UNIVERSITEIT
GENT

FACULTY OF SCIENCES

# Integration of Genomic Data to Study Genome Evolution in Plants

Sebastian Proost

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

Research funded by a PhD grant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT Vlaanderen).

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

Academic year: 2011-2012

VIB

Plant Systems Biology

bioinformatics
Gent

iwt

# Examination Commitee

**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. Olivier De Clerck**

Faculty of Sciences, Department of Marine Biology, Ghent University

**Prof. Dr. Jeroen Raes**

Faculty of Sciences, Department of Molecular and Cellular Interactions, VUB

**Prof. Dr. Jan Fostier**[*]

Faculty of Engineering, Department of Information Technology, Ghent University

**Prof. Dr. Koen Geuten**[*]

Faculty of Sciences, Department of Biology, Catholic University of Leuven

**Dr. Alessandro Cestaro**[*]

Research and Innovation Centre - Fondazione Edmund Mach, Comparative Genomics Research Group

**Dr. Steven Maere**[*]

Faculty of Sciences, Department of Plant Biotechnology and Bioinformatics, Ghent University

[*] Member of Reading Commission

# Thanks to ...

...Yves and Klaas for their guidance. Without your efforts and patience this work wouldn't have had the quality nor the quantity it has today.

...my jury members for taking the time to evaluate this work.

...Jan and Michiel[*] for the many pleasant, though highly productive, collaborations we've had through the years. By combining our complementary expertise, I feel we've created something outstanding.

...Pierre, Lieven, Jeffrey and Stephane for their help, suggestions and support on the genome projects I've been involved in.

...Elisabeth, Dieter, Dries, Koen, Pedro, Sara, Stefanie, Ying and Yvan for the work we've done together.

...my foreign collaborators. The Apple guys: Riccardo, Alessandro and Marco. Working as part of the apple consortium was an interesting experience and I genuinely enjoyed my time at your institute last summer. Nevin, Steven C, Christopher T and many others for their stamina to see through to the end of the *Medicago truncatula* genome project.

...our IT support and technicians Frederik, Kenny, Marijn, Thomas VP, Hendrik,

[*] Obligatory quote page 53

Dany, Raf and Luc for keeping the servers, websites, cluster, ... I've depended on so much during this thesis up and running.

...my students Marieke, Steven T and Christopher S, I hope I have been able to teach you something useful.

...fellow binaries: Thomas A, Nathalie, Bram V, Bram S, Ken, Zuzanna and Sofie; Post-Docs, Riet, Yao-Cheng, Jens, Cindy, Steven M and Vanessa and fellow PhD-students Evangelia, Sandra, Bing, Phuong, Kevin, Emilio and Yao; Greg, Esther-Kristin, Tom, Eric, Anaga, Guy and Jonathan who remain in academia and those who turned to the *dark side* Steven R, Cedric, Tine B, Tine C and Xin-Ying. Without all of you the atmosphere in the group wouldn't have been the same.

...the IWT and UGent for supporting this work financially.

...Carola for providing a distraction-free place to write this thesis, your support and understanding.

...my brother and Nele, for providing distraction when needed.

...my parents and grandparents for their continuous support throughout my PhD and my studies leading up to it.

# Contents

*"Before beginning, prepare carefully"*

Marcus Tullius Cicero

# 0

# Research Purpose and Scope

As this research was initiated mid 2007, it is important to consider the state of plant comparative genomics at the time to fully understand our research goals. At the start of this thesis very few plant genomes were sequenced, though *Next Generation Sequencing* technologies were on the rise and a wave of novel plant genomes was expected in the near future. Furthermore, the available data sets were scattered here and there. To assist further research centralizing all genomic data was necessary.

Additionally, it is of importance to understand that during the course of this work new opportunities (within the general scope of this thesis) presented themselves, some of which worth pursuing and justified adjusting the initial goals.

## 0.1 Building a Comparative Genomics Framework

In order to efficiently deal with new plant genomes in a comparative analysis, a (semi-)automatic pipeline needs to be developed that chains standard tools (like BLAST[1], Tribe-MCL[2], MUSCLE[3] , ...), after adding all annotated plant genomes in a database. By importing the output of these tools into this database, various data types can be combined using relatively simple queries. As such, this database will become a means to start a wide variety of analyses not only for myself, but for group members as well.

Furthermore, this resource could contain valuable information for experimental biologists, therefore a user-friendly, web based, interface will be designed. From this website, information becomes retrievable without knowledge of MySQL queries and visualizations will be provided for complex data types. This will drastically lower the learning curve necessary to use comparative genomics data.

Finally, to convincingly show the merits of the platform, appealing case studies would need to be designed and performed.

## 0.2 Improved Detection of Collinearity

While not included in our initial set of goals, it became apparent (during the development of our pipeline) that i-ADHoRe 2.0[4], a tool essential to study genome evolution, was no longer usable given the large set of plant genomes available. As this tool is an essential step in our comparative genomics analysis pipeline, improving this tool to be able to handle several dozens of genomes was a requirement to include additional organisms in our platform.

The statistics at the core of i-ADHoRe 2.0 will need to be revised, as well as the alignment algorithm that is essential for the sensitive detection of collinearity. Furthermore, to increase the speed, support for modern hardware (like multithreading and message passing interface) needs to be implemented. Additionally, if the speed can be sufficiently increased, various parameter combinations can be evaluated to determine optimal settings for different organisms. Finally, a case study will be performed on the largest dataset available.

## 0.3 Study of Genome Evolution

From the start the study of genome evolution was a major objective of this thesis. Initially, we envisioned a study of gene loss after a large-scale duplication and the effects on speciation. Though, given the opportunity to study genome evolution in the framework of a new genome project, this goal was shifted to studying the effects of relatively recent whole genome duplications in these newly sequenced plant species.

*"Biology has at least 50 more interesting years."*

James D. Watson

# 1 Introduction

# “Biology has at least 50 more interesting years.”

James D. Watson mentioned this in 1984, about 30 years after he discovered the structure of DNA together with Francis Crick[5]. However just 50 years seems an underestimation, could he have foreseen the enormous diversity and complexity hidden in genomes of species around us? Scientists worldwide have had access to the human genome for over a decade, and so far it raised more questions than it gave answers[6]. Still many agree that, to date, the human genome continues to contribute significantly to various branches of biology[7]. We have now entered the post-genomic era, determining the genomic sequence of an organism is no longer an obstacle, but making sense out of it has become the next grand challenge. Numerous projects have appeared to fill the gaps in our knowledge using various approaches. The ENCODE project[8] tries to generate comprehensive functional data for a well defined portion of the genome of human and several model organisms[9,10]. Genome Wide Association Studies (GWAS) focus on linking variations (such as Single Nucleotide Polymorphisms (SNP)) with phenotypic traits[11]. Here we approach the problem using comparative genomics, using related genomes from different species, we attempt to gradually gain insights in how species evolve on the genomic level and thus how genomic changes ultimately allow them to adapt to their specific niche.

In this chapter, first a brief overview is given how, by comparative genomics, a better understanding can be gained in the evolution of genes and gene families. The next section goes into detail on what these techniques have revealed in plant genome evolution since the genome sequence of *Arabidopsis thaliana* was released in 2000[12]. The last section highlights what can be found in the other chapters of this thesis.

Section 1.3 has been redrafted from Proost et al.[13]. See page 21 for author contributions.

## 1.1 Comparative Genomics

Theodosius Dobzhansky boldly titled his famous 1973 essay 'Nothing in biology makes sense except in the light of evolution'[14]. However while this title might seem intentionally provocative at first, it harbors a great deal of truth. As differences observed in sequenced genomes contain adaptations to a specific ecological niche, the evolutionary background should never be neglected in order to ultimately understand the processes shaping genomes.

For instance, the evolutionary distance between species compared is of major importance for the features that can be studied[15]. Eg. while comparing yeasts and humans (that have a large phylogenetic distance), only basal eukaryotic features such as cell division and DNA replication can be studied. However, more recent adaptations of humans, can be studied by comparing the human genome to a closely related species such as Chimpanzee[16] or Orang-utan[17]. To find out what genomic features are associated with certain phenotypical marks (such as eye color, ...) one would need to compare genomes from several individuals to see which changes are usually associated with the trait of interest.

Most comparative studies will first attempt to detect similar features and thus conserved features in the genomes. This doesn't only form the basis for the comparative analyses itself, but it also allows unknown functional elements to be found based on the their conservation in various organisms (as conservation implies necessity and thus function). For most purposes one however can start from an annotated genome, where functional parts of the genome (such as genes with their introns and exons, RNA genes, transposable elements...) are already detected (aka. the structural annotation). A logical first step is to find similar genes in and between all compared organisms. This can efficiently can be done using BLAST[1]. While one could assume that gene-pairs, displaying sufficient similarity, are homologous (derived from a common ancestor), better results can be obtained by clustering the BLAST output using tribe-MCL[2] (details can be found in Section 2.3.1). The contents of such gene clusters is often referred to as a gene family. If one species has more genes in a gene family than others this point toward a duplication and retention of genes that might be well suited for that species' lifestyle. However to obtain detailed information in which lineage such expansions happened, the construction of a phylogenetic tree, that describes accurate relations between homologous genes is necessary[18](Section 2.3.3 discusses in detail how phylogenetic trees can be constructed). Using such trees, homologous genes can be further divided into orthologs and paralogs. Orthologs, on the one hand, are homologous genes derived from a speciation event. For instance the gene ADH1 (alcohol dehydrogenase 1A, the enzyme that breaks down ethanol) in human and mouse was present in the common ancestor of both and thus the human gene and its mouse counterpart were separated due to the speciation of both species. Also note that orthologous genes are usually assumed to have similar functions in different species. Hence correct delineation of orthologs is crucial to transfer knowledge from one (model-)organism to other (nonmodel-)organisms (Figure 1.1A). Paralogs, on the other hand, were created by duplication of the ancestral gene. Though more complex scenarios are possible as well; a duplication predating a speciation creates out-paralogs (Figure 1.1B) while duplicates constrained to a single lineage are referred to as in-paralogs (Figure 1.1A). In-paralogs that have an orthologous gene are called co-ortholog with that gene.

Until now homology has been used in the context of genes, however the term can be used in a broader context as well. As such, morphological structures derived from the same structure in the common ancestor, are defined as homologous. Indeed, the wings of a bat, the flippers of a whale and our own arms can be considered as homologous. In this work we've focused on homology at the genomic level, as large stretches of DNA, chromosome arms and even complete chromosomes can be homologous to each other. Though in this context the term hom**e**ologous is often used. Unlike at the gene level, where through BLAST and tribe-MCL homology is inferred based on sequence similarity, this measure is unsuited for homology on a larger scale. The sequences here are up to several orders of magnitude longer and, as they contain intergenic regions, more diverged. Therefore two other measures are used to detect homology between

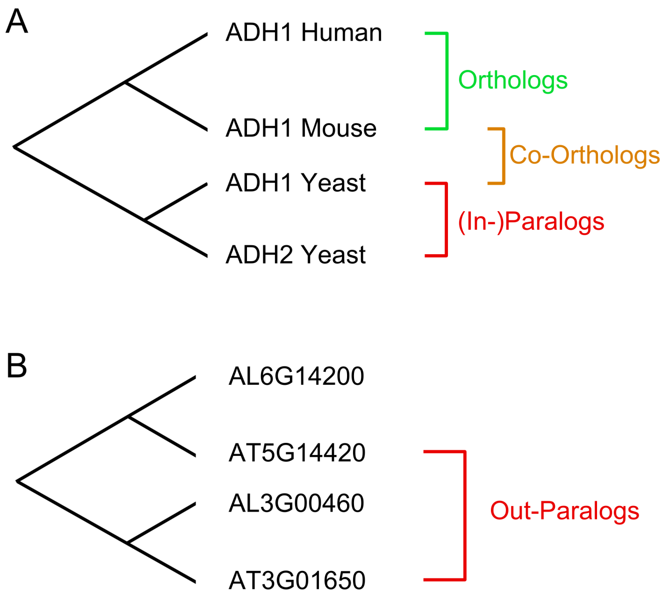


**Figure 1.1:** (A) Tree of alcohol dehydrogenase genes (pruned to clearly illustrate the relevant concepts) that shows an orthologous gene pair between human and mouse, a pair of (in-)paralogous genes in yeast that by definition can be considered co-orthologs with the human and mouse genes. (B) This example shows *Arabidopsis thaliana* and *Arabidopsis lyrata* genes that have been duplicated in their common ancestor, hence the two *Arabidopsis thaliana* genes now should be considered out-paralogs.

genomic regions, namely synteny and collinearity. Synteny requires a similar content in regions in and between genomes, while collinearity requires similar order as well. Phrased differently, two regions that have sufficient homologous genes in common can be considered syntenic. If those genes occur in (more or less) the same order they can be considered collinear as well.

Numerous tools emerged to detect both syntenic and collinear regions within and between genomes[19–22]. However few go beyond pairwise comparisons[4,23,24] and those who did weren't up to the task of analyzing dozens of genomes (See Section 3.3.3 and Figure 3.6). Therefore there was a clear need for a faster, efficient tool to detect collinearity that would use the additional genomes to its advantage to perform a more sensitive detection.

## 1.2 Platforms to Study Gene and Genome Evolution

At the end of 2007, when the development of PLAZA[25] started, few other comparative genomics resources for plants were available. Most contained only a subset of the available species (eg. GreenPhylDB[26] and Genome cluster database[27] which contained at the time only rice and *Arabidopsis thaliana*) or offered few data types (eg. Plant Genome Duplication Database[24,28]). Gramene[29], while being at the time rather comprehensive, was based on Ensembl[30] and lacked some plant specific functionality (eg. tools to study whole genome duplications (see Section 1.3), find orthologs in the presence of multiple duplications, ...).

So, despite the information available to the plant community, there was a clear need to centralize the available data and integrate it with tools specifically designed towards exploring plant genomes. From this need the development of PLAZA started (discussed in Chapter 2). An additional requirement was that building the platform had to be automated to a high degree. At the time development started we anticipated that the number of plant genomes would, like

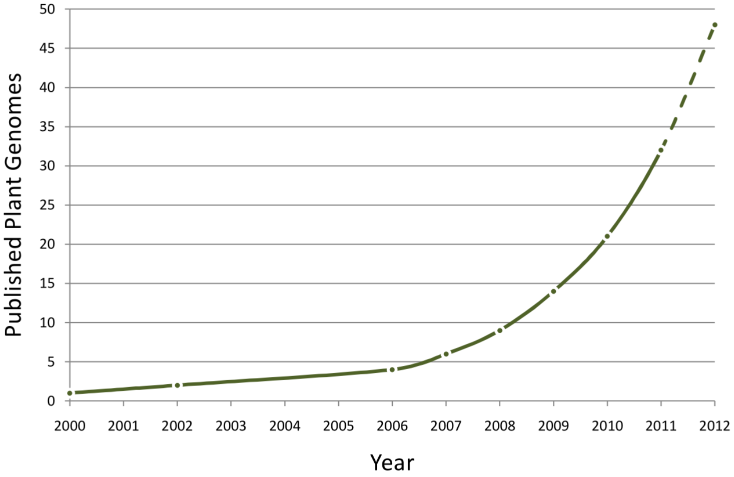


**Figure 1.2:** Overview of the number of published plant genomes (cumulative), and a prospect what to expect in the next year.

the number of sequenced prokaryotic genomes, grow at an ever increasing rate. Now we know this is indeed the case and the number of published plant genomes rises nearly exponentially, doubling in less than two years (Figure 1.2). Without a proper pipeline, rebuilding the platform to include newly sequenced organisms would be extremely laborious. Especially in the context of genome projects (See Chapter 5) a change in assembly or annotation requires restarting the analysis, hence a swift building procedure is essential.

# 1.3 Genome Evolution in Plants

## 1.3.1 The Ancestral Angiosperm Genome

### (TP1, Figure 1.3)

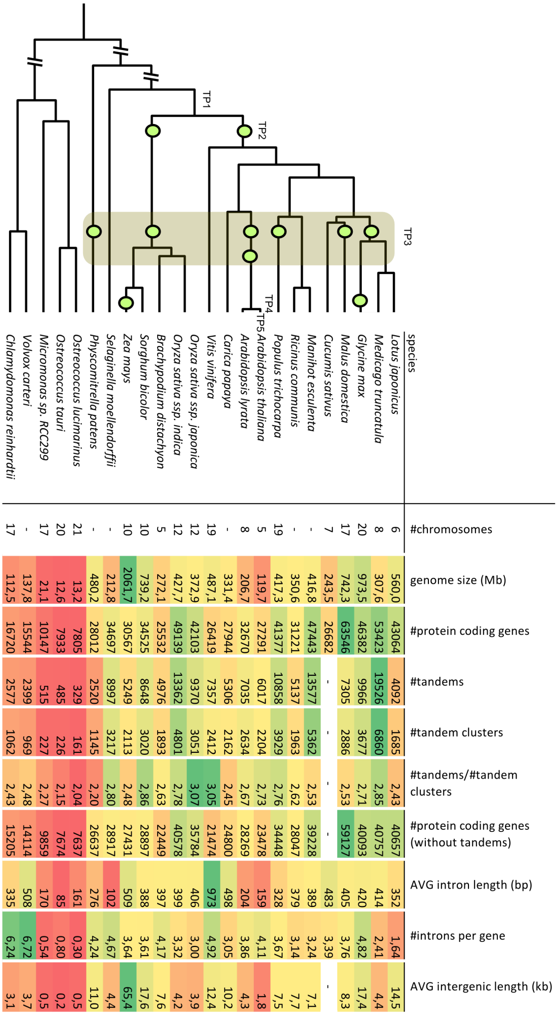
For several decades, *Arabidopsis thaliana* has been an excellent plant model organism for reasons well known[31]. Additionally, various techniques are available to genetically engineer *Arabidopsis thaliana*[32]. Furthermore, within the family of the Brassicaceae, many species are of major economical value. Important food crops include broccoli, cabbage (both *Brassica oleracea* ssp.) and mustard (*Brassica rapa/nigra*, *Sinapis alba*), while rapeseed (*Brassica napus*) is used to produce oils and more recently became a source for biodiesel. All this contributed to the popularity of *Arabidopsis thaliana* in plant laboratories worldwide. Last but not least, there is the small genome size of *Arabidopsis thaliana*. With the recent advances in sequencing technologies, determining a genome sequence can be considered almost routine. A decade ago, however, genome sequencing was still a daunting, very expensive and laborious task and the size of the genome to be sequenced was a major determinant in whether or not a genome project was initiated. Coincidentally, with a size of about 125 Mb, the genome of *Arabidopsis thaliana* was also one of the smallest plant genomes known and therefore an ideal target for sequencing[12].

Analysis of the *Arabidopsis thaliana* genome, and comparison with other plant genomes that have been determined subsequently, unveiled a very complex evolutionary history of the genome and that of its dicot ancestors. Although being a superb model system for plant geneticists, the genome of *Arabidopsis thaliana* actually might be rather exceptional, with its many genome duplications, huge amount of gene losses, and recent genome shrinkage (TP4–TP5, Figure 1.3). In this section, covering some 150 million years of angiosperm evolution, we discuss some milestones in the evolution of the *Arabidopsis thaliana* genome and that of its ancestors, which eventually have led to the genome we know today.

In earlier studies, based on a mathematical model that simulates the birth and death of genes through small- and large-scale gene duplication events, we estimated that the ancestral angiosperm genome contained no more than 14 000 genes[33]. Although this was solely based on the analysis of the *Arabidopsis thaliana* genome, similar values have been obtained through the comparison of different plant genomes. For instance, comparing the *Arabidopsis thaliana* and poplar (*Populus trichocarpa*) gene sets suggested an ancestral gene count of 12 000[34], whereas clustering of homologous genes from *Arabidopsis thaliana*, rice and 32 other plant species delineated 12 400 ancestral genes[35]. Recently, counting the number of genes that show cross-species synteny between the genomes of *Arabidopsis thaliana*, grapevine (*Vitis vinifera*), papaya (*Carica papaya*) and poplar, suggested 10 000 – 13 000 ancestral angiosperm genes[24]. In conclusion, it is probably safe to say that the ancestral angiosperm genome contained around 12 000 – 14 000 genes. Gene counts in extant angiosperm genomes are all considerably larger (data derived from PLAZA 2.0[a] [25]), due to the continuous process of gene duplication[36] and, in numerous cases, genome duplications (see further).

Although the moss *Physcomitrella patens* seems to contain a number of genes that is comparable to that of many angiosperms, probably also due to a genome duplication event, the gene content is considerably different[37]. Unicellular green algae on the other hand contain much fewer genes, as might be expected from their much simpler morphology, lifestyle, and ecology. *Volvox carteri*[38] and *Chlamydomonas reinhardtii*[39] contain more than 15 000 and 16 000 genes, respectively, while the picoeukaryotic algae *Micromonas* and *Ostreococcus* contain about 10 000 and 8000 genes, respectively. It is interesting to note that the difference in gene count between the prasinophytes *Ostreococcus* sp.[40] and *Micromonas* sp.[41] and the Chlorophyceae *Volvox carteri* and *Chlamydomonas reinhardtii* seems to be mainly due to duplicated genes present in the latter two species, but generally missing in the former ones.

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



| species | #chromosomes | genome size (Mb) | #protein coding genes | #tandems | #tandem clusters | #tandems/#tandem clusters | #protein coding genes (without tandems) | AVG intron length (bp) | #introns per gene | AVG intergenic length (kb) |
|---|---|---|---|---|---|---|---|---|---|---|
| *Lotus japonicus* | 6 | 560,0 | 43064 | 4092 | 1685 | 2,43 | 40657 | 352 | 1,64 | 14,5 |
| *Medicago truncatula* | 8 | 307,6 | 53423 | 19526 | 6860 | 2,85 | 40757 | 414 | 2,41 | 4,4 |
| *Glycine max* | 20 | 973,5 | 46382 | 9966 | 3677 | 2,71 | 40093 | 420 | 4,82 | 17,4 |
| *Malus domestica* | 17 | 742,3 | 63546 | 7305 | 2886 | 2,53 | 59127 | 405 | 3,76 | 8,3 |
| *Cucumis sativus* | 7 | 243,5 | 26682 | - | - | - | - | 483 | 3,39 | - |
| *Manihot esculenta* | - | 416,8 | 47443 | 13577 | 5362 | 2,53 | 39228 | 389 | 3,24 | 7,1 |
| *Ricinus communis* | - | 350,6 | 31221 | 5137 | 1963 | 2,62 | 28047 | 379 | 3,14 | 7,7 |
| *Populus trichocarpa* | 19 | 417,3 | 41377 | 10858 | 3929 | 2,76 | 34448 | 328 | 3,67 | 7,5 |
| *Arabidopsis thaliana* | 5 | 119,7 | 27291 | 6017 | 2204 | 2,73 | 23478 | 159 | 4,11 | 1,8 |
| *Arabidopsis lyrata* | 8 | 206,7 | 32670 | 7035 | 2634 | 2,67 | 28269 | 204 | 3,86 | 4,3 |
| *Carica papaya* | - | 331,4 | 27944 | 5306 | 2162 | 2,45 | 24800 | 498 | 3,05 | 10,2 |
| *Vitis vinifera* | 19 | 487,1 | 26419 | 7357 | 2412 | 3,05 | 21474 | 973 | 4,92 | 12,4 |
| *Oryza sativa ssp. japonica* | 12 | 372,9 | 42103 | 9370 | 3051 | 3,07 | 35784 | 406 | 3,00 | 3,9 |
| *Oryza sativa ssp. indica* | 12 | 427,7 | 49139 | 13362 | 4801 | 2,78 | 40578 | 399 | 3,32 | 4,2 |
| *Brachypodium distachyon* | 5 | 272,1 | 25532 | 4976 | 1893 | 2,63 | 22449 | 397 | 4,17 | 7,6 |
| *Sorghum bicolor* | 10 | 739,2 | 34525 | 8648 | 3020 | 2,86 | 28897 | 388 | 3,61 | 17,6 |
| *Zea mays* | 10 | 2061,7 | 30567 | 5249 | 2113 | 2,48 | 27431 | 509 | 3,64 | 65,4 |
| *Selaginella moellendorffii* | - | 212,8 | 34697 | 8997 | 3217 | 2,80 | 28917 | 102 | 4,67 | 4,4 |
| *Physcomitrella patens* | - | 480,2 | 28012 | 2520 | 1145 | 2,20 | 26637 | 276 | 4,24 | 11,0 |
| *Ostreococcus lucimarinus* | 21 | 13,2 | 7805 | 329 | 161 | 2,04 | 7637 | 161 | 0,30 | 0,5 |
| *Ostreococcus tauri* | 20 | 12,6 | 7933 | 485 | 226 | 2,15 | 7674 | 85 | 0,80 | 0,2 |
| *Micromonas sp. RCC299* | 17 | 21,1 | 10147 | 515 | 227 | 2,27 | 9859 | 170 | 0,54 | 0,5 |
| *Volvox carteri* | - | 137,8 | 15544 | 2399 | 969 | 2,48 | 14114 | 508 | 6,72 | 3,7 |
| *Chlamydomonas reinhardtii* | 17 | 112,5 | 16720 | 2577 | 1062 | 2,43 | 15205 | 335 | 6,24 | 3,1 |

**Figure 1.3:** Schematic and highly pruned phylogenetic tree of green algae and land plants for which the genome sequence has been determined. The background colour of cells indicates whether values are small (red), intermediate (yellow) or high (green) compared to the average value in the same column. Dots on the tree denote whole genome duplications. TPx denote specific time points discussed in the text. Raw data is derived from PLAZA 2.0[25]. (AVG, Average).

### 1.3.2 The Hexaploid Ancestor of Core Eudicot Plants

#### (TP2, Figure 1.3)

Early analysis of the *Arabidopsis thaliana* genome unveiled several rounds of Whole Genome Duplications (WGDs), although the exact number and timing has been disputed[42–45]. For instance, it was initially suggested that one of the WGDs detected in *Arabidopsis thaliana* occurred before the radiation of most eudicots, and that the oldest WGD predated the divergence of dicots and monocots[43,44]. By comparison with additional whole plant genomes however, a more complete picture has emerged. In particular the genomes of grapevine and papaya revealed conclusive evidence regarding the exact number of WGDs that occurred early in the history of eudicots[46,47], furthermore these finding added further constraints on the possible time of the duplications. Grapevine is an early-diverging rosid and regions in the grapevine genome typically show homology with two other regions elsewhere in the same genome. Because of this triplicate genome structure, it was concluded that, most likely, three ancestral genomes had contributed to the grapevine lineage[46]. The recently released papaya genome shows a similar triplicate genome structure, although papaya is not closely related to grapevine[47]. Instead, it belongs to the order Brassicales and is more closely related to *Arabidopsis thaliana* from which it diverged 70 Million Years Ago (MYA)[47,48]. Therefore, the most plausible and parsimonious explanation would be that the triplicate genome structure is ancient and shared between many, if not all eudicots. This is further supported by analysis of partial genome data of the asterid Coffea[49] and EST data of several other Asteraceae, as well as by the recent completion of two additional rosid genomes, soybean (*Glycine max*)[50] and apple (*Malus domestica*)[51]. By comparing the pattern of gene losses in homeologous segments in papaya and grapevine, it was observed that two of three were more fractionated, suggesting that a first duplication event generated a tetraploid, which then hybridized with a diploid to generate a triploid. This triploid then underwent yet another WGD event to generate a hexaploid, giving rise to the triplicate genome structure we still find in species such as grapevine and papaya[52]. Uncovering the triplicate genome structure in other plant genomes is more difficult because of additional

WGD events that have occurred in several of these lineages[53].

The extant grapevine genome consists of 19 chromosomes, most of which are clearly syntenic to two other chromosomes, hence the triplicate genome structure. Furthermore, two chromosomes show synteny to two different chromosomes, indicating chromosome fusions[46]. This particular structure would suggest that, about 120 MYA, the ancestral pre-hexaploid genome from which all dicots have evolved, consisted of seven chromosomes. This would also suggest that, subsequent to the hexaploidy event, the ancestral post-hexaploid genome would have consisted of 21 chromosomes[46,54]. Possibly, amongst the ones available at this moment, the grapevine genome is the genome that still resembles that ancestral chromosomal state most, due to its slow rate of evolution[46].

There is some evidence, albeit mostly circumstantial, that these early duplications can be linked to the origin and fast diversification of angiosperms[55–57]. Gene and genome duplications potentially facilitate reproductive isolation[36,58–60] and increase the diversifying potential of species thereby providing putative selective advantages over their diploid progenitors[61–64]. Although their exact timing is uncertain, the hexaploidization event early in the evolution of flowering plants might have facilitated the emergence of new, more complex, flower morphologies and specialized pollination strategies[65]. This in turn might have been one of the crucial factors in the rapid diversification and speciation of flowering plants in the Early Cretaceous[55,56,66,67] and, if true, make the abominable mystery Darwin referred to somewhat less of a mystery.

### 1.3.3 Two More Genome Duplications for Arabidopsis

#### (TP3, Figure 1.3)

Apart from the hexaploidy shared by most eudicots, many plant lineages show traces of additional, independent and more recent genome duplications[68–73]. Interestingly, many independent WGDs, such as those in the cereals, the legumes, the Solanaceae, the Compositae, cotton (*Gossypium hirsutum*), poplar, banana (Musa sp.), and apple (*Malus domestica*) appear to have occurred somewhere between 50 and 70 MYA[34,72,74,75]. Recently, it has been suggested that these duplication events might have coincided with the Cretaceous-Tertiary (K-T) extinction, the most recent large-scale mass extinction that wiped out around 80% of plant and animal species, including the dinosaurs[57,74].

Also the ancestors of *Arabidopsis thaliana* seem to have undergone two additional genome duplications. Again, this has been uncovered through comparison with a close(r) relative, namely papaya, which has not shared these genome duplications. Figure 1.4 shows several sets of homologous regions in the genomes of papaya and *Arabidopsis thaliana*. As can be observed, the one genome copy in papaya corresponds with four copies in *Arabidopsis thaliana*, providing convincing support for two genome duplications in the lineage leading to *Arabidopsis thaliana* since their divergence from papaya, about 70 MYA[24,28,47,48]. These findings were unexpected as other methods, relying on fossil evidence and phylogenetic trees to calibrate molecular clocks, placed both duplications considerably earlier. However, fossils for the Brassicales are rare and therefore few reliable age constraints could be used. Only recently, more advanced methods have been developed that can account for uncertainties in tree topology and allow evolutionary rates to be uncorrelated across the tree[76]. Recent age estimates now also place one WGD very close to the divergence from papaya and the most recent WGD within a window of 23 to 43 MYA[68,74].

From Figure 1.3, it also becomes clear why inferring the number of WGDs proved difficult using only the *Arabidopsis thaliana* genome. Homologous segments in *Arabidopsis thaliana* are often highly degenerated due to extensive gene loss. Indeed, as previously noted, high frequencies of gene loss (or gene fractionation sensu Freeling et al.[77]) reduce collinearity resulting in duplicated regions that share very few, if any, homologous genes[78]. Nevertheless, by comparing chromosomal segments across multiple genomes, and in particular with genomes that have not shared the duplication event(s), such highly degenerated regions

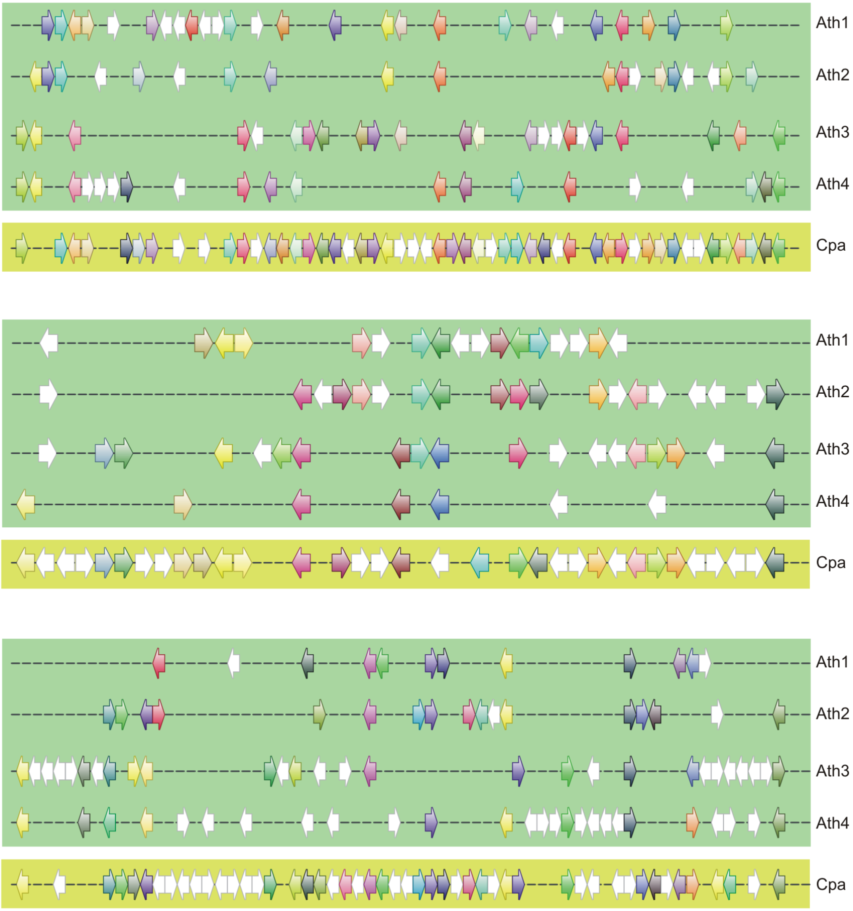


**Figure 1.4:** Collinearity between papaya and duplicated regions in *Arabidopsis thaliana*. In general, one region in papaya corresponds with four homologous regions in *Arabidopsis thaliana*, providing strong evidence for two WGDs in *Arabidopsis thaliana* since its divergence from papaya, approximately 70 MYA. Ath: *Arabidopsis thaliana*; Cpa: *Carica papaya*

can often still be unveiled to be homologous[28,52,79].

