich, D.E.; Buckler, E.S. Association mapping: critical considerations shift from genotyping to experimental design. *Plant Cell*, **2009**, *21* (8), 2194-2202.

[109] Horton, M.W.; Hancock, A.M.; Huang, Y.S.; Toomajian, C.; Atwell, S.; Auton, A.; Muliyati, N.W.; Platt, A.; Sperone, F.G.; Vilhjálmsson, B.J. Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. *Nat. Genet.*, **2012**, *44* (2), 212-216.

[110] Atwell, S.; Huang, Y.S.; Vilhjálmsson, B.J., Willems, G.; Horton, M.; Li, Y.; Meng, D.; Platt, A.; Tarone, A.M.; Hu, T.T. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, **2010**, *465* (7298), 627-631.

[111] Chan, E. K.; Rowe H.C.; Corwin J.A.; Joseph B.; Kliebenstein D.J. Combining genome-wide association mapping and transcriptional networks to identify novel genes controlling glucosinolates in Arabidopsis thaliana. *PLoS Biol.*, **2011**, *9* (8), 1713.

[112] Filiault, D.L.; Maloof J.N. A genome-wide association study identifies variants underlying the Arabidopsis thaliana shade avoidance response. *PLoS Genet*., **2012**, *8* (3), e1002589.

[113] Chao, D.-Y.; Silva A.; Baxter I.; Huang Y.S.; Nordborg M.; Danku J.; Lahner B.; Yakubova E.; Salt, D.E. Genome-wide association studies identify heavy metal ATPase3 as the primary determinant of natural variation in leaf cadmium in Arabidopsis thaliana. *PLoS Genet.,* **2012**, *8* (9), e1002923.

[114] Baxter, I.; Brazelton J.N.; Yu D.; Huang Y.S.; Lahner B.; Yakubova E.; Li Y.; Bergelson J.; Borevitz J.O.; Nordborg M. A coastal cline in sodium accumulation in Arabidopsis thaliana is driven by natural variation of the sodium transporter AtHKT1. *PLoS Genet.,* **2010**, *6* (11), e1001193.

[115] Li, Y.; Huang Y.; Bergelson J.; Nordborg M.; Borevitz J.O. Association mapping of local climate-sensitive quantitative trait loci in Arabidopsis thaliana. *Proc. Natl. Acad. Sci.*, **2010**, *107* (49), 21199-

21204.

[116] Huang, X.; Wei X.; Sang T.; Zhao Q.; Feng Q.; Zhao Y.; Li C.; Zhu C.; Lu T.; Zhang Z. Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.*, **2010**, *42* (11), 961-967.

[117] Famoso, A. N.; Zhao K.; Clark R.T.; Tung C.-W. ; Wright M.H.; Bustamante C.; Kochian L.V.; McCouchS. R. Genetic architecture of aluminum tolerance in rice (Oryza sativa) determined through genome-wide association analysis and QTL mapping. *PLoS Genet.*, **2011**, *7* (8), e1002221.

[118] Tian, F.; Bradbury P.J.; Brown P.J.; Hung H.; Sun Q.; Flint- Garcia S.; Rocheford T.R.; McMullen M.D.; Holland J.B.; Buckler E.S. Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat. Genet.*, **2011**, *43* (2), 159-162.

[119] Poland, J.A., Bradbury P.J.; Buckler E.S.; Nelson R.J. Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize. *Proc. Natl. Acad. Sci.*, **2011**, *108* (17), 6893-6898.

[120] Li, H.; Peng Z.; Yang, X.; Wang, W.; Fu, J.; Wang, J.; Han, Y.; Chai, Y.; Guo, T.; Yang, N. Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. *Nat. Genet.*, **2013**, *45* (1), 43-50.

[121] Hwang, E.-Y.; Song, Q.; Jia, G.; Specht, J. E.; Hyten, D. L.; Costa, J.; Cregan, P.B. A genome-wide association study of seed protein and oil content in soybean. *BMC Genomics*, **2014**, *15* (1), 1.

