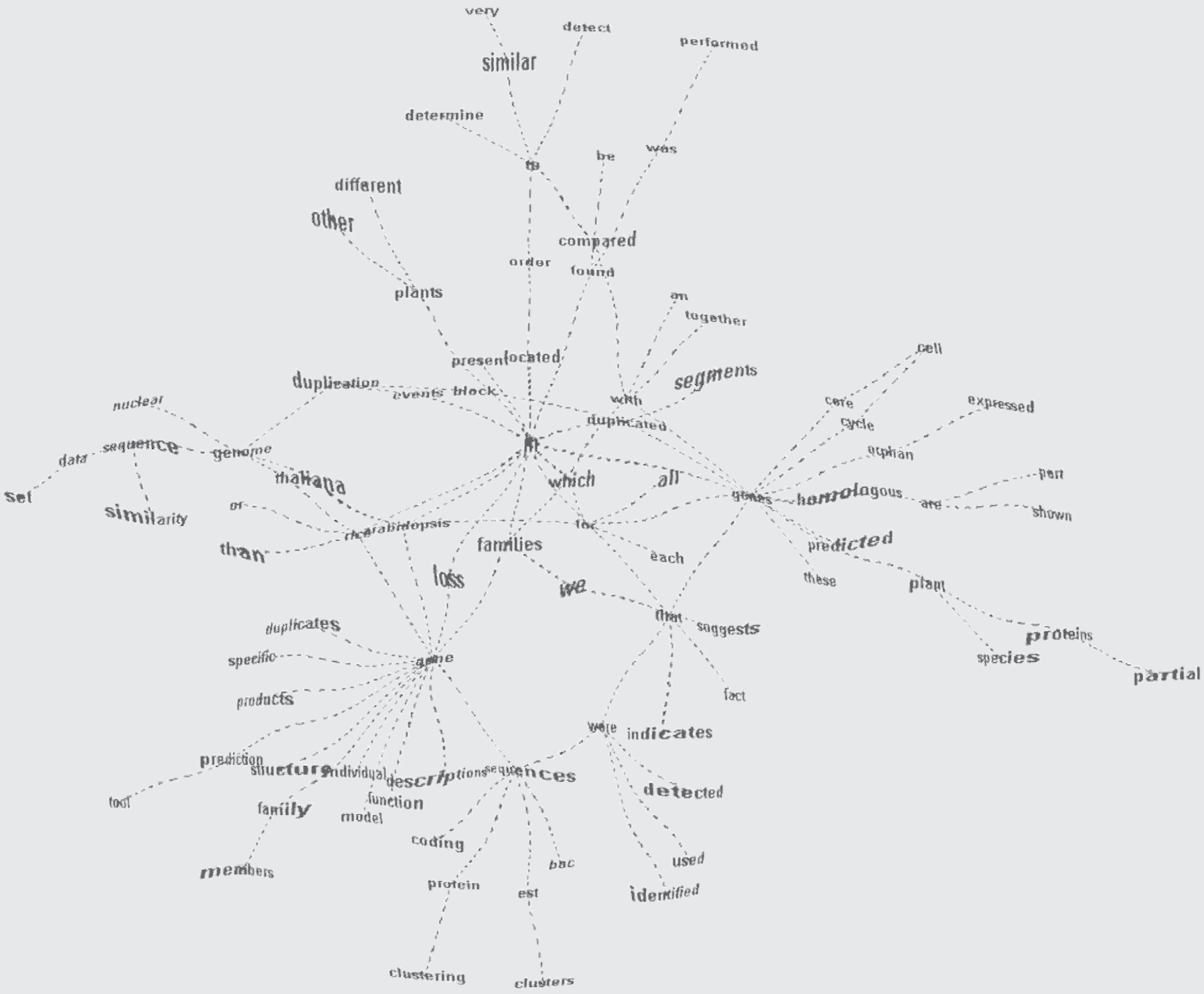
Ghent University
Faculty of Science
Department of Molecular Genetics
Bioinformatics and Evolutionary Genomics division

Mode and tempo of gene and genome evolution in plants



Klaas Vandepoele

Dissertation submitted in fulfillment of the requirements for the degree of Doctor (Ph.D.) of Science: Biotechnology

Promotor: Prof. Dr. Y. Van de Peer
March 2005

About the cover: *The network on the cover shows a basic linguistic summary of this thesis. The frequency of all word pairs was determined and the 200 pairs with the highest occurrence were selected. Subsequently, the individual words were represented as nodes in a graph, whereas the lines connecting the words indicate all different word pairs initially selected.*

Examination committee

Prof. Dr. Erik Remaut (chairman) [1]

Prof. Dr. Yves Van de Peer (promotor) [1]

Dr. Klaus Mayer [2]

Dr. Luc Krols [3]

Prof. Dr. Dirk Inzé [1]

Prof. Dr. Marc Zabeau [1]

Prof. Dr. Ann Depicker [1]

Pierre Rouzé [4]

Dr. Lieven De Veylder [1]

1 Faculty of Science, Ghent University
2 MIPS, Germany
3 Peakadilly nv
4 INRA - Ghent University

Acknowledgements

I can believe anything, provided that it is quite incredible.

Oscar Wilde, 1891

Desondanks het feit dat dit citaat de wereld een beetje op zijn kop zet, beschrijft het behoorlijk goed de uitzonderlijke momenten van wetenschappelijke extase gedurende mijn doctoraat. Momenten, die we moeten koesteren.

Alhoewel het starten van een doctoraat eigenlijk nog het best te vergelijken is met een grote sprong in het onbekende, kwam ik al vrij vlug tot de geruststellende conclusie dat deze wereld mij wel beviel. Daarom wens ik dan ook in eerste instantie Prof. Marc Zabeau en Prof. Dirk Inzé te bedanken, die het voor mij mogelijk maakten om een doctoraat in dit departement te starten en mij doorheen de eerste woelige IWT jaren loodsten.

Yves, zonder jou zou deze thesis er helemaal anders (lees: niet zo lijvig) hebben uitgezien. Ik moet je dan ook uitvoerig bedanken voor de uitstekende begeleiding die je me al die jaren hebt gegeven en voor je aanstekelijk enthousiasme. Desondanks je helse werkritme (waarvan iedereen, inclusief mezelf, zich nog steeds afvragen hoe je dat klaarspeelt), stond je altijd klaar om (de overvloed aan) nieuwe ideeën te aanhoren, en dat heb ik dan ook altijd ten zeerste geapprecieerd. Ook Cedric, met wie ik vele fantastische momenten tijdens (“ghost duplications!?”) en na de werkuren (“Rosie, these customers are waiting! [Union Square, NYC]”) heb gedeeld, wil ik bedanken voor de uiterst productieve samenwerking.

Of course a big thanks to all my great colleagues who assisted me through this scientific adventure. You are all amazing people and every one of you is indispensable for the great atmosphere in the team! Some special words of gratitude for Pierre Rouzé (The person who introduced me to the world of bioinformatics), Stephane (What would we be without structural annotation?), Jeroen (Evo=Tree-vo?), Yvan (Never doubt the pseudo-distance!), Wouter De Vos (fishes need new habitats sometimes...) and my thesis students Steven and Cindy. For all other people who helped me in any kind of way (papers, administration & amusement): thank you very much.

Uiterst belangrijk, maar ó zo zelden in de *spot-light*, zijn natuurlijk mijn vrienden en familieleden, die mij altijd hebben gesteund tijdens mijn periode aan *den unief*. Vooral mijn ouders, grootouders en (bijna) schoonouders wil ik speciaal bedanken: jullie zijn allemaal onmisbaar!

Natuurlijk moet ik het schoonste volk voor laatst houden: bedankt Stefanie, voor je steun en je lach, voor je begrip en je liefde.

Klaas

Table of contents

Section I

Introduction

The structure and evolution of plant nuclear genomes

Multiple layers of information are embedded in the nucleoprotein structure of chromosomes. The information content of the duplex DNA molecule within each chromosome contains signals for nucleosome positioning, transcription, gene splicing and amino acid selection. Apart from the diversity present in the construction and organization of DNA sequences in different species, molecular and evolutionary processes are continuously reshaping genome structures. At the macromolecular level, genomes primarily evolve through translocation, inversion, duplication, unequal recombination, deletion and substitution. Although the divergence of major angiosperms into monocotyledons and dicotyledons occurred some 130-200 million years ago (Yang et al., 1999; Wikström et al., 2001), there is considerable interest in the comparative analysis of plant genomes because of the expectation that information gained from one or more taxa may be extrapolated to a wide range of more complex or valuable crop genomes (Messing and Llaca, 1998). Therefore, basic knowledge about the constraints on genome structure and evolution is required, together with data describing the genetic and functional implications associated with any kind of modification.

The composition of angiosperm genomes

Apart from differences in chromosome number, size is the most basic feature that can be compared between nuclear genomes. Plants vary tremendously in genome size, from the 125Mb of *Arabidopsis thaliana* to some lily genomes (*Lilium*) that are about 1000-fold larger. Many crops, including cereals and legumes, possess large genomes although there is considerable variation in genome size within these plant families (Figure I). The lack of correlation between organism complexity and genome size, the ‘C-value paradox’ (Callan, 1972), has been debated for many years. The current view is that some of this variation is caused by differences in ploidy levels, although the major differences can be attributed to higher amounts of mobile and tandem repetitive elements (Doolittle and Sapienza, 1980; Orgel and Crick, 1980; Cavalier-Smith, 1985). All plant genomes appear to have many different types of transposable elements, but larger genomes seem to accumulate some subsets of these elements at very high copy numbers (Grandbastien, 1992; SanMiguel and Bennetzen, 1998). Whether small plant genomes have less of these elements because they are better able to inhibit their amplification or because they have some unknown mechanism for removal of these repeats, is unclear (Bennetzen and Kellogg, 1997). Nevertheless, accumulating evidence suggests that genome size contraction, through a diversity of DNA deletion mechanisms, may also be a common evolutionary process in eukaryotic genomes (Petrov et al., 1996; Petrov, 1997; Kirik et al., 2000; Shirasu et al., 2000). Therefore, it is possible that a bidirectional model combining DNA content increase and decrease operates on a more extensive scale in plants than previously thought (Bennetzen and Kellogg, 1997; Wendel et al., 2002).

In general, plant genomes appear to comprise a mosaic of different amounts of genic and non-gene-coding DNA. Euchromatin tends to be transcriptionally competent, while in heterochromatin transcription is predominantly repressed or inactive. Furthermore, there seem to be different constraints on the evolution of repetitive DNA and genes, which causes that the amount of repetitive DNA varies significantly between different plant genomes (see Table I). Like all other eukaryotic species, standard plant chromosomes contain, apart from genes, mobile repetitive DNAs and various classes of tandemly repeated sequences. A majority of tandem sequences are essential for the survival of the organism, because they are required for the organization and functioning of centromeres and telomeres. Other types of repeats, like minisatellites, microsatellites and transposable elements may represent selfish DNA, although low-copy number transposons that integrate near genes can serve as the raw material for the evolution of new *cis*-regulatory elements (White et al., 1994). Recently, additional roles for transposable elements and repeats in gene and genome evolution have been described (Devos et al., 2002; Jiang et al., 2004), confirming their importance for the evolution of plant genome structures.