Using the papaya genome, it also became possible to estimate how much gene translocation has occurred in *Arabidopsis thaliana*, since their divergence. Starting from collinear regions between both species, the chromosomal positions of *Arabidopsis thaliana* genes were scored based on the conservation of homologous neighboring genes in papaya[77]. Although the frequency of translocation varied among different gene families and functional categories, Freeling and coauthors estimated that about 25% of all *Arabidopsis thaliana* genes had translocated since the origin of the Brassicales. Therefore, both massive gene loss and gene translocations seem to be responsible for the highly degenerated patterns of collinearity observed in intra-genome *Arabidopsis thaliana* comparisons (Figure 1.3).

Previously, we estimated that the number of genes created by the hexaploidy event and surviving until today amounted to about 800. Furthermore, we estimated that the number of genes that have survived both of the more recent WGDs in *Arabidopsis thaliana* is about 6700. The number of genes created through continuous small-scale duplications since the core eudicot ancestor has been estimated to be about 5300[33]. Given the current size of the *Arabidopsis thaliana* genome, with about 27 000 genes annotated (Figure 1.3), we estimate that the ancestor of the core eudicots had around 14 000 genes. It should be noted though that these values will be different for different plant species, dependent on the rate genes get duplicated and lost again.

## 1.4 Chapter Overview

At the start of this work plant comparative genomics was, with only five genomes available (three angiosperms, one moss and one alga), still in its infancy. Therefore, a considerable amount of effort was put into developing a comparative genomics platform, that could be used as a basis for future analyses. This platform, coined PLAZA, is described in Chapter 2. However as time progressed more and more genomes were sequenced and additional updates of the platform were done. For an overview of new features and the species currently included in PLAZA, please visit http://bioinformatics.psb.ugent.be/plaza/.

As new genomes were continuously being released during the course of this research it quickly became apparent that the tool used to study genome evolution, i-ADHoRe, needed to be updated to be able to cope with the increase in data. After implementing a more memory efficient way to store homologous genes (version 2.4) i-ADHoRe could be used for the first public release of PLAZA containing nine species. However as development of PLAZA continued and the number of sequences surpassed this improvement alone proved to be insufficient to cope with the increase in data. In Chapter 3 is describe how various improvements ultimately reduced runtimes of i-ADHoRe significantly, while Chapter 4 introduced a novel alignment algorithm that was necessary as previous implementations did not perform adequately when then number of gene lists aligned grew too large.

In Chapter 5 several case studies, illustrating the merits of both PLAZA and i-ADHoRe are presented. As the opportunity arose to show the full potential of our tools on two newly sequenced genomes, the relevant sections of the resulting publications have also been included in this chapter.

Finally conclusions about this research as a whole and thoughts of how we can progress in the future are described in Chapter 6. In addition some of the most recent novelties in the field, and their potential implications are discussed here.

## 1.5 Author Contribution

As first author I wrote this review together with Yves Van de Peer. All images shown in this chapter are made by me.

*“If we knew what it was we were doing, it wouldn’t be called* ***research****, would it?”*

Albert Einstein

# 2

# PLAZA Comparative Genomics 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 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.[25]. Author contribution, see page 52.

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

## 2.1 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[80], 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[35,81]. 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[12,82]. 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[83] have been published, and new genome initiatives will at least double the number of plant genome sequences by the end of this decade[84,85].

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[86,87]. 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[88–90]. 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[91]. 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[34]. 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[46]. 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[37,39].

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)[92,93]. 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[94,95].

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[96], 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[a]. 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.

## 2.2 Results

### 2.2.1 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

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

by the different sequencing centers yielded a data set of 295 865 gene models, of which 92.6% represent protein-coding genes (Table 2.1). 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[a] genome browser[97]. 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 2.1). 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 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 2.1).

Complementary to the structural annotation, we also retrieved, apart from free-text gene descriptions, functional information through Gene Ontology (GO) associations[98], InterPro domain annotations[99], and Arabidopsis Reactome[b] pathway data[100]. 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 pro-

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

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

**Table 2.1:** Summary of the Gene Content in PLAZA v1. (a) Size assembled (sequencing method). PAC, phague artificial chromosome; TAC, transformation-competent artificial chromosome; WGS, whole-genome shotgun. (b) percentage of protein coding genes. (c) 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. (d) Percentages in parentheses include projected GO annotations, while the first value only reports original primary GO data.

| Species | Genome Size (a) | Genes (b) | Scaffolds (c) | Coding | GO (d) | InterPro |
|---|---|---|---|---|---|---|
| *Arabidopsis thaliana* | 115 Mb (BAC, PAC, TAC) | 33,284 (88.81%) | 5 | 27,228 | 63.62% (66.21%) | 56.49% |
| *Carica papaya* | 271 Mb (3x WGS) | 28,072 (99.84%) | 1,898 | 28,072 | 0.00% (22.88%) | 57.75% |
| *Populus trichocarpa* | 410 Mb (7.5x WGS) | 45,699 (99.90%) | 19+1 (5,724) | 45,654 | 44.69% (52.89%) | 61.91% |
| *Vitis vinifera* | 468 Mb (8.4x WGS) | 38,127 (99.63%) | 19+1 (35) | 37,987 | 40.09% (45.90%) | 57.62% |
| *Oryza sativa* | 371 Mb (BAC, PAC) | 57,955 (72.32%) | 12 | 41,912 | 30.42% (30.91%) | 63.69% |
| *Sorghum bicolor* | 626 Mb (WGS) | 34,686 (99.78%) | 10+1 (217) | 34,609 | 44.44% (48.13%) | 67.79% |
| *Physcomitrella patens* | 480 Mb (8.6x WGS) | 36,137 (99.80%) | 1,446 | 36,065 | 33.20% 42.44% | |
| *Chlamydomonas reinhardtii* | 121 Mb (13x WGS) | 14,731 (99.64%) | 552 | 14,678 | 34.99% | 49.29% |
| *Ostreococcus lucimarinus* | 13 Mb (WGS) | 7,805 (100.00%) | 21 | 7,805 | 47.94% | 62.86% |
| Total | | 295,865 (92.60%) | | 273,965 | 39.36% | 44.88% |

vides 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 2.1).

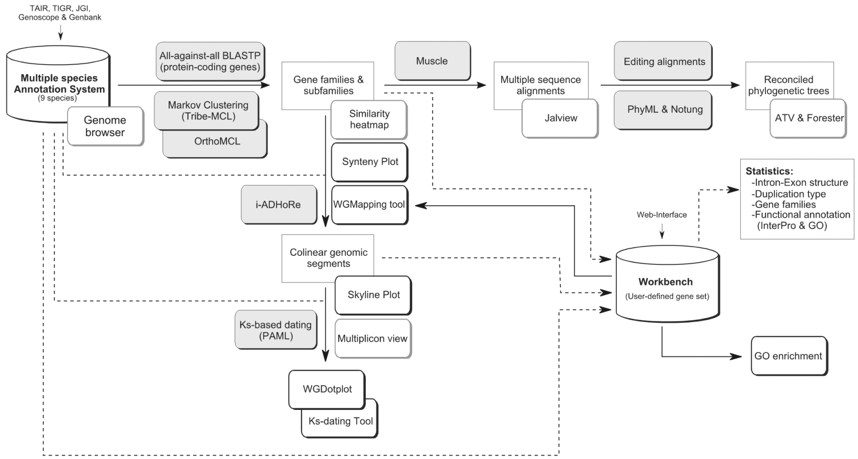

TAIR, TIGR, JGI, Genoscope & Genbank
Multiple species Annotation System (9 species)
Genome browser
All-against-all BLASTP (protein-coding genes)
Markov Clustering (Tribe-MCL)
OrthoMCL
Gene families & subfamilies
Similarity heatmap
Synteny Plot
WGMapping tool
i-ADHoRe
Colinear genomic segments
Skyline Plot
Multiplicon view
Ks-based dating (PAML)
WGDotplot
Ks-dating Tool
Muscle
Multiple sequence alignments
Jalview
Editing alignments
PhyML & Notung
Reconciled phylogenetic trees
ATV & Forester
Statistics:
-Intron-Exon structure
-Duplication type
-Gene families
-Functional annotation (InterPro & GO)
Web-Interface
Workbench (User-defined gene set)
GO enrichment

**Figure 2.1:** Structure of the first 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.

### 2.2.2 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[1]. 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[2]) 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 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[94]. 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 2.3.1). The OrthoMCL method[101] 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[94]. 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 2.3.2,[102]). 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 2.2). 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 2.2). 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.

### 2.2.3 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[103]. Based on the gene families generated using protein clustering, a phylogenetic pipeline was applied to construct 20 781 phylogenetic trees covering ~172 000 protein-coding genes. Bootstrapped phylogenetic trees were constructed using the maximum likelihood method PhyML[104] based on pro-

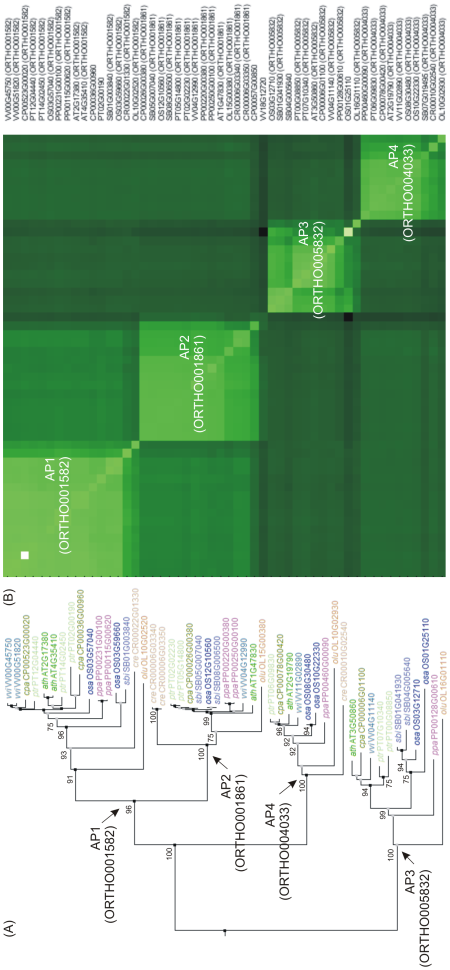


**Figure 2.2:** Gene Family Delineation Using Protein Clustering, Phylogenetic Tree Construction, and Similarity Heat Maps.

(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 $\geq$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. Species abbreviations as in Table 2.2.

tein multiple sequence alignments generated using MUSCLE[3] (see Section 2.3.3). 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[105]. 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[106]. Therefore, we calculated a duplication consistency score, originally developed by Ensembl[107], to identify erroneously inferred duplication events (see Section 2.3.3). 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[100,108,109]. 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 2.3.4 and Figure 2.3). 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 2.1). 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 informa-

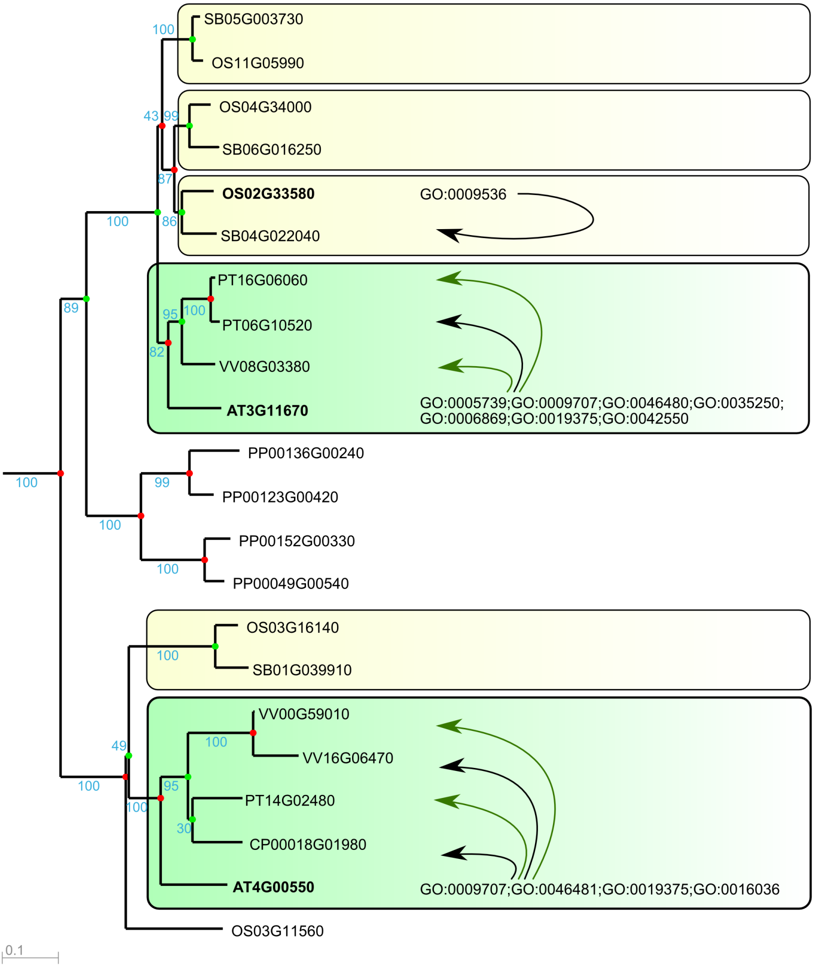


**Figure 2.3:** GO projection using eudicot and monocot orthologous groups. The rounded boxes indicate the orthologous groups extracted from the phylogenetic tree while green and yellow shadings refer to eudicot and monocot clades, respectively. If for genes in an orthologous group functional annotation 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).

tion for 36 473 genes. This projected information covers ∼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 ∼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[25]. Whereas Blast2GO, a high-throughput and automatic functional annotation tool[110], 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 evidence 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).

### 2.2.4 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[4]. 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[78]. 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[78,111].

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 2.4A). 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 2.3.6). $K_S$ 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 underesti-

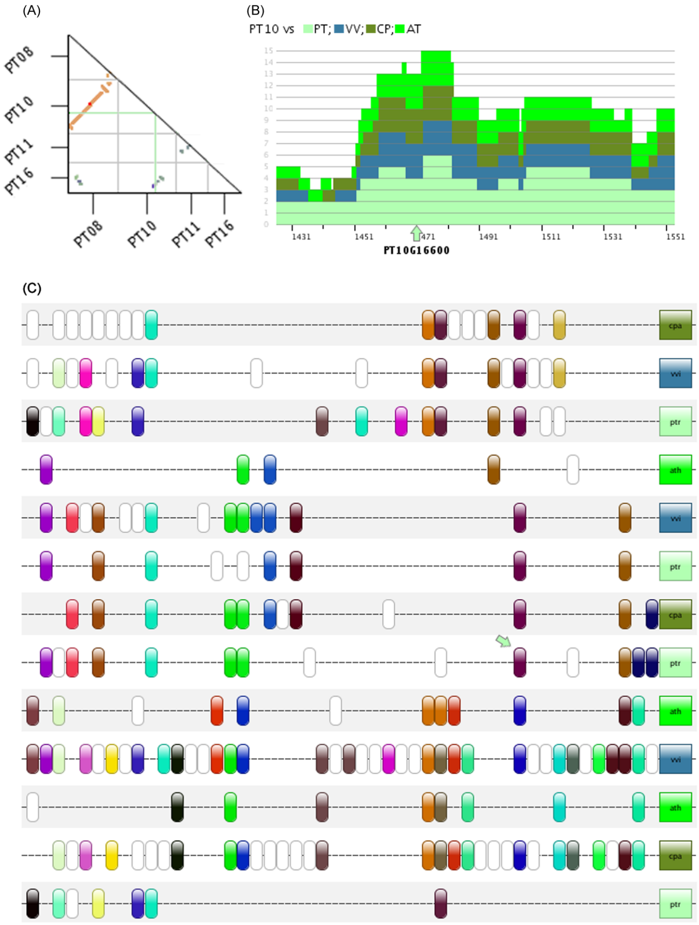


**Figure 2.4:** Overview of Different Collinearity-Based Visualizations of the Genomic Region around Poplar Gene PT10G16600. (A) The WGDotplot shows that the gene of interest, indicated by the light-green line, is located in a duplicated block between chromosomes PT08 and PT10. The orange color refers 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).

mation of the actual age[112]. The average $K_S$ for a collinear (or duplicated) block is calculated and colored accordingly in the WGDotplots (Figure 2.4A). 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. As shown in Chapter 5, 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[28,53].

When investigating genomic homology between more than two genomes, the Skyline plot provides a rapid and flexible way to browse multiple homologous genomic segments (Figure 2.4B). 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 2.3.5). 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 2.4C). The gene alignment algorithm maintains the original gene order but will introduce gaps to place homologous genes in the same column (if possible).

### 2.2.5 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 2.1). 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 2.5), 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.

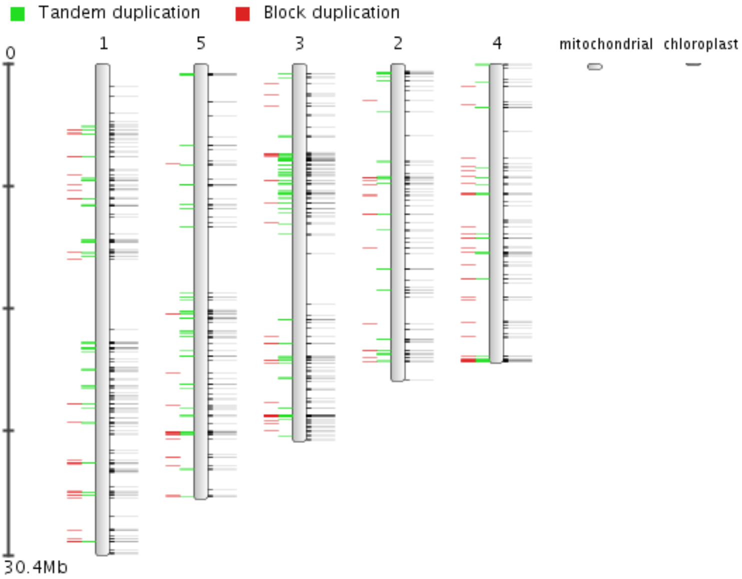


**Figure 2.5:** Whole Genome Mapping tool. Overview of 664 *Arabidopsis thaliana* genes with a Cyclin-like F-box domain (IPR001810).

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 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.

### 2.2.6 Comparison with Other 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[87]. There has been a rapid growth in the number of plant genomics databases (Table 2.2). A major difference between these databases is the number of organisms included: whereas the Genome Cluster Database[27] and GreenPhylDB[a][26] only include Arabidopsis and rice, Gramene[b][29], PLAZA, and CoGe[c][113] have the most comprehensive set of species. CoGe includes, besides fully sequenced plant genomes, a large collection of viral, bacterial, fungal, and animal genomes. Comparing the data types, a noticeable trend is that most platforms focus on either gene families or genomic homology. Genome Cluster Database, GreenPhylDB, OrthologID[d][114], and PlantTribes[e][115] all provide detailed information about gene families and phylogenetic trees but do not have any means to study genomic homology. By contrast, Plant Genome Duplication Database[f][28], SynBrowse[g][116], and CoGe provide meth-

[a] http://greenphyl.cirad.fr/

[b] http://www.gramene.org/

[c] http://synteny.cnr.berkeley.edu/CoGe/

[d] http://nypg.bio.nyu.edu/orthologid/

[e] http://fgp.bio.psu.edu/tribedb/index.pl

[f] http://chibba.agtec.uga.edu/duplication/

[g] http://www.synbrowse.org/

ods to study synteny and collinearity but do not include information about gene families. Phytozome[a] and Gramene partially combine gene family and genome evolution data types. Whereas the former provides family-based local synteny plots, the collinearity framework in Gramene is based solely on genetic markers. Intraspecies dot plots are available in the Plant Genome Duplication Database, CoGe, and PLAZA and make it possible to investigate genes originating fromWGD events. Finally, only Gramene, CoGe, and PLAZA provide a genome browser to obtain a general overview of a genomic region of interest.

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[b] [117], AGRIS[c] [118], and GRASSIUS[d] [119]. The complementary platforms Phytome [120] and SPPG[e] [35] are hybrid systems integrating gene information from genome sequencing projects with EST data for a comprehensive set of plant species.

## 2.3 Methods

### 2.3.1 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

[a] http://www.phytozome.net

[b] http://plntfdb.bio.uni-potsdam.de/

[c] http://arabidopsis.med.ohio-state.edu/

[d] http://grassius.org/plantgenome.html

[e] http://bioinformatics.psb.ugent.be/cgi-bin/SPPG/index.htpl

**Table 2.2:** Features of Plant Comparative Genomics Tools (a) Species names are abbreviated: *Arabidopsis lyrata* (Aly), *Arabidopsis thaliana* (Ath), *Brachypodium distachyon* (Bdi), *Carica papaya* (Cpa), *Chlamydomonas reinhardtii* (Cre), *Glycine max* (Gma), *Lotus japonica* (Lja), *Medicago trunculata* (Mtr), *Ostreococcus lucimarinus* (Olu), *Oryza sativa* (Osa), *Physcomitrella patens* (Ppa), *Populus trichocarpa* (ptr), *Sorghum bicolor* (Sbi), *Selaginella moellendorffi* (Smo), *Vitis vinifera* (Vvi), *Volvox carteri* (Vca), and *Zea mays* (Zma).(b) Phytozome has a synteny viewer instead of a genuine colinearity pipeline.(c) CoGe includes also viral, prokaryotic, and other, nonplant, eukaryotic genomes.(d) Gramene has some features to visualize macrocolinearity based on marker maps.

| Tool | Species (a) | Gene Families | Phylogenetic Trees | WGDotplots | Inter Species Colinearity | Functional Annotation | Genome Browser | Comments |
|---|---|---|---|---|---|---|---|---|
| **PLAZA** | 9 (Ath, Cpa, Ptr, Vvi, Osa, Sbi, Ppa, Olu & Cre) | X | X | X | X | X | X | Multi-species collinearity views (Skyline Plot & Multiplicon view), $K_s$ -dating tool, Family-wise similarity heatmap and Workbench. |
| **Genome Cluster Database** | 2 (Ath & Osa) | X | X |  |  | X |  | Chromosome map and link with *Arabidopsis* expression data. |
| **GreenPhylDB** | 2 (Ath & Osa) | X | X |  |  | X |  | Manual curation of a subset of families. |
| **OrthologID** | 3+2 (Ath, Ptr & Osa + Ppa and Cre as outgroup) | X | X |  |  |  |  | Diagnostic characters per orthologous group. |
| **Plant Genome Duplication Database** | 7 (Ath, Cpa, Ptr, Mtr, Vvi, Osa & Sbi) |  |  | X | X |  |  | Genome-wide mapping tool for homologous sequences and syntenic locus search. |
| **Phytozome (b)** | 14 (Ath,Aly,Cpa, Ptr, Vvi, Mtr, Gma, Osa, Bdi, Sbi, Zma, Smo, Ppa & Cre) | X |  |  | +/- | X |  |  |
| **PlantTribes** | 5 (Ath, Cpa, Ptr, Mtr & Osa) | X |  |  |  | X |  | Link with *Arabidopsis* expression data. |
| **CoGe (c)** | 14 (Ath, Cpa, Ptr, Mtr, Lja, Vvi, Osa, Sbi, Zma, Ppa, Smo, Olu, Cre, Vca,...) |  |  | X | X | X | X | DNA based sequence comparisons (Conserved Non-coding Sequences). |
| **SynBrowse** | 3 (Ath, Mtr, Lja) |  |  |  | X |  |  | Synteny browser based on GBrowse (no intra-species colinearity). |
| **Gramene (d)** | 6 (Ath, Osa, Ptr, Vvi, Sbi & Zma) | X | X | +/- | +/- | X | X | Based on the Ensembl pipeline. |

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[1]. 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.

### 2.3.2 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.[25]). 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[102].

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.

### 2.3.3 Alignments and Phylogenetic Trees

For all gene families multiple sequence alignments were created using MUSCLE[3]. 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[a] 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[b] is used as multiple sequence editor[121] to view and transfer sequence data to the user's PC. ATV/Archaeopteryx[c] is used for tree visualization[122].

### 2.3.4 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 2.3). 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

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

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

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

all primary gene annotations Inferred from Electronic Annotation (evidence code IEA) were excluded as information source (see Supplemental Table 5 accompanying Proost et al.[25]). 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 distribuiton and Bonferroni correction for multiple hypothesis testing.

### 2.3.5 Detection of Collinearity

To detect collinearity within and between species i-ADHoRe 2.4 was used[4]. Whereas the algorithm is identical to the i-ADHoRe 2.0 version, a more efficient way to store gene pairs in memory was implemented 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.

### 2.3.6 Relative Dating using Synonymous Substitutions

Only collinear gene pairs were dated using $K_S$. Compared to dating all pairwise 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[33] 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)[123] 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)[124] 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.

## 2.4 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[125] 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.

## 2.5 Author Contribution

As a first author, I had a lead role (along with Michiel Van Bel and Klaas Vandepoele) in the development of the PLAZA platform. Several visualizations shown throughout this chapter are based on tools I designed and wrote (see Figure 2.2 and Figure 2.4). Development of new methods, such as GO-projection, was done by myself as well was . Finally, the manuscript was written by Klaas Vandepoele and myself.

*"I'll be honest - we're throwing science at the wall here to see what sticks. No idea what it'll do. Probably nothing."*

Cave Johnson

# 3

# i-ADHoRe 3.0 - Fast and Sensitive Detection of Genomic Homology in Extremely Large Data Sets

## Abstract

Comparative genomics is a powerful means to gain insight into the evolutionary processes that shape the genomes of related species. As the number of sequenced genomes increases, the development of software to perform accurate cross-species analyses becomes indispensable. However, many implementations that have the ability to compare multiple genomes exhibit unfavorable computational and memory requirements, limiting the number of genomes that can be analyzed in one run. Here, we present a software package to unveil genomic homology based on the identification of conservation of gene content and gene order (collinearity), i-ADHoRe 3.0, and its application to eukaryotic genomes. The use of efficient algorithms and support for parallel computing enable the analysis of large-scale data sets. Unlike other tools, i-ADHoRe can process the Ensembl data set, containing 49 species, in less than one hour. Furthermore, the profile search is more sensitive to detect degenerate genomic homology than chaining pairwise collinearity information based on transitive homology. From ultra-conserved collinear regions between mammals and birds, by integrating coexpression information and protein-protein interactions, we identified more than 400 regions in the human genome showing significant functional coherence. The different algorithmical improvements ensure that i-ADHoRe 3.0 will remain a powerful tool to study genome evolution.

This chapter is based on Proost et al.[126]. Author contribution, see page 77.

# 3.1 Introduction

During their evolution, genomes have been altered at various levels. At the smallest scale, point mutations and small insertions and deletions[127] affect only a few nucleotides. Larger modifications include duplication, deletion, translocation or inversion of a single gene or genomic segment[128]. At the largest scale, the entire genome can be doubled via genome duplication or merging[53,57,61]. Identification of these structural rearrangements provides insight into how genomes have evolved and diverged over time. It is therefore of crucial importance to correctly determine chromosomal regions that are homologous (i.e. derived from a common ancestor), either within a genome, or between genomes of related species. Genomic homology can be inferred from collinearity, namely the conservation of both gene content and gene order. Synteny, though initially defined as 'the property of being located on the same chromosome'[129], is often used to indicate the conservation of gene content but not necessarily gene order[28]. Like collinearity, synteny also points to homology between different genomic regions based on a number of shared genes[79,130].

Detection of collinear regions between the genomes of related species allows for the identification of chromosomal fusions and fissions, along with inverted or translocated regions. Additionally, gene loss and gain can be efficiently estimated, and cross-species genome analysis provides a framework for transferring gene annotation and biological information to newly sequenced genomes. Finally, orthologous intergenic sequences derived from collinear regions can be screened for conserved non-coding regions as a way to detect regulatory motifs and to identify various types of RNA genes[90]. As both gene loss and different types of rearrangements accumulate over time, the resulting genome erosion gradually reduces the degree of collinearity between species. Therefore, gene order preserved over a large phylogenetic distance can imply a biological constraint[131].

Collinear regions within a genome can also hint at the occurrence of one or more rounds of whole-genome duplications (WGDs)[44,79]. Based on within-genome collinearity, the loss of gene duplicates created during a WGD can be estimated[25,132,133], whereas the functions of genes retained in duplicate can be linked to lineage or species-specific adaptations, including specific pathways and biological processes. WGDs appear to have played a crucial role in the evolution of all major eukaryotic lineages and, particularly in plants, they are often associated with key events during evolution including fast adaptive radiation[53,134] and survival of mass extinction events[74]. Additionally, gene family expansions critical for the pome fruit development in apple (*Malus domestica*)[51] have been linked to a recent WGD, whereas expansions in genes producing aromatic compounds have been observed in grapevine (*Vitis vinifera*)[46]. Although remnants of several recent WGDs are abundant in the plant kingdom, WGDs in land vertebrates and fishes are seemingly much older[135,136]. In vertebrates, the complex body plan is often attributed to the duplication of developmental genes during two WGDs 450 Million Years Ago (MYA)[136]. The first traces of a WGD have been unveiled in *Saccharomyces cerevisiae* based on comparative approaches[137]. Additional proof for the WGD in brewer's yeast has been provided later by comparison with the genome of an unduplicated outgroup species, *Kluyveromyces waltii*[138]. The more complex carbohydrate metabolism of *Saccharomyces cerevisiae* and other post-duplication yeast species is probably a direct consequence of this duplication[139]. Therefore, the discovery of large-scale duplications, through the study of collinear regions, has provided a remarkably detailed view on the genomic evolution and adaptation of various species.

Here, we focus on the accurate detection of homologous chromosomal segments both within and between the genomes of related species. Specifically, sensitive and accurate algorithms are needed for the identification and evolutionary analysis of duplicated regions that have undergone massive gene loss. Several tools, by means of various approaches, have recently been proposed[19–21,140]. Whereas most tools only perform pairwise comparisons, the iterative Automatic Detection of Homologous Regions (i-ADHoRe)[4] was one of the first that simultaneously analyzed genomes of multiple species and allowed for the detection of highly diverged collinear regions. On the one hand, i-ADHoRe has been used in several genome projects to uncover the remnants of large-scale duplications (e.g., apple (*Malus domestica*)[51], soybean (*Glycine Max*)[50], *Arabidopsis lyrata*[141] and black cottonwood (*Populus trichocarpa*)[34]), and, on the other hand, to detect inter-species collinearity in yeasts[142] and Archaea[143]. In contrast to tools that infer genomic homology through a multiple sequence alignment of complete genomic DNA sequences[144–147], i-ADHoRe detects genomic homology through the identification of gene collinearity and/or synteny. The core feature of i-ADHoRe 3.0, which is based on a new alignment algorithm[148] and improved statistical evaluation, is the ability to handle large numbers of genomes. Due to the further optimization of many algorithmic steps, the current version of i-ADHoRe 3.0 is roughly 30 times faster than the previous version. In addition, i-ADHoRe 3.0 can now take advantage of a parallel computing platform, reducing the runtime even further. For large data sets, the combination of improvements in the sequential algorithm and the parallelization results in overall speedup of a factor of 1000. Here, we demonstrate that i-ADHoRe is capable of processing much larger datasets than the current state-of-the-art tools. In particular, the complete Ensembl release 57[149] data set that contains 49 eukaryotic genomes can be analyzed in less than one hour (using 64 CPU cores), while producing highly accurate results.

## 3.2 Material and Methods

### 3.2.1 Data Sets

The *Arabidopsis thaliana* and *Vitis vinifera* genomes together with gene family information were retrieved from PLAZA, an on-line plant comparative genomics resource[25] that provides gene families constructed with Tribe-MCL clustering[2] starting from an all-against-all BLAST[1] protein similarity search. The E-values and bit-scores were saved, because these values are necessary for Cyntenator[23] and MCScan[24]. The lengths for *Carica papaya* gene lists were also obtained via PLAZA. Animal genomes and families were downloaded from Ensembl (release

57) with the Ensembl Perl API[150]. An all-against-all BLAST protein similarity search was done to obtain bit-scores and E-values[a].

### 3.2.2 Detection of Collinearity

The initial steps of the algorithm (Figure 3.1) are identical to i-ADHoRe 2.0; tandem duplicated genes are mapped to a single representative and for each pair of gene lists a gene homology matrix (GHM) is generated (Figure 3.2A). In this sparse matrix pairs of homologous genes are represented as dots and as such collinear regions will appear as dense diagonals. Compared to the previous i-ADHoRe version, several major components of the algorithm were re-implemented for a better performance. First, the statistical validation of the clusters in the GHM was improved. To avoid inclusion of diagonals in the GHM generated merely by chance, the significance of each cluster is now estimated with a statistical model that takes into account the overall background density of the matrix. When multiple seeds (i.e. clusters with at least three homologous gene pairs that meet the initial criteria) were found, a correction for multiple hypothesis testing was done either with the Bonferroni or False Discovery Rate (FDR)[151,152] method.

Significant collinear regions found during this initial detection were converted into a profile, both collinear regions were aligned, i.e. homologous genes are placed in the same column adding gaps where necessary (Figure 3.2B). Like in previous versions of i-ADHoRe this alignment can be done by progressively applying the Needleman-Wunsch (pNW) algorithm or a greedy graph (GG) based alignment strategy. In version 3.0, a novel greedy graph based alignment algorithm (GG2), described in Fostier et al.[148], was implemented. Using this aligned profile a new search is performed (Figure 3.2C), here a GHM is created with the profile and all gene lists in the dataset. Significant regions are added to a new profile and the profile search is repeated (Figure 3.2D). With a single profile, multiple segments

[a] An overview of all included species in PLAZA and Ensembl can be found on http://bioinformatics.psb.ugent.be/plaza/ and http://www.ensembl.org/ respectively.

