UNIVERSITEIT
GENT

FACULTY OF SCIENCES

Comparative Analysis of Plant Genomes through Data Integration

Michiel Van Bel

Promoter: Prof. Dr. Yves Van de Peer
Co-Promoter: Prof. Dr. Klaas Vandepoele

Ghent University
Faculty of Sciences
Department of Plant Biotechnology and Bioinformatics
VIB Department of Plant Systems Biology
Bioinformatics and Systems Biology

Dissertation submitted in fulfillment of the requirements for the degree of Doctor (PhD) in Sciences, Bioinformatics).

Academic year: 2012-2013

Examination Committee

Prof. Dr. Geert De Jaeger (chair)
Faculty of Sciences, Department of Plant Biotechnology and Bioinformatics, Ghent University
Prof. Dr. Yves Van de Peer (promoter)
Faculty of Sciences, Department of Plant Biotechnology and Bioinformatics, Ghent University
Prof. Dr. Klaas Vandepoele (co-promoter)
Faculty of Sciences, Department of Plant Biotechnology and Bioinformatics, Ghent University
Prof. Dr. Jan Fostier
Faculty of Engineering, Department of Information Technology, Ghent University
Prof. Dr. Peter Dawyndt
Faculty of Science, Department of Applied Mathematics and Computer Science, Ghent University
Dr. Steven Robbens
Bayer Cropscience, Belgium
Dr. Matthieu Conte
Syngenta Seeds, France

Acknowledgements

While the cover of this book carries my name, this thesis did not come to fruition by my hand only. These past years have been a great experience, for which I would like to express my gratitude to several people.

First of all, I would like to thank Thomas Abeel, for getting me in touch with Yves' research group, and encouraging me to start a PhD in bioinformatics. Without a chance encounter with him, I never would have dreamed obtaining a PhD would be possible.

Secondly, I would like to thank my promoter and co-promoter, Yves Van de Peer and Klaas Vandepoele. The opportunity Yves has given to me to pursue a PhD and the great research environment of Yves' lab have proven to be invaluable. The constant support and patience of Klaas in guiding me form the fundaments of this PhD. Our numerous discussions on how to proceed with our shared research were definitely instrumental in my growth as a researcher.

Thirdly, I would like to express my gratitude to the members of my PhD jury, for reading my thesis and evaluating my work.

Next up in line to be thanked is Sebastian Proost: a great colleague and flat-mate. Most of my research was done in collaboration with him, and the results need to be seen as such.

A big thank you as well for my fellow IT knowledgeable colleagues at the lab: Sofie Van Landeghem, Marijn Vandevoorde, Thomas Van Parys, Frederik Delaere and Kenny Billeau. Their beacon of computer related jokes, general fun and laughter, and overall support in the darkness of the biology department were definitely important in keeping me happy and working.

I also want to thank Yvan Saeys, for guiding me through the rough first year of my PhD. Though our shared research didn't pan out, it was a good learning experience on what to do and not to do.

All Binari people, present and past, need to be thanked, as well as all the people within the BSB and biocomp group. The overall fun and interesting discussions we had will be remembered. Honorary mention goes to Lieven Sterck, whose constant presence provides a great atmosphere of collegiality within the lab.

Another group to be thanked consists of the people from the IT staff. Without their unwavering dedication in keeping our servers running and our hard drives spinning, together with their tacit approval of my development skills on the web server, this PhD would have been a lot more difficult.

Outside of the PSB building I would give a big thank you to all my friends, especially my former school mates. The efforts most of you put into obtaining a PhD really gave me that extra boost in confidence to continue my own research.

And last but definitely not least I would like to thank my family: my brother, sisters and mother. Their constant support, interest and love gave me the strength these past 6 years to carry on.

Table of Contents

List of Figures

List of Tables

"And here we go ..."

The Joker – The Dark Knight

1
Research Purpose and Scope

1.1 Overview

When we started our research in 2008, several online resources for genomics existed, each with a different focus. **TAIR** (The Arabidopsis Information Resource) has a focus on the plant model species *Arabidopsis thaliana*, with (at that time) little or no support for evolutionary or comparative genomics. **Ensemble** provided some basic tools and functions as a data warehouse, but it would only start incorporating plant genomes in 2010. There was no online resource at that time however, that provided the necessary data content and tools for plant comparative and evolutionary genomics that we required.

As such, the plant community was missing an essential component to get their research at the same level as the biomedicine oriented research communities. We started to work on PLAZA in order to provide such a data resource that could be accessed by the plant community, and which also contained the necessary data content to help our research group's focus on evolutionary genomics.

1.2 Creation of a Platform for Comparative and Evolutionary Genomics

The platform for comparative and evolutionary genomics, which we named PLAZA, was developed from scratch (i.e. not based on an existing database scheme, such as Ensemble). Gathering the data for all species, parsing this data into a common format and then uploading it into the database was the next step. We developed a processing pipeline, based on sequence similarity measurements, to group genes into gene families and sub families. Functional annotation was gathered through both the original data providers and through InterPro scans, combined with Interpro2GO. This primary data information was then ready to be used in every subsequent analysis.

Building such a database was good enough for research within our bioinformatics group, but the target goal was to provide a comprehensive resource for all plant biologists with an interest in comparative and evolutionary genomics. Designing and creating a user-friendly, visually appealing web interface, connected to our database, was the next step. While the most detailed information is commonly presented in data tables, aesthetically pleasing graphics, images and charts are often used to visualize trends, general statistics and also used in specific tools. Design and development of these tools and visualizations is thus one of the core elements within my PhD.

The PLAZA platform was designed as a gene-centric data resource, which is easily navigated when a biologist wants to study a relative small number of genes. However, using the default PLAZA website to retrieve information for dozens of genes quickly becomes very tedious. Therefore a 'gene set'-centric extra layer was developed where user-defined gene sets could be quickly analyzed. This extra layer, called the PLAZA workbench, functions on top of the normal PLAZA website, implicating that only gene sets from species present within the PLAZA database can be directly analyzed.

1.3 Creation of a Platform for Transcriptome Analysis

The PLAZA resource for comparative and evolutionary genomics was a major success, but it still had several issues. We tried to solve at least two of these problems at the same time by creating a new platform. The first issue was the building procedure of PLAZA: adding a single species, or updating the structural annotation of an existing one, requires the total re-computation of the database content. The second issue was the restrictiveness of the PLAZA workbench: through a mapping procedure gene sets could be entered for species not present in the PLAZA database, but for species without a phylogenetic close relative this approach did not always yield satisfying results. Furthermore, the research in question might just focus on the difference between a species present in PLAZA and a close relative not present in PLAZA (e.g. to study adaptation to a different ecological niche). In such a case, the mapping procedure is in itself useless. With the advent of NGS transcriptome data sets for a growing number of species, it was clear that a next challenge had presented itself.

We designed and developed a new platform, named TRAPID, which could automatically process entire transcriptome data sets, using a reference database. The target goal was to have the processing done quickly with the results containing both gene family oriented data (such as multiple sequence alignments and phylogenetic trees) and functional characterization of the transcripts. Major efforts went into designing the processing pipeline so it could be reliable, fast and accurate.

“What is it that makes natural selection succeed as a solution to the problem of improbability, whereas chance and design both fail at the starting gate? The answer is that natural selection is a cumulative process, which breaks the problem of improbability up into small pieces. Each of the small pieces is slightly improbable, but not prohibitively so.”

Richard Dawkins

2 Introduction

2.1 Abstract: A history of genetics

With the discovery of the structure of DNA in 1953 by Watson and Crick[1], the underlying mechanisms for genetic inheritance became tangible for biologists. Standing on the shoulders of giants in the fields of genetics and evolution, such as Mendel[2], Morgan[3] and Darwin[4], Watson and Crick took the first steps in a new and exciting field in natural sciences with their discovery. The logical next step was the determination of the actual content of a nucleotide sequence, a step which took the better part of two decades and was achieved first by sequencing RNA from viral strains[5], followed shortly by DNA sequencing[6]. While the initial procedures for DNA sequencing were error-prone, labor-intensive and quite costly, recent technological improvements[7,8] have led to an extreme drop in price, allowing the field of genetics to move forward at an unprecedented pace. Indeed, the cost of the Human Genome Project[9], which took more than a decade to be finally completed in 2000, is estimated at 3 billion dollar. A decade later, the sequencing cost for the same feat has dropped to 10,000 dollar[10], with a projected goal of 1000 dollar[11].

The completion of the Human Genome project was met with great expectations, but it did not, however, lead to an immediate and total insight into the inner workings of human genetics[12]. Meanwhile, in the field of plant genomics, the model species *Arabidopsis thaliana* was the first plant to have its genome fully sequenced in 2000[13] . With Arabidopsis being the model-species for plant research, this was a solid choice, especially considering its rather small genome size compared to human (130Mb vs 3Gb). However, the multitude of whole genome duplications which occured during the evolutionary history of Arabidopsis[13,14], highlighted some new difficulties during the sequencing and assembly phase[15]. The particular structure and properties of plant genomes would present more problems in the years to come[16].

In this chapter we will introduce key concepts from comparative, functional and evolutionary genomics, as well as a general overview of bioinformatics development and platforms to study genomic features. As such, the necessary background information will be provided to understand the gist of the following chapters in this PhD thesis.

For the author contributions, see page 2-14.

2.2 Comparative and Evolutionary Genomics in Plants

Comparison of living organisms has been performed since ancient times[17,18], yet the focus was always on phenotypic features such as the form and size of leafs and flowers, in the case of plants. These classifications were only natural as it allowed humans to characterize the world around them and help them survive, as differentiating between various shapes and colors helped them to distinguish edible from poisonous berries. The focus shifted from phenotype to genotype long before the arrival of DNA sequencing[19], but the shift intensified as more insights into the evolutionary history of plants could be derived this way. When the entire genome of multiple plant species became available, Arabidopsis in 2000[13], rice in 2002[20] and poplar in 2006[21], the resulting interest in comparative and evolutionary genomics grew accordingly. The studies of comparative and evolutionary genomics show considerable overlap: comparison of genomes can best be performed while considering how the genomes evolved and adapted through time, and when species diverged.

Comparison of plant genomes is not a strictly academic exercise, with a variety of real-world applications making extensive use of the knowledge gained through comparative genomics: transferring information of genetic pathways from one organism to another[22,23], understanding the adaptation of organisms to their environments[24], delineating clade specific genes[24,25], etc. The first of these applications has a direct economic impact by trying to infer shared (and divergent) properties and characteristics between model plants and others, mostly crops of economic value. As such, the broad knowledge obtained for one specific model organism (in casu *Arabidopsis thaliana*) can be used to understand how vital crop species, such as rice and corn, can be made more resistant to several stress factors and conditions: drought[26], bacterial/fungal infections[27,28] and herbivorous activity[29].

Applying the gained knowledge in the crop organism can be done in several ways, for example through the creation of Genetically Modified organisms (GMOs) and through marker assisted breeding. Multiple possibilities exist when creating GMOs: from enhancing the expression of interesting genes (such as resistance genes) to the introduction of new genes into the genome[30]. The public and governmental resistance to GMO crop species have made the introduction of these GMOs in Europe (and other parts of the world) very difficult, as the scientific community has failed in convincing the public of the potential benefits and reducing the fear of the potential risks[31]. Marker assisted breeding on the other hand is a technological improvement of the cultivation process employed throughout the centuries[32].

To perform these comparative genomics studies, some key concepts should be defined first. Two different seminal publications introduced some of the necessary terminology and ideas on how to describe basic genetic evolutionary principles:

- *Distinguishing homologous from analogous proteins*, written by Walter Fitch in 1970[33], defines homologous sequences as sequences derived from a common ancestor. Orthologous sequences are homologs derived through a speciation event, while paralogous sequences are homologs derived through a duplication event. To distinguish species-specific duplications from shared duplication events, the terms in-paralog and out-paralog were coined in the following years.
- *Evolution by gene duplication*, written by Susumu Ohno in 1970[34], describes how evolution and speciation are mainly driven by genomic duplications, with different fates for the resulting duplicated genes. Most duplication events result in one of the genes becoming a pseudogene. However, duplication events can also result in either neo-functionalization or sub-functionalization, where

> one of the genes is given a new function, or where the original functions (of the non-duplicated gene) are distributed among its duplicate progeny. Lastly, duplicates can also both retain the same functions, resulting in the amplification of the transcript expression.

The different types of homologous genes can easily be visualized using a phylogenetic tree[35] (see Figure 2.1), and their corresponding sequences using a multiple sequence alignment (MSA). These phylogenetic trees offer researchers a visual way to interpret, annotate and if necessary correct sets of homologous genes (also refered to as gene families). In case a gene family contains only a single gene for each species, then every internal node within the phylogenetic tree will correspond with a speciation event. If multiple genes are present per species, then the internal nodes correspond with either a speciation or a duplication event.

2.2.1 Duplications in Plant Genomes

Discovering the necessary knowledge in a model species and the transfer of this information to crop organisms can be seriously hampered by the multitude of duplications present in various plant genomes[24]. Gene and genome duplications were described as one of the basic forces in evolution by Ohno[34], and special attention needs to be given to plant genomes in this respect. In contrast to vertebrate species, plant genomes (and especially flowering plants) have a very rich and continuous history of Whole Genome Duplications (WGD)[14,36,37] . As the name implies, these duplications effectively multiply the number of chromosomes within a cell, changing the nature of an organism from a diploid tot a polyploid. Tandem duplications on the other hand, where only a single sequence is duplicated within the genome, are prevalent throughout the eukaryota[38–40]. The patterns in evolution for both types of duplication are different however, as they differ in which gene types are preserved as duplicate pairs[41]. However, the multiplicity of various duplication events in plant genomes tend to complicate the orthologous relationships between species (see also section 2.2.2).

2.2.1.1 Whole Genome Duplications

The evolutionary history of plants is rife with examples of WGDs[14], with up to 70% of the flowering plants having a polyploid history[42]. Despite the powerful biomolecular mechanisms in place to enforce the reduction of gametes[43], despite the following problematic fertility bottleneck known as the triploid block[44], and despite the following biomolecular problems such as dosage effects, WGDs are a widespread phenomenon and several rounds of WGD events can be detected in the genomes of multiple plant species[45]. Indeed, research indicates that exactly these WGD events could be responsible for the massive explosion in the number of flowering species which happened during the last 60 million years[14,45], an observation which was initially called *an abominable mystery* by Charles Darwin. These WGDs are typically followed by diploidization, in the form of severe gene loss, neo- and subfunctionalization of genes, and chromosomal rearrangments[34].

Detecting these WGDs can be done through the study of both synteny and colinearity. Synteny is the conservation of gene content within homologous regions, and colinearity is the conservation of both gene content and order within homologous regions[46]. When large colinear regions are detected within a single genome (see also Figure 2.3) this is evidence for either large-scale segmental duplications or a WGD. The detection of these WGDs and colinear regions between multiple species in general, can be a computational challenge[47,48].

2.2.1.2 Tandem Duplications

Tandem duplications are another way of increasing the protein diversity within a genome, and are also very prevalent in plant genomes[49]. Genes which are tandemly duplicated and retained often serve a very different role than their WGD counterparts[41]. Several studies have shown that, in plants, these tandem duplicates are very often involved in response to stress conditions[49,50]. Due to their sesile nature, plants need to adapt to both biotic and abiotic (e.g. cold, drought) stress conditions in different ways than metazoa. Tandem duplication events thus allow for a rapid introduction of gene variation to combat these stress conditions[49].

2.2.1.3 Allopolyploidy and Hybridization

Besides the multitude of duplications within various plant species, the presence of hybridization between (closely related) different organisms provides yet another source of complexity with regards to genomic studies. Hybrids typically result in allopolyploid genomes, with the number of haploid chromosomes equal to the sum of its haploid progenitors. Although hybridization does not occur in every plant lineage, many of the crop species (e.g. Brassica[51,52] and wheat[53]) that are of prime importance in today's agriculture are allopolyploid. Sequencing and assembly of these allopolyploid genomes can prove to be very challenging, because the differences between its homeologous chromosomes can be very small[54].

2.2.2 Orthology

One of the main interest in comparative plant genomics is the study of orthologous genes, genes which have a common ancestor and originated through a speciation event. An important tenet within the comparative genomics field is the so-called *Orthology Conjecture*, which states that orthologs more often retain the same function than paralogs[55]. While this conjecture has not gone unchallenged[56,57], more recent studies have shown that the *Orthology Conjecture* does hold true in all probability[58,59]. As such, this conjecture validates approaches to transfer functional knowledge between plants. The origin of this controversy lies within the definition given by Fitch[33], which does not distinguish between *functional orthologs* and orthologs with no shared function[60].

The detection of homologs, and orthologs in particular, is a complicated matter (see also chapter 4). Most methods start from using protein sequence similarity to measure the evolutionary distance between sequences, and based on those measurements genes are categorized as being orthologous or not[61,62]. Other methods use phylogenetic trees to infer the different types of homologous relationships[63]. These methods imply of course that orthologs with very little conserved sequence similarity can be miscategorized[64], as are genes with no common ancestor but with the same protein domains. Indeed, protein domains tend to show a high level of rearrangement in plant genomes[65], making this a potential issue. The different rates in evolution between genes[66] and between species[24,67] further complicate these problems.

Given the polyploid nature of many plant genomes (see section 2.2.1), finding orthologs between several species becomes even more difficult, as a multitude of one-to-many or many-to-many relationships (also 1-N and M-N orthologous relationships) is often introduced (see Figure 2.1). While the idea of finding perfect one-to-one orthology between species may in itself may be flawed due to possible subfunctionalizations[34], the goal to reduce the number of potential ortholog candidates is still valid. Testing and confirming the functionality of a reduced gene set is the target goal in many studies involving transfer of functional knowledge. Comparing only the protein coding sequences between orthologs may not be

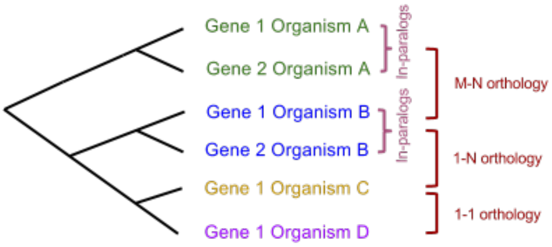


Figure 2.1: Possible orthologous relationships between organisms. *By example of in-paralogs in species A en B, three different orthologous relationships are displayed: an easy 1-1 orthologous relationship between the genes of species C en D, a 1-N orthologous relationship between the genes of species C en B (also D and B, C and A, D and A), and the complex M-N orthologous relationship between the genes of species A and B.*

sufficient to fully resolve the complex many-to-many orthologous relationships. Both expression data (see also section 5.2) and transcription factor binding site information can be used to solve these knotty problems.

2.3 Functional Genomics

Following the centuries-old habit of committing all discoveries to plain text, the resulting body of books and articles to be read before being well-versed in the functionality of a particular gene quickly became very large in the wake of the growing interest in genetics[68]. With different notations used between species and research areas when documenting the same features and characteristics, and with notations changing over time, this problem intensified only more over time. The classification of genes became quickly hampered by this text-only approach, and as such several solutions were devised.

2.3.1 Gene Ontology

The first solution to the stated problem is the Gene Ontology (GO)[68]. It consists of a well-structured vocabulary of terms which can be assigned to any gene. Due to the nature of the directed acyclic graph in which all GO terms are structured, a gene annotated with a particular GO term is also automatically annotated with the parental terms (see Figure 2.2). Within the GO there are three main domains[68], which are used to annotate genes at a different level. The *Biological Process* domain describes series of events, or a collection of molecular events with defined beginning and end. The *Molecular Function* domain contains molecular activities, without information on the used entities or context. Finally, the *Cellular Component* domain describes locations at the levels of subcellular structures.

The Gene Ontology Consortium (GOC) is constantly updating the Gene Ontology resource[69], extending the GO graph with new GO terms, replacing obsolete GO terms, and establishing new relationships between these GO terms. Care must be taken, however, to not over-interprete the available GO data because GO suffers from the *Open World Assumption*: not all data is known, and as such the absence of a gene-GO annotation is not equal to evidence that this gene-GO relationship does not exist[58].

The power of GO is not only due to its hierarchical graph structure, but also due to the variety of ev-

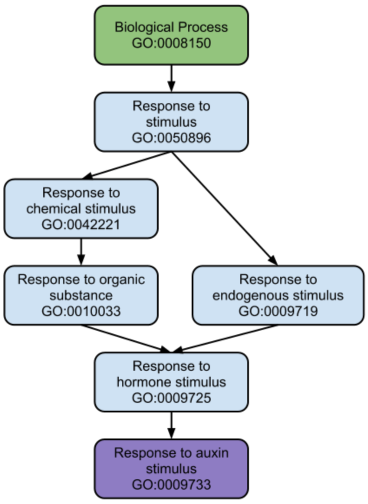

Biological Process
GO:0008150
Response to stimulus
GO:0050896
Response to chemical stimulus
GO:0042221
Response to organic substance
GO:0010033
Response to endogenous stimulus
GO:0009719
Response to hormone stimulus
GO:0009725
Response to auxin stimulus
GO:0009733

Figure 2.2: Gene Ontology graph for GO term 'Response to auxin stimulus'. *Parent-child relationship graph for the GO term 'Response to auxin stimulus', indicating the cyclic nature of the directed graph. The green GO term is the top entry for the category (Biological Process).*

idence types that can be associated with each gene-GO annotation[68,70]. We can easily associate some measure of trustworthiness with each evidence type, with major differences between the automatically assigned evidence types such as IEA (Inferred from Electronic Annotation) and those that are the result of painstaking experiments such as EXP (Inferred from Experiment). The gene-GO annotations can as such easily be sorted or filtered by their evidence types.

2.3.2 Protein Domains

Protein domains, on the other hand, are motifs from a protein sequence capable of evolving independently from the rest of the protein. These protein domains are commonly conserved through evolution, as their respective functions are often vital to the survival of the organism. As such, orthologous genes often contain strictly conserved protein domains[71]. Prediction of these protein domains, solely based on the amino acid sequence, can be done, although the actual function of the domains is often dependend on its 3d-structure. Yet, various programs exist for this very purpose[72,73], most of them based on Hidden Markov Models (HMM) . One of the more promising efforts is the InterPro database[74], which merges together the results of various other protein domain databases such as PFAM[72] and PANTHER[73]. One of the key differences with the Gene Ontology is that these protein domains are useful only for protein coding genes. As such, other gene types, such as the variety of RNA genes, cannot be described in terms protein domains. At the same time, a major effort has been put forward to map InterPro domains to associated GO terms, merging the efforts of both research fields[74].

2.3.3 Molecular Interactions

Although the Gene Ontology and protein domains annotations can provide valuable information about a single gene, no interaction information between genes or gene products is captured this way. To remediate this problem, several solutions have been put forward:

- PlantCyc[75] (with AraCyc[76] being the *Arabidopsis* specific version) is an effort to standardize information about metabolic pathways, reactions and compounds. The study of biochemical pathways within plants is augmented by this approach, with the different Cyc-versions (MediCyc for *Medicago*, CornCyc for Maize, GrapeCyc for *Vitis*, etc.) being helpful in understanding the evolution of pathways.

- AraNet[77] is a network-based approach to study genes from an *omics* point of view. By aggregating multiple data sources from different plants and using a bayesian approach to integrate this data, the number of genes in an annotated network in increased drastically. By using the AraNet resource users can make informed decisions on, for example, trait-association of genes.

2.3.4 Text Mining

Despite the stated efforts to categorize the gained scientific knowledge, several issues still remain: the content from older publications becomes part of the databases at a slow pace (if at all), not all knowledge can be captured by the given ontologies, changing annotations and unstable gene identifiers are not acknowledged, etc. This all points to a global problem with data integration. To remediate some of these issues, data mining the scientific publications can result in extra information for researchers[78]. This automated approach is commonly based on machine-learning methods (ML) , and thus offers some of its advantages (such as being able to deal with natural language processing) and its disadvantages (such as being dependent on high-quality training data sets).

2.4 Bioinformatics Tools and Platforms

The importance of bioinformatics tools and platforms can be studied by considering that a growing number of scientific journals (in the field of genetics) are dedicated to publishing these tools (e.g. Bioinformatics), or have dedicated sections for publishing these tools (Nucleic Acids Research, Genome Biology). With ever-growing amounts of data, two needs arise:

1. Tools and algorithms which can handle this data. The focus of the publications here are other bioinformaticians who will use these tools to perform custom analyses on their own datasets. Examples include software for clustering proteins into gene families (e.g. TribeMCL[79] and OrthoMCL[61]), software for delineating colinearity between genomic regions (e.g. MCScan[80] and I-ADHoRe[48]) and software for assembling reads from Next Generation Sequencing (NGS) technologies (e.g. Velvet[81] and Abyss[82]). These tools are most often accessed through the command line interface, an interface unfamiliar to the majority of life science researchers[83]. Efficiency and reliability are often the primary concerns, with user-friendliness a distant afterthought[83].

2. Platforms through which pre-processed data can be accessed, with varying degrees of user-supplied data. The focus here are biologists who will use the platforms as a reference, to guide wetlab experiments. Examples include online platforms for genomics (e.g. PLAZA[84] and TAIR[85]), platforms

for protein domains (e.g. PFAM[72] and PANTHER[73]), etc. These platforms are accessed through a Graphical User Interface (GUI) , and are often online and thus accessible through a web browser. User-friendly and intuitive interfaces are of primary concern here, as a steep learning curve will often drive inexperienced users away.

The two purposes are not fully mutually exclusive, but when developing software the target user-base must be clearly defined, otherwise one risks targeting neither. In this thesis, the focus was on the development of bioinformatics platforms, and as such a lot of thought and care was put into defining the correct user-computer interactions, and providing ample documentation and background material[86]. Besides the aspects visible to the end-user, namely data content and representation, various server-side factors can also influence the success of a web resource: portability[86], software design principles such as modularity[87], maintenance[88], etc.

2.4.1 Web Visualizations and Technologies

Scientific visualizations[89] are, even in the automated world of today, very important, especially in the fields of genetics and molecular biology[90]: browsing through genomes[91], visualizing phylogenetic trees[35], making sense of a gene interaction network[92], etc. Understanding multi-dimensional data, when one is not even fully sure what he's looking at, is the first purpose[93]. The second one is to communicate ideas and information[93] to other people. Using well-designed visualizations in online platforms is thus a way to convey key concepts and information to users.

Differences exist between data representations and visualizations on paper versus website[93]. Not only is there in the latter an expectancy of interactivity (zooming, data selection data filtering, linking), but the cost and space limitation in scientific journals implies that printed images should contain as much information as possible. Another distinction is the time required for generating the visualization: intra-species colinear regions are commonly visualized in genome papers using the Circos software[94]. This software package can generate very high quality infographics using multiple extra datatypes besides colinearity, but it can take multiple minutes to render a non-interactive illustration. This can be compared with the instant-rendered Circle Plot for colinearity in PLAZA, albeit with less features (see Figure 2.3).

When creating online visualizations and charts, several technologies are available, each with its own advantages and disadvantages:

- *Static images*, often combined with a clickable map to facilitate linking, are the most basic form of representation. No animations are possible, but its rather low memory requirements and ease of implementation result in a strong presence in many websites.
- *Flash* applications, often combined with snippets of JavaScript, are executed by a browser plugin and are used in very different kinds of web applications: from simple browser games to intricate and complex applications for online shopping. Many different charting tools[96,97] depend on the use of Flash, as it is capable of visually attractive and interactive graphics.
- *Java Applets* are a rather old technology, and also one of the first attempts to offer interactive applications through the web browser. The rather slow loading time of the initial Java Virtual Machines (JVM) resulted in its low adaptation, and many applets have been phased out over time in favor of other technologies. Still, most computers have a Java plugin installed, thus making it a feasible choice.

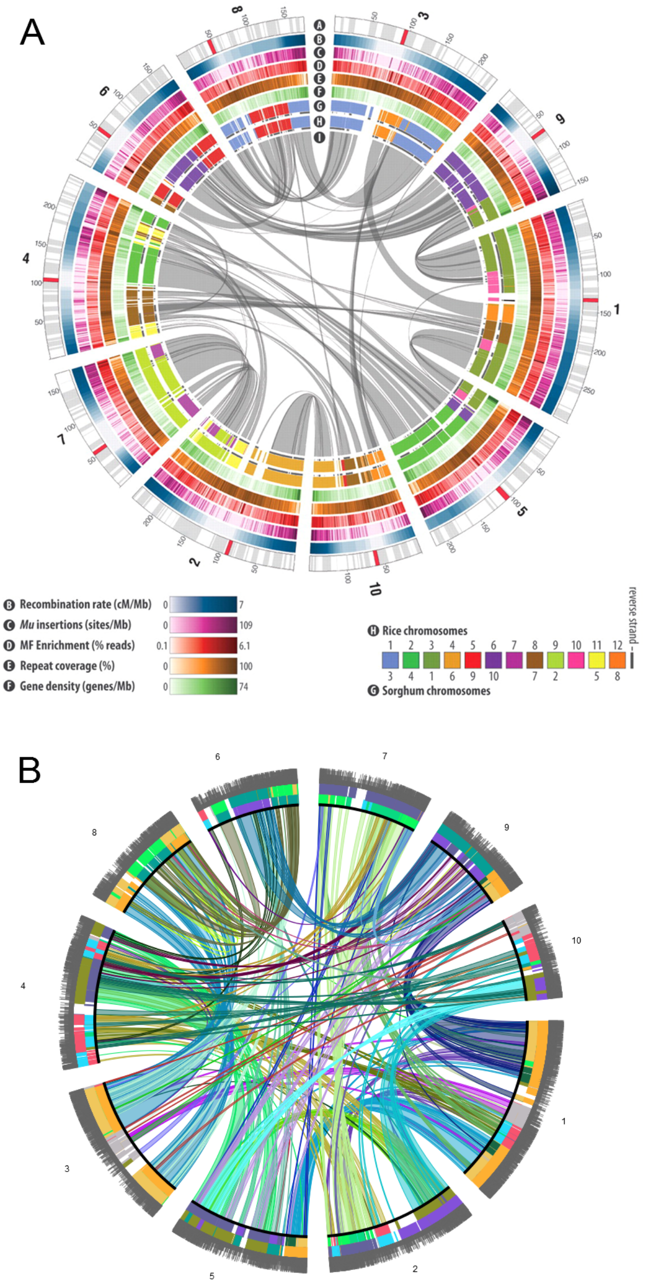


Figure 2.3: Circos and PLAZA colinearity of the maize genome. *Circleplots showing intra-species colinearity within the maize genome, together with inter-species colinearity with rice and sorghum. (A) The Circos plot, also showing general genome statistics*[95]*. (B) PLAZA plot.*

- *Scalable Vector Graphics (SVG)* are XML-like representations of graphics. All content within an SVG are objects, and can thus be accessed as such: animation, linking, and other actions can be defined per object. However, this flexibility comes with a performance penalty, and most implementations which support the SVG standard have memory troubles when several thousands of objects are present on a single web page. As the name implies these graphics are vector based, and can thus be rendered at any resolution.

- *JavaScript* graphics, more specifically the graphics associated with the *HTML5 Canvas object*, are a more modern approach to visualizations. While they can be considered a step back (compared to the SVG standard) as these graphics do not provide a standard interactivity, their use is becoming more widespread. The increased availability of flexible JavaScript libraries is further strengthening its presence in the online (and increasingly offline as well[98]) world.

Not all web browsers are capable of handling the same content however, which is unfortunately also a point to consider before implementation. Other technology restrictions are brought forth by the shift in focus towards mobile computing devices: Apple's IPad for example, does not support either *Flash* or *Java*.

2.4.2 Online Plant Genomics Platforms

The availability of online sequence databases and genome browsers provides an easy entry point for researchers to immediately investigate genome information without having to install any software. Furthermore, such services usually provide the possibility to link with an assembly of other web-based resources[99].

The development of any software or platform should be preceded by looking at currently available solutions. During the initial planning phase of PLAZA[a][100] (in 2008), very few online solutions for plant genomics existed, and none combined comparative, evolutionary and functional analyses (see Table 2.1). Now, in 2012, it is important to review whether any other platforms have been created, and whether platforms which existed in 2008 have been further developed (see Table 2.2). One of the major differences between the various platforms is how many organisms are included. Some platforms focus only on a single organism, such as TAIR[b][85] for *Arabidopsis thaliana* and Chlamydomonas Connection[c] for *Chlamydomonas*, and provide no or very limited tools for comparative studies. While these platforms certainly do cater to a specific niche, they fall outside the scope of this comparison. Other platforms focus more on a relative small group of phylogenetically close organisms, such as LegumeIP[d][101], targetting breeder-specific questions to improve legume crops. More general plant genomics platforms include plants from various lineages, such as GreenPhylDB[e][102], Phytozome[f][103] and EnsemblPlants[g][104]. Finally, the CoGe platform[h][105] does not focus on any lineage or even kingdom, but rather includes a vast collection of species that range from bacteria and viruses, to metazao and plants.

[a] http://bioinformatics.psb.ugent.be/plaza
[b] http://www.arabidopsis.org
[c] http://www.chlamy.org/
[d] http://plantgrn.noble.org/LegumeIP/
[e] http://greenphyl.cirad.fr/
[f] http://www.phytozome.net
[g] http://plants.ensembl.org/index.html
[h] http://synteny.cnr.berkeley.edu/CoGe/

Comparison of the features and data included in these platforms is not always straightforward, as the various target audiences require specific needs. However, some basic questions can still be answered:

- How many species are present?
- What kind of functional annotation is available?
- Are the genes clustered in homologous or orthologous groups?
- What tools and visualizations are available?
- Is all data available for download?

Since we are comparing these platforms with our own PLAZA platform, it is possible though that a certain bias may present itself, as we may feel that certain tools are more important than others due to our different research focus. New tools and data types have become available in the period 2008-2012 and as such a strict comparison between Table 2.1 and Table 2.2 is not completely possible. Finally, some online platforms that were available in 2008 have not been updated since then (e.g. Genome Cluster Database[i] [106], OrthologID[j] [107], PlantTribes[k] [108] and SynBrowse[l] [109]) and are thus not included in the comparison of 2012 (Table 2.2). EnsemblPlants is the succesor of Gramene, and as such Gramene is also not included anymore in the comparison of 2012.

The most obvious improvement in most updated platforms is the steep rise in number of available genomes. In 2008 only a limited number of plant genomes were publicly available, while in 2012 more than 30 plant genomes have been published and are freely available for academic research. An important factor to consider is thus how well the platforms and their database schemes scale with a non-linear increase in available genome data (see also chapter 8).
Other platforms provide data focused on specific gene functions or sequence types but are not extensively described here. Plant transcription factors can be studied using PlnTFDB[m] [112], AGRIS[n] [113], and GRASSIUS[o] [114]. The complementary platforms Phytome [115] and SPPG[p] [116] are hybrid systems integrating gene information from genome sequencing projects with EST data for a comprehensive set of plant species.

2.5 Author Contribution

All content was written by myself, except for the comparison of platforms in section 2.4.2 which was partially retrieved from the PLAZA v1.0 paper [100] and thus written by Sebastian Proost, Klaas Vandepoele and myself.

[i] http://bioweb.ucr.edu/databaseWeb/index.jsp
[j] http://nypg.bio.nyu.edu/orthologid/
[k] http://fgp.bio.psu.edu/tribedb/index.pl
[l] http://www.synbrowse.org/
[m] http://plntfdb.bio.uni-potsdam.de/
[n] http://arabidopsis.med.ohio-state.edu/
[o] http://grassius.org/plantgenome.html
[p] http://bioinformatics.psb.ugent.be/cgi-bin/SPPG/index.htpl

Name	#Species	Gene Families	Phylogenetic trees	WGDotplot	Inter species colinearity	Functional annotation	Genome Browser	Comments
PLAZA[100]	9	X	X	X	X	X	X	Multi-species colinearity views, K_S-dating tool, family-wise similarity heatmap and workbench
Genome Cluster Database[106]	2	X	X			X		Chromosome map and link with *Arabidopsis* expression data
GreenPhylDB[102]	2	X	X			X		Manual curation of a subset of families
OrthologID[107]	3+2	X	X					Diagnostic characters per orthologous group
Plant Genome Duplication Database[46]	7			X	X			Genome-wide mapping tool for homologous sequences and syntenic locus search
Phytozome[103]	14	X			+/-	X		
PlantTribes[108]	5	X				X		Link with *Arabidopsis* expression data
CoGe[105]	14			X	X	X	X	DNA-based sequence comparisons (Conserved Non-coding Sequences)
SynBrowse[109]	3				X			Synteny browser based on GBrowse (n intra-species colinearity)
Gramene[110]	6	X	X	+/-	+/-	X	X	Based on Ensembl pipeline

Table 2.1: Comparison of online tools for plant genomics in 2008.

Name	#Species	Gene Families	Phylogenetic trees	WGDotplot	Inter species colinearity	Functional annotation	Genome Browser	Expression data	Genetic Variation	Functional clusters
PLAZA v2.5[84]	25	X	X	X	X	X	X			X
GreenPhylDB v3[111]	22	X	X			X				
Plant Genome Duplication Database[46]	26			X	X					
Phytozome v8[103]	31	X			+/-	X	X			
CoGe[105]	433 (?)			X	X	X	X			
EnsemblPlants[104]	19	X	X	+/-	+/-	X	X	X	X	
LegumeIP[104]	5	X	X		X	X	X	X		

Table 2.2: Comparison of online tools for plant genomics in 2012.

"With insufficient data it is easy to go wrong."

Carl Sagan

3

PLAZA: a Comparative Genomics Resource to Study Gene and Genome Evolution in Plants

Abstract

The number of sequenced genomes of representatives within the green lineage is rapidly increasing. Consequently, comparative sequence analysis has significantly altered our view on the complexity of genome organization, gene function, 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. Here, we describe PLAZA, an online platform for plant comparative genomics[a]. This resource integrates structural and functional annotation of published plant genomes together with a large set of interactive tools to study gene function and gene and genome evolution. Precomputed data sets cover homologous gene families, multiple sequence alignments, phylogenetic trees, intraspecies whole-genome dot plots, and genomic collinearity between species. Through the integration of high confidence Gene Ontology annotations (selected based on GO evidence codes) and tree-based orthology between related species, thousands of genes lacking any functional description are functionally annotated. Advanced query systems, as well as multiple interactive visualization tools, are available through a user-friendly and intuitive web interface. In addition, detailed documentation and tutorials introduce the different tools, while the workbench provides an efficient means to analyze user-defined gene sets through PLAZA's interface. In conclusion, PLAZA provides a comprehensible and up-to-date research environment to aid researchers in the exploration of genome information within the green plant lineage.

This chapter is based on Proost et al. [100]. For the author contributions, see page 3-19.

[a] http://bioinformatics.psb.ugent.be/plaza/

3.1 Introduction

The availability of complete genome sequences has significantly altered our view on the complexity of genome organization, genome evolution, gene function, and regulation in plants. Whereas large-scale cDNA sequencing projects have generated detailed information about gene catalogs expressed in different tissues or during specific developmental stages[117], the application of genome sequencing combined with high-throughput expression profiling has revealed the existence of thousands of unknown expressed genes conserved within the green plant lineage[116,118]. The generation of high-quality complete genome sequences for the model species *Arabidopsis thaliana* and rice (*Oryza sativa*) required large international consortia and took several years before completion[13,119]. Facilitated by whole-genome shotgun and next-generation sequencing technologies, genome information for multiple plant species is now rapidly expanding. The genomes of four eudicots, *Arabidopsis thaliana*, poplar (*Populus trichocarpa*), grapevine (*Vitis vinifera*), and papaya (*Carica papaya*), two monocots, rice and *Sorghum bicolor*, the moss *Physcomitrella patens*, and several green algae[120] have been published, and new genome initiatives will at least double the number of plant genome sequences by the end of this decade[121,122].

Although the genomes of some of these species provide invaluable resources as economical model systems, comparative analysis makes it possible to learn more about the different characteristics of each organism and to link phenotypic with genotypic properties. Hanada and coworkers demonstrated how the integration of expression data and multiple plant sequences combined with evolutionary conservation can greatly improve gene discovery[99,123]. Whereas a detailed gene catalog provides a starting point to study growth and development in model organisms, sequencing species from different taxonomic clades generates an evolutionary framework to study how changes in coding and noncoding DNA affect the evolution of genes, resulting in expression divergence and species-specific adaptations[124–126]. Based on orthologous genes (i.e., genes sharing common ancestry evolved through speciation), comparative genomics provides a powerful approach to exploit mapping data, sequence information, and functional information across various species[127]. Similarly, the analysis of genes or pathways in a phylogenetic context allows scientists to better understand how complex biological processes are regulated and how morphological innovations evolve at the molecular level. For example, studying gene duplicates in poplar has revealed specific expansions in gene families related to cell wall formation covering cellulose and lignin biosynthesis genes and genes associated with disease and insect resistance[21]. Similarly, amplifications of genes belonging to the metabolic pathways of terpenes and tannins in grapevine directly relate the diversity of wine flavors with gene content[128]. Besides the comparative analysis of specific gene families in higher plants, comparisons with other members of the green lineage provide additional information about the evolutionary processes that have changed gene content during hundreds of millions of years. Although the genomes of, for instance, moss and green algae contain a smaller number of genes compared with flowering plants, they provide an excellent starting point to reconstruct the ancestral set of genes at different time points during plant evolution and to trace back the origin of newly acquired genes[129,130].

Gene duplication has been extensive in plant genomes. In addition, detailed comparison of gene organization and genome structure has identified multiple whole-genome duplication (WGD) events in different land plants. From a biological point of view, the large number of small- and large-scale duplication events in flowering plants has had a great influence on the evolution of gene function and regulation. For instance, between 64 and 79% of all protein-coding genes in *Arabidopsis thaliana*, poplar, and rice are part of multigene families, compared with 40% for the green alga *Chlamydomonas reinhardtii*. Paralogs are generally considered to evolve through nonfunctionalization (silencing of one copy), neofunctionalization (acquisition of a novel function for one copy), or subfunctionalization (partitioning of tissue-specific patterns of expression of the ancestral gene between the two copies)[131,132]. The impact of the large number of duplicates on the complexity, redundancy, and evolution of regulatory networks in multicellular organisms is currently far from being well understood[133,134].

Performing evolutionary and comparative analyses to study gene families and genome organization requires a centralized plant genomics infrastructure where all information generated by different sequencing initiatives is integrated, in combination with advanced methods for data mining. Even though general formats have been developed to store and exchange gene annotation[135], the properties of available plant genomic data (i.e., structural annotation of protein-coding genes, RNAs, transposable elements, pseudogenes, or functional annotations through protein domains or ontologies) vary greatly between different sequencing centers, impeding comparative analyses for nonexpert users. Additionally, large-scale comparisons between multiple eukaryotic species require huge computational resources to process the large amounts of data. Here, we present PLAZA, a new online resource for plant comparative genomics[b]. We show how PLAZA provides a versatile platform for integrating published plant genomes to study gene function and genome evolution. Precomputed comparative genomics data sets cover homologous gene families, multiple sequence alignments, phylogenetic trees, intraspecies whole-genome dot plots, and genomic collinearity information between species. Multiple visualization tools that are available through a user-friendly web interface make PLAZA an excellent starting point to translate sequence information into biological knowledge.

3.2 Results

3.2.1 Data Assembly

The first version of PLAZA contained the nuclear and organelle genomes of nine species within the Viridiplantae kingdom: the four eudicots *Arabidopsis thaliana*, papaya, poplar, and grapevine, the two monocots rice and sorghum, the moss *Physcomitrella patens*, and the unicellular green algae *Chlamydomonas reinhardtii* and *Ostreococcus lucimarinus*. The integration of all gene annotations provided by the different sequencing centers yielded a data set of 295 865 gene models, of which 92.6% represent protein-coding genes (Table 3.1). The remaining genes are classified as transposable elements, RNA, and pseudogenes (6.5, 0.6, and 0.3%, respectively). Whereas most of the genes are encoded in the nuclear genomes, a small set are from chloroplast and mitochondrial origin (0.4 and 0.2%, respectively). For all genes showing alternative splicing, the longest transcript was selected as a reference for all downstream comparative genomics analyses. Detailed gene annotation, including information about alternative splicing variants is displayed using the AnnoJ[c] genome browser[136]. Whereas genomes from model species like *Arabidopsis thaliana* and rice are characterized by high sequence coverage and a set of contiguous genomic sequences resembling the actual number of chromosomes, other genome sequences, such as those of *Physcomitrella patens* and papaya, are produced by the whole-genome shotgun sequencing method and contain more than 1000 genomic scaffolds (Table 3.1). For poplar, grape, and sorghum, a large fraction of the genome is assembled into chromosomes, but several scaffolds that could not be anchored physically are still present in the data set. In this case, we allocated the genes that were not assigned to a

[b] http:// bioinformatics.psb.ugent.be/plaza/

[c] http://www.annoj.org/

Species	Size	Genes (a)	Scaffolds (b)	Coding	GO (c)	InterPro
Arabidopsis thaliana	115 Mb	33,284 (88.81%)	5	27,228	63.62% (66.21%)	56.49%
Carica papaya	271 Mb	28,072 (99.84%)	1,898	28,072	0.00% (22.88%)	57.75%
Populus trichocarpa	410 Mb	45,699 (99.90%)	19+1 (5,724)	45,654	44.69% (52.89%)	61.91%
Vitis vinifera	468 Mb	38,127 (99.63%)	19+1 (35)	37,987	40.09% (45.90%)	57.62%
Oryza sativa	371 Mb	57,955 (72.32%)	12	41,912	30.42% (30.91%)	63.69%
Sorghum bicolor	626 Mb	34,686 (99.78%)	10+1 (217)	34,609	44.44% (48.13%)	67.79%
Physcomitrella patens	480 Mb	36,137 (99.80%)	1,446	36,065	33.20%	42.44%
Chlamydomonas reinhardtii	121 Mb	14,731 (99.64%)	552	14,678	34.99%	49.29%
Ostreococcus lucimarinus	13 Mb	7,805 (100.00%)	21	7,805	47.94%	62.86%
Total		295,865 (92.60%)		273,965	39.36%	44.88%

Table 3.1: Summary of the Gene Content in PLAZA v1. *(a) percentage of protein coding genes. (b) Numbers in parentheses refer to the number of genomic sequences in the original annotation; "+1" indicates the creation of a virtual chromosome zero to group scaffolds. (c) Percentages in parentheses include projected GO annotations, while the first value only reports original primary GO data.*

chromosome in the original annotation to a virtual chromosome zero. This procedure reduces the number of pseudomolecules when applying genome evolution studies while preserving the correct proteome size (i.e., the total number of proteins per species) and the relative gene positions on the genomic scaffolds (Table 3.1).

Complementary to the structural annotation, we also retrieved, apart from free-text gene descriptions, functional information through Gene Ontology (GO) associations[68], InterPro domain annotations[74], and Arabidopsis Reactome[d] pathway data[137]. Whereas GO provides a controlled vocabulary to describe gene and gene product attributes (using Cellular Component, Biological Process, and Molecular Function), the InterPro database provides an annotation system in which identifiable features found in known proteins (i.e., protein families, domains, and functional sites) can be applied to new protein sequences. GO provides a set of different evidence codes that indicate the nature of the evidence that supports a particular annotation. The Arabidopsis Reactome is a curated resource for pathways where enzymatic reactions are added to genes and a set of reactions is grouped into a pathway.

Apart from the basic information related to gene structure and function (e.g., genome coordinates, mRNA coding and protein sequences, protein domains, and gene description), different types of comparative genomics information are provided through a variety of web tools. In general, these data and methods can be classified as approaches to study gene homology and genome structure within and between species. Whereas the former focuses on the organization and evolution of families covering homologous genes, the latter exploits gene collinearity, or the conservation of gene content and order, to study the evolution of plant genomes (Figure 3.1).