Repetitive sequences and transposons

Plant centromeres are required for correct chromosomal segregation in mitosis and meiosis. Although centromeric chromatin is highly condensed, it can comprise more than 50% of an entire chromosome. *In situ* hybridizations and sequence analysis have identified sequences that are tandemly repeated in all cereal centromeres and thus might be required for correct centromeric function (Jiang et

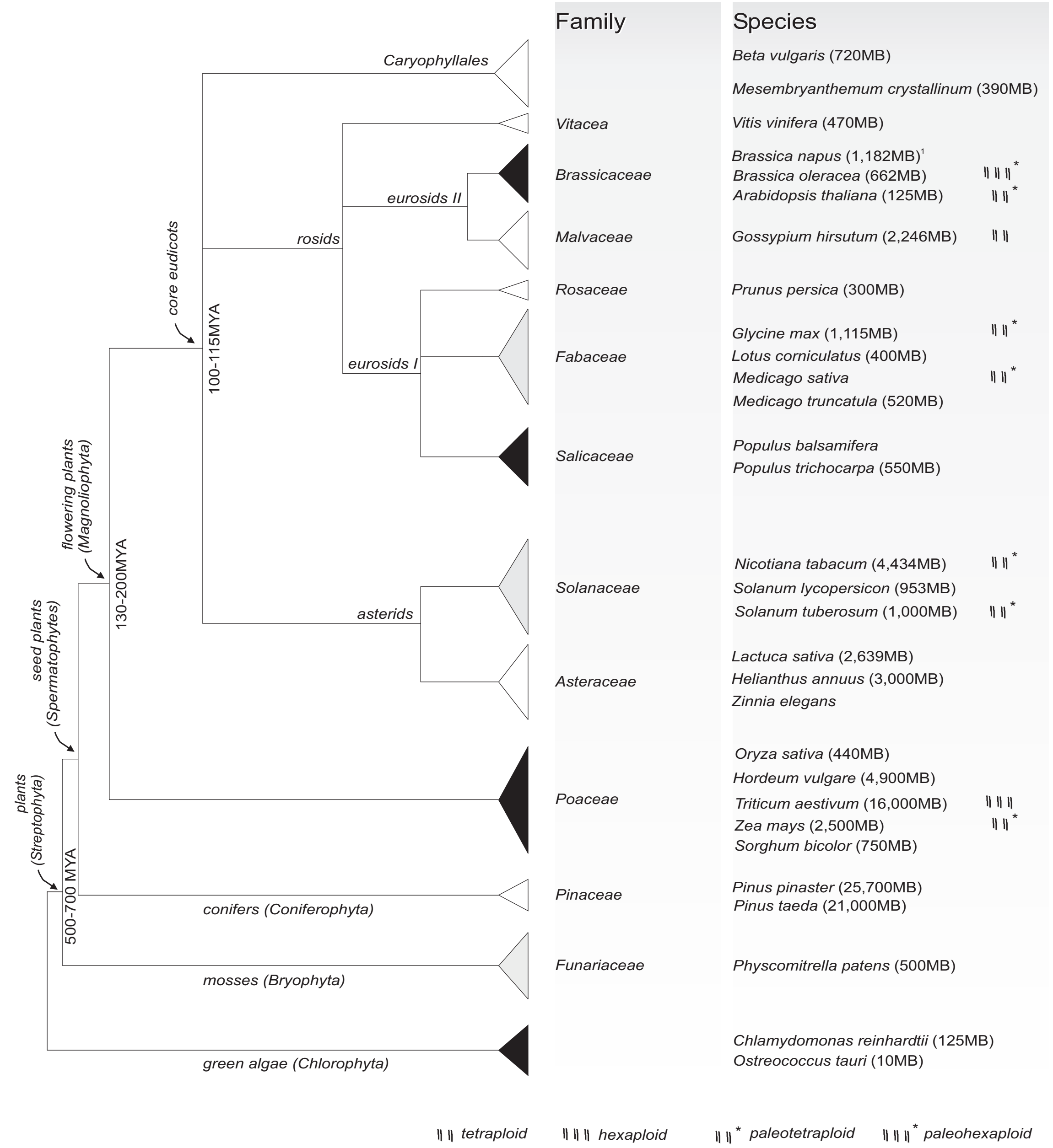

Family
Species
Caryophyllales
Beta vulgaris (720MB)
Mesembryanthemum crystallinum (390MB)
Vitacea
Vitis vinifera (470MB)
Brassicaceae
Brassica napus (1,182MB)1
Brassica oleracea (662MB)
Arabidopsis thaliana (125MB)
eurosids II
rosids
Malvaceae
Gossypium hirsutum (2,246MB)
Rosaceae
Prunus persica (300MB)
core eudicots
100-115MYA
eurosids I
Fabaceae
Glycine max (1,115MB)
Lotus corniculatus (400MB)
Medicago sativa
Medicago truncatula (520MB)
Salicaceae
Populus balsamifera
Populus trichocarpa (550MB)
flowering plants (Magnoliophyta)
130-200MYA
seed plants (Spermatophytes)
Solanaceae
Nicotiana tabacum (4,434MB)
Solanum lycopersicon (953MB)
Solanum tuberosum (1,000MB)
asterids
Asteraceae
Lactuca sativa (2,639MB)
Helianthus annuus (3,000MB)
Zinnia elegans
Poaceae
Oryza sativa (440MB)
Hordeum vulgare (4,900MB)
Triticum aestivum (16,000MB)
Zea mays (2,500MB)
Sorghum bicolor (750MB)
plants (Streptophyta)
500-700 MYA
Pinaceae
conifers (Coniferophyta)
Pinus pinaster (25,700MB)
Pinus taeda (21,000MB)
Funariaceae
mosses (Bryophyta)
Physcomitrella patens (500MB)
green algae (Chlorophyta)
Chlamydomonas reinhardtii (125MB)
Ostreococcus tauri (10MB)
tetraploid
hexaploid
* paleotetraploid
* paleohexaploid

Figure I Phylogenetic relationships among major lineages of green plants for which substantial genomic or EST data is available. Black triangles indicate plant families in which the nuclear genome of a model plant species is fully or nearly fully sequenced, whereas grey triangles indicate families in which genome initiatives have been started (e.g. *P. patens*, *S. lycopersicon* and *M. truncatula*). ^{1}*B. napus* is an amphidiploid species composed of homoeologous A and C genomes which are thought to have derived from the recent progenitors of extant *B. rapa* and *B. oleracea*, respectively (U, 1935). Based on: Wendel, 2000; Paterson et al., 2000; Wikström et al., 2001 and Heckman et al., 2001.

al., 1996). This is confirmed by the observation that all standard centromeres share several features, including tandem repeats of approximately 180bp, together with a highly heterochromatic state (Arabidopsis Genome Initiative, 2000). Like centromeres, telomeres also contain short tandem repeats, which are located at the termini of the linear plant chromosomes. Monomeric minisatellite repeats of 180-220bp are commonly present in thousands of tandem copies, where they form a large and fairly homogeneous knob of heterochromatin. Although these knobs are shared by all seed plant genomes, their sizes and locations show extreme interspecies variation (Bennetzen, 1998). Other types of tandemly repeated sequences like microsatellites or simple sequence repeats are hypervariable and scattered throughout genomes. Although they are found on all chromosomes in large numbers, their small size indicates that they only cover a small fraction of a total plant genome (Table I).

All transposable elements share two basic characteristics. The first is the ability to move from place to place in the genome and the second is their ability to amplify their copy number within the genome through transposition (Kumar and Bennetzen, 1999; Bennetzen, 2000a). Mobile elements fall into two major categories: those that transpose as DNA molecules (Class II) and those that transpose through an RNA intermediate (Class I). Well-studied elements like *Ac* and *En/Spm* are class II elements and comprise at most a few percent of any plant genome (Bennetzen, 1998). Whereas the copy number of the active element encoding the mobilizing transposase (e.g. *Ac*) is usually low (<5 copies), a few hundred copies of the defective element (e.g. *Ds*) that responds to the transposase can be found. One exception are the miniature inverted repeat transposable elements (MITEs), sometimes called Class III transposons, which are derived from DNA transposons. They can be present in thousands of copies per genome and are mainly found in or near genes or putative matrix attachment regions (Bureau and Wessler, 1995; Wessler et al., 1995; Avramova et al., 1998). However, because of their general small size (~200 bp), MITEs represent only a small part of plant genomes despite their often high copy number (Table I). In *Arabidopsis*, other class II transposons like CACTA elements and mutator-like elements (MULEs) are clustered near centromeres and heterochromatic knobs (Arabidopsis Genome Initiative, 2000).

Class I elements that move through an RNA intermediate are called retrotransposons. Since these elements use a *copy-and-paste* mechanism in contrast to the class II elements, which jump through a *cut-and-paste* mechanism,

Table I Summary statistics on nuclear plant genomes

Feature		*Arabidopsis thaliana*	*Oryza sativa*	*Zea mays*
Genome size	(MB)	125	430	2,500
%GC		36	43	47
Number of genes		25,498	46,022-55,615	50,000-70,000
Gene density	(1 gene per x kb)	4.5	9	40-45
Average gene length	(kb)	2.0	4.5	-
Average exon length	(bp)	214	201	-
Average intron length	(bp)	164	356	-
Simple sequence repeats		-	1.7%	1.1%
Transposable elements		10%	25%	73%
	Class I	1,385	(41%)	(92%)
	Class II	1,209	(18%)	(0.9%)
	MITEs	818	(40%)	(0.06%)

Based on: Gaut and Doebley, 1997; Bennetzen, 1998; Arabidopsis Genome Inititative, 2000; Meyers et al., 2001; Goff et al., 2002; Yu et al., 2002 and Whitelaw et al., 2003. Numbers in parenthesis indicate the percentage by length for different classes of transposable elements.

they make up the majority of the DNA in the nuclear genomes of large-genome plants like barley, lily and maize. Retrotransposons are the most abundant and widespread class of eukaryotic transposable elements, consisting of the long terminal repeat (LTR) and the non-LTR retrotransposons (Kumar and Bennetzen, 1999). LTR retrotransposons are further classified into the Ty1-*copia* and Ty3-*gypsy* groups that differ from each other in both their degree of sequence similarity and their order of encoded gene products. LTR-retrotransposons vary in size from several hundred bases to over 10 kb, with LTRs that are usually a few hundred bases to several thousand bases in length, and make up over 70% of the nuclear DNA in maize (SanMiguel and Bennetzen, 1998). The non-LTR retrotransposons, LINEs (long interspersed repetitive elements; 1-8kb) and SINEs (short interspersed repetitive elements; 100-300bp) can also be found in high copy numbers (up to 250,000) in different plant species. In many cases, Ty1-copia, Ty3-gypsy, LINE and SINE retrotransposons are dispersed widely throughout plant chromosomes. However, detailed sequence analysis in *Arabidopsis* and maize suggests that retrotransposons are highly enriched near centromeres, and often arranged as in nested series between genes, suggesting a preference for insertion or retention within inactive and methylated regions (SanMiguel et al., 1996; Arabidopsis Genome Initiative, 2000).