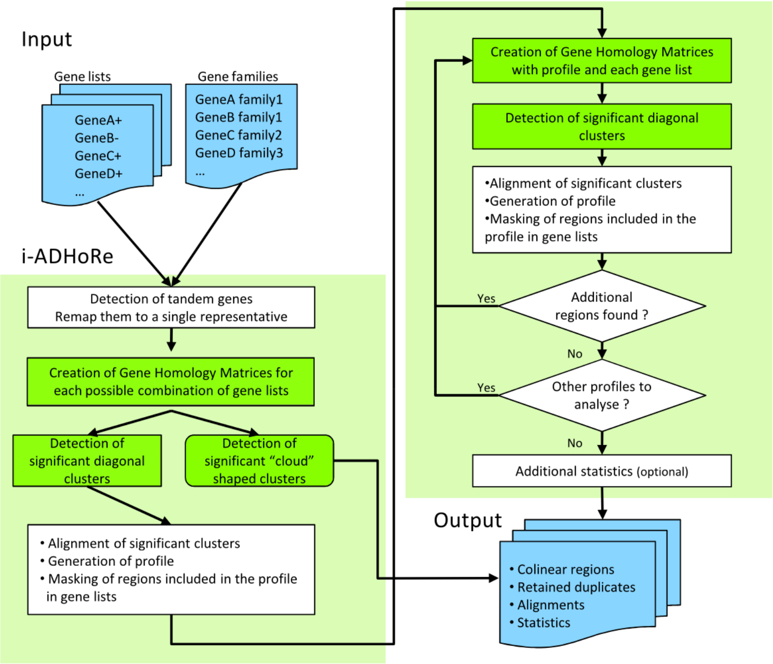

Input
Gene lists
GeneA+
GeneB-
GeneC+
GeneD+
...
Gene families
GeneA family1
GeneB family1
GeneC family2
GeneD family3
...
i-ADHoRe
Detection of tandem genes
Remap them to a single representative
Creation of Gene Homology Matrices for each possible combination of gene lists
Detection of significant diagonal clusters
Detection of significant “cloud” shaped clusters
• Alignment of significant clusters
• Generation of profile
• Masking of regions included in the profile in gene lists
Creation of Gene Homology Matrices with profile and each gene list
Detection of significant diagonal clusters
•Alignment of significant clusters
•Generation of profile
•Masking of regions included in the profile in gene lists
Yes
Additional regions found ?
No
Yes
Other profiles to analyse ?
No
Additional statistics (optional)
Output
• Colinear regions
• Retained duplicates
• Alignments
• Statistics

**Figure 3.1:** Flowchart of the i-ADHoRe detection strategy. Steps indicated by green filled boxes can be executed in parallel.

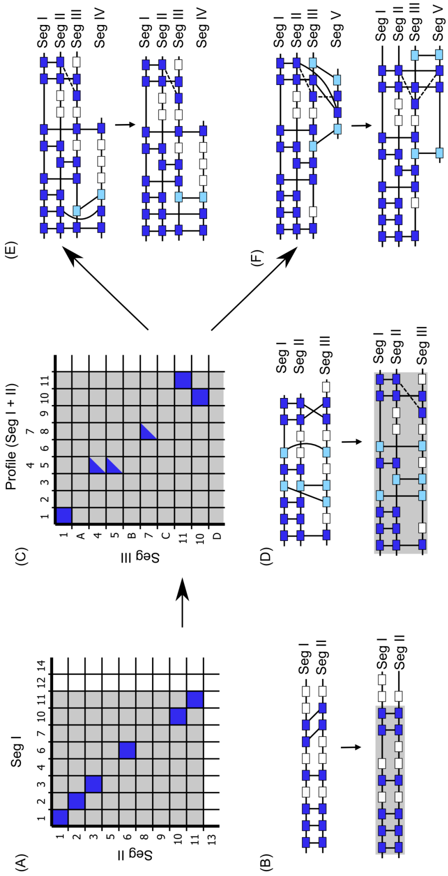


**Figure 3.2:** Gene homology matrix (GHM) for the initial two segments (Seg I and Seg II). In a GHM collinear regions will appear as dense diagonals. (B) Alignment of shared homologs between collinear regions; gaps are introduced to place as many homologous pairs in the same column as possible (35). The alignment (or 'profile') now contains the information of both segments. (C) Start of the iterative process, GHMs are now created with the profile and additional collinear regions can be found, e.g. Seg III. (D) Generation of a new profile. As long as additional segments can be found steps (C) and (D) are repeated. In this example (E) and (F) show how a single profile detects two additional segments that are mutually non-homologous (Seg IV and V), leading to a split in the detection process.

can be found that are homologous to the profile but not necessarily to each other (Figure 3.2 E and F). In this case, several profiles are generated and the detection algorithm continues detection with the longest profile first. Once no additional segments can be found the search continues with the next profile. Additionally, the initial pairwise and profile searches can now be executed on a parallel computing platform (a multiprocessor/multicore systems or a computational cluster of networked computers). If N denoted the number of gene lists provided as an input to i-ADHoRe, the $N(N+1)/2$ pairwise comparisons could be processed independently and, hence, distributed over different processes. The size of each gene list was taken into account to ensure a good load balance between the processes. At the end of this step, the detected collinear regions are communicated (using the Message Passing Inteface (MPI)) among the processes. Similarly, a single profile search can be parallelized by distributing the N gene lists among the different processes, again taking the size of the chromosomes into account. At the end of every profile search, the detected collinear regions were again communicated between the processes. However, due to the much smaller task granularity of one single profile search, a good load balancing was more difficult to achieve.

### 3.2.3 Synteny Mode

The input and initial steps were identical to the collinearity search mode. Only after the pairwise GHMs had been built, a different clustering algorithm was used (Figure 3.1). Clouds of dots contained by a bounding box were detected. Initially, the method started by considering all the dots of the GHM as potential cloud seeds. Subsequently the seeding algorithm searched a rectangular area, defined by the *cloud_gap*, for additional dots, and all dots in this window formed the seed cloud. Next, all seed clouds would grow by adding all dots present in a frame with a thickness equal to the *cloud_gap* to the current cloud. This process was repeated as long as additional genes could be included in the cloud. Finally, clouds within each other's range, defined by the *cluster_cloud_gap*, were merged into one single large cluster.

In a final step, the statistical significance of the clouds was calculated. Two methods were available. One method used a binomial distribution in which the probability density was set to the number of dots divided by the area of the dot matrix. The other method took into account the removal of the tandem duplicates during a pre-processing step. Therefore, one dot per column and row was assumed to be present. For boxes larger than the *tandem_gap*, this assumption might be broken and the significance might be slightly overestimated. The binomial distribution supposed that one dot might be present at every position in the box. Hence, the second distribution would seemingly be a more realistic measure for the statistical significance of a cloud. As clouds could not be aligned, the profile search was automatically disabled with the cloud search.

### 3.2.4 Empirical Estimation of False Positive Rates

False positive (FP) rates were calculated with permutation tests in which 100 randomized data sets were compared with a real reference data set. Tandem duplicated genes (homologs within a window of 70 genes) were removed prior to shuffling the reference data set to generate a randomized version. This pre-processing step guaranteed a comparable density in the randomized run because breaking up tandem-clusters artificially increased the GHM background density. All genes had their original orientation replaced with a randomly assigned one. The lengths of the original gene lists were maintained during the randomization, but genes could be moved from one gene list to another. To estimate the performance with different settings, a permutation test was carried out for each of the desired settings, generating parameter landscapes for *Arabidopsis thaliana*, human and yeast, with various combinations of *q_value* and *gap_size* parameters. Settings that yielded the maximum amount of anchor points, while maintaining a FP rate near the selected cut-off value were considered as optimal.

### 3.2.5 Comparison with MCScan and Cyntenator

BLASTP pairs for MCScan were filtered and only the best five hits in each species were retained[24]. Because in MCScan first proteins are clustered to group homologous genes in gene families, this step was excluded when monitoring runtimes for the different tools. Cyntenator was also run with filtered BLASTP output, retaining only the top five hits for each species if their bit score was within 95% of the highest bit-score (as described in Rödelsperger et al.[22]). The gap and mismatch penalties were set to -0.3, the threshold to 2 and the filter to 1000. i-ADHoRe was run with a *gap_size* of 30 and *cluster_gap* of 35, while keeping the *prob_cutoff* on 0.01 and the *q_value* on 0.75. GG2 was used as the alignment algorithm and correction for multiple hypothesis testing was done with FDR. The minimal number of anchor points in a cluster was set to five.

### 3.2.6 Detection of Highly Conserved Regions Enriched for Coexpressing and Interacting Gene Pairs

Phylogenetic profiles, describing the number of homologous regions per species present in a multiplicon, a set of mutually collinear regions, were generated for all multiplicons in the output from the high-quality Ensembl subset. Multiplicons with one human and one bird (either chicken or zebra finch) segment and with conserved segments of at least five other mammals were selected. From these regions the human segment was identified and the genes collinear with genes from other segments were stored. Expression data were derived from COXPRESdb version c3.1[153] and highly expressed gene pairs were selected based on a mutual rank below or equal to 50. Experimentally characterized interacting protein pairs (41 088 binary interactions for 9142 human genes) were downloaded from IntAct[154]. Using Ensembl's BioMart tool, a conversion table was generated to map all gene identifiers in these data sets to the Ensembl genes. For each selected multiplicon, the length of the human segment and number of human collinear genes were determined. Then, the number of coexpressed or interacting pairs was counted. When at least one human gene pair was found, the statistical significance was tested with a permutation test. Over 10 000 iterations, a random segment from the human genome (with the tandem duplicated genes removed) was sampled with the same length as the selected multiplicon. From the random region, an equal number of genes was randomly selected as collinear and, the number of coexpressed or protein-protein interaction pairs in this gene set was established. The number of iterations in which a number of pairs was equal or larger than that found in the real data set were counted and used to calculate a p-value for each multiplicon. All regions with a p-value $< 0.05$ were considered significant.

### 3.2.7 Evaluation of Low-Quality Genomes

To artificially reduce the quality of the *Arabidopsis thaliana* genome, the gene list length distribution of the papaya genome was used as a template to split the *Arabidopsis thaliana* gene lists in fragments resembling a draft assembly. i-ADHoRe was executed on both the *Arabidopsis thaliana* genome and the artificial low-quality version. The collinear fractions were measured by enabling the *write_stats* option in i-ADHoRe.

## 3.3 Results and Discussion

### 3.3.1 The i-ADHoRe 3.0 algorithm

The detection strategy of i-ADHoRe 3.0 is shown in Figure 3.1 [4,155]. First, tandem duplicated genes are mapped onto one single representative gene, because tandem clusters can hinder the detection of diagonals (see further). Next, for each pair of chromosomes or scaffolds, a so-called gene homology matrix (GHM) is generated. A GHM is a sparse matrix in which homologous gene pairs are marked by dots and collinear regions appear as diagonals. For each detected diagonal, the statistical significance is evaluated (Figure 3.2A). Significant collinear regions are aligned into a profile (Figure 3.2B) that contains the combined gene content of the two collinear regions and can hence be used as a more sensitive probe to scan for additional collinear regions (Figure 3.2 C and D). This step

**Table 3.1:** FP-rate on the basecluster level on the Arabidopsis dataset. Parameter combinations with a FP-rate equal or below the p-value cutoff specified ($10^{-2}$) are indicated in bold.

| | $r^2$ **0,5** | $r^2$ **0,6** | $r^2$ **0,7** | $r^2$ **0,8** | $r^2$ **0,9** | $r^2$ **1,0** |
|---|---|---|---|---|---|---|
| **gap 15** | **2,74E-04** | **2,74E-04** | **2,74E-04** | **2,75E-04** | **3,33E-04** | **0,00E+00** |
| **gap 20** | **1,33E-03** | **1,34E-03** | **1,40E-03** | **1,39E-03** | **1,52E-03** | **0,00E+00** |
| **gap 25** | **4,72E-03** | **4,70E-03** | **4,54E-03** | **4,58E-03** | **4,33E-03** | **0,00E+00** |
| **gap 30** | 1,02E-02 | **1,00E-02** | **9,84E-03** | **9,21E-03** | **9,36E-03** | **0,00E+00** |
| **gap 35** | 2,17E-02 | 2,10E-02 | 1,98E-02 | 1,86E-02 | 1,81E-02 | **0,00E+00** |
| **gap 40** | 3,97E-02 | 3,76E-02 | 3,50E-02 | 3,22E-02 | 2,89E-02 | **0,00E+00** |
| **gap 45** | 7,43E-02 | 7,00E-02 | 6,53E-02 | 5,92E-02 | 5,01E-02 | **0,00E+00** |
| **gap 50** | 1,27E-01 | 1,18E-01 | 1,11E-01 | 9,94E-02 | 8,36E-02 | **0,00E+00** |
| **gap 55** | 1,97E-01 | 1,81E-01 | 1,65E-01 | 1,46E-01 | 1,21E-01 | **0,00E+00** |

is iterated as long as new collinear regions are found and mutually homologous regions are grouped into a multiplicon. Even though the profile search requires an increased computational cost, it has proven its merits as a means to detect more degenerate genomic homology[4,44,78].

In order to deal with increasingly large data sets, various parts of the original i-ADHoRe code[155] have been replaced by equivalent algorithms with a reduced computational complexity. A first major improvement was the development of an efficient statistical model to estimate the significance of diagonals in the GHM, because the computational cost to calculate the exact p-value[156] increases exponentially with the number of gene pairs that shape the diagonal. The Arabidopsis thaliana data set was analyzed with different p-value thresholds and an empirical false positive (FP) rate for each threshold was determined using permutation tests (Figure 3.3). The combination of better heuristics and the implementation of a correction for multiple hypothesis testing (Bonferroni or FDR) resulted in a more realistic estimation of p-values and consequently improved the control of the FP rate compared to the previous statistical model. The effects of using different parameter settings on *Arabidopsis thaliana* are reported in Tables 3.1 but similar results were obtained for other human and yeast.

In the iterative search procedure, additional collinear regions are identified and the corresponding profiles are updated in every step. Therefore, an accu-

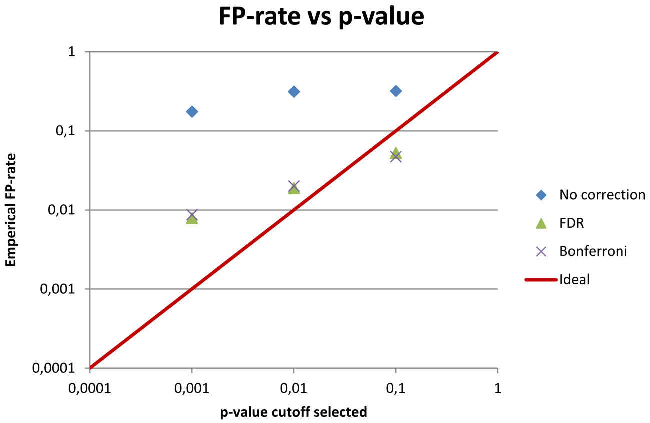


**Figure 3.3:** Correlation between empirical FP rate and the selected p-value cut-off. Inclusion of a correction for multiple-hypothesis testing (FDR and Bonferroni) results in an observed p-value closely reflecting the selected value. The recommended p-value range is $10^{-3}$ to $10^{-1}$.

rate alignment algorithm is imperative for the sensitive discovery of more degenerate collinear regions (Figure 3.2 C and D). Originally, i-ADHoRe relied on the progressive application of the pairwise Needleman-Wunsch (pNW) algorithm to align multiple homologous segments into profiles[155]. Whereas with the Needleman-Wunsch algorithm an optimal pairwise alignment of two segments can be obtained, its quality quickly degrades due to the propagation of erroneous decisions in early alignment steps when additional segments are added[157]. To resolve this issue, a greedy, graph-based (GG) aligner had been introduced into i-ADHoRe 2.0 that converted the alignment problem into a cycle-canceling problem in a graph[4]. Whereas this implementation provided a viable solution for the 'once a gap, always a gap' problem, it was unable to outperform the pNW aligner in terms of number of correctly aligned homologous genes. In i-ADHoRe 3.0, a novel greedy, graph-based aligner (GG2) was featured that, by means of maximum flow calculations in the graph, resolved efficiently unalignable sections in the graph (conflicts). Even though this graph-based method is computationally more intensive than the application of the pNW aligner, fast heuristics allow this algorithm to be efficiently used[148] (see Chapter 4).

Finally, two practical issues arise when multiple genomes are compared: the processing time and the memory requirements. Whereas the runtime increases super-linearly with the size of the data set, i.e. faster than the number of genomes that are analyzed, the memory requirements are mainly determined by the number of homologous gene pairs. To limit the runtime and, hence, facilitate the analysis of large-scale data sets, the two most time-consuming parts of the algorithm were parallelized (Figure  3.2, green boxes): the initial all-to-all pairwise comparison (every gene list versus every gene list) and the iterative profile searches (one profile versus every gene list). The parallelization of the all-to-all pairwise step revealed that by using a dataset of 31 high-quality genomes (Ensembl release 57, all genomes sequenced up to 6x using WGS or better quality) and 64 CPU cores, a 46-fold increase in speed (Figure  3.4) was observed. Searching additional collinear regions in a gene list using a profile is more difficult to parallelize, because of more intense communication requirements between the subtasks and hence a larger communication overhead. Overall, the runtime for the complete algorithm was reduced 32-fold on 64 cores, corresponding to a parallel efficiency (relative reduction in runtime compared to one with one single core, over the number of cores used) of approximately 50%.

### 3.3.2 The synteny mode and ancient duplications

Whereas collinearity is excellent to detect remnants of relatively recent duplications and homologous regions between closely related species, more ancient homologous regions may remain undetected[79]. Synteny is a valid, albeit less stringent, alternative than collinearity to detect ancient homology between regions that experienced severe rearrangements, such as, for example, paralogous regions that originated from the whole-genome duplication (WGD) in the common ancestor of all vertebrates 350–450 MYA[135]. i-ADHoRe 3.0 features an additional clustering algorithm to detect genomic homology based on shared gene content, coined the synteny mode.

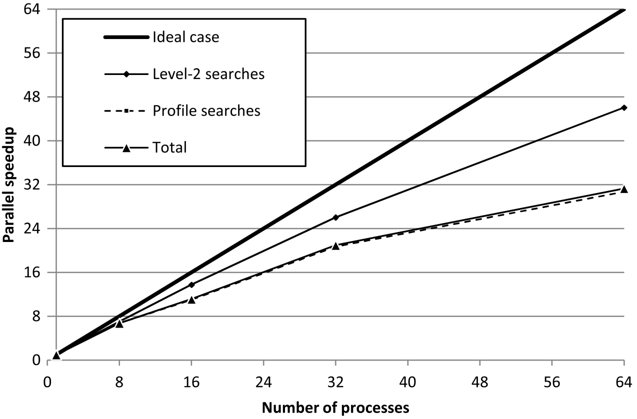


**Figure 3.4:** Parallel speed-up in function of the number of processes used. The profile searches have a harder load to balance (smaller granularity) and therefore are not as efficient to run in parallel compared to the level 2 detection.

For the human data set, the empirical FP rate was determined with a permutation test on several datasets, every time gradually increasing the *cloud_gap* from 5 to 55 and setting the *cloud_cluster_gap* to the *cloud_gap* plus five. We found that a *cloud_gap* of 15 (and thus a *cloud_cluster_gap* of 20) was the closest to the selected p-value and thus optimal for the human genome. Therefore, these settings were applied for further evaluation of the synteny mode. (Table 3.2). Note that a wrong choice of parameters might cause an avalanche effect in which the bounding box keeps growing because new dots are found in the window frame.

On this data set with the collinear search, 544 anchor point pairs were reported and while using the cloud search 2215 were found. For both runs the p-value was set to 0.01. Comparison of the number of block-duplicated genes revealed that the synteny mode detected nearly 4-fold more genes in significant syntenic blocks. Therefore, the synteny mode is recommended to detect highly diverged homologous regions, such as those derived from the WGD in vertebrates but also between species with a large evolutionary distance.

**Table 3.2:** Empirical Estimated FP-Rates using the synteny mode on the human dataset (p-value cutoff $10^{-2}$) In bold settings with a p-value near or better than the selected value are indicated.

| *Cloud_gap* Settings | **FP-Rate** |
|---|---|
| 5 | **4,27E-04** |
| 10 | **5,25E-03** |
| 15 | **2,14E-02** |
| 20 | 5,42E-02 |
| 25 | 1,06E-01 |
| 30 | 1,78E-01 |
| 35 | 2,77E-01 |
| 40 | 3,98E-01 |
| 45 | 5,22E-01 |
| 50 | 6,47E-01 |
| 55 | 7,90E-01 |

### 3.3.3 Evaluation of gene-based collinearity detection tools

When genomes with remnants of WGDs are dealt with or when highly diverged genomes are compared, gene loss and different types of rearrangements can interfere with the accurate detection of duplicated or homologous collinear regions[28,79]. To the best of our knowledge, only Cyntenator[23], MCScan[24] and i-ADHoRe go beyond simple pairwise comparison and combine, via different approaches, information to find additional homologous regions. Cyntenator performs progressive pairwise combinations based on a user-defined species tree that strictly imposes the order in which genomes are compared. Only valid alignments including homologous regions from all species are retained to find collinearity with the next genome in line. Unlike the profile search of i-ADHoRe, in MCScan each chromosome is used as a reference and all pairwise collinear segments are mapped, followed by a multiple alignment procedure of homologous genes, inspired by the threaded blockset aligner[147]. MCScan allows pairing regions that had initially not been detected based on their collinearity with the reference, a method referred to as "transitive homology"[155]. Unlike some tools[19–21,140], Cyntenator, MCScan and i-ADHoRe use ordered gene lists rather than the actual genome sequence. This level of abstraction allows for an efficient detection of collinearity. An additional advantage is that more diverged intergenic sequences do not interfere with

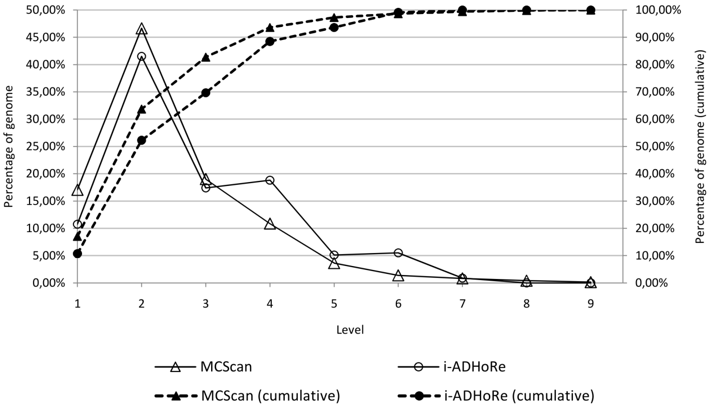


**Figure 3.5:** Distribution of the fraction of genes (*n*) found in sets of homologous genomic segments (multiplicons) with different levels (*m*) by MCScan and i-ADHoRe, respectively. Level 1 indicates the fraction of genes that was not found in any collinear region. The cumulative curve (i.e. the sum of all genes with the indicated level or lower) remains lower for i-ADHoRe, indicating that a larger fraction of the genome could be grouped into higher level multiplicons.

the discovery of ancient collinearity or synteny.

To benchmark the application of a profile search versus transitive homology mapping of pairwise collinear segments, i-ADHoRe and MCScan were executed on the *Arabidopsis thaliana* data set to identify degenerated duplicated segments. Cyntenator was excluded from this experiment, because it does not allow detection of internal duplications. Figure 3.5 shows the number of genes present in regions with a certain level, indicating the total number of homologous segments. Although i-ADHoRe and MCScan use very different approaches, the number of genes in collinear regions was comparable (23 912 and 24 559, respectively), but the profile search enabled i-ADHoRe to group more genes in regions with level four (4499 versus 2669 genes), five (1223 versus 891) and six (1318 versus 340). This result implies that the more advanced profile search allows for a more sensitive detection of collinear regions compared to the progressive chaining in MCScan.

To evaluate the discovery of inter-species collinearity, the three tools were applied to analyze a small subset of the genomes available in Ensembl, namely human[158] (*Homo sapiens*), chimpanzee[16] (*Pan troglodytes*), mouse[159] (*Mus musculus*), chicken[160] (*Gallus gallus*) and pufferfish[161] (*Tetraodon nigroviridis*). For each gene, all overlapping homologous segments were retrieved and the highest number of species found in one single alignment (or multiplicon) was scored. In contrast to Cyntenator, MCScan and i-ADHoRe collapse tandem genes into one single representative and, therefore, reported fewer genes. The predefined species order applied by Cyntenator to compare genomes forms a major drawback for large-scale analyses including multiple species. For instance, a region that is collinear between human and mouse, but for which the homologous counterpart in the chimpanzee lineage was lost, will not be reported because only collinear regions from the first pairwise comparison (i.e. human and chimpanzee) are retained to identify additional collinearity in mouse. Therefore, a fair comparison was possible only for regions in which collinearity was conserved in all five species. Whereas using MCScan and Cyntenator, 416 and 498 genes were assigned to such regions, respectively, the profile search applied by i-ADHoRe allocated 3296 genes in multiplicons containing regions conserved in all five species (Figure 3.7).

Fast algorithms that exhibit a favorable computational complexity are imperative to keep pace with the ever-increasing number of available genomes. Therefore, the runtime of all three programs was first monitored on the data set of the five species. i-ADHoRe, the only tool that takes advantage of a parallel environment, was executed using a single and eight threads respectively on a multicore machine. Because MCScan first clusters proteins into gene families, a step is not part of the actual collinearity detection algorithm, the program runtime was measured without this pre-processing step (Figure 3.6). Whereas Cyntenator required 6.25 hours to analyze the five genomes, MCScan and i-ADHoRe were considerably faster, analyzing the dataset in 19 and 14 minutes respectively. When i-ADHoRe was run with eight cores, the runtime was reduced to only 3 minutes.

In a second experiment, the maximum number of genomes that could be analyzed was determined by processing data sets of gradually increased size (Figure

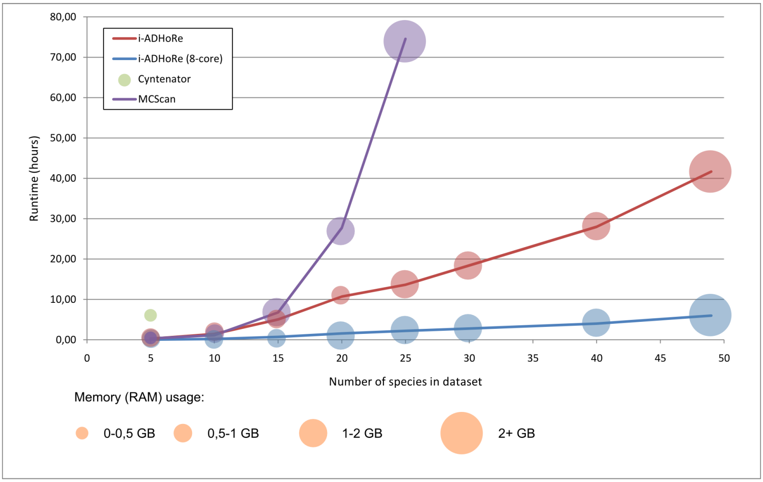


**Figure 3.6:** Runtime and memory usage comparison of Cyntenator, MCScan and i-ADHoRe (this study). Each tool was run on subsets of the Ensembl data set each including a different number of species.

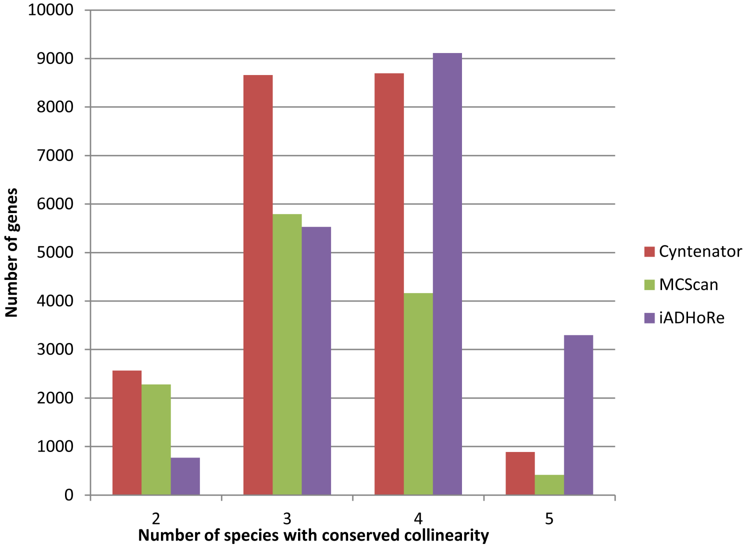


**Figure 3.7:** Number of genes found in regions conserved in *n* species, as detected with Cyntenator, MCScan and i-ADHoRe. For a fair comparison with Cyntenator, only the counts conserved in all five species were considered.

3.6). Only i-ADHoRe succeeded in analyzing the complete Ensembl data set covering 49 species (832 666 genes). Although Cyntenator could analyze up to 17 high-coverage genomes[23], the detection approach based on the strict usage of a guidance species tree posed a problem for data sets that include genomes sequenced at low coverage. As a result, inclusion of low-coverage or fractionated genomes into a large data set quickly eroded the amount of collinearity found, abruptly terminating the algorithm and leading to missing data when 10 or more genomes were included in the benchmark data set. For MCScan, the largest possible data set that could be analyzed in less than 48 hours included 20 species (Figure 3.6); although within 168 hours also 30 species could be covered, this duration however is impractically long for the efficient processing of extremely large data sets. In contrast, i-ADHoRe finished the full Ensembl data set covering 49 genomes within 42 hours using a single CPU core. This runtime could be reduced to less than six hours using the eight cores (88% efficiency). Finally, when using eight such machines (i.e. 64 cores in total) that are connected through a fast communication network (Infiniband), the runtime decreased to 40 minutes (50% efficiency; Figure 3.4). An additional advantage of i-ADHoRe is that gene families rather than individual homologous gene pairs can be used to construct the GHM, whereas in both Cyntenator and MCScan, per query gene, a limit of five homologous genes in each other species (based on BLAST hits) is suggested. Furthermore, the usage of gene families is a more memory-efficient alternative than storage of all homologous gene pairs covering multiple genomes. Although for small data sets, i-ADHoRe utilizes more memory than MCScan and Cyntentor, the required memory scales linearly with the total number of genes and remains below that of MCScan once the data sets include 20 or more genomes (Figure 3.6).

## 3.4 Conclusion

We show that the novel version of i-ADHoRe represents a major improvement over the current state-of-the-art algorithms and can be successfully applied to some of the largest data sets currently available.

As new sequencing initiatives such as the 1000 human genome project[162], the 1001 Arabidopsis genomes[125] and the 10k vertebrate genomes[163] will continue to generate many more genome sequences, the improved scalability of i-ADHoRe is imperative to keep runtimes acceptable. The support for parallel computing platforms ensures that i-ADHoRe 3.0 will efficiently detect genomic homology and will be instrumental to unveil genome evolution in the different kingdoms of life.

## 3.5 Availability

The i-ADHoRe 3.0 software package is free for academic use. Source code, documentation and example data sets are provided in the package.
Download from: http://bioinformatics.psb.ugent.be/software

## 3.6 Funding

This work was supported by Ghent University [Multidisciplinary Research Partnership "Bioinformatics: from nucleotides to networks"]; the Interuniversity Attraction Poles Programme [IUAP P6/25], initiated by the Belgian State, Science Policy Office (BioMaGNet); S.P. is indebted to the Agency for Innovation by Science and Technology in Flanders for a predoctoral fellowship; and K.V. is a Postdoctoral Fellow of the Research Foundation-Flanders.

## 3.7 Acknowledgments

The authors wish to thank Martine De Cock for help preparing the manuscript. The computational resources (Stevin Supercomputer Infrastructure) and services

used in this work were provided by Ghent University.

## 3.8 Author Contribution

Being first author I played the lead role in designing and implementing benchmark experiments and comparisons. Support for MPI and integration of the novel alignment algorithm was done by Jan Fostier, while the synteny mode was implemented by Dieter De Witte. I wrote the manuscript, though considerable contributions were made by Jan Fostier and Klaas Vandepoele.

*"It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts."*

Sherlock Holmes

# 4

# A Greedy, Graph-Based Algorithm for the Alignment of Multiple Homologous Gene Lists

# Abstract

**Motivation:** Many comparative genomics studies rely on the correct identification of homologous genomic regions using accurate alignment tools. In such case, the alphabet of the input sequences consists of complete genes, rather than nucleotides or amino acids. As optimal multiple sequence alignment is computationally impractical, a progressive alignment strategy is often employed. However, such an approach is susceptible to the propagation of alignment errors in early pairwise alignment steps, especially when dealing with strongly diverged genomic regions. In this paper, we present a novel accurate and efficient greedy, graph-based algorithm for the alignment of multiple homologous genomic segments, represented as ordered gene lists.

**Results:** Based on provable properties of the graph structure, several heuristics are developed to resolve local alignment conflicts that occur due to gene duplication and/or rearrangement events on the different genomic segments. The performance of the algorithm is assessed by comparing the alignment results of homologous genomic segments in *Arabidopsis thaliana* to those obtained by using both a progressive alignment method and an earlier graph-based implementation. Especially for datasets that contain strongly diverged segments, the proposed method achieves a substantially higher alignment accuracy, and proves to be sufficiently fast for large datasets including a few dozens of eukaryotic genomes.

**Availability:** http://bioinformatics.psb.ugent.be/software. The algorithm is implemented as a part of the i-ADHoRe 3.0 package.

**Contact:** yves.vandepeer@psb.vib-ugent.be

This chapter is based on Fostier et al.[148]. Author contribution, see page 105.