3.2.2 Delineating Gene Families and Subfamilies

As a starting point to study gene function and evolution, all protein-coding genes are stored in gene families based on sequence similarity inferred through BLAST[138]. A gene family is defined as a group of two or more homologous genes. A graph-based clustering method (Markov clustering implemented in Tribe-MCL[79]) was used to delineate gene families based on BLAST protein similarities in a process that is sensitive to the density and the strength of the BLAST hits between proteins. Although this method

[d] http://www.arabidopsisreactome.org/

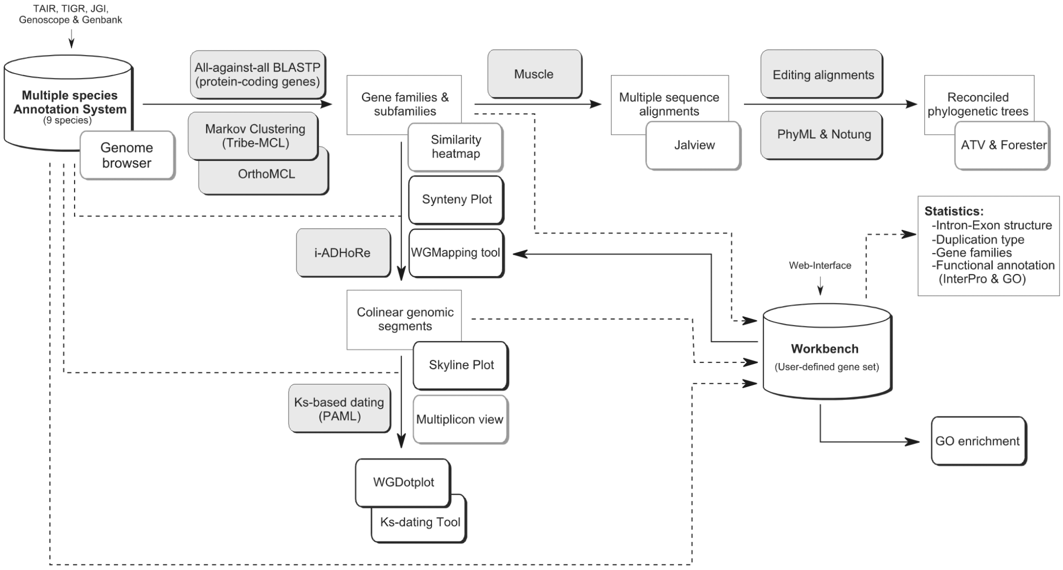


Figure 3.1: Structure of the PLAZA Platform. *Outline of the different data types (white boxes) and tools (gray rounded boxes) integrated in the PLAZA platform. White rounded boxes indicate the different tools implemented to explore the different types of data available through the website.*

is very well suited for clustering large sets of proteins derived from multiple species, high false-positive rates caused by the potential inclusion of spurious BLAST hits have been reported[133]. Therefore, we applied a postprocessing procedure by tagging genes as outliers if they showed sequence similarity to only a minority of all family members (see Methods 3.3.1). The OrthoMCL method[61] was applied to build subfamilies based on the same protein similarity graph. Benchmark experiments have shown that OrthoMCL yields fewer false positives compared with the Tribe-MCL method and that, overall, it generates tighter clusters containing a smaller number of genes[133]. Because OrthoMCL models orthology and in-paralogy (duplication events after dating speciation) based on a reciprocal-best hit strategy, the final protein clusters will be smaller than Tribe-MCL clusters because out-paralogs (homologs from duplication events predating speciation) will not be grouped. Therefore, from a biological point of view, subfamilies or out-paralogs can be considered as different subtypes within a large protein family. In total, 77.62% of all protein-coding genes (212 653 genes) are grouped in 14 742 multigene families, leaving 61 312 singleton genes. Sixty-two percent of these families cover genes from multiple species, and for approximately one-fifth, multiple subfamilies were identified. Manual inspection and phylogenetic analysis of multiple families revealed that in many cases, OrthoMCL correctly identified outparalogous groups that can be linked with distinct biological subtypes or functions (see Section 3.3.2,[49]). Examples of identified subfamilies are different clathrin adaptors (Adaptor Protein complex subunits), minichromosome maintenance subunits, ATP binding GCN transporters, cullin components of SCF ubiquitin ligase complexes, replication factors, and a/b/g tubulins (Figure 3.2). Although fast-evolving genes or homologs showing only limited sequence similarity can lead to incorrect families, a similarity heat map tool was developed to explore all pairwise sequence similarities per family (Figure 3.2). This visualization provides an intuitive approach, complementary to the automatic protein clustering and phylogenetic trees, to explore gene homology. In addition, a BLAST interface is available that provides a flexible entry point to search for homologous genes using user-defined sequences and parameter settings.

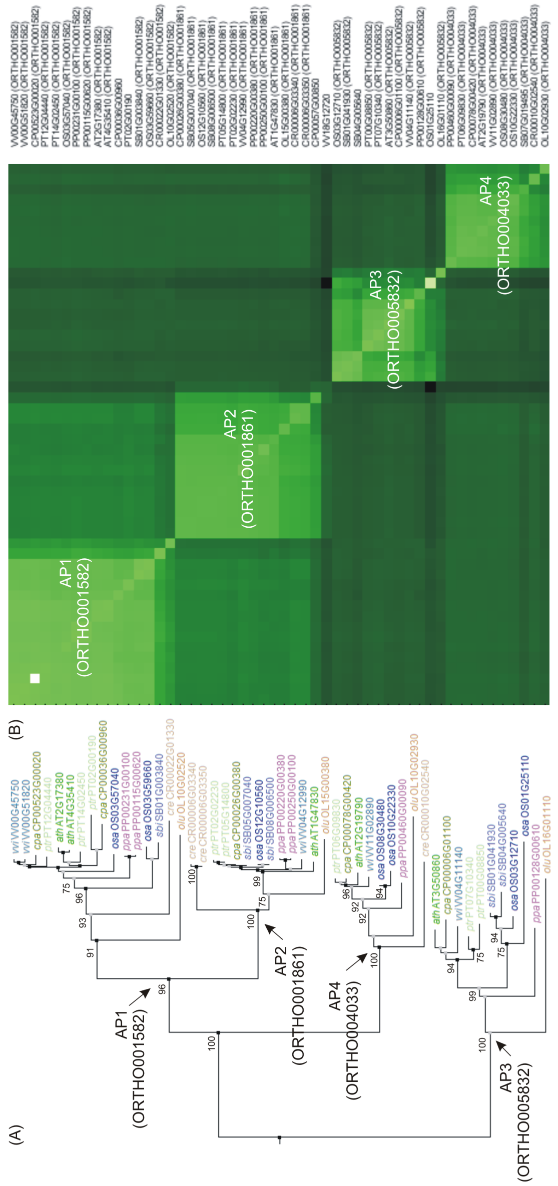


Figure 3.2: Gene Family Delineation Using Protein Clustering, Phylogenetic Tree Construction, and Similarity Heat Maps. *(A) Phylogenetic tree of clathrin adaptors (HOM000575) with the AP1–4 subfamilies delineated using OrthoMCL. Black and gray squares on the tree nodes indicate duplication and speciation events identified using tree reconciliation, respectively. Only bootstrap values ≥70% are shown. (B) Similarity heat map displaying all pairwise similarity scores for all gene family members. BLAST bit scores were converted to a color gradient with white/bright green and dark green indicating high and low scores, respectively. Clustering of the sequence similarities supports the existence of the four AP subfamilies that were identified using protein clustering and confirmed using phylogenetic inference. Note that subfamilies AP3 and AP4 are inverted in the heat map compared with the tree.*

3.2.3 Projection of Functional Annotation Using Orthology

Phylogenetic studies generate valuable information on the evolutionary and functional relationships between genes of different species, genomic complexity, and lineage-specific adaptations. In addition, they provide an excellent basis to infer orthology and paralogy[60]. Based on the gene families generated using protein clustering, a phylogenetic pipeline was applied to construct 20 781 phylogenetic trees covering $\sim$172 000 protein-coding genes. Bootstrapped phylogenetic trees were constructed using the maximum likelihood method PhyML[139] based on protein multiple sequence alignments generated using MUSCLE[140] (see Section 3.3.3). In order to extract biological information from all phylogenies, we applied the NOTUNG tree reconciliation method to annotate, based on parsimony and a species tree, tree nodes as duplication/speciation events together with a time estimate[141]. Detailed inspection of tree topologies revealed that, even for well-supported nodes with high bootstrap values, a high number of nodes (53 to 64%) correspond with falsely inferred duplication events. This problem is caused by the different rates of amino acid evolution in different species, potentially leading to incorrect evolutionary reconstructions[142]. Therefore, we calculated a duplication consistency score, originally developed by Ensembl[143], to identify erroneously inferred duplication events (see Section 3.3.3). This score reports, for a duplication node, the intersection of the number of postduplication species over the union and is typically high for tree nodes denoting a real duplication event. Consequently, the reconciled phylogenetic trees provide a reliable means to identify biologically relevant duplication and speciation events (or paralogs and orthologs, respectively). In addition, the time estimates at each node make it possible to infer the age of paralogs and correlate duplications with evolutionary adaptations.

Since speciation events inferred through phylogenetic tree construction provide a reliable way to identify orthologous genes, these orthology relationships can be used to transfer functional annotation between related organisms[137,144,145]. We applied a stringent set of rules to identify a set of eudicot and monocot tree-based orthologous groups and used GO projection to exchange functional annotation between species (see Section 3.3.4 and Figure 3.3). Whereas in the original annotation, 39% of all proteins were annotated with at least one GO term, this fraction greatly varies for different species (Table 3.1). Model species like *Arabidopsis thaliana* and rice have a large set of functionally annotated genes with GO terms supported by various experimentally derived evidence codes. In contrast, other organisms only have annotations inferred through electronic annotation (e.g., grapevine and popular) or completely lack functional annotation (e.g., papaya; see data overview on PLAZA website). Application of GO projection using eudicot and monocot orthologous groups resulted in new or improved functional information for 36 473 genes. This projected information covers $\sim$105 000 new annotations, of which one-fifth is supported by evidence from multiple genes. Overall, 11.8% of all genes lacking GO information in flowering plants could be annotated based on functional data of related genes/species and for $\sim$22 000 genes (17% of protein-coding genes in angiosperms already annotated using GO) new or more specific GO terms could be assigned. For papaya, initially lacking functional GO data, 39% of all genes for which a phylogenetic tree exists have now one or more associated GO term. To estimate the specificity of the functional annotations, we used the GO depth (i.e., the number of shortest-path-to-root steps in the GO hierarchy) as a measure for the information content for the different annotations. Distributions per species reveal that the projected annotations are as detailed as the original primary GO data and that for species initially lacking GO information, detailed GO terms can be associated to most genes[100]. Whereas Blast2GO, a high-throughput and automatic functional annotation tool[146], applies sequence similarity to identify homologous genes and collect primary GO data, GO projection uses phylogenetic inference to identify orthologous genes prior to transfer of functional annotation. Both methods incorporate information from different GO evi-

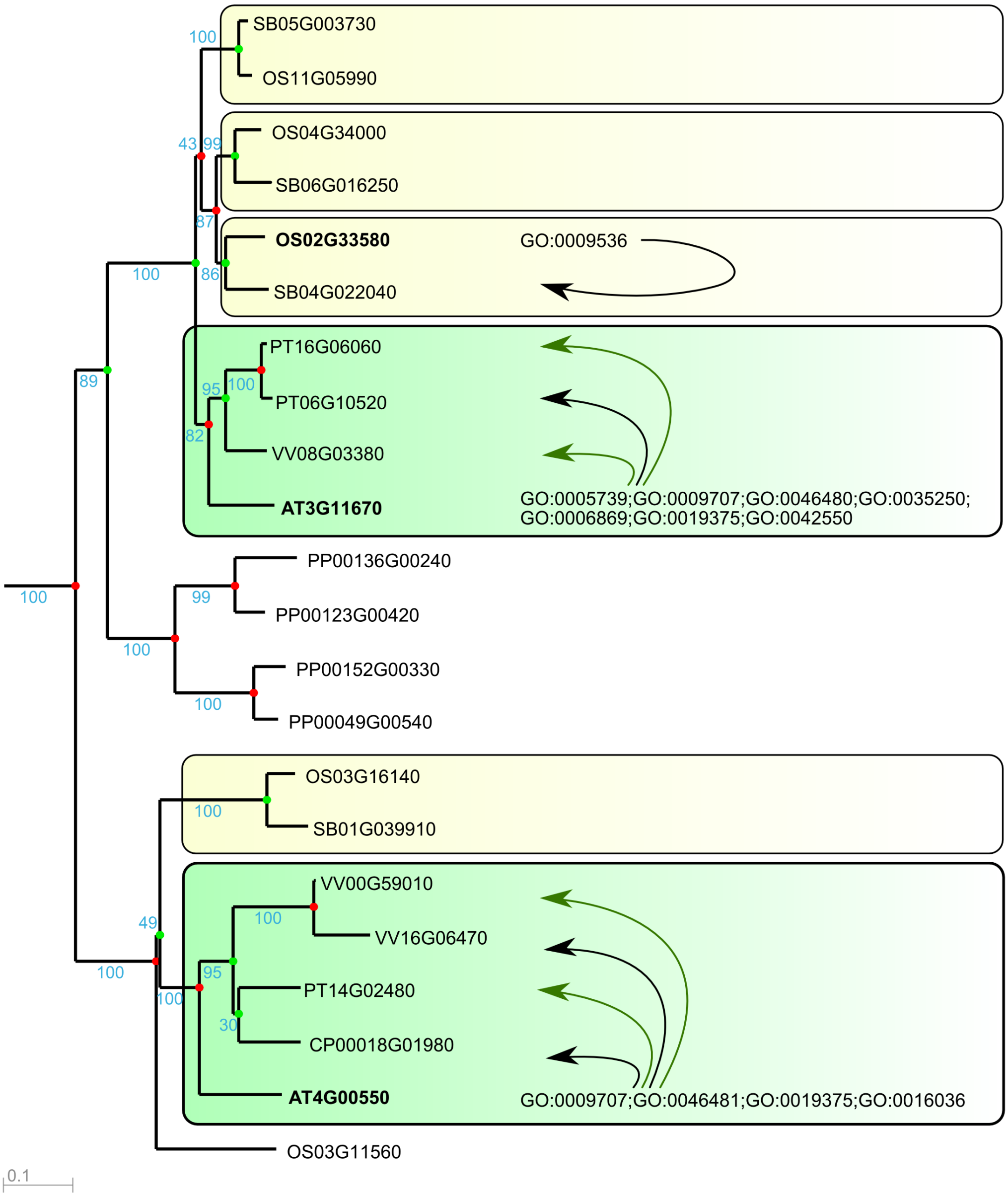


Figure 3.3: GO projection using eudicot and monocot orthologous groups. *The rounded boxes indicate the orthologous groups extracted from the phylogenetic tree while green and yellow shadings refer to eudicot and monocot clades, respectively. If for genes in an orthologous group functional annotation was available (excluding GO annotations with an IEA evidence tag), these terms were transferred to all other genes (with ISS evidence tag) in that group keeping track of the source gene(s). Consequently, some un-annotated genes received new functional annotations while other genes were re-annotated with a more specific GO term (black and green arrows, respectively). In this example the green arrow denotes the re-annotation of the GO term 'biosynthetic process' (GO:0009058, depth 2) using 'galactolipid biosynthetic process' (GO:0019375, depth 6).*

dence tags to avoid the inclusion of low-quality annotations while generating functional information for uncharacterized proteins. It is important to note that all pages and tools presenting functional annotation through the PLAZA website can be used, including either all GO data or only the primary GO annotations (i.e., excluding projected GO terms).

3.2.4 Exploring Genome Evolution in Plants

To study plant genome evolution, PLAZA provides various tools to browse genomic homology data, ranging from local synteny to gene-based collinearity views. Whereas collinearity refers to the conservation of gene content and order, synteny is more loosely defined as the conservation of similar genes over two or more genomic regions. Moreover, genome organization can be explored at different levels, making it possible to easily navigate from chromosome-based views to detailed gene-centric information for one or multiple species. Based on gene family delineation and the conservation of gene order, homologous genomic regions were detected using i-ADHoRe[147]. The i-ADHoRe algorithm combines gene content and gene order information within a statistical framework to find significant microcollinearity taking into account different types of local rearrangements[148]. Subsequently, these collinear regions are used to build genomic profiles that allow the identification of additional homologous segments. As a result, sets of homologous genomic segments are grouped into what is referred to as a multiplicon. The multiplication level indicates the number of homologous segments for a given genomic region. The advantage of profile searches (also known as top-down approaches) is that degenerate collinearity (or ancient duplications) can still be detected[148,149].

The Synteny plot is the most basic tool to study gene-centric genomic homology. This feature shows all genes from the specified gene family with their surrounding genes, providing a less stringent criterion to study genomic homology compared with collinearity. To ensure the fast exploration of positional orthologs, gene family members have been clustered based on their flanking gene content. Investigating collinearity on a genome-wide scale can be done using the WGDotplot (Figure 3.4A). This tool can be applied to identify large-scale duplications within a genome or to study genomic rearrangements within or between species (e.g., after genome doubling or speciation, respectively). In a first view, a genome-wide plot displays inter- or intraspecies collinearity, while various features are available to zoom in to chromosomewide plots or the underlying multiplicon gene order alignment. Intraspecies comparisons can also be visualized using circular plots that depict all duplicated blocks physically mapped on the chromosomes.

All collinear gene pairs (or block duplicates) have been dated using K_S, the synonymous substitution rate (see Section 3.3.6). K_S is considered to evolve at a nearly constant neutral rate since synonymous substitutions do not alter the encoded amino acid sequence. As a consequence, these values can be used as a molecular clock for dating, although saturation (i.e., when synonymous sites have been substituted multiple times, resulting in K_S-values >1) can lead to underestimation of the actual age[150]. The average K_S for a collinear (or duplicated) block is calculated and colored accordingly in the WGDotplots (Figure 3.4A). Based on the K_S-distributions of block paralogs, the K_S-dating tool can be employed to date one or more large-scale duplication events relative to a speciation event considering multiple species. Ancient and more recent WGDs can be identified in several plants species, although varying evolutionary rates in different lineages due to, for instance, different generation times, might interfere with the accurate dating of these events[14,46].

When investigating genomic homology between more than two genomes, the Skyline plot provides a

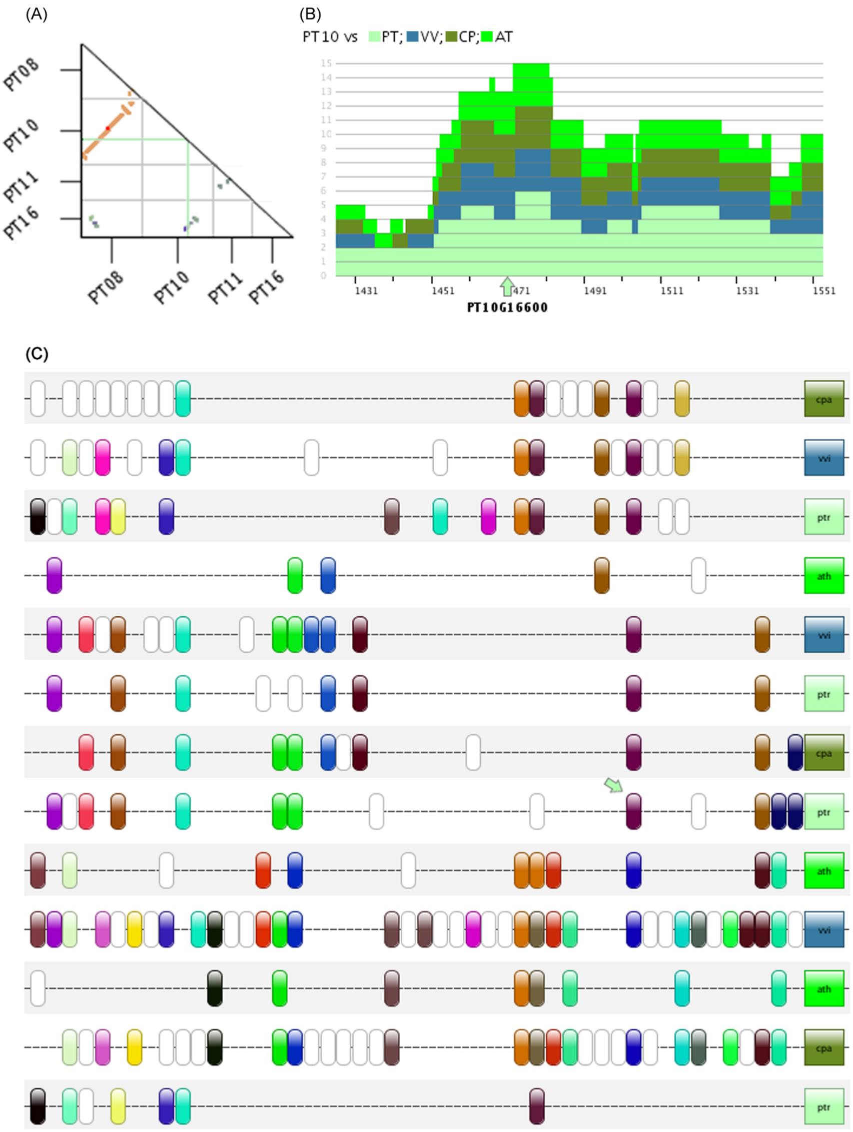


Figure 3.4: Overview of Different Collinearity-Based Visualizations of the Genomic Region around Poplar Gene PT10G16600. *(A) The WGDotplot shows that the gene of interest, indicated by the light-green line, is located in a duplicated block between chromosomes PT08 and PT10. The orange color refers to a K_S value of 0.2 to 0.3, indicating the most recent WGD in poplar. (B) The Skyline plot shows the number of collinear segments in different organisms detected using i-ADHoRe. (C) The Multiplicon view depicts the gene order alignment of the homologous segments indicated in (B). Whereas the rounded boxes represent the different genes color-coded according to the gene family they belong to, the square boxes at the right indicate the species the genomic segment was sampled from. The reference gene is indicated by the light-green arrow in (B) and (C).*

rapid and flexible way to browse multiple homologous genomic segments (Figure 3.4B). For a region centered around a reference gene, all collinear segments (from the selected set of organisms) are determined and visualized using color-coded stacked segments. The Skyline plot offers a comprehensive view of the number of regions that are collinear in the species selected (see Section 3.3.5). Navigation buttons allow the user to scroll left and right, whereas a window size parameter setting provides a zooming function to focus either on a small region around the reference gene or on the full chromosome. Clicking on one of the regions of interest shows a more detailed view (Multiplicon view; see Figure 3.4C). The gene alignment algorithm maintains the original gene order but will introduce gaps to place homologous genes in the same column (if possible).

3.2.5 Database Access, User Interface, and Documentation

An advanced query system has been developed to access the different data types and research tools and to quickly retrieve relevant information. Starting from a keyword search on gene descriptions, GO terms, InterPro domains, Reactome pathways, or a gene identifier, relevant genes and gene families can be fetched. Apart from the internal PLAZA gene identifiers, the original gene names provided by the data provider are supported as well. When multiple genes are returned using the search function, the *view-associated gene families* option makes it possible to link all matching genes to their corresponding gene families, reducing the complexity of the number of returned items. When searching for genes related to a specific biological process using GO, this function makes it possible to directly identify all relevant gene families and analyze the evolution of these genes in the different species. Although for some species the functional annotation is limited, even after GO projection, mapping genes related to a specific functional category to the corresponding families makes it possible to rapidly explore functional annotations in different species through gene homology.

To analyze multiple genes in batch, we have developed a Workbench where, for user-defined gene sets, different genome statistics can be calculated (Figure 3.1). Genes can be uploaded through a list of (internal or external) gene identifiers or based on a sequence similarity search. For example, this last option enables users to map an EST data set from a nonmodel organism to a reference genome annotation present in PLAZA. For gene sets saved by the user in the Workbench detailed information about functional annotation (InterPro and GO), associated gene families, block and tandem gene duplicates, and gene structure are provided. In addition, the GO enrichment tool allows for determination of whether a user-defined gene set is overrepresented for one or more GO terms (see the Workbench tutorial on the PLAZA documentation page). This feature makes it possible to rapidly explore functional biases present in, for example, differentially expressed genes or EST libraries.

The organization of a gene set of interest (e.g., gene family homologs, genes with a specific InterPro domain, GO term, or from a Reactome pathway, a Workbench gene set) in a genomewide context can reveal interesting information about genomic clustering. The Whole Genome Mapping tool can be used to display a selection of genes on the chromosomes (Figure 3.5), and additional information about the duplication type of these genes (i.e., tandem or block duplicate) is provided. Furthermore, the Whole Genome Mapping tool allows users to view the distribution of different gene types (protein-coding, RNA, pseudogene, or transposable element) per species.

An extensive set of documentation pages describes the sources of all primary gene annotations, the different methods and parameters used to build all comparative genomics data, and instructions on how to use

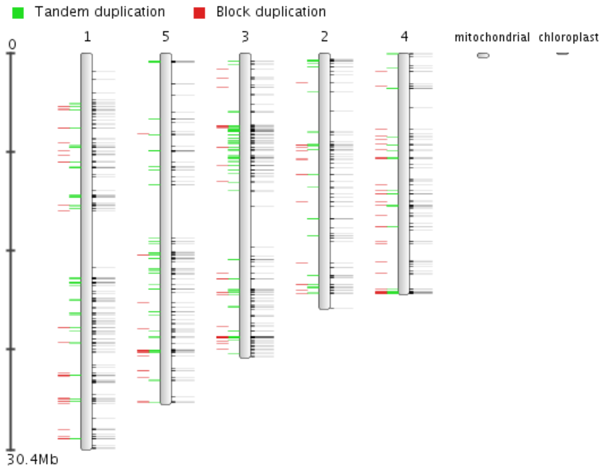


Figure 3.5: Whole Genome Mapping tool. *Overview of 664 Arabidopsis thaliana genes with a Cyclin-like F-box domain (IPR001810).*

the different tools. We also provide a set of tutorials introducing the different data types and interactive research tools. An extensive glossary has been compiled that interactively is shown on all pages when hovering over specific terms. Finally, for each data type (e.g., gene family and GO term) or analysis tool, all data can be downloaded as simple tab-delimited text files. Bulk downloads covering sequence or annotation data from one or more species are available through an FTP server.

3.3 Methods

3.3.1 Data Retrieval and Delineation of Gene Families

All gene annotation is retrieved from the different data providers (for details, see section Data content in PLAZA Documentation) and stored according to their gene type (coding, RNA, pseudo and TE). When parsing the structural gene annotation we verify if the original gene coordinates do generate the correct transcript and protein sequence (as reported by the primary data) and flag incorrect gene models. Starting from all protein-coding genes, only retaining the longest transcript if alternative splicing variants exist, protein sequences were used to construct homologous gene families by applying sequence based protein clustering. First, an all against all sequence comparison was performed using BLASTP applying an E-value threshold of 1e-05 and retaining the best 500 hits[138]. Note that applying less stringent E-value thresholds overall result in the inclusion of more outliers genes. Next, the complete sequence similarity graph was processed using Tribe-MCL (mclblastline, default parameters except I = 2 and scheme = 4) and OrthoMCL to identify gene families and sub-families, respectively. In post-processing, all genes assigned to a gene family but showing similarity (through BLASTP) to less than 25% of the median number of within-family similarity hits were annotated as outliers. The median number of within-family similarity hits is defined by first counting for each gene within a family the number of family members it shows similarity to and then determining the median number of hits per family. Manual verification of multiple sequence alignments in combination with similarity heat maps of all family members revealed that this threshold of 25% performs best to remove non-homologous false positive genes from the family. Only sub-families delineated by OrthoMCL are retained if they overlap for 95% or more with a single gene family and if two or more sub-families can be found for a given gene family defined by the Markov clustering. Thus, OrthoMCL clusters that are identical to Tribe-MCL clusters are discarded since they represent redundant information.

3.3.2 Comparison of OrthoMCL with Phylogenetic Trees

To verify the assumption that out-paralogs can correctly be identified using OrthoMCL, we validated a set 372 sub-families covering 129 large gene families using phylogenetic tree construction and reconciliation (Supplemental Table 2 accompanying Proost et al. [100]). Typically, phylogeny-based methods exhibit very low false positive rates (but also low coverage) because of the stringent criteria used to construct trees and provide a robust approach to evaluate the quality of the sub-families. Since these selected families contain multiple sub-families covering genes from all species in the dataset, they provide a good benchmark set to evaluate the accuracy of the sub-families defined by OrthoMCL. Tree reconciliation reveals that 92% (251/273) of the OrthoMCL sub-families are dated as originating in the ancestor of green plants, confirming that they represent ancient sub-types. Comparing the gene content between both methods shows that 70% (134/193) of all sub-families, for which a bootstrap supported ($\geq$70%) tree exists, are fully covered by the orthologous groups delineated using phylogenetics. This fraction increases to 76% (81/107) when considering only tree nodes with bootstrap values $\geq$99%. Similar results were obtained by Hanada and co-workers who found an overlap of 80% between similarity- and tree-based orthologous groups when clustering proteins from *Arabidopsis thaliana*, poplar, rice and moss[49].

An additional control experiment was performed to determine whether sub-families were formed by OrthoMCL that do not represent ancient sub-types. First, we assigned phylogenetic labels to the different sub-families (e.g. contains only genes from moss, algae, eudicots, monocots, all land plants or all plants). When studying the taxonomic range of the labels for the different sub-families within a family, we observed that only rarely false sub-families were defined. For example, when considering a set of 333 gene families having at least two sub-families, one annotated with 'monocot' and one with 'eudicot', respectively, only 16 cases (5%) were found where the family was erroneously split in a eudicot and monocot sub-family not representing out-paralogs.

3.3.3 Alignments and Phylogenetic Trees

For all gene families multiple sequence alignments were created using MUSCLE[140]. Alignment columns containing gaps were removed when a gap was present in >10% of the sequences. To reduce the chance of including misaligned amino acids, all positions in the alignment left or right from the gap were also removed until a column in the sequence alignment was found where the residues were conserved in all genes included in our analyses. This was determined as follows: for every pair of residues in the column, the BLOSUM62 value was retrieved. Next, the median value for all these values was calculated. If this median was $\geq$0, the column was considered as containing homologous amino acids. To prevent the emergence of low-branch attraction or badly-supported nodes yielding uninformative trees, highly divergent and partial sequences were removed from the alignment prior to phylogenetic tree construction if they contained in more than 50% of the alignment columns gaps or two times or more gaps than the average sequence in the alignment. Phylogenetic trees were constructed using PhyML applying the JTT substitution model, 100 bootstrap samples, estimated proportion of invariable sites, four substitution categories, estimated gamma distribution parameter, the BIONJ distance-based tree as starting tree and without tree optimization (default parameters for protein sequences). Notung 2.6[e] was used to root the trees and to infer speciation and duplication events using the tree reconciliation mode and applying the Duplication/Loss Score to evaluate alternate hypotheses. In the website JalView[f] is used as multiple sequence editor[151] to view and transfer sequence data to the user's PC. ATV/Archaeopteryx[g] is used for tree visualization[152].

3.3.4 Functional Annotation

Delineating correct othologous relations is a daunting task in plants due to many ancient and species-specific WGD creating many paralogous genes. A main issue for orthology projection is that an orthologous group covering for example genes from different land plants will include many paralogs that originated before/after the radiation of these species and that these duplicates might have diverged in function or regulation. Consequently, sub-or neo-functionalization of ancient duplicates makes transfer of functional annotation at the 'land plant' level heavily unreliable. Therefore, we selected eudicot and monocot orthologous groups to project functional annotation (Figure 3.3). The inherent drawback of this approach is that functional annotation from Arabidopsis cannot be transferred to rice and sorghum and vice versa. This limitation however will result in a smaller, but more reliable set of orthologous groups for projection. For the GO projection all primary gene annotations Inferred from Electronic Annotation (evidence code IEA) were excluded as information source (see Supplemental Table 5 accompanying Proost et al. [100]). Finally, all new gene-GO associations inferred through projection were labeled with evidence tag Inferred from Sequence or Structural Similarity (ISS).

The delineation of eudicot/monocot orthologous groups was done based on the phylogenetic trees. A recursive algorithm was developed which traverses the tree topology and checks each node based on its reconciled date and bootstrap value ($>$ 70). The consistency score (in case the node was labeled as a duplication node) was used to determine if the node was a genuine duplication (consistency score $>$ 0.30 for duplication). Note that the last criterion prevents the inclusion of ancient paralogous sub-types in the orthologous groups. Nodes that met this set of criteria were extracted as valid orthologous groups (18 513 and 13 216 groups for eudicots and monocots, respectively) and all GO terms from genes within such a group were collected. Redundancy caused by parent-child relations between related GO terms was removed and this extended set of labels was projected to all genes in the group recording the source gene(s) for newly inferred gene annotations. Consequently, some un-annotated genes received new functional annotations while other genes were re-annotated with a more specific GO term. Note that GO parent-child redundancy between primary and projected GO annotations was not removed in order to keep both data sources clearly distinguishable.

GO and family enrichment analysis was performed using the hypergeometric distribution and Bonferroni correction for multiple hypothesis testing.

3.3.5 Detection of Collinearity

To detect collinearity within and between species i-ADHoRe 2.4 was used[147]. Whereas the algorithm is identical to the i-ADHoRe 2.0 version, a more efficient way to store gene pairs in memory was im-

[e] http://www.cs.cmu.edu/ durand/Notung/

[f] http://www.jalview.org/

[g] http://www.phylosoft.org/archaeopteryx/

plemented allowing the program to be executed with up to 11 species on a machine with 2 gigabytes of RAM. Collinear regions can be used to study the conservation of genome organization between different species or to study duplicated blocks within one organism. Initially, all chromosomes from all species are compared against each other and significant collinear regions are identified. i-ADHoRe was run with the settings *alignment_method* gg, *gap_size* 30, *cluster_gap* 35, *q_value* 0.9, *prob_cutoff* 0.0001, *anchor_points* 4 and *level_2_only* false. The default run was done including all organisms. For optimal results however it is recommended to limit the number of species. Hence several other runs, with a subset of species, were done and stored in the database. Where relevant the website will allow users to pick to subset of species they're interested in (i.e. brassicales, eudicots, monocots, moss and algae).

Whereas the Multiplicon View and WGDotplot present raw i-ADHoRe output, the Skyline plot performs an additional processing step where several multiplicons are combined to show as many collinear regions as possible. For genes in the shown region all segments containing this gene are extracted and each of these segments belongs to a certain multiplicon which is accessible through the Multiplicon View by clicking the segment. For each selected organism the highest number of segments from this organism in one of these multiplicons will be determined and stored. This process is repeated for every gene in the reference region and the stored values will be used to build the graph depicted in the Skyline plot.

3.3.6 Relative Dating using Synonymous Substitutions

Only collinear gene pairs were dated using K_S. Compared to dating all pair-wise combinations of gene homologs per family, this has several advantages. First, as tandem duplications are filtered out when detecting collinearity, the L shaped curve caused by tandems isn't superimposed on K_S-plots obscuring peaks from large-scale duplications. Second, no correction for the number of K_S-measurements versus the number of real duplications has to be applied[153] and lastly, a reduction in the number of gene pairs to date results in a reduction of computational time. The coding sequences for the gene pairs were aligned with CLUSTALW (version 1.83)[154] using the protein sequences as alignment guides. From this alignment bad positions were stripped as described for the gene family alignments. The actual dating using synonymous substitutions was done using codeml (part of PAML package)[155] with the settings *verbose* 0, *noisy* 0, *runmode* -2, *seqtype* 1, *model* 0, *NSsites* 0, *icode* 0, *fix_alpha* 0, *fix_kappa* 0 and *RateAncestor* 0.

3.4 Summary and Future Prospects

The PLAZA platform integrates genome information from a wide range of species within the green plant lineage and allows users to extract biological knowledge about gene functions and genome organization. Besides the availability of different comparative genomics data types, a set of interactive research tools, together with detailed documentation pages and tutorials, are accessible through a user-friendly website. Sequence similarity is used to assign protein-coding genes to homologous gene families, and phylogenetic trees allow the reliable identification of paralogs and orthologs. Through the integration of high confidence GO annotations and tree-based orthology between related plant species, we could (re-)annotate thousands of genes in multiple eudicot and monocot plants. Apart from local synteny plots that facilitate the identification of positional orthologs, gene-based collinearity is calculated between all chromosomes from all species and can be browsed using the so-called Skyline plots. The WGDotplot visualizes all duplicated segments within one genome and dating based on synonymous substitutions generates an evolutionary framework to study large-scale duplication events. In addition, PLAZA's Workbench provides an easy access point to study user-defined gene sets or to process genes derived from high-throughput experiments. Based on a sequence similarity search or a list of gene identifiers, custom gene sets can rapidly be created and detailed information about functional annotations, associated gene families, genome-wide organization, or duplication events can be extracted. Consequently, this tool opens perspectives for researchers generating EST libraries from nonmodel species as these can easily be mapped onto a model organism. PLAZA hosts a diverse set of data types as well as an extensive set of tools to explore plant genome information.

Future efforts will be made to extend the number of available plant species and to include novel types of data to further explore gene function and regulation. Newly published plant genomes will be added on a regular basis to enlarge the evolutionary scope of PLAZA. The availability of genome information from more closely related organisms[156] will make it possible to explore the similarities and differences between species at the DNA level and to identify, for example, conserved cis-regulatory elements on a genome-wide scale.

In conclusion, PLAZA will be a useful toolkit to aid plant researchers in the exploration of genome information through a comprehensive web-based research environment.

3.5 Author Contribution

I was the only developer of the PLAZA webplatform, with the entire frontend programmed by me, with design help from other co-authors. As (shared) first author, I made substantial contributions to the manuscript, together with Sebastian Proost (first author) and Klaas Vandepoele (last author). Multiple visualizations used on the website and included in the manuscript were conceived and implemented by me. The entire workbench and its associated pipeline was also implemented by me. For a more technical overview of issues encountered during the development, see chapter 7.

“Science isn’t about why, it’s about why not. ”

Cave Johnson

4

Dissecting Plant Genomes with the PLAZA Comparative Genomics Platform

Abstract

With the arrival of low-cost, next-generation sequencing a multitude of new plant genomes is being publicly released, providing unseen opportunities and challenges for comparative genomics studies. Here, we present PLAZA 2.5, a user-friendly online research environment to explore genomic information from different plants. Compared to the previous published PLAZA version (PLAZA 1.0, see chapter 3), this new release features updates to previous genome annotations and a substantial number of newly available plant genomes, as well as various new interactive tools and visualizations. These additions mimic the growth in sequencing performance as seen in the period 2009 – 2012. A more detailed analysis of the differences in data content can be found in section 4.2.1.

Currently, PLAZA hosts 25 organisms covering a broad taxonomic range, including 13 eudicots, five monocots, one Lycopod, one moss, and five algae. The available data consist of structural and functional gene annotations, homologous gene families, multiple sequence alignments, phylogenetic trees, and colinear regions within and between species. A new Integrative Orthology Viewer, combining information from different orthology prediction methodologies, was developed to efficiently investigate complex orthology relationships. Cross-species expression analysis revealed that the integration of complementary data types extended the scope of complex orthology relationships, especially between more distantly related species. Finally, based on phylogenetic profiling, we propose a set of core gene families within the green plant lineage that will be instrumental to assess the gene space of draft or newly sequenced plant genomes during the assembly or annotation phase.

This chapter is based on Van Bel et al.[84]. For the author contributions, see page 4-18.

4.1 Introduction

Thanks to recent advances in sequencing technologies[157], the price per base pair has dropped sharply[158]. Therefore, genome sequencing is no longer restricted to model organisms and a variety of species of ecological, agricultural and/or economical importance are sequenced by several laboratories around the world[128,159,160]. Recently, re-sequencing additional genomes of a reference species has become feasible as well[161], improving the understanding of genomic variation. Whereas a single genome provides a basic catalog of all genes it encodes, comparison of genomes gives insights into the evolution and adaptation of species to specific environments[162]. However, comparative genomics studies come at an extra cost: as the number of available genomes increases, large-scale analyses become increasingly difficult for non-experts, whereas the computational requirements to extract biological information grow rapidly. Furthermore, biological variation between species and differences in sequence quality enhance the complexity of evolutionary analyses. Therefore, platforms for comparative genomics[100,104,111,163], that take care of some of these challenges, are valuable resources for experimental biologists.

A key challenge in comparative genomics is the reliable grouping of homologous genes (derived from a common ancestor) and orthologous genes (homologs separated by a speciation event) into gene families[33,64,164,165]. Orthology is generally considered a good proxy to identify genes performing a similar function in different species[60]. Consequently, orthologs are frequently used as a mean to transfer functional information from well-studied model systems, such as *Arabidopsis thaliana* or *Oryza sativa* (rice), to non-model organisms. In plants, utilization of orthology is not trivial, due to a wealth of paralogs (homologous genes created through a duplication event) in almost all plant lineages. Ancient duplication events preceding speciation lead to outparalogs, which are frequently considered as subtypes within large gene families. In contrast to this are inparalogs, genes that originated through duplication events occurring after a speciation event[33]. Besides continuous duplication events (for instance via tandem duplication), many plant paralogs are remnants of whole genome duplications (WGDs). In flowering plants, the frequent WGDs in several lineages[14] result in the establishment of one-to-many and many-to-many orthologs (or co-orthologs). Other modes of duplication, such as retro-transposition, also introduce co-orthologous relationships, but the duplicated copy ends up in a different genomic context and is probably regulated differently due to the absence of its original promoter. As such, transfer of functional information between organisms is a non-trivial operation[166]. Various algorithms for orthology detection have been developed and benchmarked[167], and, overall, can be catalogued as graph-based and tree-based methods, with the latter closer to the original orthology definition[33], because they are based on the reconciliation of a family tree with a species tree.

PLAZA, an online resource for plant genomics, had been developed to integrate and distribute comparative genomics data for both computational and experimental plant biologists[100]. The first release, based on nine sequenced plant genomes, included various tools to easily retrieve specific data types, such as gene families, multiple sequence alignments, phylogenetic trees, and genomic homology. To accommodate the evolutionary analysis of an increasing number of available plant genomes, more powerful and streamlined computational pipelines were required as well as new tools to visualize genome information from multiple species. Here, we present version 2.5 of PLAZA, a major update of the comparative genomics platform, which currently hosts twenty-five species together with a variety of new tools to browse gene families, study functional clustering, and explore multispecies colinearity data. In addition to the development of a new tool to identify complex gene orthology relationships, different prediction methods were also evaluated by means of expression context conservation.

4.2 Results and Discussion

4.2.1 Gene Annotation and Gene Families

Parsing the 25 genomes present in PLAZA 2.5 resulted in 909,850 genes, covering 85.8% protein coding genes, 13.7% transposable elements, 0.3% RNA genes and 0.1% pseudogenes (Table 4.1). Besides nuclear gene annotations, chloroplast and/or mitochondrial gene information was included was well, when available. In total, 13 eudicots, five monocots (Liliopsida), one Lycopod, one moss, and five algae were integrated, of which 16 are new species compared to the previous release. The functional annotation pipeline resulted in 462,958 (419,028 without GO projection) genes with at least one associated Gene Ontology (GO) term, and 519,047 protein coding genes with at least one InterPro domain (Table 4.1). Overall, projected functional information inferred through sequence orthology[100] covered 10% of the available gene-GO annotations (43,930 genes from 18 different species have only GO annotations based on projection).

Protein clustering based on all-against-all sequence similarity searches resulted in 32,294 gene families, covering 87.8% of all the protein coding genes, and 22,350 multispecies gene families, covering 82.6% of all protein coding genes, with a gene family defined as a cluster of two or more homologous genes. This coverage represents a considerable increase compared to PLAZA 1.0, in which only 77.6% and 68.1% of the coding genes where assigned to gene families and multispecies gene families, respectively. Due to a variety of problems (changed annotations, split/merged gene families, ...) the gene family identifiers were not kept stable between PLAZA versions. Multispecies gene families are commonly applied for improving, through homology, the structural annotation of gene models[168]. The increase in gene number assigned to both classes of gene families demonstrates the importance of sequencing additional species to obtain a better gene coverage within specific phylogenetic clades. Only a relatively small fraction of gene families contains proteins from all species (see Figure F.2).

Reliably transfer of known functional descriptions from the gene level to the gene family level was achieved by calculating GO enrichment statistics for each family (see Methods). Through the website, this functional information, together with protein domain information, is displayed per family. Although this family GO annotation procedure yielded information for only 8,606 gene families and 28,281 subfamilies, they cover more than 70% of the protein-coding genes present in gene families.

4.2.2 Core Plant Gene Families and Detection of Clade-specific or Expanded Gene Families

Most new genome sequences generated by next-generation sequencing methods do not provide the full genomic sequence[185], but rather aim at providing sequences containing the majority of the proteome, potentially missing noncoding genes or intergenic regions. The extremely large genome sizes associated with some organisms prevent full-genome sequencing and enforce the application of transcriptome sequencing to build gene catalogs[186]. A key challenge in comparative gene family analysis is discerning whether the absence of a species within a gene family is functionally and evolutionarily relevant or rather an artifact from the assembly and/or annotation procedures. As a consequence, the reliable assessment of the gene space provides an important measure to determine the quality of genome sequencing and

Species	Genes (a)	Scaffolds (b)	GO (c)	InterPro (d)	Version	Reference
Arabidopsis lyrata	32,670 (100%)	8+1 (429)	53.8% (65.6%)	72.1%	JGI 1.0*	Hu et al. [169]
Arabidopsis thaliana	33,602 (81.6%)	$5^{C,M}$	77.3% (80.2%)	78.3%	TAIR10	Initiative [13]
Brachypodium distachyon	26,678 (99.8%)	$5+1\ (15)^{C}$	56.6% (66.9%)	78.2%	MIPS 1.2*	Initiative [170]
Carica papaya	28,072 (99.8%)	4635^{C}	43.4% (49.6%)	58%	Hawaii ARC	Ming et al. [171]
Chlamydomonas reinhardtii	16,841 (99.7%)	$88^{C,M}$	50.7% (50.7%)	53.4%	JGI 4.0	Merchant et al. [130]
Fragaria vesca	34,809 (100%)	7+1 (1080)	43.5% (49%)	61.4%	Strawberry Genome 1.0*	Shulaev et al. [172]
Glycine max	46,509 (99.9%)	$20+1\ (97)^{C}$	61.3% (70.2%)	82.9%	JGI 1.0*	Schmutz et al. [173]
Lotus japonicus	69,647 (61.9%)	$6+1\ (22048)^{C}$	42.2% (45.8%)	57.3%	Kazusa 1.0*	Sato et al. [159]
Malus domestica	95,230 (66.7%)	17+1 (23653)	61.8% (66.4%)	69.3%	IASMA*	Velasco et al. [160]
Manihot esculenta	30,800 (99.8%)	3142^{C}	57.6% (66.5%)	78.8%	Cassava4*	Not published
Medicago truncatula	57,587 (78.5%)	$8+1\ (145)^{C}$	35.4% (39.5%)	48.7%	Mt3.5*	Young et al. [174]
Micromonas sp. RCC299	10,276 (99.3%)	$17^{C,M}$	58.3% (58.3%)	69.8%	JGI 3.0*	Worden et al. [175]
Oryza sativa ssp. indica	59,430 (82.8%)	$12+1\ (2217)^{C,M}$	44.1% (53.9%)	59.6%	9311_BGF_2005*	Yu et al. [176]
Oryza sativa ssp. japonica	57,874 (72.9%)	$12^{C,M}$	55.2% (58.6%)	58.6%	MSU RGAP 6.1	Ouyang et al. [177]
Ostreococcus lucimarinus	7,805 (100%)	21	60.7% (60.7%)	74.4%	JGI 2.0	Palenik et al. [178]
Ostreococcus tauri	8,116 (98.5%)	$20^{C,M}$	49.6% (49.6%)	63.7%	Ghent University*	Derelle et al. [179]
Physcomitrella patens	36,137 (77.8%)	$1121^{C,M}$	47.8% (47.8%)	57.9%	JGI 1.1,cosmoss.org 1.2	Rensing et al. [129]
Populus trichocarpa	41,521 (99.9%)	$19+1\ (957)^{C}$	54.6% (61.8%)	73.7%	JGI 2.0	Tuskan et al. [21]
Ricinus communis	31,221 (100%)	4962	48.3% (54.1%)	65%	JCVI 1.0*	Chan et al. [180]
Selaginella moellendorffii	22,285 (100%)	361	55.7% (55.7%)	71.8%	JGI 1.0*	Banks et al. [181]
Sorghum bicolor	34,686 (99.8%)	$10+1\ (207)^{C,M}$	54.8% (62.1%)	71.1%	JGI 1.4	Paterson et al. [182]
Theobroma cacao	46,269 (62.4%)	11^{C}	50.7% (57.7%)	69.4%	CocoaGen v1.0*	Argout et al. [183]
Vitis vinifera	26,644 (99.5%)	$19+1\ (14)^{C,M}$	72.6% (76.4%)	71.8%	Genoscope v1	Jaillon et al. [128]
Volvox carteri	15,544 (100%)	762	39.1% (39.1%)	54.1%	JGI 1.0*	Prochnik et al. [184]
Zea mays	39,597 (99%)	$11^{C,M}$	48.1% (55.9%)	65.6%	Version 5.60*	Schnable et al. [95]

Table 4.1: Data content PLAZA v2.5. *(a) Numbers in parentheses refer to the fraction of protein coding genes. (b) Numbers in parentheses refer to the number of genomic sequences in the original annotation (assembly) containing genes. The* +1 *tag indicates the creation of a virtual chromosome zero to group scaffolds together whereas* C *and* M *indicate the inclusion of chloroplast and mitochondrial genomes, respectively. (c) Percentage of coding genes with an associated GO term. The fraction after the GO projection is displayed between the parentheses. (d) Percentage of coding genes with at an associated InterPro domain. (*) New species compared to PLAZA 1.0*

annotation projects.