Plant transposable elements have a range of activities, all of them associated with possible alterations in gene and/or genome structure and function. Chromosomal modifications (e.g. breakage, rearrangement), insertional mutation, altered gene regulation, gene creation, sequence deletion and amplification are all identified effects of the transpositional and/or recombinational potential of retrotransposons (Bennetzen, 2000b; Devos et al., 2002). In addition, transposable elements carry with them regulatory sequences that can alter the expression of adjacent loci. An insertion of such an element into a promoter of a gene can bring that gene’s regulation under the control of the transposable element (Martienssen et al., 1998). Finally, some transposable elements can amplify DNA sequences from other parts of the genome. The action of the reverse transcriptase complex from retroelements can potentially turn any RNA into a DNA that can be integrated into the genome. Hence, *trans*-acting retroelement activity can convert a mRNA into an intronless pseudogene (Doring and Starlinger, 1986). Other elements can also take up portions of other sequences (e.g. genes) within the elements themselves. Consequently, transposition then amplifies these acquired segments, along with the rest of the element, thereby leading to an increased amount of raw material that can serve for the creation of new genes (e.g. Jiang et al., 2004).

Gene distribution

The mosaic pattern of plant genomes was initially discovered through the presence of different compartments having a dissimilar GC content (Salinas et al., 1988). One of the features distinguishing monocot and dicot genomes is the contrast of GC and dinucleotide content in exon and intron sequences (White et al., 1992; Carels et al., 1998; Yu et al., 2002). Furthermore, monocots in general have a higher GC content compared to dicots (Table I). Preliminary data also suggested that cereal genomes might display some features of compartimentalization, with gene-rich and gene-poor regions characterized by different GC composition (Barakat et al., 1997). This gene-cluster model was later confirmed in maize, where experimentally determined distances revealed a dense packing of genes in islands, separated by long stretches of apparently non-genic sequences (Panstruga et al., 1998; Tikhonov et al., 1999). Therefore, it seems that especially large plant genomes have managed to differentiate between desirable repeats (e.g. gene families) and potentially damaging repeats (e.g. transposable elements), and keep mobile and other repetitive DNA inactive through epigenetic control (see below).

Overall, plant genes are relatively compact, with average intron sizes of less than 200bp in *Arabidopsis* and less than 400bp in rice (Table I). Both in *Arabidopsis* and rice, the gene space occupies about 50% of the genome (Arabidopsis Genome Initiative, 2000; Yu et al., 2002). Within gene clusters, typically found in large plant genomes, the gene density approaches one gene per 5 kb, which is close to the average value of one gene per 4,5 kb for the sequenced portion of the *Arabidopsis* genome (Arabidopsis Genome Initiative, 2000). Upstream and downstream regulatory sequences are usually small as well, covering no more than a few hundred additional bases in most genes (Kaplinsky et al., 2002; Guo and Moose, 2003; Inada et al., 2003; Hong et al., 2003). Regulation at a distance, a feature commonly found in many animal genes, appears to be rare in plants, which makes that the average gene plus its regulatory components normally occupies only 1 to 5kb of genomic space (Bennetzen, 2000b).

The evolution of genome organization: colinearity

In the late 1990s, comparative sequence analysis revealed that for large genomes containing enormous amounts of retrotransposon DNA, a similar structure of relative gene order conservation seemed to exist within a varying distribution of repeats (Feuillet and Keller, 1999). Later, more evidence for both conservation of synteny (conserved clustering of genes or markers) and colinearity (conserved content and order of genes or markers) over different levels of divergence amongst different plant species was found (e.g. Devos et al., 1999; van Dodeweerd, 1999; Grant et al., 2000; Ku et al., 2000). The overall extent of synteny and colinearity appears to be correlated with evolutionary distance, although rate differences in specific lineages have been reported (Gale and Devos, 1998a; Schmidt, 2002). A detailed sequence comparison of a 60 kb genomic region in *Arabidopsis* with its counterpart in the closely related *Capsella rubella* revealed complete conservation of gene content, order and transcriptional orientation (Acarkan et al., 2000). In contrast, differential divergence patterns in different regions of the genome were observed between *Arabidopsis* and *Brassica oleracea*, an ancient hexaploid (O'Neill and Bancroft, 2000). Comparing rosids and asterids, Ku et al. (2000) found a network of microsynteny between a genomic region of tomato and its multiple homologous segments in *Arabidopsis*. Similarly, a high degree of microsynteny between related grass species (e.g. rice, sorghum, maize), which diverged 50-70 million years ago, was found (Chen et al., 1998; Tikhonov et al., 1999). A synthesis of data generated by several comparative mapping studies demonstrated that all cereal species could be represented by a small number of linkage blocks (Gale and Devos, 1998b), indicating that the grasses could be studied as variants on a single experimental genome (Bennetzen and Freeling, 1993).

Although this finding led to the perception that the grasses, compared to related eudicot species, were exceptional in their degree of genome conservation, it became also clear that several studies might be biased towards promoting colinearity and ignoring exceptions, using the argument that paralogous instead of orthologous markers were mapped (Bennetzen, 2000b; Bennetzen and Ramalrishna, 2002). Consequently, the availability of methods based on robust statistical analysis for assessing colinearity between genomes became essential (King, 2002). In order to resolve complex colinear genome relationships, Gaut (2001) developed a statistical method to assign a statistical significance to the detection of colinearity, using a simple Monte-Carlo simulation. After re-analyzing the maize genome, he concluded that the homology in the maize genome is more complex than initially thought, based on comparative maps, and revealed that 80% of the genome is duplicated, confirming that maize is an ancient tetraploid (Gaut and Doebley, 1997).

The prevalence of gene duplication

Simultaneously with comparative mapping experiments and genome sequence comparisons between different related plant species, detailed sequence analysis on large *Arabidopsis* contigs provided evidence for ancient large-scale duplication events (Lin et al., 1999; Terryn et al., 1999). Although unexpected for the small-sized “innocent” diploid *Arabidopsis thaliana*, this finding confirmed the role of genome doubling in the evolutionary history of flowering plants (Stebbins, 1971). As reported by Wendel (2000), it is difficult to overstate the importance of genome doubling in plants, since 50-70% of all angiosperms have experienced one or more episodes of polyploidy at some point in their evolutionary history (Figure I). In addition, very ancient doubling events may be difficult to discern due to potentially rapid evolutionary restoration of diploid-like chromosomal behaviour. Therefore, many angiosperms are considered to have paleopolyploid genomes. Logically, the significance of polyploidy in flowering plants implies that it also has some adaptive significance. Although novel phenotypes in polyploids, such as increased organ size and biomass, drought tolerance, pest resistance and asexual seed production have been described extensively (Levin, 1983), pioneering studies in the early 70’s revealed that chromosome doubling by itself is not a help but a hindrance to the evolutionary success of higher plants (Stebbins, 1971). Therefore, it was assumed that the success of polyploidy in nature must have been accompanied by other genetic-evolutionary processes, which compensated for the initial disadvantages of raw polyploids.

Although the mechanisms by which polyploidy contributes to novel variation are not well understood, it is clear that the genomic redundancy may lead to novel functions through divergence of gene duplicates (Stephens, 1951; Ohno, 1970). In 1970, Ohno predicted that after duplication, one copy is released from functional constraints and through mutation either will decay (nonfunctionalization) or acquire a new function (neofunctionalization). This concept of “evolutionary opportunity through divergence after duplication” has become widely embraced, although there are relatively few examples that demonstrate convincingly divergence after duplication in plants (Adams et al., 2003). Although functional divergence is a potential consequence of gene duplication, it is clear that only few duplicates escape the accumulation of deleterious mutations, which leads to pseudogene formation and subsequent gene loss or silencing (Stephens 1951; Ferris and Whitt, 1977; Wagner, 1998). Although population genetics modelling studies revealed that the occurrence of gene loss might be an order of magnitude higher than that of functional divergence, the observations that rates of gene silencing are much lower than predicted, induced the potential significance of other possible fates of duplicated genes, such as long-term maintenance of similar if not identical function (Hughes and Hughes, 1993; Pickett and Meeks-Wagner, 1995) or sub-functionalization. The latter model, also known as the duplication-degeneration-complementation (DDC) model, predicts that degenerative mutations, apart from creating pseudogenes, may also preserve duplicated genes by changing or specializing their functions, either at the protein level or at the regulatory level (Force, 1999). Since more and more data provide evidence for different evolutionary outcomes after gene duplication, it seems that several models describing the fate of a gene duplicate are valid (Wendel, 2000).

Apart from modifications in the set of duplicated genes, many potentially important processes in polyploid genome evolution operate above the gene level. Although the current knowledge about these aspects is very rudimentary, it seems that a set of different processes collectively lead to genome stabilization and evolution in polyploids. Based on mapping data, more and more examples of ancient cryptic cycles of genome doubling and chromosomal diploidization in currently diploid plants are found (e.g. *Brassica*; Lagercrantz, 1998, *Glycine*; Shoemaker et al., 1996, *Gossypium*; Brubaker et al., 1999 and *Zea*; Helentjaris et al., 1988 [Figure I]). Interestingly, apart from the extensive colinearity and retention of synteny, chromosomal rearrangements such as inversions and translocations are commonly observed in these diploid plants (Moore et al., 1995). Therefore, it seems that recombination between homoeologous chromosomes (i.e. sister chromosomes created by polyploidy that are partially homologous) is responsible for the inter-genomic translocations in polyploidy and diploid lineages (Zohary and Feldman, 1962; Wendel, 2000). Analysis of synthetic *Brassica* allopolyploids, created by the hybridization of two differentiated genomes, revealed patterns of non-Mendelian genomic change and rapid sequence elimination, indicating the dynamic nature of ployploid genomes (Song et al., 1995; Soltis and Soltis, 1999).

Although the functional significance of sequence elimination is not fully understood, Feldman et al. (1997) noted that this process, converting sequences of initially homoeologous chromosomes into chromosome-specific sequences, might provide a physical basis for the rapid restoration of diploid-like chromosome pairing following polyploidization.