# 4.1 Introduction

In the past decades, considerable effort has been devoted to the development of algorithms for the alignment of biological sequences at the nucleotide or amino acid level. Using dynamic programming techniques, optimal pairwise global[164] and local[165] alignments can be obtained in $O(l^2)$ time, where $l$ denotes the length of the sequences. A straightforward extension of these algorithms to $N > 2$ sequences results in a computational complexity of $O(l^N)$, which renders the handling of sequences of realistic length impractical. Therefore, most Multiple Sequence Alignment (MSA) tools are based on progressive alignment, in which $N$ sequences are aligned through $N - 1$ applications of a pairwise alignment algorithm, usually guided by a tree which determines the order in which the sequences are combined. Many MSA tools that build on this principle have been implemented such as the well-known programs CLUSTAL(W)[166,167], T-COFFEE[168], MUSCLE[3] and MAFFT[169]. Almost without exception, MSA tools target the alignment of amino acid or nucleotide sequences.

In this paper, we focus on the alignment of multiple, mutually homologous (i.e. derived from a common ancestor) genomic segments, represented as *gene lists*. This means that the alphabet of the input sequences consists of individual genes, rather than nucleotides or amino acids. Similarly to MSA at the nucleotide or amino acid level, the goal is to align homologous genes, i.e. place genes that belong to the same gene family in the same column. The homology relationships between the individual genes have been established in a pre-processing step using sequence similarity searches and protein clustering[18]. Whereas ancestral gene order reconstruction[170] starts from homologous genomic segments to infer ancestral genome states and quantify genome dynamics, the objective of our graph-based approach is to create accurate *alignments* of homologous segments, in order to facilitate the detection of additional homologous genomic segments.

The multiple sequence alignment of gene lists differs significantly from the alignment of sequences at the nucleotide or amino acid level. First, the size of the alphabet of different nucleotides (four) or amino acids (twenty) is much smaller than the typical number of different gene families that occur in the genome of an organism. This means that a certain gene only has a very limited number of homologous genes in other gene lists. Second, through evolution, nucleotide and amino acid sequences mainly undergo character substitutions, whereas chromosomes mainly undergo gene loss/insertion, inversion and other types of rearrangements (e.g. reciprocal translocation). These two major differences allow for the development of a graph-based alignment approach, which will be demonstrated to have a higher accuracy than a progressive approach, in terms of the number of correctly aligned homologous genes.

We propose an algorithm similar to the so-called *segment-based* alignment approach that is used in e.g. DIALIGN[171]. The first step in DIALIGN consists of the identification of corresponding gap-free local alignments or '*fragments*' between pairs of sequences. The alignment of some of these fragments can prohibit the alignment of others. Finding the largest (weighted) subset of fragments that can be incorporated into a multiple alignment is a difficult task, often referred to as the *consistency* problem[172]. In the context of the gene list alignment problem, the 'fragments' correspond to homologous genes. The consistency problem then is to find the maximal number of homologous genes that can be included in a multiple alignment. Optimal solution methods to this consistency problem exist[173], but are NP-hard and therefore in general computationally impractical. Here, fast heuristic methods are developed to remove inconsistent or *conflicting* homology relationships between genes, from a graph-theoretic perspective. Similar ideas have been developed by Pitschi et al.[174].

The proposed alignment algorithm is part of the i-ADHoRe (iterative Automatic Detection of Homologous Regions) software[4,155], a map-based method to detect homologous genomic segments within or between the genomes of related organisms. Rather than identifying primary sequence similarity, map-based methods look for statistically significant conservation of gene content and gene order (collinearity). One of the key features of i-ADHoRe is the capability to uncover segmental homology, even between highly diverged segments. When two homologous segments have been identified, a so-called profile is constructed by aligning both segments, hence combining the gene order and content of both homologous segments. This genomic profile can then be used by i-ADHoRe as a more sensitive probe to scan the genome, in order to identify additional homologous segments[155]. This iterative process of alignment and detection continues, until no additional statistically significant genomic segments can be found.

It is clear that a high-quality alignment of the homologous gene lists within a profile is imperative for a sensitive detection of additional homologous genomic segments within the i-ADHoRe software. The original i-ADHoRe[155] implementation relied on profile construction using a progressive application of the Needleman-Wunsch (pNW) aligner. Especially when dealing with strongly diverged segments, one of the biggest problems with the pNW method is that erroneous alignment decisions in early pairwise steps propagate to the final alignment, causing the alignment quality to degrade significantly when more segments are added. This problem was already partially addressed in i-ADHoRe 2.0, through the introduction of a greedy, graph-based (GG) aligner[4]. Rather than relying on a progressive adding of segments, the GG-aligner considers the $N$ segments 'simultaneously'. Although this GG-aligner has its merits compared to the pNW-aligner (e.g. it avoids the 'once a gap, always a gap' problem), it was unable to outperform the latter in terms of the number of correctly aligned genes. This paper introduces a new greedy, graph-based algorithm (called GG2), that builds on the original GG-aligner. First, the basic graph-based alignment algorithm will be explained, followed by the development of a heuristic to resolve consistency problems in this graph, so-called conflicts. In later sections, we demonstrate that the new GG2-aligner outperforms both the pNW method and the original GG-aligner in terms of alignment accuracy. The new GG2-aligner has been implemented in the latest 3.0 version of i-ADHoRe and its C++ source code can be downloaded for academic purposes (http://bioinformatics.psb.ugent.be/software).

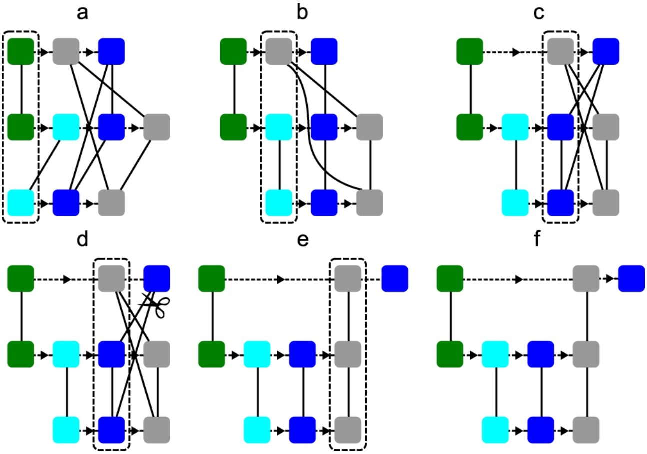


**Figure 4.1:** Example of the graph-based aligner for three simple gene lists. (a)–(e) Basic alignment algorithm. The active nodes are contained in the dashed rectangle. Note that the basic alignment procedure is 'stalled' in (c) and that two conflicting links have to be removed first (d), to allow the aligner to continue. (f) Resulting
alignment.

## 4.2 Algorithm

### 4.2.1 Graph Structure

Consider a set of $N$ unaligned genomic segments that are known to be mutually homologous. Each of the segments is represented by an ordered list that contains the genes in the same order as they appear on the corresponding segment. The number of genes in the $i^{\text{th}}$ list is denoted by $l_i$. Every gene in a list is homologous to zero or more genes in other lists. Although tandem duplicated genes on a genomic segment are largely filtered from the input by i-ADHoRe (see Section 3.1), their presence within the unaligned segments does not interfere with the alignment procedure. The gene lists can be represented together as a single graph $G(V, E, w)$ as follows. First, the genes are represented by vertices (or nodes) $V$. The $j^{\text{th}}$ node ($j = 1 \dots l_i$) on the $i^{\text{th}}$ gene list ($i = 1 \dots N$) is referred to by $n_{i,j}$. The functions seg(.) and ind(.) return the gene list and the position index of a node respectively, i.e. seg$(n_{i,j}) = i$ and ind$(n_{i,j}) = j$. Second, consecutive genes on a segment (i.e. $n_{i,j}$ and $n_{i,j+1}$) are connected through a directed arc or so-called *edge* pointed towards the gene with the highest index (the right-most gene). These directed edges simply connect the genes on a segment in a linear fashion. Finally, homologous genes located on different segments are connected through an undirected arc or so-called *link*. No links are created between homologous genes on the same segment (tandem duplicates). A weight $w$ can be attributed to each link. The higher this weight is taken, the more likely it is that the two nodes connected by this link, will show up in the same column in the final alignment. This will be explained in later sections. The graph corresponds to the 'extended alignment graph' as introduced by Lenhof et al.[173], although a slightly different terminology has been adopted here.

### 4.2.2 Basic Alignment Procedure

The basic workflow of the alignment algorithm is illustrated in Figure 4.1. Figure 4.1a depicts three simple unaligned gene lists. The undirected links are represented by a solid line, the directed edges by a dashed line.

At any time, the basic alignment algorithm considers a set of $N$ nodes, one node from each segment. These nodes are referred to as *active nodes*. For each segment $i$, the index $a_i$ refers to the active node $n_{i,a_i}$. At any time, all nodes on segment $i$, located to the left of the active node $n_{i,a_i}$ have already been aligned, the nodes $n_{i,j}$ with an index $j \geq a_i$ still have to be processed. Links that are incident to active nodes are called *active links*. The algorithm starts by considering the leftmost node from each segment, i.e. nodes $\{n_{1,1}, \ldots, n_{N,1}\}$.

If, among the $N$ active nodes, a minimal set of nodes $S = \{n_{k,a_k}\}$ can be found, for which each node in $S$ is linked *only* to other nodes within $S$, this set can be aligned. We say that $S$ is *alignable*. Note that $S$ can be a singleton, and that more than one (disjoint) set can be found at a given time. The term *minimal* therefore refers to the fact that $S$ should not be the union of two other alignable sets. Hence, all nodes within a minimal, alignable set $S$, correspond to genes that are homologous to each other.

The next set of active nodes is obtained by incrementing the index $a_i$ for each segment $i$ that has a node contained within one of the detected alignable sets. In other words, on those segments, the subsequent node is considered. At the corresponding position of all *other* segments, a gap is introduced. This is illustrated in Figure 4.1a and 4.1b. This process continues until either the end of all segments is reached, or a so-called *conflict* is encountered. A conflict is immediately detected when no alignable set $S$ can be found among the active nodes, as illustrated in Figure 4.1c. Conflicts can only be resolved by removing one or more links (see Figure 4.1d). This procedure will be explained in sections 4.2.3 – 4.2.6. Once a conflict has been resolved, the basic alignment procedure can be resumed (Figure 4.1e). Note that aligning all segments 'simultaneously' differs conceptually from progressive alignment, where first two complete segments are aligned before considering a third one, and so on. Finally, the resulting alignment is obtained as shown in Figure 4.1f.

### 4.2.3 Conflicts and Cycles in the Graph

The basic alignment procedure described above is straightforward, as long as no alignment conflicts are encountered. We define a conflict as a set of links that cannot be aligned, i.e. the alignment of some links in the set prohibits the alignment of other links. By the expression '*alignment of a link*', we mean the alignment of the two nodes connected by the link. Sources for alignment conflicts are gene duplications, local inversions, translocations and false positive homology assignment between genes.

In Section 4.2.5, we will be prove that if no alignable set $S$ can be found among the active nodes, such a conflict is always present. Conflicts can only be resolved by removing one or more links that contribute to the conflict. This means that certain homologous genes will not be placed in the same column in the final alignment. Because the goal of the algorithm is to minimize this number of misaligned (taking the weight $w$ of the links into account), it is imperative to carefully select which links are removed.

The presence of links and edges induces an ordering of the nodes in the graph $G$. Consider two nodes $u$ and $v$, for which a path $P$ in $G$ exists from $u$ to $v$. In general, such a path consists of both links and edges. The latter can only be traversed in the sense indicated by their arrow, i.e. from left to right. If a path from vertex $u$ to vertex $v$ contains at least one edge, then the order relationship $u \prec v$ holds. This means that, if all links in $P$ were to be aligned (suppose that this is possible), node $u$ would necessarily end up in a column left to the column containing node $v$ in the final result. We call such a path a *blocking* path $P_B$ with respect to to the nodes $u$ and $v$, as opposed to a *direct* path $P_D$, that contains only links and hence implies that nodes $u$ and $v$ should be aligned. This is indicated by $u \sim v$. A path from node $u$ to node $v$ imposes a direction on the links that are part of that path. In this context, the functions tail(.) and head(.) return the initial and terminal vertex of a such a link, respectively. This directional property of links exists only in the context of the specified path. A path $P$ from node $u$ to node $v$ can unambiguously be described by only listing the links –and not the directed edges (if any)– in the order of appearance in the path, i.e. $P = \{L_i\}$ $(i = 1 \ldots p)$, where seg($u$) = seg(tail($L_1$)), ind($u$) $\leq$ ind(tail($L_1$)), seg(head($L_i$)) = seg(tail($L_{i+1}$)), ind(head($L_i$)) $\leq$ ind(tail($L_{i+1}$)), $\forall i = 1 \ldots p-1$, seg($v$) = seg(head($L_p$)) and ind(head($L_p$)) $\leq$ ind($v$).

Given a link $L_1$ between nodes $u$ and $v$, an alignment conflict occurs, when there is at least one blocking path $P_B = \{L_i\}$ $(i = 2 \ldots p)$ from $u$ to $v$. Indeed, the presence of $L_1$ implies that $u \sim v$, whereas the presence of $P_B$ implies that $u \prec v$, a contradiction. Clearly, it is impossible to align all links in the set $\{L_i\}$ $(i = 1 \ldots p)$, hence they generate a conflict. The union $C_C = \{L_1 \cup P_B\}$ is a so-called *conflicting* cycle in the graph $G$. We define a conflicting cycle as a closed path in $G$ that contains at least one (directed) edge. By this reasoning, one can immediately see that alignment conflicts correspond to conflicting cycles in $G$ and vice versa. We define the number of links $p$ in the cycle $C_C$ as the

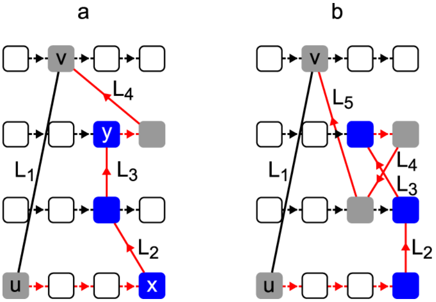


**Figure 4.2:** (a) The path $P_B = \{L_2, L_3, L_4\}$ defines an elementary blocking path from node $u$ to $v$. Similarly, $P_D = \{L_2, L_3\}$ is an elementary direct path between nodes $x$ and $y$. The cycle $C_C = \{L_1 \cup P_B\}$ is an elementary blocking cycle, corresponding to a minimal conflict of degree 4. Removing any link from $C_C$ will resolve the conflict. (b) The path $P_B = \{L_2, L_3, L_4, L_5\}$ is a blocking path from $u$ to $v$, however, the path is not elementary since both links $L_3$ and $L_5$ originate from nodes on the same segment. Indeed, even though $C_C = \{L_1 \cup P_B\}$ is a blocking cycle in $G$, the removal of e.g. $L_1$ does not resolve the conflict. The cycle $C'_C = \{L_3, L_4\}$ (hence $C'_C \subset C_C$) on the other hand is an elementary blocking cycle. Removing either one of the two links in $C'_C$ resolves the conflict.

*degree* of the conflict. Clearly, the degree is at least two. Also, note that the link $L_1$ does not play a special role in the conflict. Indeed, if we consider an arbitrary link $L_i$ $(i = 1 \dots p)$ in $C_C$, the links $\{L_{i+1}, \dots, L_p, L_1, \dots, L_{i-1}\}$ also define a blocking path from node head$(L_i)$ to node tail$(L_i)$.

As mentioned before, a conflict can only be resolved by removing one or more links that contribute to the conflict. If the removal of any link $L_i$ $(i = 1 \dots p)$ from its corresponding cycle $C_C$ resolves *all* conflicts between the remaining links (i.e. there are no conflicting cycles left in $C_C \setminus \{L_i\}$), we say that the conflict is *minimal*.

For any given cycle in the graph $G$, the number of links that terminate in nodes on a certain segment $s$ is equal to the number of links that originate from nodes on the same segment $s$. If at most one link in the cycle originates from each segment, we call it *elementary*. The maximum number of links in an elementary cycle is therefore given by $N$. Similarly, an *elementary path* is defined as a path where at most one link originates from each segment. The maximum number of links in such a path is $N - 1$.

**Proposition 1:** Minimal conflicts correspond to elementary conflicting cycles $C_C$ and vice versa.

Proof: see supplementary data accompanying Fostier et al.[148].

As an immediate consequence, the maximum degree of a minimal conflict is given by the number of segments $N$.

The importance of the concept of minimal conflicts stems from the fact that such conflicts can be resolved by removing any link involved in the conflict. This is not the case for conflicts associated with non-elementary blocking cycles (compare e.g. the examples in Figures 4.2a and 4.2b). Also, from the proof of Proposition 1, it follows that any non-elementary conflicting cycle $C_C$ corresponds to one or more *minimal* conflicts, either by removing superfluous links from $C_C$, or by decomposing $C_C$ into several elementary conflicting cycles. Therefore, in what follows, we only consider elementary paths, elementary cycles and minimal conflicts, without explicitly mentioning the terms *elementary* and *minimal*.

### 4.2.4 Conflict Detection and Resolution

If the basic alignment procedure is stalled because of conflicts (i.e. no alignable set can be found among the active nodes), we want to determine which links are involved in these conflicts and which links are to be removed from $G$. For now, we only consider the active links as candidates for removal. In the next section, we will prove that this approach is indeed a valid one.

Consider an active link between nodes $n_{i,a_i}$ and $n_{j,k}$, with $i \neq j$ and $k \geq a_j$ (with $a_j$ the index of the active node $n_{j,a_j}$). For the simplicity of notation, these nodes are referred to as $s$ and $t$ respectively, the active link is denoted by $L_{st}$. The link $L_{st}$ contributes to a conflict, if there is a blocking path $P_B$ between $s$ and $t$ or vice versa, between $t$ and $s$. Indeed, the alignment of the link $L_{st}$ (or

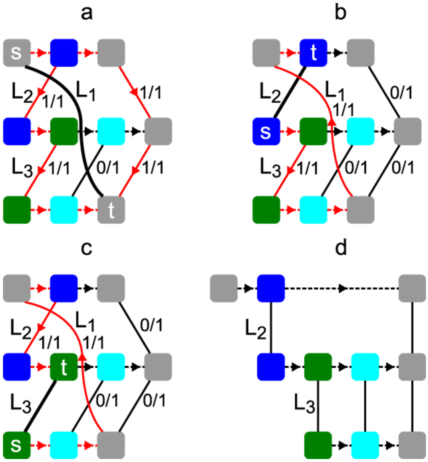


**Figure 4.3:** Example of a simple alignment conflict and its solution. In (a), (b) and (c), the link scores $S_L$ are calculated for the active links $L_1$, $L_2$ and $L_3$ (indicated by a bold line) respectively. All links have weight (and hence capacity) $w = 1$, the edges have unlimited capacity. The numbers near the links denote the flow/capacity of that link. The maximum flow from node $s$ to $t$ through elementary blocking paths (indicated in red) is (a) $f_{st}^C = 2$ for link $L_1$, (b) $f_{st}^C = 1$ for link $L_2$, (c) $f_{st}^C = 1$ for link $L_3$. In all three cases, no conflicting flows are possible from node $t$ to $s$, i.e. $f_{ts}^C = 0$. The maximum flow from node $s$ to $t$ through direct paths is $f_{st}^D = 1$, in all three cases. Therefore, $S_{L_1} = -1$, $S_{L_2} = 0$ and $S_{L_3} = 0$. (d) Resulting alignment after the link with the lowest score (i.e. $L_1$) is removed from $G$.

any other direct path $P_D$ between $s$ and $t$) implies the ordering $s \sim t$. A blocking path $P_B$ from $s$ to $t$ implies $s \prec t$, and similarly, a blocking path from $t$ and $s$ implies $s \succ t$. We refer to these conflicts as *st-conflicts* and *ts-conflicts* respectively.

To quantitatively investigate the number of conflicts that $L_{st}$ is involved in, we want to assess to which degree $s$ and $t$ are connected through blocking paths. In graph theory, such problems can be addressed by solving the well-known *maximum flow* problem. For a formal definition of the maximum flow problem, we refer to Ford and Fulkerson[175]. Intuitively, the maximum flow is the largest amount of 'flow' (e.g. fluid or current) that can be transported between two given nodes, called *source* and *sink* respectively. Let $f_{st}$ be the maximum flow from node $s$ to node $t$ acting as the source and sink respectively. As an extra restriction, it is imposed that a valid flow can only pass through *elementary* paths (either blocking or direct) from $s$ to $t$. The edges have unlimited flow capacity, however, the flow can only pass in the sense indicated by the direction of the edge (from left to right). The links have a capacity equal to their weight $w$, but impose no direction on the flow. There exist many polynomial algorithms for the solution of the maximum flow problem. This is more thoroughly discussed in the supplementary text included in Fostier et al.[148].

Similarly, let $f_{st}^D$ be the maximum flow from $s$ to $t$ through *direct* elementary paths. Note that this includes the flow through the link $L_{st}$. Then clearly, $f_{st}^C = f_{st} - f_{st}^D$ is the maximum flow from $s$ to $t$ through elementary *blocking* paths. As a consequence of the max-flow, min-cut theorem[176], $f_{st}^C$ is the minimum link capacity that has to be removed from $G$ to disconnect $s$ and $t$ through elementary blocking $st$-paths, i.e. to resolve all $st$-conflicts. Similarly, $f_{ts}^C$ can be calculated as the maximum flow through elementary *blocking* paths from $t$ to $s$. We then use the following score to evaluate link $L_{st}$:

$$S_{L_{st}} = f_{st}^D - |f_{st}^C - f_{ts}^C|$$

Since $st$-conflicts and $ts$-conflicts mutually conflict, at least $\min(f_{st}^C, f_{ts}^C)$ capacity will need to be removed from $G$, regardless whether or not $s$ and $t$ are aligned. The term $|f^C_{st} - f^C_{ts}|$ therefore denotes the minimal, net capacity that will need to be removed from $G$ if $s$ and $t$ are aligned. Similarly, $f^D_{st}$ denotes the minimal capacity that will need to be removed from $G$ if $s$ and $t$ will *not* be aligned. Clearly, a positive score for $S_{L_{st}}$ indicates that it is probably best to align $s$ and $t$, whereas a negative score for $S_{L_{st}}$ indicates that it is probably best to remove the link $L_{st}$ from the graph.

Note that if there are multiple $st$-paths, that these paths might be in mutual conflict. This fact is not taken into account by the link score $S_{L_{st}}$. In other words, there is no guarantee that the capacity $|f^C_{st} - f^C_{ts}|$ will effectively be aligned, even if $s$ and $t$ are disconnected through direct paths.

The algorithm for conflict resolution can now be described as follows. If, during the basic alignment procedure, described in Section 4.2.2, no alignable set $S$ can be found among the active nodes, all active links are considered. For each of these links $L$, the score $S_L$ is calculated, and the link with the lowest score (i.e. the most problematic link), is removed from $G$. This process is repeated, until again, an alignable set $S$ can be found among the active nodes. Figure 4.3 presents a simple example.

### 4.2.5 Active Conflicts

In this section, a refinement to the conflict resolution algorithm is developed. Consider a conflict situation in the graph $G(V, E)$ (i.e. no alignable set $S$ can be found among the $N$ active nodes in $G$). Next, consider the subgraph $G'(V, E') \subset G(V, E)$ that only contains the active links (i.e. the links incident to the active nodes). We show that even in the reduced graph $G'(V, E')$, no alignable set can be found among the active nodes.

**Proposition 2:** If, during the basic alignment procedure, no alignable set can be found among the $N$ active nodes in the graph $G(V, E)$, at least one conflict is present among the active links. Furthermore, no alignable set can be found

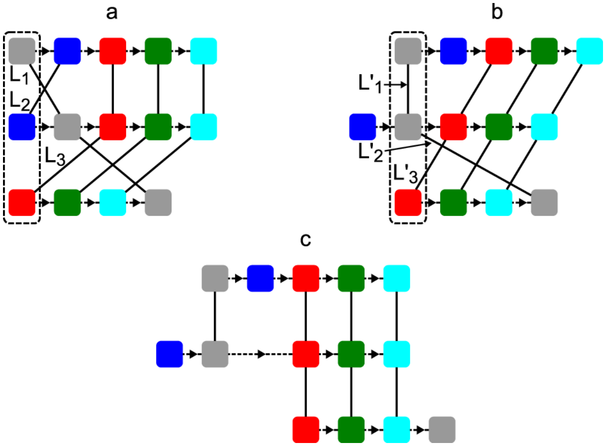


**Figure 4.4:** Example of active conflicts and the improved heuristic. All links have weight $w = 1$. (a) Although all active links have an equal score ($S_{L_1} = S_{L_2} = S_{L_3} = 0$), only links $L_1$ and $L_2$ are involved in an active conflict. The alignment of the upper two segments can not progress as long as this conflict exists. (b) Situation after $L_2$ has been removed. Now, links $L'_2$ and $L'_3$ are involved in an active conflict, with scores $S_{L'_2} = -2$ and $S_{L'_3} = 0$. (c) Final alignment after $L'_2$ has been removed.

among the same active nodes in the subgraph $G'(V, E')$.

Proof: see supplementary data accompanying Fostier et al.[148].

Proposition 2 provides a more fundamental understanding of alignment conflicts. First, it shows that active links are indeed good candidates for removal. Indeed, even the removal of *all* non-active links would still not allow for the alignment of *any* of the active nodes.

Second, it shows that if the basic alignment procedure is stalled, at least one conflict, consisting of active links, is present. Such a conflict is called an *active* conflict. None of the active nodes that are incident to a link in an active conflict can be aligned, as long as this conflict exists. Active conflicts are therefore high-priority conflicts that need to be resolved instantaneously. In the case of conflicts, the active links can therefore be grouped into three categories: active links involved in at least one active conflict, active links involved in non-active conflicts and active links that are not involved in a conflict. The conflict resolution algorithm is therefore modified as follows. For each of the links $L$ involved in at least one active conflict, calculate the score $S_L$. The link with the lowest score is removed from $G$. It is easily determined whether or not an active link is involved in an active conflict, by calculating $f^C_{st}$ and $f^C_{ts}$ in the reduced graph $G'$. If any of the two flows is nonzero, an active conflict is present. A simple example of the improved heuristic is illustrated in Figure 4.4.

In the results section, it will be demonstrated that this approach indeed improves the alignment quality.

### 4.2.6 Faster Heuristics for Conflict Resolution

The calculation of the maximum flow between two nodes in the graph is computationally expensive. However, upper bounds to the maximum flow can easily be derived. Given an active link $L_{st}$ between source node $s = n_{i,a_i}$ and sink node $t = n_{j,k}$, one can immediately notice that the final link in a blocking $st$-path necessarily ends in a node $n_{j,k'}$ with $k' \leq k$. Therefore, an upper bound to $f^C_{st,\mathrm{UB}}$ can be found by summing the weights $w(L)$ of all links incident to these nodes:

$$f^C_{st,\mathrm{UB}} = \sum_{k'=a_j}^{k} \sum_{L \in n_{j,k'}} w(L),$$

with $j$ =seg$(t)$ and $k$ =ind$(t)$. Similarly, blocking $ts$-paths necessarily end in the source node $s$ and an upper bound $f^C_{ts,\mathrm{UB}}$ can therefore be established by summing the weights of the links $L \neq L_{st}$ incident to $s$.

$$f^C_{ts,\mathrm{UB}} = \sum_{L \in n_{i,a_i}, L \neq L_{st}} w(L)$$

Finally, a *lower* bound to the direct flow $f^D_{st}$ is simply given by $f^D_{st,\mathrm{LB}} = w(L_{st})$. Therefore, a lower bound to the link score is given by

$$S_{L_{st},\mathrm{LB}} = f^{D}_{st,\mathrm{LB}} - \max(f^{C}_{st,\mathrm{UB}}, f^{C}_{ts,\mathrm{UB}}).$$

Selecting the active link $L$ with the lowest lower bound score $S_{L,\mathrm{LB}}$ yields a much faster heuristic. Indeed, the calculation of $S_{L,\mathrm{LB}}$ requires no maximum flow problems to be solved. Even though this lower bound estimation may be significantly underestimating the actual score $S_{L_{st}}$, it still provides a powerful method to select a link for removal, if we assume that the link with the lowest lower bound score is also the link with the lowest score.

Taking this reasoning even a step further, the heuristic can even be further simplified, if one assumes that the sum of the weights of the links, incident to a node, is constant for all nodes, i.e. that the links are evenly distributed among the nodes. Given a link $L_{st}$ between source node $s = n_{i,a_i}$ and sink node $t = n_{j,k}$, this means that $f^{C}_{st,\mathrm{UB}}$ is proportional to $(k - a_j)$, while $f^{C}_{ts,\mathrm{UB}}$ and $f^{C}_{st,\mathrm{LB}}$ are constant. This is clearly a rather rough approximation, however, it leads to the very simple and fast heuristic: select the active link incident to node $n_{j,k}$ for which the '*length*' of the link $(k - a_j)$ is maximal (the 'longest' link). Such a link has the most possibilities for conflicting $st$-paths, and is hence a good candidate for removal.

We now summarize the heuristics for conflict resolution and introduce three random methods. These random methods are not of any particular interest, but it is always interesting to compare the more mathematically supported methods to random methods.

Select, in the case of a conflict, among the *active* links, the following link for removal:

- RA (RAndom): a random link.
- RC (Random Conflict): a random link that is involved in at least one (active or non-active) conflict.

**Table 4.1:** Comparison of the number of correctly aligned homologous genes in dataset I. The scores of the random methods are averaged over 1000 runs.

| | | $\sum$ # correctly aligned homologous genes | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | # input | | | GG2: the proposed greedy, graph-based aligner | | | | | |
| N | sets | pNW | GG | RA | RC | RAC | LL | LLBS | LS |
| 2 | 447 | 4877 | 4719 | 4108 | 4109 | 4109 | 4843 | 4871 | 4874 |
| 3 | 169 | 2823 | 2704 | 2477 | 2544 | 2574 | 2810 | 2817 | 2852 |
| 4 | 119 | 2924 | 2769 | 2502 | 2611 | 2684 | 2921 | 2971 | 3008 |
| 5 | 49 | 1634 | 1533 | 1375 | 1454 | 1514 | 1627 | 1655 | 1697 |
| 6 | 41 | 1715 | 1516 | 1375 | 1485 | 1577 | 1773 | 1747 | 1814 |
| 7 | 39 | 2114 | 1803 | 1572 | 1732 | 1884 | 2149 | 2152 | 2275 |
| 8 | 24 | 1229 | 1062 | 882 | 995 | 1094 | 1278 | 1263 | 1358 |
| 9 | 16 | 807 | 713 | 623 | 725 | 773 | 880 | 885 | 921 |
| 10 | 13 | 703 | 670 | 602 | 704 | 741 | 810 | 822 | 825 |
| 11 | 4 | 211 | 228 | 203 | 239 | 246 | 263 | 259 | 259 |
| $\sum$ | 921 | 19 037 | 17 717 | 15 719 | 16 596 | 17 196 | 19 354 | 19 442 | 19 883 |

- RAC (Random Active Conflict): a random link that is involved in at least one active conflict.
- LL (Longest Link): the link $L$, involved in at least one active conflict, incident to node $n_{j,k}$ for which $(k - a_j)$ is maximum.
- LLBS (Lowest Lower Bound Score): the link $L$, involved in at least one active conflict, with the lowest lower bound score $S_{L,\mathrm{LB}}$.
- LS (Lowest Score): the link $L$, involved in at least one active conflict, with the lowest score $S_L$.

## 4.3 Results and Discussion

### 4.3.1 Datasets

To test the performance of multiple sequence alignment tools, a number of benchmarks have been introduced for both protein sequences (such as BALiBASE[177], OXBench[178], PREFAB[169] and SMART[179]) and DNA sequences[180]. Because no similar benchmark exists to test the performance of gene list alignment tools, two ad hoc input datasets were generated by running the i-ADHoRe tool on the *Arabidopsis thaliana*[12] genome separately (dataset I) and on the *Arabidopsis thaliana*, poplar(*Populus trichocarpa*)[34] and grapevine(*Vitis vinifera*)[46] genomes (dataset II). *Arabidopsis thaliana* is a good candidate to validate the aligners and heuristics, since its genome contains both strongly diverged and more closely related homologous chromosomal regions[28,79]. Using the profile searches (see Simillion et al.[4]), the i-ADHoRe algorithm produces 921 and 7 821 sets of homologous genomic segments for datasets I and II respectively. The number of genomic segments $N$ in these sets varies from 2 to 11 (dataset I) and from 2 to 15 (dataset II). For both datasets, the i-ADHoRe settings were *gap size* = 30, *cluster gap* = 35, *q* = *0.75* and *p* = *0.01*. Tandem duplicates within a distance of *gap size / 2* were remapped onto the representative gene with the lowest index.

### 4.3.2 Alignment accuracy

To detect homologous segments, i-ADHoRe looks for statistically significant conservation of gene content and order. When two homologous segments are visualized in a dot-plot, their collinearity shows up as a 'diagonal'. The homologous gene pairs between the two segments that are used by i-ADHoRe to detect these 'diagonals', are called *anchors*. These anchors are a subset of all homologous gene pairs between the two segments. By giving a higher weight to the links associated with anchors, they can effectively be used as an *alignment guide* to improve the overall alignment quality. In all simulations, the weight $w$ of the links corresponding to anchors was set to 1, whereas the weight of the other links was set to 0.1. These other links correspond to homologous genes that are further off-diagonal, and therefore less likely to be aligned in the final result. Note that more complex weight schemes could be incorporated, possibly improving alignment results. For example, the link weights could represent the probability that two genes are truly homologous. In this work, such schemes were not investigated.

The proposed greedy, graph-based aligner (GG2) is compared to both a progressive application of the Needleman-Wunsch method (pNW) and the original greedy, graph-based aligner (GG). The pNW-aligner first performs a pairwise alignment of the two genomic segments that share the most anchor points, i.e. the two most closely related segments. Subsequently, a third segment is added to this intermediate result and so on. It should be noted that more advanced improvements to this basic progressive approach have been implemented, e.g. by using a guide tree to determine the order in which the segments are added[167], or by incorporating consistency-objective functions[181].