Based on families conserved in a specific set of species, core gene families were created by means of PLAZA 2.5. Families were selected on the basis of their gene content in phylogenetic subclades from the PLAZA species tree, tolerating missing homologs in a small subset of species. Three sets of core gene families were built based on the subclades rosids, monocots, and green plants. This phylogenetic approach resulted in 6,316, 7,076 and 2,928 core gene families for the rosids, monocots, and green plants, respectively. As expected, the core gene families cover, among others, housekeeping genes and genes involved in primary metabolism. For each gene family a representative gene was selected from the rosids and monocots (with a preference for genes from either Arabidopsis and rice, respectively) that could be used as a probe to quantify genome completeness. Assessment of the gene space of each species included in the platform using the weighted core gene family scores revealed the relatively low gene coverage for some species (Figure 4.1). Especially Lotus japonica and Medicago truncatula within the eudicot species, and Selaginella moellendorffii within the primitive land plants, showed a high number of potentially missing genes. We propose these lists of core gene families as a reference set to quantify the gene space in future genome projects.

Whereas core gene families are a useful tool for asserting proteome completeness, the study of species- (or lineage-) specific (expanded) gene families is equally important to understand how species can adapt to particular niches. Tandem gene duplications are a known mechanism used by plants to rapidly increase the expression rate of a gene[49], instead of the transcription rate. Two new tools were implemented to facilitate the detection of gene families based on phylogenetic profiles (presence or absence of a gene family in a species) or expansion statistics. Whereas the Gene Family Finder tool enables the identification of

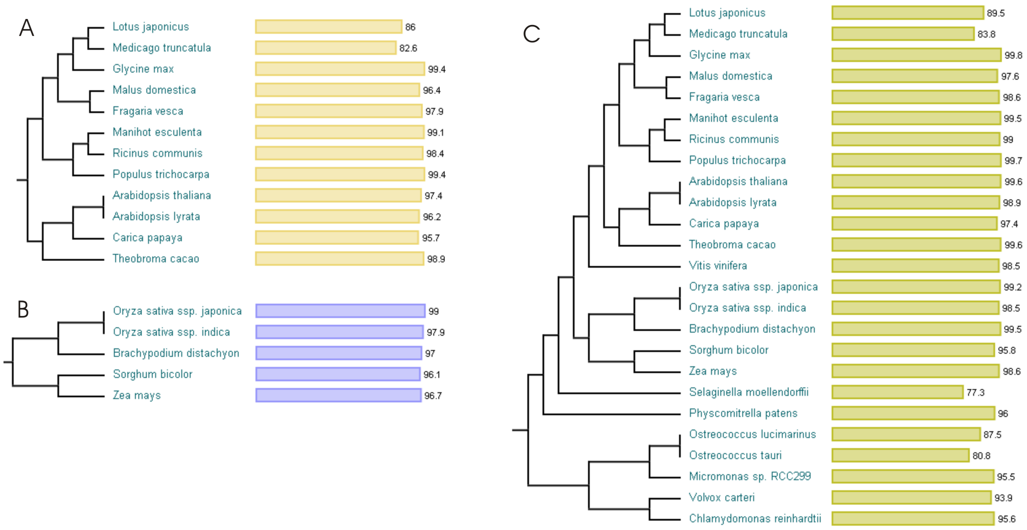


Figure 4.1: Core gene family coverage in all PLAZA organisms. *Core gene family coverage in all PLAZA organisms, using the 6,316 rosid (A), 7,076 monocot (B), and 2,928 green plant core gene families (C). Coverage is expressed as percentage of the core gene families having the indicated organism.*

(expanded) gene families specific to one or more species, the Gene Family Expansion Plot displays gene family expansions patterns between two (sets of) organisms (Figure 4.2).

4.2.3 Integrative Orthology Viewer: an Ensemble Approach to Detect Orthology Relationships

Several methods for finding orthologs between two or more species have been described, each with its own strengths and weaknesses[164]. Whereas Reciprocal best BLAST-Hit (RBH) detection[187] between closely related species provides a practical solution to identify orthologs, it cannot deal with complex one-to-many or many-to-many orthologous relationships between more distantly related species. Although the construction of phylogenetic trees[188,189] should offer the highest confidence to identify speciation events in gene family trees, it has a relatively low gene coverage compared to sequence-based clustering methods, as trees could not be generated for all gene families. In PLAZA 2.5, 46,651 phylogenetic trees were constructed covering 81% of all protein-coding genes assigned to gene families. Besides heavy computational requirements, the method is also hampered by its sensitivity to differences in the topology of the gene tree compared to the species tree, which are used for reconciliation[142].

To detect orthologous gene relationships in plants with an enhanced robustness, an integrative approach was developed to identify orthologs on a gene-by-gene basis. The developed ensemble approach consists of four distinct orthology prediction methods: orthologous gene families inferred through sequence-based clustering with OrthoMCL[61] (including modeling of RBH orthology and inparalogy), reconciled phylogenetic trees, colinearity information and multispecies Best-Hits-and-Inparalogs (BHI) families. The latter are based on the best BLAST hit for each species, extended with the inparalogous genes in each species[190]. The integration of gene colinearity facilitates the detection of positional orthologs, namely genes with conserved genome organization between species. The combination of different methods for

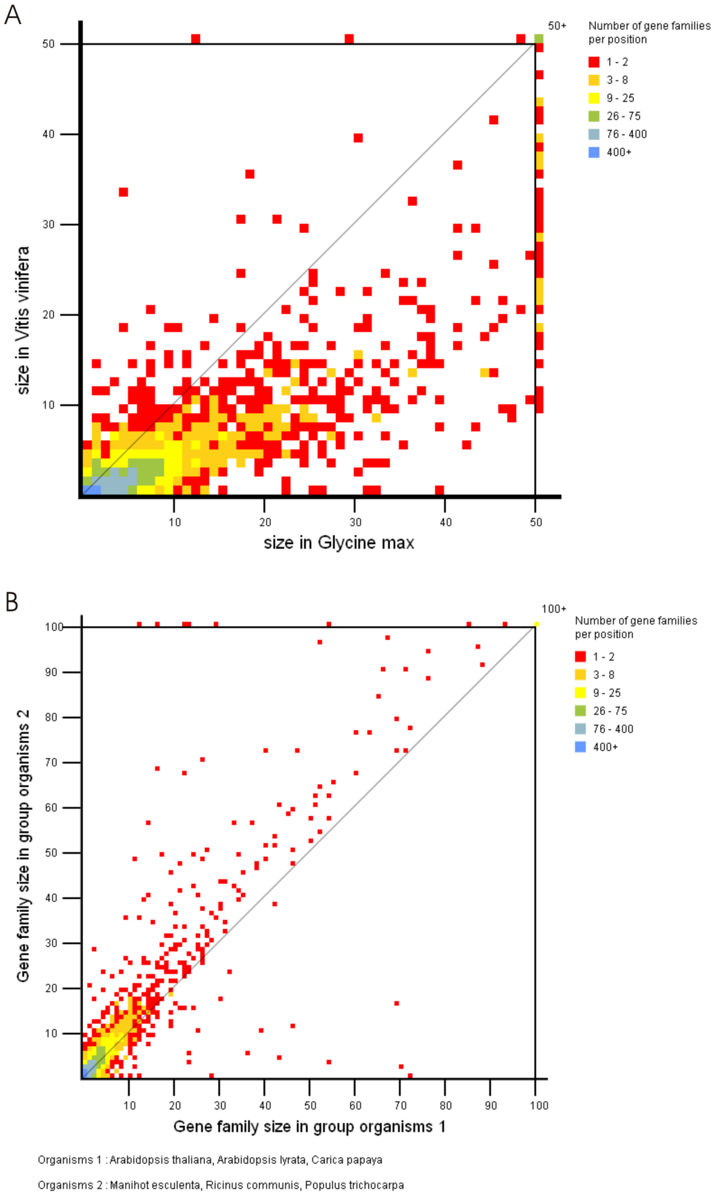
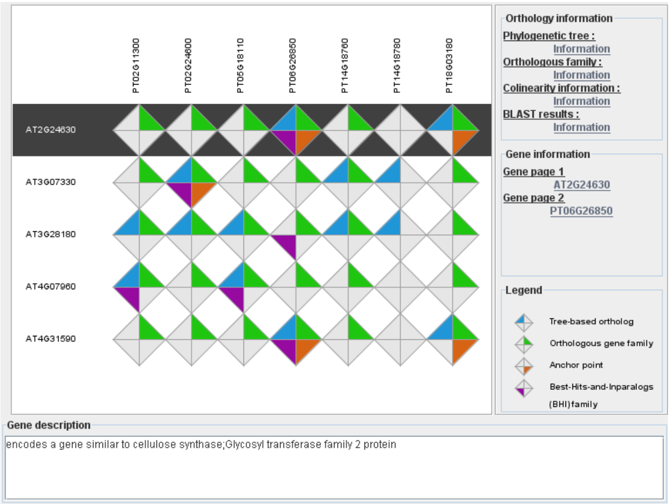


Figure 4.2: Gene family expansion plot. *(A) The gene copy number in Vitis vinifera and Glycine max, within each gene family, is indicated by the position of a dot, and the color indicates the number of gene families with these gene copy numbers. (B) Density plot between two sets of organisms, Brassicales (Arabidopsis thaliana, Arabidopsis lyrata, and Carica papaya) versus Malpighiales (Manihot esculenta, Ricinus communis,and Populus trichocarpa).*



encodes a gene similar to cellulose synthase;Glycosyl transferase family 2 protein

Figure 4.3: Integrative Orthology Viewer. *Orthology overview for the Arabidopsis thaliana gene AT2G24630, its paralogs and orthologs in Populus trichorpa. The selected query gene is marked with a black border.*

orthology detection, as implemented in the PLAZA platform, allows for the more accurate selection of orthologs, for example using majority voting[191] or through the application of a weighted voting scheme based on the sensitivity of individual tools. Other plant comparative genomics database like GreenPhylDB[111] and Phytozome[103] only group homologous genes into families using clustering, the latter also including synteny information to identify putative positional orthology. PlantEnsembl[104] performs orthology and paralogy predictions solely based on reconciled gene family phylogenetic trees.

The Integrative Orthology Viewer displays for a query gene and its predicted inparalogs the associated orthologs, including the support from the different orthology methods (Figure 4.3). In addition, all links are provided to explore the supporting evidence and specific details of the individual predictions. For instance, the phylogenetic trees that served as the primary data source for the tree-based orthologs can be viewed and the user can evaluate the support of a specific speciation node.

To compare the performance of individual methods, as well as of an integrative approach, we first generated basic statistics about the number of inferred orthology relationships. With focus on the model species Arabidopsis as query species and any other species as target, the gene coverage was highest for the BHI families and OrthoMCL (25,862 and 23,932 genes with at least one ortholog, respectively)(see also Figure F.4). As expected, reconciled phylogenetic trees only provided orthology information for 18,415 Arabidopsis genes. To evaluate the quality of these predictions, the percentage of orthologous gene pairs with conserved expression was determined by using the Expression Context Conservation (ECC)[192]. The expression context was based on the expression similarity between a query gene and all other genes in that species (gene-centric coexpression cluster). The ECC was obtained by starting from a predicted orthologous gene pair, retrieving all coexpressed genes per species, and calculating how many homologs were coexpressed in both species. Significant ECC values indicate that the orthologous genes show share coexpression with several other genes in both species. Consequently, conserved ECC gene pairs can be used as a proxy to measure conserved gene functions between putative orthologs, based on spatial-temporal expression information.

Based on a random sample of 9,319 orthologous Arabidopsis - rice gene pairs, ECC scores for the different orthology prediction methods indicated that gene pairs uniquely predicted by individual methods overall contain less gene pairs with conserved coexpression compared to predictions supported by multiple tools: 44%, 41% and 41% for OrthoMCL, BHI families and trees, respectively, versus 60% (supported by OrthoMCL and BHI families), 41% (supported by BHI families and trees), 57% (supported by OrthoMCL and trees) and 68% (supported by OrthoMCL, BHI families and trees). Although these results indicated that multiple evidences increase the reliability of orthology prediction, application of a majority voting system (i.e. only selecting orthologs with the highest number of evidences) could miss true orthologs with less support types (i.e. false negatives). To compare the performance of a majority-voting protocol with a selection procedure only requiring two support types, we evaluated ECC scores for orthologous gene pairs supported by only two evidences with those confirmed by three prediction methods. Despite majority voting orthologs having a higher fraction of ECC conserved genes (66%), 51% of the gene pairs with only two evidences also showed conserved expression between Arabidopsis and rice (based on the same reference query Arabidopsis gene set). Therefore, we retained all orthologous predictions supported by two or more evidences in the integrative orthology method.

Although OrthoMCL has been shown to have a good tradeoff between false positives and false negatives[133], we observed that 3,506 Arabidopsis genes (13% of the proteome) had a predicted orthologous rice gene based on the Integrative method, whereas no ortholog was found using OrthoMCL. Of the 3,506 Arabidopsis genes having one or more rice ortholog(s) (covering 3,874 rice genes in total), 40% exhibited conserved expression conservation. This result indicates that a considerable fraction of gene pairs not reported by OrthoMCL represents conservatively coexpressed orthologs, revealing the complementary nature of both approaches. Application of the integrative method (requiring at least two support types) to predict orthologs from Arabidopsis in other species, revealed overall 30% more predictions compared to OrthoMCL (Figure 4.4). Although the difference in the number of one-to-one orthologs is minor for most species, the number of complex orthology relationships (one-to-many and many-to-many) is higher for the integrative method. The frequent occurrence of WGD is an important driver responsible for the high frequency of complex orthology gene relationships in plant genomes.

4.2.4 Clusters of Functionally Related Genes in Eukaryotic Genomes

Whereas in many prokaryotic genomes genes are organized in operons, this is relatively rare in eukaryotes[193]. However, the overall absence of polycistronic mRNAs in eukaryotic genomes does not imply a random gene organization within chromosomes[193–195]. In several eukaryotic species clusters, with genes sharing similar expression patterns, members of the same pathway or genes with related functions, have been described, indicating that the null-hypothesis of random gene order is incorrect[196]. Recent studies have suggested that the chromatin state, either euchromatin or heterochromatin, is one of the contributing factors to the coexpression of neighboring genes[194,196], and bidirectional promoters as well[197].

To study the clustering of functionally related genes, the C-Hunter program[198] was used for a genome-wide analysis. This tool detects statistically significant clusters of neighboring genes based on the similarity of GO annotations. The standard C-Hunter run (no tandem gene removal, minimum genes 2, maximum genes 30) resulted in 5408 significant clusters covering 34,407 genes from 25 different species. As the majority of these clusters (68%) are composed uniquely of tandemly duplicated genes, an extra data set was created to detect clustering of nonhomologous genes[194]. In this data set every set of tan-

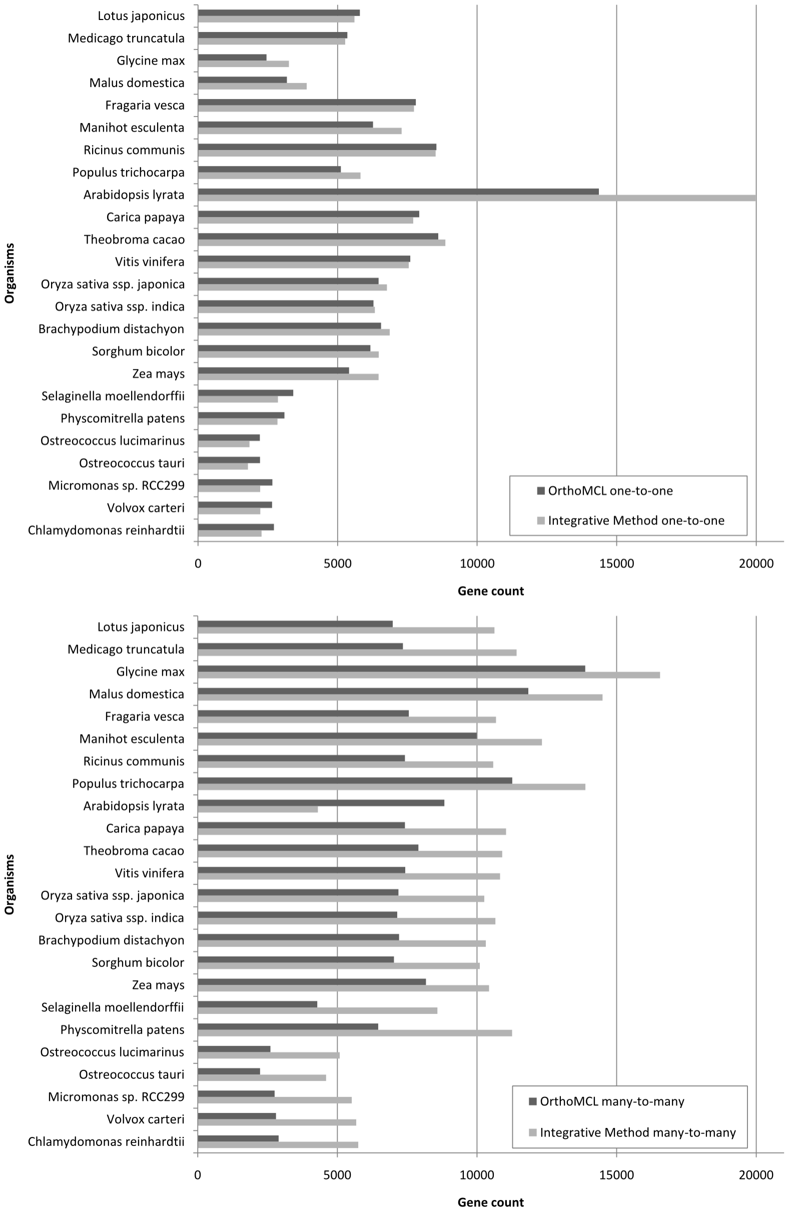


Figure 4.4: Quantification of Arabidopsis orthologs. *Summary of the different orthologous relationships predicted by the PLAZA integrative method and OrthoMCL, between Arabidopsis thaliana and all other PLAZA 2.5 species, respectively. The integrative method requires at least two support types to retain orthologous genes. Species are ordered per phylogenetic clade. The top panel displays results for one-to-one orthologs and the bottom panel shows many-to-many orthologs.*

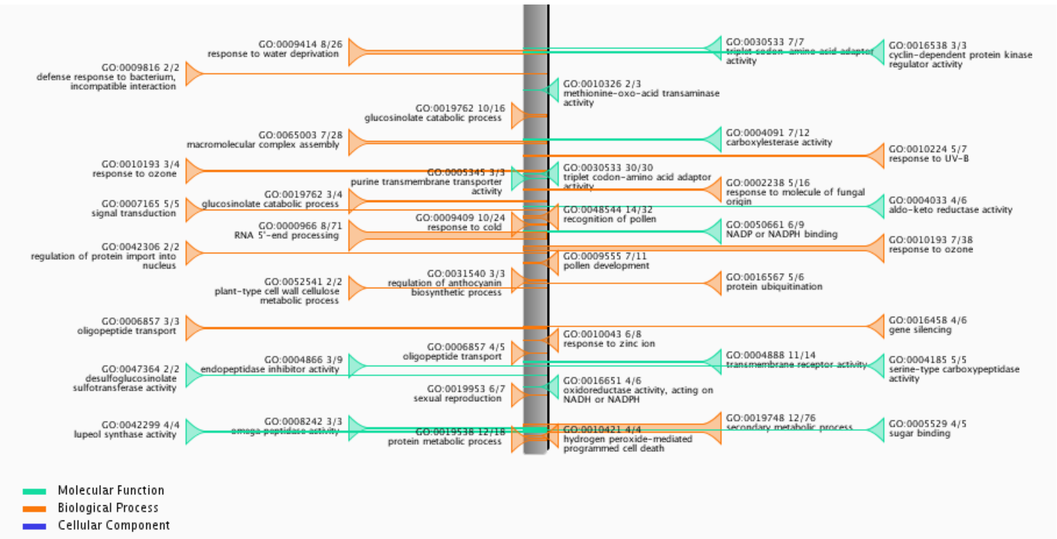

GO:0009414 8/26 response to water deprivation
GO:0009816 2/2 defense response to bacterium, incompatible interaction
GO:0019762 10/16 glucosinolate catabolic process
GO:0065003 7/28 macromolecular complex assembly
GO:0010193 3/4 response to ozone
GO:0005345 3/3 purine transmembrane transporter activity
GO:0019762 3/4 glucosinolate catabolic process
GO:0007165 5/5 signal transduction
GO:0009409 10/24 response to cold
GO:0000966 8/71 RNA 5'-end processing
GO:0042306 2/2 regulation of protein import into nucleus
GO:0031540 3/3 regulation of anthocyanin biosynthetic process
GO:0052541 2/2 plant-type cell wall cellulose metabolic process
GO:0006857 3/3 oligopeptide transport
GO:0006857 4/5 oligopeptide transport
GO:0004866 3/9 endopeptidase inhibitor activity
GO:0047364 2/2 desulfoglucosinolate sulfotransferase activity
GO:0019953 6/7 sexual reproduction
GO:0008242 3/3 omega peptidase activity
GO:0042299 4/4 lupeol synthase activity
GO:0019538 12/18 protein metabolic process
GO:0030533 7/7 triplet codon-amino acid adaptor activity
GO:0016538 3/3 cyclin-dependent protein kinase regulator activity
GO:0010326 2/3 methionine-oxo-acid transaminase activity
GO:0004091 7/12 carboxylesterase activity
GO:0010224 5/7 response to UV-B
GO:0030533 30/30 triplet codon-amino acid adaptor activity
GO:0002238 5/16 response to molecule of fungal origin
GO:0004033 4/6 aldo-keto reductase activity
GO:0048544 14/32 recognition of pollen
GO:0050661 6/9 NADP or NADPH binding
GO:0010193 7/38 response to ozone
GO:0009555 7/11 pollen development
GO:0016567 5/6 protein ubiquitination
GO:0016458 4/6 gene silencing
GO:0010043 6/8 response to zinc ion
GO:0004888 11/14 transmembrane receptor activity
GO:0004185 5/5 serine-type carboxypeptidase activity
GO:0016651 4/6 oxidoreductase activity, acting on NADH or NADPH
GO:0019748 12/76 secondary metabolic process
GO:0005529 4/5 sugar binding
GO:0010421 4/4 hydrogen peroxide-mediated programmed cell death
Molecular Function
Biological Process
Cellular Component

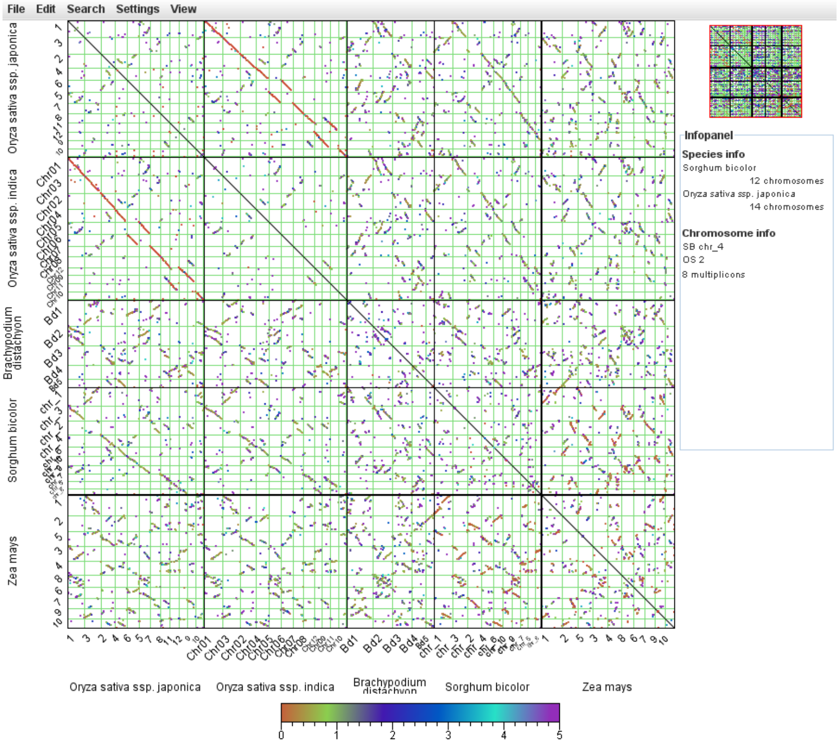
Figure 4.5: Functional clusters based on GO annotation. *Functional clusters in Arabidopsis thaliana chromosome 1, detected with the C-Hunter software package. Data in the text fields include GO term and description, cluster size, and the number of genes within a cluster with a specific GO term.*

dem genes was represented by a single gene representative (see Methods). The number of clusters varied widely among the different species, suggesting that both the quality and quantity of the structural and GO annotations of genes played a major role, as well as the assembly of scaffolds in the chromosomes. More compact genomes, such as those of the algal species, had a smaller number of functional clusters, whereas the number of detected functional clusters in larger genomes correlated with the number of genes per scaffold and the number of genes with a GO term. The resulting clusters are included in the database and can be browsed from both gene pages and GO pages on the PLAZA website. Furthermore, a visualization presenting the significant functional clusters per chromosome (Figure 4.5) was created with a GO domain based coloring. Using these functional clusters in combination with colinearity information can also give clues to the origin of these clusters, as the most parsimonious explanation for shared functional clusters would be that the common ancestor already contained these clusters (see Figure F.3).

4.2.5 Colinearity-based Genome Analysis

As a means to study genome organization and evolution, i-ADHoRe[48] is used to discover genomic homology based on gene colinearity within and among species. Colinearity information can be applied to analyze segmental and WGD events, whereas cross-species genome conservation facilitates the analysis of chromosomal rearrangements, such as inversions, chromosomal fissions/fusions, and translocations. As the increase in number of species resulted in more complex genomic homology relationships, two new tools, the WGDotplot applet (Figure 4.6) and the CirclePlot (Figure 4.7), were developed which provide more advanced and configurable visualizations. For both tools the dating of colinear regions, based on the fraction of synonymous substitutions over all synonymous sites (Ks), is visualized by color coding.

The WGDotplot applet was implemented in Java and designed to be an interactive extension of the static



Infopanel

Species info
Sorghum bicolor
12 chromosomes
Oryza sativa ssp. japonica
14 chromosomes

Chromosome info
SB chr_4
OS 2
8 multiplicons

Figure 4.6: WGDotplot applet. *Visualization of the colinearity between five monocot species: Oryza sativa ssp. japonica, Oryza sativa ssp. indica, Brachypodium distachyon, Sorghum bicolor and Zea mays. Green lines indicate different chromosomes, and the colors of the colinear regions reflect the Ks values.*

visualizations present in PLAZA 1.0, also allowing the visualization of colinear regions between more than two species. At the same time, the functionalities were extended to encompass a rich palette of visualization options, such as hiding chromosomes and rearranging chromosomal positions, adapting color usage, and using stepless zoom features to browse the genomic colinearity between multiple species.

The Circle Plot tool was developed as a lightweight and interactive circular visualization tool, similar to the popular Circos software[94]. Fully written in Javascript, this program runs natively on most modern browsers. Whereas the primary use of the Circle Plot is the study of intraspecies colinear regions, the ability to map interspecies colinear regions on the circumference of the plot closes the gap between the capabilities of the Circle Plot and the WGDotplot applet. Extra features, such as coding gene density and InterPro and GO terms, can also be displayed on the circumference of the Circle Plot. Another main difference is the mode of chromosome size determination. Whereas the Circle Plot uses nucleotide-based coordinates, the WGDotplot uses genes as the smallest units (retaining protein-coding genes only). Consequently, the former can display information in low-coding regions, such as centromeres or telomeres, and the latter facilitates the comparison of colinear regions from species with differences in gene density.

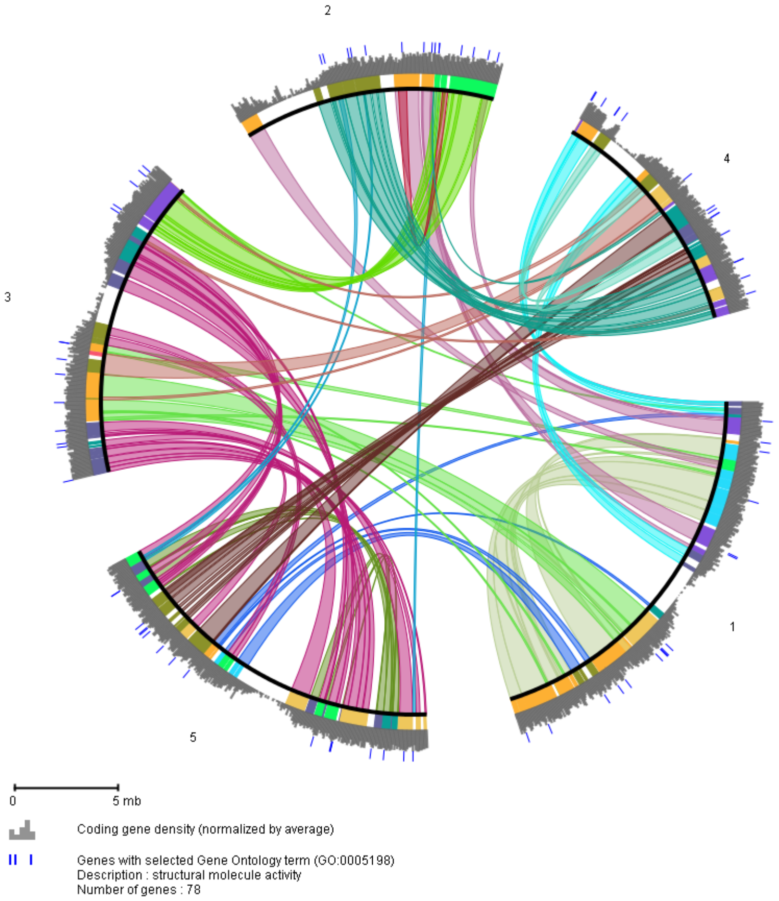


Figure 4.7: Circle Plot. *Plot showing colinear regions within Arabidopsis thaliana (inner circle) and between Arabidopsis thaliana and Arabidopsis lyrata (colored border of circle, indicating different Arabidopsis lyrata chromosomes). Also displayed are the coding gene density (grey blocks on the border of the circle) and a selected GO term (GO:0005198) (blue stripes on border of the circle). Coloring of colinear regions within Arabidopsis thaliana is based on start/end chromosomes, and only those colinear regions (both intra- and inter-species) with a Ks-value between 0.3 and 2 (corresponding with 3R duplication event) are shown.*

4.2.6 User Interactivity via Workbench and Bulk Downloads

PLAZA offers a versatile resource for easy data mining of homologous genes, sequence alignments and phylogenetic trees, genome organization, and functional annotation in different plants. However, large-scale analyses with a web-based user interface quickly become tedious and time consuming. To overcome this problem, a user-oriented Workbench was implemented in which specific gene sets can be analyzed. Different collections of user-provided gene lists are stored as separate experiments and genes can be added to an experiment based on internal/external gene identifiers or sequence similarity searches (see Tutorial on the PLAZA website), providing a versatile approach to map genes across species or to summarize sequencing data on a reference genome annotation. Whereas the initial workbench contained tools to explore the functional annotation of sets of genes, in PLAZA 2.5 multiple improvements were made for an easier and more comprehensive user experience. The GO enrichment tool is extended, bulk detection of orthologs on a gene-by-gene basis is possible and multiple workbench experiments can be compared. In addition, based on the outcome of a first analysis (such as gene filters in an experiment with GO), a new workbench experiment can easily be created or, reversely, genes can be removed from the initial experiment. The export function allows the user to retrieve general gene information (functional annotation, gene family data, orthologs, duplication data) as well as various sequence features (e.g. coding sequence, intron, and upstream and downstream sequences) for large gene sets covering all 25 genomes. Overall, the workbench offers a user-friendly solution for the efficient analysis of multiple data sets containing hundreds of genes. In addition, bulk downloads of most data sets in PLAZA are available through the FTP-site.

In conclusion, the PLAZA platform is a user-friendly platform for small- and large-scale comparative sequence analyses of plant genomes. This new version includes sixteen new genomes and implements new methods for colinearity and orthology detection.

4.3 Material and Methods

4.3.1 Gene Models and Gene Families

An overview of all primary sources supplying gene annotation data is presented in Table 4.1. All genomes, and their associated gene models, where first parsed into a uniform format and stored in a relational database. The association of a gene model with one of the four different gene types (coding, RNA, transposable element and pseudo gene) was extracted from the primary data sources. For species lacking chloroplast and mitochondrial DNA sequences, organellar genomes were, when available, obtained from the European Bioinformatics Institute[a] (EBI) .

The gene models, DNA sequences and protein sequences were tested for consistency, and irregular results (such as mismatches between translated DNA sequences and the provided protein sequences) were flagged in the database. The longest splicing variant was selected as representative for genes with alternative transcripts and, in turn, used in subsequent analyses focusing on gene family delineation and colinearity detection. Splice variants, if annotated, could be explored with the genome browsers AnnoJ[136] or GenomeView[199]. Gene families were delineated by first computing the protein sequence similarity through an all-against-all BLAST (e-value cutoff 1e-05, retaining the top 500 hits), and then by applying

[a] http://www.ebi.ac.uk/genomes/organelle.html

TribeMCL[79] and OrthoMCL[61] to cluster genes in families and subfamilies, respectively.

The PLAZA species tree was manually constructed using information from NCBI taxonomy[200] and literature[201] to resolve trifurcations.

The core gene families were selected by a phylogenetic approach: the clades with at least two non-leaf subclades were retained from the PLAZA species tree, and to be considered as a core family for these clades, at least one organism within each of the subclades had to possess a representative gene. This approach inferred, based on parsimony, that a gene family was present at ancestral nodes with tolerance of potential annotation errors in a limited number of species. The total set of core gene families for a given clade consisted of the intersection gene family sets generated by subclades. For each core gene family representative genes were selected, using BLASTP scores with other gene family members as evaluation metric. To assess the gene space in available plant genomes, each core family was counted with a weight equal to 1/average family size. The average gene family size was defined by the total number of genes in a gene family divided by the number of species within that family. The weighting scheme corrected for the observation that the probability of finding a homolog is higher for large families compared to single-copy or small families.

4.3.2 Colinearity

Homologous genomic regions were detected with i-ADHoRe 3.0[47,48], that identified colinear regions based on conserved gene order and content. i-ADHoRe was run with the settings: alignment_method gg2, gap_size 30, tandem_gap 30, cluster_gap 35, q_value 0.85, prob_cutoff 0.01, multiple_hypothesis_correction FDR, anchor_points 5 and level_2_only false. Tandem gene duplicates were also determined with i-ADHoRe, whereas the relative dating of duplicated genes in colinear regions was done with PAML (settings: verbose 0, noisy 0, runmode -2, seqtype 1, CodonFreq 2, model 0, NSites 0, icode 0, fix_alpha 0, fix_kappa 0 and RateAncestor 0).

4.3.3 Functional Annotation

Gene ontology (GO) annotation, when available, was downloaded along with the gene models. Furthermore InterPro scan[74] was run on all protein-coding gene models and additional GO annotations were inferred with InterPro to GO mapping. Redundant GO annotations were merged according to the GO evidence code rank[202]. To avoid the inclusion of obsolete GO terms, a filter was applied using the set of valid GO terms derived from http://geneontology.org[68]. The GO annotation was also projected between orthologs from eudicots and monocots[100]. GO enrichment was analyzed for each gene family, with only the organisms with genes in the gene family under investigation being used as background model for the statistical analysis (hypergeometric distribution with Bonferonni correction for multiple testing). Only GO terms covering at least half of the annotated genes in a family and with corrected p-values < 0.05 were retained.

4.3.4 Functional Gene Clusters

Clusters of functionally related genes (functional clusters) were detected using C-Hunter[198] on two different datasets that differed by whether the tandemly duplicated genes had been collapsed to a single representative or not. Two different runs were performed on each dataset, with different minimum (2/30) and maximum (10/150) cluster sizes. The e-value cutoff (0.001) and maximum cluster overlap (50%) were the same for the different runs. When multiple clusters spanning the same location were detected, because of GO terms organization as a directed acyclic graph[68], only the most significant cluster was retained.

4.3.5 Orthology Prediction and Evaluation

The PLAZA integrative approach for orthology detection was based on four methods: orthologous gene families, phylogenetic trees, colinear regions and multispecies best BLAST hits. For the gene families OrthoMCL clusters were used, the phylogenetic trees were constructed based on gene families inferred with TribeMCL, the colinear regions were detected with i-ADHoRe, and the best BLAST hits (with inparalogs), namely Best-Hits-and-Inparalogs (BHI) families, were detected by an OrthoInspector-like approach[190]. Briefly, interspecies best BLAST hits were first retrieved for each gene and in a second phase inparalogs were included, defined as the intraspecies BLAST hits that were more similar than the best interspecies BLAST hits.

For all gene families, phylogenetic trees were constructed with PhyML[139] based on multiple sequence alignments generated by MUSCLE[140]. Duplication and speciation events in the phylogenetic trees were identified by applying the NOTUNG tree reconciliation method[141]. Based on a duplication consistency score, erroneous duplications due to incongruences between the gene family and species tree, were determined[100].

The reliability of the different orthology predictions was scored with the Expression Context Conservation score (ECC)[192]. ECC compared the expression profile conservation between two species by a statistical framework evaluating shared homologous relationships between coexpressed genes. The retrieved expression compendia[192], consisted of 76 *Arabidopsis thaliana* and 63 *Oryza sativa ssp. Japonica* (rice) Affymetrix non-redundant microarray experiments. These expression data sets were constructed starting from 322 Arabidopsis and 203 rice experiments using data normalization, collapsing of redundant conditions and removal of transgenic or mutant experiments. A total of 19,937 *Arabidopsis* and 32,004 rice genes were present on the microarrays for expression analysis (based on a custom Chip Description File[192]). Pearson correlation coefficient thresholds for Arabidopsis and rice were 0.48 and 0.41 respectively. The evaluation of the different orthology predictions using ECC was performed using homologous gene relationships based on TribeMCL clusters.

4.4 Author Contribution

As the first author, I had the lead role in the both the original development of PLAZA (see chapter 3), the further development of the platform, as well as the writing the manuscript. All new visualizations (except for the WGDotplot applet) were designed and implemented by me, although the data retrieval backend of the WGDotplot applet was reimplemented by me due to performance issues. The evaluation of the Integrative Orthology method through expression data was also performed by me. For a more technical overview of issues encountered during the development, see chapter 7.

"Extraordinary claims require extraordinary evidence"

Carl Sagan

5
PLAZA Applications

Abstract

The development of the PLAZA[84,100] platform has played an instrumental role in further research, as can be observed by the multitude of research papers (65 for PLAZA 1.0, 9 for PLAZA 2.5 at the time of writing) citing and using PLAZA. With the usage of PLAZA ranging from a data warehouse to actively browse the website and its tools to infer information. Instrumental in understanding how to browse and use the PLAZA platform are the various use cases published by our group.

Examples include the study of duplicated genes and regions, using the Integrative Orthology method to help quantify conserved co-expression and using the GO enrichment tool to infer functional biases in expression data. Such examples clearly indicate how the PLAZA platform can be used to study a wide range of problems. With the power provided by having direct database access, more advanced use cases become possible.

The content of this chapter is based on various published research papers, to which I made significant contributions[100,203,204]. For the author contributions, see page 5-15.

5.1 The Study of Gene Duplicates Using the PLAZA Platform

To illustrate the applicability of PLAZA for comparative genomics studies, a combination of tools was used to characterize in detail the mode and tempo of gene duplications in plants. In the first case study, tree-based dating and GO enrichment analysis were used to analyze the gene functions of species-specific paralogs. Initially, gene duplicates were extracted from the reconciled phylogenetic trees for all organisms. To ensure the reliability of the selected duplication nodes, we only retained nodes with good bootstrap support ($\geq$70%) and consistency scores ($>$0.30) (calculated as described in Vilella et al.[143]). By cross-referencing all returned genes with the colinearity information included in PLAZA, all species-specific duplicates were further divided into tandem and block duplicates. Subsequently, enriched GO terms were calculated for each of those gene sets using PLAZA's workbench. Whereas in the green alga *Ostreococcus lucimarinus*, 45% of all species-specific duplicates are derived from a recent segmental duplication between chromosomes 13 and 21, nearly half of all grapevine-specific duplicates correspond with tandem duplications (see Supplemental Table 5 accompanying Proost et al.[100]). For many species, tandem duplications account for the largest fraction (34 to 50%) of species-specific paralogs. The GO enrichment analysis provides an efficient approach to directly relate duplication modes in different species with specific biological processes or evolutionary adaptations. Browsing the associated gene families makes it possible to explore the functions of the different genes (Figure 5.1).

5.1.1 Duplicated Resistence Genes in Arabidopsis and Poplar

The GO term *response to biotic stimulus* (GO:0009607) was enriched for the tandem duplicates of *Arabidopsis thaliana*, poplar, and grapevine. When focusing on the duplicated genes causing this enrichment, we observed that different gene families involved in biotic response are expanded in different species (Figure 5.1B). Whereas in *Arabidopsis thaliana*, the Avirulence-Induced Gene and anthranilate synthase family are associated with bacterial response, genes from expanded families in poplar, covering a/b hydrolases, a set of proteins with a currently unknown function (DUF567), and proteinase inhibitors, have been reported to be involved in response to fungal infection. Quantification of fungus-host distributions based on the fungal databases from the USDA Agricultural Research Service and literature[205] reveals, for different regions worldwide, 1.5 to 106 times more fungal interactions for poplar compared with *Arabidopsis thaliana*. These findings indicate a strong correlation between the wide distribution of poplar - fungal interactions and the adaptive expansion of specific responsive gene families.

5.1.2 Tandem and Block Duplicates in *Chlamydomonas reinhardtii*

In *Chlamydomonas reinhardtii*, both tandem and block duplicates exhibit a strong GO enrichment for the term *chromatin assembly or disassembly.* Inspection of the gene families responsible for this GO enrichment revealed that the four major types of histones (H2A, H2B, H3, and H4) are included. When analyzing other plant genomes, we observed that the histone family expansions were specific for *Chlamydomonas reinhardtii*. Detailed analysis of these genes reveals that there are 28 clusters that are composed of at least three different core histones (Figure 5.2). During the S-phase of the cell cycle, large amounts of histones need to be produced to pack the newly synthesized DNA. In order to increase histone protein abundance, gene duplication, as also observed in mammalian genomes, provides a biological alternative compared with increased rates of transcription[206–208]. Apart from sufficient histone proteins in rapidly dividing cells, exact quantities also are required for correct nucleosome formation. The assembly of hi-

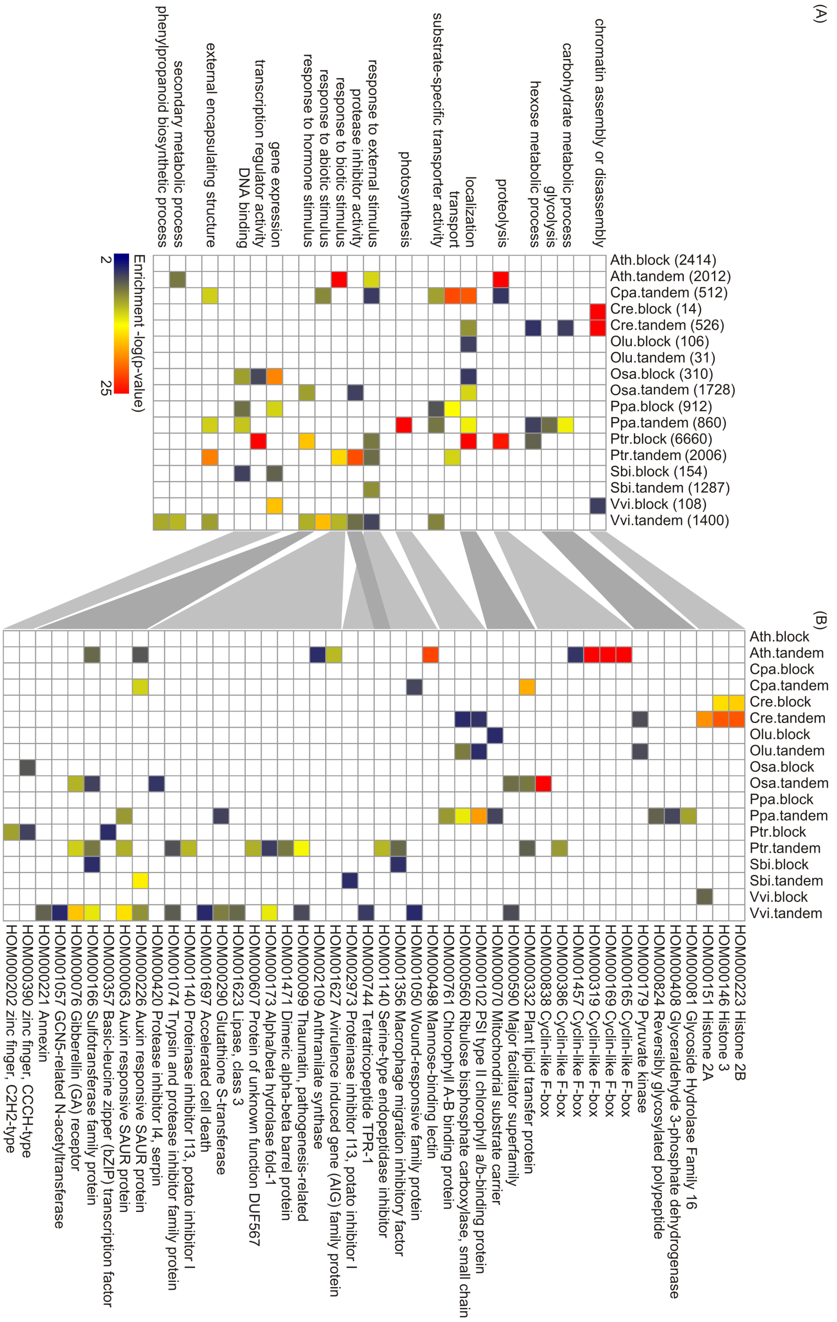


Figure 5.1: GO Enrichment Analysis of Species-Specific Gene Duplicates. *(A) The GO enrichment for species-specific block and tandem duplicates in different species is visualized using heat maps. Colors indicate the significance of the functional enrichment, while nonenriched cells are left blank. The number of genes per set is indicated in parentheses. (B) Family enrichments indicate expanded gene families for different species. The gene sets are identical as in (A). The gray bands link the enriched GO terms with the corresponding gene family expansions.*

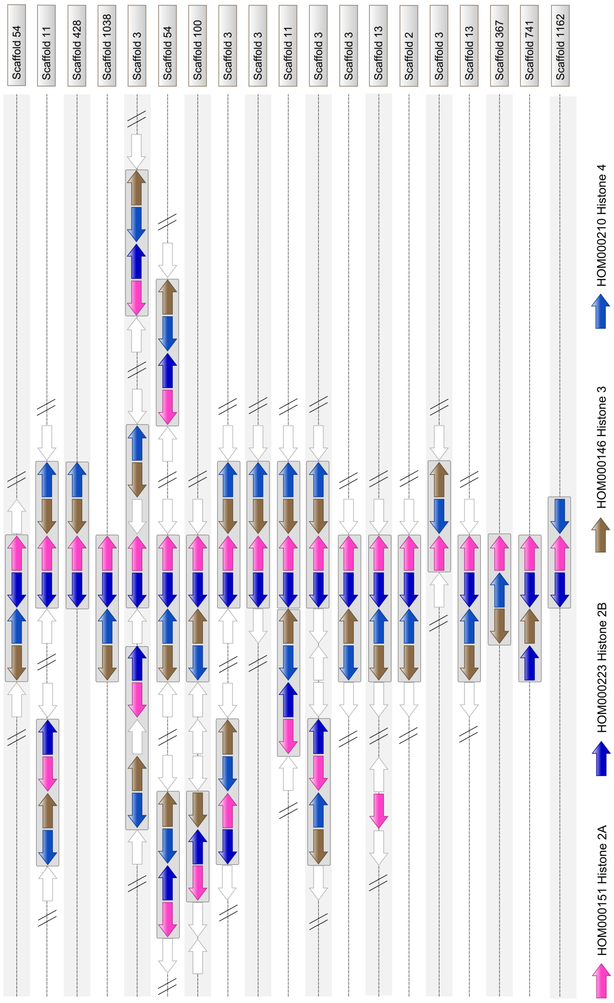


Figure 5.2: Duplicated histon genes in *Chlamydomonas reinhardtii. The genomic organization of the core histone genes in Chlamydomonas reveals a pattern of dense clustering (indicated by gray boxes). Genes are shown as arrows; the direction indicates the transcriptional orientation and colors refer to the gene family a gene belongs to (families occurring only once are not colored for simplicity).*

stones occurs in a highly coordinated fashion: two H3/H4 heterodimers will first form a tetramer that binds the newly synthesized DNA and subsequently the addition of two H2A/ H2B dimers completes the histone bead[209,210]. As shown in Figure 5.2, the histone pairs that form dimers, which therefore should be present in equimolar amounts, occur very frequently in a divergent configuration (>95% of the histone genes occur in head-to-head pairs with their dimerization partner). This specific gene clustering suggests that bidirectional promoters guarantee equal transcription levels for the flanking genes[197].