Finally, increased or altered levels of DNA methylation have been observed when monitoring the early stages after polyploidy (Song et al., 1995; Liu et al., 1998). Cytosine methylation in CpG dinucleotides and CpNpG trinucleotides is common in plants and plays a role in the regulation of gene expression and DNA replication (Finnegan et al., 1998). DNA methylation is also able to repress the activity of transposable elements (Yoder et al., 1997). Since during polyploidy two genomes are united into a single nucleus, this signal of foreign DNA might be responsible for altered patterns of epigenetic silencing, during which transposable elements are released from suppression (McClintock, 1984). Therefore, it is likely that the burst of genic and regulatory evolution through transposable element insertion is an important feature of the early stages of polyploid formation (Flavell, 1994). In addition, the epigenetic response associated with polyploidy formation may also lead to silencing (and reactivation) of gene duplicates (Comai et al., 2000) and increased mutation rates, which further enlarge the process of rapid genomic change (Wendel, 2000).

Conceptual framework

In the early days of the *Arabidopsis* genome sequence, some pioneering studies already described complex patterns of genome evolution in *Arabidopsis* (e.g. Lynch and Connery, 2000; Ku et al., 2000; Vision et al., 2000). However, it became clear that advanced and high-throughput tools were required in order to fully explore the gene content and structure of the first plant nuclear genome sequence. The first part of the results section will focus on the different methodologies that can be applied for the investigation of genome structure and evolution, and introduces some basic concepts and terminologies. In addition, a newly developed software tool for the detection of genomic homology, incorporating a robust statistical validation, is presented. The final chapter of part one illustrates the power of comparative interspecies strategies for the detection of ancient and heavily degraded genomic homology.

Part two presents several analyses that provide evidence for the importance of large-scale duplication events in the evolution of plant nuclear genomes. Chapter 2.1 describes a detailed analysis of a limited number of duplicated segments in the *Arabidopsis* genome, focusing on the origin of these duplicated blocks applying different dating strategies. Chapters 2.2 and 2.3 report on the occurrence of large-scale duplication events in the sequenced genomes of *Arabidopsis* and rice, and illustrate the significance of polyploidy in the evolution of angiosperm plants. Part two ends with chapter 2.4, which describes the evolutionary consequences of gene duplication in *Arabidopsis.* This study focuses on two major topics: microcolinear networks between different eurosid plant species and the analysis of *cis*-regulatory evolution in gene duplicates using a comparative approach.

Finally, part three discusses the annotation, delineation and organization of seven gene families controlling cell cycle regulation in the *Arabidopsis* genome, together with an interspecies comparison of gene families within the green plant lineage. Whereas the results of chapter 3.1 reveal the complexity and the large number of genes controlling the plant cell cycle machinery, chapter 3.2 illustrates the plasticity in gene content between 32 different plant taxa and discusses the major differences in copy number and gene family organization between *Arabidopsis* and rice.

Section II

Results

- Part 1 -

Developing automatic approaches for the detection of homologous chromosomal regions

1.1 The quest for genomic homology

Klaas Vandepoele, Cedric Simillion and Yves Van de Peer

Redrafted from: *Current Genomics* **5**, 299-308 (2004)

New initiatives to sequence complete genomes of related organisms have introduced a new era of large-scale evolutionary genomics. The comparative analysis of these genomes allows us to obtain a comprehensive view of many aspects of eukaryotic genome evolution. Consequently, new computational methods and approaches are being developed in order to investigate chromosomal organization, rearrangements and segmental homology. Here, we review the different techniques currently available to identify homologous chromosomal segments in closely and more distantly related species and highlight some of the difficulties inherent to the statistical validation of putative genomic homology. In addition, advantages of cross-species genome analysis are discussed as well as novel approaches to study large-scale gene duplications.

Introduction

Comparative genomics provides an efficient way to detect functional elements in genomic sequences. The observation that functional regions are conserved throughout evolution, in contrast to their non-functional counterparts, has triggered the sequencing of (at least parts of) genomes of closely related animals, plants and fungi (Hardison et al., 1997; Thacker et al., 1999; Waterson et al., 2002; Wortman et al., 2003, Cliften et al., 2001; Kellis et al., 2003). Such large-scale sequencing projects offer an integrated framework for comparative sequence analysis and greatly enlarge our knowledge about gene structure, function and regulation. Perhaps the most illustrative example is the sequencing of the mouse genome, which, in comparison with the human genome, has allowed the identification of many regulatory elements and has improved gene annotation in both human and mouse (Levy et al., 2001; Dermitzakis et al., 2002; Dieterich et al., 2002; Alexandersson et al., 2003; Pedersen and Hein, 2003; Flicek et al., 2003; Collins et al., 2003; Clamp et al., 2003). Moreover, the detection of signals that are conserved but cannot be recognized in the absence of a cross-species comparison makes it possible to discover new functional elements, such as non-coding RNAs (McCutcheon and Eddy, 2003; Lagos-Quintana et al., 2003), and hint to their importance in biological systems.

Apart from the improved detection of conserved elements and a better understanding of the complexity embedded in biological processes through the comparative analysis of the genes involved, the availability of an increasing amount of genome sequences from a large variety of organisms makes it possible to study the organization of genes in a genome. Especially the characterization of different types of rearrangements (e.g. inversions, translocations and transpositions), duplications and gene loss exposes the actual impact of genome evolution on the complete catalogue of genes encoded by the genome (Nadeau and Taylor, 1984; Seoighe et al., 2000; Bennetzen, 2000b; Ranz et al., 2001; Coghlan and Wolfe, 2002; Pevzner and Tesler, 2003). However, in order to study genome organization and genome evolution, it is essential that conserved regions between and within genomes can be correctly identified. Since these homologous regions, derived from a common ancestral region, may have been extensively rearranged, their identification is not always obvious. In this review, we discuss the different techniques currently applied for the detection of homologous chromosomal regions and their application to the analysis of large-scale duplication events. Furthermore, we highlight some of the advantages of having access to related genomes when unraveling a genome's evolutionary past.

The detection of homologous chromosomal segments

Choosing the best method for the detection of homology at the genomic level highly depends on the resolution one wants to obtain and on the nature of the genomic information that is available. If complete genomic sequences of closely related species are available, the most straightforward way to detect homology is by comparing the sequences at the DNA level using a standard sequence similarity search tool such as BLAST (Altschul et al., 1997) or FASTA (Pearson and Lipman, 1988). Similarly, a DNA-based sequence comparison can be applied to identify recently duplicated and thus paralogous chromosomal regions within the same genome. For the comparison of very long stretches of DNA, both pairwise alignment tools (e.g., Smith-Waterman (Smith and Waterman, 1981), DOTTER (Sonnhammer and Durbin, 1995), MUMmer (Delcher et al., 1999), PipMaker (Schwartz et al., 2000), SSAHA (Ning et al., 2001), BLAT (Kent, 2002), BLASTZ (Schwartz et al., 2003a), AVID (Bray et al., 2003), LAGAN (Brudno et al., 2003)) and multiple alignment tools (Multi-LAGAN (Brudno et al., 2003), MultiPipmaker (Schwartz et al., 2003b)) have been developed. If both input sequences are closely related, large-scale alignments can be generated, which show a detailed base-to-base mapping between the two genomic sequences. Although some of the programs listed above are able to cope with genomic sequences from more distantly related organisms, the increasing amount of sequence dissimilarity between such genomes, or alternatively between anciently duplicated regions within the same genome, seriously complicates the detection of significant homology over long genomic distances (e.g. 100-1,000kb). Rather, small conserved fragments, typically conserved exons or non-coding conserved sequences might be recovered, but these provide little overall information on the evolution of chromosomes or complete genomes.

When the amount of sequence similarity at the DNA level is too low to determine homology between or within genomes, the inference of conserved gene content and order (i.e. colinearity) provides an elegant alternative to unravel common ancestry of chromosomal regions. The advantage of this method, compared to DNA sequence alignment methods, is that similarity that has faded away at the DNA level still can be detected at the protein level. This is demonstrated in Figure 1.1.1 showing a comparison of two highly similar and two degenerated paralogous chromosomal regions in the genome of *Arabidopsis thaliana*, both at the DNA level and at the protein level. Where for recently duplicated (and thus highly similar) regions homology can still clearly be inferred by both methods (i.e. DNA-based alignments and colinearity at the protein level), the homology between the degenerated paralogous regions is only visible through the detection of colinearity at the protein level.

The map-based approach: detection of conserved content and order

The identification of homologous chromosomal regions between distantly related organisms is thus usually based on a genome-wide comparison that aims at delineating regions of conserved gene content and order in different parts of the genome. The same is true for the detection of duplicated chromosomal regions within the same genome. Although the map-based approach can be applied on the basis of different types of genomic information (e.g. genes, molecular markers or local DNA similarities, see further), we will explain the general concept of this method with genes as the genomic units of a chromosome. Essential in the map-based approach is that the (absolute or relative) chromosomal locations of all genes (or in general the units describing the chromosome under investigation) are known.

Although the detection of colinearity seems a fairly simple way to detect genomic homology, the dynamic nature of genomes, responsible for the duplication, deletion, and rearrangements of genomic DNA, results in a degraded pattern of colinearity that makes it difficult to detect more ancient homology. Nevertheless, the correct identification of homologous segments remains an important issue. Regarding large-scale gene duplication, several studies already applied a map-based approach for the detection of duplicated segments in fully sequenced genomes (Wolfe and Shields, 1997, McLysaght et al., 2002; Li et al., 2003; Blanc et al., 2003). Recently, we developed a publicly available software tool, called ADHoRe, for the automatic detection of homologous regions combined with a robust statistical validation (Vandepoele et al., 2002a). The general concept of ADHoRe makes it possible to use the software tool for the analysis within one genome, i.e. to look for paralogous regions with duplicated genes, or for comparisons between genomes of different organisms, i.e. to look for orthologous