The original graph-based GG-aligner relies on the same 'basic alignment procedure' as the GG2-aligner, however, conflicts are handled in a more primitive fashion. In short, based on the number of links and their *lengths* (cfr. Section 4.2.6), the GG-aligner calculates a score for each active node (as opposed to for each active link in the GG2-aligner). Instead of removing a single link, the GG-aligners removes *all* links incident to the active node with the lowest score. In the GG-aligner's heuristic, no thorough analysis of conflicting paths or links is conducted.

In Table 4.1, the number of correctly aligned homologous genes for the profiles generated by the different aligners are compared for dataset I. We consider two homologous genes to be correctly aligned if they are placed in the same column in the final result. This omits the need for a reference alignment. The numbers in Table 4.1 therefore correspond to the sum-of-pairs metric. Each row shows the accumulated sum-of-pairs scores for all input sets with a specified number of segments $N$ ($N = 2 \ldots 11$). The final row represents the sum-of-pairs over all profiles, and can therefore be seen as a quality benchmark for the complete dataset.

First, it is immediately clear that all random methods perform significantly worse than the more mathematically supported heuristics. The score of the RA-aligner is an indication of the number of homologous genes that can be aligned 'for free' by the basic alignment procedure. The fact that this score is rather high means that a fairly large number of links is not involved in any conflict. Indeed, if for example all input segments were identical (perfect collinearity and hence no conflicts), all methods would produce identical (and optimal) results. When comparing the numbers of the other aligners, it is important to keep this consideration in mind. The RC-aligner improves the RA score, by making sure that no active links are removed that do not contribute to any conflict. Interestingly, the alignment score is again significantly improved by using the RAC-aligner, which selects a random active link, involved in at least one *active* conflict. This provides experimental evidence for the observations made in Section 4.2.5, i.e. that active conflicts are high-priority conflicts. Note that in the case of a conflict for $N = 2$, all active links are necessarily involved in an active conflict. Therefore, the RA, RC and RAC heuristics perform equally well for $N = 2$.

The LL, LLBS and LS heuristics strongly outperform both the random methods and the original GG-aligner, and, albeit to a somewhat lesser extent, also the pNW-method. Unsurprisingly, the pNW-aligner is best for $N = 2$, since it produces optimal results. The LL, LLBS and LS methods however, also obtain close to optimal results. For larger $N$, the relative difference in score between pNW on the one hand and LL, LLBS and LS one the other hand increases. This is to be expected: erroneous alignment decisions in early pairwise steps of the pNW-aligner propagate when more segments are added. The graph-based methods are more robust in the sense that they take the links on all segments into consideration. For higher $N$, the difference in score between LS and pNW is larger than 10%. In total, the LS-method is able to align 846 (4.4%) more homologous genes than the pNW methods. An alignment example of six homologous genomic segments in the *A. thaliana* genome as produced by the pNW and the LS heuristic is given in Figure 4.5.

The difference in alignment quality between the LL, LLBS and LS heuristics is rather modest, however, it can be observed that LL $<$ LLBS $<$ LS, for nearly all $N$. Despite the simplicity of the LL heuristic, this method still performs remarkably well, and even outperforms the pNW method on this dataset. The main difference between these methods lies in the alignment speed. This will be discussed in more detail in the next section.

It is important to mention that the relative difference in alignment quality between the pNW on the one hand and the LL, LLBS and LS heuristics associated with the GG2-aligner on the other hand, decreases for datasets that consist of ge-

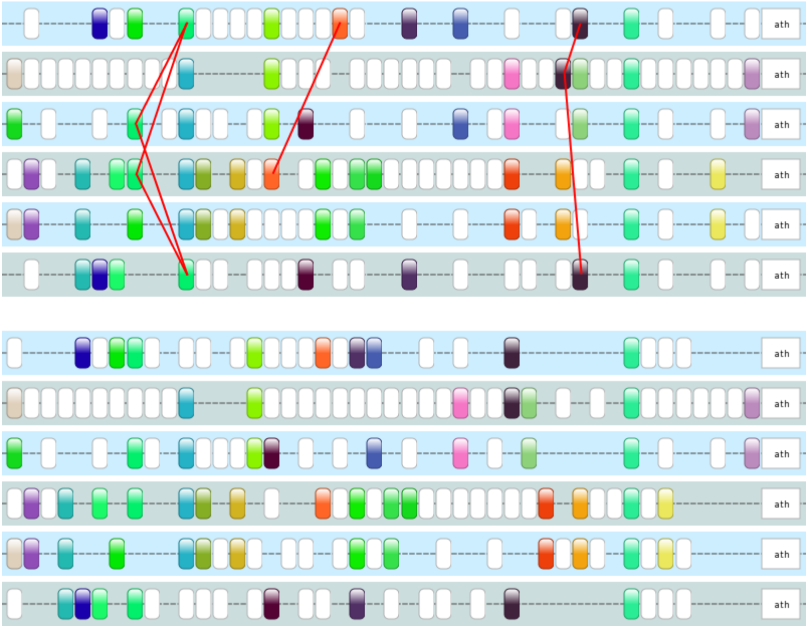


**Figure 4.5:** The alignment of six homologous genomic segments in the *A. thaliana* genome as produced by the progressive Needleman-Wunsch (top) and the proposed greedy, graph-based aligner (bottom). The genes that are misaligned by the progressive Needleman-Wunsch approach are indicated by a red line.

nomic segments that are less diverged. For instance, Table 4.2 lists the alignment scores for dataset II. Even though the ranking of the different alignment methods remains the same, the relative difference in alignment score is smaller. This is due to the fact that relatively fewer alignment conflicts exists in this dataset, which can be seen from the high score of the random aligners.

### 4.3.3 Program runtime

A comparison for the alignment times of the different heuristic methods can be found in Table 4.2 for the larger dataset II. Except for the LS heuristic, the total alignment times are very low. The unfavorable time complexity of the LS heuristic prohibits the alignment of sequences with larger $N$. In practice, when $N > 10$, the CPU time for the LS method rapidly increases. In i-ADHoRe, we therefore offer the LLBS heuristic by default. Experiments on an extremely large dataset consisting of dozens of eukaryotic species[149] have shown that this method can easily handle N=50, enough for most practical problems.

A detailed analysis of the computational complexity of the algorithm is given in supplementary data of Fostier et al.[148].

### 4.3.4 Comparison of i-ADHoRe to related tools

The GG2-aligner is an important component of i-ADHoRe, which detects evolutionary related genomic regions within or between related species through sensitive iterative profile searches. In contrast to this approach, CYNTENATOR[23] and the method described by[182] compute multiple gene order alignments progressively using initial pairwise alignments and a guide tree. DRIMM-Synteny[183] detects non-overlapping synteny blocks to perform rearrangement analysis in duplicated genomes and reconstruct ancestral genomes. Although our method does not infer likely evolutionary paths of genome evolution events, the application of the profile search on the *Arabidopsis thaliana* genome (dataset I) identifies a much larger

| alignment method | alignment score | alignment time (s) |
|---|---|---|
| pNW | 518 247 | 1.7 |
| GG | 497 771 | 5.9 |
| RA | 489 020 | 6.2 |
| RC | 501 241 | 5.2 |
| RAC | 505 447 | 3.3 |
| LL | 525 665 | 2.3 |
| LLBS | 525 652 | 2.2 |
| LS | 529 633 | 6 742 |

**Table 4.2:** Comparison of alignment scores and runtimes for dataset II.

fraction of the genome in duplicated blocks, compared to DRIMM. Fractions of 66% and 8% are reported in duplicated blocks with a multiplicity of at least two and at least four respectively[183], whereas i-ADHoRe detects 90.3% and 25.8% respectively. In agreement with the yeast results reported by DRIMM, including a more ancestral genome lacking a recent whole-genome duplication (e.g. grapevine in dataset II) serves as a reference backbone to identify and align highly diverged *Arabidopsis thaliana* genomic segments[53].

## 4.4 Conclusion

We have developed a greedy, graph-based algorithm for the alignment of multiple, homologous gene lists. Several properties of conflicts within the alignment graph have been derived and proved. Three heuristics for conflict resolution were developed on these theoretical grounds, and have been demonstrated to be able to outperform an older graph-based algorithm and a progressive approach in terms of alignment accuracy. As is often the case, a trade-off between computational requirements and alignment accuracy can be observed. The algorithm has been implemented in the latest version of i-ADHore 3.0.

## 4.5 Acknowledgments

This project is funded by the Research Foundation Flanders and the Belgian Federal Science Policy Office: IUAP P6/25 (BioMaGNet). We acknowledge the support of Ghent University (Multidisciplinary Research Partnership "Bioinformatics: from nucleotides to networks"). KV and YS are Postdoctoral Fellows of the Research Foundation-Flanders (FWO). SP thanks the Institute for the Promotion of Innovation by Science and Technology in Flanders for a predoctoral fellowship.

## 4.6 Author Contribution

The algorithm was designed and implemented by Jan Fostier, while I (as first author), constructed the input datasets and rigorously evaluated the output of different variants of the algorithm during development.

*"Although nature commences with reason and ends in experience it is necessary for us to do the opposite, that is to commence with experience and from this to proceed to investigate the reason."*

Leonardo Da Vinci

# 5

# Using PLAZA and i-ADHoRe to Dissect Eukaryotic Genomes

# Abstract

Any tool, no matter how well designed, is useless without any real-world applications. In previous chapters we have introduced PLAZA[25] and i-ADHoRe 3.0[126], here we will present various case-studies how the application of these tools can lead to novel insights in biological processes. Using these tools, as can be seen in this chapter, duplicated genes can be further classified into tandem or block duplicates. By combining duplication data with GO, the effects of small, local duplications can be compared with the impact of large-scale duplications. Such Whole Genome Duplications (WGD) are know to be a driving force for evolution and adaptation.

The second part of this chapter focusses on the application of PLAZA and i-ADHoRe on two newly sequenced genomes, namely that of apple[51] and *Medicago truncatula*[184]. In both cases the tools were shown to be of seminal importance in unveiling the evolutionary history key processes, such as fruit development in apple and nodulation in *Medicago truncatula*, along with the changes in genome structure.

In conclusion, here we demonstrate the potential of PLAZA and i-ADHoRe as a starting point for a wide variety of analyses, that lead to novel and interesting biological insights.

This chapter is redrafted from the Proost et al. [25],[126] along with sections from Velasco et al. [51] and Young et al. [184] resulting from analyses using PLAZA[25] and i-ADHoRe 3.0[126]. Author contributions can be found on page 136.

## 5.1 Introduction

While earlier chapters of this thesis focus on the development of tools and techniques to study various aspects of genome evolution. In this last research chapter convincing evidence is provided of how powerful such tools exactly are while dissecting Eukaryotic genomes. Both PLAZA[25] and i-ADHoRe 3.0[126] were published along with some case studies to show potential users example applications. In this chapter these case studies are presented along with the results obtained on two newly sequenced genomes, namely that of apple[51] and *Medicago truncatula*[184]. As a whole these case studies show that both PLAZA and i-ADHoRe can be used as a base for a wide range of different studies.

## 5.2 Analysis of Gene Duplicates Using PLAZA

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.[107]). 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.[25]). 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).

### 5.2.1 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[185] 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.

### 5.2.2 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

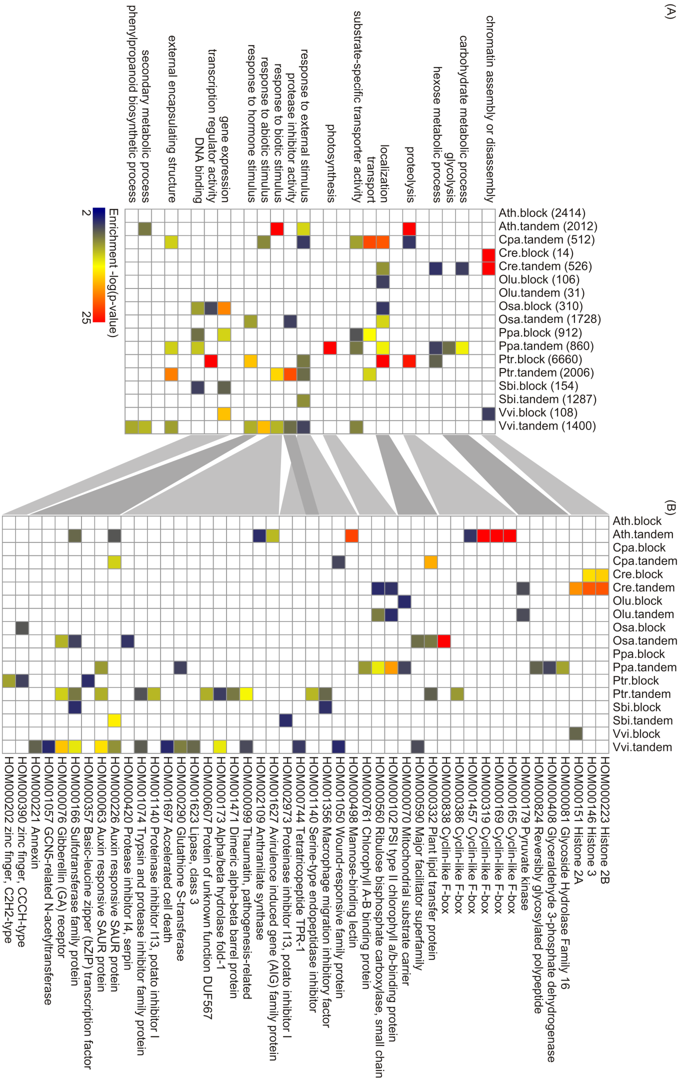


**Figure 5.1:** GO Enrichment Analysis of Species-Specific Gene Duplicates. (A) The GO enrichment for species-specific block and tandem duplicates in different species is visualized using heat maps. Colors indicate the significance of the functional enrichment, while nonenriched cells are left blank. 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.

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[186–188]. Apart from sufficient histone proteins in rapidly dividing cells, exact quantities also are required for correct nucleosome formation. The assembly of histones 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[189,190]. 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[191].

### 5.2.3 Studying the Effects of WGDs

As a second case study, we used PLAZA to study large-scale duplication events in different lineages. Counting all gene duplication events for the different organisms confirms the presence of one or more WGD in *Arabidopsis thaliana*, moss, and monocots[25]. Interestingly, when analyzing the inferred ages of the different duplication nodes using the reconciled phylogenetic trees, we observed that the number of duplication events in the ancestor of angiosperms is larger than those in the eudicot ancestor (1880 and 1146 duplication nodes, respectively). In addition, these ancestral angiosperm duplications cover a larger number of gene families compared with the eudicot duplications (1141 and 757 families, respectively). This pattern suggests that, apart from the ancient hexaploidy detectable in all sequenced eudicot plant genomes[24], older gene duplications have also significantly contributed to the expansion of the ancestral angiosperm proteome.

It is now generally accepted that, after the divergence of papaya and *Arabidopsis thaliana*, the latter species has undergone two rounds of WGD[28,46,53].

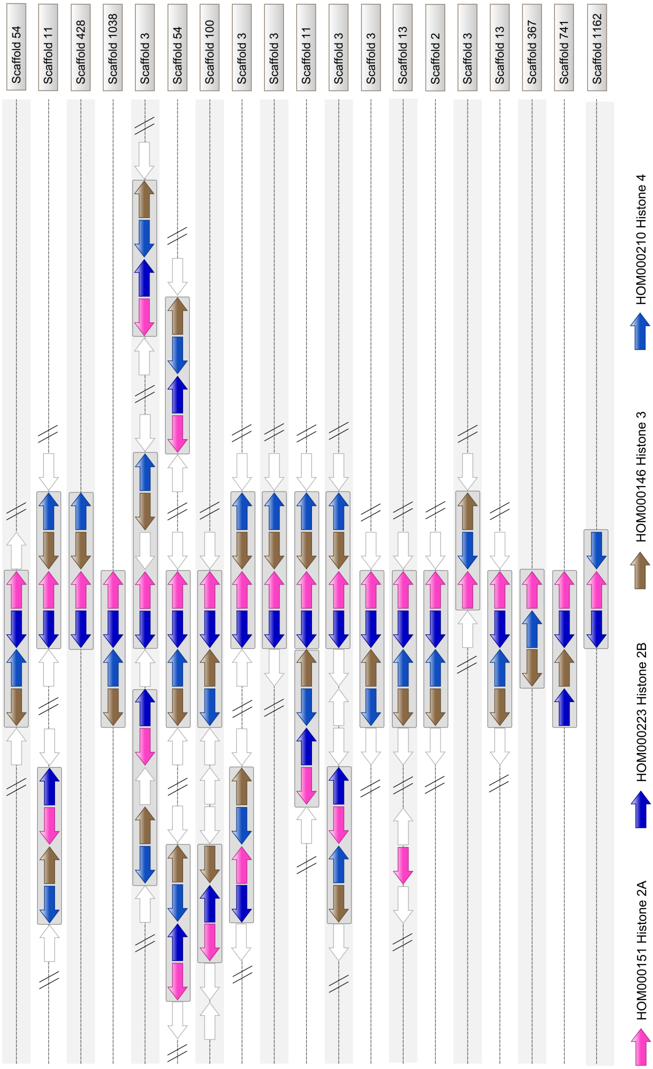


**Figure 5.2:** Duplicated histon genes in *Chlamydomonas reinhardtii*. The genomic organization of the core histone genes in *Chlamydomonas* reveals a pattern of dense clustering (indicated by gray boxes). Genes are shown as arrows; the direction indicates the transcriptional orientation and colors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

**Table 5.1:** Counting gene loss in *Arabidopsis thaliana* segment generated by the alpha and beta WGD

| Counts for the analyzed multiplicons | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | multiplicon ID | | | | | | | |
| | 15 | 69 | 83 | 185 | 227 | 542 | 915 | Total |
| No Genes Lost (a) | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 2 |
| One Gene Lost (b) | 2 | 1 | 1 | 1 | 1 | 0 | 0 | 6 |
| Two Genes Lost (beta) (c) | 5 | 7 | 4 | 4 | 7 | 5 | 5 | 37 |
| Two Genes Lost (alpha) (d) | 0 | 0 | 0 | 0 | 2 | 1 | 1 | 4 |
| Three Genes Lost (e) | 15 | 15 | 5 | 20 | 20 | 8 | 7 | 90 |
| Loci before beta | | | | | | | | 139 |
| Loci before alpha | | | | | | | | 151 |

**Table 5.2:** Summary counts of the gene loss in *Arabidopsis thaliana* since the alpha and beta WGD

| Summary counts for Table 5.1 | | | | | |
|---|---|---|---|---|---|
| | #genes lost | #genes retained | total | %lost | %retained |
| Beta WGD | 127 | 12 | 139 | 91.37% | 8.63% |
| Alpha WGD | 104 | 47 | 151 | 68.87% | 31.13% |

PLAZA colinearity data were used to determine if levels of gene loss were different after the first (oldest) and second (youngest) WGD (also referred to as beta and alpha, respectively). To this end, we selected multiplicons grouping four aligned *Arabidopsis thaliana* duplicated regions with an unduplicated outgroup region from either grapevine or papaya to count gene loss based on parsimony. Grapevine/papaya-*Arabidopsis thaliana* 1:4 alignments reveal that massive gene loss within *Arabidopsis thaliana* makes it very hard to link the homoeologous segments without aligning them to either grape or papaya (Figure 5.3)[53]. Manual inspection identified 26 reliable nonredundant multiplicons of which, in seven cases, the *Arabidopsis thaliana* segments could, based on K$_S$, unambiguously be grouped in two pairs that originated during the youngest duplication. Analyzing all different patterns of gene loss using 139 ancestral loci (Table 5.1 and 5.2) revealed that 3.6 times more genes have been retained after the youngest a than after the oldest beta Arabidopsis-specific WGD (31.13 and 8.63% retention, respectively). Consequently, this massive amount of gene loss masks most traces of the oldest WGD and explains why, with only the *Arabidopsis thaliana* genome available, the existence and timing of an older beta duplication was debated[42–44].

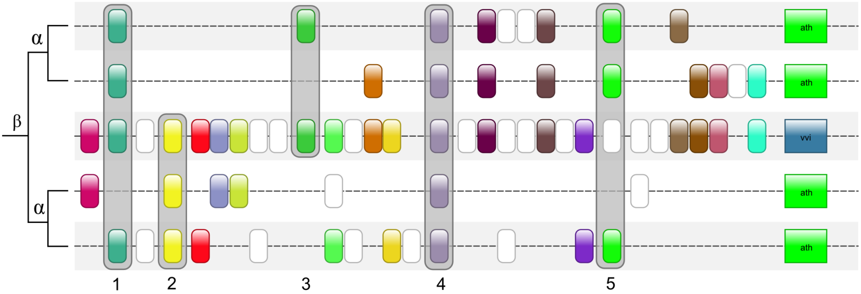


**Figure 5.3:** The gene order alignment of a grapevine region and four corresponding alpha/beta WGD *Aradidopsis thaliana* regions. Example of a multiplicon used to estimate the gene loss after the alpha and beta WGD. Whereas the rounded boxes represent the different genes color-coded according to the gene family they belong to, the square boxes at the right indicate the species the genomic segment was sampled from. Based on the presence of paralogs over the different segments, the different patterns of gene loss were quantified starting from loci present on the grapevine outgroup segment (Table 5.1). Consequently, some loci were excluded for further analysis (e.g. locus 5). Whereas locus 1 indicates loss of an alpha paralog, locus 2 refers to a loss after the beta WGD. Locus 3 reveals both a beta and alpha loss and locus 4 represent complete retention of all duplicates.

## 5.3 Biological Significance of Ultra-Conserved Multispecies Collinearity

Starting from 25 293 genomic scaffolds present in the Ensembl data set, 319 245 multiplicons were identified, some of which contained homologous regions from more than 20 species. The 'largest' multiplicon contained 33 segments from 22 species and included several homeobox Dlx proteins. Several HOX gene clusters including homeobox transcription factors were also found in a few high-level multiplicons (HOX D, level 28; HOX C and HOX D, level 22; HOX A and HOX D, level 25; HOX B and HOX D, level 20). This region is known to be highly conserved across species because these genes, involved in development of the body plan, require correct order to function[192]. The HOX cluster was duplicated and retained during two rounds of WGD in the ancestor of the vertebrates over 450 Mya[136], and since then the HOX A, HOX B, HOX C and HOX D clusters diverged significantly[193]. Many genes coding for interacting proteins are robust against rearrangements[131] and clusters of coexpressed genes conserved between human and mouse have been reported[194]. Given the large set of species, regions where gene order is strongly conserved over a large phylogenetic distance were delineated. Next, we assessed whether genes in these strongly conserved regions showed significant functional clustering. Briefly, experimental protein-protein interaction data and coexpression information were used to determine whether a highly conserved region was significantly enriched for interacting proteins or genes showing coordinated expression profiles. Coexpression is frequently used as a strong indicator for functionally related genes ("guilt by association"). Also, interacting protein pairs are known to have a high chance to be involved in the same biological process[195]. From the output of the high-quality subset, multiplicons with a strong conservation between either chicken or the songbird zebra finch (*Taeniopygia guttata*)[196], human, and at least five other mammals were extracted. Out of these 2863 multiplicons, 466 regions containing 2424 human genes, were significantly enriched (p-value $< 0.05$) for coexpressed pairs and/or gene pairs coding for interacting proteins (Figure 5.4). Mapping of these regions to a chromosome conservation plot depicting collinearity with all the 23 species included revealed that these regions are often among the most conserved in the genome (Figure 5.5).

Significant enrichment of coexpressed and interacting protein pairs points toward an evolutionary constraint to conserve gene order in these regions. These results provide further evidence that gene order in vertebrates is non-random and might play a considerable role in regulation of gene expression. However, the precise mechanism of the observed coexpression remains an open question, because both transcription factors, chromatin modifications[197], and long range enhancers are likely candidates to play a role in this process.

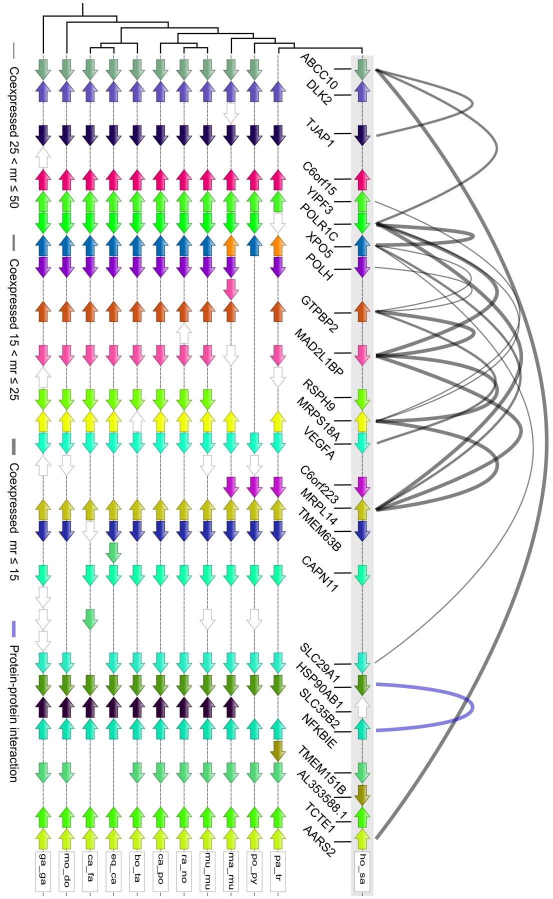


**Figure 5.4:** Gene order alignment of collinear regions conserved over a large phylogenetic distance (human-chicken). Species to which the segments belong are given on every line by the boxes on the right: *Homo sapiens* (ho_sa), *Pan troglodytes* (pa_tr), *Pongo pygmaeus* (po_py), *Macaca mulatta* (ma_mu), *Mus musculus* (mu_mu), *Rattus norvegicus* (ra_no), *Cavia porcellus* (ca_po), *Bos taurus* (bo_ta), *Equus caballus* (eq_ca), *Canis familiaris* (ca_fa), *Monodelphis domestica* (mo_do) and *Gallus gallus* (ga_ga). Arrows indicate coding genes and their orientation. Homologous genes are depicted in the same color. Coexpressed gene pairs are linked by black curved lines, of which the thickness of the line corresponds to the coexpression level (based on the mutual rank of the human genes in CoXPRESDB). Blue curved lines link pairs of genes coding for interacting proteins (in human). This region was found to be significantly enriched for coexpressed genes and, therefore, a biological constraint might cause gene order in this region to be retained.

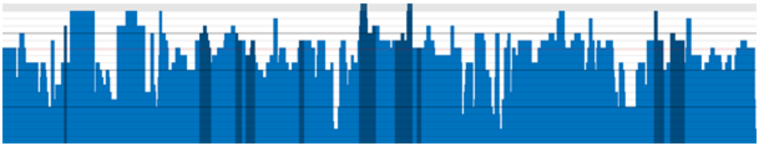


**Figure 5.5:** Conservation plot of the human chromosome 3. The height of the bars marks the number of species showing collinearity with that part of the chromosome. The areas in dark blue correspond to multiplicons significantly enriched (p-value $< 0.05$) for coexpressing or interacting gene pairs. The red line indicates the average conservation level whereas the horizontal gray bar gives the 5% of genes with the highest conservation level.

## 5.4 The Genomic Evolution of the Domesticated Apple

The domesticated apple (*Malus x domestica* Borkh., family Rosaceae, tribe Pyreae) is the main fruit crop of temperate regions of the world. Therefore, to assist breeding programs and the development of novel cultivars , recently a high-quality draft genome sequence of the diploid apple cultivar 'Golden Delicious' was generated[51].

Pairwise comparison of 17 apple chromosomes highlighted strong collinearity between large segments of chromosomes 3 and 11, 5 and 10, 9 and 17, and 13 and 16, and between shorter segments of chromosomes 1 and 7, 2 and 7, 2 and 15, 4 and 12, 12 and 14, 6 and 14, and 8 and 15 (Figure 5.6A). The distribution of synonymous substitution rates ($K_S$), an indication of the relative age of duplication, based on the number of synonymous substitutions in the coding sequences-peaked around 0.2 for recently duplicated genes (Figure 5.6B), indicating that a (recent) WGD has shaped the genome of the domesticated apple.

Dating of this WGD was based on the construction of penalized likelihood trees, as described in Fawcett et al.[74]. Given a node of grape to rosids fixed at 115 MYA, the WGD has been dated to between 30 and 45 MYA. If similar rates of protein evolution are assumed for apple and poplar (Figure 5.6C), the recent apple WGD may be as old as that of poplar, about 60 to 65 MYA[34].

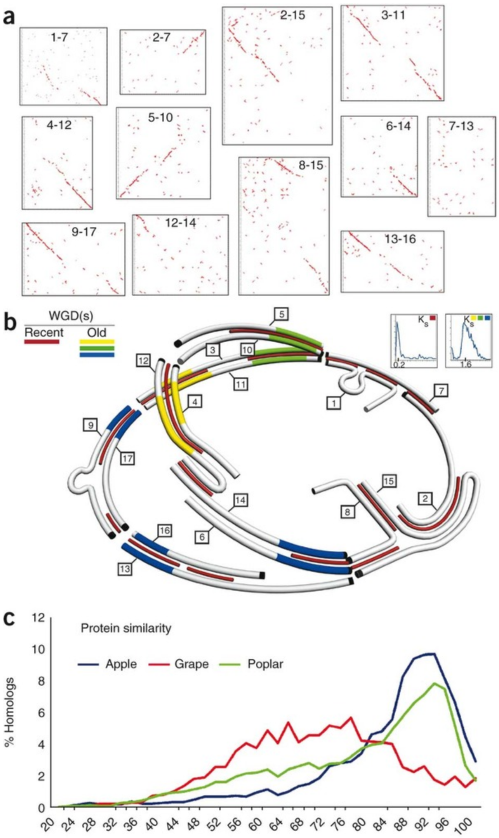


**Figure 5.6:** (a) Alignment of apple chromosomes shown by pairwise dot plots based on gene homology. Strong collinearity of members of chromosome doublets, or of large chromosome segments, indicates a recent WGD (red dots and bars in a and b, respectively). Unrelated chromosomes 7 and 13 were compared as a negative control. (b) Reconstruction of the relationships among apple chromosomes based on the most recent and the older WGD. The chromosomes ends represented at bottom right corners in a are marked in black in b. Red bars, regions of synteny that support the recent WGD. Size of chromosomes is proportional to their DNA content in megabases. Segments of chromosomes 1, 5, 6, 8, 10, 13, 14 and 15 have no syntenic counterparts. Chromosome segments predicted to be the outcome of the older duplication are highlighted with blue, green and yellow. Chromosomes 1, 2, 7, 8 and 15 do not show obvious signs of the older duplications, although they may contain short blocks of genes that reveal old paleopolyploid events. Inset graphs show that $K_S$ from the comparisons between paralogous genes has a peak at 0.2 when the recent duplication is considered, and between 1.4 and 1.6 for the older paleopolyploid events. (c) Distributions of protein similarities for duplicated genes in duplicated segments compared with grape (red), poplar (green) and apple (blue).

Remnants of older large-scale gene duplications or WGDs were also evident (Figure 5.7). Genes in these duplicated regions had average K$_S$-values around 1.6, as expected for paleoduplication events (Figure 5.6 B). Most remnants of these older duplications are found between chromosomes 5 and 10 and chromosomes 3 and 11, between chromosomes 3 and 11 and chromosomes 4 and 12, and between chromosomes 6 and 14, 13 and 16, and 9 and 17 (Figure 5.6 A and B). Chromosomes 1, 2, 7, 8 and 15 seem relatively devoid of older duplicated blocks; however, short blocks of genes showing old polyploidy events were found on all chromosomes. One region in the apple genome with an approximate size of 4 to 7 Mbp seems to be clearly present in six copies (regions in blue, Figure 5.6 A and B). Remapping those to the ancestral state reveals a triplicate structure among parts of chromosomes 9 and 17, 6 and 14 and 13 and 16. Notably, we found that these regions are collinear with chromosomes 1, 14 and 17 of grape (Figure 5.7), which have been demonstrated to be homologous because of an ancient hexaploidy[46]. Additional chromosomal fragments that we found to be duplicated in apple (green and yellow bars in Figure 5.6B) can also be interpreted as remains of a paleohexaploid state of the eudicot progenitor on the basis of dot-plot comparisons among other grape and apple chromosomes. This provides further evidence for a paleohexaploid state shared by most eudicots[28,53].

The chromosome homologies derived from the recent WGD allow inference of the cytological events that have led to the number and composition of the extant apple chromosomes, starting from a putative nine-chromosome ancestor (Figure 5.8). Each doublet of the eight apple chromosomes (3–11, 5–10, 9–17 and 13–16) is derived principally from one ancestor, although minor interchromosomal rearrangements have occurred. Chromosomes 4, 6, 12 and 14 originate from duplications of the ancient chromosomes V and VI, followed by a translocation and a deletion event. Similar events have generated chromosomes 1, 2, 7, 8 and 15 from chromosomes VII, VIII and IX. Chromosome 15 could have been produced from the translocation of an entire copy of chromosome IX into the centromeric region of chromosome VIII, following a model of dysploidy (reduction of chromosome number) common in cereals[198]. The second copy of ancient chro-

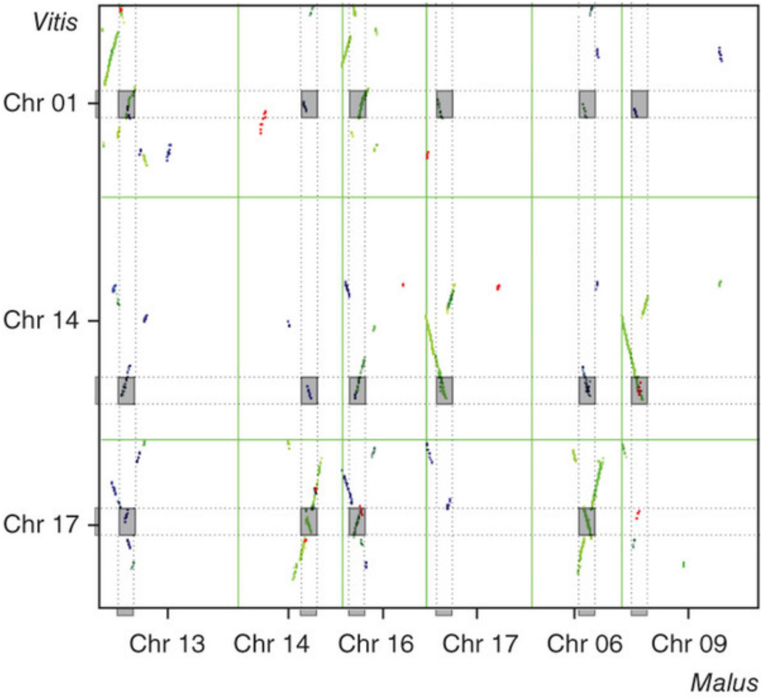


**Figure 5.7:** Dot plots are based on gene homology. The apple chromosomes are those with the segment triplication deriving from an old WGD (shown in blue in Figure 5.6B). Grape chromosomes 1, 14 and 17 constitute a triplet having the same ancestor in common7. Chromosome segments with homologous genes common both to grape and apple (16 of a total of 18 comparisons) are indicated by gray boxes connected with dashed lines. Green, red and blue dots indicate increasing $K_S$-values, in that order. Perpendicular lines on the x and y axes mark the middle of each chromosome. Green grid separates chromosomes.