[122] Dhanapal, A.P.; Ray, J.D.; Singh, S.K.; Hoyos-Villegas, V.; Smith, J.R.; Purcell, L.C.; King, C.A.; Cregan, P.B.; Song, Q.; Fritschi, F. B. Genome-wide association study (GWAS) of carbon isotope ratio (613C) in diverse soybean [Glycine max (L.) Merr.] genotypes. *Theor. Appl. Genet.*, **2015**, *128* (1), 73-91.

[123] Korte, A.; Farlow, A. The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods*, **2013**, *9* (1), 29.

[124] Kumar, S.; Garrick D.J.; Bink M.C.; Whitworth, C.; Chagné, D.; Volz, R.K. Novel genomic approaches unravel genetic architecture of complex traits in apple. *BMC Genomics*, **2013**, 14 (1), 393.

[125] Furbank, R.T.; Tester, M. Phenomics–technologies to relieve the phenotyping bottleneck. *Trends Plant Sci.*, **2011**, *16* (12), 635-644.

[126] Fahlgren, N.; Gehan, M.A.; Baxter, I. Lights, camera, action: high-throughput plant phenotyping is ready for a close-up. *Curr. Opin. Plant Biol.*, **2015**, *24*, 93-99.

[127] Klukas, C.; Chen D.; Pape, J.-M. Integrated analysis platform: an open-source information system for high-throughput plant phenotyping. *Plant Physiol.*, **2014**, *165* (2), 506-518.

[128] Junker, A.; Muraya M.M.; Weigelt-Fischer, K.; Arana-Ceballos, F.; Klukas, C.; Melchinger, A.E.; Meyer, R.C.; Riewe D.; Altmann, T. Optimizing experimental procedures for quantitative evaluation of crop plant performance in high throughput phenotyping systems. *Front. Plant Sci.*, **2014**, *5*, 770.

[129] Yang, W.; Guo, Z.; Huang, C.; Duan, L.; Chen, G.; Jiang, N.; Fang, W.; Feng, H.; Xie, W.; Lian; X. Combining high-throughput phenotyping and genome-wide association studies to reveal natural genetic variation in rice. *Nat. Commun.*, **2014**, *5*, ncomms6087.

[130] Huang, M. In *Biogpu: A High Performance Computing Tool for Genome-Wide Association Studies*, Plant and Animal Genome XXIII Conference, Plant and Animal Genome, **2015**.

[131] Kobayashi, M. In *Heap: A SNPs Detection Tool for NGS Data with Special Reference to GWAS and Genomic Prediction*, Plant and Animal Genome XXIII Conference, Plant and Animal Genome, **2015**.

[132] Steinbach, D. In GnpIS-Asso: A Generic Database for Managing and Exploiting Plant Genetic Association Studies Results Using High Throughput Genotyping and Phenotyping Data, Plant and Animal Genome XXIII Conference, Plant and Animal Genome, **2015**.

[133] Wang, J. In A Bayesian Model for Detection of High-Order Interactions Among Genetic Variants in Genome-Wide Association Studies and Its Application on Soybean Oil/Protein Traits, Plant and Animal Genome XXIII Conference, Plant and Animal Genome, **2015**.

[134] Cantor, R. M.; Lange, K.; Sinsheimer J.S. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *Am. J. Hum. Genet.*, **2010**, *86* (1), 6-22.

[135] Yu, J.; Pressoir, G.; Briggs, W.H.; Bi, I.V.; Yamasaki, M.; Doebley, J.F.; McMullen, M.D.; Gaut, B.S.; Nielsen, D.M.; Holland, J.B. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.*, **2006**, *38* (2), 203-208.

[136] Seren, Ü.; Vilhjálmsson, B.J.; Horton, M.W.; Meng, D.; Forai, P.; Huang, Y.S.; Long, Q.; Segura, V.; Nordborg, M. GWAPP: a web application for genome-wide association mapping in Arabidopsis. *Plant Cell*, **2012**, *24* (12), 4793-4805.