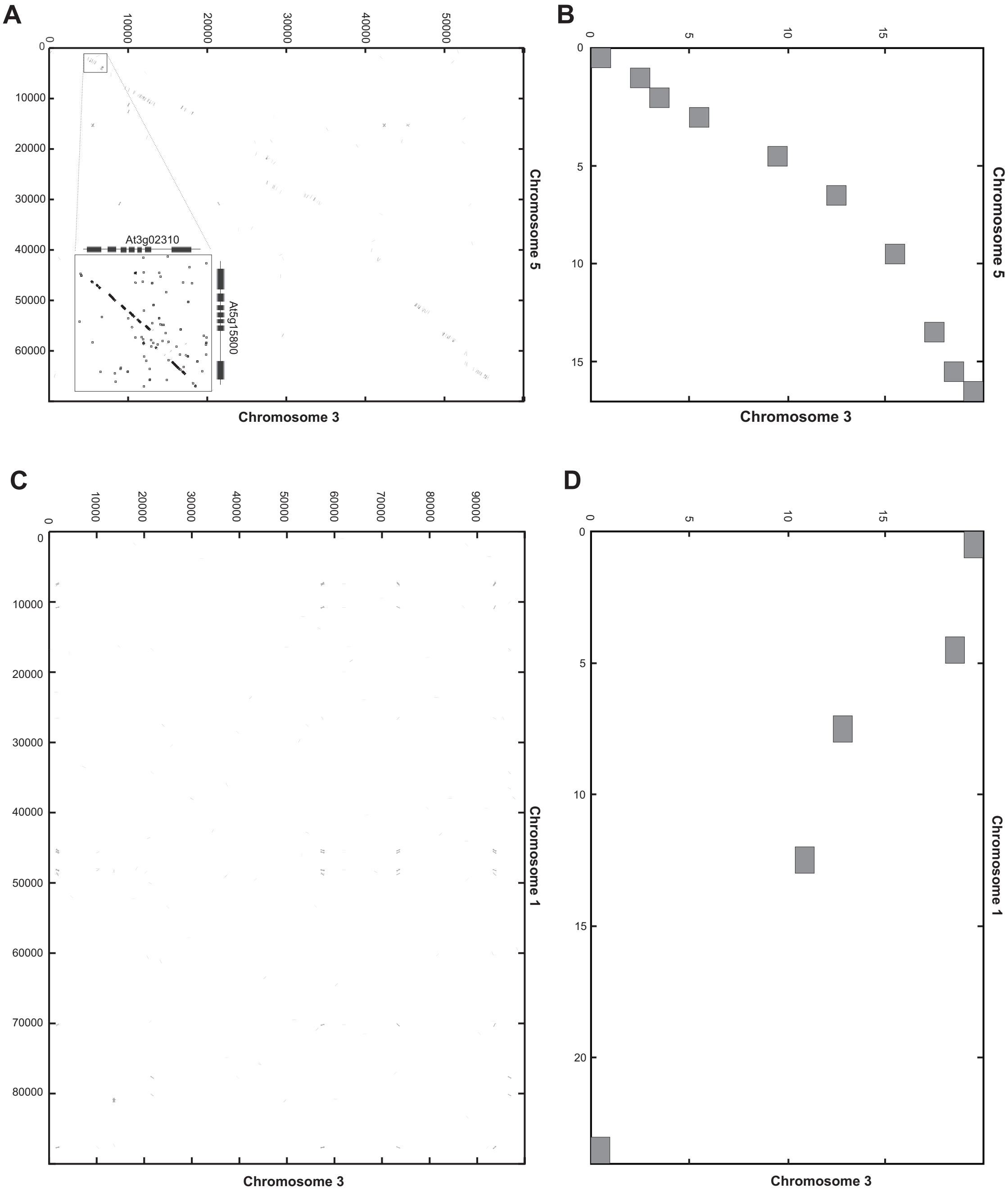


Figure 1.1.1 Comparison of duplicated regions in *Arabidopsis* through both DNA-based alignments and the detection of colinearity (conserved gene content and order). Panels A and B show a recently duplicated chromosomal segment between chromosome 3 (size 55,6 kb or 21 annotated genes) and chromosome 5 (size 65.5 kb or 20 annotated genes) that can be detected by DNA-based alignments and by colinearity at the protein level, respectively. DNA-based alignments were created using MuMmer (parameters: -l 15 –b –c; Delcher et al., 1999). The zoom-in, created with DOTTER (Sonnhammer and Durbin, 1995), shows the conserved exon-intron structure at the DNA level of a paralogous gene pair. Panels C and D show an ancient duplication event between chromosome 1 (size 89,7 kb or 26 annotated genes) and chromosome 3 (size 100,4 kb or 24 annotated genes). Whereas colinearity at the protein level enables the detection of this

anciently duplicated segment (D), no similarity at the DNA level can be found (C). Note that in panels A and C the axes of the graph represent the base pairs of the chromosome, where in panels B and D the graph represent genes positioned along the chromosome.

regions. Moreover, events such as inversions, deletions and tandem duplications that complicate the detection of homology, can be taken into account. Based on similar principles, Gaut and coworkers recently published the LineUp package that aims at detecting significant chromosomal homology based on molecular marker information, even if substantial rearrangements of marker order have occurred (Hampson et al., 2003).

The ADHoRe algorithm

In the map-based approach as implemented in ADHoRe, the information on homologous gene pairs is stored in a matrix of (m.n) elements (m and n being the total number of genes on each genomic fragment), each non-zero element (x, y) being a pair of homologous genes (x and y denote the coordinates of these homologous genes or anchor points). Figure 1.1.2a shows a small hypothetical gene homology matrix (GHM). In the matrix, colinear segments are represented as diagonal lines, while tandem duplications form horizontal or vertical lines, inversions can be detected by considering the orientation of the elements, and gaps in diagonal regions refer to gene loss or gene insertions in duplicated blocks. To detect colinearity, one has to find more or less diagonal series of elements (i.e. homologous genes) in the matrix. This way of presenting the organization of genes on genomic segments reduces finding colinearity to a clustering problem. During construction of the GHM, ADHoRe subjects it to a number of procedures. For example, after identification of the homologous genes, irrelevant data points need to be removed, a process we refer to as negative filtering. During this step, all elements that cannot belong to a cluster because they are too far away from other elements in the homology matrix – i.e. homologous genes that most probably have not been created by the block duplication - are removed. Also tandem duplications are removed from the matrix. Since we are looking for diagonal regions in the GHM, purely horizontal or vertical regions due to tandem duplications are remapped by collapsing all tandem duplications. This way it is easier to detect diagonal regions, since they are no longer interrupted by horizontal or vertical elements. The end result is a matrix that has been cleaned up by filtering and a colinear region is now defined in the matrix representation as a number of elements

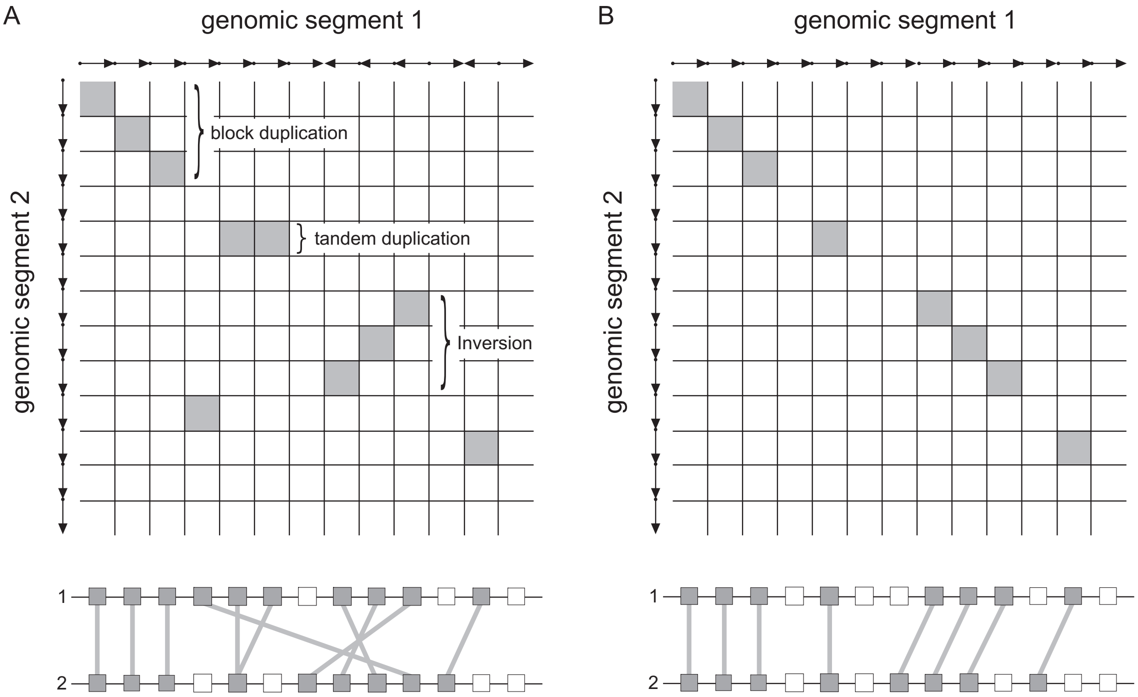


Figure 1.1.2 Hypothetical gene homology matrix. Arrows on the axes of both segments represent genes on the genomic segments. Grey cells illustrate homologous genes (anchor points). In panel A, the original organization of all genes, including tandem duplications and inversions, is shown. Panel B shows the same gene homology matrix after remapping of tandem duplications and the removal of irrelevant single data points, i.e., homologous genes that are most likely not part of the block duplication. In addition, the small inverted colinear segment of 3 anchor points was restored to its original orientation, in order to create a larger colinear region.

(which we refer to as anchor points) showing clear diagonal proximity (Figure 1.1.2b). In order to find such diagonals on a mathematical basis, we have developed a special distance function that yields a shorter distance for elements that are in diagonally close proximity than points that are in horizontal or vertical proximity (see chapter 1.2). Figure 1.1.3 shows the application of this distance function to a hypothetical example. Briefly, all elements in the GHM that are in close proximity are grouped into clusters. Subsequently, the quality of each cluster is examined and can be used to remove non-colinear homologous regions (see Figure 1.1.3). Finally, it is investigated whether detected clusters can be combined into larger homologous regions (see chapter 1.2).