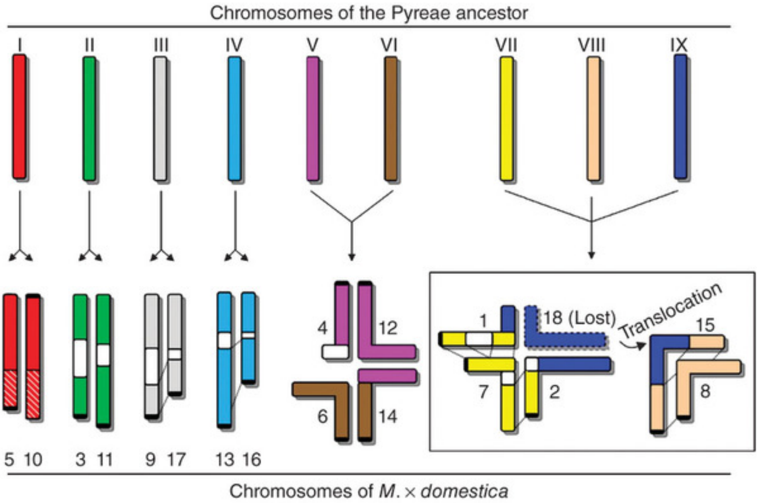


**Figure 5.8:** A WGD followed by a parsimony model of chromosome rearrangements is postulated. Shared colors indicate homology between extant chromosomes. White fragments of chromosomes indicate lack of a duplicated counterpart. The white-hatched portions of chromosomes 5 and 10 indicate partial homology. Black marks at chromosome ends correspond to those in Figure 5.6B.

mosome VIII has evolved into the extant chromosome 8. A conservative estimate of the number of large chromosome rearrangements since the divergence of the Pyreae subtribe, corresponding to the recent chromosome duplication, includes one chromosome fusion (extant chromosome 15), three translocations (involving extant chromosomes 1, 2 and 14), six deletions defined by telomeres that are not currently duplicated (chromosomes 4, 6, 8, 10, 11 and 13), one intrachromosome deletion (within chromosome 7, according to the chromosome 1-chromosome 7 comparison) and a deletion of a centromere (from ancient chromosome IX).

An intriguing aspect of the apple's biology concerns its characteristic fruit, the pome, which is found only in the Pyreae tribe[199]. This indicates that the pome probably evolved after a relatively recent Pyreae-specific WGD, a polyploidization step that we hypothesize has contributed to the apple's developmental and metabolic specificity. Pome fruit is derived by enlargement of the receptacle, which is the region below the whorl of sepals in the apple flower. MADS-box genes may regulate pome development, as they determine the eventual fate of floral tissues in all plant species analyzed so far[200]. For example, it has recently been shown that an apple MADS-box gene that is a member of the AP1 clade, common to all flowering plants[201] and closely related to *Arabidopsis thaliana* FRUITFULL (FUL), is differentially expressed during pome development[202]. In addition, a substantial number of apple type II MADS-box genes belong, phylogenetically, to the StMADS11 subclade, a group named for its first reported member, which was isolated from potato[203]. This subclade includes only two Arabidopsis genes, SVP and AGL24. Ectopic overexpression of SVP and related genes in *Arabidopsis thaliana* leads to foliose sepal syndrome-that is, the formation of large sepals[204]. In apple, this specific subclade not only includes two genes expressed in the pome but is also expanded to include 15 other genes.

A number of models have been proposed to explain the uniquely high number of chromosomes in Pyreae, the most popular being the 'wide-hybridization' hypothesis based on an allopolyploidization event between spireoid ($x = 9$) and amygdaloid ($x = 8$) ancestors[205,206]. More recent molecular phylogeny studies point to the possibility that Pyreae originated by autopolyploidization or by hybridization between two sister taxa with $x = 9$ (similar to extant Gillenia), followed by diploidization and aneuploidization[207] to $x = 17$. This hypothesis takes into account that Gillenia and related taxa are New World species and that the earliest fossil evidence of specimens belonging to extant genera of Pyreae are from North America.

Our results support the autopolyploidization hypothesis[207], as the derivation from a Gillenia-like taxon best fits the available data. First, the apple genome derives from a relatively recent duplication. Relationships between its homologous chromosomes based on genome sequence extend observations based on synteny and collinearity of molecular markers[208,209]. The timing of such a WGD, as estimated from our genomic data (Figure 5.6C), agrees with archeobotanical dates of 48–50 MYA[210].

In addition, a simple and parsimonious pattern of chromosome breakage and fusion explains the derivation of the current x = 17 Pyreae karyotype from a polyploidization event of two x = 9 genomes (Figure 5.8). The rate of chromosome rearrangements after polyploidization (12 chromosome events in 60 My) is similar to that for poplar (∼16 events in 60 My)[34] and lower than in maize (at least 17 chromosome fusion events in 5 My)[211] or in artificial neopolyploids[212]. In this sense, molecular clocks of perennial woody species seem slower than those of annual species, in terms of both nucleotide substitutions and chromosome rearrangements. For the genus Helianthus, a similar observation that only some of the ancestor chromosomes are rearranged in the extant chromosomes has been discussed in detail. In this genus, such rearrangement was associated with chromosomal differences between two sister species contributing to a WGD allopolyploid event[213].

## 5.5 Medicago truncatula Genome Evolution

Legumes (Fabaceae or Leguminosae) are unique among cultivated plants for their ability to carry out endosymbiotic nitrogen fixation with rhizobial bacteria, a process that takes place in a specialized structure known as the nodule. Legumes belong to one of the two main groups of eurosids, the Fabidae, which includes most species capable of endosymbiotic nitrogen fixation[214]. Legumes comprise several evolutionary lineages derived from a common ancestor 60 million years ago. Papilionoids are the largest clade, dating nearly to the origin of legumes and containing most cultivated species[215]. *Medicago truncatula* is a long-established model for the study of legume biology.

### 5.5.1 A WGD Shaped Legume Genomes

Recent analyses of plant genomes indicate a shared whole-genome triplication preceding the rosid–asterid split at 140–150 MYA[24]. Duplication patterns and genomic comparisons strongly suggest an additional WGD approximately 58 MYA in the papilionoids[216,217]. Near the time of this WGD, papilionoids radiated into several clades, the largest of which split quickly into two subclades, the Hologalegina (including *M. truncatula* and *L. japonicus*) and the milletioids (including *G. max* and other phaseoloids) at about 54 MYA[215].We therefore compared M. truncatula pseudomolecules with other sequenced plant genomes to learn more about shared synteny and genome duplication history.

There is significant macrosynteny among *M. truncatula*, *L. japonicus* and *G. max* (Figure 5.9 and 5.10). Conserved blocks, sometimes as large as chromosome arms, span most euchromatin in all three genomes. A given *M. truncatula* region is typically syntenic with one other *M. truncatula* region as a result of the WGD approximately 58 MYA, usually in small blocks showing degraded synteny (Figure 5.11 and 5.10). A given *M. truncatula* region is most similar to two *G. max* regions via speciation at about 54 MYA and the Glycine WGD at, $<$ 13 MYA[50] and less similar to two other G. max regions resulting from the ∼58 MYA and $<$ 13 MYA WGD events. A *M. truncatula* region is likewise most similar to one *L. japonicus* region via speciation at about 50 MYA and less similar to a second *L. japonicus* region as a result of the ∼58 MYA WGD. Finally, each *M. truncatula* region and its homeologue typically show similarity to three *Vitis vinifera* regions via the pre-rosid triplication. Exceptions to these patterns could be due to gene losses, gains, or rearrangements specific to the *M. truncatula* lineage, resulting in synteny being more evident between *M. truncatula* and other genomes than in self-comparisons. Indeed, self-comparisons within *M. truncatula* reveal few remnants of the legume-specific WGD (Figure 5.11 and 5.10). Whereas this seems paradoxical, it is probably explained by extensive gene fractionation between WGD-derived homeologues in *M. truncatula*. In Figure 5.13, two short regions on Mt1 and Mt3 resulting from the ∼58 MYA WGD are displayed beside microsyntenic regions of *G. max* and *V. vinifera*. As expected, many genes are microsyntenic between *M. truncatula* and *G. max* (ranging from 7/19 between Mt3 and Gm14 to 10/20 between Mt1 and Gm17). Between the two *M. truncatula* homeologues, however, only 6 out of 33 genes (or collapsed gene families)

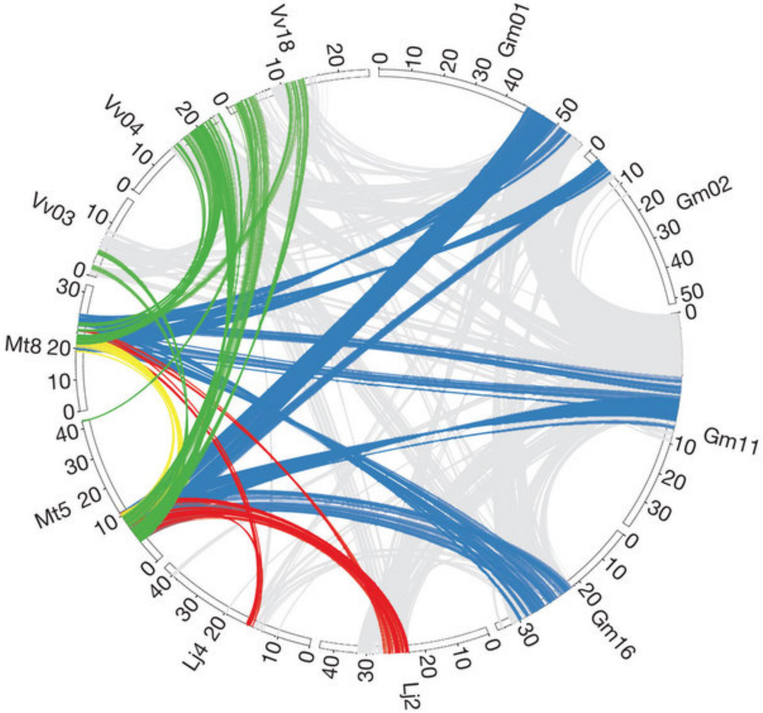


**Figure 5.9:** Circos diagram illustrating syntenic relationships between *Medicago*, *Glycine*, *Lotus* and *Vitis*. Homologous gene pairs were identified for all pairwise comparisons between *M. truncatula*, *G. max*, *L. japonicus* and *V. vinifera* genomes. Syntenic regions associated with the ancestral WGD events were identified by visually inspection of corresponding dot-plots. The large Mt5–Mt8 synteny block (yellow) was found to have two syntenic regions in *L. japonicus* (red), four syntenic regions in *G. max* (blue) and three in *V. vinifera* (green).

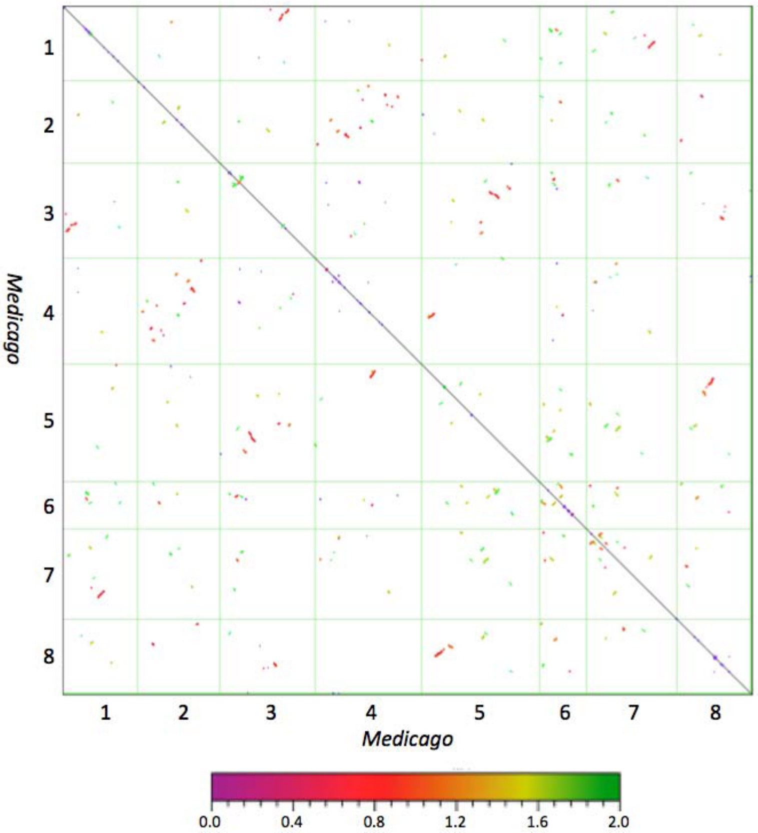


**Figure 5.10:** *Medicago X Medicago* Self-Comparison Dot-Plot. Axes represent the Mt genome compared with itself with gridlines separating different chromosomes. In the matrix, homologous gene pairs found in significant collinear regions (using i-ADHoRe) are indicated by a dot, colored according to the average $K_S$-value of all pairs between collinear segments. Hence, collinear regions appear as diagonal lines in the matrix, with their lengths and density roughly indicating the conservation of collinearity.

are microsyntenic, with a homeologue missing from one or the other duplicate. Apparently, there have been many more changes, large and small, in *M. truncatula* than in *G. max* since the legume WGD. This is borne out by the fact that synteny blocks in *M. truncatula* are one-third the length of those remaining from the papilionoid WGD in *G. max* (524 kb against 1503 kb) with the average number of homologous gene pairs per block correspondingly lower (12.4 against 31.0).

The *M. truncatula* genome also has undergone high rates of local gene duplication. The ratio of related genes within local clusters compared to all genes in families is 0.339 in *M. truncatula*, 3.1-fold higher than in *G.max* and 1.6-fold higher than in *A. thaliana* or *P. trichocarpa*. ('Local clusters' are defined as genes in a family all within 100 gene models of one another.) The excess of local gene duplications in *M. truncatula* is observed genome-wide and affects many families. There are 2.63 times as many gene families with local duplications in *M. truncatula* compared with *G. max* (2980 against 1131), an excess that also is seen in detailed comparisons of syntenic regions in *M. truncatula* and *G. max*. We examined 16.3Mb of Mt05 showing synteny to two large regions of Gm01 plus homeologous blocks on Gm02, Gm09 and Gm11. In these regions, 25.8% of *M. truncatula* genes are locally duplicated compared with just 8.0% in *G. max*. Local gene duplications and losses have contributed both to synteny disruptions (Figure 5.13) and to high gene count (62 388) in *M. truncatula* - a value nearly as high as the 65 781 total gene models in *G. max* despite its additional (<13 MYA) WGD. Local gene duplications are evident in certain gene families, such as F-box genes, which have undergone pronounced expansions. *M. truncatula* also has experienced higher rates of base substitution compared to other plant genomes (Figure 5.12). Assuming 58 Mya as the date of the legume WGD, then the rate of synonymous substitutions per site per year in *M. truncatula* is $1.08 \times 10^{-8}$, 1.8 times faster than estimates in other vascular plants[36]. Higher rates of mutation and greater levels of rearrangement in *M. truncatula* following the papilionoid duplication may have been driven by factors including short generation times, high selfing rates or small effective population sizes, although these characteristics are not unique to *M. truncatula*.

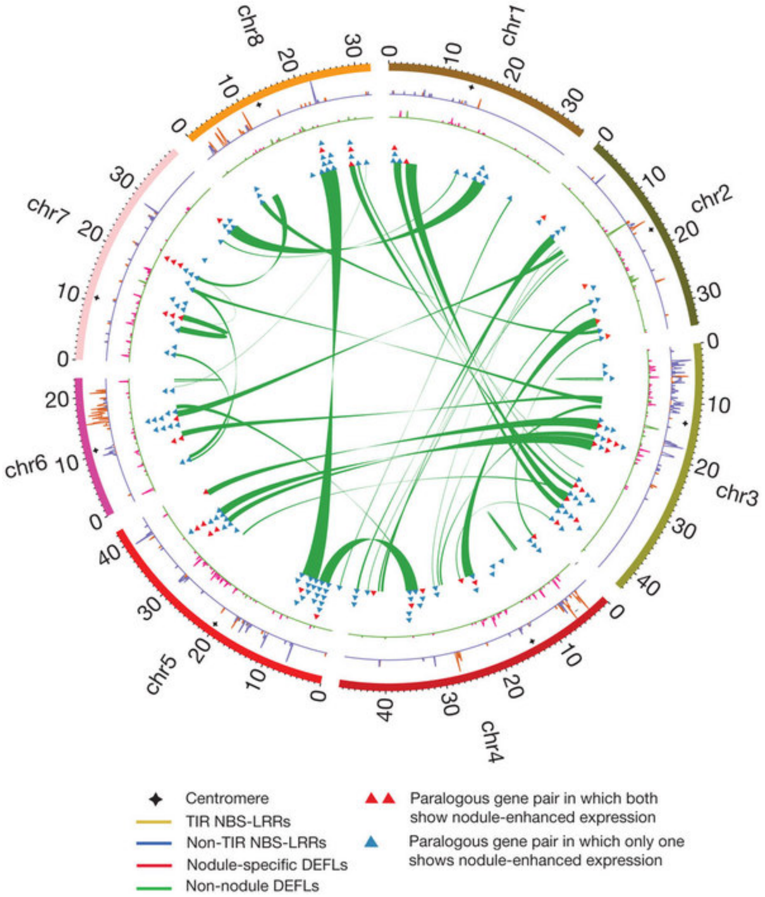


**Figure 5.11:** Circos diagram illustrating the *Medicago* WGD and selected gene families. The 963 WGD-derived paralogous gene pairs were examined for overlap with the nodule-enhanced gene list (Supplementary Data 2 accompanying Young et al.[184]). Resulting gene pairs were joined and plotted as either blue triangles (only one of the duplicates is nodule-enhanced) or red (both nodule enhanced). Gene densities of NBS-LRRs, NCRs and other defensin-like proteins are plotted against chromosome position. Density was calculated using a sliding window (100 kb window with 50kb steps).

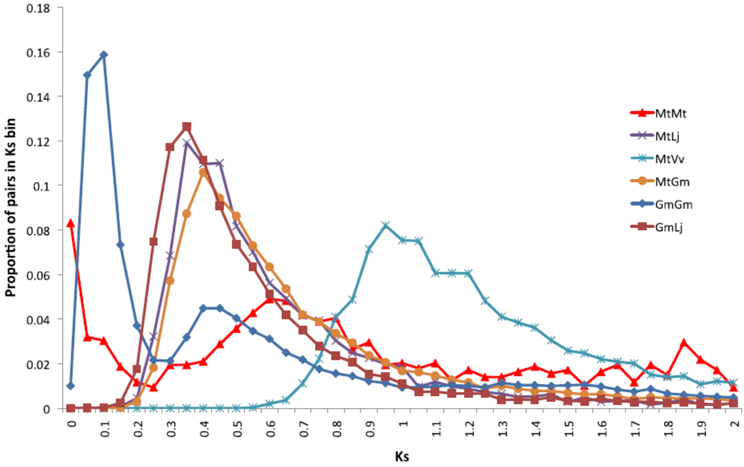


**Figure 5.12:** $K_S$-Analysis of Legume Species. Proportions of gene pairs per $K_S$-range for indicated species pairings. The species are Mt: *Medicago truncatula*; Lj: *Lotus japonicus*; Vv: *Vitis vinifera*; Gm: *Glycine max*. Proportions are taken as the counts of synteny-block paralogs within a $K_S$-range (with bin sizes of 0.05), divided by the count of all such paralogs with $K_S$-values less than or equal to two. Detection of collinear synteny blocks and $K_S$-dating are described in Proost et al. 2009[25]. For Mt–Mt comparison, $K_S$-values of 0 were removed (likely representing uncollapsed duplicate sequence contigs), as were paralogs values presumably resulting from local duplications (occurring on the same chromosome, within 100 gene models of one another). The shift of the legume WGD $K_S$-peaks in Mt-Mt vs. Gm–Gm is noteworthy (0.6 vs. 0.4), indicating more rapid accumulation of point mutations in Mt than in Gm.

### 5.5.2 The Legume-Specific WGD and the Evolution of Nodulation

Legumes and actinorhizal species are capable of forming a specialized organ, the root nodule, a highly differentiated structure hosting nitrogenfixing symbionts. Phylogenetic studies suggest that nodulation may have evolved multiple times in the Fabidae, but the observation that all nodulating species are contained within this single clade indicates that a predisposition to nodulate evolved in their common ancestor[218]. It is unknown whether nodulation with rhizobia preceded the divergence of the three legume subfamilies or evolved on multiple occassions[219].

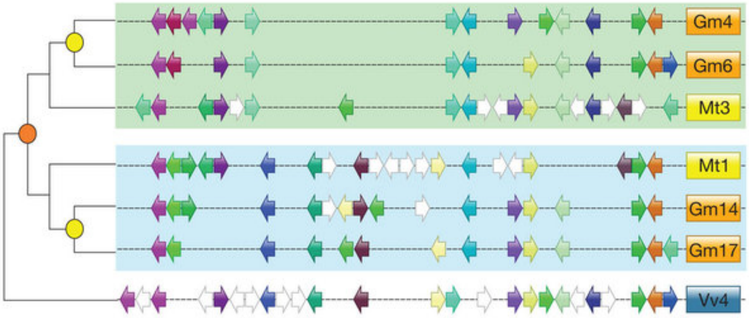


**Figure 5.13:** Microsynteny comparison between *Medicago* homeologues and corresponding regions of *Glycine* and *Vitis*. Microsyntenic genome segments are centred around Medtr3g104510/Medtr1g015890, a duplicated region derived from the ~58 MYA WGD event noted in orange. The <13 MYA *G. max*-specific WGD is coloured yellow. Orthologous/paralogous gene pairs are indicated through use of a common colour. White arrows represent genes with no syntenic homologue(s) in this genome region. Some of these genes may actually have a syntenic sequence in soybean but no corresponding model reported in the current annotation (http://www.phytozome.net/soybean).

Nevertheless, rhizobial nodulation and the 58 MYA WGD are features common to most papilionoid legumes and both occurred early in the emergence of the group[215]. Given that WGDs generate genetic redundancy that potentially facilitates the emergence of novel gene functions without compromising existing ones[220], we examined the *M. truncatula* genome to ask whether the 58 MYA WGD might have had a role in the evolution of rhizobial nodulation in *M. truncatula* and its relatives.

Nod factors are bacterial signalling molecules that initiate nodulation. Previous studies have shown that several of the plant components involved in the response to Nod factors also function in mycorrhizal signalling[221]. However, some Nod factor receptors and transcription factors have distinctly nodulation-specific functions. Among these nodulation-specific components, we found that the Nod factor receptor, NFP, and the transcription factor, ERN1, each have paralogues, LYR1 and ERN2 respectively, that trace back to the papilionoid WGD based on genome location and synonymous substitution rate values.

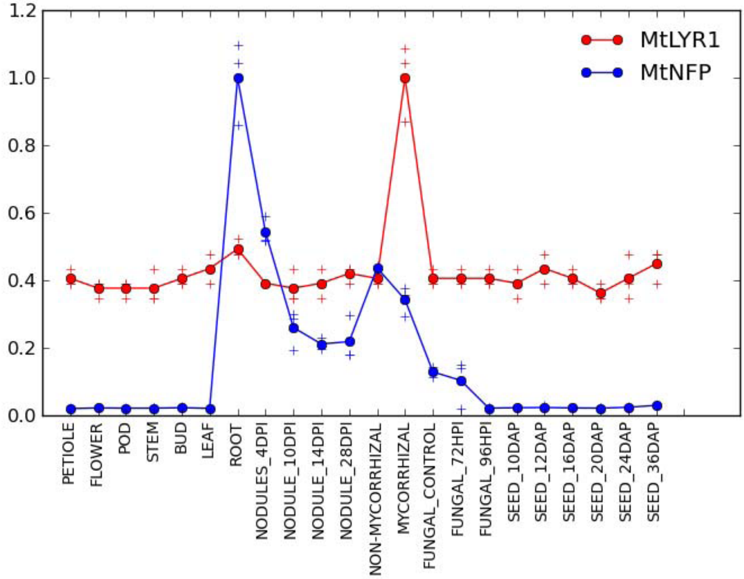


**Figure 5.14:** Expression profile NFP/LYR. Scaled transcript level of three replicates (+) and mean over replicates (dots connected by line) are shown. Scaling was performed by dividing each data point by maximum mean transcript level across experiments, independently for both paralogs in order to cope with overall differences in gene expression between them. Genes were mapped to the following probesets at http://bioinfo.noble.org/gateway/: Mtr.15789.1.S1_at (MtNFP), Mtr.19870.1.S1_at (MtLYR1), Mtr.7556.1.S1_at (MtERN1), Mtr.43947.1.S1_at (MtERN2).

Both sets of gene pairs also show contrasting expression patterns and functional specialization. NFP and ERN1 are expressed predominantly in the nodule and are known to function in nodulation[222,223], whereas LYR1 and ERN2 are highly expressed during mycorrhizal colonization (Figure 5.14). These observations indicate that two important nodulation-specific signalling components in *M. truncatula* might have evolved from more ancient genes originally functioning in mycorrhizal signalling and then duplicated by the 58 MYA WGD. In the case of *M. truncatula* NFP/LYR1, this conclusion is supported by the observation that the apparent orthologue of NFP in the nodulating non-legume *Parasponia andersonii* functions in both nodule and mycorrhizal signalling[224]. Thus, the 58 MYA WGD seems to have led to sub-functionalization of an ancestral gene participating in both interactions, resulting in two homeologous genes that each performs just one of the original functions.

To assess further the contribution of the WGD to *M. truncatula* nodulation, we analysed expression of paralogous gene pairs using RNA-seq data from six different organs. A total of 963 WGD-derived gene pairs were found with 618 pairs (1046 genes) having RNA-seq data for one or both homeologue. We then determined the number of genes showing organ-enhanced expression (defined as genes with expression level in a single organ at least twice the level in any other) within the pseudomolecule and the WGD-derived gene sets. In both cases, different organs contained markedly different numbers of genes with enhanced expression ($\chi^2$ with 5 degrees of freedom, P=$10^{-272}$); however, the rank order among the organs was identical. Roots had the largest number of genes with enhanced expression followed by flower, nodule, leaf, seed/pod and bud. Among gene pairs with nodule-enhanced expression, both paralogs were nodule-enhanced in eight pairs, whereas just a single paralog was nodule-enhanced in the other 43 pairs. This is consistent with nodulation pre-dating the WGD and further sub- and neo-functionalization emerging afterwards. We went on to examine transcription factors because they can act as regulators of plant growth and development. A total of 3692 putative TF genes were discovered (Supplementary Data 3 accompanying Young et al.[184]), representing 5.9% of all *M. truncatula* gene models. Of the 1513 TF genes on pseudomolecules with RNA-seq data, 142 genes (9.4%) derived from the 58 MYA WGD (Supplementary Data 4 accompanying Young et al.[184]), consistent with previous observations indicating greater retention of transcription factors following polyploidy[133]. Nodule-enhanced expression was significantly higher among transcription factors (92 out of 1513 or 6.1%) than among all pseudomolecule genes (1111 out of 23 478 or 4.7%) ($\chi^2$ with 1 degree of freedom, P=0.024). Nodule-enhanced expression was even higher in WGD-derived transcription factors (11 out of 142 or 7.7%), although this enrichment did not reach statistical significance (P=0.113). As expected, ERN1 is found within this group of WGD-retained, nodule-enhanced transcription factors.

## 5.6 Conclusion

From the numerous examples shown in this chapter it is clear both PLAZA and i-ADHoRe are highly versatile tools and can be used for a wide variety of studies. From literature, additional examples can be found where data derived from PLAZA was used in one way or another[225–228]. Hence the tools presented here are not just tailored to our own needs, but are well received in the scientific community as well.

## 5.7 Author Contribution

In this chapter case studies from Proost et al.[25],[126] are presented, these analyses have been designed and performed by me. As a first author on these publications, images and text were also generated by myself. From Velasco et al.[51], Young et al.[184] sections that reflect my contributions have been extracted and redrafted in this chapter. Additionally, downstream analyses based on data I generated are included.

*“Science is simply common sense at its best - that is, rigidly accurate in observation, and merciless to fallacy in logic.”*

Thomas H. Huxley

# 6 Conclusion and Future Prospects

Genomics is currently among the fastest evolving fields. So called *Next Generation Sequencing* (NGS) technologies, like Roche's 454 FLX[229], Illumina's HiSeq[230,231] and Life Technologies' SOLiD are now common practice. Meanwhile, third generation techniques, promising even faster,cheaper and/or longer reads are emerging. Helicos[232] and PacBio[233] being two examples that can do single molecule sequencing, while Ion Torrent[234] technology reduces the size of the machines to fit comfortable on a desktop. For electronics progress seems to follow Moore's Law, which states that 'the number of transistors that can be placed inexpensively on an integrated circuit doubles every two years', genomic data however grows faster. Hence bioinformaticians worldwide are confronted with a daunting amount of data, that grows at an ever increasing rate, while improvements in hardware cannot keep up. For bioinformaticians this can be seen as a challenge, however for experimental biologists this potentially becomes a major issue as the biological interpretation of such large data sets becomes problematic.

In this work we've successfully shown that using clever implementations and support for modern hardware can significantly improve existing bioinformatics tools. While integration of various datatypes in combination with a user-friendly web-interface allows non-expert users to browse comparative genomics data efficiently.

## 6.1 More Data Types

While more data, consisting out of extra genomes, is great for a comparative analysis, additional data types can help explain some findings and therefore should not be neglected. Quite recently a whole arsenal of novel, high-throughput methods have become available, many based on the latest sequencing technologies mentioned at the beginning of this chapter, that can give insights in different processes[235]. Bisulfite sequencing is providing us a genome wide view of epigenetic modifications[236], RNAseq[237] povides us a remarkably detailed view of the transcriptome and Chip-Seq[238] allows for transcription factor binding sites

(TFBS) to be found. Through integration of these new datatypes with existing structural and functional information various relations between for instance epigenetic changes and transcription levels can be observed, or if methylation patterns influence the binding of transcription factors, ...

Especially RNAseq has increased tremendously in popularity and is well on its way to replace microarrays entirely. The fact it doesn't rely on prior genomic knowledge of the organism (eg. Vera et al.[239]) is a significant advantage over traditional microarrays, as species-specific microarrays are bypassed. Additionally RNAseq has a remarkable sensitivity, this has two main consequences. First it is able to measure expression of genes expressed at very low levels, and thus gives a more complete image of the genes transcribed in a sample. Secondly, it allows samples with very little biological material to be analysed. Hence, in combination with laser capture microdissection, this allows expression to be measured in a single cell-type rather than a tissue[240]. Pushing technology even further would allow analysis of one cell, so called *single-cell genomics*. While useful for studying rare bacterias in microbial communities[241] in eukaryotic species this would allow hypervariable regions (which play for instance a considerable role in the development of cancer) of the genome to be studied[242].

Currently it's difficult to foresee how these additional datatypes should be integrated. Given a comprehensive dataset of epigenetic modifications, TFBS and expression levels, several biological questions can be answered. For instance, epigenetic modifications associated with high expression levels can be detected (and potently used for breeders to improve traits), by combining genes expressed in specific tissue- or cell-types with information of their TFBS detailed information on how this tissue- or cell-specificity is regulated can be obtained. However, while extremely interesting (and certainly the subject of currently ongoing research) this is still within a single species and does not contain a comparative aspect.

A first step towards using this novel data in a comparative way, would be to integrate an assembled transcriptome into PLAZA. Raw RNAseq reads (usually short reads, about 30 bp in length) can be assembled into longer transcript sequences using Velvet/Oases[243], ABySS[244] or Trinity[245]. Once the transcripome has been assembled, Open Reading Frames (ORF) can be discovered using FrameD[246] (if a training set is available) or FrameDP[247] (in the absence of training data). Once assembled transcripts have an valid ORF assigned, these can be seen as a short contig/scaffold with a single gene and thus entered as a new "genome" in PLAZA (Figure 6.1). Evidently i-ADHoRe cannot be used here to study the genome evolution, the gene families can be generated and studies much in the same way one would do with sequences derived from a genome. In this case some issues could occur due to read lengths (many transcripts only contain a fragment of a gene) and coverage (here it's difficult to assess the completeness), that might degrade the quality of the gene families generated and data generated downstream like the multiple sequence alignments and phylogenetic trees. Additionally, as a simple way to have an impression about the presence or absence of any WGDs in the species where only a transcriptome is available, a K$_S$-based dating system can be done.