5.2 Comparative Co-expression Analysis in Plants

Comparative sequence analysis is a successful tool to study homologous gene families, define conserved gene functions between orthologs, and identify lineage- and species-specific genes. Apart from conserved sequences, inter-species differences, quantified through gene expression data, provide important clues about evolutionary history and species-specific adaptations[211]. Consequently, the integration of functional genomics information provides, apart from gene sequence data, an additional layer of information to study gene function and regulation across species[212]. Depending on the availability of expression profiling technologies and the evolutionary distances between the species under investigation, a number of different approaches can be applied to study expression profiles between organisms[213].

Although comparative expression analysis is most straightforward when compatible expression data sets are used that cover equivalent conditions for all species, only a small fraction of all available data in different species can be utilized in this approach[212]. To overcome these limitations, pioneering comparative transcriptomics studies have shown that comparing co-expression, instead of the raw expression values, provides a valid alternative to identify gene modules (set of co-expressed genes potentially sharing similar function and regulation) and study their evolution[214,215]. An advanced case-study to systematically compare microarray expression data across species is presented here. Apart from the retrieval, normalization and annotation of microarray expression information, challenges related to the detection of co-expressed genes, the accurate delineation of gene orthology and the integration of expression networks and homology data are highlighted.

5.2.1 Construction of Co-expression Networks and Comparison Across Species of Co-expression

In order to compare genome-wide expression profiles between different species, most studies apply a clustering algorithm to search, based on a large-scale expression compendium, for groups of highly co-expressed genes per species. The idea of clustering is to study groups of genes, sharing similar expression patterns, instead of individual ones. There are many different gene expression clustering tools available, with each its own advantages and disadvantages and with most applying a similarity or a distance measure to construct gene co-expression clusters [216]. Here we use the Pearson correlation coefficient (PCC) , one of the most commonly used similarity measures. Clusters of genes showing expression similarity are derived using either rank-based (gene-centric) or graph-based methods.

A major objective in comparative expression studies is the systematic comparison of gene clusters across species using homologous or orthologous genes. Defining sequence-based orthologs is a powerful approach to link expression datasets across species and to identify genes with conserved gene functions or conserved modules that participate in similar biological processes[213–215]. Different approaches are available to identify homologous and orthologous genes[60], with most of them starting from the output of a global all-against-all sequence similarity search. Examples include reciprocal best hits (RBH), clustering methods such as OrthoMCL and phylogenetic trees construction, with the last one theoretically providing the highest confidence. Each one of these methods has its own strengths and weaknesses, but by using an ensemble method (e.g. PLAZA Integrative Orthology[84]), a consensus can be reached from multiple orthology predictions.

Gene expression is typically compared between species in a pairwise manner and, optionally, information about conserved genes in multiple species is combined[217]. Although this approach provides a first glimpse on the co-expressed genes that are conserved between different species[218], recently developed methods also apply statistical tests to verify if the number of shared orthologs between two expression clusters is significant[192,217,219]. Indeed, a potential bias exists with regards to the size of the co-expression clusters, the nature of the orthology relationships and the size of the orthologous groups. One can correct for this bias through random permutation sampling to test whether the number of shared orthologs is significantly higher than expected by chance.

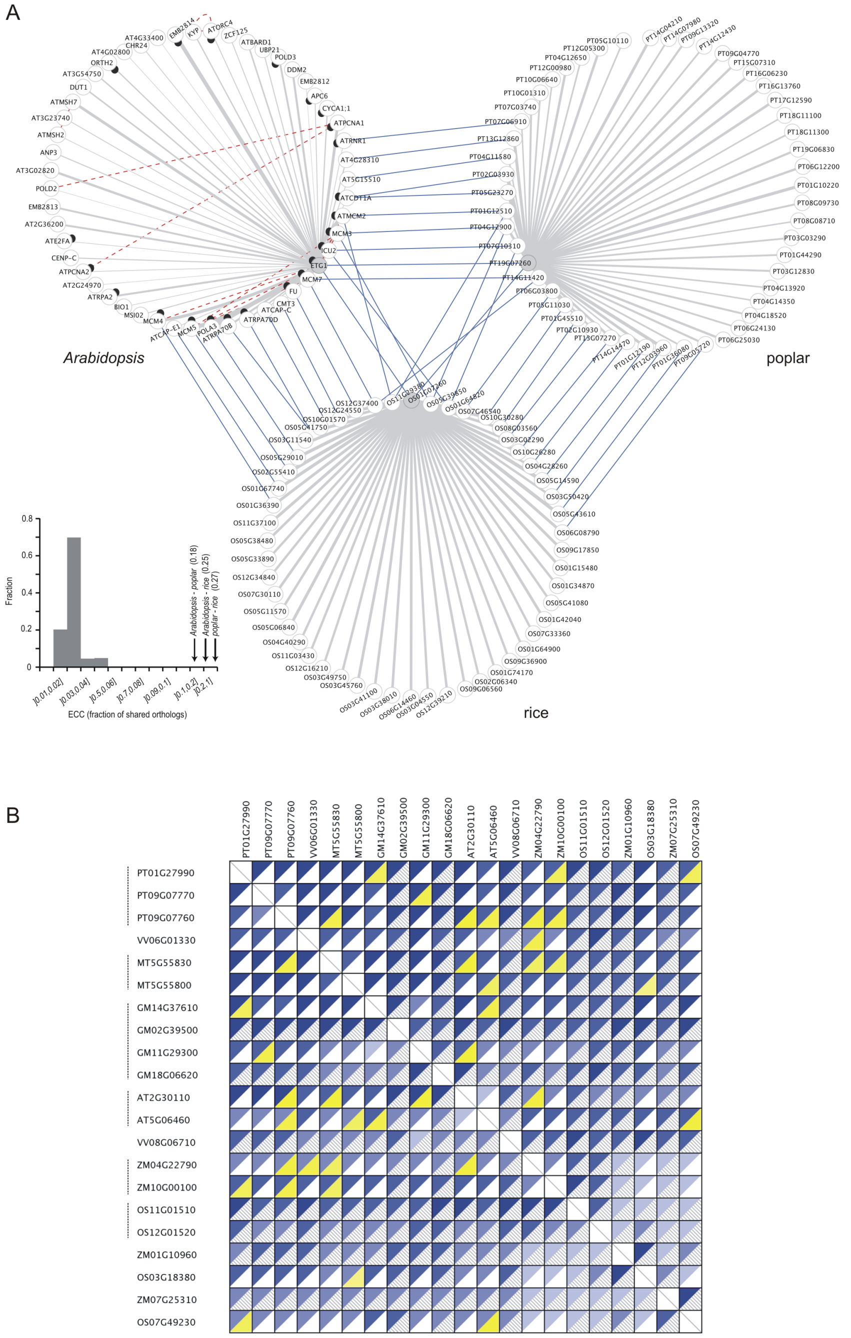
5.2.2 Functional Annotation

Gene annotation enrichment analysis is a high-throughput strategy that increases the likelihood for investigators to identify biological processes most pertinent to their study, based on an underlying enrichment algorithm[220]. The integration of known protein-protein interactions, tissue specific expression or phenotypic information from mutant lines provides an additional level of experimental information that can be used to characterize conserved modules[192,217,221].

5.2.3 Studying Conserved Gene Functions Using Comparative Co-expression Analysis

To demonstrate the power of comparative co-expression methods to study gene functions across species, Figure 5.3 displays the result of a comparative transcriptomics analysis for the Arabidopsis gene ETG1 (AT2G40550). Whereas this gene was previously described as a conserved E2F target gene with unknown function[116], recent experimental work revealed it has an essential role in sister chromatin cohesion during DNA replication[222]. To identify the biological role of ETG1 and verify whether it is part of a conserved co-expression module in plants, we first characterized the gene's co-expression context based on a general Arabidopsis expression compendium from CORNET[223]. Retrieval of the 50 most co-expressed genes based on the PCC yielded a set of genes showing a strong GO enrichment towards 'cellular DNA replication' (90-fold enrichment, p-value 19 1.33e-36). Enrichment analysis for known plant cis-regulatory elements using ATCOECIS[224] yielded enrichment for the E2F binding site TTTCCCGC (18-fold enrichment, p-value 21 1.41e-18), confirming that ETG1 is a putative E2F target gene. To explore whether this functional enrichment is evolutionary conserved, we first searched for ETG1 orthologs using the PLAZA 2.0 Integrative Orthology Viewer in species for which microarray data is publicly available. Whereas poplar, maize and rice have one ETG1 ortholog (PT19G07260, ZM03G04050 and OS01G07260, respectively), two copies were found in soybean (GM04G39990 and GM06G14860). Next, for each species a general expression compendium was compiled using Affymetrix experiments from GEO and the top-50 co-expressed genes were isolated in these organisms as well. Finally, the number of shared orthologs between the different co-expression clusters was determined and the resulting conserved modules were delineated (Figure 5.3). Based on the ETG1 Arabidopsis co-expression cluster, 9 and 13 orthologous genes were conserved with the co-expression clusters for poplar and rice, respectively. Whereas for both species the fraction of conserved orthologs is much higher than expected by chance (p-value ¡1e-5, see inset Figure 5.3), the functions of these orthologs (MCM2-5, MCM7, RPA70B, RPA70D and POLA3) as well as the expression context conservation in both monocots and dicots lend support for the conserved role of ETG1 in DNA replication.

Based on the frequent nature of many-to-many gene orthology relationships in plants, mediated by large-scale duplication events[14], comparative transcriptomics also offers a practical solution to identify


A
Arabidopsis
poplar
rice
ATPCNA1
ATRNR1
AT4G28310
AT5G15510
ATCDT1A
ATMCM2
MCM3
ICU2
ETG1
MCM7
FU
CMT3
ATCAP-C
ATRPA70D
ATRPA70B
POLA3
MCM5
ATCAP-E1
MCM4
MSI02
BIO1
ATRPA2
AT2G24970
ATPCNA2
CENP-C
ATE2FA
AT2G36200
EMB2813
POLD2
AT3G02820
ANP3
ATMSH2
AT3G23740
ATMSH7
DUT1
AT3G54750
ORTH2
AT4G02800
CHR24
AT4G33400
EMB2814
KYP
ATORC4
ZCF125
AT8ARD1
UBP21
POLD3
DDM2
EMB2812
APC6
CYCA1;1
Fraction
0.8
0.6
0.4
0.2
0
Arabidopsis - poplar (0.18)
Arabidopsis - rice (0.25)
poplar - rice (0.27)
]0.01,0.02]
]0.03,0.04]
]0.5,0.06]
]0.7,0.08]
]0.09,0.1]
]0.1,0.2]
]0.2,1]
ECC (fraction of shared orthologs)
B
PT01G27990
PT09G07770
PT09G07760
VV06G01330
MT5G55830
MT5G55800
GM14G37610
GM02G39500
GM11G29300
GM18G06620
AT2G30110
AT5G06460
VV08G06710
ZM04G22790
ZM10G00100
OS11G01510
OS12G01520
ZM01G10960
OS03G18380
ZM07G25310
OS07G49230

functional homologs in multi-gene families[219]. Apart from detecting conserved gene modules, the ECC method can also be applied to identify orthologs and inparalogs with conserved co-expression between different species for which large-scale expression data is available. For a set of 21 ubiquitin-activating enzyme homologs from seven species (Figure 3B), the systematic examination of conserved co-expression between all family members makes it possible to explore whether duplicates show different conservation patterns. Application of the ECC method using the 50 most co-expressed genes revealed that, for those orthologs which have expression data, in poplar, Medicago, soybean, Arabidopsis and maize ECC patterns with orthologs from other species were different between inparalogs. This result reveals that for at least five species both co-orthologs with conserved and non-conserved co-expression contexts exist, making the transfer of biological information between different species challenging.

5.3 Studying Algal Genomics Using the pico-PLAZA Platform

The PLAZA 2.5 platform contains a sample of the available algal genomes, which can be used as outgroups in various studies. The study of the core genes[84] present in all plant species (see Figure F.1) gives a clear indication on which genes were present in the ancestral genome shared by all green plants. However, with more and more microbial eukaryotes being sequenced, and with a clear industrial and ecological interest in these species[204,225,226], a more dedicated resource was needed.

The development of the pico-PLAZA platform went through multiple stages, with the latest version containing not only 10 algal genomes from within the green lineage, but also a number of red and brown algae, as well as diatoms (see Table 5.1). While superficially similar, the phylogenetic distance between unicellular algae from the green and brown lineage is extremely large, more than 500 Mya (according to NCBI Taxonomy). However, by including these species several goals can be obtained: comparing the adaptations of unicellular organisms to aquatic environments, and studying the core eukaryotic machinery.

Here we present some case studies to clarify how the pico-PLAZA platform can be used to study the evolutionary history of algal genomes.

Figure 5.3 (preceding page): Plant orthologs with conserved co-expression. *(A) Co-expression context 1 analysis for the Arabidopsis ETG1 gene and its orthologs in poplar and rice (based on PLAZA 2.0 annotations). Grey edges represent co-expression links between ETG1 (query gene) and its top 50 coexpressed genes, weighted by the PCC value. Red dashed edges denote protein-protein interactions, black add-ons are used to indicate genes with known GO annotations for cell cycle and/or DNA replication, and blue edges depict orthology. The inset displays a histogram of the Expression Context Conservation (ECC) background model (expected number of shared orthologs for random clusters with equal sizes as real co-expression clusters) while the arrows indicate the ECC scores for the different ETG1 co-expression context comparisons. (B) Systematic evaluation of orthology and conserved co-expression using the ECC method for a set of 21 homologs (encoding ubiquitin-activating enzyme E1) from Arabidopsis, grape, Medicago, maize, poplar, rice and soybean (AT, VV, MT, ZM, PT, OS and GM prefixes, respectively). Groups of inparalogous genes are indicated using dashed vertical lines. Upper-left triangles denote the sequence-based orthologous relationship between the genes, with a darker shade of blue indicating a higher number of evidence types reported by the PLAZA 2.0 Integrative Orthology approach. The lower-right yellow triangles denote gene pairs with significant ECC scores (p-value ¡ 0.05), white triangles represent gene pairs lacking a significant number of hared orthologs (p-value .0.05) and darker shades of yellow indicate a higher fraction of shared orthologs. Arced sections denote missing expression data for at least one of the genes. ECC scores are only computed between genes from different species.*

Species	Taxonomy	Version	Scaffolds (2)	Genes	GO	InterPro
Chlamydomonas reinhardtii	Chlorophyceae	JGI 4.0	88	16,706	6,426	8,896
Volvox carteri	Chlorophyceae	JGI 1.0	762	15,544	6,082	8,410
Chlorella sp NC64A	Trebouxiophyceae	JGI 1.0	254	9,791	5,388	7,428
Coccomyxa sp. C-169	Trebouxiophyceae	JGI 1.0	51	9,994	5,071	6,731
Bathycoccus prasinos	Mamiellophyceae	UGent ORCAE	24	7,951	4,120	5,686
Micromonas pusilla strain CCMP1545	Mamiellophyceae	JGI 2.0	22	10,587	4,985	6,869
Micromonas sp RCC299	Mamiellophyceae	JGI 3.0	18	10,197	5,124	7,081
Ostreococcus lucimarinus	Mamiellophyceae	JGI 2.0	21	7,805	4,302	5,807
Ostreococcus sp RCC809	Mamiellophyceae	JGI 2.0	20	7,492	3,893	5,305
Ostreococcus tauri	Mamiellophyceae	UGent ORCAE	21	8,036	3,782	5,054
Aureococcus anophagefferens	Pelagophyceae	JGI 1.0	1,031	11,637	7,032	8,680
Ectocarpus siliculosus	Phaeophyceae	UGent ORCAE	955	16,788	7,078	10,004
Fragilariopsis cylindrus	Bacillariophyceae	JGI 1.0	202	18,074	8,353	9,032
Phaeodactylum tricornutum	Bacillariophyceae	JGI 2.0	35	10,257	5,653	6,786
Thalassiosira pseudonana	Coscinodiscophyceae	JGI 3.0	29	11,632	6,501	7,402
Cyanidioschyzon merolae	Bangiophyceae	Tokyo University	74	5,178	3,058	3,889
Arabidopsis thaliana	Streptophyta (1)	TAIR10	7	33,602	22,087	21,467
Oryza sativa	Streptophyta (1)	TIGR6.1	14	57,874	24,583	24,735
Physcomitrella patens	Streptophyta (1)	JGI 1.1	1,448	36,137	14,283	16,275

Table 5.1: Data content pico-PLAZA database. *(1) Phylum (2) Including organel genomes, if available*

Clade-specific core families	#Gene families
land plants (3)	974
green algae (10)	37
- Chlorophyceae (2)	1827
- Trebouxiophyceae (2)	139
- Mamiellophyceae (6)	449
diatoms (3)	1035

Table 5.2: Clade-specific gene families in pico-PLAZA. *(1) Numbers between parentheses indicate species counts*

5.3.1 Gene Dynamics in Algal Genomes

Using the Gene Family Finder tool which allows searching gene families through phylogenetic profiles (i.e. the presence or absence of a gene family in a species), it is possible to study gene family dynamics across species and shed light on ancestral gene content[130]. Comparing the number of genes assigned to families reveals large differences for individual species: whereas O. tauri contains 6893 genes grouped into families, V. carteri has more than 13,800 genes. The identification of orphan genes (genes lacking paralogs or homologous genes in any other species present in the platform; based on BLASTP E-value 1e-05 threshold) as well as splitting up gene families based on the copy-number (single-copy versus multi-gene) and species content (species-specific or sharing homologs in other species) provides a general overview of gene distributions in the different species (Figure 5.4A). The Ostreococcus species have the most streamlined genomes characterized by a large number of multi-species single-copy gene families and a low number of multi-copy families and orphans. Reversely, the largest number of genes in multi-copy families is observed in V. carteri and C. reinhardtii. Together with the high abundance of species-specific gene families, 1827 gene families are shared and unique to the Chlorophyceae (Table 5.2). These results explain the overall high number of genes present in these two species. Counting paralogous genes for all species reveals an important role for tandem duplication explaining putative gene family expansions (Figure 5.4B).

In contrast to species-specific features, determining the number of core genes (i.e. genes shared in

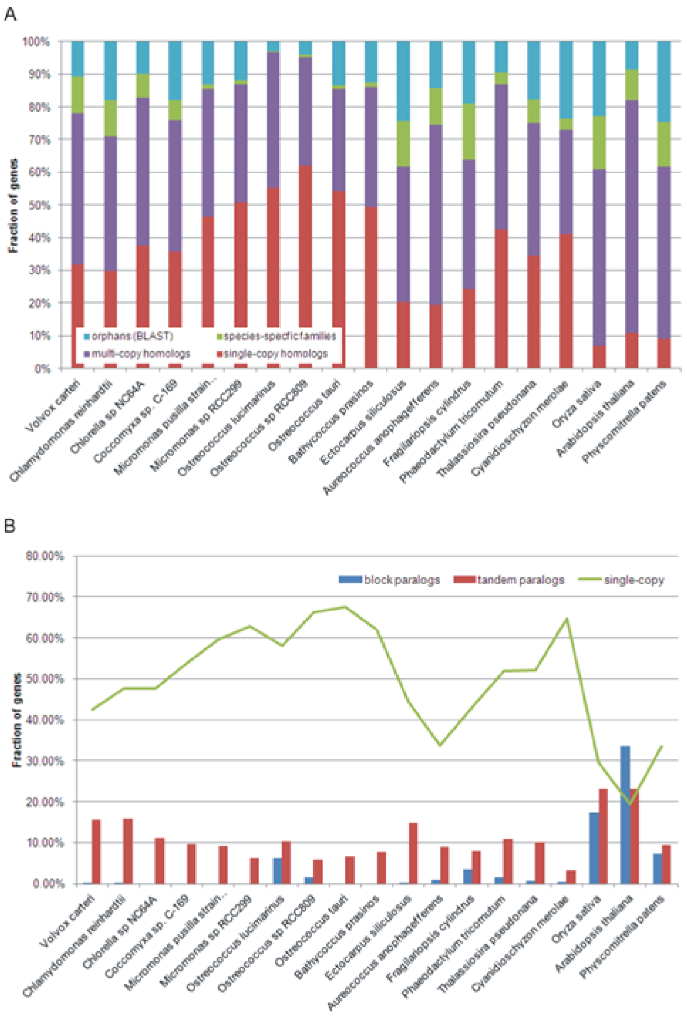


Figure 5.4: Overview gene content in different species. *(A) Fraction of protein-coding genes assigned to different categories based on homologs in other species and copy number. (B) The fraction of block and tandem duplicates is depicted using bars whereas the fraction of single-copy genes is indicated by the green line.*

all species from a clade or specific set of organisms) within green algae revealed that 2078 core families are shared between all ten species (Figure 5.4). When including brown/red algae or higher plants, this number further decreases to 1494 and 1089, respectively. Considering both the large number of new pan families (Figure 5.4) as well as clade-specific families (Table 5.2), it is clear that both the acquisition of new gene functions as well as the expansion of specific gene families plays an important role in the relationship between genotype and phenotype[204,227]. Examples of expanded functional categories include proteins with ankyrin repeat-containing domains in Ectocarpus siliculosus and Bathycoccus prasinos, protein kinases in Chlamydomonas reinhardtii, and tetratricopeptide-like helical proteins in Ectocarpus siliculosus and Aureococcus anophagefferens.

5.3.2 Functional Analysis of Large-scale Expression Data

Apart from browsing individual genes or functional categories, pico-PLAZA can also be applied as a data warehouse to analyze large gene sets or characterize new sequences. To demonstrate this feature, we performed a functional and comparative analysis of a set of >10,000 EST sequences from Phaeodactylum tricornutum using the Workbench. Based on a large-scale expression data set of >120,000 sequenced cDNAs from 16 different libraries[228], we created two Workbench experiments for each library. The first experiment comprises all sequences expressed in that condition (called condition_all), the second experiment covers sequences uniquely expressed in that condition (called condition_specific). We present a detailed analysis of sequences from the 'urea adapted (ua)' library. After mapping all 3436 'ua' sequences to the genome annotation of Phaeodactylum tricornutum using BLASTN against annotated transcripts (E-value $< 1e - 05$), a total of 2863 gene models were tagged with one or more EST sequence. 94% of these genes are associated to 1954 pico-PLAZA multi-gene families and a detailed analysis of the phylogenetic family profiles reveals that 69 and 441 families are specific to Phaeodactylum tricornutum and diatoms, respectively. Interestingly, the latter includes a family of S-adenosylmethionine decarboxylases (HOM004619) involved in spermidine biosynthesis putatively acquired through horizontal gene transfer from a bacterial donor[228]. GO enrichment analysis (Supplementary data file 4) of the 'ua_all' gene set reveals an over-representation of genes involved in nitrogen metabolism (405 genes), amino acid metabolism (117 genes) and organic acid metabolism (132 genes), confirming previous results[228]. From the set of 145 'ua_specific' transcript sequences, 46 could be mapped to gene models while the 36 associated gene families comprising a variety of functional categories (no significant GO enrichment). Interestingly, five 'ua_specific' gene families have only homologs in diatoms, therefore comprising diatom-specific genes playing a role in urea-mediating signaling.

5.3.3 Environmental Genomics

Based on the different integrated genomes, pre-computed gene families and detailed gene orthology information, pico-PLAZA enables a systematic screen of the gene content from complete genomes of microbial photosynthetic eukaryotes to identify, apart from 18S rDNA, alternative barcoding genes. These barcoding genes preferably should be single-copy genes with scalable phylogenetic spread from the genus to the order and phyla level. Although this case study is currently restricted to the lineages represented in pico-PLAZA, the number of available genomes will rapidly increase by future genome projects of microbial eukaryotes and large-scale sequencing initiatives such as the TARA ocean protist sequencing project[229] and CAMERA[230].

To identify lineage-specific genes for environmental monitoring, the Gene Family Finder tool can be used find species/clade-specific gene families and identify putative gene markers. For example, 442 protein coding genes are single-copy in all three Ostreococcus species (option 'Clade selection: Ostreococcus') and absent in Micromonas and Bathycoccus. The single-copy feature of a candidate barcoding gene is essential to avoid spurious diversity overestimation from multiple gene copies within a genome. Performing a query on single-copy genes in the order Mamiellales leads to the retrieval of 328 gene families. For each of these gene families, visual inspection of the amino-acid alignment using JalView enables the identification of conserved motifs for PCR primer design. This two-step protocol represents a practical approach for the detection of genes that can be used to investigate the prevalence of Ostreococcus (or Mamiellophyceae) in the environment. Because of relaxed selective constraint on synonymous positions, these protein-coding gene markers will enable us to investigate intraspecific alongside interspecific diversity.

As a second example, we demonstrate how pico-PLAZA can be used to identify intraspecific markers based on multi-species colinearity. The level of nucleotide polymorphism at neutrally evolving sites is a fundamental parameter in molecular evolution, as it is informative about the mutation rate and the effective population size of a species. The proportion of neutrally evolving sites is expected to be lower in protein-coding genes than in intergenic regions, where it depends on the neighboring gene orientation[231]. In Ostreococcus, intergenic regions flanked by two stop codons (called 'tail-to-tail' intergenic regions) have the highest proportion of neutral evolving sites[231]. Based on the GenomeView genome browser[199], pico-PLAZA enables the rapid identification of tail-to-tail intergenic regions in each genome. Furthermore, using the cross-species colinearity information, it provides detailed information about conserved intergenic regions that are flanked by orthologous genes. These regions are good candidates for the estimation of intraspecific diversity from environmental strains. Previously, eight of these tail-to-tail intergenic regions have been sequenced in 18 wild isolates of O. tauri in the NW Mediterranean sea to estimate the level of nucleotidic polymorphism[232]. This polymorphism pattern provided indirect evidence for meiotic recombination in natural populations.

5.4 Author contribution

The work in this chapter is based on several published papers[100,203] and one manuscript in preparation, all of which contain me as a co-author.

- Section 5.1 is based on Proost et al.[100], which I co-authored as a shared first author (see also chapter 3).
- Section 5.2 is based on Movahedi et al.[203], for which I performed a large part of the analysis: writing and testing the permutation and evaluation scripts of the expression data (with a redesigned and optimized procedure compared to Movahedi et al.[192] for generating the used background model), as well as developing the code to generate Figure 5.3. Significant contributions were also made to the manuscript by myself.
- Section 5.3 is based on a manuscript in prepation (which will present the pico-plaza platform). Here I contributed significantly by being involved in the initial conception of the platform (based on the PLAZA 2.5 code, see also chapter 4), parsing and uploading large parts of the initial structural data, creating custom visualizations for shared genomic alignments between algae, as well as small contributions to the manuscript. The enrichment of the used datasets was also automated by me.

"I have no use for people who have learned the limits of the possible."

Leonard of Quirm – The Last Hero

6

TRAPID, an Efficient Online Tool for the Functional and Comparative Analysis of *De Novo* RNA-Seq Transcriptomes

abstract

With the arrival of Next Generation Sequencing (NGS) technologies, it is now possible to acquire the entire transcriptome of non-model species for a relatively low price. The processing of these transcripts does however pose new challenges in terms of computational requirements and bioinformatics expertise. In order to mitigate some of these problems, we have developed TRAPID[a]: an efficient online tool for the fast processing of transcriptome data. TRAPID can be used to perform both functional and comparative genomics analyses, by taking homologous and orthologous relationships with species from pre-defined reference databases into account. Functional analyses consist of evaluating Gene Ontology (GO) assignments and protein domain information, while the comparative genomics analyses cover the investigation of possible gene family expansions, and the creation of multiple sequence alignments and phylogenetic trees. We benchmarked the critical components within the TRAPID pipeline against similar software, although no other tool (that we know of) offers the full functionality of TRAPID.

This chapter is based on a manuscript in preparation. For the author contributions, see page 6-14.

[a] http://bioinformatics.psb.ugent.be/webtools/trapid/

6.1 Introduction

Technological advances in sequencing have made it possible to rapidly and cost-effectively take a snapshot of gene expression in a specific tissue or condition and have led to an explosion of transcriptome RNA-Seq data. For the plant kingdom alone, more than 600 RNA-Seq experiments were available in the NCBI Short Read Archive database at the end of 2011, covering in total close to 2Tb of raw sequence data. Remarkably, more than 80% of these experiments were derived from species for which a draft or complete genome sequence was lacking, which makes the data processing and biological interpretation a challenging task. In case a reference genome is available, the short reads can be processed using alignment-first (or align-then-assemble) methods that provide a genome-guided approach to study splice site junctions, identify new or alternative transcripts, or to quantify expression levels using known gene annotations[233]. In contrast, for species without a reference genome, assemble-then-align methods require that the millions of reads are first processed using de novo assembly before the reconstructed transcriptome is further characterized[234]. Examples of downstream analysis include the remapping of the input sequence reads from the different libraries to the assembled transcripts to quantify expression levels, the remapping of all reads to assess the genetic diversity within a genotype, or the alignment of the assembled transcripts against genome or transcripts sequences from closely related species .

The development and improvement of de novo transcript assembly tools is an active research field and algorithms like OASES/Velvet, Trans-ABySS and SOAPdenovo provide efficient tools to reconstruct transcriptomes for non-model species starting from raw sequence reads[234–237]. Despite the fact that both library normalization and increasing sequencing depths (or higher coverage) will have a positive influence on the completeness of a transcriptome[238], most de novo transcriptome studies typically present gene catalogues where the number of transcripts after the assembly phase exceeds the estimated number of genes[239]. This pattern is mainly the result from redundancy caused by the presence of partial, unassembled or highly heterozygous sequences. Despite these imperfections, transcriptomes provide a sequence backbone for various non-model species and, in line with traditional genome projects, a detailed annotation of these transcript sequences is an important feature for the downstream biological analysis. As an increasing number of unassembled singleton reads is inversely correlated with the number of full-length assembled unigenes, the success of generating high quality functional annotations will highly depend on the complexity of the transcriptome being analyzed and the applied sequencing strategy[238].

Although the workflow to process transcriptome data is highly dependent on the type of analysis, functional annotation for the assembled transcripts is often derived in the same way via sequence similarity searches against a reference database. Clearly, the default application of large-scale sequence similarity searches against databases like NCBI or UniProtKB, which contain annotated proteins, drastically increases the amount of data that needs to be interpreted to derive functional annotations. Currently, systems like KEGG Automatic Annotation Server (KAAS)[240] , Blast2GO[241] and T-ACE[242] provide tools for non-expert users to perform functional characterization of transcript sequences, but both the throughput as well as the quality of the reference datasets are important factors influencing the biological knowledge that can be extracted from non-model transcriptomes. Whereas systems like KAAS and Blast2GO can be operated through a web-browser, T-ACE requires the installation of a PostgreSQL database on local hardware. Although both Blast2GO and T-ACE can derive functional annotations from a BLAST search against NCBI or through protein domain identification using InterProScan, the associated runtimes grow rapidly, hindering the efficient processing of a complete transcriptome dataset. Furthermore, the quality of the functional annotations of known sequences as well as the number of species or genes included in reference databases will have a big impact on the success of translating transcript sequences into functional gene catalogs. Tools which apply the Gene Ontology (GO) controlled vocabulary benefit from the different functional levels embedded in the ontology structure, while systems like KEGG Orthology provide detailed information but only for a limited number of genes. Apart from functional transcript annotations, the analysis of transcripts from non-model species using comparative genomics can also generate valuable information about conserved pathways, gene family expansions, species-specific genes and genetic diversity[243–245]. However, performing such evolutionary analyses for thousands of transcripts is computationally expensive and user-friendly interfaces to compare de novo transcriptomes with high-quality reference genomes are still missing.

To address some of the issues inherent to the analysis of de novo transcriptomes, we present TRAPID, a web-based and high-throughput analysis pipeline which uses predefined reference databases. Available analyses include automatically identify coding sequences in transcripts, correcting frameshifts, assigning coding-sequences to multi-species gene families, performing transcript quality control and generating functional annotations. Furthermore, detailed multiple sequence alignments and phylogenetic trees can easily be generated providing a comparative framework for the analysis of non-model transcriptomes. Finally, quantitative comparisons can be performed to study functional biases in transcriptome subsets derived from different tissues or conditions.

6.2 Results and Discussion

6.2.1 General Properties of the TRAPID Transcript Analysis Platform

To provide a web-based resource for the high-throughput processing of assembled transcriptomes derived from de novo RNA-Seq experiments or classical EST sequencing, a two-step procedure was developed. First, large-scale sequence similarity searches and open-reading frame (ORF) detection are combined to identify coding sequences, assign transcripts to gene families, identify partial/full length transcripts and generate homology-based functional annotations. In a second step, detailed sequence analysis can be performed on-the-fly to correct frameshifts and study transcripts within an evolutionary context using multiple sequence alignments and phylogenetic trees (see Figure 6.1). Although building a transcriptome analysis pipeline based on standard components for similarity searches and ORF detection is relatively straightforward, the large number of sequences and the fragmented nature of RNA-Seq data makes balancing the processing speed and quality a challenging task. This section outlines the basic features of the TRAPID system, with the following two sections focusing on the implementation and benchmarking of specific analysis components, and the last section providing a case study which illustrates how TRAPID can be used to quickly infer quality annotations for sets of transcriptomes in a transparent and reproducible manner.

After the user has created a personal account, logged in into the TRAPID system and uploaded a set of assembled transcripts (called an ’experiment’), a sequence similarity search using RAPSearch2[246] is executed against a specific protein database selected by the user (see Figure 6.1). Whereas *Reference proteome* databases refer to the full set of proteins for a given species or clade based on integrated genomes from PLAZA 2.5[84] or OrthoMCL-DB version 5[247], the *Gene family representatives* contain one representative protein for each species present in a given gene family. The former makes it possible to select

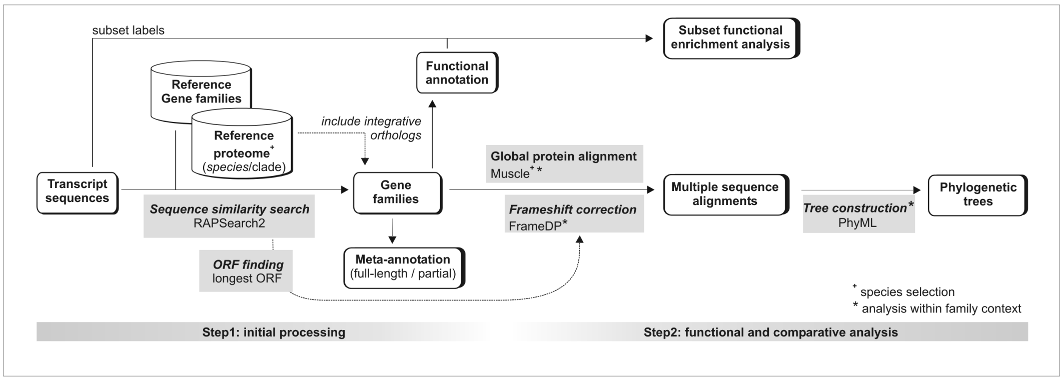


Figure 6.1: Schematic overview TRAPID pipeline. *The TRAPID pipeline consists of two separate parts. The first one is a non-interactive processing step, during which all transcripts are assigned to gene families using a similarity search, followed by functional annotation transfer and meta-annotation assignment. The second step is interactive and directly commanded through the website interface. Here, the user has the ability to analyze his data using functional enrichment analyses, multiple sequence alignments and phylogenetic trees.*

an appropriate specific taxonomic level and the full set of associated proteins based on the species the transcripts are originating from, while the latter provides a good alternative for the efficient processing of datasets within a broad taxonomic context, for example in case no closely related species with a reference proteome is available. Through the TRAPID website, we offer users a variety of reference protein databases and pre-computed gene families (see Table 6.1). Apart from 175 species-specific proteomes (25 from PLAZA and 150 from OrthoMCL-DB, including >2 million proteins) covering 25 plants, 115 eukaryotes, 36 Bacteria and 16 Archaea, 12 different clade and two Gene family representative databases were generated to assign transcripts to families in a high-throughput manner using RapSearch2.

The output of the sequence similarity searches is used to assign each transcript to a predefined gene family and to generate frame statistics to subsequently perform ORF detection. By default these frame statistics are submitted to a simple routine that extracts the associated longest ORF within the frame showing similarity with reference proteins (see section 6.3). However, this information is also used to predict whether specific transcripts contain putative frameshifts, which can in a later stage through the website be automatically corrected using FrameDP, a self-training tool to predict peptide sequences in mature mRNA sequences[248]. The association of a transcript to a specific gene family is also used to facilitate the transfer of functional consensus Gene Ontology and protein domain information to transcripts. Finally, meta-information with regards to the length of the ORF of a transcript is generated, by comparing the ORF's length to the average coding sequence length of the genes in the reference gene family.

The second phase of the pipeline is performed interactively through the website, during which the transcripts associated with each gene family are analyzed in more detail using homologous proteins from a set of reference species selected by the user. For transcripts that were flagged as potentially containing frameshifts, the user can execute FrameDP to putatively correct the transcript sequence and identify the correct ORF. Furthermore, based on the inferred coding sequences, per gene family a multiple sequence alignment is generated using MUSCLE[140] and the protein conservation of different transcripts with homologs from related species can be inspected using JalView[151]. Finally, after the application of an automatic editing routine removing non-homologous alignment positions (see section 6.3), a phylo-

Reference database	Functional annotation	Clade	#species	#proteins	GF information
OrthoMCL-DB v.5	PFAM domains	all	150	1,398,546	OrthoMCL clusters
		Alveolata	15	98,796	
		Amoebozoa	4	41,930	
		Archaea	16	30,233	
		Bacteria	36	112,059	
		Euglenozoa	9	107,034	
		Eukaryota	98	1,256,264	
		Fungi	24	680,778	
		Metazoa	29	529,788	
PLAZA v2.5	GO, InterPro	Viridiplantae	25	780,667	TribeMCL clusters, Integrative Orthologs
		Angiosperms	18	671,950	
		Eudicots	13	480,106	
		Monocots	5	191,844	

Table 6.1: Overview and content TRAPID reference databases.

genetic tree is constructed using PhyML[249] to identify orthologous and paralogous gene relationships or trace allelic transcript variants.

The evolutionary analyses are based on predefined gene families from either OrthoMCL-DB or PLAZA, and in some cases the latter, which were constructed using the TribeMCL clustering method including a quality control post processing step[100], contain multiple out-paralogous sub-types (sub-clades within a family originating from an ancient gene duplication event predating most speciation events in the tree). As a consequence, some transcripts will be assigned to big gene families covering multiple genes, which make phylogenetic analysis difficult. Therefore, in case a single species Reference proteome is selected (see Figure 6.1), it is possible to first assign transcripts to individual reference genes, e.g. from a closely related model species, and in a second phase build custom gene families through the inclusion of PLAZA integrative orthologs. These orthologs comprise for a reference gene all orthologs identified using an ensemble method combining orthologous predictions inferred through OrthoMCL, reconciled phylogenetic trees, colinearity information, and multispecies best hits and inparalogs (BHI) families[84], including inparalogs (genes from a duplication event that happened after a speciation event). In contrast to homologous gene families, families based on integrative orthology will contain a smaller number of genes covering less outparalogs which makes downstream comparative analysis more feasible and easier interpretable for large or complex families. Complementary, the user can also discard some species within a specific gene family in order to reduce the number of proteins prior that will submitted to the phylogenetic tree construction routine.

Apart from the functional annotation of individual transcripts, TRAPID also supports the quantitative analysis of experiment subsets using GO and protein domain enrichment statistics. Through the association of specific labels to sets of sequences, transcripts can be annotated with specific sample information (e.g. tissue, developmental stage, control or treatment condition) and be used to perform within-transcriptome functional analysis. Based on the integrated functional transcript annotation, GO enrichment analysis can subsequently be used to study the biological properties of specific experiment subsets or to compare the functional biases present in for example a treatment/control transcriptome experiment setup.

6.2.2 Evaluation of Homology Assignments

As shown in Figure 6.1, the first step is to assign each transcript to a predefined homologous gene family. Because transcriptome datasets for species lacking a reference genome sequence can contain more than 100,000 transcripts (with a lot of them being fragments, allelic variants, splice variants, or highly expressed non-coding genes), the efficient processing of all these transcripts is essential to provide users with results within a reasonable timeframe. Two sequence similarity tools were considered: BLASTX[250] and RapSearch2[246]. The transcript-to-family assignment results were compared using different protein reference databases with varying size, as the size of the database also influences the total runtime. BLASTX is often used to find proteins similar to a query gene in a large database, but uses a rather large amount of processing time. RapSearch2 was designed to perform the same searches but for short reads, and uses more efficient data structures to significantly speed up this process. Both tools were run using 1000 randomly selected *Arabidopsis thaliana* transcripts against different databases containing all proteins from all species within a specific clade, and the correct assignment of a transcript to a family was evaluated together with the running time. In all evaluations the protein sequences of Arabidopsis thaliana and Arabidopsis lyrata were excluded from the database and the known assignments of *Arabidopsis thaliana* gene sequences to families from the PLAZA 2.5 database were used as a golden standard. Apart from reference databases containing all proteins for a specific species or clade, the *Gene family representatives* database containing 32294 proteins was also included in the test (see section 6.3).

Assigning a transcript to a gene family was initially done with the top 10 similarity search hits using a simple majority voting rule (see section 6.3 and Table F.2). As shown in Table F.2, it is clear that both BLASTX and RapSearch2 assigned 87-98% of the transcripts to the correct gene family in all runs. For most reference databases the runtimes for RapSearch2 were approximately 10x lower compared to BLASTX, while overall the correctness of gene family assignment was equal. Increasing the reference database from one to multiple species (e.g. from the Brassicales, which only contains Carica papaya, to Eudicots, covering 11 species) quickly increases the runtimes for both tools. Overall, better results with regards to the gene family assignment can be obtained by using a larger database.

Various metrics, for example taking only one or multiple hits into account, were evaluated to assign transcripts to families (see Table F.3). The best performance was generally achieved by considering the best hit when using species/clade reference databases and majority voting using the top five hits when using the gene family representative database. To avoid over fitting of this method to *Arabidopsis thaliana transcripts*, this benchmark was repeated using *Oryza sativa spp. japonica* (*excluding Oryza sativa spp. japonica* and *Oryza sativa spp. indica* from the databases) and *Vitis vinifera* (excluding *Vitis vinifera* from the databases), yielding similar results (see Table F.2). Although one would expect the correct assignment rate of a transcript to the corresponding gene family to decrease when the assembly quality of the input transcripts deteriorates, this is not the case (see Table F.2). As such, even relatively short fragments of transcripts can be assigned to the correct gene family. Based on manual inspection of the amino acid and sequence similarity information, the user is able to modify the association between a transcript and a family in case the automatic gene family assignment was deemed incorrect.

6.2.3 Evaluation of the ORF Finding Routine

In the absence of a reference genome, transcripts generated using de novo assembly of RNA-Seq reads frequently contain errors (e.g. short insertions or deletions) and methods for the downstream analysis of coding sequences should be able to correct for potential frameshifts during ORF detection[251]. Although advanced self-learning algorithms such as FrameDP[248] exist to correct frameshifts during ORF prediction, running these tools on a complete RNA-Seq transcriptome is computationally unfeasible, even using multi-core or cluster hardware systems. Therefore, we implemented and evaluated a system to first perform the detection of putative frameshifts on all input sequences and next only process these frameshift containing sequences using FrameDP. This rationale is inspired by the observation that, when running FrameDP on complete plant transcriptomes, such as *Helianthus annuus*[252] and *Pachysandra terminalis*[253], in only 3-15% of the input sequences a frameshift was identified that could be corrected .

Apart from gene family assignments, the Rapsearch2 output is also used to estimate if a frameshift is present in an input transcript based on the output from the similarity search. For each input transcript the best hit in the reference database is selected and all alignments between this query and hit gene are evaluated. For each alignment the frame of the transcript hit is determined and if no frameshift is present, all alignments should report the same reading frame, which can immediately used to extract the corresponding longest ORF (see Figure 6.1). To evaluate this method to identify input transcripts containing frameshifts, we selected 1000 transcripts from *Arabidopsis thaliana* containing no frameshifts and an equal amount of genes where one insert or deletion was artificially introduced at a random position in the coding sequence of the transcript. Again databases of various clades, each time excluding *Arabidopsis thaliana* and *Arabidopsis lyrata*, were used along with a database containing gene family representatives, to perform similarity searches. We found that, using these alignment-based frame statistics, 72.8% of all transcripts containing a frameshift were correctly identified, with only few false positives in the dataset lacking frameshifts. To provide a good balance between global ORF quality and processing time, this method was integrated as the default procedure to identify frameshifts and subsequently correct them using FrameDP. As this benchmark experiment suggests that the applied frame statistics will miss a substantial fraction of frameshifts, the TRAPID system also provides an option for the user to run FrameDP on all transcripts within a family context.

6.2.4 Comparison of TRAPID with Blast2GO and KAAS

A feature comparison between different publicly available web-tools for transcriptome analysis reveals that TRAPID has some unique properties (see Table 6.2). The BLAST2GO[241] interface is commonly used to assign Gene Ontology functional information to DNA or protein sequences by using either BLASTX or BLASTP, respectively. Although the BLAST2GO program can also be installed locally, reducing the runtime for the user but also requiring dedicated hardware, we compared TRAPID with the online Blast2GO interface in order to only compare web-based solutions. The KAAS platform[240] provides users with KEGG pathway information for a set of given sequences based on a BLAST bit scores. This functional information is complimentary to other functional annotation systems such as GO or protein domains. Whereas all tools focus on the functional annotation of input sequences, TRAPID provides functionalities to identify and correct frameshifts, perform ORF detection and downstream sequence analysis. Especially the comparative genomics functionalities of TRAPID offer the user an intuitive interface to inspect sequence conservation using multiple sequence alignments and to identify, using phylogenetic tree construction and an extensive set of reference genomes, orthologs in related species.

We conducted a series of benchmarks to assess both runtime and the transcript coverage of functional assignments for the different web-tools reported in Table 6.2. As data-set we used 52,348 contigs from *Pogona vitticeps*, the bearded dragon lizard[245]. For TRAPID we used the Eukaryota clade from the

Features	BLAST2GO (a)	KAAS	TRAPID
Sequence similarity search	NCBI BLAST	BLAST (bi-directional)	RAPSearch2
ORF finding	no	no	yes
Frameshift correction	no	no	FrameDP
Reference database	NCBI non redundant database	Curated KEGG genes	OrthoMCL-DB version v5, PLAZA v2.5
Functional annotation	Gene Ontology, InterProScan, Enzyme codes, KEGG	KEGG (KEGG Orthology groups)	Gene Ontology, Protein domains (InterPro / PFAM)
Enrichment analysis	yes	no	yes
Protein alignments	no	no	Muscle
Phylogenetic trees	no	no	PhyML
Other	advanced stand-alone graphical user interface	graphical pathway maps	ORF length meta-annotation

Table 6.2: Feature comparison web-based transcript analysis platforms. *(a) Basic web-start version*

Dataset	KAAS(SBH)	BLAST2GO	TRAPID (Eukaryota)	TRAPID (GF representatives)
50	2 (16%)	15 (34%)	26 (28%)	3 (22%)
500	3 (21%)	146 (34%)	27 (29%)	4 (20%)
5000	6 (19%)	1460 (a)	44 (28%)	6 (18%)
52348	30 (19%)	15286 (a)	216 (28%)	42 (18%)

Table 6.3: Benchmark comparison between KAAS, BLAST2GO and TRAPID. *Benchmark between the different online platforms, giving both the used processing time (measured in minutes), and the fraction of genes which were given a functional annotation (between brackets). The dataset was derived from P. vitticeps. For TRAPID two different similarity search databases were used: the Eukaryota (1,256,264 proteins) and the gene family representatives (216,189 proteins).(a) Extrapolated time measurements.*

OrthoMCL-DB reference database (approximately 1.4 million proteins) as well as the corresponding *gene family representatives* database (216,189 proteins proteins), while for the other tools we used default parameters. The results in Table 6.3 highlight that BLAST2GO requires 70-509x more computing time to process the complete dataset, compared to TRAPID or KAAS. Comparing the fraction of sequences that received functional information reveals that BLAST2GO could annotate approximately an extra 7-15% more genes than TRAPID and KAAS, respectively. Although the functional characterization using TRAPID gene family representatives yields a 10% decrease in coverage compared to using all Eukaryota proteins, the processing is reduced drastically (216 minutes for all Eukaryota versus 42 minutes for gene family representatives).