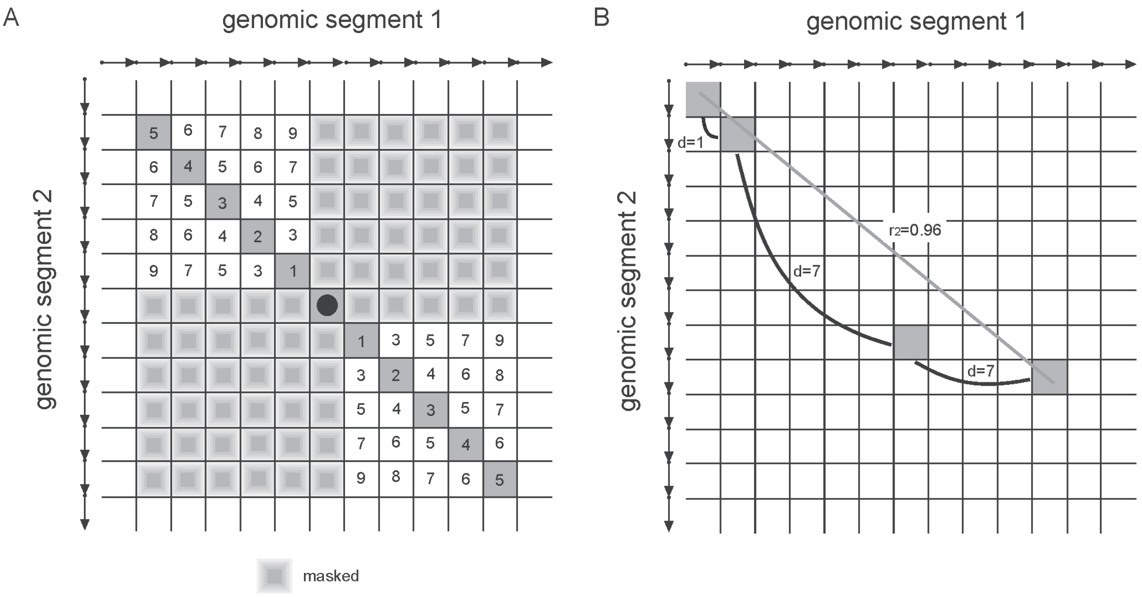


Figure 1.1.3 Application of the diagonal pseudo distance (DPD) function to the detection of elements with diagonal proximity in the gene homology matrix. Panel A shows the DPD for a given cell in the matrix to the central black dot (anchor point). The diagonal pseudo distance is smaller for diagonally orientated elements (grey boxes) than for elements deviating from the diagonal. Shaded boxes represent elements (genes) with an infinite distance to the central dot, since these elements are unlikely to be part of the duplicated segment that contains the anchor point (black dot). Panel B shows the iterative clustering of elements for a colinear region with positive orientation (i.e. from top left to down right) in the homology matrix. All genes lie within a maximum gap distance G (e.g. 30) of each other. The best-fit line and its coefficient of determination (r^2) shows the quality of the cluster, which is clearly above the predefined Q value cut-off, here set to 0.9. As a result, all four homologous genes are considered to have been arisen by a block duplication.

Statistical significance of colinearity

When all clusters (i.e. colinear regions) have been compiled as described above, colinear segments (or clusters in the homology matrix) that are not statistically significant need to be removed. The goal of this procedure is to determine which colinear regions could occur purely by chance and are therefore not biologically significant. This problem was first recognized by Gaut (2001) who introduced a statistical test to validate whether colinearity of genetic markers represented genuine homology or could be expected by chance. To this end, the number of anchor points (i.e. homologous genes) within a colinear segment together with the size of the segment was compared with colinear segments found in a large number of randomized data. If the original colinear segment contains more markers or markers in closer proximity than expected by chance, the conclusion is that both segments are indeed homologous. This is usually implemented as a statistical test (a so-called permutation test or Monte Carlo simulation), sampling a large number of reshuffled data sets and calculating the probability that a colinear region, characterized by a number of conserved genes and an average gap size, can be found by chance.

Several recently published analyses have applied statistical validation through comparisons of observed data with expected data obtained by randomization tests (Friedman and Hughes, 2001; Vision et al., 2000; McLysaght et al., 2002; Cavalcanti et al., 2003). Although frequently done, the selection of colinear segments based solely on the number of anchor points within a colinear region is not entirely correct. This is due to the fact that the significance of colinearity strongly depends on the overall distribution of the homologous genes in a colinear segment, rather than on the total number of homologous genes (see also Durand, 2002). One can easily imagine that the significance of 7 homologous genes within a colinear region of 15 genes is much higher than a colinear region of 100 genes with 7 homologous genes. Therefore, taking into account the number of anchor points in a cluster together with the average distance between all anchor points in a cluster (or reciprocal density) provides a more reliable way to calculate the probability that a cluster detected in the real dataset could have been generated by chance. This will result in small but dense clusters being retained, whereas loose small clusters will be rejected, since the chance that they were generated by chance is high.

A major drawback of the validation of colinearity through the comparison with randomized datasets is that the analysis of the large number of permutated datasets (typically 100 or 1,000) is computationally expensive and in many cases more time-consuming than analyzing the original dataset. Consequently, new methods have been developed for the validation of colinearity that do not require the presence of randomized data (Calabrese et al., 2003; Simillion et al., 2004).

Selection and identification of homologous genes

In order to identify statistically significant orthologous or paralogous colinear segments based on the gene homology matrix, it is important to use strict criteria before concluding whether two genes (or markers) are anchor points. In the case of genetic maps, information about similar units – applied for describing the chromosome - is derived from markers that cross-hybridize on different chromosomal locations, whereas in sequenced genomes anchor points are simply homologous DNA or protein sequences. In the map-based approach, usually lists of predicted genes resembling the order of the genes on the chromosome are used for comparing genomic segments. Recently, Pevzner and Tesler (2003) used local similarities at the DNA level to compare the genomes of human and mouse, bypassing problems due to possible erroneous gene annotation. Nevertheless, as discussed above, homology at evolutionary distances where only protein similarity is conserved is missed. A possible solution, not yet implemented as far as we know, would be to identify homology between two segments by combining local similarities both at the DNA level and protein level. This method would have the advantage that it offers higher resolution compared to using only protein sequences and consequently should provide a more accurate view of the actual similarities between genomic sequences, both in coding and non-coding regions.

A first crucial step in applying the map-based approach as described above is the identification of homologous genes. Usually, an all-against-all sequence similarity search (e.g. BLASTP; Altschul et al., 1997) is performed to find homologous proteins. Apart from applying an E-value or a similarity score cutoff, additional parameters such as the coverage of the alignable region on both potentially homologous genes can be applied to select ‘suitable’ homologs (for examples, see McLysaght et al., 2002; Li et al., 2003). A major problem in identifying homologous genes based on sequence similarity is the discrimination between paralogous and orthologous genes, especially if genes belong to large multigene families (Jensen, 2001). For example, finding colinearity considering gene families with only one member in each genome will provide strong evidence to define truly orthologous segments between distantly related genomes. In contrast, the inclusion of large gene families will introduce a large number of homologous anchor points in the GHM of which only a very small fraction represents genuine orthology. Therefore, prior to the construction of the GHM, one should consider to first define all gene families and their sizes using specifically designed cluster algorithms (Tatusov et al., 2000; Li et al., 2001; Remm et al., 2001; Enright et al., 2002). In order to reduce the noise created by paralogy, only small gene families could then be selected and included in the analysis.

Large-scale gene duplication events and gene loss

Often, very degenerated block duplications that originated hundreds of millions of years ago cannot be identified as such by directly comparing the duplicated segments. Differential gene loss, which is responsible for the loss of a different but complementary set of genes on both paralogous genomic segments, makes it impossible to detect significant colinearity by directly comparing anciently duplicated regions. Therefore, two genomic segments in the same genome form a ghost duplication when their homology can only be inferred through comparison with the genome of another species (see chapter 1.3). In Figure 1.1.4, the chromosomal segments A3.1 and A2.1 from *Arabidopsis* clearly show a pattern of differential gene loss when compared with the rice segment R10.1, since a number of genes located on the rice segment have been lost in one of the two paralogous segments of *Arabidopsis* (e.g. genes belonging to gene family 6733 (serine/threonine protein kinase), 4240 (bZIP leucine zipper) and 7796 (palmitoyl-protein thioesterase precursor)). Based on similar principles, hidden duplications can be inferred, which are heavily degenerated block duplications that cannot be identified by directly comparing both duplicated segments with each other, but only through comparison with a third segment of the same genome (see chapters 2.2 and 2.3). Consequently, hidden duplications are important to consider for determining the actual number of duplication events that have occurred over time, as previously demonstrated for *Arabidopsis* (Ku et al., 2000; see chapter 2.2). Indeed, by taking into account hidden duplications, one can often group additional segments in a multiplicon (a set of mutually homologous segments), as shown in Figure 1.1.4. The number of segments in a multiplicon, referred to as the multiplication level (Simillion et al., 2002; Vandepoele et al., 2003), can be used to infer the number of duplication events that must have occurred. For example, the presence of three homologous rice segments in the multiplicon shown in Figure 1.1.4 reveals that 2 duplication events must have occurred.

Apart from combining data of two genomes, Wong and coworkers (Wong et al., 2002) integrated partial sequence information of 14 related yeast strains in order to find evidence for an entire genome duplication event in *S. cerevisiae*. In their approach, the combination of a large number of chromosomal homologous segments allowed detecting heavily degraded duplicated regions, scattered throughout the genome.

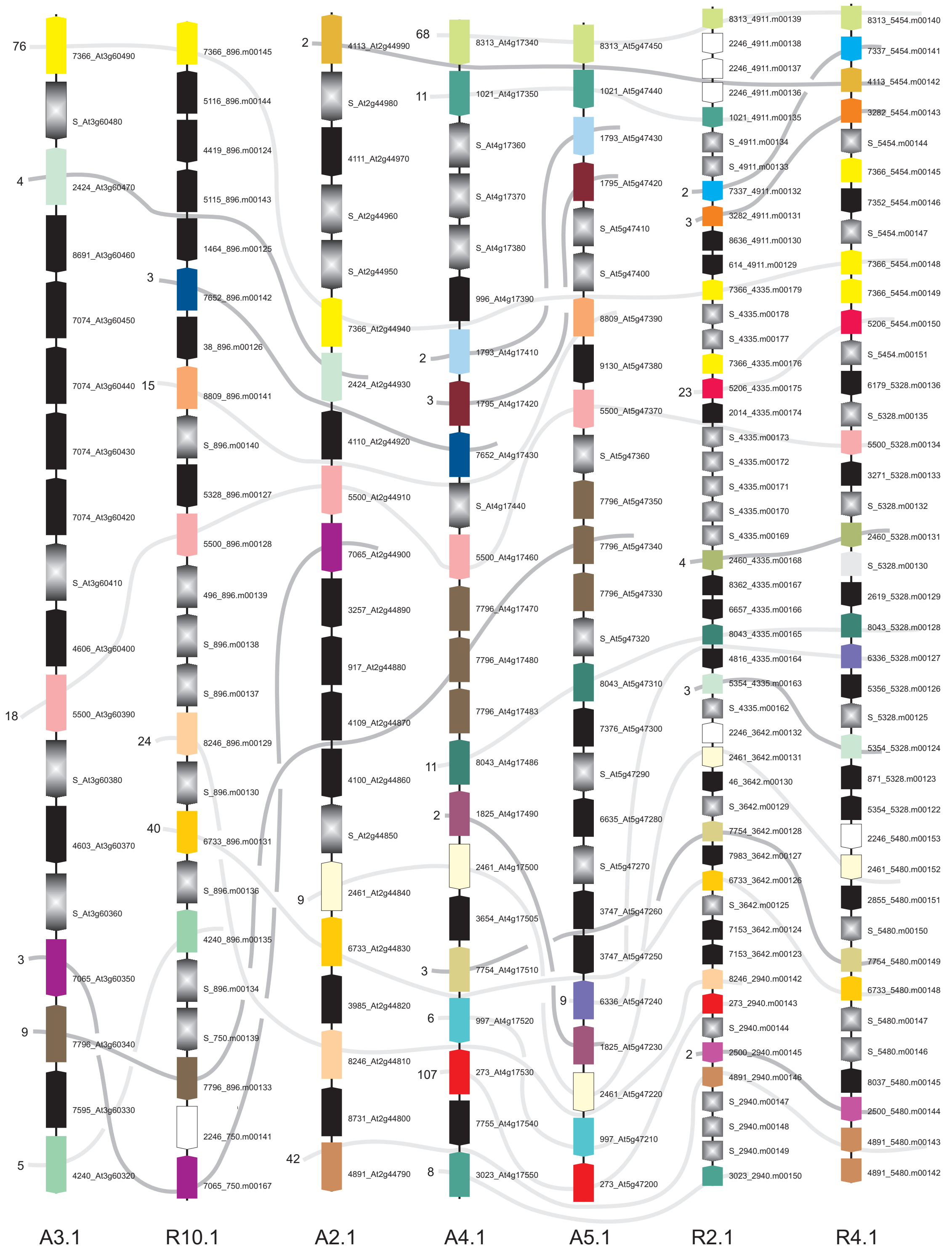


Figure 1.1.4 Set of homologous chromosomal segments (multiplicon) of *Arabidopsis* (segments A) and rice (segments R). Boxes represent the genes on the chromosomal segments whereas connecting lines indicate the anchor points (i.e. homologous genes part of the same gene family). Dark grey connecting lines show gene families of which 50% or more of all genes are present in the multiplicon shown (see text for details). Therefore, these genes provide a particularly strong case for homology. For each genomic segment, the names of the genes preceded by the gene