While not a new idea, TFBS and epigenetic information can be mapped onto the gene tree (much like, in current versions of PLAZA, InterPro domains and intron-exon structure can be shown). This would allow to see how regulation changes or becomes more complex in different lineages[88]. After duplication the most likely fate of a copied gene is deletion[36]. However duplicates that remain can either remain unchanged (a way to boost mRNA production without increasing expression of the gene), sub-functionalize (the original function is distributed over both copies) or neo-functionalize (one copy retains the original function, whereas one copy acquires a new function). These three scenarios should be reflected in the expression and thus in the TFBS present in the promoter region of duplicated genes. Hence one can imagine in case of redundancy the TFBS remain unchanged when compared the ancestral state (as can be found in an unduplicated outgroup). Neo-functionalization should be indicated by one copy retaining the ancestral set of TFBS and one acquiring a new one, sub-functionalization would show an intermediate pattern.

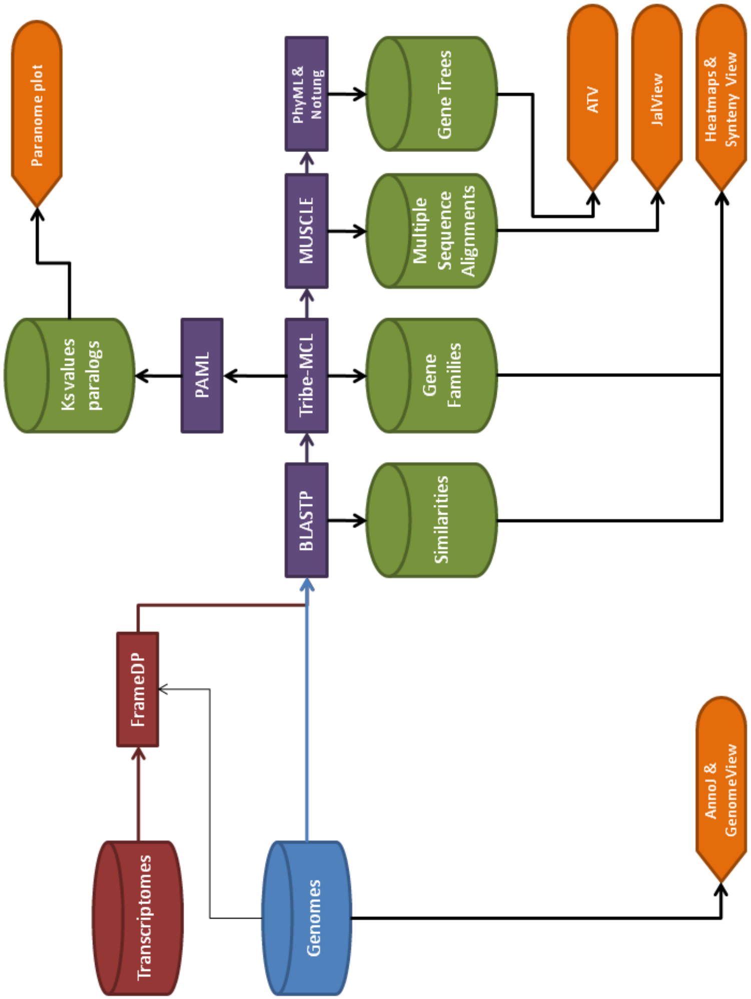


**Figure 6.1:** PLAZA flowchart with a small modification to include transcriptome data. The annotation of whole genomes can be used as a BLAST database for FrameDP to blast against. As the genome evolution pipeline cannot be used on transcriptome data, a system based on synonymous substitutions, also known as $K_S$, could be implemented to give an rough impression if and when WGDs occurred.

Similarly mapping epigenetic information on the gene tree could reveal recent adaptations to species-specific niches. Gene loss usually is a process that requires several generations, first through the accumulation of mutations the gene is reduced to a pseudogene and mutations continue to erode evidence of the gene until it can no longer be distinguished from an intergenic. Silencing through methylation however can turn expression of a gene off in a shorter timespan (and is widespread as a defense mechanism against foreign, eg. viral, DNA[248,249]), hence some of the genes we still find in genomes are in fact no longer necessary and turned off, despite not having had enough time to be cleared from the genome. Note that exceptionally methylation can also have a positive effect on the expression of a gene (eg. Makarevich et al.[250]).

A last thing worth giving some thought is how re-sequencing data needs to be handled in an comparative genomics platform. PLAZA 2.0 and higher include two distinct rice genomes, these still can be included in the ordinary fashion. But can the output of the 1001 Arabidopsis genomes[125] still be handled as such? Here another opportunity lies to expand the platform, mindlessly integrating hundreds of genomes will result in a bulk, slow and ultimately unattractive, unusable platform. One should note that in this case the similarities between the genomes are of little importance, between individuals of the same species the differences are important. Hence a preprocessing step might become important, based on all sequences available a consensus genome could be obtained and differences from various strains/individuals could be superimposed on that. This abstraction not only leads to a considerable decrease in data to analyze, but also allows for better retrieval of relevant data to answer biological questions. Though tools to perform Genome Wide Association Studies, like Hancock et al.[251] and Fournier-Level et al.[252], which should include geographical, climate, ... data, are quite beyond the scope of PLAZA.

## 6.2 Difficulties for Comparative Genomics

With the price for a draft-sequence a genome within the budget of smaller labs, new genomes are published at an ever increasing rate (see Figure 1.2). Unfortunately, the quality and completeness of these novel genomes cannot be compared to the finished reference genomes. Usually these genomes are sequenced to solve one specific question. Date palm (*Phoenix dactylifera*) is a recent example, here the genome is mainly sequenced to develop a quick test to determine the sex of seedlings[253]. Integrating such genomes and using them for comparative purposes however needs to be done with extreme caution. Genomes, such as Date palm, are not assembled into chromosomes, and as a consequence the majority of the scaffolds doesn't contain enough genes for i-ADHoRe to detect any collinear regions within the genome or with other species. Therefore only a minor fraction of the genome remains usable to study genome evolution, even large-scale events, such as WGDs, can effectively remain undetected because of this. Also the completeness of the genome should be considered an issue, if a large fraction of the genomes is missing members of some gene families will be missed and one might think this gene family to be contracted. The other way around if different alleles of the same locus are present on different scaffolds, artificially an expansion can be observed in certain gene families. Furthermore, sequence errors can lead to annotations errors, if genes are predicted too short, introns or exons might be missed and so on... All the above makes it difficult to discriminate genuine biological findings from sequencing/assembly/annotation artifacts. While in a pairwise comparison one might still account for these mistakes, when multiple species are compared this might lead to unexpected and erroneous results.

A second recurring issue that emerged while working on this topic is the lack of standards to release annotation. While in practice usually General Feature Format[a] (GFF) or eXtensible Markup Language (XML) based formats are used, these can come in different flavors. The GFF format in its pure form only defines features on a strand, what those features are or which features (such as exons)

[a] http://www.sanger.ac.uk/resources/software/gff/

belong together (to form a coding sequence) is ill defined and usually stored in the free-form description field. XML has no restriction on the tags used to describe genes and their properties, hence different sequencing initiatives do wander from the generally accepted scheme to specify certain features in a different way. While time consuming, these issues can be overcome manually (!), by writing or adjusting parsers to get the genomes in a single uniform format. A bigger concern is the lack of a gold standard to generate the sequence and perform the annotation. As described in the previous paragraph, quality is an issue, but the way the sequence is produced differs vastly from one project to another. A striking example of this can be found in the *Selaginella moellendorffii* genome[254], here both haplotypes are present in the genome sequence. The annotation is available in two versions one with all genes included (hence both haplotypes are present) or one with the different alleles hidden, retaining only those on the largest scaffold. Regardless of what version is used, this can have serious consequences when such a genome is used in a comparative environment. Retaining both alleles will show up in WGDotplots as a recent whole genome duplication, the gene counts are too large and gene families appear artificially expanded. When one allele is hidden, intergenics can no longer be correctly measured or extracted, gene density is underestimated, ... As such unconventional releases would take a considerable time to convert to a standard format, one is confronted with the decision to integrate the data as is with the risk of misinterpretation, correcting the annotation (what would mean using a version different from the official one) or excluding such genomes from the analysis entirely (and potentially loosing a whole community as users).

Finally a problem intrinsic to the clustering algorithms used to build the gene families is that for each update the whole pipeline needs to be run again. At first, with only one or two new genomes a year, this was not an issue. However, now it is no longer feasible to keep up with number genomes that are being published. While evidently this is an issue of the tools to build PLAZA and not the PLAZA pipeline itself, it is a problem we are currently confronted with. While much thought has gone into building an update procedure, where a genome is added without re-running the entire pipeline, this is hard to achieve. A better solution, is the expansion of a workbench to allow users to work with their own sequences rather then mapping their genes to a species included. This will allow users interested in working with data from a species not included in the database to add sequences to gene families, generate multiple sequence alignments and phylogenetic trees at the click of a button.

## 6.3 The Future of Comparative Genomics and Beyond

While merely five years ago, comparative genomics in plants was still in its infancy, this is evidently no longer the case. Continuous efforts to sequence additional plant species have generated a comprehensive dataset for bioinformaticians to work on. Genomics in any form is still very much in an explorative phase, where researchers aim to observe traits (in this case the genome sequences) and how ultimately they code for the wide variety of organisms that occur in nature. Here only the slight scratches have been made in the proverbial surface, there is still a large gap to bridge between changes we see in genome sequence and the phenotypes linked with them. However new projects, such as the iPlant Collaborative[a] [255], already have their minds set on closing this gap.

Such insights are of value for molecular breeders as this allows to pinpoint attractive genes to increase yield, reduce contents of unwanted components (such as lignin for biofuel crops), beautify flowers, improve resistance to pathogens, ... Using either traditional crossing with marker assisted selection or through the creation of Genetically Modified (GM) promising genes can be introduced into commercially interesting varieties. However through classic crosses, this is limited in phylogenetic distance and hence has limited applications (in addition this is also a lengthy process). Using genetic, modification genes can be introduced in other

[a] www.iplantcollaborative.org/

species regardless of the phylogenetic distance. Unfortunately these GM crops are under constant attack by various environmental groups and specifically in the EU, are bound to strict laws which limits their application. While obviously the large-scale cultivation of new GM crops has to be approached with the necessary test and studies, once found safe these new crops have significant potential to improve the food industry, provide energy, ... So without a change in both legislation and public opinion this will, quite unfortunately, stifle practical application of results obtained through comparative genomics.

*Metagenomics* is another novel field which makes heavy use of comparative genomics techniques. Here a whole community of (usually prokaryotic) organisms is sequenced at once and through computational techniques one can assess the abundance of certain species, the biodiversity of the sample, ... This is currently being used to study a wide range of bacterial communities with ecological importance to health related human gut bacteria[256,257]. As the sensitivity of sequencing technologies increases, the amount of data generated in such projects will increase (especially low abundant bacteria will appear). Hence to further support such studies comparative genomics tools need to scale with the growing amount of data.

Once a profound knowledge of the link between genotype and phenotype is obtained[258], the field of *Synthetic Biology* will boom. Pioneer studies performed in the lab of J. Craig Venter (the first human to have his genome sequenced) were already able to synthetically duplicate the DNA of a bacteria[259]. However, this effort was extremely labor intensive as it took a team of researchers 15 years to accomplish this and the selected genome is with 1 Mb extremely small. So while current DNA sequencing technology allows for fast reading of a genome, writing DNA is far behind. But if technological advances in writing DNA follow the rapid, exponential, growth we've seen in sequencing technology the last decade, it's a matter of mere years before DNA molecules in the length range of eukaryotic chromosomes can be written.

So one thing is certain, for years to come there is a dire need for comparative genomics and tools that give access to such data to researches worldwide. A major challenge here will be to design tools that are not tailored towards one specific study one would like to perform, but tools that despite the huge amount of available data remain flexible, fast and user-friendly.

*"If I have done the public any service,*
*it is due to my patient thought."*

Sir Isaac Newton

# A
# Summary

Since both the *Arabidopsis thaliana* and the rice genomes were sequenced, comparative genomics finally started within the plant field as it did in prokaryotes, fungi and animals already a few years earlier. While powerful to study genome evolution, adaptation,... , comparative genomics comes with a steep learning curve, extensive comprehension of tools involved is required as is a strong knowledge of computer programming to write the necessary programs to extract biological data from large data sets. As the number of sequenced plant genomes quickly augmented, so did the hardware requirements to perform multi-species analysis. Despite the fact many experimental biologists make frequent use of gene families, functional annotation and genomic homology to explain their observations or to exchange knowledge between different organisms, doing this without access to high memory computers and computer clusters generating the necessary data has become impossible.

During this thesis, a considerable amount of time has been spend on generating comparative genomics data in such a way it becomes accessible for plant researchers worldwide. We opted for a web based interface linked to a relational database, as such a user-friendly way to query the database can be provided, additionally this provided an access point from which more complex tools, such as visualizations and advances statistics, can be started. The main focus of this platform, called PLAZA, is gene family evolution on the one hand and genome evolution on the other. The gene families included allow users, for instance, to quickly retrieve lineage- or species-specific gene families, those can be linked to ecological adaptations. Additionally, by clearly defining homologous (derived from the same ancestor) genes, in depth knowledge of a specific gene in one species can be transfered to other organisms.

Plant genomes appear to be highly dynamic in comparison with animal genomes and, the recent evolution of plants appears to be governed by several independent whole genome duplications (WGD). To study the effects of these i-ADHoRe, a tool to detect such duplications and to find homologous regions between genomes, has been integrated in PLAZA. This allows users to retrieve genes duplicated during these large-scale duplication events as well as studying how genomes evolved after the WGD. However while building PLAZA 2.0, the i-ADHoRe version available at the time proved to be unable to cope with the number of genomes included. Thus emerged the need for an update, three major issues hindered the detection of collinear (regions with homologous genes preserved in a similar order) and duplicated regions, from the technical side the runtime and memory usage were growing out of proportion, while the alignment strategy, a key component of the detection algorithm proved to be unable to correctly align sets of ten or more collinear genomic regions. To cope with the technical issues a new version was developed which makes cunning use of optimization and support for modern hardware to achieve a significant speedup and reduction in memory footprint. This update also includes better statistical models and a novel alignment algorithm that is able to generate high quality gene order alignments for several dozens of segments.

Besides development of new and improvement of novel tools, a set of case studies is described in this thesis. These show how a wide range of studies can be performed using (a combination of) the tools developed during this thesis. Ranging from estimating the gene loss in *Arabidopsis thaliana* to elaborating on the biological significance of conserved collinear regions in vertebrates, these tools show an remarkable potential as an starting point for a variety of analyses. To further illustrate this the PLAZA pipeline was used as part of two genome projects, to dissect the genomic evolution of the domesticated apple (*Malus domestica*) and the barrel medic (*Medicago truncatula*). By integrating the newly sequenced and annotated genomes along with some reference genomes and relevant outgroups (such as the moss *Physcomitrella patens* and a few algae (like *Chlamydomonas reinhardtii* and *Ostreococcus lucimarinus*) a wide range of the topics traditionally discussed in genome papers (such as detection of WGD, presence of expanded gene families, ...) could quickly and efficiently be obtained. Furthermore by combining this knowledge with additional experimental techniques insights in how key processes of those species evolved could be detected.

To conclude, the recent increase in sequence data is a double-edged blade, that while potentially giving remarkable insights in biology and evolution of species, the analysis becomes increasingly complex. Here we present two tools that are a considerable improvement over the current state-of-the-art and show how researchers worldwide can use them on a wide variety of biological challenges.

*"I have nothing to declare except my genius."*

Oscar Wilde

# B Nederlandse Samenvatting

Zodra naast het genoom van *Arabidopsis thaliana* ook dat van rijst beschikbaar werd, kon ook in planten aan vergelijkend genoomonderzoek gedaan worden. Dit enkel jaren nadat er reeds verschillende prokaryote, schimmel en dierlijke genomen gesequeneerd waren. Hoewel enorm krachtig, heeft dit veld een zeer stijle leercurve en is een diepgaande kennis van verschillende programma's vereist. Eveneens is ervaring met programmeer- of scripting-talen noodzakelijk om uit zulke datasets biologische kennis te verwerven. Doordat het aantal gesequeneerde plantengenomen bleef toenemen, was ook steeds krachtigere hardware vereist voor zulk onderzoek te doen. Hoewel experimentele biologen veelvuldig gebruik maken van genfamilies, functionele beschrijvingen van genen en genomische homologie om hun observaties te verklaren of om kennis verworven in andere soorten over te dragen naar het organisme van interesse, is dit zonder toegang tot krachtige supercomputers of computer clusters niet meer mogelijk.

Gedurende deze thesis werd, naast het verelijken van genomen, een groot deel van de tijd gewijd aan het ontwikkelen van technieken om deze data op een toegankelijke manier te presenteren aan een breed publiek. Hiervoor werd gekozen om een platform met *web-interface* te gebruiken. Via deze weg is het mogelijk om in de bijhorende database met gerichte zoekopdrachten snel relevante data op te vragen. Bovendien kunnen zo visualisaties en complexere berekeningen eenvoudig gestart worden. Deze website, die als naam PLAZA meekreeg, bestaat uit twee luiken, ten eerste is een deel gewijd aan het bestuderen van gen families. Het tweede deel is gericht op het ontrafelen van genoom evolutie. Genfamilies laten gebruikers toe om snel expansies te vinden in soorten of groepen die gelinkt kunnen zijn aan ecologische adaptaties. Bovendien laten correct afgelijnde genfamilies het toe om kennis, verworven in een (model-)organisme, over te dragen naar een andere soort.

Recente ontdekkingen tonen aan dat plantengenomen zeer dynamisch zijn en gekenmerkt zijn door de aanwezigheid van meerdere volledige genoomduplicaties. Om de gevolgen van zulke duplicaties te bestuderen werd i-ADHoRe geïntegreerd in PLAZA. Deze integratie laat gebruikers toe om genen, gedupliceerd tijdens dergelijke grootschalige duplicaties, nader te bekijken alsook de gevolgen op het genoom verder te onderzoeken. Echter, tijdens het ontwikkelen van PLAZA 2.0, bleek dat de versie van i-ADHoRe waarover we op dat tijdstip beschikten, niet in staat was de hoeveelheid beschikbare genomen te analyseren. Een update drong zich op! Zo doken er drie problemen op, twee van technische aard; de tijd die nodig was om de detectie van homologe gebieden uit te voeren en de hoeveelheid werkgeheugen die hiervoor nodig was. Bovendien bleek het algoritme om gen lijsten te aligneren niet in staat om tientallen homologe regio's correct te aligneren. In de nieuwe versie werd, met behulp van efficiënte algoritmen en ondersteuning voor moderne hardware, de snelheid gevoelig opgedreven terwijl bovendien het gebruikte werkgeheugen binnen de perken kon worden gehouden. Verder werd ook een nieuw alignerings algoritme ontwikkeld wat, in tegenstelling tot voorgaande implementaties, wel in staat was tientallen gen lijsten correct te aligneren.

Naast de ontwikkeling van nieuwe methoden en het verbeteren van bestaande, worden in deze thesis ook meerdere toepassingen beschreven. Hiermee wordt duidelijk aangetoond hoe deze tools (of de combinatie van beide) aan de basis kunnen liggen van een brede waaier van analyses. De voorbeelden gaan van het schatten van gen verlies in *Arabidopsis thaliana* tot het bestuderen van de biologische relevantie van regio's met sterk behouden genvolgorde, in gewervelden. Om de kracht van PLAZA verder aan te tonen werd het ingeschakeld voor de analyses van de genomen van appel en *Medicago truncatula*. Door een platform te maken met de nieuwe genoomsequentie, enkele referentie genomen en *outgroups* waaronder een mos en algen, konden resultaten die doorgaans in genoompublicaties verschijnen zeer snel bekomen worden. Bovendien was het mogelijk, door deze data te combineren met experimenteel verkregen data, nieuwe inzichten te krijgen in enkele commercieel interessante processen aanwezig in deze soorten.

Samengevat, in vele opzichten is de hoeveelheid sequentie data waarover we momenteel beschikken een mes dat aan twee kanten snijdt. Enerzijds kan deze data enorme inzichten verschaffen over diverse soorten, anderzijds wordt het steeds moeilijker om biologische inzichten te destilleren uit een dergelijke hoeveelheid data. De software en technieken hier besproken zijn echter een belangrijke verbetering ten opzichte van de huidige stand van zaken in vergelijkend genoomonderzoek. Zonder twijfel kunnen onderzoekers wereldwijd deze software gebruiken om allerhande biologische vragen te beantwoorden.

*"You think you're pretty clever, don't you? I happen to know that every word in your book was published* years *ago! Perhaps you've read...the dictionary!"*

Dick Solomon

# C
# Curriculum Vitae

# Personal Information

**Name** Sebastian Proost

**Address** Steenweg 219, 3590 Diepenbeek BELGIUM

**Tel.** +32 (0)494 14 10 25

**E-mail** sebastian.proost@gmail.com

**Date of Birth** October 19, 1985

# Education

**2007-2012** PhD in Biotechnology, UGent

**2005-2007** Master in Biotechnology (Distinction), UGent

Option: Bioinformatics, Genetic Manipulation of Model Organisms and Genome Biology

Master thesis: Genomewide prediction of protein-protein interactions in *Arabidopsis thaliana*. Promoter: Prof. Dr. Yves Van de Peer, Supervisor: Dr. Stefanie De Bodt

**2003-2005** Bachelor in Biology (Distinction), UHasselt

Option: Philosophy of Biology

# Publications

**Citations**[‡]**:** 255 **H-Index**[‡]: 6

1. S. De Bodt, **S. Proost**, K. Vandepoele, P. Rouze, and Y. Van de Peer. Predicting protein-protein interactions in arabidopsis thaliana through integration of orthology, gene ontology and co-expression. *BMC Genomics*, 10:288, 2009

2. 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

[‡] measured using *Publish or Perish 3*, March 8, 2012

3. **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 (* contributed equally)

4. 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

5. 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*, 27(6):749-756, 2011 (* contributed equally)

6. **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

7. N. D. Young, F. Debellé, G. Oldroyd, R. Geurts, S. B. Cannon, M. K. Udvardi,V. A. Benedito, K. F. X. 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. Bergès, S. Bidwell, T. Bisseling, N. Choisne, A. Couloux, R. Denny, S. Deshpande, J. J. Doyle, A. Dudez, A. D. Farmer, S. Fouteau, C. Franken, C. Gibelin, J. Gish, A. J. González, P. J. Green, A. Hallab, M. Hartog, A. Hua, S. Humphray, D. Jeong, Y. Jing, A. Jöcker, S. M. Kenton, D. Kim, K. Klee, H. Lai, C. Lang, S. Lin, S. L. Macmil, G. Magdelenat, L. Matthews, J. McCorrison, E. L. Monaghan, J. Mun, F. Z. Najar, C. Nicholson, C. Noirot, C. R. Paule, J. Poulain, F. Prion, B. Qin, C. Qu, E. F. Retzel, C. Riddle, E. Sallet, S. Samain, N. Samson, O. Saurat, C. Scarpelli, T. Schiex, B. Segurens, M. Seigfried, A. Severin, D. J. Sherrier, R. Shi, S. Sims, S. Sinharoy, L. Sterck, I. Vasylenko, A. Viollet, K. Wang, B. 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. Dénarié, R. A. Dixon, G. D. May, D. C. Schwartz, J. Rogers, F. Quétier, C. D. Town, and B. A. Roe. The medicago genome provides insight into evolution of rhizobial symbiosis. *Nature*, 480(7378):520-524, 2011

8. M. Van Bel*, **S. Proost***, Elisabeth Wischnitzki, Sara Mohavedi, Christopher Scheerlinck, Yves Van de Peer, and Klaas Vandepoele. Dissecting plant genomes with the PLAZA comparative genomics platform. *Plant Phys*, 158(2):590-600, 2012 (* contributed equally)

9. **S. Proost***, J. Fostier*, D. De Witte, B. Dhoedt, P. Demeester, Y. Van de Peer, and K. Vandepoele. i-ADHoRe 3.0 - Accurate and Sensitive Detection of Genomic Homology in Extremely Large Datasets. *Nucleic Acids Res*, 40(2), 2012 (* contributed equally)

10. D. Vekemans*, **S. Proost***, H. Coenen, T. Viaene, P. Ruelens, Y. Van de Peer, K. Geuten. Gamma paleohexaploidy in the stem-lineage of core eudicots: significance for MADS-box gene and species diversification. *Manuscript submitted to Mol Biol Evol* (* contributed equally)

# Scientific Activity

## Meetings and workshops with active participation

**6th Annual RECOMB Satellite Meeting on Comparative Genomics**, October 13-15, 2008, Ecole Normale Supérieure, Paris, France. Poster presentation.

**BioMaGNet's Annual Meeting**, October 22, 2009, VIB-Ghent University, Ghent, Belgium. Oral presentation and poster presentation.

**Brassicacea Map Alignment Project - 2nd BMAP meeting**, March 22, 2010, Max Planck Institute, Tübingen, Germany. Oral presentation.

**First European Student Council Symposium (ESCS1)**, September 26, 2010, Ghent, Belgium. Poster presentation.

**9th European Conference on Computational Biology (ECCB2010)**, September 26-29, 2010, Ghent, Belgium. Poster presentation.

**BioMaGNet's Annual Meeting**, March 21, 2011, ULB Campus Plaine, Brussels, Belgium. Poster presentation.

**Comparative & Regulatory Genomics in Plants - Conference**, April 11-12, 2011, PSB/VIB, Gent, Belgium. Organizing committee, oral presentation and poster presentation.

**Comparative & Regulatory Genomics in Plants - Workshop**, April 13-15, 2011, UGent, Gent, Belgium. Organizing committee, lecturer.

**GMPF Summer School 2011: Introduction to Bioinformatics**, June 28 - July 1, 2011, IASMA, San Michele all'Adige, Italy. Invited speaker.

**Plant Genome Evolution**, September 4-6, 2011, Amsterdam, The Netherlands. Poster presentation and software demonstration.

**Bioinformatics Affinity Seminar**, February 8, 2012, MPI-MP, Potsdam-Golm, Germany. Guest Speaker.

**Vibes 2012 - 3th International VIB PhD Symposium**, September 5-7, 2012, Gent, Belgium. Organizing committee.

## Other meetings and workshops

**Ensembl Workshop** (by Bert Overduin), September 20, 2007, Ghent, Belgium.

**Benelux Bioinformatics Conference 2007**, November 12-13, 2007, Leuven, Belgium.

**Vibes 2010 - 2nd International VIB PhD Symposium**, October 14-15, 2010, Leuven, Belgium.

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

**Effective Oral Presentations** (by Dr. Jean-Luc Doumont), February 1,7,10,21, 2011, Ghent, Belgium.

**Project Management** (by Tom Jacobs), March 2,9,16, 2011, Ghent, Belgium.

**Creative Thinking** (by Karl Raats), April 20-22, 2011, Ghent, Belgium.

**Unix Systems for Bioinformatics** (by Dr. Lieven Sterck), 2011-2012, Ghent, Belgium.

# Educational Support

## Supervised students

**Marieke Dubois**, 1st Master in Biochemistry and Biotechnology, UGent, 2008-2009, Masterproject: *Analysis of pseudogenes in Arabidopsis thaliana*.

**Steven Timmermans**, 1st Master in Biochemistry and Biotechnology, UGent, 2009-2010, Masterproject: *Homology identification through sequence similarity clustering of multidomain proteins*.

**Christopher Scheerlinck**, Master in Industrial Engineering: Informatics, UGent, 2009-2010, Master thesis: *Integration and visualization of evolutionary genomics data within an online comparative genomics platform*.

## Courses and Lectures

**Functional Plant Genomics**(by Dr. Pierre Hilson), 2011-2012, Ghent, Belgium. Guest Lecture: *Introduction: Comparative Genomics Using the PLAZA Platform*.

**Comparative Genomics**(by Prof. Dr. Klaas Vandepoele), 2009-2012, Ghent, Belgium. Assisted practical courses.

# Honors and Awards

Flemish Biology Olympiad 2003 - First Place

International Biology Olympiad, July 8-16, Minsk, Belarus - Awarded Bronze Medal

# Service as Reviewer

BMC Bioinformatics

# Skills

## ICT

**Programming Languages**: PERL, JAVA and C/C++/C#

**Office**: Extensive knowledge of Microsoft Office, Openoffice.org and LaTeX; Familiar with image processing software (Photoshop and Inkscape)

**Specialties**: Bioinformatics, databases (MySQL), comparative genomics, genome evolution, genome analysis, tool development

## Languages

Dutch, mother tongue

Fluent in written and spoken English

Proficient in French

Basic knowledge of German

*“DNA is an abbreviation for deoxyribonucleicantidisestablishmentarianism, a complex string of syllables.”*

Dave Barry

# D
# List of Abbreviations

# List of Abbreviations

| | |
|---|---|
| AVG | Average |
| BLAST | Basic Local Alignment Search Tool |
| DNA | DeoxyriboNucleic Acid |
| ENCODE | ENCyclopedia Of DNA Elements |
| EST | Expressed Sequence Tag |
| FDR | False Discovery Rate |
| FP | False Positive |
| GFF | General Feature Format |
| GG | Greedy Graph |
| GHM | Gene Homology Matrix |
| GM | Genetically Modified |
| GO | Gene Ontology |
| GWAS | Genome Wide Association Studies |
| i-ADHoRe | iterative Automatic Detection of Homologous Regions |
| IEA | Inferred from Electronic Annotation |
| ISS | Inferred from Sequence or Structural Similarity |

| | |
|---|---|
| K$_S$ | Synonymous substitution rate |
| K-T | Cretaceous-Tertiary |
| LL | Longest Link |
| LLBS | Lowest Lower Bound Score |
| LS | Lowest Score |
| Mb | Megabases |
| MPI | Message Passing Interface |
| MSA | Multiple Sequence Alignment |
| MYA | Million Years Ago |
| NGS | Next Generation Sequencing |
| ORF | Open Reading Frame |
| pNW | progressive Needleman-Wunsch |
| RA | RAndom |
| RAC | Random Active Conflict |
| RC | Random Conflict |
| SNP | Single Nucleotide Polymorphisms |
| sp. | species |
| ssp. | subspecies |
| TF | Transcription Factor |
| TFBS | Transcription Factor Binding Site |
| TP | Time Point |
| WGD | Whole Genome Duplication |
| XML | eXtensible Markup Language |

*"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

# E Bibliography

# Bibliography

[1] 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.

[2] 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.

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

[4] 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.

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

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

[7] D. Butler. Science after the sequence. *Nature*, 465(7301):1000–1001, 2010.

[8] E. Birney, J.A. Stamatoyannopoulos, A. Dutta, R. Guigó, T.R. Gingeras, E.H. Margulies, Z. Weng, M. Snyder, E.T. Dermitzakis, R.E. Thurman, et al. Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature*, 447(7146):799–816, 2007.

[9] M.B. Gerstein, Z.J. Lu, E.L. Van Nostrand, C. Cheng, B.I. Arshinoff, T. Liu, K.Y. Yip, R. Robilotto, A. Rechtsteiner, K. Ikegami, et al. Integrative analysis of the caenorhabditis elegans genome by the modencode project. *Science*, 330(6012):1775, 2010.

[10] S. Roy, J. Ernst, P.V. Kharchenko, P. Kheradpour, N. Negre, M.L. Eaton, J.M. Landolin, C.A. Bristow, L. Ma, M.F. Lin, et al. Identification of functional elements and regulatory circuits by drosophila modencode. *Science*, 330(6012):1787, 2010.

[11] P.R. Burton, D.G. Clayton, L.R. Cardon, N. Craddock, P. Deloukas, A. Duncanson, D.P. Kwiatkowski, M.I. McCarthy, W.H. Ouwehand, N.J. Samani, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.

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

[13] 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.

[14] T. Dobzhansky. Nothing in biology makes sense except in the light of evolution. *American Biology Teacher*, 35(3):125–129, 1973.

[15] R.C. Hardison. Comparative genomics. *PLoS biology*, 1(2):e58, 2003.

[16] TS Mikkelsen, LW Hillier, EE Eichler, MC Zody, DB Jaffe, SP Yang, W. Enard, I. Hellmann, K. Lindblad-Toh, TK Altheide, et al. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69–87, 2005.

[17] D.P. Locke, L.D.W. Hillier, W.C. Warren, K.C. Worley, L.V. Nazareth, D.M. Muzny, S.P. Yang, Z. Wang, A.T. Chinwalla, P. Minx, et al. Comparative and demographic analysis of orang-utan genomes. *Nature*, 469(7331):529–533, 2011.

[18] A. Kuzniar, R. C. van Ham, S. Pongor, and J. A. Leunissen. The quest for orthologs: finding the corresponding gene across genomes. *Trends in genetics : TIG*, 24(11):539–551, November 2008.

[19] S. Hampson, A. McLysaght, B. Gaut, and P. Baldi. Lineup: statistical detection of chromosomal homology with application to plant comparative genomics. *Genome research*, 13(5):999–1010, 2003.

[20] X. Wang, X. Shi, Z. Li, Q. Zhu, L. Kong, W. Tang, S. Ge, and J. Luo. Statistical inference of chromosomal homology based on gene colinearity and applications to arabidopsis and rice. *BMC bioinformatics*, 7(1):447, 2006.

[21] T. Hachiya, Y. Osana, K. Popendorf, and Y. Sakakibara. Accurate identification of orthologous segments among multiple genomes. *Bioinformatics*, 25(7):853, 2009.

[22] C. Rödelsperger and C. Dieterich. Syntenator: Multiple gene order alignments with a gene-specific scoring function. *Algorithms for Molecular Biology*, 3(1):14, 2008.

[23] C. Rödelsperger and C. Dieterich. CYNTENATOR: progressive gene order alignment of 17 vertebrate genomes. *PloS one*, 5(1):e8861+, January 2010.