6.2.5 Detection of Functional Biases in Transcriptome Subsets Using Enrichment Analysis

Apart from the general characterization of a complete transcriptome using various functional annotation systems, the detailed analysis of genes expressed in specific tissues or developmental stages can provide new insights about the underlying biological processes and their regulation. Starting from a recently published transcriptome from *Panicum hallii*, a model for biofuel research, we analyzed a set of transcripts showing distinct expression profiles in eight tissues for functional biases[254]. After processing all 25,392 contigs using the *Oryza sativa ssp. japonica* proteome as a reference and including integrative orthologs from the PLAZA 2.5 database, 16,748 (66%) transcripts were assigned to 9860 gene families.

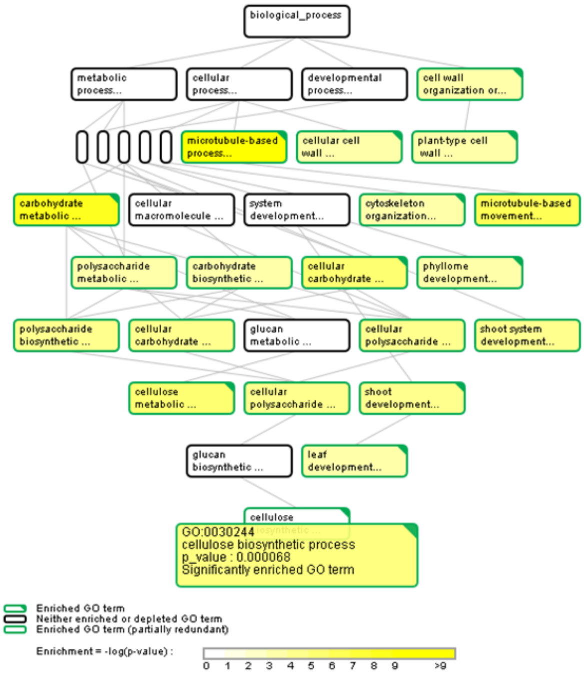

biological_process
metabolic process...
cellular process...
developmental process...
cell wall organization or...
microtubule-based process...
cellular cell wall ...
plant-type cell wall ...
carbohydrate metabolic ...
cellular macromolecule ...
system development...
cytoskeleton organization...
microtubule-based movement...
polysaccharide metabolic ...
carbohydrate biosynthetic ...
cellular carbohydrate ...
phyllome development...
polysaccharide biosynthetic ...
cellular carbohydrate ...
glucan metabolic ...
cellular polysaccharide ...
shoot system development...
cellulose metabolic ...
cellular polysaccharide ...
shoot development...
glucan biosynthetic ...
leaf development...
cellulose biosynthetic ...
GO:0030244
cellulose biosynthetic process
p_value : 0.000068
Significantly enriched GO term
Enriched GO term
Neither enriched or depleted GO term
Enriched GO term (partially redundant)
Enrichment = -log(p-value) :
0 1 2 3 4 5 6 7 8 9 >9

Figure 6.2: GO enrichment results for plant *Pamicum hallii* subset covering transcripts in stem-associated tissues. *BP enrichment plot cluster 1 (stem-associated tissues) based on iORTHO Oryza sativa run. Explain collapsed boxes and white/yellow.*

Based on the results of an expression clustering procedure reported by Meyer and co-workers, 6517 transcripts were tagged with a specific label (cluster 1-7) and GO enrichment analysis was performed for each subset. Whereas cluster 1, including transcripts with expression in stem-associated tissues, was significantly enriched for carbohydrate metabolism, cytoskeleton/cell wall organization and shoot development (see Figure 6.2), seed specific transcripts (cluster 5) included genes involved in of precursor metabolites and energy, wax metabolism and cuticle development (hypergeometric distribution, Bonferroni corrected p-value $<$0.05). Transcripts showing differential expression in root and seedling (cluster 3) were enriched for translation, ribosome biogenesis and rRNA metabolism, while leaf-specific expression (cluster 6) coincided with photosynthesis, energy metabolism and multicellular organismal development, confirming previous results[254], Finally, application of GO queries to tissue-specific subsets allows for the identification of transcriptional regulators involved in development. For example, searching for example *transcription factor activity* on subset root (cluster 4) yields 21 transcription factors showing differential expression in root, including multiple CCAAT-binding, NAM and bZIP proteins.

6.3 Material and Methods

6.3.1 Datasets, Construction Reference Protein Databases and Selection of Gene Family Representatives

The PLAZA 2.5 database was used as reference, and was also used as data source for the Arabidopsis transcripts used in the benchmark datasets. The similarity search protein databases containing clade-specific content, for both the PLAZA 2.5 and OrthoMCL reference databases where created by using NCBI taxonomy as reference. Gene family representative databases where constructed according to the procedure outlined in Van Bel et al., 2012[84].

The *Pachysandra terminalis* dataset was retrieved from Vekemans et al., 2012[253], *Helianthus annuus* and *Aquilegia formosa x Aquilegia pubescens* were retrieved from TIGR Plant Transcript Assemblies[b] [252]. *Panicum hallii* transcript sequences were retrieved from Meyer et al., 2012[254] and contig sequences showing differential expression among tissues were isolated from Supporting Information, file S8.

6.3.2 Similarity Search, Gene Family Assignment and Functional Transfer Using Homology

We used RapSearch2[246] to search for protein hits for each query transcript (comparable to BlastX). In case the selected protein database consists of either species or clade specific proteins, then only the top protein hit is retained and the associated gene family for this protein is assigned to the transcript. In case the selected protein database consists of gene family representatives, then the top 5 protein hits are retained, and the gene family for the transcript is selected based on majority voting. For each protein hit, all detected alignments are stored and used during the detection of putative frameshifts. The functional annotation for each transcript is transferred from its assigned gene family, where per gene family the GO terms and protein domains are selected which constitute 50% or more of the size of the gene family. In case not a single protein hit was detected during the similarity search, no gene family and no functional annotation is assigned to the transcript.

6.3.3 Frame Assignment and Detection of Putative Frameshifts

From each alignment of the top protein hit the strand and frame is determined for the region with sufficient similarity to the hit. If the same frame and strand was detected for each alignment the longest Open Reading Frame (ORF) within this frame is stored. In case multiple alignments occurred with the target protein in different frames, the transcript was flagged as potentially containing a frameshift and the longest ORF in all possible frames was detected and retained. For each longest ORF, additionally, it was detected if the ORF contained a start and/or stop codon. No minimum length requirement was specified to be required for each transcript.

6.3.4 Meta-annotation

The meta-annotation for all transcripts is determined by comparing the transcript length to the lengths of the coding sequences which constitute its associated gene family. In case no gene family was assigned to the transcript, or in case the associated gene family comprises less than 5 proteins, the transcript receives

[b] http://plantta.jcvi.org/

the label *No Information* as meta-annotation. Otherwise, the lengths of the coding sequences from the gene family are ordered, and the longest 10% and shortest 10% are removed in order to reduce potential outliers within the reference data. Using the remaining lengths, the average and standard deviation are computed. If the transcript length is shorter than the average minus two standard deviations, the transcript receives the label *Partial* as meta-annotation. If the transcript is longer, it receives the label *Full Length* as meta-annotation.

6.3.5 Correction Using FrameDP

Using FrameDP version 1.0.3[248] transcripts with expected frameshifts could be corrected. As a reference database all protein coding genes present in PLAZA 2.5[84] were provided. FrameDP was configured with to run with blastall 2.2.17 (settings used; Expectation value: 1e-3, Open Gap Penalty: 9, Gap Extension Penalty: 2 and retaining only the 100 best hits) while the GC3 split training with 3 iterations was used. Other parameters were left at their default values.

6.3.6 Multiple Sequence Alignments and Phylogenetic Trees

Using MUSCLE[140] translated Coding Sequences (CDS) from transcripts belonging to the same gene family were aligned with amino acid sequences of homologous genes present in the reference database. When building a phylogenetic tree, this multiple sequence alignment was edited following the same procedure as outlined in Proost et al., 2009[100]. From this final alignment phylogenetic trees were generated using PhyML, using default parameters for protein sequences.

6.3.7 Implementation

The website of the TRAPID platform was developed using CakePHP (http://www.cakephp.org), with Flash and JavaScript used for visualizations, except for the Java programs used for the multiple sequence alignments and phylogenetic trees (Jalview[151] and Archaeopteryx[152] resp.). The backend of the online tool consists of a MySQL database (http://www.mysql.com) with custom Java programs and Perl scripts. A small computing cluster is available allowing the simultaneous processing of up to four different datasets.

6.4 Conclusion

In this manuscript we have presented TRAPID, an online tool for the fast analysis of entire transcriptome datasets. TRAPID is unique in that it is fast, web-based and offers a variety of features which are not present in other comparable platforms, such as the tools for comparative genomics. We have benchmarked the most critical entities within the processing pipeline with regards to both time usage and accuracy, on real and simulated datasets, resulting in the followed rationales within the pipeline. We have compared TRAPID to other platforms (where applicable), and here it is clear that TRAPID performs equally well or better.

6.5 Author Contribution

This chapter is based on an unpublished manuscript, in which I had the lead in both writing and designing the pipeline of the platform, determining software and hardware requirements, as well as developing the web-based frontend of the platform. I'm also the lead author on the paper. Sebastian Proost provided valuable support with regards to benchmarking the platform and testing FrameDP, Klaas Vandepoele provided much-needed feedback and oversight of the platform development.

“Applications programming is a race between software engineers, who strive to produce idiot-proof programs, and the universe which strives to produce bigger idiots. So far the universe is winning. ”

Rick Cook – The Wizardry Compiled

7
Technology and Development

Abstract

The development of any software should be preceded by an assessment of the necessary requirements, the number of expected concurrent users, potential avenues of expanding the software, etc... During the development of both the PLAZA and the TRAPID platform I encountered multiple issues, which often required me to revise the chosen design decisions: database layout, software interactions, merging and splitting of functionality. In this chapter I will demonstrate some of the technological choices made, as well as parts of the software design.

A very large portion of my time as a PhD student was dedicated to designing, writing, and debugging software. The previous chapters were focused on the results of these endeavours, and how the produced software could be used to help solve biological questions. In contrast, this chapter will give a deeper understanding in the efforts and work behind the PLAZA and TRAPID platform.

For the author contributions, see page 7-11.

7.1 Data Processing

7.1.1 Data Parsing

Various challenges are encountered in the first stages of the PLAZA pipeline, which encompasses the gathering and parsing of the structural annotation of the various organisms to be added to the PLAZA platform. Not only is this structural data (which contains the gene loci, gene structures, isoforms, etc.) often presented in different file formats, but there can also be significant differences between instances of the same format.

- Some of the annotations are presented in an XML format (*Arabidopsis thaliana* and *Medicago truncatula*). Although the XML-template is clearly defined by TAIR and logically ordered (e.g. a locus containing multiple isoforms), this structure is not taken over fully for the *Medicago truncatula* annotation, making the parsing much more complicated.
- Most of the other species' annotations are in GFF3 file format. This format is a tab-delimited format, and has the unfortunate property that the last column acts as a *free format* field in which the annotators can enter any information they deem necessary. Although multiple standards have arisen for the last column, the unfortunate side-effect remains that the content of this column often complicates the data parsing, especially because identifiers and names have to be encoded at this location.

Furthermore, it is not always clear from the start whether an annotation contains all the necessary data, whether some extra data (such as isoforms) is present, and whether the data is actually consistent.

7.1.2 Data Validation

It is indeed not enough to ensure that the data is parsed correctly, an equally important step within the PLAZA pipeline is ensuring that the data is not nonsensical, through data validation. One important question which has to be answered is how inconsistent data will be dealt with: removal of the data, correction (if possible) of the data, tagging of the data, etc. are all valid options. Within the PLAZA framework we chose for the last option: if data is found to be inconsistent, we tag it as being as such in our database. This approach has the advantage that we work with the structural annotation as was intended by the providers, but the disadvantage is the possible inclusion of incorrect data in our downstream analyzes. Inconsistencies (or even errors) come in many forms:

- Missing or double identifiers
- Genes which are present multiple times in the input files
- Non-matching or impossible coordinates of genes
- Genes mapped onto non-existing scaffolds or chromosomes
- Coding genes without start and/or stop codon

Although these issues should preferentially be resolved during the parsing of the data, it became apparent that a separate validation program was needed. As such, all parsed annotation data was written to a homogenous file format for each species, and the validation program was run on these files, reporting eventual issues (see Figure 7.1). This would be an indication for the people responsible for the data parsers to change their parsers to handle the indicated inconsistencies.

```
Testing annotation.ath.public_02_5.csv
# genes : 33518
Total genes : 33518 || alternative transcript genes : 1428
Total genes : 33518 || 5’UTRs : 19537 || 3’UTRs : 19571
prot=eq        :27363
prot=ne        :7
transcript=eq  :33486
transcript=ne  :32
total genes    :33518
Coding genes      : 27379
Transposon genes  : 3901
RNA genes         : 1312
Pseudo genes      : 926
# coding genes without ATG start codon : 18   (example : AT3G29255)
```

Figure 7.1: Validation of parsed structural annotation. *Example output of the validation program, indicating potential issues. No critical errors in this annotation were detected.*

7.2 Visualizations

An introduction of the various technologies available for web visualizations has been given in section 2.4.1. Here I will describe some of the choices made during the development of the visualizations for the PLAZA platform. There are two orthogonal routes of inquiry: the used technology and whether the visualization was developed by ourselves or was retrieved from an external developer (see Table 7.1). Care was taken not to re-invent the wheel, and as such we made use of existing programs wherever we could. However, specific demands coupled with non-available source code made it unfortunately necessary to sometimes write visualizations for which already an implementation was available.

7.2.1 Graphs and Charts

The rise of JavaScript in web visualizations, at the cost of other technologies which require plugins (e.g. Flash and Java), has several implications which we are slowly addressing. During the development of the first version of the PLAZA platform the HTML5 standard was in its very early stages, and support for the *canvas*-object was missing in most web browsers. For example, the *Google Chrome* browser has been pushing the boundaries of HTML5, and has been critical in its rapid adaptation, but was only released in December 2008 (after the initial development of PLAZA had started) and had yet to gain traction. As such, the choice was initially made to create charts (such as the pie-chart on a gene family page, or the line-chart of a K_S-distribution) using the freely available *OpenFlashChart*[96] library. Later on, especially during the development of the TRAPID platform, the use of Flash for charts was abandoned in favor of JavaScript.

7.2.2 Phylogenetic Trees

The display of phylogenetic trees is a basic necessity in a comparative genomics platform, and many tree viewers have already been developed for this goal[35]. The *Archaeopteryx*[152] Java Applet (labeled *ATV* until 2010) is a well-known tree viewer, and has the additional benefit of being created with the

	Developed Program	External Program
JavaScript	CirclePlot GO Enrichment Graph GF Expansion Species Tree	Charts
Java (Applet)	WGDotplot Synteny Plot Integrative Orthology	Archaeopteryx (a) JalView
Static Image	Multiplicon View Skyline Plot Functional Clusters WGDotplot (b) Synteny Plot (b)	
Flash		Charts

Table 7.1: Web visualizations in the PLAZA platform. *(a) XML-preprocessing necessary for display of protein domains and gene structures. (b) Static image still present, but new interactive visualization available.*

PhyloXML[255] standard in mind. This PhyloXML standard (which is a template using standard XML) allows for a great freedom in extra annotations in a phylogenetic tree (compared to the ubiquitous Newick format), such as protein domain information per gene. The PhyloXML software package allows for the easy translation from Newick to PhyloXML. The extra preprocessing done to the PhyloXML files, prior to display by *Archaeopteryx*, is done by dynamically altering the XML code. This code was written by me, and allows for a wide diversity of information to be added to the PhyloXML code.

The species tree which is displayed on the main page of PLAZA is not generated using this Java Applet, and also not using the HTML5 *canvas* object. A fast-drawing method, which would in an ideal case also be dynamic, was needed for the phylogenetic tree which identifies the relations between the various species present in the PLAZA database. Because Java Applets are too slow (long initial loading time of the Java Virtual Machine), and because the HTML5 canvas object was not yet mature enough, the DHTML *jsgraphics*[256] JavaScript library was used. This library uses modified div-objects to mimick a real drawing canvas, such as seen in HTML5, although with severe limitations. I wrote the code to parse the species-newick string, and generate the associated phylogenetic tree. A key feature in the algorithm is a two-step iteration over the leaves and branches from the tree, in order to assert their respective X and Y coordinates.

7.2.3 WGDotplot

The study of genome evolution by studying by intra- and inter-species colinearity is aided by the popular Whole Genome Dotplot (WGDotplot, see section 3.2.4 and section 4.2.5). In the first release of the PLAZA platform a WGDotplot visualization was implemented using a generated static image together with a clickable map, with two *zoom-levels*: genome-vs-genome and chromosome-vs-chromosome. This approach was found to be lacking and limiting in the next years, for the following reasons:

- A maximum of two species can be compared. Although the program which generates the static image could be changed to deal with a more expansive comparison of collinearity between genomes, the associated increase in query-time and the problematic design issues such as image size, made such changes infeasible.

- Very limited extra onscreen information could be displayed, because of the used technology (clickable map).

- When changing the *zoom-level* a new instance of the image generating program was started, requiring a new set of time-expensive database queries to be executed.

Based on a comparative analysis of different available technologies (Flash, Java, SilverLight and JavaScript), the choice was made to create a Java program to resolve most of these issues, with the implementation done by a thesis student (see Figure 4.6). The original implementation retrieved its data through Remote Function Calls (RFC) with a Java server. Highly problematic performance lead us to consider other avenues for data retrieval. I changed the procedure from the Java RFC architecture to a PHP solution, where the Java Applet retrieves the necessary data from files which were generated (and cached) by the webserver. The raw data, necessary for the WGDotplot applet to function, is between 100kb and 600kb in size. As such, the caching of data is not an issue. Furthermore, all data is cached on a genome-vs-genome basis, leading to a maximum of 625 files (25*25) to be cached. When colinearity between more than two genomes is visualized, the data is merged from the indicated cache.

7.2.4 CirclePlot

While WGDotplots are very powerful, the visualizations associated with intra-species colinearity often lack a certain aesthetic appeal. Circular plots, as popularized by the Circos software, are a common visualization present in many genome publications (see also section 2.4.1 and Figure 2.3). The development of the CirclePlot within the PLAZA framework is fully based on the HTML5 canvas object, with no direct dependencies on external JavaScript libraries. The main features are its interactivity and fast drawing speed. The speed is achieved by smart caching of data (all colinearity data is cached once, irregardless of the number of chromosomes or scaffolds displayed), as well as intelligent break points in the algorithm to suppress spending computer resources on hidden colinear regions. The cached data is presented to the JavaScript program in the form of JSON data, making the subsequent processing quite efficient.

The study of genomic duplications in the evolutionary history of an organism is the focus of much of the intra-species colinear regions available within the PLAZA platform. To facilitate this research, the coloring of colinear regions based on K_S-dating was part of the initial design. With the multitude of duplications present in the history of many plant genomes (see section 2.2.1), the image was however quickly reduced to a garish display of colors for many species (see Figure 7.2 A). As such, features were added to overcome this issue:

- Filtering of colinear regions based on K_S-value (see Figure 7.2 B).

- Filtering out inter- and intra-chromosomal colinearity.

- An alternative coloring scheme which is based on chromosome combinations rather than K_S-values (see Figure 7.2 C).

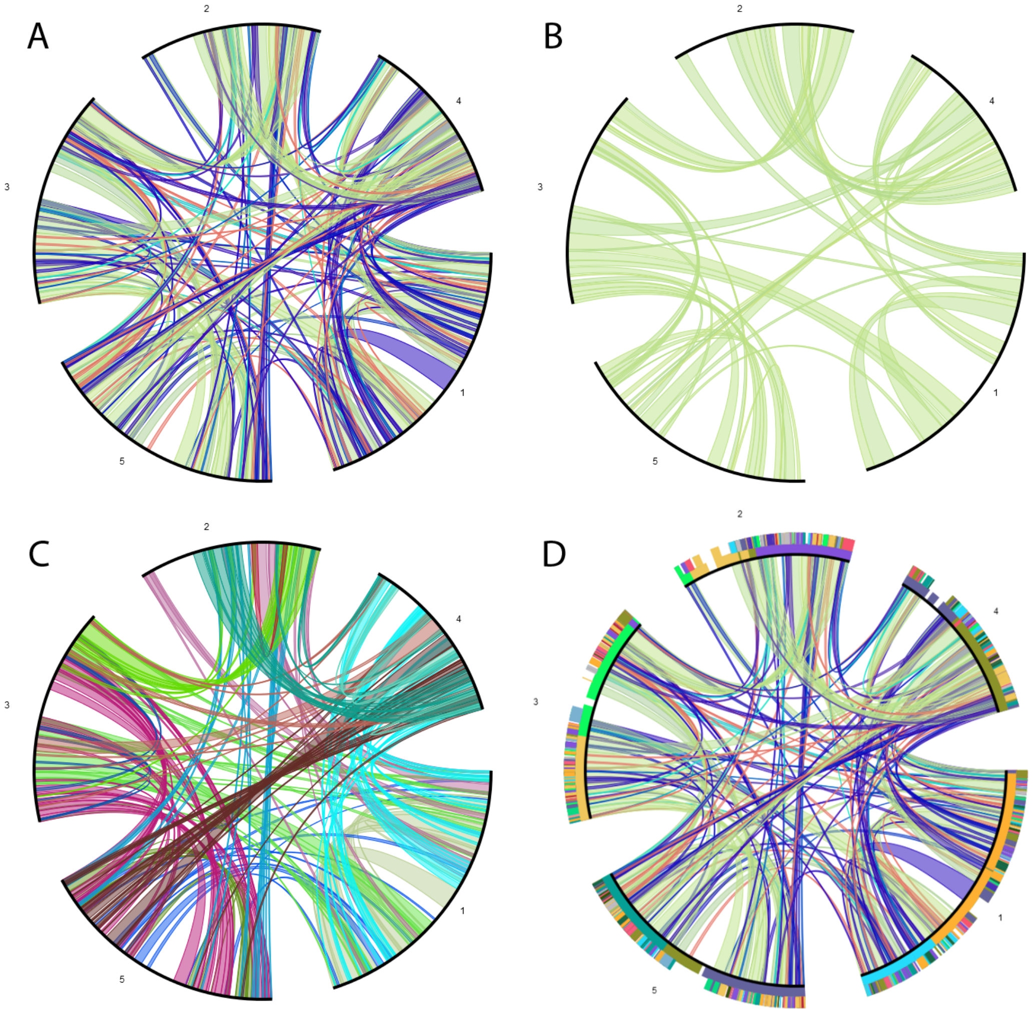


Figure 7.2: Comparison of CirclePlot features using *Arabidopsis thaliana* as reference species.*(A) Basic display of Circleplot. (B) Colinear regions with K_S-values outside the [0.1,1] interval are removed. (C) Alternative color scheme of colinear regions based on chromosomal combinations. (D) Inclusion of inter-species colinearity data.*

In order to replicate the capabilities of the multi-genome WGDotplot as much as possible, we extended the CirclePlot to have inter-species colinearity visualizations as well, mapped onto the outer rim of the displayed chromosomes/scaffolds of the reference species (see Figure 7.2 D). The various colinear regions are color-coded as well, based on chromosome combinations. The current version of the CirclePlot gives one level per extra species, but this approach is problematic when multiple colinear blocks map to the same region.

7.2.5 GO Enrichment plot

One of the most used features within the PLAZA workbench is the GO enrichment tool (later duplicated in the TRAPID platform). The code to perform the enrichment analysis is written by Sebastian Proost.

Initially the results of this enrichment were presented in a standard HTML table, but issues were raised about the interpretability of these results: in a table form the relationships between the GO terms are not discernible anymore. Therefore another JavaScript visualization was developed, which displays both the results of the enrichment and the acyclic graph structure of the GO terms (see also section 2.3.1 and Figure 6.2) and allows interactivity and linking as well.

The development of this visualization required both some inventiveness on the part of data retrieval and data structure organization, as well as on the part of image layout and rendering:

- Because GO terms can have multiple parents in the graph, they can also appear at different *depths*. This does, however, lead to the structural problem of deciding at which depths the GO terms should be drawn, further complicated by the possibility that certain levels within the GO graph will contain many more terms than others for which we would like to compensate.
- Only GO terms which appear in the results, or for which one of the child-terms appear in the results, should be displayed.
- Too many GO terms are often organized on one single depth in the GO graph, and as such some filtering needs to be applied to collapse GO terms. I solved this issue by creating an overlapping field containing all information, when hovering with the mouse pointer over any displayed GO terms (whether they are collapsed or not).

7.3 PLAZA Webserver Organization

The primary code executed when a user visits a page from the PLAZA website is written in PHP. More specifically, I have made heavy use of the open-source CakePHP[257] framework which is based on the *Model-View-Controller* (MVC) design pattern. A clear separation between code which accesses the database and the presentation logic results in far fewer potential security concerns, and enforces the structuring of data into reusable elements.

PHP is however not the only code used for generating webpages: the creation of static images and preprocessing of XML-data is delegated to Java programs, wrapped together with Perl-scripts. There is a clear-cut distinction between these 2 types of code in the architecture of the PLAZA platform (see Figure 7.3):

- PHP code, which is executed by by the Apache webserver. This handles the initial user-requests, and generates the HTML-code for presentation after performing the necessary SQL-queries.
- Other code (mostly Perl and Java), which is called by the PHP code through an XML-based webinterface. This code is generally called to generate images and/or data for display on the PLAZA website.

Both code-bases can connect to the database, and read/write to shared data storage. However, only the PHP-code is called directly by the user, while the other scripts are not accessible, mainly due to security concerns. This architecture has several advantages:

1. **Scalability**. Although at the moment both code-bases are run on the same physical webserver, it would only require a small configuration change to move the Perl and Java code to a separate

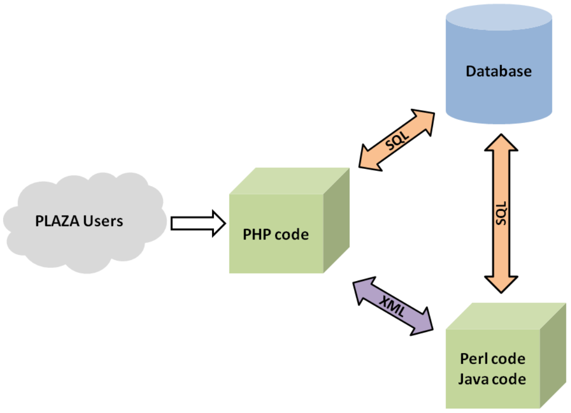


Figure 7.3: PLAZA architecture. *A pictographic representation of the possible interactions within the webserver following a user-request to generate a web page.*

machine. The communication will still be done through a normal HTTP request with XML-data. As such, profiling and balancing of the computational power becomes much easier. The database is already located on a different server, and can as such also be scaled according to potential needs.

2. **Stability**. In case of problematic behavior, or during debuging, it is easy to deduce where the errors are located using this architecture. The main pages of the website will not be affected by problems with the image generating backend of the server.

3. **Version Control and Avoiding Replication**. At the moment there are a multitude of PLAZA versions available: some are public, some are private. For each of these versions, the PHP-code has to be replicated on the server, as it corresponds with a different URL. All these PLAZA versions can however still make use of the same Perl/Java backend.

7.4 Author Contribution

This chapter is not based on any manuscript, but is rather based on the experience gained during the development of the PLAZA and TRAPID platforms. All content in this chapter was written by myself.

“Another flaw in the human character is that everybody wants to build and nobody wants to do maintenance.”

Kurt Vonnegut

8 Discussion and Future Prospects

Abstract

In this chapter we will discuss how the field of genetics will change in the future and how we can deal with these developments, as well as evaluate our work on comparative genomics. A thorough review of our developed tools and methodology can give a glimpse on how to improve some of the design decisions made during the development of the platforms.

Much of the topics will discuss the use of Next Generation Sequencing (NGS) data, because the stunning decrease in cost-per-basepair[10] has several important implications:

- The rate at which plant genomes will become available increases, taxing the comparative genomics platforms to stay up-to-date.
- More data is becoming available per species. Not only is RNA-seq being used to extensively profile the expression under many different conditions[258], in a variety of tissues[258] and in different development stages[258], but it is also revealing a much more detailed view of the transcriptional variety present in genomes[259]. Other data types are becoming increasingly available as well: ChIP-seq data for example is giving insight into the intricate transcriptional regulation of genes by measuring DNA-protein interactions.

For the author contributions, see page 8-13.

8.1 Introduction

The growth rate at which genomic data is currently being produced by the different sequencing centers, most notably JGI[a] [260] and BGI[b], is outpacing the advancements being made in the fields of data capacity and computing power. New *Next Generation Sequencing* (NGS) technologies are constantly being developed and existing ones are refined. Illumina's HiSeq[8] and Roche 454[7] are two examples of technologies which are widely used in current genome and transcriptome projects, and since their introduction (2004-2005) they have been drastically improved, with longer read-lengths making the subsequent assembly more practical and computationally feasible.

With data storage and transfer being such a challenge, this will become even more of an issue with the advent of the so-called third generation NGS technologies: Helicos[261], PacBio[262] and Ion Torrent[263]. These technologies will provide longer read-lengths, making the subsequent assembly much easier, as less memory is required to construct the *de Bruijn graphs*[264] which are currently widely used in the assembly process. However, the memory gains of the longer read-lengths are offset by the constant drive of researchers to sequence ever more large and complex genomes. Other genome sequencing efforts, such as the resequencing of a thousand *Arabidopsis thaliana*[156] genomes, put even more strain on the data storage and processing pipelines.

8.2 Regulatory Genomics

Much of the content within this thesis focuses on the protein coding genes of various genomes. With the very small difference between the coding genes of human and chimp in mind[265], it becomes clear that the regulation of these protein coding genes is as important as their respective content. The Encyclopedia of DNA Elements (ENCODE)[c] [266,267] project for example, is specifically aimed at understanding the regulation of the human genome. The phylogenetic distance between metazoa and plants is quite large, and many regulatory systems will not be translatable. However, the developed techniques and computational tools from the ENCODE project will definitely help the research within the plant community as well. Within the PLAZA framework, it would be highly advisable to not only invest in the comparative analysis of protein coding genes, but to include their regulation as well.

Comparative expression regulation can be studied from the context of how genes are transcribed. To do so, it is imperative that we can measure which transcription factors bind to the promotor of each gene, and under which conditions. With transcription varying widely between tissues[268], developmental stages[268] and stress conditions[269], the possible combinations to perform these measurements are enormous. Once again, comparative genomics can help resolve this issue by studying a single model organism and through translational research.

8.2.1 DNA Binding Motifs

With experimental detection and validation of promotor regions, transcription factors and DNA binding sites a time-consuming, costly and often incomplete undertaking, multiple approaches have been made to computationally predict possible DNA binding sites[270,271]. These sites are, just as their associated coding

[a] http://genome.jgi.doe.gov/

[b] http://www.genomics.cn/en/index

[c] http://encodeproject.org/ENCODE/

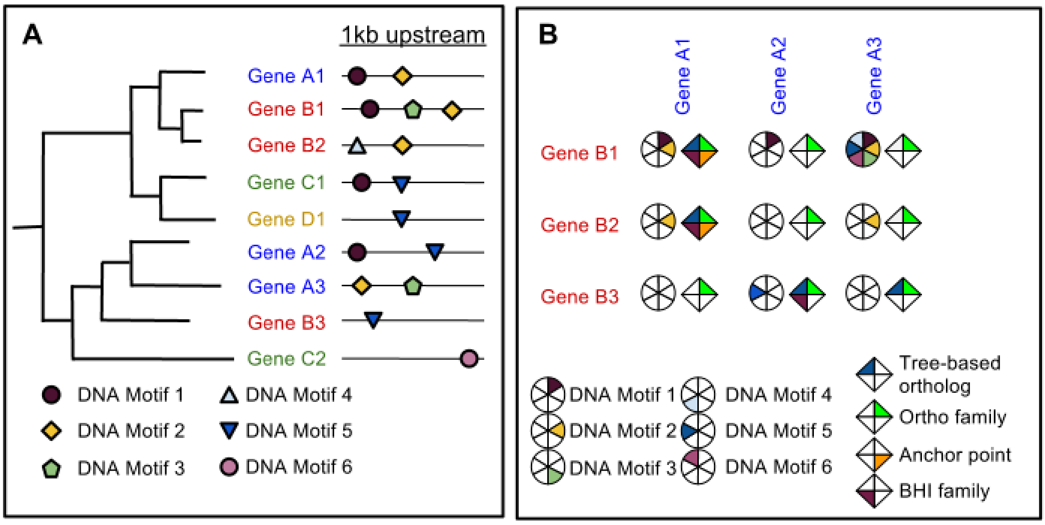


Figure 8.1: Possible use of DNA motifs within the PLAZA framework.*(A)Mapping the motifs to a phylogenetic tree. (B)Extending the Integrative Orthology view with extra DNA motif information.*

sequences, often conserved due to evolutionary pressure, and as such more recent computational techniques try to use these evolutionary constraints in their predictions[272]. The large search-space induced by motif degeneration and large motif lengths makes the task at hand computationally very challenging, and it is questionable whether DNA binding sites will be strongly conserved over very large phylogenetic distances.

The PLAZA platform would benefit greatly from the inclusion of these DNA motifs. Studying these motifs within a gene family can reveal species-specific changes, as well as differences and similarities between the transcriptional regulation of duplications. Aside from the gene family context several other approaches become feasible as well. The functional enrichment studies within a species for genes associated with a certain DNA motif (or sets of DNA motifs) might reveal shared transcriptional regulation between genes, while the reverse (genes with the same DNA motif(s) but with opposite functional annotation) can reveal on/off switches within the transcription networks.

Three possible ways to utilize and visualize these DNA motifs within the PLAZA framework easily come to mind (see Figure 8.1):

- The most familiar way is to map the DNA motifs on a phylogenetic tree of a gene family (see Figure 8.1A). Using this method, it can become apparent where binding sites were gained or lost in evolution.
- The Integrative Orthology approach (see section 4.2.3) is an ensemble method which can be used to infer the most likely functional ortholog in case M-N orthology is present between homologs of two species. The current approach is based solely on the protein sequences of the homologs. By including DNA motifs (see Figure 8.1B) in the Integrative Orthology approach, the transcriptional regulation for the genes is included as well, which should make detecting functional orthologs much easier.

- A table-representation of the overlap between the DNA motifs and functional annotation, such as Gene Ontology terms and InterPro domains. This approach would make it easy to correlate specific binding sites with genomic functions, and in a more advanced stage combinations of binding sites with genomic functions.

8.2.2 (Co-)Expression Data

The micro-array expression data which was gathered and normalized, and subsequently used in the analysis of section 5.2, could potentially be integrated within the PLAZA platform. As such, the co-expression between orthologs could also be seen as an extra evidence type within the Integrative Orthology approach described in section 4.2.3.

The main problem with this approach is that there is no expression data available for all species in the PLAZA database, and the species with expression data exhibit the issue that not all genes are represented on the micro-arrays. Also, the amount of expression experiments per species varies wildly, depending on the amount of scientific and economic interest. This leads to the conclusion that the inclusion of the expression data might present some biases, which will then be exposed to the user. With micro-arrays being slowly replaced by NGS expression technoloqies such as RNA-seq, the picture might shift in the future. It is debatable whether for all species in the PLAZA database such expression data will be made available in sufficient quantities, but for all the new genome projects there always seems to be at least one associated expression study.

Simply replacing the micro-array expression data with RNA-seq data will not be possible, as both technologies have different biases and require different post-processing. However, once such a processing pipeline is constructed, and once enough RNA-seq data has become available, the inclusion of these data types into the platform should become possible.

8.3 Alternative Splicing

Alternative Splicing (AS), the process in which different coding sequences can be produced by a single transcript, is quite prevalent in plant species[273]. Large differences exist in the prevalent types of AS between the plant and animal kingdoms, with *intron retention* being the most common in plants[273] and *exon skipping* the most common in animals[274]. This difference is very likely influenced by the difference in intron lengths between the two kingdoms[274].

In the current state of the PLAZA platform each gene locus is represented by one transcript, implemented by selection of the longest transcript in case several splice variants are present. This simplification of the true coding potential of a single non-spliced messenger RNA molecule was accepted because of several considerations:

- First and foremost is the data availabilty aspect: splice variants were annotated for only the minority (2 out of 9 species) of sequenced plant genomes during the initial development of the PLAZA platform (see chapter 3). As such, paying special attention to splice variants did not measure up to the required effort in software development. Spliced transcript variety has however become more apparent in recent years due to increased interest, and due to the availability of NGS transcriptome data and the dedicated support for splicing variants in current read-mappers[259]. As such, we see a steady increase in the amount of genomes with annotated splice variants, with 9 out of 25 species in PLAZA 2.5 having been annotated with splice variants. Also, subsequent publications[275–279] have shown a growing ratio of genes with splice variants in *Arabidopsis thaliana* (see Figure 8.2). A recent study[280] puts the ratio of genes in *Arabidopsis thaliana* at 61%, at normal growth conditions and DNA extracted from a full-grown plant. With many splice variants being characterized as stress specific or developmental stage specific, this ratio is likely still an underestimation.

- Another reason is that there is the difficulty of balancing the data within the platform. If mRNA transcripts instead of gene loci would be made the basic data entity, several major issues appear:
  - Gene families would no longer be a possible form of clustering. Instead transcript families would be created. This implies that for a single locus, its associated mRNA transcripts could be categorized into different transcript families. If a gene has two splice variants, one with and one without the retention of an intron, than the potential presence of a well-conserved protein domain with that intron might cause the two variants to be clustered differently.
  - There is still no AS data for all species available. Calculating gene family expansions (or rather transcript family expansions) as such becomes non-trivial, as one would need to correct for the absence of AS data. The used nomenclature could also be considered confusing: is a gene with no duplications but with several splice variants still to be considered as single-copy, compared to a gene with no duplications and only a single transcript?
  - Some phylogenetic trees within the PLAZA platform are already very large. Expanding these trees by introducing the splice variants could make them even larger, further reducing the interpretability of these phylogenetic trees. The creation of the multiple sequence alignments and trees might also become biased by the inclusion of multiple splice variants with very little protein sequence diversity (such as induced by an *alternative donor/acceptor* splicing event).
  - The functional annotation (such as GO terms) is in most cases currently assigned to genes, and not transcripts. While the transcriptional regulation of splice variants would have to be similar due to the same upstream DNA binding motifs, the produced proteins can be significantly different.

The Ensemble Plants[d][104] resource does provide some information with regards to splice variants, but many features (such as phylogenetic trees) are still only available at the gene level.

The inclusion of splice variants within the PLAZA platform is, given these challenges with no real solutions yet, not a task to be taken lightly. And extending the PLAZA platform to include splice variants for the sake of completeness with no scientific goal is not appropriate, as other possbile major improvements have also been described in this chapter. However, several studies offer insights into how the protein variation introduced by alternative splicing could be used. Comparative studies of alternative splicing in plants can show how alternative splicing evolves in different plants[281], and how the introduced protein diversity can be beneficial to the organisms[273]. Since both gene duplication and alternative splicing can be considered as two different evolutionary methods to induce protein diversity, the influence on one another can be studied to distinguish different types of subfunctionalization and whether genes with multiple splice variants are more likely to be retained over a gene duplication[282,283]. Tandem duplicated genes are known to be often stress-related, and splice variants are as well[284], indicating a potential avenue of research into making plants more stress resistant.

[d] http://plants.ensembl.org

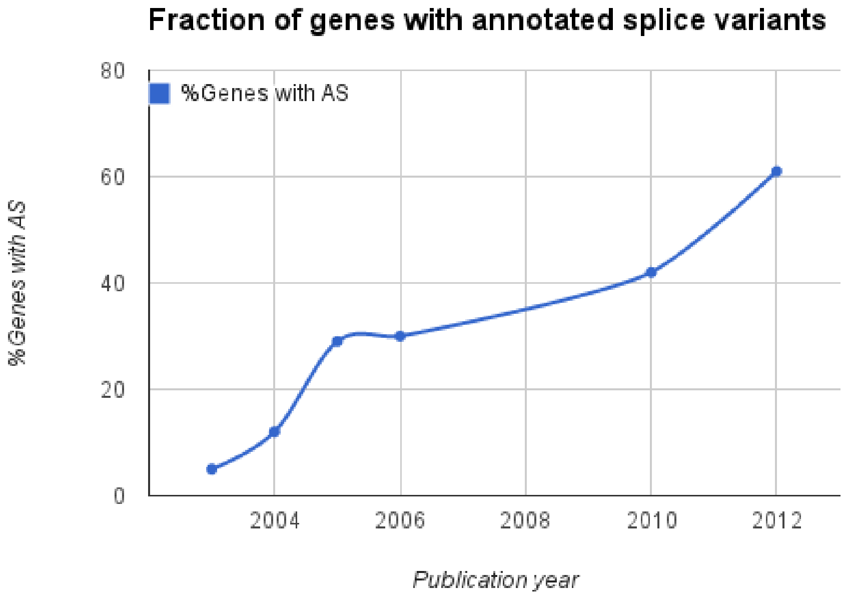


Figure 8.2: Fraction of genes of *Arabidopsis thaliana* for which multiple splice variants have been annotated. *Data collected from*[275–280]

8.4 Data Mining and Availability

8.4.1 PLAZA as a Data Resource

Data from the PLAZA platform is available through the following means on the website:

- Each page representing a specific data type (e.g. GO term, gene family, etc.) contains the functionality to download information directly related to this data type, together with basic structural annotation information.
- Through the workbench (see section 4.2.6) a variety of data types can be accessed for a given set of genes.
- Some tool pages (e.g. WGDotplot) provide the raw data used by these tools to create visualizations.
- The FTP server contains most of the content of the PLAZA database in structured files.

However, this approach indicates a scattering of specific information across the website, over multiple web pages. This is in strong contrast with the BioMart tool[285], which offers a structured and easily understood (although somewhat slow as well) interface through which users can download a wide variety of data. Through query concatenation a broad range of filters can be applied to this downloaded data, further strengthening this approach. Associated with the BioMart tool is an Application Programming Interface (API) which allows bioinformaticians to easily download custom datasets.
Over the years, the PLAZA platform has proven itself to be a valuable resource for standardized genome

information. A substantial amount of the scientific citations of the PLAZA platform are coming from researchers who downloaded the data to perform custom analyzes (e.g. [286,287]). Several requests were made with regards to a more complete access to the database, especially API access is in high demand. It is thus clear that the PLAZA platform is lacking in this regards. The time investment required for implementing such features, both an API and a BioMart-like download access point, forces us to fully review some issues before fully committing ourselves to such a development:

1. Does the PLAZA platform offer any data that is not available from other sources? Or is the data content not unique, but more readily accessible and easily processable compared to other sources?

2. Does making the entire content of the PLAZA database publicly available interfere with research from within the group? Is it ethical to publish a resource and not make the total data content readily available for download?

3. Is it worth to spend time developing these features, delaying the implementation of other data types into the platform? Do we want PLAZA to become more of a data warehouse, competing with bigger consortia such as Ensembl[288] or Phytozome[103] (JGI's plant portal) ?

The answer to the first question is two-fold: most of the data can either be downloaded from third party websites or be generated by the researchers themselves, except for the Integrative Orthologs (see section 4.2.3). However, the ease of use through the standardized data files make PLAZA an attractive solution. The second question is of course more subjective and open to debate. Bringing a next-version of the platform online always takes a respectable amount of time, certainly when new data types are being added. This next version does not need to be immediately publicly available though, giving our research group a head start to use the data for our own purposes. However, I feel that when the platform is made public, than so should all its data content be made fully accesible as well. The third question is the most difficult to answer. It is clear that our small research group cannot compete with larger institutions with personel dedicated to data integration. However, despite the large user-base of these other platforms, a surprising amount of researchers still directly refer to us, indicating that the PLAZA platform is offering something unique.

8.4.2 Automated Data Mining

Since the conception of databases, and the constant growth of the content in databases in the world, several ideas have been put forward to automatically discover either inconsistencies or hidden patterns within the data. One of the major applications is discovering financial discrepancies within and between databases, to combat fraud. Identifying inconsistencies or contradictory data entries are, besides obvious data curation purposes, of limited use within the bioinformatics community. However, finding non-obvious relationships between variables, or combination of variables, within genomic databases could result in important discoveries[289]. The search for functional clusters (see section 4.2.4) within genomes already pointed towards such an application. The complexity and multiple layers within the genomic databases make automating such a discovery procedure very difficult.

To my knowledge, no efforts have thus far been applied to the automatic detection of patterns within genomic databases, to link features and loci together, and as such predict properties for which we have currently no knowledge. All current search solutions are custom, and based on existing ideas and theories on how genetic data is used and organized within living cells. However, by limiting ourselves to these kinds of searches, we are unable to find actual hidden features. Some research has been done on a lower level, for example to extract features from DNA sequences in order to classify those sequences using machine learning methods[290,291]. This feature extraction is still not fully autonomous, although subsequently applied feature selection techniques[292] are. The main problem is still that in these cases the researchers know what they want to classify.

The dimensionality of the data presents a formidable challenge, as do the hidden dependencies such as gene order within a chromosome. As such, a fully automatic procedure will be very difficult to create. While no positive results can be promised, data mining genomic databases might result in useful and unexpected discoveries. The pitfalls of over-analyzing the data can not be overlooked, as multiple testing might result in skewed statistics[293].

8.5 Genome Challenges

With the constant publication of new plant genomes in the past years, and with no decrease in pace in sight, some challenges with the PLAZA platform, or any platform for comparative genomics, become apparent. Although the PLAZA platform and processing pipeline will be able to handle the inclusion of up to 40 genomes pretty well, we should ask the question whether this approach still makes sense. Gene families will be growing (in general) linearly with the number of genomes, making the subsequent analysis pages, such as the phylogenetic trees, quite complicated. Researchers are in most cases only interested in a small subset of genomes, but this subset varies wildly between research fields, and various researchers each have their own *pet genome* they would like to see incorporated and extended within the PLAZA platform.

8.5.1 Subversions of the PLAZA Platform

We have already demonstrated that different versions of the PLAZA platform can co-exist (see section 5.3) and it is clear that this approach might work as well to deal with the constant sequencing of new plant genomes. One solution would be to split up the PLAZA platform into multiple subversions:

- One *Core* high quality PLAZA version, containing a selection of the best annotated and assembled, and most utilized species. Species such as *Arabidopsis thaliana*, *Oryza sativa ssp. japonica*, *Zea mays* and *Populus trichocarpa* are prime examples of this. At the same time, some necessary outgroup species from outside the dicot and monocot clades are necessary, with one of the mosses and algae being the best choices. This core version should be kept relatively small in terms of number of species.

- Subversions for each of the major clades with a sufficient number of sequenced genomes: monocots, dicots and algae. Inclusion of the necessary outgroups for each subversion should still keep the number of species below 20. Except for the dicots, for which a very large number of genomes has been sequened, this limit should not really be a problem. The algae version is already created (coined pico-PLAZA, see section 5.3), and the creation of the other subversions should be relatively easy.

This approach should deal with most of the issues brought forth by the growing number of publicly available sequences. Further refinements can be made, such as benchmarking the quality of the genomes and excluding the genomes which do not meet certain standards. The *Lotus japonicus* genome assembly and gene annotation are for example notoriously sub-standard: with a total number of coding genes of 43,146 only 25,716 (or 6̃0%) of these are not singletons (see Table F.1). With the average of this ratio of all species in PLAZA v2.5 being 8̃7%, it is clear that there are serious issues with this annotation, even when taking possible extreme species-specific gene developments into consideration.

Two problems become apparent with the subversion approach: the duplication of data and the required processing time. Each subversion will be stored in a separate database, and seeing that *Arabidopsis thaliana* is the model species in plant research and will be included in all subversions, this indicates that the *Arabidopsis* genome will be stored at least four times. This may be viewed as a necessary evil, yet potential solutions should be reviewed to remediate this. Revamping the database design in order to store all subversions within one database is one solution, but this is contrary to the principle of data separation. The problem of increased processing time is less of a problem. Although for each version the entire processing pipeline needs to be run, each pipeline instance will take less processing time. And keeping in mind that some of the used algorithms within the pipeline have greater-than-linear time and memory complexity, the subversion approach might in fact keep the increase in processing time at a minimum.