family ID are shown. Grey shaded boxes represent genes with no homologs in *Arabidopsis* and rice (gene family 'S' for singleton) and white boxes represent annotated genes with high similarity to retrotransposons. By considering the colinearity between *Arabidopsis* and rice, a set of, at first sight unrelated, *Arabidopsis* segments can be joined into a multiplicon with multiplication level 4 (i.e. the number of homologous segments in a multiplicon). Vice versa, this colinearity reveals that all three rice segments are linked with each other by two duplication events.

Genomic profiles: an extension to the map-based approach

Although considering transitive homologies such as hidden and ghost duplications allows the identification of many additional, previously undetectable, homologous genomic segments, it still requires that each of the homologous segments show significant colinearity with at least one other homologous segment. However, it is possible that, within a given multiplicon, one or more segments have diverged that much from the others in gene content and gene order, that they no longer show any clear colinearity with any of the other segments. Such segments that are in the twilight zone of genomic homology cannot be detected with any of the currently available methods. Recently, we have developing new software to uncover chromosomal segments that are homologous (in respect with having common ancestry) to others but can no longer be identified as such due to extreme gene loss. This is done by aligning clearly colinear segments and using this alignment as a 'genomic profile' that combines gene content and order information from multiple segments to detect these heavily degenerated homology relationships (see Figure 1.1.5; Simillion et al., 2004).

After the initial detection of a level 2 multiplicon with the basic ADHoRe algorithm (see chapter 1.2 for details), an alignment of the two segments that form this multiplicon is created where the anchor points of the multiplicon are positioned in the same columns. Using this alignment now as a profile, a new type of homology matrix can be constructed in which the gene products of a segment are compared to the gene products of the profile. Once this homology matrix is constructed, it is again presented to the basic ADHoRe algorithm, which will again detect clusters of anchor points applying the same statistical validation method as described before. This time, however, new significant clusters will not reveal homology between two individual segments but between the two segments inside the profile (i.e. the initial level 2 multiplicon) and a third segment. Because this type of GHM combines gene content and order information of the different segments in the profile, it is possible to detect homology relationships with a third segment

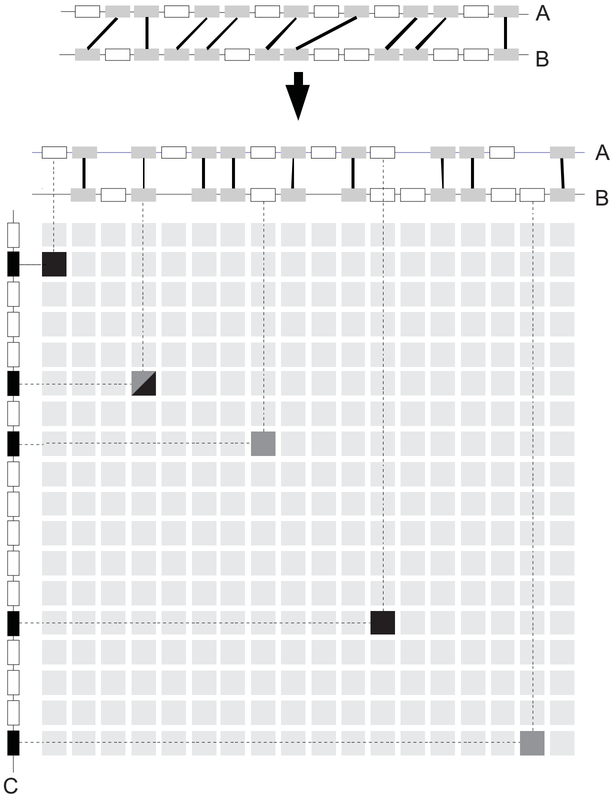


Figure 1.1.5 Detection of homology through a genomic profile. The upper section shows an initially detected level 2 multiplicon (a pair of homologous chromosomal segments). The grey boxes connected by black lines represent pairs of homologous genes (anchor points) between the two segments. The lower section shows the construction of a homology matrix using this multiplicon as a profile. To accomplish this, the multiplicon is first aligned by inserting gaps at the proper positions (depicted by empty spaces in the alignment). The homology matrix can now be constructed by comparing this profile with the genes of a chromosomal segment C (shown on the left of the matrix). Anchor points in the matrix are detected whenever a gene of this chromosomal segment belongs to the same gene family as one of the genes in any of the segments in the profile. The black squares represent homologs between segments A and C, the dark grey between B and C. The black/dark-grey square denotes a gene that has a homolog on both segment A and B. Combining segments A and B in a profile thus results in 5 anchor points with segment C, whereas the individual segments A and B only have 3 anchor points with segment C, which might be too few to decide on statistical significant homology.

that could not be recognized by directly comparing any of the segments of the multiplicon individually with this third segment. If such a third segment is detected, it is added to the multiplicon, thereby increasing its multiplication level, and the corresponding profile is updated by aligning the new segment to it. The entire detection process can now be repeated with the newly obtained profile (Simillion et al., 2004).

Biological implications of large-scale gene duplications for gene function

The widespread occurrence of large and small-scale duplication events highly complicates the extrapolation of functional relationships between homologous genes in different species (see for example chapter 3.1). Whereas one-to-one orthologous relationships suggest conservation of gene function, complex many-to-many homologous relationships offer limited information regarding gene function (Doyle and Gaut, 2000; see Figure 1.1.4). Although initially duplicated genes harbor redundant gene function, models have been formulated to explain the evolution of new functions (neofunctionalization) or preservation of both duplicates by subfunctionalization, where both members of a pair experience degenerative mutations that reduce their joint levels and patterns of activity to that of the single ancestral gene (Lynch and Force, 2000). Some biological examples of subfunctionalization have been documented (for review, see Prince and Pickett, 2002), but it remains unclear whether this model accounts for the majority of preserved gene duplicates.

One way further to understand the evolutionary mechanisms underlying the expansion of gene families is to combine segmental or tandem duplications with gene phylogenies. Recently, Cannon and Young developed a suite of programs for the detailed analysis of gene families combining comparative genomic positional information with phylogenetic reconstructions (Cannon and Young, 2003). As such, the impact of tandem and segmental duplications on gene family evolution can be inferred, which allows scientists to get deeper insights into the evolution of gene sub-families, which might be associated with functional divergence, or the acquisition of extra, potentially redundant, gene copies in particular species. Finally, this approach can provide valuable clues about conserved gene function in orthologous genes and functional divergence in paralogous genes.

Conclusion

It is clear that large-scale genome sequencing and advanced comparative sequence analysis offer a powerful combination to study the complex evolutionary forces that shape the structure of genomes. The analysis of complete genomes and the comparison of gene organization in related species finally allows scientists, at different levels of resolution from large-scale events such as translocations, duplications and segmental deletions to single-base pair differences, to unravel processes that drive gene and genome evolution (Eichler and Sankoff, 2003). Moreover, through the development of novel computational methods that allow the reliable detection of remnants of ancient large-scale gene duplication events, the evolutionary past of many eukaryotic genomes starts to reveal its secrets.

1.2 The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between *Arabidopsis* and rice

Klaas Vandepoele[+], Yvan Saeys[+], Cedric Simillion, Jeroen Raes and
Yves Van de Peer

Redrafted from: *Genome Research* **12**, 1792-1801 (2002)

It is expected that one of the merits of comparative genomics lies in the transfer of structural and functional information from one genome to another. This is based on the observation that, although the number of chromosomal rearrangements that occur in genomes is extensive, different species still exhibit a certain degree of conservation regarding gene content and gene order. It is in this respect that we have developed a new software tool for the Automatic Detection of Homologous Regions (ADHoRe). ADHoRe was primarily developed to find large regions of microcolinearity, taking into account different types of micro-rearrangements such as tandem duplications, gene loss and translocations, and inversions. Such rearrangements often complicate the detection of colinearity, in particular when comparing more anciently diverged species. Application of ADHoRe to the complete genome of *Arabidopsis* and a large collection of concatenated rice BACs yields more than 20 regions showing statistically significant microcolinearity between both plant species. These regions comprise from 4 up to 11 conserved homologous gene pairs. We predict the number of homologous regions and the extent of micro-colinearity to increase significantly once better annotations of the rice genome become available.

[+] both authors contributed equally

Introduction

Comparative genome analysis has demonstrated that across different plant species, which diverged from a common ancestor but currently tend to vary largely in genome sizes, gene content and order are often conserved. Especially, comparative genetic mapping in the grasses revealed a high degree of conservation of markers within large chromosomal segments (for reviews, see Gale and Devos, 1998b; Keller and Feuillet, 2000). Because, in general, different plant species use homologous genes for similar functions, these observations have great potential. Comparative genome mapping experiments can be a powerful and efficient tool to transfer biological information from a well-studied reference genome to related plant species. However, there are some serious drawbacks when using comparative genetic maps based on recombinational mapping of DNA markers. First, when the marker density is low, small exceptions to colinearity will not be observed, and second, the fact that most genes are organized in multigene families makes it difficult to determine whether real orthologous loci are being compared. Consequently, one can imagine that many experiments suffer from a bias toward promoting colinear regions and miss exceptions to colinearity (Bennetzen, 2000b).