[24] H. Tang, X. Wang, J. E. Bowers, R. Ming, M. Alam, and A. H. Paterson. Unraveling ancient hexaploidy through multiply aligned angiosperm gene maps. *Genome Res*, 18(12):1944–54, 2008.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] Ewan Birney, T. Daniel Andrews, Paul Bevan, Mario Caccamo, Yuan Chen, Laura Clarke, Guy Coates, James Cuff, Val Curwen, Tim Cutts, Thomas Down, Eduardo Eyras, Xose M. Fernandez-Suarez, Paul Gane, Brian Gibbins, James Gilbert, Martin Hammond, Hans-Rudolf Hotz, Vivek Iyer, Kerstin Jekosch, Andreas Kahari, Arek Kasprzyk, Damian Keefe, Stephen Keenan, Heikki Lehvaslaiho, Graham McVicker, Craig Melsopp, Patrick Meidl, Emmanuel Mongin, Roger Pettett, Simon Potter, Glenn Proctor, Mark Rae, Steve Searle, Guy Slater, Damian Smedley, James Smith, Will Spooner, Arne Stabenau, James Stalker, Roy Storey, Abel Ureta-Vidal, K. Cara Woodwark, Graham Cameron, Richard Durbin, Anthony Cox, Tim Hubbard, and Michele Clamp. An overview of ensembl. *Genome Research*, 14(5):925–928, May 2004.

[31] M. Koornneef and D. Meinke. The development of arabidopsis as a model plant. *Plant J*, 61(6):909–21, 2010.

[32] A. F. Bent. Arabidopsis in planta transformation. uses, mechanisms, and prospects for transformation of other species. *Plant Physiol*, 124(4):1540–7, 2000.

[33] 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.

[34] 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.

[35] K. Vandepoele and Y. Van de Peer. Exploring the plant transcriptome through phylogenetic profiling. *Plant Physiol*, 137(1):31–42, 2005.

[36] M. Lynch and J. S. Conery. The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494):1151–5, 2000.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] G. Blanc, K. Hokamp, and K. H. Wolfe. A recent polyploidy superimposed on older large-scale duplications in the arabidopsis genome. *Genome Res*, 13(2):137–44, 2003.

[43] J. E. Bowers, B. A. Chapman, J. Rong, and A. H. Paterson. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, 422(6930):433–8, 2003.

[44] C. Simillion, K. Vandepoele, M. C. Van Montagu, M. Zabeau, and

Y. Van de Peer. The hidden duplication past of arabidopsis thaliana. *Proc Natl Acad Sci U S A*, 99(21):13627–32, 2002.

[45] T. J. Vision, D. G. Brown, and S. D. Tanksley. The origins of genomic duplications in arabidopsis. *Science*, 290(5499):2114–7, 2000.

[46] 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.

[47] 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.

[48] N. Wikstrom, V. Savolainen, and M. W. Chase. Evolution of the angiosperms: calibrating the family tree. *Proc Biol Sci*, 268(1482):2211–20, 2001.

[49] A. Cenci, M. C. Combes, and P. Lashermes. Comparative sequence analyses indicate that coffea (asterids) and vitis (rosids) derive from the same paleo-hexaploid ancestral genome. *Mol Genet Genomics*, 283(5):493–501, 2010.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] M. Abrouk, F. Murat, C. Pont, J. Messing, S. Jackson, T. Faraut, E. Tannier, C. Plomion, R. Cooke, C. Feuillet, and J. Salse. Palaeogenomics of plants: synteny-based modelling of extinct ancestors. *Trends Plant Sci*, 15 (9):479–87, 2010.

[55] 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.

[56] D. E. Soltis, C. D. Bell, S. Kim, and P. S. Soltis. Origin and early evolution of angiosperms. *Ann N Y Acad Sci*, 1133:3–25, 2008.

[57] Y. Van de Peer, S. Maere, and A. Meyer. The evolutionary significance of ancient genome duplications. *Nat Rev Genet*, 10(10):725–32, 2009.

[58] D. Bikard, D. Patel, C. Le Mette, V. Giorgi, C. Camilleri, M. J. Bennett, and O. Loudet. Divergent evolution of duplicate genes leads to genetic incompatibilities within a. thaliana. *Science*, 323(5914):623–6, 2009.

[59] D. R. Scannell, K. P. Byrne, J. L. Gordon, S. Wong, and K. H. Wolfe. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature*, 440(7082):341–5, 2006.

[60] M. Semon and K. H. Wolfe. Reciprocal gene loss between tetraodon and zebrafish after whole genome duplication in their ancestor. *Trends Genet*, 23(3):108–12, 2007.

[61] L. Comai. The advantages and disadvantages of being polyploid. *Nat Rev Genet*, 6(11):836–46, 2005.

[62] T. C. Osborn, J. C. Pires, J. A. Birchler, D. L. Auger, Z. J. Chen, H. S. Lee, L. Comai, A. Madlung, R. W. Doerge, V. Colot, and R. A. Martienssen. Understanding mechanisms of novel gene expression in polyploids. *Trends Genet*, 19(3):141–7, 2003.

[63] L. H. Rieseberg, S. C. Kim, R. A. Randell, K. D. Whitney, B. L. Gross, C. Lexer, and K. Clay. Hybridization and the colonization of novel habitats by annual sunflowers. *Genetica*, 129(2):149–65, 2007.

[64] L. H. Rieseberg, O. Raymond, D. M. Rosenthal, Z. Lai, K. Livingstone, T. Nakazato, J. L. Durphy, A. E. Schwarzbach, L. A. Donovan, and C. Lexer. Major ecological transitions in wild sunflowers facilitated by hybridization. *Science*, 301(5637):1211–6, 2003.

[65] T. F. Stuessy. a transitional-combinatorial theory for the origin of angiosperms. *Taxon*, 53:3–16, 2004.

[66] W. L. Crepet. Progress in understanding angiosperm history, success, and relationships: Darwin's abominably "perplexing phenomenon". *Proc Natl Acad Sci U S A*, 97(24):12939–41, 2000.

[67] D. E. Soltis, V. A. Albert, J. Leebens-Mack, C. D. Bell, A. H. Paterson, C. Zheng, D. Sankoff, C. W. dePamphilis, P. K. Wall, and P. S. Soltis. Polyploidy and angiosperm diversification. *Am. J. Bot.*, 96:336–348, 2009.

[68] M. S. Barker, H. Vogel, and M. E. Schranz. Paleopolyploidy in the brassicales: analyses of the cleome transcriptome elucidate the history of genome duplications in arabidopsis and other brassicales. *Genome Biol Evol*, 1:391–9, 2009.

[69] M. S. Barker, N. C. Kane, M. Matvienko, A. Kozik, R. W. Michelmore, S. J. Knapp, and L. H. Rieseberg. Multiple paleopolyploidizations during the evolution of the compositae reveal parallel patterns of duplicate gene retention after millions of years. *Mol Biol Evol*, 25:2445–2455, 2008.

[70] G. Blanc and K. H. Wolfe. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell*, 16 (7):1667–78, 2004.

[71] 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.

[72] M. Lescot, P. Piffanelli, A. Y. Ciampi, M. Ruiz, G. Blanc, J. Leebens-Mack, F. R. da Silva, C. M. Santos, A. D'Hont, O. Garsmeur, A. D. Vilarinhos, H. Kanamori, T. Matsumoto, C. M. Ronning, F. Cheung, B. J. Haas, R. Althoff, T. Arbogast, E. Hine, Jr. Pappas, G. J., T. Sasaki, Jr. Souza, M. T., R. N. Miller, J. C. Glaszmann, and C. D. Town. Insights into the musa genome: syntenic relationships to rice and between musa species. *BMC Genomics*, 9:58, 2008.

[73] J. A. Schlueter, P. Dixon, C. Granger, D. Grant, L. Clark, J. J. Doyle, and R. C. Shoemaker. Mining est databases to resolve evolutionary events in major crop species. *Genome*, 47(5):868–76, 2004.

[74] J. A. Fawcett, S. Maere, and Y. Van de Peer. Plants with double genomes might have had a better chance to survive the cretaceous-tertiary extinction event. *Proc Natl Acad Sci U S A*, 106(14):5737–42, 2009.

[75] A. H. Paterson, J. E. Bowers, and B. A. Chapman. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc Natl Acad Sci U S A*, 101(26):9903–8, 2004.

[76] M. A. Beilstein, N. S. Nagalingum, M. D. Clements, S. R. Manchester, and S. Mathews. Dated molecular phylogenies indicate a miocene origin for arabidopsis thaliana. *Proc Natl Acad Sci U S A*, 107(43):18724–8, 2010.

[77] M. Freeling, E. Lyons, B. Pedersen, M. Alam, R. Ming, and D. Lisch. Many or most genes in arabidopsis transposed after the origin of the order brassicales. *Genome Res*, 18(12):1924–37, 2008.

[78] 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.

[79] Y. Van de Peer. Computational approaches to unveiling ancient genome duplications. *Nat Rev Genet*, 5(10):752–63, 2004.

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

[81] 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.

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

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

[84] 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.

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

[86] 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.

[87] 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.

[88] 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.

[89] 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.

[90] 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.

[91] 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.

[92] 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.

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

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

[95] 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.

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

[97] 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.

[98] 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.

[99] 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.

[100] 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.

[101] 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.

[102] 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.

[103] E. V. Koonin. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet*, 39:309–38, 2005.

[104] 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.

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

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

[107] 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.

[108] 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.

[109] 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.

[110] 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.

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

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

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

[114] 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.

[115] 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.

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

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

[118] 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.

[119] 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.

[120] 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.

[121] 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.

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

[123] 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.

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

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

[126] 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.

[127] M. Garcia-Diaz and T. A. Kunkel. Mechanism of a genetic glissando: structural biology of indel mutations. *Trends in Biochemical Sciences*, 31(4): 206 – 214, 2006.

[128] M. Hurles. Gene duplication: the genomic trade in spare parts. *PLoS biology*, 2(7):e206, 2004.

[129] E. Passarge, B. Horsthemke, and R.A. Farber. Incorrect use of the term synteny. *Nat. Genet*, 23(4):387, 1999.

[130] K.H. Wolfe. Yesterday's polyploids and the mystery of diploidization. *Nature Reviews Genetics*, 2(5):333–341, 2001.

[131] T. Makino and A. McLysaght. Interacting gene clusters and the evolution of the vertebrate immune system. *Molecular biology and evolution*, 25(9): 1855, 2008.

[132] K.P. Byrne and K.H. Wolfe. Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication. *Genetics*, 175(3):1341, 2007.

[133] B.C. Thomas, B. Pedersen, and M. Freeling. Following tetraploidy in an arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome research*, 16(7): 934, 2006.

[134] Y. Jiao, N.J. Wickett, S. Ayyampalayam, A.S. Chanderbali, L. Landherr, P.E. Ralph, L.P. Tomsho, Y. Hu, H. Liang, P.S. Soltis, et al. Ancestral polyploidy in seed plants and angiosperms. *Nature*, 2011.

[135] K. Vandepoele, W. De Vos, J. S. Taylor, A. Meyer, and Y. Van de Peer. Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc Natl Acad Sci U S A*, 101(6):1638–43, 2004.

[136] P. Dehal and J.L. Boore. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS biology*, 3(10):e314, 2005.

[137] K.H. Wolfe and D.C. Shields. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387(6634):708–713, 1997.

[138] M. Kellis, B.W. Birren, and E.S. Lander. Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisiae. *Nature*, 428(6983):617–624, 2004.

[139] D.R. Scannell, G. Butler, and K.H. Wolfe. Yeast genome evolutionŮthe origin of the species. *Yeast*, 24(11):929–942, 2007.

[140] C. Soderlund, M. Bomhoff, and W.M. Nelson. Symap v3. 4: a turnkey synteny system with application to plant genomes. *Nucleic Acids Research*, 39(10):e68, 2011.

[141] 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.

[142] B. Dujon, D. Sherman, G. Fischer, P. Durrens, S. Casaregola, I. Lafontaine, J. De Montigny, C. Marck, C. Neuvéglise, E. Talla, et al. Genome evolution in yeasts. *Nature*, 430(6995):35–44, 2004.

[143] N.S. Baliga, R. Bonneau, M.T. Facciotti, M. Pan, G. Glusman, E.W. Deutsch, P. Shannon, Y. Chiu, R.S. Weng, R.R. Gan, et al. Genome sequence of haloarcula marismortui: a halophilic archaeon from the dead sea. *Genome research*, 14(11):2221, 2004.

[144] A.C.E. Darling, B. Mau, F.R. Blattner, and N.T. Perna. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research*, 14(7):1394, 2004.

[145] C.N. Dewey and L. Pachter. Evolution at the nucleotide level: the problem of multiple whole-genome alignment. *Human molecular genetics*, 15(suppl 1):R51, 2006.

[146] C.N. Dewey. Aligning multiple whole genomes with mercator and mavid. *Methods in Molecular Biology*, 395:221, 2007.

[147] M. Blanchette, W.J. Kent, C. Riemer, L. Elnitski, A.F.A. Smit, K.M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E.D. Green, et al. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome research*, 14(4):708, 2004.

[148] 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.

[149] T. J. P. Hubbard, B. L. Aken, S. Ayling, et al. Ensembl 2009. *Nucl. Acids Res.*, 37(suppl_1):D690–697, January 2009.

[150] 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.

[151] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[152] S. Dudoit and M.J. van der Laan. *Multiple testing procedures with applications to genomics*. Springer Verlag, 2008.

[153] T. Obayashi, S. Hayashi, M. Shibaoka, M. Saeki, H. Ohta, and K. Kinoshita. Coxpresdb: a database of coexpressed gene networks in mammals. *Nucleic acids research*, 36(suppl 1):D77, 2008.

[154] B. Aranda, P. Achuthan, Y. Alam-Faruque, I. Armean, A. Bridge, C. Derow, M. Feuermann, AT Ghanbarian, S. Kerrien, J. Khadake, et al. The intact molecular interaction database in 2010. *Nucleic acids research*, 38(suppl 1):D525, 2010.

[155] C. Simillion, K. Vandepoele, Y. Saeys, and Y. Van de Peer. Building genomic profiles for uncovering segmental homology in the twilight zone. *Genome research*, 14(6):1095–1106, June 2004.

[156] D. Durand and D. Sankoff. Tests for gene clustering. *Journal of Computational Biology*, 10(3-4):453–482, 2003.

[157] D.F. Feng and R.F. Doolittle. Progressive sequence alignment as a prerequisitetto correct phylogenetic trees. *Journal of molecular evolution*, 25(4): 351–360, 1987.

[158] E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[159] A.T. Chinwalla, L.L. Cook, K.D. Delehaunty, G.A. Fewell, L.A. Fulton, R.S. Fulton, T.A. Graves, L.D.W. Hillier, E.R. Mardis, J.D. McPherson, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562, 2002.

[160] L.D.W. Hillier, W. Miller, E. Birney, W. Warren, R.C. Hardison, C.P. Ponting, P. Bork, D.W. Burt, M.A.M. Groenen, M.E. Delany, et al. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, 432(7018):695–716, 2004.

[161] O. Jaillon, J.M. Aury, F. Brunet, J.L. Petit, N. Stange-Thomann, E. Mauceli, L. Bouneau, C. Fischer, C. Ozouf-Costaz, A. Bernot, et al. Genome duplication in the teleost fish tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*, 431(7011):946–957, 2004.

[162] D.L. Altshuler, R.M. Durbin, G. Abecasis, D.R. Bentley, A. Chakravarti, A.G. Clark, F.S. Collins, F.M. de La Vega, P. Donnelly, M. Egholm, et al. A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073, 2010.

[163] D. Haussler, S.J. O'Brien, O.A. Ryder, F.K. Barker, M. Clamp, A.J. Crawford, R. Hanner, O. Hanotte, W.E. Johnson, J.A. McGuire, et al. Genome 10k: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. *J Hered*, 100(6):659–674, 2009.

[164] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, March 1970. ISSN 0022-2836.

[165] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, March 1981. ISSN 0022-2836.

[166] Desmond G. Higgins and Paul M. Sharp. Clustal: a package for performing multiple sequence alignment on a microcomputer. *Gene*, 73(1):237–244, December 1988.

[167] J. D. Thompson, D. G. Higgins, and T. J. Gibson. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, 22(22):4673–4680, November 1994. ISSN 0305-1048.

[168] C. Notredame, L. Holm, and D. G. Higgins. Coffee: an objective function for multiple sequence alignments. *Bioinformatics*, 14(5):407–422, June 1998. ISSN 1367-4803.

[169] K. Katoh, K. Misawa, K. Kuma, and T. Miyata. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic acids research*, 30(14):3059–3066, July 2002.

[170] D. Sankoff and M. Blanchette. Multiple genome rearrangement and breakpoint phylogeny. *Journal of Computational Biology*, 5:555–570, 1998.

[171] B. Morgenstern. Dialign 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, 15(3):211–218, March 1999.

[172] E. Corel, F. Pitschi, and B. Morgenstern. A min-cut algorithm for the consistency problem in multiple sequence alignment. *Bioinformatics*, 26 (8):1015–1021, April 2010.

[173] H. Lenhof, B. Morgenstern, and K. Reinert. An exact solution for the segment-to-segment multiple sequence alignment problem. *Bioinformatics*, 15:203–210, 1999.

[174] F. Pitschi, C. Devauchelle, and E. Corel. Automatic detection of anchor points for multiple sequence alignment. *BMC bioinformatics*, 11:445+, September 2010.

[175] L.R. Ford and D.R. Fulkerson. *Flows in networks*. Princeton University Press, Princeton, NJ, 1962.

[176] P. Elias, A. Feinstein, and C. Shannon. A note on the maximum flow through a network. *Information Theory, IRE Transactions on*, 2(4):117–119, 1956.

[177] J. D. Thompson, P. Koehl, R. Ripp, and O. Poch. Balibase 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins*, 61 (1):127–136, October 2005.

[178] G. P. S. Raghava, S. Searle, P. Audley, J. Barber, and G. Barton. Oxbench: A benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics*, 4(1):47+, October 2003.

[179] J. Schultz, F. Milpetz, P. Bork, and C. P. Ponting. Smart, a simple modular architecture research tool: Identification of signaling domains. *Proceedings of the National Academy of Sciences of the United States of America*, 95 (11):5857–5864, May 1998.

[180] H. Carroll, W. Beckstead, T. O'Connor, M. Ebbert, M. Clement, Q. Snell, and D. Mcclellan. Dna reference alignment benchmarks based on tertiary structure of encoded proteins. *Bioinformatics*, 23(19):2648–2649, October 2007.

[181] C. Notredame, D. G. Higgins, and J. Heringa. T-coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, 302(1):205–217, September 2000. ISSN 0022-2836.

[182] G. Fritzsch, M. Schlegel, and P. F. Stadler. Alignments of mitochondrial genome arrangements: Applications to metazoan phylogeny. *Journal of Theoretical Biology*, 240:511–520, 2006.

[183] S. K. Pham and P. A. Pevzner. DRIMM-Synteny: decomposing genomes into evolutionary conserved segments. *Bioinformatics*, 26:2509–2516, 2010.

[184] N.D. Young, F. Debellé, G.E.D. Oldroyd, R. Geurts, S.B. Cannon, M.K. Udvardi, V.A. Benedito, K.F.X. Mayer, J. Gouzy, H. Schoof, et al. The medicago genome provides insight into the evolution of rhizobial symbioses. *Nature*, 480(7378):520–524, 2011.

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

[186] 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.

[187] 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.

[188] 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.

[189] 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.

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

[191] 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 (4):333–45, 1995.

[192] E.B. Lewis et al. A gene complex controlling segmentation in drosophila. *Nature*, 276(5688):565, 1978.

[193] D. Lemons and W. McGinnis. Genomic evolution of hox gene clusters. *Science*, 313(5795):1918, 2006.

[194] G.A.C. Singer, A.T. Lloyd, L.B. Huminiecki, and K.H. Wolfe. Clusters of co-expressed genes in mammalian genomes are conserved by natural selection. *Molecular biology and evolution*, 22(3):767–775, 2005.

[195] S. De Bodt, S. Proost, K. Vandepoele, P. Rouze, and Y. Van de Peer. Predicting protein-protein interactions in arabidopsis thaliana through integration of orthology, gene ontology and co-expression. *BMC Genomics*, 10:288, 2009.

[196] W.C. Warren, D.F. Clayton, H. Ellegren, A.P. Arnold, L.D.W. Hillier, A. Künstner, S. Searle, S. White, A.J. Vilella, S. Fairley, et al. The genome of a songbird. *Nature*, 464(7289):757–762, 2010.

[197] C. Wu. Chromatin remodeling and the control of gene expression. *Journal of Biological Chemistry*, 272(45):28171, 1997.

[198] MC Luo, KR Deal, ED Akhunov, AR Akhunova, OD Anderson, JA Anderson, N. Blake, MT Clegg, D. Coleman-Derr, EJ Conley, et al. Genome comparisons reveal a dominant mechanism of chromosome number reduction in grasses and accelerated genome evolution in triticeae. *Proceedings of the National Academy of Sciences*, 106(37):15780–15785, 2009.

[199] D. Potter, T. Eriksson, RC Evans, S. Oh, JEE Smedmark, DR Morgan, M. Kerr, KR Robertson, M. Arsenault, TA Dickinson, et al. Phylogeny and classification of rosaceae. *Plant Systematics and Evolution*, 266(1):5–43, 2007.

[200] M. Ng and M.F. Yanofsky. Activation of the arabidopsis b class homeotic genes by apetala1. *The Plant Cell Online*, 13(4):739–754, 2001.

[201] H. Shan, L. Zahn, S. Guindon, P.K. Wall, H. Kong, H. Ma, J. Leebens-Mack, et al. Evolution of plant mads box transcription factors: evidence for shifts in selection associated with early angiosperm diversification and concerted gene duplications. *Molecular biology and evolution*, 26(10):2229–2244, 2009.

[202] B. Janssen, K. Thodey, R. Schaffer, R. Alba, L. Balakrishnan, R. Bishop, J. Bowen, R. Crowhurst, A. Gleave, S. Ledger, et al. Global gene expression analysis of apple fruit development from the floral bud to ripe fruit. *BMC plant biology*, 8(1):16, 2008.

[203] A. Becker and G. Theißen. The major clades of mads-box genes and their role in the development and evolution of flowering plants. *Molecular phylogenetics and evolution*, 29(3):464–489, 2003.

[204] S. Masiero, M.A. Li, I. Will, U. Hartmann, H. Saedler, P. Huijser, Z. Schwarz-Sommer, and H. Sommer. Incomposita: a mads-box gene controlling prophyll development and floral meristem identity in antirrhinum. *Development*, 131(23):5981–5990, 2004.

[205] E. Chevreau, Y. Lespinasse, and M. Gallet. Inheritance of pollen enzymes and polyploid origin of apple (malus x domestica borkh.). *TAG Theoretical and Applied Genetics*, 71(2):268–277, 1985.

[206] J.B. Phipps, K.R. Robertson, J.R. Rohrer, and P.G. Smith. Origins and evolution of subfam. maloideae (rosaceae). *Systematic botany*, pages 303–332, 1991.

[207] R.C. Evans and C.S. Campbell. The origin of the apple subfamily (maloideae; rosaceae) is clarified by dna sequence data from duplicated gbssi genes. *American journal of botany*, 89(9):1478–1484, 2002.

[208] C. Maliepaard, FH Alston, G. Van Arkel, LM Brown, E. Chevreau, F. Dunemann, KM Evans, S. Gardiner, P. Guilford, AW Van Heusden, et al. Aligning male and female linkage maps of apple (malus pumila mill.) using multi-allelic markers. *TAG Theoretical and Applied Genetics*, 97(1):60–73, 1998.

[209] J.M. Celton, DS Tustin, D. Chagné, and SE Gardiner. Construction of a dense genetic linkage map for apple rootstocks using ssrs developed from malus ests and pyrus genomic sequences. *Tree Genetics & Genomes*, 5(1): 93–107, 2009.

[210] J.A. Wolfe and W. Wehr. Rosaceous chamaebatiaria-like foliage from the paleogene of western north america. *Aliso*, 12, 1988.

[211] J. Salse, S. Bolot, M. Throude, V. Jouffe, B. Piegu, U.M. Quraishi, T. Calcagno, R. Cooke, M. Delseny, and C. Feuillet. Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *The Plant Cell*, 20(1):11–24, 2008.

[212] J.J. Doyle, L.E. Flagel, A.H. Paterson, R.A. Rapp, D.E. Soltis, P.S. Soltis, and J.F. Wendel. Evolutionary genetics of genome merger and doubling in plants. *Annual Review of Genetics*, 42:443–461, 2008.

[213] L.H. Rieseberg, B. Sinervo, C.R. Linder, M.C. Ungerer, and D.M. Arias. Role of gene interactions in hybrid speciation: evidence from ancient and experimental hybrids. *Science*, 272(5262):741, 1996.

[214] H. Wang, M.J. Moore, P.S. Soltis, C.D. Bell, S.F. Brockington, R. Alexandre, C.C. Davis, M. Latvis, S.R. Manchester, and D.E. Soltis. Rosid radiation and the rapid rise of angiosperm-dominated forests. *Proceedings of the National Academy of Sciences*, 106(10):3853, 2009.

[215] M. Lavin, P.S. Herendeen, and M.F. Wojciechowski. Evolutionary rates analysis of leguminosae implicates a rapid diversification of lineages during the tertiary. *Systematic Biology*, 54(4):575–594, 2005.

[216] BE Pfeil, JA Schlueter, RC Shoemaker, and JJ Doyle. Placing paleopolyploidy in relation to taxon divergence: a phylogenetic analysis in legumes using 39 gene families. *Systematic biology*, 54(3):441–454, 2005.

[217] S.B. Cannon, D. Ilut, A.D. Farmer, S.L. Maki, G.D. May, S.R. Singer, and J.J. Doyle. Polyploidy did not predate the evolution of nodulation in all legumes. *PloS one*, 5(7):e11630, 2010.

[218] D.E. Soltis, P.S. Soltis, D.R. Morgan, S.M. Swensen, B.C. Mullin, J.M. Dowd, and P.G. Martin. Chloroplast gene sequence data suggest a single origin of the predisposition for symbiotic nitrogen fixation in angiosperms. *Proceedings of the National Academy of Sciences*, 92(7):2647, 1995.

[219] J.J. Doyle and M.A. Luckow. The rest of the iceberg. legume diversity and evolution in a phylogenetic context. *Plant Physiology*, 131(3):900–910, 2003.

[220] M. Freeling and B.C. Thomas. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome research*, 16(7):805–814, 2006.

[221] G.E.D. Oldroyd and J.A. Downie. Coordinating nodule morphogenesis with rhizobial infection in legumes. *Annu. Rev. Plant Biol.*, 59:519–546, 2008.

[222] J.F. Arrighi, A. Barre, B. Ben Amor, A. Bersoult, L.C. Soriano, R. Mirabella, F. de Carvalho-Niebel, E.P. Journet, M. Ghérardi, T. Huguet, et al. The medicago truncatula lysine motif-receptor-like kinase gene family includes nfp and new nodule-expressed genes. *Plant physiology*, 142(1):265, 2006.

[223] P.H. Middleton, J. Jakab, R.V. Penmetsa, C.G. Starker, J. Doll, P. Kaló, R. Prabhu, J.F. Marsh, R.M. Mitra, A. Kereszt, et al. An erf transcription factor in medicago truncatula that is essential for nod factor signal transduction. *The Plant Cell Online*, 19(4):1221–1234, 2007.

[224] R. Op den Camp, A. Streng, S. De Mita, Q. Cao, E. Polone, W. Liu, J.S.S. Ammiraju, D. Kudrna, R. Wing, A. Untergasser, et al. Lysm-type mycorrhizal receptor recruited for rhizobium symbiosis in nonlegume parasponia. *Science*, 331(6019):909, 2011.

[225] R. Mittler, S. Vanderauwera, N. Suzuki, G. Miller, V. Tognetti, K. Vandepoele, M. Gollery, V. Shulaev, and F. Van Breusegem. Ros signaling: the new wave? *Trends in Plant Science*, 16(6):300, 2011.

[226] S.L. Liu and K.L. Adams. Dramatic change in function and expression pattern of a gene duplicated by polyploidy created a paternal effect gene in the brassicaceae. *Molecular biology and evolution*, 27(12):2817, 2010.

[227] S.L. Liu, G.J. Baute, and K.L. Adams. Organ and cell type-specific complementary expression patterns and regulatory neofunctionalization between duplicated genes in arabidopsis thaliana. *Genome Biology and Evolution*, 2011.

[228] E. Babiychuk, K. Vandepoele, J. Wissing, M. Garcia-Diaz, R. De Rycke, H. Akbari, J. Joubès, T. Beeckman, L. Jänsch, M. Frentzen, et al. Plastid gene expression and plant development require a plastidic protein of the mitochondrial transcription termination factor family. *Proceedings of the National Academy of Sciences*, 108(16):6674, 2011.

[229] 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.

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

[231] S.T. Bennett, C. Barnes, A. Cox, L. Davies, and C. Brown. Toward the $1000 human genome. *Pharmacogenomics*, 6(4):373–382, 2005.

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

[233] 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.

[234] 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.

[235] R. Lister, B.D. Gregory, and J.R. Ecker. Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. *Current opinion in plant biology*, 12(2):107–118, 2009.

[236] S. J. Cokus, S. Feng, X. Zhang, Z. Chen, B. Merriman, C. D. Haudenschild, S. Pradhan, S. F. Nelson, M. Pellegrini, and S. E. Jacobsen. Shotgun bisulphite sequencing of the arabidopsis genome reveals dna methylation patterning. *Nature*, 452(7184):215–9, 2008.

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

[238] E.R. Mardis. Chip-seq: welcome to the new frontier. *Nature methods*, 4 (8):613–614, 2007.

[239] J. C. Vera, C. W. Wheat, H. W. Fescemyer, M. J. Frilander, D. L. Crawford, I. Hanski, and J. H. Marden. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Mol Ecol*, 17(7):1636–47, 2008.

[240] A. J. Matas, T. H. Yeats, G. J. Buda, Y. Zheng, S. Chatterjee, T. Tohge, L. Ponnala, A. Adato, A. Aharoni, R. Stark, A. R. Fernie, Z. Fei, J. J. Giovannoni, and J. K. Rose. Tissue- and cell-type specific transcriptome profiling of expanding tomato fruit provides insights into metabolic and regulatory specialization and cuticle formation. *Plant Cell*, 23(11):3893–910, 2011.

[241] T. Woyke, G. Xie, A. Copeland, J. M. Gonzalez, C. Han, H. Kiss, J. H. Saw, P. Senin, C. Yang, S. Chatterji, J. F. Cheng, J. A. Eisen, M. E. Sieracki, and R. Stepanauskas. Assembling the marine metagenome, one cell at a time. *PLoS One*, 4(4):e5299, 2009.

[242] T. Kalisky and S. R. Quake. Single-cell genomics. *Nat Methods*, 8(4): 311–4, 2011.

[243] 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.

[244] 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.

[245] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev. Full-length transcriptome assembly from rna-seq data without a reference genome. *Nat Biotechnol*, 29(7):644–52, 2011.

[246] T. Schiex, J. Gouzy, A. Moisan, and Y. de Oliveira. Framed: A flexible program for quality check and gene prediction in prokaryotic genomes and noisy matured eukaryotic sequences. *Nucleic Acids Res*, 31(13):3738–41, 2003.

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

[248] Y. Pollack, R. Stein, A. Razin, and H. Cedar. Methylation of foreign dna sequences in eukaryotic cells. *Proc Natl Acad Sci U S A*, 77(11):6463–7, 1980.

[249] B. F. Vanyushin and V. V. Ashapkin. Dna methylation in higher plants: past, present and future. *Biochim Biophys Acta*, 1809(8):360–8, 2011.

[250] G. Makarevich, C. B. Villar, A. Erilova, and C. Kohler. Mechanism of pheres1 imprinting in arabidopsis. *J Cell Sci*, 121(Pt 6):906–12, 2008.

[251] A. M. Hancock, B. Brachi, N. Faure, M. W. Horton, L. B. Jarymowycz, F. G. Sperone, C. Toomajian, F. Roux, and J. Bergelson. Adaptation to climate across the arabidopsis thaliana genome. *Science*, 334(6052):83–6, 2011.

[252] A. Fournier-Level, A. Korte, M. D. Cooper, M. Nordborg, J. Schmitt, and A. M. Wilczek. A map of local adaptation in arabidopsis thaliana. *Science*, 334(6052):86–9, 2011.

[253] 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.

[254] J.A. Banks, T. Nishiyama, M. Hasebe, J.L. Bowman, M. Gribskov, C. dePamphilis, V.A. Albert, N. Aono, T. Aoyama, B.A. Ambrose, et al. The selaginella genome identifies genetic changes associated with the evolution of vascular plants. *science*, 332(6032):960, 2011.

[255] S.A. Goff, M. Vaughn, S. McKay, E. Lyons, A.E. Stapleton, D. Gessler, N. Matasci, L. Wang, M. Hanlon, A. Lenards, et al. Frontiers: The iplant collaborative: Cyberinfrastructure for plant biology. *Frontiers in Plant Genetics And Genomics*, 2, 2011.

[256] J. Qin, R. Li, J. Raes, M. Arumugam, K.S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65, 2010.

[257] M. Arumugam, J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D.R. Mende, G.R. Fernandes, J. Tap, T. Bruls, J.M. Batto, et al. Enterotypes of the human gut microbiome. *Nature*, 473(7346):174–180, 2011.

[258] R. Kwok. Five hard truths for synthetic biology. *Nature*, 463(7279):288–90, 2010.

[259] D. G. Gibson, J. I. Glass, C. Lartigue, V. N. Noskov, R. Y. Chuang, M. A. Algire, G. A. Benders, M. G. Montague, L. Ma, M. M. Moodie, C. Merryman, S. Vashee, R. Krishnakumar, N. Assad-Garcia, C. Andrews-Pfannkoch, E. A. Denisova, L. Young, Z. Q. Qi, T. H. Segall-Shapiro, C. H. Calvey, P. P. Parmar, 3rd Hutchison, C. A., H. O. Smith, and J. C. Venter.

Creation of a bacterial cell controlled by a chemically synthesized genome. *Science*, 329(5987):52–6, 2010.