8.5.2 Use of TRAPID Pipeline

The complexity of phylogenetic trees and multiple sequence alignments, induced by the the increase in number of species, can also be reduced by allowing users to create the phylogenetic trees on-the-fly by making use of the developed TRAPID pipelines. The default phylogenetic tree containing all proteins from the gene family should remain available. However, by implementing an intermediate selection screen in which the user can select the species he wants to investigate, the phylogenetic trees can become markable smaller, and thus easier to analyze.

One potential issue is the processing time required to construct these custom phylogenetic trees. Creating the multiple sequence alignment and the tree can take up to several minutes for a medium sized gene family of 200 proteins, depending on options such as bootstraps. If several users launch such requests concurrently, the available power of the web cluster can become exhausted. A solution to this problem would be to move the processing to the client, for example by implementing the used algorithms in Java or JavaScript. This would, however, require quite an investment in development time.

8.6 Conclusion

The scientific world is in constant change, with new technologies rising and old technologies being send to the scrapyard to be forgotten. The data integration aspect of genomic research is of clear importance, both from an academic and economic point of view, yet multiple challenges still remain unsolved. The PLAZA platform for comparative and evolutionary genomics which we created is definitely a good start, but possible extensions still remain to be implemented, in order to be prepared for the challenges of tomorrow. Predicting which features will be of most use to the resarch community is a difficult task, but by communicating extensively with the users a ranking could be made in the development plants.

One thing is certain though, in the foreseeable future there will definitely be work for people working on genomic data integration.

8.7 Author Contribution

All content within this chapter was written by myself.

“Truth is ever to be found in simplicity, and not in the multiplicity and confusion of things.”

Isaac Newton

A
Summary

The accelerating creation within the biological sciences of more data and new data types by innovative technologies gives rise to unparalleled opportunities and challenges. This leads to further specialization in a lot of areas, as more and more background information is required to understand how to process and interpret the growing mountain of data in a correct fashion. Thus, in order to prevent the same data being systematically analyzed over and over again by different experts, to serve the need of scientists over the world, it makes absolute sense to centralize knowledge and processed data. Enhancing the collective cognitive abilities of the scientific community can be achieved by providing centralized repositories which give access to expertly processed data to non-expert users, who can use this data as a starting point for further research.

A.1 Creating a Platform for Comparative and Evolutionary Genomics

In this thesis we have demonstrated how the integration, processing and presentation of genomic data can be achieved, to the benefit of non-expert users. The PLAZA[a] platform is such an implementation, with a core focus on compararive and evolutionary genomics in plants. By presenting structural and functional plant genomic data in a visually appealing way that is easily searchable by interested plant biologists, both academic and applied plant research is strengthened. Furthermore, by making as much of the raw and pre-computed data available as possible in a common file format, bioinformaticians from around the globe can make use of the platform to perform their own large-scale analyzes.

The design of such a platform is no easy task, as is the need for constant support and updates: there is a continuous stream of new and updated genomes and genome annotations which has to be dealt with to stay relevant in the field. Indeed, while pure academic research may focus on only a few model species, the real world applications target economically important crop species, for which there is evermore data available. The PLAZA platform we developed can be considered to be at the forefront of plant genomic research. With a large, returning and growing user-community, our platform is widely known and recognized for its value.

The various tools we have developed for the PLAZA platform, going from colinearity research to gene families, from functional overrepresentation to orthology detection, each have their own research niche, but are also interconnected within the website where necessary.

A.2 Applications of Comparative Genomics

In this thesis we have also shown how we can apply the PLAZA platform and the data contained within to solve meaningful use-cases. By studying gene family expansions we have demonstrated how species can adapt themselves to a particular ecological niche. With two major modes for gene duplication available, tandem and WGD, it is important to note whether and how each type of duplication provides an evolutionary benefit to the organism. Though most genes become pseudogenes after a duplication event, gene family expansions are still a common occurrence within the plant kingdom. As such, the study of the retention of gene duplicates can provide insights into how gene networks can be expanded through duplications without suffering from gene dosage effects.

[a] http://bioinformatics.psb.ugent.be/plaza

Combining PLAZA orthology and expression data gives us the ability to study conserved co-expression between various species, giving insight into how the evolution of orthologs and facilitating translational research. By using co-expression instead of raw expression values the pitfall of incompatible expression data can be circumvented.

By creating a custom algae-oriented PLAZA database, the focused analysis of important microbial eukaryotes becomes possible. The rise of marine environmental metagenomics also requires the necessary reference databases to process the samples. We have shown that our PLAZA platform is also capable of playing a role in these analyses.

A.3 A Look into the Transcriptome NGS Future

Finally, we have shown that there is merit in developing new systems and provide the necessary interactions, rather than over burden existing platforms. Indeed, the TRAPID[b] platform for fast transcriptome analysis clearly shows the value of the PLAZA platform as a reference, while at the same time bringing its own tools to the front to help users analyze their data. This way, we were not restricted by certain conventions of the PLAZA platform, allowing us to gain the necessary speed improvements to very quickly process entire transcriptome datasets, which are becoming ever more commonplace due to the constant rise in NGS production capabilities.

[b] http://bioinformatics.psb.ugent.be/webtools/trapid

*"Iedereen kan nummer één zijn, dat is geen kunst.
De kunst is uit te vinden waarin."*

Midas Dekkers

B
Samenvatting

De toenemende hoeveelheid data die wordt gecreëerd door nieuwe en innovatieve technologiën binnen de genetica, maakt ongekende mogelijkheden en uitdagingen waar. Eén van de problemen bestaat er echter uit dat er ook steeds meer specializatie nodig is, aangezien er meer kennis en training nodig is om de groeiende hoeveelheid data correct te verwerken en te interpreteren. Het is dus nodig om een gecentraliseerde punt te maken waar bepaalde kennis en data wordt opgeslagen, om te voorkomen dat dezelfde data steeds opnieuw wordt verwerkt door experts ten bate van zichzelf en niet-experts. Het verbeteren van de mogelijkheden binnen de wetenschappelijke wereld door deze verwerkte data aan te bieden wordt op deze manier mogelijk, en wetenschappers die geen expert zijn op het gebied van data integratie en verwerking kunnen zo hun eigen onderzoek sneller voortzetten.

B.1 Een Online Platform voor Vergelijkende en Evolutionaire Genoom Studies

In deze thesis hebben we aangetoond hoe de integratie, verwerking en presentatie van genoom information can bereikt worden, ten bate van wetenschappers die geen expert zijn op het gebied van data integratie. Het PLAZA[a] platform is een implementatie voor vergelijkende en evolutionaire planten genoom studies. Zowel academisch als toegepast onderzoek wordt ondersteund door het platform, dat zowel structurele als functionele plant genoom data presenteert op een visueel aantrekkelijke wijze. Het gemakkelijk doorzoekbaar maken van de data is hierbij een extra pluspunt. Door het beschikbaar maken van zoveel mogelijke verwerkte data, kunnen ook andere bioinformatici van de door ons verwerkte data gebruik maken om hun eigen analyzes te doen.

Het ontwerpen van een dergelijk platform dat met de nodige functionaliteit is uitgerust, is geen gemakkelijke opdracht. Een constante toevoer van nieuwe genomen en nieuwe annotaties van reeds verwerkte genomen, maken een constante support en update procedure nodig. De academische wereld blijft voorlopig inderdaad misschien gefocused op model organismen, maar er komt steeds meer en meer data beschikbaar voor economische belangrijke voedsel gewassen. Het PLAZA platform kan beschouwd worden als één van de betere platformen op het gebied van vergelijkende en evolutionaire genoom studies. Er is een groeiende groep gebruikers van over de wereld, waarvan de meesten ook terugkeren om verdere analyzes te doen, hetgeen aangeeft dat ons platform de wijdverspreid gekend is.

B.2 Toepassingen van Vergelijkende Genoom Studies

In deze thesis hebben we aangetoond hoe het PLAZA platform kan gebruikt worden om bepaalde usecases op te lossen. Door het bestuderen van gene families, en hoe deze geëvolueerd en uitgebreid zijn doorheen de tijd, hebben we aangetoond hoe organismen zich hebben aangepast aan hun leefomstandigheden. Door de twee belangrijkste mogelijkheden van gen duplicatie te beschouwen (tandem en WGD), kunnen we aantonen of deze duplicaties een evolutionair voordeel bieden aan het organisme, en hoe we dit kunnen verklaren. Ondanks het feit dat de meeste genen een pseudogen worden na een duplicatie, zijn gene family expansies toch een veelvuldige gebeurtenis bij planten. Het onderzoek naar hoe beide genen van een duplicatie paar bewaard blijven kan us dus inzicht geven hoe gen netwerken zich aanpassen aan deze veranderingen.

[a] http://bioinformatics.psb.ugent.be/plaza

Het combineren van de PLAZA orthologie met expressie data geeft ons de mogelijkheid om evolutionair bewaarde co-expressie te bestuderen. Dit kan ons inzicht verschaffen met betrekking tot de evolutie van orthologen en het kan ons extra mogelijkheden geven met betrekking tot het vertalen van kennis binnen een model organisme naar voedsel gewassen. Door het gebruik van co-expressie kan veel meer expressie data gebruikt worden vergeleken met het gebruik van expressie data waarbij de experimenten moeten overeenkomen op het gebied van weefsel en omstandigheden.

De creatie van een PLAZA versie die gefocused is op algen, maken we het onderzoek naar micro eukaryoten in de oceaan mogelijk. Door het opkomen van *marine metagenomics*, waarbij random samples uit de oceaan worden gesequenced, is er nood aan de nodige verwerking van deze data. We hebben aangetoond dat het gebruik van het pico-PLAZA platform hier een belangrijke rol bij kan spelen.

B.3 Transcriptome Data van NGS Technologiën

We hebben aangetoond dat het mogelijk is een platform te maken voor het vereenvoudigen van transcriptome studies, door PLAZA als een referentie database te gebruiken. Het niet integreren van dit nieuwe platform binnen PLAZA, maar eerder een nieuw platform aanmaken met de nodige connecties, heeft duidelijk zijn voordeel aangetoond. Volledige transcriptoom datasets kunnen nu op korte tijd verwerkt werden door ons online platform, zowel op functioneel gebied als binnen de context van gen families. Op deze wijze dient het PLAZA platform zelf ook niet verandert te worden om met transcriptoom datasets om te kunnen gaan.

"Why waste time learning,
when ignorance is instantaneous? "

Calvin and Hobbes

C
Publications

Publications

6. Moreau, H., Verhelst, B., Couloux, A., Derelle, E., Rombauts, S., Grimsley, N., **Van Bel, M.**, Poulain, J., Katinka, M., Hohmann-Marriott, M., Piganeau, G., Rouz, P., Da Silva, C., Wincker, P., Van de Peer, Y., Vandepoele, K. (2012) *Gene functionalities and genome structure in Bathycoccus prasinos reflect cellular specializations at the base of the green lineage.* Genome Biology 13, R74.

5. Movahedi, S., **Van Bel, M.**, Heyndrickx, KS., Vandepoele, K. (2012) *Comparative co-expression analysis in plant biology.* Plant, Cell & Environment 35, 1787-98.

4. * **Van Bel, M.**, * Proost, S., Wischnitzki, E., Movahedi, S., Scheerlinck, C., Van de Peer, Y., Vandepoele, K. (2012) *Dissecting plant genomes with the PLAZA comparative genomics platform.* Plant Physiol. 158, 590-600. *contributed equally

3. Dessimoz, C., Gabaldon, T., Roos, D. S., Sonnhammer, E., Herrero, J., Altenhoff, A., Apweiler, R., Blake, J., Boeckmann, B., Bridge, A., Bruford, E., Cherry, M., Conte, M., Dannie, D., Datta, R., Entfellner, J., Ebersberger, I., Galperin, M., Joseph, J., Koestler, T., Kriventseva, E., Lecompte, O., Leunissen, J., Lewis, S., Linard, B., Livstone, M., Lu, H., Martin, F., Mazumder, R., Miele, V., Muffato, M., Perriere, G., Punta, M., Rouard, M., Schmitt, T., Schreiber, F., Silva, E.P., Sjolander, K., Skunca, N., Stanley, M., Szklarczyk, R., Thomas, E., Uchiyama, I., **Van Bel, M.**, Vandepoele, K., Vilella, A., Yates, J.R., Zdobnov, E. (2012) *Toward Community Standards in the Quest for Orthologs.* Bioinformatics 28, 900-4.

2. * Proost, S., * **Van Bel, M.**, Sterck, L., Van Parys, T., Van de Peer, Y., Vandepoele, K. (2009) *PLAZA: a comparative genomics resource to study gene and genome evolution in plants.* The Plant Cell 21, 3718-3731. *contributed equally

1. **Van Bel, M.**, Saeys, Y., Van de Peer, Y. (2008) *FunSiP : A Modular and Extensible Classifier for the Prediction of Functional Sites in DNA.* Bioinformatics 24, 1532-3.

“Man has three ways of acting wisely.
First on meditation; that is the noblest.
Second on imitation; that is the easiest.
Thirdly on experience; that is the bitterest.”

Confucius

D
Curriculum Vitae

Contact Information

Michiel Van Bel
Oudenaardsesteenweg 32f
9000 Gent
Belgium

michiel.vanbel@gmail.com or michiel.vanbel@psb.vib-ugent.be
+32 (0)485 41 34 98

Date of birth: September 29, 1982
Place of birth: Antwerp, Belgium

Research Interests

Comparative and Evolutionary Genomics, Processing and Integration of (Biological) Data, Tool and Web development

Education

Ghent University, Ghent, Belgium
Doctor of Science, Bioinformatics October 2006 – December 2012
Promoter: Prof. Dr. Yves Van de Peer
Co-Promoter: Prof. Dr. Klaas Vandepoele

Ghent University, Ghent, Belgium
Master of Science, Computer Science October 2000 – June 2006
Option: Software development
Graduated with Distinction

Websites

PLAZA, an on-line comparative genomics resource
http://bioinformatics.psb.ugent.be/plaza/

TRAPID, rapid analysis of transcriptome data
http://bioinformatics.psb.ugent.be/webtools/trapid/

Selected Conferences

NGS workshop
November 25 – December 05, 2012, UGent/VIB, Ghent, Belgium.
Organizing committee.

SPICY sympozium
March 7, 2012, Wageningen, The Netherlands.
Invited speaker, lecture (workshop).

Quest for Orthologs
June 17 – 19, 2011, EBI, Hinxton, Great Britain.
Presentation.

Comparative & Regulatory Genomics in Plants
April 11 – 15, 2011, UGent, Ghent, Belgium.
Organizing committee, lecture (workshop).

Selected Training

Effective Writing for Life Sciences (by Dr. Jane Fraser)
October 19 – 20, 2010, Ghent, Belgium.

Computer Skills

Familiar with both Microsoft Windows and Linux based systems
Skilled in various programming languages: Perl, Java, C/C++, xHTML, PHP, Javascript, Linux shell scripting, SQL and, LaTeX 2$_{\varepsilon}$

Language Skills

Dutch (mother tongue), English (fluent), French(basic) and, German(basic)

References

Prof. Dr. Yves Van de Peer
Professor
Ghent University
Ghent, BELGIUM
phone: + *32 (0) 9 331 3807*
e-mail:
yves.vandepeer@psb.vib-ugent.be

Prof. Dr. Klaas Vandepoele
Professor
Ghent University
Ghent, BELGIUM
phone: *+32 (0) 9 331 3822*
e-mail:
klaas.vandepoele@psb.vib-ugent.be

“Writers should not fear jargon”

Trevor Quirk

E

List of Abbreviations

List of Abbreviations

API	Application Programming Interface
AS	Alternative Splicing
BHI	Best-Hits-and-Inparalogs
BLAST	Basic Local Alignment Search Tool
DNA	DeoxyriboNucleic Acid
EBI	European Bioinformatics Institute
ECC	Expression Context Conservation
ENCODE	Encyclopedia of DNA Elements
EST	Expressed Sequence Tag
EXP	Inferred from Experiment
Gb	Giga-base: 1 billion basepairs
GMO	Genetically Modified Organism
GOC	Gene Ontology Consortium
GUI	Graphical User Interface
HMM	Hidden Markov Model
IEA	Inferred from Electronic Annotation
IEA	Inferred from Electronic Annotation
ISS	Inferred from Sequence or Structural Similarity
JGI	Joint Genome Initiative
JVM	Java Virtual Machine
K_S	Synonymous substitution rate
KAAS	KEGG Automatic Annotation Server
KEGG	Kyoto Encyclopedia of Genes and Genomes

Mb	Mega-base: 1 million basepairs
ML	Machine Learning
MSA	Multiple Sequence Alignment
MVC	Model View Controller
NCBI	National Center for Biotechnology Information http://www.ncbi.nlm.nih.gov/
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
ORF	Open Reading Frame
PCC	Pearson correlation coefficient
PWM	Positional Weight Matrix
RBH	Reciprocal best BLAST-hit
RFC	Remote Function Calls
RNA	RiboNucleic Acid
SVG	Scalable Vector Graphics
WGD	Whole Genome Duplication

"A picture is worth a thousand words"

proverb

F

Supplementary Figures and Tables

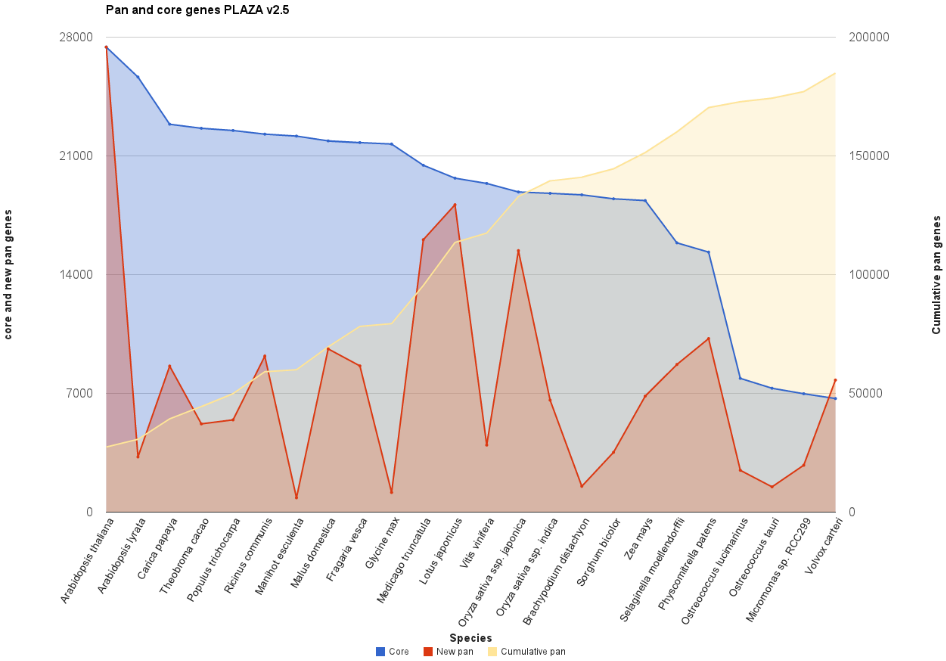


Figure F.1: Pan and core genes of all organisms present in PLAZA v2.5. *The pan genes (homologs present in a at least one species) and the core genes (homologs present in all species) for all species from PLAZA v2.5.*

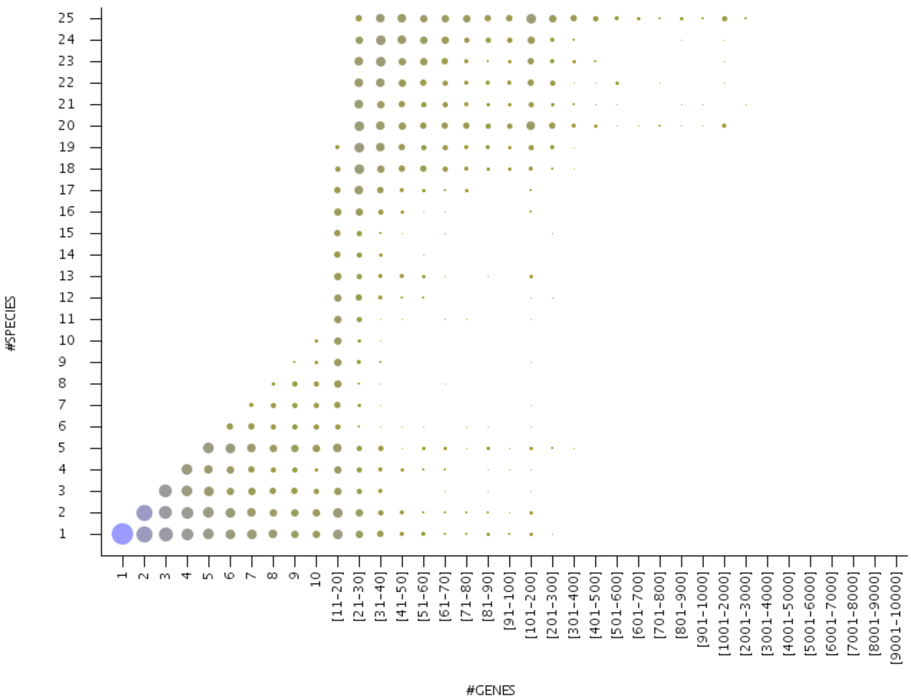


Figure F.2: Gene family sizes in PLAZA v2.5 *The size of the circles indicate the number of gene families with X genes from Y species. The color indicates the fraction of the total amount of gene families.*

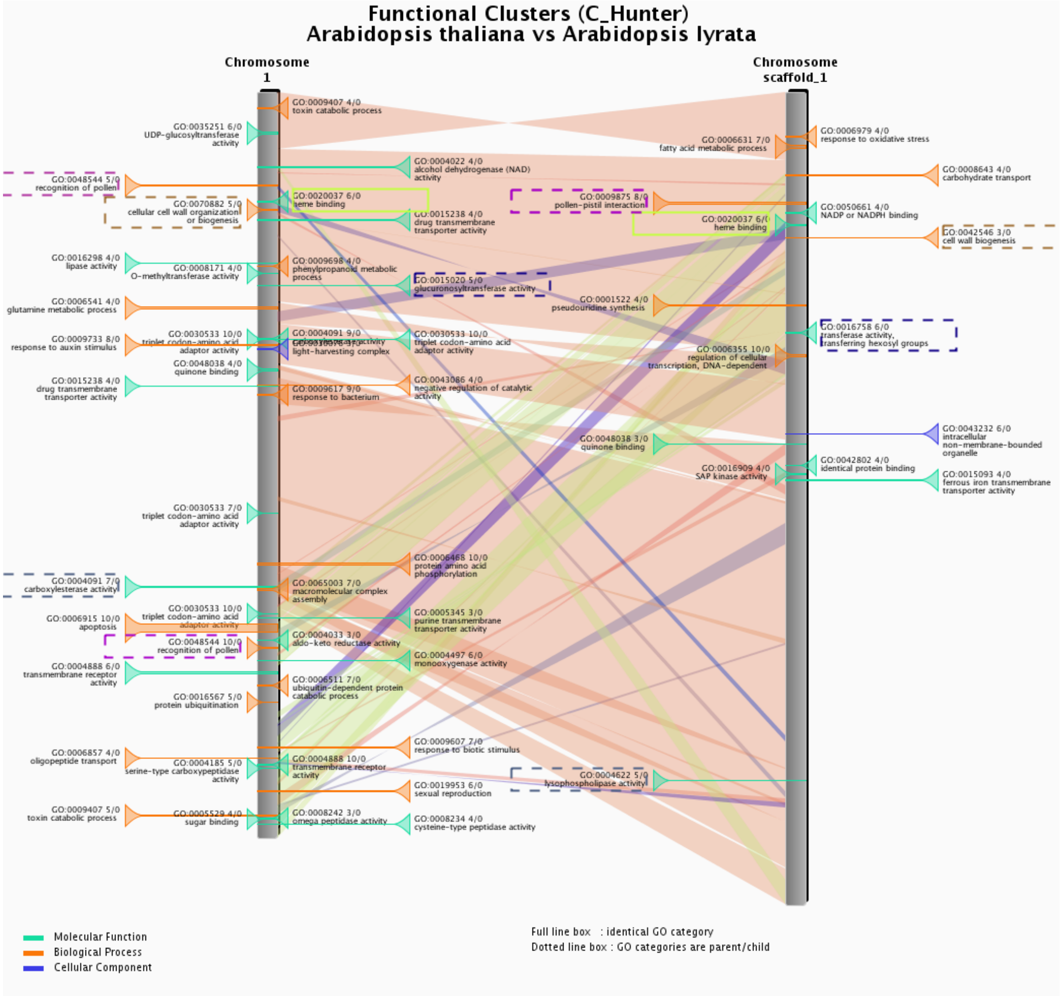

Functional Clusters (C_Hunter)
Arabidopsis thaliana vs Arabidopsis lyrata
Chromosome 1
Chromosome scaffold_1
GO:0009407 4/0 toxin catabolic process
GO:0035251 6/0 UDP-glucosyltransferase activity
GO:0004022 4/0 alcohol dehydrogenase (NAD) activity
GO:0048544 5/0 recognition of pollen
GO:0070882 5/0 cellular cell wall organization or biogenesis
GO:0020037 6/0 heme binding
GO:0015238 4/0 drug transmembrane transporter activity
GO:0016298 4/0 lipase activity
GO:0008171 4/0 O-methyltransferase activity
GO:0009698 4/0 phenylpropanoid metabolic process
GO:0015020 5/0 glucuronosyltransferase activity
GO:0006541 4/0 glutamine metabolic process
GO:0009733 8/0 response to auxin stimulus
GO:0030533 10/0 triplet codon-amino acid adaptor activity
GO:0004091 9/0 carboxylesterase activity
GO:0030076 3/0 light-harvesting complex
GO:0030533 10/0 triplet codon-amino acid adaptor activity
GO:0048038 4/0 quinone binding
GO:0015238 4/0 drug transmembrane transporter activity
GO:0009617 9/0 response to bacterium
GO:0043086 4/0 negative regulation of catalytic activity
GO:0030533 7/0 triplet codon-amino acid adaptor activity
GO:0006468 10/0 protein amino acid phosphorylation
GO:0004091 7/0 carboxylesterase activity
GO:0065003 7/0 macromolecular complex assembly
GO:0006915 10/0 apoptosis
GO:0030533 10/0 triplet codon-amino acid adaptor activity
GO:0005345 3/0 purine transmembrane transporter activity
GO:0048544 10/0 recognition of pollen
GO:0004033 3/0 aldo-keto reductase activity
GO:0004888 6/0 transmembrane receptor activity
GO:0004497 6/0 monooxygenase activity
GO:0006511 7/0 ubiquitin-dependent protein catabolic process
GO:0016567 5/0 protein ubiquitination
GO:0009607 7/0 response to biotic stimulus
GO:0006857 4/0 oligopeptide transport
GO:0004185 5/0 serine-type carboxypeptidase activity
GO:0004888 10/0 transmembrane receptor activity
GO:0019953 6/0 sexual reproduction
GO:0009407 5/0 toxin catabolic process
GO:0005529 4/0 sugar binding
GO:0008242 3/0 omega peptidase activity
GO:0008234 4/0 cysteine-type peptidase activity
GO:0006979 4/0 response to oxidative stress
GO:0006631 7/0 fatty acid metabolic process
GO:0008643 4/0 carbohydrate transport
GO:0009875 8/0 pollen-pistil interaction
GO:0050661 4/0 NADP or NADPH binding
GO:0020037 6/0 heme binding
GO:0042546 3/0 cell wall biogenesis
GO:0001522 4/0 pseudouridine synthesis
GO:0016758 6/0 transferase activity, transferring hexosyl groups
GO:0006355 10/0 regulation of cellular transcription, DNA-dependent
GO:0043232 6/0 intracellular non-membrane-bounded organelle
GO:0048038 3/0 quinone binding
GO:0042802 4/0 identical protein binding
GO:0016909 4/0 SAP kinase activity
GO:0015093 4/0 ferrous iron transmembrane transporter activity
GO:0004622 5/0 lysophospholipase activity
Molecular Function
Biological Process
Cellular Component
Full line box : identical GO category
Dotted line box : GO categories are parent/child

Figure F.3: Functional clusters and colinearity *The colinear regions between Arabidopsis thaliana chromosome 1 and Arabidopsis lyrata scaffold 1, together with their respective functional clusters. The conserved clusters are indicated by dotted boxes, indicating the presence of the cluster in the common ancestor.*

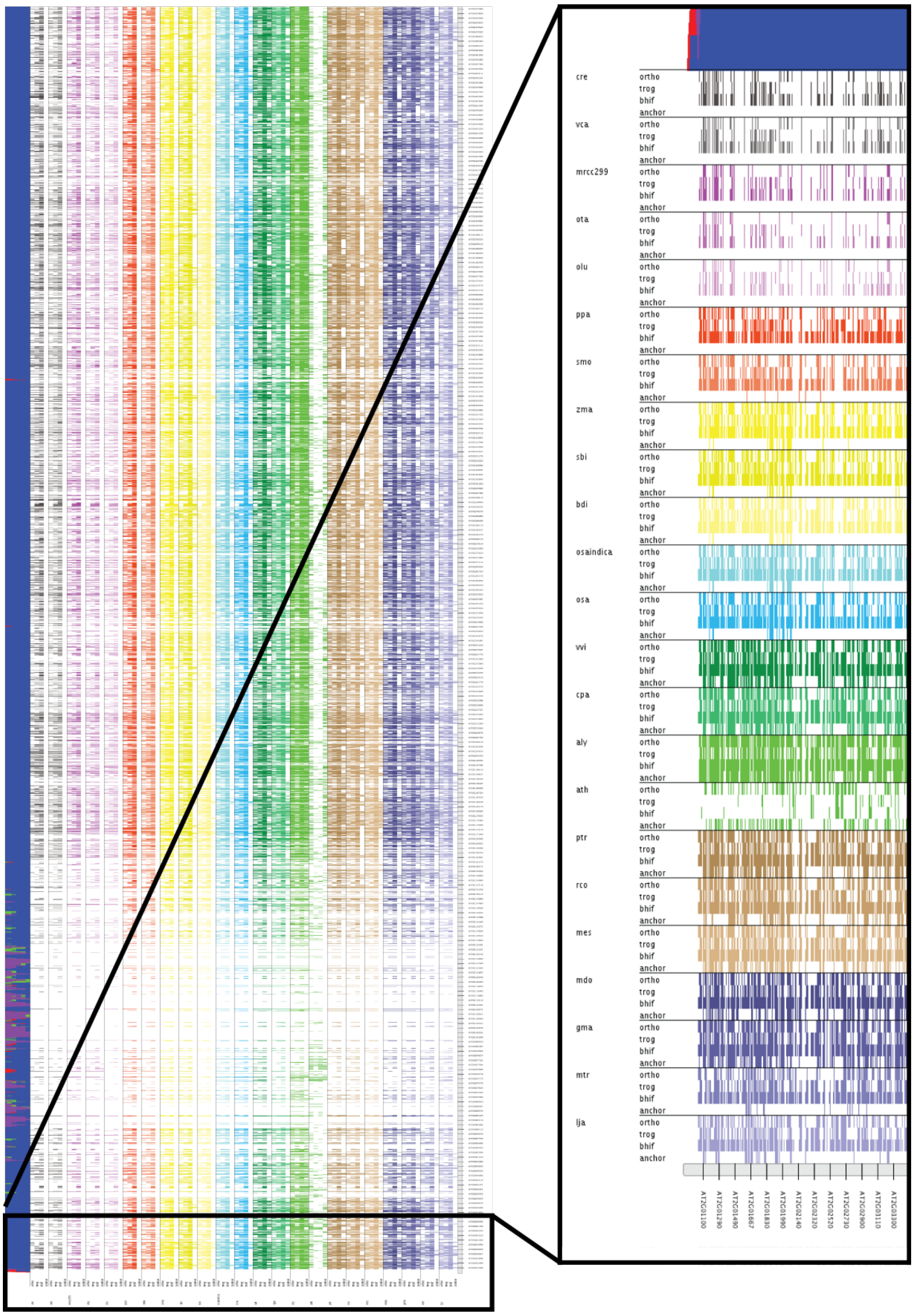
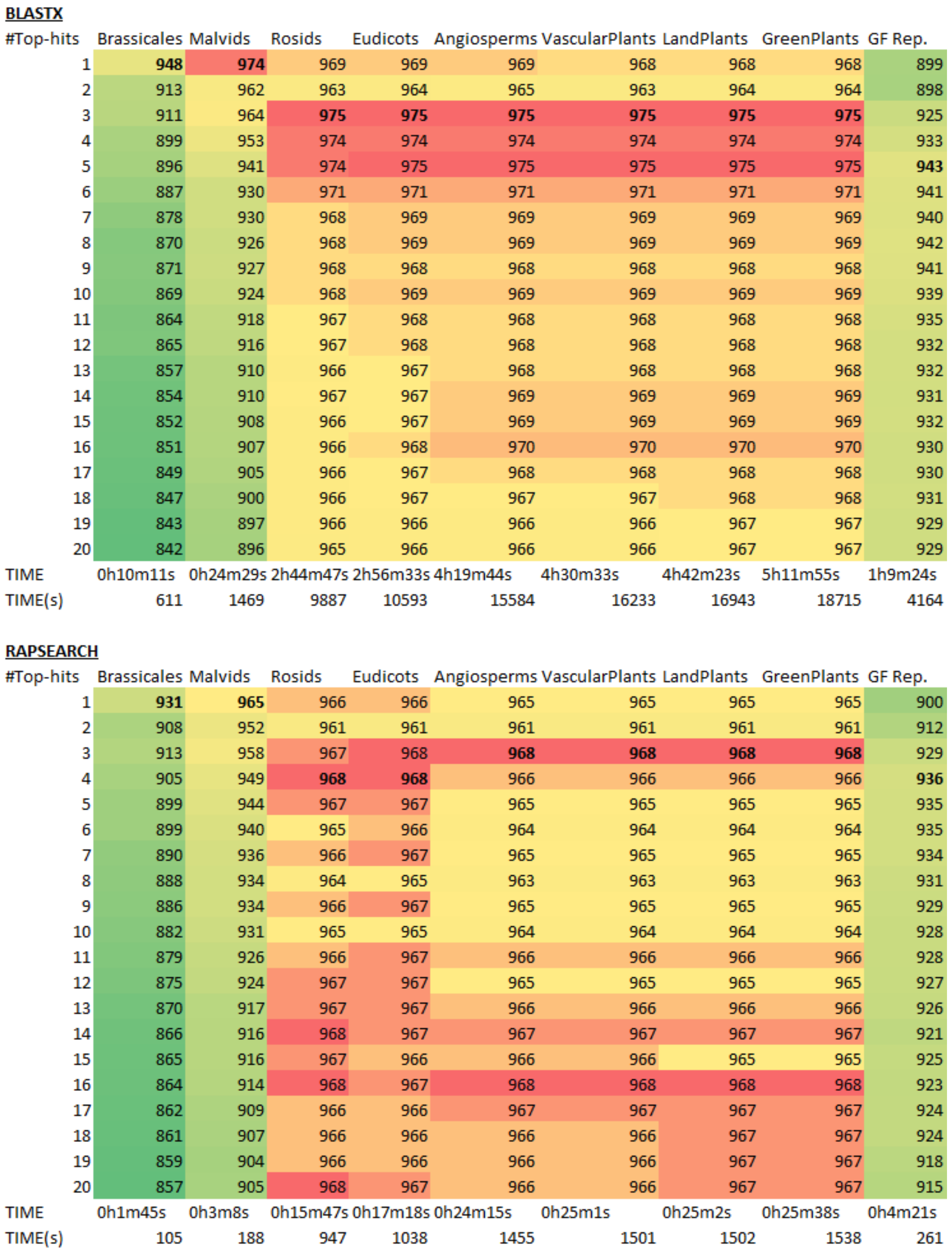


Figure F.4: Integrative Orthology support in *Arabidopsis thaliana* chromosome 2. *The Integrative Orthology support per gene in chromosome 2 of Arabidopsis thaliana. Each species has its own color, and each species track is subdivided in four lanes which correspond with the evidence types in the Integrative Orthology method. The top track indicates the gene type fraction within a 100 gene window on that position: blue for coding genes, purple for transposons, red for RNA genes and green for pseudo genes.*

Species	Genes	Coding	RNA	Pseudo	TE (a)	Genes in non singleton GF (b)
Arabidopsis lyrata	32,670	32,670	0	0	0	30,870 (94.5%)
Arabidopsis thaliana	33,602	27,416	1,359	924	3,903	26,118 (95.3%)
Brachypodium distachyon	26,678	26,632	46	0	0	25,687 (96.5%)
Carica papaya	28,072	28,027	45	0	0	22,531 (80.4%)
Chlamydomonas reinhardtii	16,841	16,788	53	0	0	13,666 (81.4%)
Fragaria vesca	34,809	34,809	0	0	0	30,833 (88.6%)
Glycine max	46,509	46,464	45	0	0	45,982 (98.9%)
Lotus japonicus	69,647	43,146	45	0	26,456	25,716 (59.6%)
Malus domestica	95,23	63,546	0	0	31,684	58,790 (92.5%)
Manihot esculenta	30,800	30,748	52	0	0	30,132 (98.0%)
Medicago truncatula	57,587	45,197	776	0	11,614	38,494 (85.2%)
Micromonas sp. RCC299	10,276	10,204	72	0	0	8,144 (79.8%)
Oryza sativa ssp. indica	59,43	49,202	39	0	10,189	44,310 (90.1%)
Oryza sativa ssp. japonica	57,874	42,211	92	0	15,571	37,391 (88.6%)
Ostreococcus lucimarinus	7,805	7,805	0	0	0	7,408 (94.9%)
Ostreococcus tauri	8,116	7,994	122	0	0	6,797 (85.0%)
Physcomitrella patens	36,137	28,097	72	0	7,968	21,287 (75.8%)
Populus trichocarpa	41,521	41,476	45	0	0	37,777 (91.1%)
Ricinus communis	31,221	31,221	0	0	0	24,455 (78.3%)
Selaginella moellendorffii	22,285	22,285	0	0	0	17,392 (78.0%)
Sorghum bicolor	34,686	34,609	77	0	0	31,921 (92.2%)
Theobroma cacao	46,269	28,882	45	0	17,342	27,575 (95.5%)
Vitis vinifera	26,644	26,504	88	52	0	23,268 (87.8%)
Volvox carteri	15,544	15,544	0	0	0	13,782 (88.7%)
Zea mays	39,597	39,19	0	323	84	35,221 (89.9%)

Table F.1: Gene types PLAZA v2.5. *(a) Transposable Elements. (b) GF=Gene Families. Values between brackets indicate percentage (using number of coding genes as denominator).*



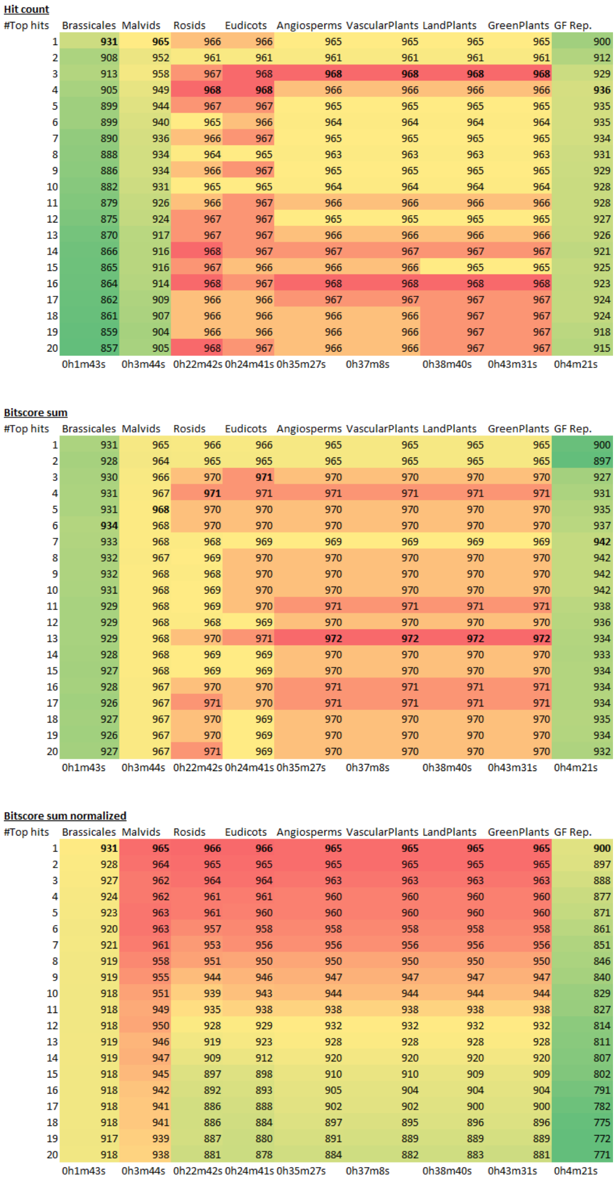
BLASTX

#Top-hits	Brassicales	Malvids	Rosids	Eudicots	Angiosperms	VascularPlants	LandPlants	GreenPlants	GF Rep.
1	**948**	**974**	969	969	969	968	968	968	899
2	913	962	963	964	965	963	964	964	898
3	911	964	**975**	**975**	**975**	**975**	**975**	**975**	925
4	899	953	974	974	974	974	974	974	933
5	896	941	974	975	975	975	975	975	**943**
6	887	930	971	971	971	971	971	971	941
7	878	930	968	969	969	969	969	969	940
8	870	926	968	969	969	969	969	969	942
9	871	927	968	968	968	968	968	968	941
10	869	924	968	969	969	969	969	969	939
11	864	918	967	968	968	968	968	968	935
12	865	916	967	968	968	968	968	968	932
13	857	910	966	967	968	968	968	968	932
14	854	910	967	967	969	969	969	969	931
15	852	908	966	967	969	969	969	969	932
16	851	907	966	968	970	970	970	970	930
17	849	905	966	967	968	968	968	968	930
18	847	900	966	967	967	967	968	968	931
19	843	897	966	966	966	966	967	967	929
20	842	896	965	966	966	966	967	967	929
TIME	0h10m11s	0h24m29s	2h44m47s	2h56m33s	4h19m44s	4h30m33s	4h42m23s	5h11m55s	1h9m24s
TIME(s)	611	1469	9887	10593	15584	16233	16943	18715	4164

RAPSEARCH

#Top-hits	Brassicales	Malvids	Rosids	Eudicots	Angiosperms	VascularPlants	LandPlants	GreenPlants	GF Rep.
1	**931**	**965**	966	966	965	965	965	965	900
2	908	952	961	961	961	961	961	961	912
3	913	958	967	968	**968**	**968**	**968**	**968**	929
4	905	949	**968**	**968**	966	966	966	966	**936**
5	899	944	967	967	965	965	965	965	935
6	899	940	965	966	964	964	964	964	935
7	890	936	966	967	965	965	965	965	934
8	888	934	964	965	963	963	963	963	931
9	886	934	966	967	965	965	965	965	929
10	882	931	965	965	964	964	964	964	928
11	879	926	966	967	966	966	966	966	928
12	875	924	967	967	965	965	965	965	927
13	870	917	967	967	966	966	966	966	926
14	866	916	968	967	967	967	967	967	921
15	865	916	967	966	966	966	965	965	925
16	864	914	968	967	968	968	968	968	923
17	862	909	966	966	967	967	967	967	924
18	861	907	966	966	966	966	967	967	924
19	859	904	966	966	966	966	967	967	918
20	857	905	968	967	966	966	967	967	915
TIME	0h1m45s	0h3m8s	0h15m47s	0h17m18s	0h24m15s	0h25m1s	0h25m2s	0h25m38s	0h4m21s
TIME(s)	105	188	947	1038	1455	1501	1502	1538	261

Table F.2: BLASTX vs. RAPSearch2 comparison. *Dataset consists of a set of 1000 full-length Arabidopsis thaliana CDS sequences. The similarity search databases are based on PLAZA 2.5 phylogenetic clades, but do not contain Arabidopsis thaliana or Arabidopsis lyrata genes. Performance is measured by comparing how well the assigned gene families coincide with the default PLAZA gene families. The comparison was performed on the same single core CPU. Red colors indicate the best values, green colors the worst.*



Hit count

#Top hits	Brassicales	Malvids	Rosids	Eudicots	Angiosperms	VascularPlants	LandPlants	GreenPlants	GF Rep.
1	**931**	**965**	966	966	965	965	965	965	900
2	908	952	961	961	961	961	961	961	912
3	913	958	967	968	**968**	**968**	**968**	**968**	929
4	905	949	**968**	**968**	966	966	966	966	**936**
5	899	944	967	967	965	965	965	965	935
6	899	940	965	966	964	964	964	964	935
7	890	936	966	967	965	965	965	965	934
8	888	934	964	965	963	963	963	963	931
9	886	934	966	967	965	965	965	965	929
10	882	931	965	965	964	964	964	964	928
11	879	926	966	967	966	966	966	966	928
12	875	924	967	967	965	965	965	965	927
13	870	917	967	967	966	966	966	966	926
14	866	916	968	967	967	967	967	967	921
15	865	916	967	966	966	966	965	965	925
16	864	914	968	967	968	968	968	968	923
17	862	909	966	966	967	967	967	967	924
18	861	907	966	966	966	966	967	967	924
19	859	904	966	966	966	966	967	967	918
20	857	905	968	967	966	966	967	967	915
	0h1m43s	0h3m44s	0h22m42s	0h24m41s	0h35m27s	0h37m8s	0h38m40s	0h43m31s	0h4m21s

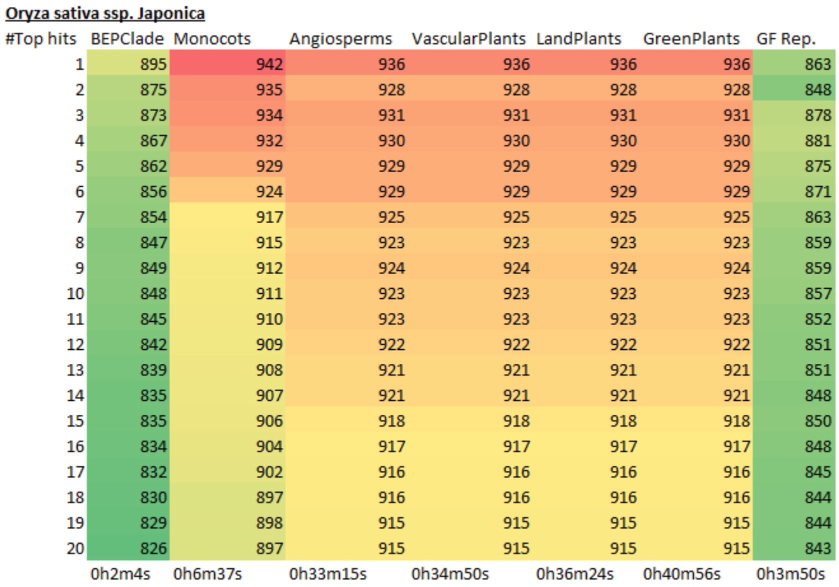
Bitscore sum

#Top hits	Brassicales	Malvids	Rosids	Eudicots	Angiosperms	VascularPlants	LandPlants	GreenPlants	GF Rep.
1	931	965	966	966	965	965	965	965	900
2	928	964	965	965	965	965	965	965	897
3	930	966	970	**971**	970	970	970	970	927
4	931	967	**971**	971	971	971	971	971	931
5	931	**968**	970	970	970	970	970	970	935
6	**934**	968	970	970	970	970	970	970	937
7	933	968	968	969	969	969	969	969	**942**
8	932	967	969	970	970	970	970	970	942
9	932	968	968	970	970	970	970	970	942
10	931	968	969	970	970	970	970	970	942
11	929	968	969	970	971	971	971	971	938
12	929	968	968	969	970	970	970	970	936
13	929	968	970	971	**972**	**972**	**972**	**972**	934
14	928	968	969	969	970	970	970	970	933
15	927	968	969	969	970	970	970	970	934
16	928	967	970	970	971	971	971	971	934
17	926	967	971	970	971	971	971	971	934
18	927	967	970	969	970	970	970	970	935
19	926	967	970	969	970	970	970	970	934
20	927	967	971	969	970	970	970	970	932
	0h1m43s	0h3m44s	0h22m42s	0h24m41s	0h35m27s	0h37m8s	0h38m40s	0h43m31s	0h4m21s