[137] Huang, X.; Han, B. Natural variations and genome-wide association studies in crop plants. *Annu. Rev. Plant Biol.*, **2014**, *65*, 531-551.

[138] Xu, H.; Jiang, B.; Cao, Y.; Zhang, Y.; Zhan, X.; Shen, X.; Cheng, S.; Lou, X., Cao, L. Detection of Epistatic and Gene-Environment Interactions Underlying Three Quality Traits in Rice Using High-Throughput Genome-Wide Data. *BioMed Res. Int.*, **2015**, *2015*, 135782.

[139] Cowling, W.; Balazs, E. Prospects and challenges for genome-wide association and genomic selection in oilseed Brassica species This article is one of a selection of papers from the conference "Exploiting Genome-wide Association in Oilseed Brassicas: a model for genetic improvement of major OECD crops for sustainable farming". *Genome*, **2010**, *53* (11), 1024-1028.

[140] Dobson, R.; Ramagopalan, S.V.; Giovannoni, G. Genome-wide association studies: will we ever predict susceptibility to multiple sclerosis through genetics? *Expert Rev. Neurother.*, **2013**, *13* (3), 235-237.

[141] Murcray, C.E.; Lewinger, J.P.; Gauderman W.J. Gene-environment interaction in genome-wide association studies. *Am. J. Epidemiol.*, **2009**, *169* (2), 219-226.

[142] Zhu, Z.; Tong, X.; Zhu, Z.; Liang, M.; Cui, W.; Su, K.; Li, M.D.; Zhu, J. Development of GMDR-GPU for gene–gene interaction analysis and its application to WTCCC GWAS data for type 2 diabetes. *PloS one*, **2013**, *8* (4), e61943.

[143] Lü, H.-Y.; Liu, X.-F.; Wei, S.-P.; Zhang; Y.-M. Epistatic association mapping in homozygous crop cultivars. *PloS one*, **2011**, *6* (3), e17773.

[144] Carlborg, Ö.; Haley, C.S. Epistasis: too often neglected in complex trait studies? *Nat. Rev. Genet.*, **2004**, *5* (8), 618-625.

[145] Phillips, P. C. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.*, **2008**, *9* (11), 855-867.

[146] van Os, J.; Rutten, B.P. Gene-environment-wide interaction studies in psychiatry. *Am. J. Psychiatry*, **2009**, *166* (9), 964-966.

[147] Mochida, K.; Shinozaki, K. Advances in omics and bioinformatics tools for systems analyses of plant functions. *Plant Cell Physiol.*, **2011**, *52* (12), 2017-2038.

[148] Lister, R.; Gregory, B.D.; Ecker, J.R. Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. *Curr. Opin. Plant Biol.,* **2009**, 12 (2), 107-118.

[149] Dreze, M.; Carvunis, A.-R.; Charloteaux, B.; Galli, M.; Pevzner,S.J.; Tasan, M.; Ahn, Y.-Y.; Balumuri, P.; Barabási, A.-L.; Bautista, V. Evidence for network evolution in an Arabidopsis interactome map. *Science*, **2011**, *333* (6042), 601-607.

[150] Kojima, M.; Kamada-Nobusada, T.; Komatsu, H.; Takei, K.; Kuroha, T.; Mizutani, M.; Ashikari, M.; Ueguchi-Tanaka, M.; Matsuoka, M.; Suzuki, K. Highly sensitive and high-throughput analysis of plant hormones using MS-probe modification and liquid chromatography–tandem mass spectrometry: an application for hormone profiling in Oryza sativa. *Plant Cell Physiol.*, **2009**, *50* (7), 1201-1214.

[151] Saito, K.; Matsuda, F. Metabolomics for functional genomics, systems biology, and biotechnology. *Annu. Rev. Plant Biol*., **2010**, *61*, 463-489.