Bitscore sum normalized

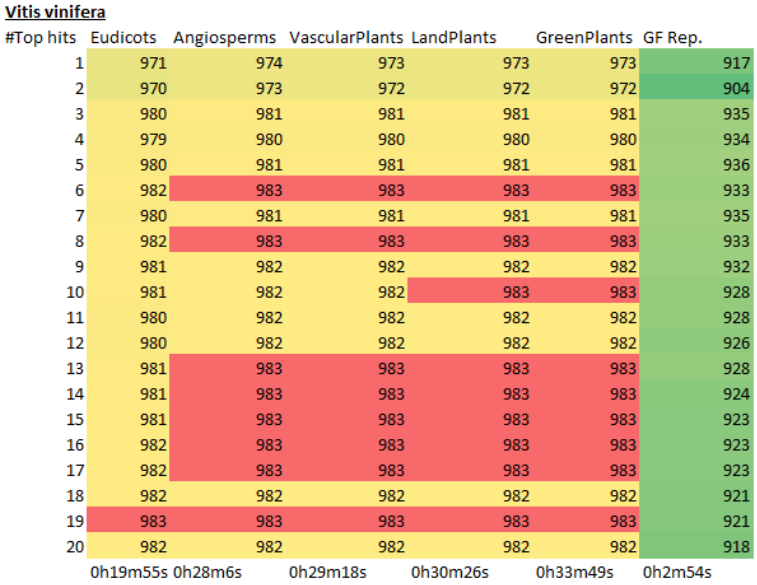
#Top hits	Brassicales	Malvids	Rosids	Eudicots	Angiosperms	VascularPlants	LandPlants	GreenPlants	GF Rep.
1	**931**	**965**	**966**	**966**	**965**	**965**	**965**	**965**	**900**
2	928	964	965	965	965	965	965	965	897
3	927	962	964	964	963	963	963	963	888
4	924	962	961	961	960	960	960	960	877
5	923	963	961	960	960	960	960	960	871
6	920	963	957	958	958	958	958	958	861
7	921	961	953	956	956	956	956	956	851
8	919	958	951	950	950	950	950	950	846
9	919	955	944	946	947	947	947	947	840
10	918	951	939	943	944	944	944	944	829
11	918	949	935	938	938	938	938	938	827
12	918	950	928	929	932	932	932	932	814
13	919	946	919	923	928	928	928	928	811
14	919	947	909	912	920	920	920	920	807
15	918	945	897	898	910	910	909	909	802
16	918	942	892	893	905	904	904	904	791
17	918	941	886	888	902	902	900	900	782
18	918	941	886	884	897	895	896	896	775
19	917	939	887	880	891	889	889	889	772
20	918	938	881	878	884	882	883	881	771
	0h1m43s	0h3m44s	0h22m42s	0h24m41s	0h35m27s	0h37m8s	0h38m40s	0h43m31s	0h4m21s

Table F.3: Evaluation metrics for gene family assignment. *Different evaluation metrics where used to assign transcripts to gene families, using RAPSearch2. The* Hit count *metric assigns the transcript to the gene family with most similarity hits, while the bitscore metrics use the sum of the bitscores of the similarity hits to reach a conclusion. Dataset consists of a set of 1000 full-length Arabidopsis thaliana CDS sequences. The similarity search databases are based on PLAZA 2.5 phylogenetic clades, but do not contain Arabidopsis thaliana or Arabidopsis lyrata genes. Performance is measured by comparing how well the assigned gene families coincide with the default PLAZA gene families. The comparison was performed on the same single core CPU. Red colors indicate the best values, green colors the worst.*



Oryza sativa ssp. Japonica

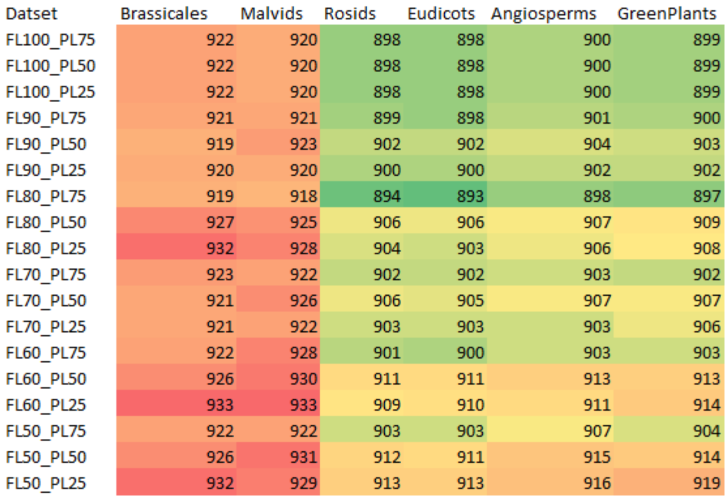
#Top hits	BEPClade	Monocots	Angiosperms	VascularPlants	LandPlants	GreenPlants	GF Rep.
1	895	942	936	936	936	936	863
2	875	935	928	928	928	928	848
3	873	934	931	931	931	931	878
4	867	932	930	930	930	930	881
5	862	929	929	929	929	929	875
6	856	924	929	929	929	929	871
7	854	917	925	925	925	925	863
8	847	915	923	923	923	923	859
9	849	912	924	924	924	924	859
10	848	911	923	923	923	923	857
11	845	910	923	923	923	923	852
12	842	909	922	922	922	922	851
13	839	908	921	921	921	921	851
14	835	907	921	921	921	921	848
15	835	906	918	918	918	918	850
16	834	904	917	917	917	917	848
17	832	902	916	916	916	916	845
18	830	897	916	916	916	916	844
19	829	898	915	915	915	915	844
20	826	897	915	915	915	915	843
	0h2m4s	0h6m37s	0h33m15s	0h34m50s	0h36m24s	0h40m56s	0h3m50s



Vitis vinifera

#Top hits	Eudicots	Angiosperms	VascularPlants	LandPlants	GreenPlants	GF Rep.
1	971	974	973	973	973	917
2	970	973	972	972	972	904
3	980	981	981	981	981	935
4	979	980	980	980	980	934
5	980	981	981	981	981	936
6	982	983	983	983	983	933
7	980	981	981	981	981	935
8	982	983	983	983	983	933
9	981	982	982	982	982	932
10	981	982	982	983	983	928
11	980	982	982	982	982	928
12	980	982	982	982	982	926
13	981	983	983	983	983	928
14	981	983	983	983	983	924
15	981	983	983	983	983	923
16	982	983	983	983	983	923
17	982	983	983	983	983	923
18	982	982	982	982	982	921
19	983	983	983	983	983	921
20	982	982	982	982	982	918
	0h19m55s	0h28m6s	0h29m18s	0h30m26s	0h33m49s	0h2m54s

Table F.4: Evaluation of the gene family assignments for other species. *Using the RAPSearch2 software, different data sets using different species were tested for the gene family assignments. Dataset consists of a set of 1000 full-length Oryza sativa ssp. japonica and a 1000 full-length Vitis vinifera CDS sequences. The similarity search databases are based on PLAZA 2.5 phylogenetic clades, but do not contain Oryza sativa ssp. japonica or Oryza sativa ssp. indica genes for the first data set, and no Vitis vinifera for the second data set . Performance is measured by comparing how well the assigned gene families coincide with the default PLAZA gene families. The comparison was performed on the same single core CPU. Red colors indicate the best values, green colors the worst.*



Datset	Brassicales	Malvids	Rosids	Eudicots	Angiosperms	GreenPlants
FL100_PL75	922	920	898	898	900	899
FL100_PL50	922	920	898	898	900	899
FL100_PL25	922	920	898	898	900	899
FL90_PL75	921	921	899	898	901	900
FL90_PL50	919	923	902	902	904	903
FL90_PL25	920	920	900	900	902	902
FL80_PL75	919	918	894	893	898	897
FL80_PL50	927	925	906	906	907	909
FL80_PL25	932	928	904	903	906	908
FL70_PL75	923	922	902	902	903	902
FL70_PL50	921	926	906	905	907	907
FL70_PL25	921	922	903	903	903	906
FL60_PL75	922	928	901	900	903	903
FL60_PL50	926	930	911	911	913	913
FL60_PL25	933	933	909	910	911	914
FL50_PL75	922	922	903	903	907	904
FL50_PL50	926	931	912	911	915	914
FL50_PL25	932	929	913	913	916	919

Table F.5: Evaluation of gene family assignments for partial transcripts. *Using six different data sets, each comprising 1000 Arabidopsis thaliana CDS sequences, the sensitivity to partial transcripts was measured. The first data set comprises 1000 full-length CDS sequences, the second data set 900 full-length CDS sequences and 100 partial sequences, the third data set 800 full-length CDS sequences and 200 partial sequences, ... , the last data set 500 full-length CDS sequences and 500 partial sequences. These data sets are indicated with the prefix FL (Full Length). For each data set, three different types of partial sequences where generated, where 75%, 50% or 25% of the original CDS length is retained, indicated with the prefix PL (Partial Length). The similarity search databases are based on PLAZA 2.5 phylogenetic clades, but do not contain Arabidopsis thaliana or Arabidopsis lyrata genes. Performance is measured by comparing how well the assigned gene families coincide with the default PLAZA gene families. The comparison was performed on the same single core CPU. Red colors indicate the best values, green colors the worst.*

> *"I received the fundamentals of my education in school, but that was not enough. My real education, the superstructure, the details, the true architecture, I got out of the public library."*
>
> Isaac Asimov

G
Bibliography

Bibliography

[1] J.D. Watson and F.H.C. Crick. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.

[2] G. Mendel. Versuche ber pflanzen-hybriden. *Verh. Naturforsch. Ver. Brnn*, 4:3–47, 1866.

[3] T. H. Morgan. *Evolution and genetics*. Princeton University Press, 1925.

[4] C. Darwin. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, 1859.

[5] W. Min Jou, G. Haegeman, M. Ysebaert, and W. Fiers. Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature*, 237(5350):82–88, May 1972.

[6] F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, C. A. Fiddes, C. A. Hutchison, P. M. Slocombe, and M. Smith. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, 265(5596):687–695, Feb 1977.

[7] M. Margulies, M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, 2005.

[8] S. Bennett. Solexa ltd. *Pharmacogenomics*, 5(4):433–438, 2004.

[9] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, and J. Szustakowki. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 2001.

[10] National Institute of Health. Nih genome sequencing costs page, 2012. URL `http://www.genome.gov/sequencingcosts/`.

[11] E. R. Mardis. Anticipating the 1,000 dollar genome. *Genome Biol.*, 7(7):112, 2006.

[12] H.E. Check et al. Human genome at ten: Life is complicated. *Nature*, 464(7289):664, 2010.

[13] Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *Nature*, 408 (6814):796–815, 2000.

[14] Y. Van de Peer, J. A. Fawcett, S. Proost, L. Sterck, and K. Vandepoele. The flowering world: a tale of duplications. *Trends Plant Sci*, 14(12):680–8, 2009.

[15] J. Lihova, K. K. Shimizu, and K. Marhold. Allopolyploid origin of Cardamine asarifolia (Brassicaceae): incongruence between plastid and nuclear ribosomal DNA sequences solved by a single-copy nuclear gene. *Mol. Phylogenet. Evol.*, 39(3): 759–786, Jun 2006.

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

[17] Aristotle. *Historia Animalium*. 330BC.

[18] Aristotle. *De Plantis*. 330BC.

[19] N. Roll-Hansen. The genotype theory of wilhelm johannsen and its relation to plant breeding and the study of evolution. *Centaurus*, 22:201–235, 1979.

[20] S. A. Goff, D. Ricke, T. H. Lan, G. Presting, R. Wang, M. Dunn, J. Glazebrook, A. Sessions, P. Oeller, H. Varma, D. Hadley, D. Hutchison, C. Martin, F. Katagiri, B. M. Lange, T. Moughamer, Y. Xia, P. Budworth, J. Zhong, T. Miguel, U. Paszkowski, S. Zhang, M. Colbert, W. L. Sun, L. Chen, B. Cooper, S. Park, T. C. Wood, L. Mao, P. Quail, R. Wing, R. Dean, Y. Yu, A. Zharkikh, R. Shen, S. Sahasrabudhe, A. Thomas, R. Cannings, A. Gutin, D. Pruss, J. Reid, S. Tavtigian, J. Mitchell, G. Eldredge, T. Scholl, R. M. Miller, S. Bhatnagar, N. Adey, T. Rubano, N. Tusneem, R. Robinson, J. Feldhaus, T. Macalma, A. Oliphant, and S. Briggs. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science*, 296(5565):92–100, Apr 2002.

[21] G. A. Tuskan, S. Difazio, S. Jansson, J. Bohlmann, I. Grigoriev, U. Hellsten, N. Putnam, S. Ralph, S. Rombauts, A. Salamov, J. Schein, L. Sterck, A. Aerts, R. R. Bhalerao, R. P. Bhalerao, D. Blaudez, W. Boerjan, A. Brun, A. Brunner, V. Busov, M. Campbell, J. Carlson, M. Chalot, J. Chapman, G. L. Chen, D. Cooper, P. M. Coutinho, J. Couturier, S. Covert, Q. Cronk, R. Cunningham, J. Davis, S. Degroeve, A. Dejardin, C. Depamphilis, J. Detter, B. Dirks, I. Dubchak, S. Duplessis, J. Ehlting, B. Ellis, K. Gendler, D. Goodstein, M. Gribskov, J. Grimwood, A. Groover, L. Gunter, B. Hamberger, B. Heinze, Y. Helariutta, B. Henrissat, D. Holligan, R. Holt, W. Huang, N. Islam-Faridi, S. Jones, M. Jones-Rhoades, R. Jorgensen, C. Joshi, J. Kangasjarvi, J. Karlsson, C. Kelleher, R. Kirkpatrick, M. Kirst, A. Kohler, U. Kalluri, F. Larimer, J. Leebens-Mack, J. C. Leple, P. Locascio, Y. Lou, S. Lucas, F. Martin, B. Montanini, C. Napoli, D. R. Nelson, C. Nelson, K. Nieminen, O. Nilsson, V. Pereda, G. Peter, R. Philippe, G. Pilate, A. Poliakov, J. Razumovskaya, P. Richardson, C. Rinaldi, K. Ritland, P. Rouze, D. Ryaboy, J. Schmutz, J. Schrader, B. Segerman, H. Shin, A. Siddiqui, F. Sterky, A. Terry, C. J. Tsai, E. Uberbacher, P. Unneberg, et al. The genome of black cottonwood, populus trichocarpa (torr. & gray). *Science*, 313(5793):1596–604, 2006.

[22] M. Spannagl, K. Mayer, J. Durner, G. Haberer, and A. Frohlich. Exploring the genomes: from Arabidopsis to crops. *J. Plant Physiol.*, 168(1):3–8, Jan 2011.

[23] F. Chardon, V. Noel, and C. Masclaux-Daubresse. Exploring NUE in crops and in Arabidopsis ideotypes to improve yield and seed quality. *J. Exp. Bot.*, 63(9):3401–3412, May 2012.

[24] A. H. Paterson, M. Freeling, H. Tang, and X. Wang. Insights from the comparison of plant genome sequences. *Annu Rev Plant Biol*, 61:349–372, 2010.

[25] S. J. Karpowicz, S. E. Prochnik, A. R. Grossman, and S. S. Merchant. The GreenCut2 resource, a phylogenomically derived inventory of proteins specific to the plant lineage. *J. Biol. Chem.*, 286(24):21427–21439, Jun 2011.

[26] E. A. Bray. Genes commonly regulated by water-deficit stress in Arabidopsis thaliana. *J. Exp. Bot.*, 55(407):2331–2341, Nov 2004.

[27] J. Hu, X. Barlet, L. Deslandes, J. Hirsch, D. X. Feng, I. Somssich, and Y. Marco. Transcriptional responses of Arabidopsis thaliana during wilt disease caused by the soil-borne phytopathogenic bacterium, Ralstonia solanacearum. *PLoS ONE*, 3(7): e2589, 2008.

[28] E. Moran-Diez, B. Rubio, S. Dominguez, R. Hermosa, E. Monte, and C. Nicolas. Transcriptomic response of Arabidopsis thaliana after 24 h incubation with the biocontrol fungus Trichoderma harzianum. *J. Plant Physiol.*, 169(6):614–620, Apr 2012.

[29] R. M. van Poecke and M. Dicke. Indirect defence of plants against herbivores: using Arabidopsis thaliana as a model plant. *Plant Biol (Stuttg)*, 6(4):387–401, Jul 2004.

[30] T. Umezawa, M. Fujita, Y. Fujita, K. Yamaguchi-Shinozaki, and K. Shinozaki. Engineering drought tolerance in plants: discovering and tailoring genes to unlock the future. *Curr. Opin. Biotechnol.*, 17(2):113–122, Apr 2006.

[31] F. Wu. Explaining public resistance to genetically modified corn: an analysis of the distribution of benefits and risks. *Risk Anal.*, 24(3):715–726, Jun 2004.

[32] S. P. Moose and R. H. Mumm. Molecular plant breeding as the foundation for 21st century crop improvement. *Plant Physiol.*, 147(3):969–977, Jul 2008.

[33] W. M. Fitch. Distinguishing homologous from analogous proteins. *Syst Zool*, 19(2):99–113, 1970.

[34] S. Ohno. *Evolution by gene duplication*. Springer-Verlag, 1970.

[35] J. B. Procter, J. Thompson, I. Letunic, C. Creevey, F. Jossinet, and G. J. Barton. Visualization of multiple alignments, phylogenies and gene family evolution. *Nat. Methods*, 7(3 Suppl):16–25, Mar 2010.

[36] S. De Bodt, S. Maere, and Y. Van de Peer. Genome duplication and the origin of angiosperms. *Trends Ecol Evol*, 20(11): 591–7, 2005.

[37] L. A. Meyers and D. A. Levin. On the abundance of polyploids in flowering plants. *Evolution*, 60(6):1198–1206, Jun 2006.

[38] S. B. Cannon, A. Mitra, A. Baumgarten, N. D. Young, and G. May. The roles of segmental and tandem gene duplication in the evolution of large gene families in Arabidopsis thaliana. *BMC Plant Biol.*, 4:10, Jun 2004.

[39] V. Shoja and L. Zhang. A roadmap of tandemly arrayed genes in the genomes of human, mouse, and rat. *Mol. Biol. Evol.*, 23(11):2134–2141, Nov 2006.

[40] C. Fan, Y. Chen, and M. Long. Recurrent tandem gene duplication gave rise to functionally divergent genes in Drosophila. *Mol. Biol. Evol.*, 25(7):1451–1458, Jul 2008.

[41] E. Rodgers-Melnick, S. P. Mane, P. Dharmawardhana, G. T. Slavov, O. R. Crasta, S. H. Strauss, A. M. Brunner, and S. P. Difazio. Contrasting patterns of evolution following whole genome versus tandem duplication events in Populus. *Genome Res.*, 22(1):95–105, Jan 2012.

[42] L. Cui, P. K. Wall, J. H. Leebens-Mack, B. G. Lindsay, D. E. Soltis, J. J. Doyle, P. S. Soltis, J. E. Carlson, K. Arumuganathan, A. Barakat, V. A. Albert, H. Ma, and C. W. dePamphilis. Widespread genome duplications throughout the history of flowering plants. *Genome Res*, 16(6):738–49, 2006.

[43] F. Bretagnolle and J.D. Thompson. Gametes with the somatic chromosome number: mechanisms of their formation and role in the evolution of autopolyploid plants. *New Phytol*, 129:1–22, 1995.

[44] C. Kohler, O. Mittelsten Scheid, and A. Erilova. The impact of the triploid block on the origin and evolution of polyploid plants. *Trends Genet.*, 26(3):142–148, Mar 2010.

[45] S. Proost, P. Pattyn, T. Gerats, and Y. Van de Peer. Journey through the past: 150 million years of plant genome evolution. *Plant J*, 66(1):58–65, 2011.

[46] H. Tang, J. E. Bowers, X. Wang, R. Ming, M. Alam, and A. H. Paterson. Synteny and collinearity in plant genomes. *Science*, 320(5875):486–8, 2008.

[47] J. Fostier, S. Proost, B. Dhoedt, Y. Saeys, P. Demeester, Y. Van de Peer, and K. Vandepoele. A greedy, graph-based algorithm for the alignment of multiple homologous gene lists. *Bioinformatics*, 2011.

[48] S. Proost, J. Fostier, D. De Witte, B. Dhoedt, P. Demeester, Y. Van de Peer, and K. Vandepoele. i-adhore 3.0–fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Res*, 40(2), 2012.

[49] K. Hanada, C. Zou, M. D. Lehti-Shiu, K. Shinozaki, and S. H. Shiu. Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. *Plant Physiol*, 148(2):993–1003, 2008.

[50] D. Leister. Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance gene. *Trends Genet.*, 20(3):116–122, Mar 2004.

[51] M. E. Schranz, M. A. Lysak, and T. Mitchell-Olds. The abc's of comparative genomics in the brassicaceae: building blocks of crucifer genomes. *Trends Plant Sci*, 11(11):535–42, 2006.

[52] M. A. Lysak, K. Cheung, M. Kitschke, and P. Bures. Ancestral chromosomal blocks are triplicated in Brassiceae species with varying chromosome number and genome size. *Plant Physiol.*, 145(2):402–410, Oct 2007.

[53] M. Feldman and A. A. Levy. Genome evolution due to allopolyploidization in wheat. *Genetics*, 192(3):763–774, Nov 2012.

[54] R. J. Buggs, S. Renny-Byfield, M. Chester, I. E. Jordon-Thaden, L. F. Viccini, S. Chamala, A. R. Leitch, P. S. Schnable, W. B. Barbazuk, P. S. Soltis, and D. E. Soltis. Next-generation sequencing and genome evolution in allopolyploids. *Am. J. Bot.*, 99(2):372–382, Feb 2012.

[55] R. L. Tatusov, E. V. Koonin, and D. J. Lipman. A genomic perspective on protein families. *Science*, 278(5338):631–637, Oct 1997.

[56] R. A. Studer and M. Robinson-Rechavi. How confident can we be that orthologs are similar, but paralogs differ? *Trends Genet.*, 25(5):210–216, May 2009.

[57] N. L. Nehrt, W. T. Clark, P. Radivojac, and M. W. Hahn. Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS Comput. Biol.*, 7(6):e1002073, Jun 2011.

[58] P. D. Thomas, V. Wood, C. J. Mungall, S. E. Lewis, and J. A. Blake. On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report. *PLoS Comput. Biol.*, 8(2):e1002386, 2012.

[59] A. M. Altenhoff, R. A. Studer, M. Robinson-Rechavi, and C. Dessimoz. Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Comput. Biol.*, 8(5):e1002514, 2012.

[60] E. V. Koonin. Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, 39:309–338, 2005.

[61] L. Li, Jr. Stoeckert, C. J., and D. S. Roos. Orthomcl: identification of ortholog groups for eukaryotic genomes. *Genome Res*, 13(9):2178–89, 2003.

[62] K. P. O'Brien, M. Remm, and E. L. Sonnhammer. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res*, 33(Database issue):D476–80, 2005.

[63] C. M. Zmasek and S. R. Eddy. RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, 3:14, May 2002.

[64] F. Chen, A. J. Mackey, J. K. Vermunt, and D. S. Roos. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS One*, 2(4):e383, 2007.

[65] A. R. Kersting, E. Bornberg-Bauer, A. D. Moore, and S. Grath. Dynamics and adaptive benefits of protein domain emergence and arrangements during plant genome evolution. *Genome Biol Evol*, 4(3):316–329, 2012.

[66] A. S. Warren, R. Anandakrishnan, and L. Zhang. Functional bias in molecular evolution rate of Arabidopsis thaliana. *BMC Evol. Biol.*, 10:125, 2010.

[67] J. X. Yue, J. Li, D. Wang, H. Araki, D. Tian, and S. Yang. Genome-wide investigation reveals high evolutionary rates in annual model plants. *BMC Plant Biol.*, 10:242, 2010.

[68] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25 (1):25–9, 2000.

[69] J. A. Blake, M. Dolan, H. Drabkin, D. P. Hill, L. Ni, D. Sitnikov, S. Burgess, T. Buza, C. Gresham, F. McCarthy, L. Pillai, H. Wang, S. Carbon, S. E. Lewis, C. J. Mungall, P. Gaudet, R. L. Chisholm, P. Fey, W. A. Kibbe, S. Basu, D. A. Siegele, B. K. McIntosh, D. P. Renfro, A. E. Zweifel, J. C. Hu, N. H. Brown, S. Tweedie, Y. Alam-Faruque, R. Apweiler, A. Auchinchloss, K. Axelsen, G. Argoud-Puy, B. Bely, M. Blatter, L. Bougueleret, E. Boutet, S. Branconi, L. Breuza, A. Bridge, P. Browne, W. M. Chan, E. Coudert, I. Cusin, E. Dimmer, P. Duek-Roggli, R. Eberhardt, A. Estreicher, L. Famiglietti, S. Ferro-Rojas, M. Feuermann, M. Gardner, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, R. Huntley, J. James, S. Jimenez, F. Jungo, G. Keller, K. Laiho, D. Legge, P. Lemercier, D. Lieberherr, M. Magrane, M. J. Martin, P. Masson, M. Moinat, C. O'Donovan, I. Pedruzzi, K. Pichler, D. Poggioli, P. Porras Millan, S. Poux, C. Rivoire, B. Roechert, T. Sawford, M. Schneider, H. Sehra, E. Stanley, A. Stutz, S. Sundaram, M. Tognolli, I. Xenarios, R. Foulger, J. Lomax, P. Roncaglia, E. Camon, V. K. Khodiyar, R. C. Lovering, P. J. Talmud, M. Chibucos, M. Gwinn Giglio, K. Dolinski, S. Heinicke, M. S. Livstone, R. Stephan,

M. A. Harris, S. G. Oliver, K. Rutherford, V. Wood, J. Bahler, A. Lock, P. J. Kersey, M. D. McDowall, D. M. Staines, M. Dwinell, M. Shimoyama, S. Laulederkind, T. Hayman, S. Wang, V. Petri, T. Lowry, P. D'Eustachio, L. Matthews, C. D. Amundsen, R. Balakrishnan, G. Binkley, J. M. Cherry, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. R. Engel, D. G. Fisk, J. E. Hirschman, B. C. Hitz, E. L. Hong, K. Karra, C. J. Krieger, S. R. Miyasato, R. S. Nash, J. Park, M. S. Skrzypek, S. Weng, E. D. Wong, T. Z. Berardini, D. Li, E. Huala, D. Slonim, H. Wick, P. Thomas, J. Chan, R. Kishore, P. Sternberg, K. Van Auken, D. Howe, and M. Westerfield. The Gene Ontology: enhancements for 2011. *Nucleic Acids Res.*, 40(Database issue):D559–564, Jan 2012.

[70] Gene Ontology consortium. Gene ontology website, 2012. URL `http://www.geneontology.org/GO.evidence.shtml`.

[71] K. Forslund, I. Pekkari, and E. L. Sonnhammer. Domain architecture conservation in orthologs. *BMC Bioinformatics*, 12: 326, 2011.

[72] R. D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. Sonnhammer, S. R. Eddy, and A. Bateman. The Pfam protein families database. *Nucleic Acids Res.*, 38 (Database issue):D211–222, Jan 2010.

[73] P. D. Thomas, A. Kejariwal, M. J. Campbell, H. Mi, K. Diemer, N. Guo, I. Ladunga, B. Ulitsky-Lazareva, A. Muruganujan, S. Rabkin, J. A. Vandergriff, and O. Doremieux. PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acids Res.*, 31(1):334–341, Jan 2003.

[74] S. Hunter, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, A. Laugraud, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, N. Mulder, D. Natale, C. Orengo, A. F. Quinn, J. D. Selengut, C. J. Sigrist, M. Thimma, P. D. Thomas, F. Valentin, D. Wilson, C. H. Wu, and C. Yeats. Interpro: the integrative protein signature database. *Nucleic Acids Res*, 37(Database issue):D211–5, 2009.

[75] L. Chae, I. Lee, J. Shin, and S. Y. Rhee. Towards understanding how molecular networks evolve in plants. *Curr. Opin. Plant Biol.*, 15(2):177–184, Apr 2012.

[76] P. Zhang, H. Foerster, C. P. Tissier, L. Mueller, S. Paley, P. D. Karp, and S. Y. Rhee. MetaCyc and AraCyc. Metabolic pathway databases for plant research. *Plant Physiol.*, 138(1):27–37, May 2005.

[77] I. Lee, B. Ambaru, P. Thakkar, E. M. Marcotte, and S. Y. Rhee. Rational association of genes with traits using a genome-scale gene network for Arabidopsis thaliana. *Nat. Biotechnol.*, 28(2):149–156, Feb 2010.

[78] M. Krallinger and A. Valencia. Text-mining and information-retrieval services for molecular biology. *Genome Biol.*, 6(7): 224, 2005.

[79] A. J. Enright, S. Van Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*, 30(7):1575–84, 2002.

[80] Y. Wang, H. Tang, J. D. Debarry, X. Tan, J. Li, X. Wang, T. H. Lee, H. Jin, B. Marler, H. Guo, J. C. Kissinger, and A. H. Paterson. Mcscanx: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res*, 2012.

[81] D. R. Zerbino and E. Birney. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome Res*, 18 (5):821–9, 2008.

[82] I. Birol, S. D. Jackman, C. B. Nielsen, J. Q. Qian, R. Varhol, G. Stazyk, R. D. Morin, Y. Zhao, M. Hirst, J. E. Schein, D. E. Horsman, J. M. Connors, R. D. Gascoyne, M. A. Marra, and S. J. Jones. De novo transcriptome assembly with abyss. *Bioinformatics*, 25(21):2872–7, 2009.

[83] M. Corpas, S. Fatumo, and R. Scheider. How not to be a bioinformatician. *Source Code Biol Med*, 7:3, 2012.

[84] M. Van Bel, S. Proost, E. Wischnitzki, S. Movahedi, C. Scheerlinck, Y. Van de Peer, and K. Vandepoele. Dissecting plant genomes with the plaza comparative genomics platform. *Plant Physiol*, 158(2):590–600, 2012.

[85] P. Lamesch, T. Z. Berardini, D. Li, D. Swarbreck, C. Wilks, R. Sasidharan, R. Muller, K. Dreher, D. L. Alexander, M. Garcia-Hernandez, A. S. Karthikeyan, C. H. Lee, W. D. Nelson, L. Ploetz, S. Singh, A. Wensel, and E. Huala. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.*, 40(Database issue):D1202–1210, Jan 2012.

[86] S. J. Schultheiss. Ten simple rules for providing a scientific Web resource. *PLoS Comput. Biol.*, 7(5):e1001126, May 2011.

[87] E. Gamma, R. Helm, R. Johnson, and J. Vlissides. *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley, 1994.

[88] P. Grubb and A. A. Takang. *Software Maintenance: Concepts and Practice*. World Scientific Publishing Company, 2003.

[89] L.J. Rosenblum. *Scientific Visualization: Advances and challenges*. Academic Press, 1994.

[90] EMBO and NIH. Visualizing biological data website, 2012. URL `http://vizbi.org/`.

[91] C. B. Nielsen, M. Cantor, I. Dubchak, D. Gordon, and T. Wang. Visualizing genomes: techniques and challenges. *Nat. Methods*, 7(3 Suppl):S5–S15, Mar 2010.

[92] N. Gehlenborg, S. I. O'Donoghue, N. S. Baliga, A. Goesmann, M. A. Hibbs, H. Kitano, O. Kohlbacher, H. Neuweger, R. Schneider, D. Tenenbaum, and A. C. Gavin. Visualization of omics data for systems biology. *Nat. Methods*, 7(3 Suppl): 56–68, Mar 2010.

[93] S. Few. *Encyclopedia of Human-Computer Interaction: Data Visualization for Human Perception*. The Interaction Design Foundation, 2010.

[94] M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra. Circos: an information aesthetic for comparative genomics. *Genome Res*, 19(9):1639–45, 2009.

[95] P. S. Schnable, D. Ware, R. S. Fulton, J. C. Stein, F. Wei, S. Pasternak, C. Liang, J. Zhang, L. Fulton, T. A. Graves, P. Minx, A. D. Reily, L. Courtney, S. S. Kruchowski, C. Tomlinson, C. Strong, K. Delehaunty, C. Fronick, B. Courtney, S. M. Rock, E. Belter, F. Du, K. Kim, R. M. Abbott, M. Cotton, A. Levy, P. Marchetto, K. Ochoa, S. M. Jackson, B. Gillam, W. Chen, L. Yan, J. Higginbotham, M. Cardenas, J. Waligorski, E. Applebaum, L. Phelps, J. Falcone, K. Kanchi, T. Thane, A. Scimone, N. Thane, J. Henke, T. Wang, J. Ruppert, N. Shah, K. Rotter, J. Hodges, E. Ingenthron, M. Cordes, S. Kohlberg, J. Sgro, B. Delgado, K. Mead, A. Chinwalla, S. Leonard, K. Crouse, K. Collura, D. Kudrna, J. Currie, R. He, A. Angelova, S. Rajasekar, T. Mueller, R. Lomeli, G. Scara, A. Ko, K. Delaney, M. Wissotski, G. Lopez, D. Campos, M. Braidotti, E. Ashley, W. Golser, H. Kim, S. Lee, J. Lin, Z. Dujmic, W. Kim, J. Talag, A. Zuccolo, C. Fan, A. Sebastian, M. Kramer, L. Spiegel, L. Nascimento, T. Zutavern, B. Miller, C. Ambroise, S. Muller, W. Spooner, A. Narechania, L. Ren, S. Wei, S. Kumari, B. Faga, M. J. Levy, L. McMahan, P. Van Buren, M. W. Vaughn, et al. The b73 maize genome: complexity, diversity, and dynamics. *Science*, 326(5956):1112–5, 2009.

[96] John Glazebrook. Open flash chart website, 2012. URL `http://teethgrinder.co.uk/open-flash-chart/`.

[97] FusionCharts Technologies LLP. Fusioncharts, 2012. URL `http://www.fusioncharts.com/`.

[98] Microsoft. Windows metro javascript, 2012. URL `http://msdn.microsoft.com/en-us/library/windows/apps/br211385.aspx`.

[99] S. M. Brady and N. J. Provart. Web-queryable large-scale data sets for hypothesis generation in plant biology. *Plant Cell*, 21(4):1034–51, 2009.

[100] S. Proost, M. Van Bel, L. Sterck, K. Billiau, T. Van Parys, Y. Van de Peer, and K. Vandepoele. Plaza: a comparative genomics resource to study gene and genome evolution in plants. *Plant Cell*, 21(12):3718–31, 2009.

[101] J. Li, X. Dai, T. Liu, and P. X. Zhao. LegumeIP: an integrative database for comparative genomics and transcriptomics of model legumes. *Nucleic Acids Res.*, 40(Database issue):D1221–1229, Jan 2012.

[102] M. G. Conte, S. Gaillard, N. Lanau, M. Rouard, and C. Perin. Greenphyldb: a database for plant comparative genomics. *Nucleic Acids Res*, 36(Database issue):D991–8, 2008.

[103] D. M. Goodstein, S. Shu, R. Howson, R. Neupane, R. D. Hayes, J. Fazo, T. Mitros, W. Dirks, U. Hellsten, N. Putnam, and D. S. Rokhsar. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res*, 40:D1178–86, 2012.

[104] P. J. Kersey, D. Lawson, E. Birney, P. S. Derwent, M. Haimel, J. Herrero, S. Keenan, A. Kerhornou, G. Koscielny, A. Kahari, R. J. Kinsella, E. Kulesha, U. Maheswari, K. Megy, M. Nuhn, G. Proctor, D. Staines, F. Valentin, A. J. Vilella, and A. Yates. Ensembl genomes: extending ensembl across the taxonomic space. *Nucleic Acids Res*, 38(Database issue):D563–9, 2010.

[105] E. Lyons and M. Freeling. How to usefully compare homologous plant genes and chromosomes as dna sequences. *Plant J*, 53(4):661–73, 2008.

[106] K. Horan, J. Lauricha, J. Bailey-Serres, N. Raikhel, and T. Girke. Genome cluster database. a sequence family analysis platform for arabidopsis and rice. *Plant Physiol*, 138(1):47–54, 2005.

[107] J. C. Chiu, E. K. Lee, M. G. Egan, I. N. Sarkar, G. M. Coruzzi, and R. DeSalle. Orthologid: automation of genome-scale ortholog identification within a parsimony framework. *Bioinformatics*, 22(6):699–707, 2006.

[108] P. K. Wall, J. Leebens-Mack, K. F. Muller, D. Field, N. S. Altman, and C. W. dePamphilis. Planttribes: a gene and gene family resource for comparative genomics in plants. *Nucleic Acids Res*, 36(Database issue):D970–6, 2008.

[109] X. Pan, L. Stein, and V. Brendel. Synbrowse: a synteny browser for comparative sequence analysis. *Bioinformatics*, 21(17): 3461–8, 2005.

[110] C. Liang, P. Jaiswal, C. Hebbard, S. Avraham, E. S. Buckler, T. Casstevens, B. Hurwitz, S. McCouch, J. Ni, A. Pujar, D. Ravenscroft, L. Ren, W. Spooner, I. Tecle, J. Thomason, C. W. Tung, X. Wei, I. Yap, K. Youens-Clark, D. Ware, and L. Stein. Gramene: a growing plant comparative genomics resource. *Nucleic Acids Res*, 36(Database issue):D947–53, 2008.

[111] M. Rouard, V. Guignon, C. Aluome, M. A. Laporte, G. Droc, C. Walde, C. M. Zmasek, C. Perin, and M. G. Conte. Greenphyldb v2.0: comparative and functional genomics in plants. *Nucleic Acids Res*, 39:D1095–102, 2011.

[112] D. M. Riano-Pachon, S. Ruzicic, I. Dreyer, and B. Mueller-Roeber. Plntfdb: an integrative plant transcription factor database. *BMC Bioinformatics*, 8:42, 2007.

[113] S. K. Palaniswamy, S. James, H. Sun, R. S. Lamb, R. V. Davuluri, and E. Grotewold. Agris and atregnet. a platform to link cis-regulatory elements and transcription factors into regulatory networks. *Plant Physiol*, 140(3):818–29, 2006.

[114] A. Yilmaz, Jr. Nishiyama, M. Y., B. G. Fuentes, G. M. Souza, D. Janies, J. Gray, and E. Grotewold. Grassius: A platform for comparative regulatory genomics across the grasses. *Plant Physiol*, 149(1):171–80, 2009.

[115] S. Hartmann, D. Lu, J. Phillips, and T. J. Vision. Phytome: a platform for plant comparative genomics. *Nucleic Acids Res*, 34(Database issue):D724–30, 2006.

[116] K. Vandepoele, K. Vlieghe, K. Florquin, L. Hennig, G. T. Beemster, W. Gruissem, Y. Van de Peer, D. Inze, and L. De Veylder. Genome-wide identification of potential plant E2F target genes. *Plant Physiol.*, 139(1):316–328, Sep 2005.

[117] S. Rudd. Expressed sequence tags: alternative or complement to whole genome sequences? *Trends Plant Sci*, 8(7):321–9, 2003.

[118] R. A. Gutierrez, P. J. Green, K. Keegstra, and J. B. Ohlrogge. Phylogenetic profiling of the arabidopsis thaliana proteome: what proteins distinguish plants from other organisms? *Genome Biol*, 5(8):R53, 2004.

[119] International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature*, 436(7052):793–800, 2005.

[120] M. S. Parker, T. Mock, and E. V. Armbrust. Genomic insights into marine microalgae. *Annu Rev Genet*, 42:619–45, 2008.

[121] A. H. Paterson. Leafing through the genomes of our major crop plants: strategies for capturing unique information. *Nat Rev Genet*, 7(3):174–84, 2006.

[122] E. Pennisi. Genome sequencing. the greening of plant genomics. *Science*, 317(5836):317, 2007.

[123] K. Hanada, X. Zhang, J. O. Borevitz, W. H. Li, and S. H. Shiu. A large number of novel coding small open reading frames in the intergenic regions of the arabidopsis thaliana genome are transcribed and/or under purifying selection. *Genome Res*, 17(5):632–40, 2007.

[124] A. Tanay, A. Regev, and R. Shamir. Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. *Proc Natl Acad Sci U S A*, 102(20):7203–8, 2005.

[125] T. Blomme, K. Vandepoele, S. De Bodt, C. Simillion, S. Maere, and Y. Van de Peer. The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol*, 7(5):R43, 2006.

[126] A. Stark, M. F. Lin, P. Kheradpour, J. S. Pedersen, L. Parts, J. W. Carlson, M. A. Crosby, M. D. Rasmussen, S. Roy, A. N. Deoras, J. G. Ruby, J. Brennecke, E. Hodges, A. S. Hinrichs, A. Caspi, B. Paten, S. W. Park, M. V. Han, M. L. Maeder, B. J. Polansky, B. E. Robson, S. Aerts, J. van Helden, B. Hassan, D. G. Gilbert, D. A. Eastman, M. Rice, M. Weir, M. W. Hahn, Y. Park, C. N. Dewey, L. Pachter, W. J. Kent, D. Haussler, E. C. Lai, D. P. Bartel, G. J. Hannon, T. C. Kaufman, M. B. Eisen, A. G. Clark, D. Smith, S. E. Celniker, W. M. Gelbart, and M. Kellis. Discovery of functional elements in 12 drosophila genomes using evolutionary signatures. *Nature*, 450(7167):219–32, 2007.

[127] T. M. Fulton, R. Van der Hoeven, N. T. Eannetta, and S. D. Tanksley. Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. *Plant Cell*, 14(7):1457–67, 2002.

[128] O. Jaillon, J. M. Aury, B. Noel, A. Policriti, C. Clepet, A. Casagrande, N. Choisne, S. Aubourg, N. Vitulo, C. Jubin, A. Vezzi, F. Legeai, P. Hugueney, C. Dasilva, D. Horner, E. Mica, D. Jublot, J. Poulain, C. Bruyere, A. Billault, B. Segurens, M. Gouyvenoux, E. Ugarte, F. Cattonaro, V. Anthouard, V. Vico, C. Del Fabbro, M. Alaux, G. Di Gaspero, V. Dumas, N. Felice, S. Paillard, I. Juman, M. Moroldo, S. Scalabrin, A. Canaguier, I. Le Clainche, G. Malacrida, E. Durand, G. Pesole, V. Laucou, P. Chatelet, D. Merdinoglu, M. Delledonne, M. Pezzotti, A. Lecharny, C. Scarpelli, F. Artiguenave, M. E. Pe, G. Valle, M. Morgante, M. Caboche, A. F. Adam-Blondon, J. Weissenbach, F. Quetier, and P. Wincker. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 449(7161):463–7, 2007.

[129] S. A. Rensing, D. Lang, A. D. Zimmer, A. Terry, A. Salamov, H. Shapiro, T. Nishiyama, P. F. Perroud, E. A. Lindquist, Y. Kamisugi, T. Tanahashi, K. Sakakibara, T. Fujita, K. Oishi, I. T. Shin, Y. Kuroki, A. Toyoda, Y. Suzuki, S. Hashimoto, K. Yamaguchi, S. Sugano, Y. Kohara, A. Fujiyama, A. Anterola, S. Aoki, N. Ashton, W. B. Barbazuk, E. Barker, J. L. Bennetzen, R. Blankenship, S. H. Cho, S. K. Dutcher, M. Estelle, J. A. Fawcett, H. Gundlach, K. Hanada, A. Heyl, K. A. Hicks, J. Hughes, M. Lohr, K. Mayer, A. Melkozernov, T. Murata, D. R. Nelson, B. Pils, M. Prigge, B. Reiss, T. Renner, S. Rombauts, P. J. Rushton, A. Sanderfoot, G. Schween, S. H. Shiu, K. Stueber, F. L. Theodoulou, H. Tu, Y. Van de Peer, P. J. Verrier, E. Waters, A. Wood, L. Yang, D. Cove, A. C. Cuming, M. Hasebe, S. Lucas, B. D. Mishler, R. Reski, I. V. Grigoriev, R. S. Quatrano, and J. L. Boore. The physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science*, 319(5859):64–9, 2008.

[130] S. S. Merchant, S. E. Prochnik, O. Vallon, E. H. Harris, S. J. Karpowicz, G. B. Witman, A. Terry, A. Salamov, L. K. Fritz-Laylin, L. Marechal-Drouard, W. F. Marshall, L. H. Qu, D. R. Nelson, A. A. Sanderfoot, M. H. Spalding, V. V. Kapitonov, Q. Ren, P. Ferris, E. Lindquist, H. Shapiro, S. M. Lucas, J. Grimwood, J. Schmutz, P. Cardol, H. Cerutti, G. Chanfreau, C. L. Chen, V. Cognat, M. T. Croft, R. Dent, S. Dutcher, E. Fernandez, H. Fukuzawa, D. Gonzalez-Ballester, D. Gonzalez-Halphen, A. Hallmann, M. Hanikenne, M. Hippler, W. Inwood, K. Jabbari, M. Kalanon, R. Kuras, P. A. Lefebvre, S. D. Lemaire, A. V. Lobanov, M. Lohr, A. Manuell, I. Meier, L. Mets, M. Mittag, T. Mittelmeier, J. V. Moroney, J. Moseley, C. Napoli, A. M. Nedelcu, K. Niyogi, S. V. Novoselov, I. T. Paulsen, G. Pazour, S. Purton, J. P. Ral, D. M. Riano-Pachon, W. Riekhof, L. Rymarquis, M. Schroda, D. Stern, J. Umen, R. Willows, N. Wilson, S. L. Zimmer, J. Allmer, J. Balk, K. Bisova, C. J. Chen, M. Elias, K. Gendler, C. Hauser, M. R. Lamb, H. Ledford, J. C. Long, J. Minagawa, M. D. Page, J. Pan, W. Pootakham, S. Roje, A. Rose, E. Stahlberg, A. M. Terauchi, P. Yang, S. Ball, C. Bowler, C. L. Dieckmann, V. N. Gladyshev, P. Green, R. Jorgensen, S. Mayfield, B. Mueller-Roeber, S. Rajamani, R. T. Sayre, P. Brokstein, et al. The chlamydomonas genome reveals the evolution of key animal and plant functions. *Science*, 318(5848):245–50, 2007.

[131] G. C. Conant and K. H. Wolfe. Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet*, 9(12): 938–50, 2008.

[132] M. Freeling and S. Subramaniam. Conserved noncoding sequences (cnss) in higher plants. *Curr Opin Plant Biol*, 12(2): 126–32, 2009.

[133] Z. J. Chen. Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. *Annu Rev Plant Biol*, 58:377–406, 2007.

[134] F. M. Rosin and E. M. Kramer. Old dogs, new tricks: regulatory evolution in conserved genetic modules leads to novel morphologies in plants. *Dev Biol*, 332(1):25–35, 2009.

[135] L. Stein. Genome annotation: from sequence to biology. *Nat Rev Genet*, 2(7):493–503, 2001.

[136] R. Lister, R. C. O'Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry, A. H. Millar, and J. R. Ecker. Highly integrated single-base resolution maps of the epigenome in arabidopsis. *Cell*, 133(3):523–36, 2008.

[137] N. Tsesmetzis, M. Couchman, J. Higgins, A. Smith, J. H. Doonan, G. J. Seifert, E. E. Schmidt, I. Vastrik, E. Birney, G. Wu, P. D'Eustachio, L. D. Stein, R. J. Morris, M. W. Bevan, and S. V. Walsh. Arabidopsis reactome: a foundation knowledgebase for plant systems biology. *Plant Cell*, 20(6):1426–36, 2008.

[138] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402, 1997.

[139] S. Guindon and O. Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*, 52(5):696–704, 2003.

[140] R. C. Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32(5): 1792–7, 2004.

[141] B. Vernot, M. Stolzer, A. Goldman, and D. Durand. Reconciliation with non-binary species trees. *J Comput Biol*, 15(8): 981–1006, 2008.

[142] M. W. Hahn. Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biol*, 8(7):R141, 2007.

[143] A. J. Vilella, J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin, and E. Birney. Ensemblcompara genetrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res*, 19(2):327–35, 2009.

[144] T. Hubbard, D. Andrews, M. Caccamo, G. Cameron, Y. Chen, M. Clamp, L. Clarke, G. Coates, T. Cox, F. Cunningham, V. Curwen, T. Cutts, T. Down, R. Durbin, X. M. Fernandez-Suarez, J. Gilbert, M. Hammond, J. Herrero, H. Hotz, K. Howe, V. Iyer, K. Jekosch, A. Kahari, A. Kasprzyk, D. Keefe, S. Keenan, F. Kokocinsci, D. London, I. Longden, G. McVicker, C. Melsopp, P. Meidl, S. Potter, G. Proctor, M. Rae, D. Rios, M. Schuster, S. Searle, J. Severin, G. Slater, D. Smedley, J. Smith, W. Spooner, A. Stabenau, J. Stalker, R. Storey, S. Trevanion, A. Ureta-Vidal, J. Vogel, S. White, C. Woodwark, and E. Birney. Ensembl 2005. *Nucleic Acids Res*, 33(Database issue):D447–53, 2005.

[145] The Reference Genome Group of the Gene Ontology Consortium. The gene ontology's reference genome project: a unified framework for functional annotation across species. *PLoS Comput Biol*, 5(7):e1000431, 2009.

[146] S. Gotz, J. M. Garcia-Gomez, J. Terol, T. D. Williams, S. H. Nagaraj, M. J. Nueda, M. Robles, M. Talon, J. Dopazo, and A. Conesa. High-throughput functional annotation and data mining with the blast2go suite. *Nucleic Acids Res*, 36(10): 3420–35, 2008.

[147] C. Simillion, K. Janssens, L. Sterck, and Y. Van de Peer. i-adhore 2.0: an improved tool to detect degenerated genomic homology using genomic profiles. *Bioinformatics*, 24(1):127–8, 2008.

[148] K. Vandepoele, C. Simillion, and Y. Van de Peer. Detecting the undetectable: uncovering duplicated segments in arabidopsis by comparison with rice. *Trends Genet*, 18(12):606–8, 2002.

[149] C. Simillion, K. Vandepoele, and Y. Van de Peer. Recent developments in computational approaches for uncovering genomic homology. *Bioessays*, 26(11):1225–35, 2004.

[150] J. M. Smith and N. H. Smith. Synonymous nucleotide divergence: what is "saturation"? *Genetics*, 142(3):1033–6, 1996.

[151] A. M. Waterhouse, J. B. Procter, D. M. Martin, M. Clamp, and G. J. Barton. Jalview version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9):1189–91, 2009.

[152] C. M. Zmasek and S. R. Eddy. Atv: display and manipulation of annotated phylogenetic trees. *Bioinformatics*, 17(4):383–4, 2001.

[153] S. Maere, S. De Bodt, J. Raes, T. Casneuf, M. Van Montagu, M. Kuiper, and Y. Van de Peer. Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci U S A*, 102(15):5454–9, 2005.

[154] J. D. Thompson, T. J. Gibson, and D. G. Higgins. Multiple sequence alignment using clustalw and clustalx. *Curr Protoc Bioinformatics*, Chapter 2:Unit 2 3, 2002.

[155] Z. Yang. Paml: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*, 13(5):555–6, 1997.

[156] D. Weigel and R. Mott. The 1001 genomes project for arabidopsis thaliana. *Genome Biol*, 10(5):107, 2009.

[157] D. A. Martinez and M. A. Nelson. The next generation becomes the now generation. *PLoS Genet*, 6(4):e1000906, 2010.

[158] S. C. Schuster. Next-generation sequencing transforms today's biology. *Nat Methods*, 5(1):16–8, 2008.

[159] S. Sato, Y. Nakamura, T. Kaneko, E. Asamizu, T. Kato, M. Nakao, S. Sasamoto, A. Watanabe, A. Ono, K. Kawashima, T. Fujishiro, M. Katoh, M. Kohara, Y. Kishida, C. Minami, S. Nakayama, N. Nakazaki, Y. Shimizu, S. Shinpo, C. Takahashi, T. Wada, M. Yamada, N. Ohmido, M. Hayashi, K. Fukui, T. Baba, T. Nakamichi, H. Mori, and S. Tabata. Genome structure of the legume, lotus japonicus. *DNA Res*, 15(4):227–39, 2008.

[160] R. Velasco, A. Zharkikh, J. Affourtit, A. Dhingra, A. Cestaro, A. Kalyanaraman, P. Fontana, S. K. Bhatnagar, M. Troggio, D. Pruss, S. Salvi, M. Pindo, P. Baldi, S. Castelletti, M. Cavaiuolo, G. Coppola, F. Costa, V. Cova, A. Dal Ri, V. Goremykin, M. Komjanc, S. Longhi, P. Magnago, G. Malacarne, M. Malnoy, D. Micheletti, M. Moretto, M. Perazzolli, A. Si-Ammour, S. Vezzulli, E. Zini, G. Eldredge, L. M. Fitzgerald, N. Gutin, J. Lanchbury, T. Macalma, J. T. Mitchell, J. Reid, B. Wardell, C. Kodira, Z. Chen, B. Desany, F. Niazi, M. Palmer, T. Koepke, D. Jiwan, S. Schaeffer, V. Krishnan, C. Wu, V. T. Chu, S. T. King, J. Vick, Q. Tao, A. Mraz, A. Stormo, K. Stormo, R. Bogden, D. Ederle, A. Stella, A. Vecchietti, M. M. Kater, S. Masiero, P. Lasserre, Y. Lespinasse, A. C. Allan, V. Bus, D. Chagne, R. N. Crowhurst, A. P. Gleave, E. Lavezzo, J. A. Fawcett, S. Proost, P. Rouze, L. Sterck, S. Toppo, B. Lazzari, R. P. Hellens, C. E. Durel, A. Gutin, R. E. Bumgarner, S. E. Gardiner, M. Skolnick, M. Egholm, Y. Van de Peer, F. Salamini, and R. Viola. The genome of the domesticated apple (malus x domestica borkh.). *Nat Genet*, 42(10):833–9, 2010.

[161] A. J. Garris, T. H. Tai, J. Coburn, S. Kresovich, and S. McCouch. Genetic structure and diversity in oryza sativa l. *Genetics*, 169(3):1631–8, 2005.

[162] M. Dassanayake, D. H. Oh, J. S. Haas, A. Hernandez, H. Hong, S. Ali, D. J. Yun, R. A. Bressan, J. K. Zhu, H. J. Bohnert, and J. M. Cheeseman. The genome of the extremophile crucifer thellungiella parvula. *Nat Genet*, 43(9):913–8, 2011.

[163] Eric Lyons, Brent Pedersen, Josh Kane, and Michael Freeling. The value of nonmodel genomes and an example using synmap within coge to dissect the hexaploidy that predates the rosids. *Tropical Plant Biology*, 1(3):181–190, 2008.

[164] T. Gabaldon. Large-scale assignment of orthology: back to phylogenetics? *Genome Biol*, 9(10):235, 2008.

[165] A. Kuzniar, R. C. van Ham, S. Pongor, and J. A. Leunissen. The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet*, 24(11):539–51, 2008.

[166] L. J. Jensen, P. Julien, M. Kuhn, C. von Mering, J. Muller, T. Doerks, and P. Bork. eggnog: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res*, 36(Database issue):D250–4, 2008.

[167] K. Trachana, T. A. Larsson, S. Powell, W. H. Chen, T. Doerks, J. Muller, and P. Bork. Orthology prediction methods: A quality assessment using curated protein families. *Bioessays*, 33(10):769–80, 2011.

[168] I. M. Meyer and R. Durbin. Gene structure conservation aids similarity based gene prediction. *Nucleic Acids Res*, 32(2): 776–83, 2004.

[169] T.T. Hu, P. Pattyn, E.G. Bakker, J. Cao, J.F. Cheng, R.M. Clark, N. Fahlgren, J.A. Fawcett, J. Grimwood, H. Gundlach, et al. The arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nature genetics*, 43(5):476–481, 2011.

[170] The International Brachypodium Initiative. Genome sequencing and analysis of the model grass brachypodium distachyon. *Nature*, 463(7282):763–8, 2010.

[171] R. Ming, S. Hou, Y. Feng, Q. Yu, A. Dionne-Laporte, J. H. Saw, P. Senin, W. Wang, B. V. Ly, K. L. Lewis, S. L. Salzberg, L. Feng, M. R. Jones, R. L. Skelton, J. E. Murray, C. Chen, W. Qian, J. Shen, P. Du, M. Eustice, E. Tong, H. Tang, E. Lyons, R. E. Paull, T. P. Michael, K. Wall, D. W. Rice, H. Albert, M. L. Wang, Y. J. Zhu, M. Schatz, N. Nagarajan, R. A. Acob, P. Guan, A. Blas, C. M. Wai, C. M. Ackerman, Y. Ren, C. Liu, J. Wang, J. Wang, J. K. Na, E. V. Shakirov, B. Haas, J. Thimmapuram, D. Nelson, X. Wang, J. E. Bowers, A. R. Gschwend, A. L. Delcher, R. Singh, J. Y. Suzuki, S. Tripathi, K. Neupane, H. Wei, B. Irikura, M. Paidi, N. Jiang, W. Zhang, G. Presting, A. Windsor, R. Navajas-Perez, M. J. Torres, F. A. Feltus, B. Porter, Y. Li, A. M. Burroughs, M. C. Luo, L. Liu, D. A. Christopher, S. M. Mount, P. H. Moore, T. Sugimura, J. Jiang, M. A. Schuler, V. Friedman, T. Mitchell-Olds, D. E. Shippen, C. W. dePamphilis, J. D. Palmer, M. Freeling, A. H. Paterson, D. Gonsalves, L. Wang, and M. Alam. The draft genome of the transgenic tropical fruit tree papaya (carica papaya linnaeus). *Nature*, 452(7190):991–6, 2008.

[172] V. Shulaev, D. J. Sargent, R. N. Crowhurst, T. C. Mockler, O. Folkerts, A. L. Delcher, P. Jaiswal, K. Mockaitis, A. Liston, S. P. Mane, P. Burns, T. M. Davis, J. P. Slovin, N. Bassil, R. P. Hellens, C. Evans, T. Harkins, C. Kodira, B. Desany, O. R. Crasta, R. V. Jensen, A. C. Allan, T. P. Michael, J. C. Setubal, J. M. Celton, D. J. Rees, K. P. Williams, S. H. Holt, J. J. Ruiz Rojas, M. Chatterjee, B. Liu, H. Silva, L. Meisel, A. Adato, S. A. Filichkin, M. Troggio, R. Viola, T. L. Ashman, H. Wang, P. Dharmawardhana, J. Elser, R. Raja, H. D. Priest, Jr. Bryant, D. W., S. E. Fox, S. A. Givan, L. J. Wilhelm, S. Naithani, A. Christoffels, D. Y. Salama, J. Carter, E. Lopez Girona, A. Zdepski, W. Wang, R. A. Kerstetter, W. Schwab, S. S. Korban, J. Davik, A. Monfort, B. Denoyes-Rothan, P. Arus, R. Mittler, B. Flinn, A. Aharoni, J. L. Bennetzen, S. L. Salzberg, A. W. Dickerman, R. Velasco, M. Borodovsky, R. E. Veilleux, and K. M. Folta. The genome of woodland strawberry (fragaria vesca). *Nat Genet*, 43(2):109–16, 2011.

[173] J. Schmutz, S. B. Cannon, J. Schlueter, J. Ma, T. Mitros, W. Nelson, D. L. Hyten, Q. Song, J. J. Thelen, J. Cheng, D. Xu, U. Hellsten, G. D. May, Y. Yu, T. Sakurai, T. Umezawa, M. K. Bhattacharyya, D. Sandhu, B. Valliyodan, E. Lindquist, M. Peto, D. Grant, S. Shu, D. Goodstein, K. Barry, M. Futrell-Griggs, B. Abernathy, J. Du, Z. Tian, L. Zhu, N. Gill, T. Joshi, M. Libault, A. Sethuraman, X. C. Zhang, K. Shinozaki, H. T. Nguyen, R. A. Wing, P. Cregan, J. Specht, J. Grimwood, D. Rokhsar, G. Stacey, R. C. Shoemaker, and S. A. Jackson. Genome sequence of the palaeopolyploid soybean. *Nature*, 463 (7278):178–83, 2010.

[174] N. D. Young, F. Debelle, G. E. Oldroyd, R. Geurts, S. B. Cannon, M. K. Udvardi, V. A. Benedito, K. F. Mayer, J. Gouzy, H. Schoof, Y. Van de Peer, S. Proost, D. R. Cook, B. C. Meyers, M. Spannagl, F. Cheung, S. De Mita, V. Krishnakumar, H. Gundlach, S. Zhou, J. Mudge, A. K. Bharti, J. D. Murray, M. A. Naoumkina, B. Rosen, K. A. Silverstein, H. Tang, S. Rombauts, P. X. Zhao, P. Zhou, V. Barbe, P. Bardou, M. Bechner, A. Bellec, A. Berger, H. Berges, S. Bidwell, T. Bisseling, N. Choisne, A. Couloux, R. Denny, S. Deshpande, X. Dai, J. J. Doyle, A. M. Dudez, A. D. Farmer, S. Fouteau, C. Franken, C. Gibelin, J. Gish, S. Goldstein, A. J. Gonzalez, P. J. Green, A. Hallab, M. Hartog, A. Hua, S. J. Humphray, D. H. Jeong, Y. Jing, A. Jocker, S. M. Kenton, D. J. Kim, K. Klee, H. Lai, C. Lang, S. Lin, S. L. Macmil, G. Magdelenat, L. Matthews,

J. McCorrison, E. L. Monaghan, J. H. Mun, F. Z. Najar, C. Nicholson, C. Noirot, M. O'Bleness, C. R. Paule, J. Poulain, F. Prion, B. Qin, C. Qu, E. F. Retzel, C. Riddle, E. Sallet, S. Samain, N. Samson, I. Sanders, O. Saurat, C. Scarpelli, T. Schiex, B. Segurens, A. J. Severin, D. J. Sherrier, R. Shi, S. Sims, S. R. Singer, S. Sinharoy, L. Sterck, A. Viollet, B. B. Wang, K. Wang, M. Wang, X. Wang, J. Warfsmann, J. Weissenbach, D. D. White, J. D. White, G. B. Wiley, P. Wincker, Y. Xing, L. Yang, Z. Yao, F. Ying, J. Zhai, L. Zhou, A. Zuber, J. Denarie, R. A. Dixon, G. D. May, D. C. Schwartz, J. Rogers, F. Quetier, C. D. Town, and B. A. Roe. The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature*, 480(7378):520–524, Dec 2011.

[175] A. Z. Worden, J. H. Lee, T. Mock, P. Rouze, M. P. Simmons, A. L. Aerts, A. E. Allen, M. L. Cuvelier, E. Derelle, M. V. Everett, E. Foulon, J. Grimwood, H. Gundlach, B. Henrissat, C. Napoli, S. M. McDonald, M. S. Parker, S. Rombauts, A. Salamov, P. Von Dassow, J. H. Badger, P. M. Coutinho, E. Demir, I. Dubchak, C. Gentemann, W. Eikrem, J. E. Gready, U. John, W. Lanier, E. A. Lindquist, S. Lucas, K. F. Mayer, H. Moreau, F. Not, R. Otillar, O. Panaud, J. Pangilinan, I. Paulsen, B. Piegu, A. Poliakov, S. Robbens, J. Schmutz, E. Toulza, T. Wyss, A. Zelensky, K. Zhou, E. V. Armbrust, D. Bhattacharya, U. W. Goodenough, Y. Van de Peer, and I. V. Grigoriev. Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes micromonas. *Science*, 324(5924):268–72, 2009.

[176] J. Yu, S. Hu, J. Wang, G. K. Wong, S. Li, B. Liu, Y. Deng, L. Dai, Y. Zhou, X. Zhang, M. Cao, J. Liu, J. Sun, J. Tang, Y. Chen, X. Huang, W. Lin, C. Ye, W. Tong, L. Cong, J. Geng, Y. Han, L. Li, W. Li, G. Hu, J. Li, Z. Liu, Q. Qi, T. Li, X. Wang, H. Lu, T. Wu, M. Zhu, P. Ni, H. Han, W. Dong, X. Ren, X. Feng, P. Cui, X. Li, H. Wang, X. Xu, W. Zhai, Z. Xu, J. Zhang, S. He, J. Xu, K. Zhang, X. Zheng, J. Dong, W. Zeng, L. Tao, J. Ye, J. Tan, X. Chen, J. He, D. Liu, W. Tian, C. Tian, H. Xia, Q. Bao, G. Li, H. Gao, T. Cao, W. Zhao, P. Li, W. Chen, Y. Zhang, J. Hu, S. Liu, J. Yang, G. Zhang, Y. Xiong, Z. Li, L. Mao, C. Zhou, Z. Zhu, R. Chen, B. Hao, W. Zheng, S. Chen, W. Guo, M. Tao, L. Zhu, L. Yuan, and H. Yang. A draft sequence of the rice genome (oryza sativa l. ssp. indica). *Science*, 296(5565):79–92, 2002.

[177] S. Ouyang, W. Zhu, J. Hamilton, H. Lin, M. Campbell, K. Childs, F. Thibaud-Nissen, R. L. Malek, Y. Lee, L. Zheng, J. Orvis, B. Haas, J. Wortman, and C. R. Buell. The tigr rice genome annotation resource: improvements and new features. *Nucleic Acids Res*, 35(Database issue):D883–7, 2007.

[178] B. Palenik, J. Grimwood, A. Aerts, P. Rouze, A. Salamov, N. Putnam, C. Dupont, R. Jorgensen, E. Derelle, S. Rombauts, K. Zhou, R. Otillar, S. S. Merchant, S. Podell, T. Gaasterland, C. Napoli, K. Gendler, A. Manuell, V. Tai, O. Vallon, G. Piganeau, S. Jancek, M. Heijde, K. Jabbari, C. Bowler, M. Lohr, S. Robbens, G. Werner, I. Dubchak, G. J. Pazour, Q. Ren, I. Paulsen, C. Delwiche, J. Schmutz, D. Rokhsar, Y. Van de Peer, H. Moreau, and I. V. Grigoriev. The tiny eukaryote ostreococcus provides genomic insights into the paradox of plankton speciation. *Proc Natl Acad Sci U S A*, 104(18):7705–10, 2007.

[179] E. Derelle, C. Ferraz, S. Rombauts, P. Rouze, A. Z. Worden, S. Robbens, F. Partensky, S. Degroeve, S. Echeynie, R. Cooke, Y. Saeys, J. Wuyts, K. Jabbari, C. Bowler, O. Panaud, B. Piegu, S. G. Ball, J. P. Ral, F. Y. Bouget, G. Piganeau, B. De Baets, A. Picard, M. Delseny, J. Demaille, Y. Van de Peer, and H. Moreau. Genome analysis of the smallest free-living eukaryote ostreococcus tauri unveils many unique features. *Proc Natl Acad Sci U S A*, 103(31):11647–52, 2006.

[180] A. P. Chan, J. Crabtree, Q. Zhao, H. Lorenzi, J. Orvis, D. Puiu, A. Melake-Berhan, K. M. Jones, J. Redman, G. Chen, E. B. Cahoon, M. Gedil, M. Stanke, B. J. Haas, J. R. Wortman, C. M. Fraser-Liggett, J. Ravel, and P. D. Rabinowicz. Draft genome sequence of the oilseed species ricinus communis. *Nat Biotechnol*, 28(9):951–6, 2010.

[181] J. A. Banks, T. Nishiyama, M. Hasebe, J. L. Bowman, M. Gribskov, C. dePamphilis, V. A. Albert, N. Aono, T. Aoyama, B. A. Ambrose, N. W. Ashton, M. J. Axtell, E. Barker, M. S. Barker, J. L. Bennetzen, N. D. Bonawitz, C. Chapple, C. Cheng, L. G. Correa, M. Dacre, J. DeBarry, I. Dreyer, M. Elias, E. M. Engstrom, M. Estelle, L. Feng, C. Finet, S. K. Floyd, W. B. Frommer, T. Fujita, L. Gramzow, M. Gutensohn, J. Harholt, M. Hattori, A. Heyl, T. Hirai, Y. Hiwatashi, M. Ishikawa, M. Iwata, K. G. Karol, B. Koehler, U. Kolukisaoglu, M. Kubo, T. Kurata, S. Lalonde, K. Li, Y. Li, A. Litt, E. Lyons, G. Manning, T. Maruyama, T. P. Michael, K. Mikami, S. Miyazaki, S. Morinaga, T. Murata, B. Mueller-Roeber, D. R. Nelson, M. Obara, Y. Oguri, R. G. Olmstead, N. Onodera, B. L. Petersen, B. Pils, M. Prigge, S. A. Rensing, D. M. Riano-Pachon, A. W. Roberts, Y. Sato, H. V. Scheller, B. Schulz, C. Schulz, E. V. Shakirov, N. Shibagaki, N. Shinohara, D. E. Shippen, I. S?rensen, R. Sotooka, N. Sugimoto, M. Sugita, N. Sumikawa, M. Tanurdzic, G. Theissen, P. Ulvskov, S. Wakazuki, J. K. Weng, W. W. Willats, D. Wipf, P. G. Wolf, L. Yang, A. D. Zimmer, Q. Zhu, T. Mitros, U. Hellsten, D. Loque, R. Otillar, A. Salamov, J. Schmutz, H. Shapiro, E. Lindquist, S. Lucas, D. Rokhsar, and I. V. Grigoriev. The Selaginella genome identifies genetic changes associated with the evolution of vascular plants. *Science*, 332(6032):960–963, May 2011.

[182] A. H. Paterson, J. E. Bowers, R. Bruggmann, I. Dubchak, J. Grimwood, H. Gundlach, G. Haberer, U. Hellsten, T. Mitros, A. Poliakov, J. Schmutz, M. Spannagl, H. Tang, X. Wang, T. Wicker, A. K. Bharti, J. Chapman, F. A. Feltus, U. Gowik, I. V. Grigoriev, E. Lyons, C. A. Maher, M. Martis, A. Narechania, R. P. Otillar, B. W. Penning, A. A. Salamov, Y. Wang, L. Zhang, N. C. Carpita, M. Freeling, A. R. Gingle, C. T. Hash, B. Keller, P. Klein, S. Kresovich, M. C. McCann, R. Ming, D. G. Peterson, Rahman Mehboob ur, D. Ware, P. Westhoff, K. F. Mayer, J. Messing, and D. S. Rokhsar. The sorghum bicolor genome and the diversification of grasses. *Nature*, 457(7229):551–6, 2009.

[183] X. Argout, J. Salse, J. M. Aury, M. J. Guiltinan, G. Droc, J. Gouzy, M. Allegre, C. Chaparro, T. Legavre, S. N. Maximova, M. Abrouk, F. Murat, O. Fouet, J. Poulain, M. Ruiz, Y. Roguet, M. Rodier-Goud, J. F. Barbosa-Neto, F. Sabot, D. Kudrna, J. S. Ammiraju, S. C. Schuster, J. E. Carlson, E. Sallet, T. Schiex, A. Dievart, M. Kramer, L. Gelley, Z. Shi, A. Berard, C. Viot, M. Boccara, A. M. Risterucci, V. Guignon, X. Sabau, M. J. Axtell, Z. Ma, Y. Zhang, S. Brown, M. Bourge, W. Golser, X. Song, D. Clement, R. Rivallan, M. Tahi, J. M. Akaza, B. Pitollat, K. Gramacho, A. D'Hont, D. Brunel, D. Infante, I. Kebe, P. Costet, R. Wing, W. R. McCombie, E. Guiderdoni, F. Quetier, O. Panaud, P. Wincker, S. Bocs, and C. Lanaud. The genome of theobroma cacao. *Nat Genet*, 43(2):101–8, 2011.

[184] S. E. Prochnik, J. Umen, A. M. Nedelcu, A. Hallmann, S. M. Miller, I. Nishii, P. Ferris, A. Kuo, T. Mitros, L. K. Fritz-Laylin, U. Hellsten, J. Chapman, O. Simakov, S. A. Rensing, A. Terry, J. Pangilinan, V. Kapitonov, J. Jurka, A. Salamov, H. Shapiro, J. Schmutz, J. Grimwood, E. Lindquist, S. Lucas, I. V. Grigoriev, R. Schmitt, D. Kirk, and D. S. Rokhsar. Genomic analysis of organismal complexity in the multicellular green alga volvox carteri. *Science*, 329(5988):223–6, 2010.

[185] E.K. Al-Dous, B. George, M.E. Al-Mahmoud, M.Y. Al-Jaber, H. Wang, Y.M. Salameh, E.K. Al-Azwani, S. Chaluvadi, A.C. Pontaroli, J. DeBarry, et al. De novo genome sequencing and comparative genomics of date palm (phoenix dactylifera). *Nature biotechnology*, 29(6):521–527, 2011.

[186] M. D. Bennett and I. J. Leitch. Nuclear dna amounts in angiosperms: progress, problems and prospects. *Ann Bot*, 95(1): 45–90, 2005.

[187] M. A. Huynen and P. Bork. Measuring genome evolution. *Proc Natl Acad Sci U S A*, 95(11):5849–56, 1998.

[188] R. D. Page and M. A. Charleston. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Mol Phylogenet Evol*, 7(2):231–40, 1997.

[189] C. M. Zmasek and S. R. Eddy. A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics*, 17(9):821–8, 2001.

[190] B. Linard, J. D. Thompson, O. Poch, and O. Lecompte. Orthoinspector: comprehensive orthology analysis and visual exploration. *BMC Bioinformatics*, 12:11, 2011.

[191] L. P. Pryszcz, J. Huerta-Cepas, and T. Gabaldon. Metaphors: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. *Nucleic Acids Res*, 39(2):231–40, 2011.

[192] S. Movahedi, Y. Van de Peer, and K. Vandepoele. Comparative network analysis reveals that tissue specificity and gene function are important factors influencing the mode of expression evolution in arabidopsis and rice. *Plant Physiol*, 156(3): 1316–30, 2011.

[193] A. E. Osbourn and B. Field. Operons. *Cell Mol Life Sci*, 66(23):3755–75, 2009.

[194] P. Michalak. Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. *Genomics*, 91 (3):243–8, 2008.

[195] E. V. Koonin. Evolution of genome architecture. *Int J Biochem Cell Biol*, 41(2):298–306, 2009.

[196] L. D. Hurst, C. Pal, and M. J. Lercher. The evolutionary dynamics of eukaryotic gene order. *Nat Rev Genet*, 5(4):299–310, 2004.

[197] S. Fabry, K. Muller, A. Lindauer, P. B. Park, T. Cornelius, and R. Schmitt. The organization structure and regulatory elements of Chlamydomonas histone genes reveal features linking plant and animal genes. *Curr. Genet.*, 28:333–345, 1995.

[198] G. Yi, S. H. Sze, and M. R. Thon. Identifying clusters of functionally related genes in genomes. *Bioinformatics*, 23(9): 1053–60, 2007.

[199] T. Abeel, T. Van Parys, Y. Saeys, J. Galagan, and Y. Van de Peer. Genomeview: a next-generation genome browser. *Nucleic Acids Res*, 2011.

[200] S. Federhen. The ncbi taxonomy database. *Nucleic Acids Res*, 2011. Journal article Nucleic acids research Nucleic Acids Res. 2011 Dec 1.

[201] M. J. Moore, P. S. Soltis, C. D. Bell, J. G. Burleigh, and D. E. Soltis. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proc Natl Acad Sci U S A*, 107(10):4623–8, 2010.

[202] T. J. Buza, F. M. McCarthy, N. Wang, S. M. Bridges, and S. C. Burgess. Gene ontology annotation quality analysis in model eukaryotes. *Nucleic Acids Res*, 36(2):e12, 2008.

[203] S. Movahedi, M. Van Bel, K. S. Heyndrickx, and K. Vandepoele. Comparative co-expression analysis in plant biology. *Plant Cell Environ*, Apr 2012.

[204] H. Moreau, B. Verhelst, A. Couloux, E. Derelle, S. Rombauts, N. Grimsley, M. Van Bel, J. Poulain, M. Katinka, M. Hohmann-Marriott, G. Piganeau, P. Rouze, C. Da Silva, P. Wincker, Y. Van de Peer, and K. Vandapoele. Gene functionalities and genome structure in Bathycoccus prasinos reflect cellular specializations at the base of the green lineage. *Genome Biol.*, 13(8):R74, Aug 2012.

[205] J. A. Lucas. *Plant Pathology and Plant Pathogens*. Wiley-Blackwell, 1998.

[206] R. A. Graves, S. E. Wellman, I. M. Chiu, and W. F. Marzluff. Differential expression of two clusters of mouse histone genes. *J Mol Biol*, 183(2):179–94, 1985.

[207] P. Tripputi, B. S. Emanuel, C. M. Croce, L. G. Green, G. S. Stein, and J. L. Stein. Human histone genes map to multiple chromosomes. *Proc Natl Acad Sci U S A*, 83(10):3185–8, 1986.

[208] B. S. Allen, J. L. Stein, G. S. Stein, and H. Ostrer. Single-copy flanking sequences in human histone gene clusters map to chromosomes 1 and 6. *Genomics*, 10(2):486–8, 1991.

[209] M. R. Parthun, J. Widom, and D. E. Gottschling. The major cytoplasmic histone acetyltransferase in yeast: links to chromatin replication and histone metabolism. *Cell*, 87(1):85–94, 1996.

[210] M. Grunstein. Histone acetylation in chromatin structure and transcription. *Nature*, 389(6649):349–52, 1997.

[211] R. C. Hardison. Comparative genomics. *PLoS Biol.*, 1(2):E58, Nov 2003.

[212] I. Tirosh, Y. Bilu, and N. Barkai. Comparative biology: beyond sequence analysis. *Curr. Opin. Biotechnol.*, 18(4):371–377, Aug 2007.

[213] Y. Lu, P. Huggins, and Z. Bar-Joseph. Cross species analysis of microarray expression data. *Bioinformatics*, 25(12):1476–1483, Jun 2009.

[214] S. Bergmann, J. Ihmels, and N. Barkai. Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol.*, 2(1):E9, Jan 2004.

[215] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–255, Oct 2003.

[216] R. Xu and D. Wunsch. Survey of clustering algorithms. *IEEE Trans Neural Netw*, 16(3):645–678, May 2005.

[217] M. Mutwil, S. Klie, T. Tohge, F. M. Giorgi, O. Wilkins, M. M. Campbell, A. R. Fernie, B. Usadel, Z. Nikoloski, and S. Persson. PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. *Plant Cell*, 23(3):895–910, Mar 2011.

[218] M. Humphry, P. Bednarek, B. Kemmerling, S. Koh, M. Stein, U. Gobel, K. Stuber, M. Pislewska-Bednarek, A. Loraine, P. Schulze-Lefert, S. Somerville, and R. Panstruga. A regulon conserved in monocot and dicot plants defines a functional module in antifungal plant immunity. *Proc. Natl. Acad. Sci. U.S.A.*, 107(50):21896–21901, Dec 2010.

[219] M. D. Chikina and O. G. Troyanskaya. Accurate quantification of functional analogy among close homologs. *PLoS Comput. Biol.*, 7(2):e1001074, 2011.

[220] d. a. W. Huang, B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, 37(1):1–13, Jan 2009.

[221] S. P. Ficklin and F. A. Feltus. Gene coexpression network alignment and conservation of gene modules between two grass species: maize and rice. *Plant Physiol.*, 156(3):1244–1256, Jul 2011.

[222] N. Takahashi, M. Quimbaya, V. Schubert, T. Lammens, K. Vandepoele, I. Schubert, M. Matsui, D. Inze, G. Berx, and L. De Veylder. The MCM-binding protein ETG1 aids sister chromatid cohesion required for postreplicative homologous recombination repair. *PLoS Genet.*, 6(1):e1000817, Jan 2010.

[223] S. De Bodt, D. Carvajal, J. Hollunder, J. Van den Cruyce, S. Movahedi, and D. Inze. CORNET: a user-friendly tool for data mining and integration. *Plant Physiol.*, 152(3):1167–1179, Mar 2010.

[224] K. Vandepoele, M. Quimbaya, T. Casneuf, L. De Veylder, and Y. Van de Peer. Unraveling transcriptional control in Arabidopsis using cis-regulatory elements and coexpression networks. *Plant Physiol.*, 150(2):535–546, Jun 2009.

[225] E. V. Armbrust, J. A. Berges, C. Bowler, B. R. Green, D. Martinez, N. H. Putnam, S. Zhou, A. E. Allen, K. E. Apt, M. Bechner, M. A. Brzezinski, B. K. Chaal, A. Chiovitti, A. K. Davis, M. S. Demarest, J. C. Detter, T. Glavina, D. Goodstein, M. Z. Hadi, U. Hellsten, M. Hildebrand, B. D. Jenkins, J. Jurka, V. V. Kapitonov, N. Kroger, W. W. Lau, T. W. Lane, F. W. Larimer, J. C. Lippmeier, S. Lucas, M. Medina, A. Montsant, M. Obornik, M. S. Parker, B. Palenik, G. J. Pazour, P. M. Richardson, T. A. Rynearson, M. A. Saito, D. C. Schwartz, K. Thamatrakoln, K. Valentin, A. Vardi, F. P. Wilkerson, and D. S. Rokhsar. The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. *Science*, 306 (5693):79–86, Oct 2004.

[226] C. J. Gobler, D. L. Berry, S. T. Dyhrman, S. W. Wilhelm, A. Salamov, A. V. Lobanov, Y. Zhang, J. L. Collier, L. L. Wurch, A. B. Kustka, B. D. Dill, M. Shah, N. C. VerBerkmoes, A. Kuo, A. Terry, J. Pangilinan, E. A. Lindquist, S. Lucas, I. T. Paulsen, T. K. Hattenrath-Lehmann, S. C. Talmage, E. A. Walker, F. Koch, A. M. Burson, M. A. Marcoval, Y. Z. Tang, G. R. Lecleir, K. J. Coyne, G. M. Berg, E. M. Bertrand, M. A. Saito, V. N. Gladyshev, and I. V. Grigoriev. Niche of harmful alga Aureococcus anophagefferens revealed through ecogenomics. *Proc. Natl. Acad. Sci. U.S.A.*, 108(11):4352–4357, Mar 2011.

[227] G. Blanc, I. Agarkova, J. Grimwood, A. Kuo, A. Brueggeman, D. D. Dunigan, J. Gurnon, I. Ladunga, E. Lindquist, S. Lucas, J. Pangilinan, T. Proschold, A. Salamov, J. Schmutz, D. Weeks, T. Yamada, A. Lomsadze, M. Borodovsky, J. M. Claverie, I. V. Grigoriev, and J. L. Van Etten. The genome of the polar eukaryotic microalga Coccomyxa subellipsoidea reveals traits of cold adaptation. *Genome Biol.*, 13(5):R39, 2012.

[228] U. Maheswari, K. Jabbari, J. L. Petit, B. M. Porcel, A. E. Allen, J. P. Cadoret, A. De Martino, M. Heijde, R. Kaas, J. La Roche, P. J. Lopez, V. Martin-Jezequel, A. Meichenin, T. Mock, M. Schnitzler Parker, A. Vardi, E. V. Armbrust, J. Weissenbach, M. Katinka, and C. Bowler. Digital expression profiling of novel diatom transcripts provides insight into their biological functions. *Genome Biol.*, 11(8):R85, 2010.

[229] E. Karsenti, S. G. Acinas, P. Bork, C. Bowler, C. De Vargas, J. Raes, M. Sullivan, D. Arendt, F. Benzoni, J. M. Claverie, M. Follows, G. Gorsky, P. Hingamp, D. Iudicone, O. Jaillon, S. Kandels-Lewis, U. Krzic, F. Not, H. Ogata, S. Pesant, E. G. Reynaud, C. Sardet, M. E. Sieracki, S. Speich, D. Velayoudon, J. Weissenbach, and P. Wincker. A holistic approach to marine eco-systems biology. *PLoS Biol.*, 9(10):e1001177, Oct 2011.

[230] S. Sun, J. Chen, W. Li, I. Altintas, A. Lin, S. Peltier, K. Stocks, E. E. Allen, M. Ellisman, J. Grethe, and J. Wooley. Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic Acids Res.*, 39(Database issue):D546–551, Jan 2011.

[231] G. Piganeau, K. Vandepoele, S. Gourbiere, Y. Van de Peer, and H. Moreau. Unravelling cis-regulatory elements in the genome of the smallest photosynthetic eukaryote: phylogenetic footprinting in Ostreococcus. *J. Mol. Evol.*, 69(3):249–259, Sep 2009.

[232] N. Grimsley, B. Pequin, C. Bachy, H. Moreau, and G. Piganeau. Cryptic sex in the smallest eukaryotic marine green alga. *Mol. Biol. Evol.*, 27(1):47–54, Jan 2010.

[233] M. Garber, M. G. Grabherr, M. Guttman, and C. Trapnell. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods*, 8(6):469–477, Jun 2011.

[234] J. A. Martin and Z. Wang. Next-generation transcriptome assembly. *Nat. Rev. Genet.*, 12(10):671–682, Oct 2011.

[235] R. Li, C. Yu, Y. Li, T. W. Lam, S. M. Yiu, K. Kristiansen, and J. Wang. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966–1967, Aug 2009.

[236] G. Robertson, J. Schein, R. Chiu, R. Corbett, M. Field, S. D. Jackman, K. Mungall, S. Lee, H. M. Okada, J. Q. Qian, M. Griffith, A. Raymond, N. Thiessen, T. Cezard, Y. S. Butterfield, R. Newsome, S. K. Chan, R. She, R. Varhol, B. Kamoh, A. L. Prabhu, A. Tam, Y. Zhao, R. A. Moore, M. Hirst, M. A. Marra, S. J. Jones, P. A. Hoodless, and I. Birol. De novo assembly and analysis of RNA-seq data. *Nat. Methods*, 7(11):909–912, Nov 2010.

[237] M. H. Schulz, D. R. Zerbino, M. Vingron, and E. Birney. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, 28(8):1086–1092, Apr 2012.

[238] P. K. Wall, J. Leebens-Mack, A. S. Chanderbali, A. Barakat, E. Wolcott, H. Liang, L. Landherr, L. P. Tomsho, Y. Hu, J. E. Carlson, H. Ma, S. C. Schuster, D. E. Soltis, P. S. Soltis, N. Altman, and C. W. dePamphilis. Comparison of next generation sequencing technologies for transcriptome characterization. *BMC Genomics*, 10:347, 2009.

[239] Q. Y. Zhao, Y. Wang, Y. M. Kong, D. Luo, X. Li, and P. Hao. Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC Bioinformatics*, 12 Suppl 14:S2, 2011.

[240] Y. Moriya, M. Itoh, S. Okuda, A. C. Yoshizawa, and M. Kanehisa. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.*, 35(Web Server issue):W182–185, Jul 2007.

[241] S. Gotz, R. Arnold, P. Sebastian-Leon, S. Martin-Rodriguez, P. Tischler, M. A. Jehl, J. Dopazo, T. Rattei, and A. Conesa. B2G-FAR, a species-centered GO annotation repository. *Bioinformatics*, 27(7):919–924, Apr 2011.

[242] E. E. Philipp, L. Kraemer, D. Mountfort, M. Schilhabel, S. Schreiber, and P. Rosenstiel. The Transcriptome Analysis and Comparison Explorer–T-ACE: a platform-independent, graphical tool to process large RNAseq datasets of non-model organisms. *Bioinformatics*, 28(6):777–783, Mar 2012.

[243] L. Baldo, M. E. Santos, and W. Salzburger. Comparative transcriptomics of Eastern African cichlid fishes shows signs of positive selection and a large contribution of untranslated regions to genetic diversity. *Genome Biol Evol*, 3:443–455, 2011.

[244] M. F. Poelchau, J. A. Reynolds, D. L. Denlinger, C. G. Elsik, and P. A. Armbruster. A de novo transcriptome of the Asian tiger mosquito, Aedes albopictus, to identify candidate transcripts for diapause preparation. *BMC Genomics*, 12:619, 2011.

[245] A. C. Tzika, R. Helaers, G. Schramm, and M. C. Milinkovitch. Reptilian-transcriptome v1.0, a glimpse in the brain transcriptome of five divergent Sauropsida lineages and the phylogenetic position of turtles. *Evodevo*, 2(1):19, 2011.

[246] Y. Zhao, H. Tang, and Y. Ye. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics*, 28(1):125–126, Jan 2012.

[247] F. Chen, A. J. Mackey, Jr. Stoeckert, C. J., and D. S. Roos. Orthomcl-db: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res*, 34(Database issue):D363–8, 2006.

[248] J. Gouzy, S. Carrere, and T. Schiex. Framedp: sensitive peptide detection on noisy matured sequences. *Bioinformatics*, 25 (5):670–1, 2009.

[249] S. Guindon, J. F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.*, 59(3):307–321, May 2010.

[250] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, October 1990.

[251] J. D. Wasmuth and M. L. Blaxter. prot4EST: translating expressed sequence tags from neglected genomes. *BMC Bioinformatics*, 5:187, Nov 2004.

[252] K. L. Childs, J. P. Hamilton, W. Zhu, E. Ly, F. Cheung, H. Wu, P. D. Rabinowicz, C. D. Town, C. R. Buell, and A. P. Chan. The TIGR Plant Transcript Assemblies database. *Nucleic Acids Res.*, 35(Database issue):D846–851, Jan 2007.

[253] D. Vekemans, S. Proost, K. Vanneste, H. Coenen, T. Viaene, P. Ruelens, S. Maere, Y. Van de Peer, and K. Geuten. Gamma Paleohexaploidy in the Stem Lineage of Core Eudicots: Significance for MADS-Box Gene and Species Diversification. *Mol. Biol. Evol.*, Aug 2012.

[254] E. Meyer, T. L. Logan, and T. E. Juenger. Transcriptome analysis and gene expression atlas for Panicum hallii var. filipes, a diploid model for biofuel research. *Plant J.*, 70(5):879–890, Jun 2012.

[255] M. V. Han and C. M. Zmasek. phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics*, 10:356, 2009.

[256] Walter Zorn. Dhtml javascript graphics, 2012. URL `http://www.walterzorn.de/en/jsgraphics/jsgraphics_e.htm`.

[257] CakePHP Community. Cakephp, 2012. URL `http://cakephp.org/`.

[258] Z. Wang, M. Gerstein, and M. Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1): 57–63, 2009.

[259] MF. Rogers, J. Thomas, AS. Reddy, and Ben-Hur A. Splicegrapher: detecting patterns of alternative splicing from rna-seq data in the context of gene models and est data. *Genome Biol.*, 13(1):R4, 2012.

[260] I. V. Grigoriev, H. Nordberg, I. Shabalov, A. Aerts, M. Cantor, D. Goodstein, A. Kuo, S. Minovitsky, R. Nikitin, R. A. Ohm, R. Otillar, A. Poliakov, I. Ratnere, R. Riley, T. Smirnova, D. Rokhsar, and I. Dubchak. The genome portal of the Department of Energy Joint Genome Institute. *Nucleic Acids Res.*, 40(Database issue):26–32, Jan 2012.

[261] P. Milos. Helicos biosciences. *Pharmacogenomics*, 9(4):477–480, 2008.

[262] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, et al. Real-time dna sequencing from single polymerase molecules. *Science*, 323(5910):133, 2009.

[263] J.M. Rothberg, W. Hinz, T.M. Rearick, J. Schultz, W. Mileski, M. Davey, J.H. Leamon, K. Johnson, M.J. Milgrew, M. Edwards, et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356):348–352, 2011.

[264] P. Compeau, P.A. Pevzner, and Tesler G. How to apply de bruijn graphs to genome assembly. *Nature Biotechnology*, 29: 987–991, 2011.

[265] No authors listed. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055): 69–87, Sep 2005.

[266] No authors listed. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, 306(5696):636–640, Oct 2004.

[267] S. Chanock. Toward mapping the biology of the genome. *Genome Res.*, 22(9):1612–1615, Sep 2012.

[268] A. Maizel and D. Weigel. Temporally and spatially controlled induction of gene expression in Arabidopsis thaliana. *Plant J.*, 38(1):164–171, Apr 2004.

[269] C. Zou, M. D. Lehti-Shiu, M. Thomashow, and S. H. Shiu. Evolution of stress-regulated gene expression in duplicate genes of Arabidopsis thaliana. *PLoS Genet.*, 5(7):e1000581, Jul 2009.

[270] M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. De Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, V. J. Makeev, A. A. Mironov, W. S. Noble, G. Pavesi, G. Pesole, M. Regnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu. Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, 23(1):137–144, Jan 2005.

[271] M. K. Das and H. K. Dai. A survey of DNA motif finding algorithms. *BMC Bioinformatics*, 8 Suppl 7:S21, 2007.

[272] P. Kheradpour, A. Stark, S. Roy, and M. Kellis. Reliable prediction of regulator targets using 12 Drosophila genomes. *Genome Res.*, 17(12):1919–1931, Dec 2007.

[273] N. H. Syed, M. Kalyna, Y. Marquez, A. Barta, and J. W. Brown. Alternative splicing in plants - coming of age. *Trends Plant Sci.*, 17(10):616–623, Oct 2012.

[274] M. Sammeth, S. Foissac, and R. Guigo. A general definition and nomenclature for alternative splicing events. *PLoS Comput. Biol.*, 4(8):e1000147, 2008.

[275] W. Zhu, S. D. Schlueter, and V. Brendel. Refined annotation of the arabidopsis genome by complete expressed sequence tag mapping. *Plant Physiol*, 132(2):469–84, 2003.

[276] K. Iida, M. Seki, T. Sakurai, M. Satou, K. Akiyama, T. Toyoda, A. Konagaya, and K. Shinozaki. Genome-wide analysis of alternative pre-mrna splicing in arabidopsis thaliana based on full-length cdna sequences. *Nucleic Acids Res*, 32(17): 5096–103, 2004.

[277] Y. L. Xiao, S. R. Smith, N. Ishmael, J. C. Redman, N. Kumar, E. L. Monaghan, M. Ayele, B. J. Haas, H. C. Wu, and C. D. Town. Analysis of the cdnas of hypothetical genes on arabidopsis chromosome 2 reveals numerous transcript variants. *Plant Physiol*, 139(3):1323–37, 2005.

[278] M. A. Campbell, B. J. Haas, J. P. Hamilton, S. M. Mount, and C. R. Buell. Comprehensive analysis of alternative splicing in rice and comparative analyses with arabidopsis. *BMC Genomics*, 7:327, 2006.

[279] S. A. Filichkin, H. D. Priest, S. A. Givan, R. Shen, D. W. Bryant, S. E. Fox, W. K. Wong, and T. C. Mockler. Genome-wide mapping of alternative splicing in arabidopsis thaliana. *Genome Res*, 20(1):45–58, 2010.

[280] D. Marquez, J.W.S. Brown, C. Simpson, A. Barta, and M. Kalyna. Transcriptome survey reveals increased complexity of the alternative splicing landscape in arabidopsis. *Genome Research*, 22:1184–1195, 2012.

[281] H. Ner-Gaon, N. Leviatan, E. Rubin, and R. Fluhr. Comparative cross-species alternative splicing in plants. *Plant Physiol.*, 144(3):1632–1641, Jul 2007.

[282] Z. Su, J. Wang, J. Yu, X. Huang, and X. Gu. Evolution of alternative splicing after gene duplication. *Genome Res.*, 16(2): 182–189, Feb 2006.

[283] Z. Su and X. Gu. Revisit on the evolutionary relationship between alternative splicing and gene duplication. *Gene*, 504(1): 102–106, Aug 2012.

[284] W. B. Barbazuk, Y. Fu, and K. M. McGinnis. Genome-wide analyses of alternative splicing in plants: opportunities and challenges. *Genome Res.*, 18(9):1381–1392, Sep 2008.

[285] A. Kasprzyk. BioMart: driving a paradigm change in biological data management. *Database (Oxford)*, 2011:bar049, 2011.

[286] F. Palmieri, C. L. Pierri, A. De Grassi, A. Nunes-Nesi, and A. R. Fernie. Evolution, structure and function of mitochondrial carriers: a review with new insights. *Plant J.*, 66(1):161–181, Apr 2011.

[287] A. J. Brueggeman, D. S. Gangadharaiah, M. F. Cserhati, D. Casero, D. P. Weeks, and I. Ladunga. Activation of the carbon concentrating mechanism by CO2 deprivation coincides with massive transcriptional restructuring in Chlamydomonas reinhardtii. *Plant Cell*, 24(5):1860–1875, May 2012.

[288] P. Flicek, M.R. Amode, D. Barrell, K. Beal, S. Brent, Y. Chen, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, et al. Ensembl 2011. *Nucleic acids research*, 39(suppl 1):D800, 2011.

[289] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti. Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524, Dec 2011.

[290] Y. Saeys, S. Degroeve, D. Aeyels, P. Rouze, and Y. Van de Peer. Feature selection for splice site prediction: a new method using EDA-based feature ranking. *BMC Bioinformatics*, 5:64, May 2004.

[291] T. Abeel, Y. Saeys, E. Bonnet, P. Rouze, and Y. Van de Peer. Generic eukaryotic core promoter prediction using structural features of DNA. *Genome Res.*, 18(2):310–323, Feb 2008.

[292] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, 26(3):392–398, Feb 2010.

[293] B. Han, H. M. Kang, and E. Eskin. Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS Genet.*, 5(4):e1000456, Apr 2009.