The various sequencing efforts over the last few years, such as the complete genome sequence of the model plant *Arabidopsis thaliana* (Arabidopsis Genome Initiative, 2000), the YAC and BAC insert libraries of several grass genomes (Panstruga et al., 1998; Feuillet and Keller, 1999) and the International Rice Genome Sequencing Project (Sasaki and Burr, 2000), make it possible to investigate whether the degree of colinearity found in comparative genetic mapping experiments is also observed at the gene level. The existence of colinearity between model species and other plant species, even in a limited number of small regions, could provide the opportunity to use these model systems to identify candidate genes in other plants. Comparative sequence analysis at the sub-megabase level indicates that microcolinearity is abundant between closely related plant species, although exceptions do appear (Chen and Bennetzen, 1996; Kilian et al., 1997; Tikhonov et al., 1999; Tarchini et al., 2000). A high degree of conservation of gene content and order between orthologous loci of rice, maize, and sorghum has been reported (Chen et al., 1997). These grass species diverged from a common ancestor ~50 million years ago. Also, within related dicots, microcolinearity can be observed. For example, conserved gene content and order have been demonstrated between tomato and *Arabidopsis*, which diverged ~112 million years ago (Ku et al., 2000), between *Arabidopsis* and soybean (Grant et al., 2000), and between tomato, *Arabidopsis* and *Capsella* (Rossberg et al., 2001). All of these comparative studies revealed that rearrangements, such as inversions, deletions, insertions, and tandem duplications, are an important mechanism responsible for breaking up colinearity, and consequently, make it hard to detect the remnants of colinearity. In addition, these rearrangement processes appear to be more active in some plant lineages than in others (Devos et al., 1993; Devos and Gale, 1997; Schmidt, 2000).

When comparing more anciently diverged plant species, such as monocots and dicots, more rearrangements are expected to have occurred and, consequently, gene content and order to be less conserved. Recent DNA sequence analysis seems to confirm this assumption and several lines of evidence result in a plastic model in which the modern plant genome is characterized by a series of nested duplications in addition to the species-specific levels of rearrangements (Arabidopsis Genome Initiative, 2000; Vision et al., 2000; Wendel, 2000). Whether these currently observed large-scale gene duplications are the result of polyploidization or a large number of iteration events (entire genome duplication, entire chromosome duplication, and generic duplications of unspecific DNA regions within the same or between two chromosomes, respectively) is still highly debated. Nevertheless, all of the different actors identified so far in playing a role in the evolution of plant nuclear genomes make the picture rather complicated. Consequently, solid conclusions about genetic colinearity between *Arabidopsis* and rice, both expected to have a great value as a model system for dicots and monocots, respectively, are still missing, although several examples showing traces of microcolinearity have been reported (Devos et al., 1999; Van Dodeweerd et al., 1999; Liu et al., 2001; Mayer et al., 2001).

To carefully study genome evolution using the massive amount of sequence data that becomes available, we have developed a flexible tool, called ADHoRe (Automatic Detection of Homologous Regions), that detects genomic regions with statistically significant conserved gene content and order. Particularly, ADHoRe was developed to find large regions of colinearity, taking into account phenomena such as gene loss, inversions, and tandem duplications. This general concept makes it possible to use ADHoRe for analysis within one genome, that is, to look for paralogous regions with duplicated genes (Raes et al., 2002), or for comparisons between genomes of different organisms, that is, to look for synteny.

Results

In this study, we have applied a new tool to estimate the frequency and significance of microcolinearity between distantly related plant species such as *Arabidopsis* and rice. Therefore, publicly available rice genomic sequences (as a series of BACs) from seven different chromosomes were used to compare with the complete *Arabidopsis* genome sequence. For both plant species, gene annotation was retrieved from public resources (see Materials and methods). Important to note is that no prior information of macrocolinearity was incorporated into this analysis.

In total, using ADHoRe, we detected 105 cases of microcolinearity between *Arabidopsis* and rice before removing non-significant colinear regions, from which 75 are between individual rice BACs and a segment of the *Arabidopsis* genome and 30 are between overlapping rice clones and an *Arabidopsis* genomic segment. Applying the default 99% cut-off level, which retains all colinear regions that have a probability to be generated by chance of <1%, 24 segments showing conserved gene content and order between *Arabidopsis* and rice remain (listed in Table 1.2.1). Of these statistically significant regions, 18 (69%) show colinearity between an individual rice BAC and an *Arabidopsis* genomic segment, whereas 8 (31%) show colinearity between *Arabidopsis* and overlapping rice BACs. The distributions of the number of conserved genes within these homologous regions between *Arabidopsis* and rice for the different significance levels are shown in Figure 1.2.1. As expected for these classes of colinear regions characterized by a small number of conserved genes and a large number of non-homologous intervening genes, the probability that they were generated by chance is the highest. Consequently, applying more stringent conditions reduces the number of these colinear regions. For all significance levels, most of the statistically significant colinear segments are characterized by four conserved genes (referred to as anchor points hereafter).

The largest homologous segment between *Arabidopsis* and rice that ADHoRe could detect contained 11 conserved genes and is shown in Figure 1.2.2a. Detailed analysis showed that within this rice region on chromosome 1(326.8 kb), originally 64 genes have been predicted, resulting in a gene density of one gene per 5.1kb. The homologous *Arabidopsis* segment on chromosome 3 shows a gene density of one gene per 3.4 kb. However, validating the automatic rice gene prediction using Expressed Sequence Tag (EST) information and comparisons with putative homologs (see Materials and methods) shows that only ~32 genes are present, resulting in a gene density of one gene per 10 kb. As a result, the number of non-

Table 1.2.1 Overview of the colinear regions detected between *Arabidopsis* and rice (99% significance level)

Rice[a]	*Arabidopsis*[b]	Anchor points	BAC type	Clone name	*Arabidopsis* ORF[d]	Q^e (%)	$P_{chance}{}^f$
1	3	11	O	P0529H11, P0005H10, P0414E03	At3g54100	0.988	0.00
1	4	10	O	P0481E12, P0046E05	At4g18870	0.880	0.25
1	2	10	O	P0439E11, P0031D02, B1088C09 P0485B12	At2g30300	0.964	0.03
1	3	8	O	P0506B12, P0031D11	At3g55180	0.973	0.99
1	2	7	O	P0506B12, P0031D11	At2g39400	0.989	0.20
1	1	6	O	P0480C01, B1131B07	At1g34060	0.889	0.89
1	3	5	O	P0454H12, OJ1529_G03	At3g08670	0.986	1.00
4	5	5	I	OSJNBa0038O10	At5g23280	0.984	0.45
4	2	5	I	OSJNBa0042L16	At2g23380	0.977	0.80
6	3	5	I	P0698A06	At3g14230	0.926	0.64
1	2	5	O	P0518C01	At5g59480	0.945	0.57
4	5	4	I	OSJNBa0088H09	At5g06340	0.969	0.63
8	5	4	I	P0543D10	At5g43420	0.919	0.44
8	5	4	I	P0705A05	At5g43420	0.929	0.63
8	5	4	I	P0690C12	At4g08100	0.929	0.63
4	2	4	I	OSJNBa0084K20	At2g43230	1.000	0.44
10	1	4	I	OSJNBa0026O12	At1g03900	0.963	0.30
4	3	4	I	OSJNBa0033G16	At3g11630	0.985	0.63
10	1	4	I	OSJNBb0044B19	At1g03900	0.987	0.25
4	3	4	I	OJ1661_E06	At3g11630	0.985	0.63
6	5	4	I	P0468G03	At5g57140	0.999	0.89
4	3	4	I	OSJNBa0088H09	At3g52470	0.995	0.63
8	4	4	I	OJ1005_B05	At4g22730	0.983	0.20
2	1	4	I	OJ1288_G09	At1g78080	0.980	0.63

[a] Rice chromosome.

[b] *Arabidopsis* chromosome.

[c] O = overlapping BACs, I = individual BAC clone.

[d] Gene indicating the position of the homologous *Arabidopsis* segment.

[e] Score obtained by quality filtering (see text for details).

[f] Probability to be generated by chance.

homologous intervening genes between the anchor points drastically decreases, and consequently, the biological significance or quality of the colinear region to be homologous increases (see Materials and methods). An analogous approach was applied to determine whether all non-homologous intervening *Arabidopsis* genes were real genes. If not, removing genes in the *Arabidopsis* genome could also result in a higher degree of conservation within a colinear area. However, no indications were found that some of these intervening non-homologous *Arabidopsis* genes were falsely predicted.

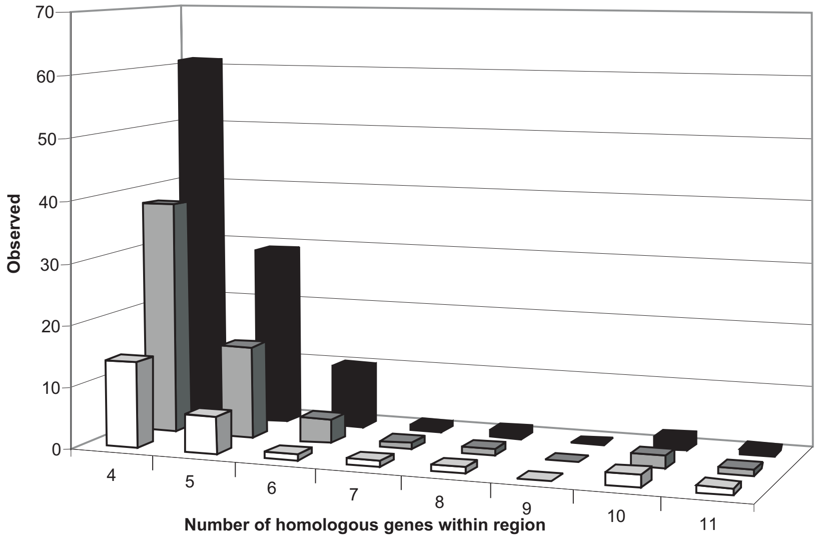


Figure 1.2.1 Distribution of the number of conserved genes within colinear regions of *Arabidopsis* and rice. The black, gray, and white histograms show the distribution of the blocks emerged by maximally 100%, 5%, and 1% chance, or 0%, 95%, and 99% significance levels, respectively. We propose to use the 99% significance level (i.e., maximally 1% probability to be generated by chance) as default setting.

Careful analysis of the long stretch of genomic sequence within the rice BAC clone P0414E03, characterized by a low gene density and no conservation with *Arabidopsis*, showed that multiple transposable elements have been integrated into this particular region (Figure 1.2.2a). Analysis of putative genes and ORFs revealed high similarities with proteins encoded by transposable elements (e.g., gag protein, reverse transcriptase, integrases, RNaseH). In addition, different sets of long repetitive elements were discovered, which allowed us to reconstruct a number of distinct transposable elements involved in plant gene and genome evolution (Grandbastien, 1992; Vicient et al., 2001). On the basis of organization of the proteins encoded in these transposons, three *gypsy*-like LTR-retrotransposons (Bennetzen, 2000a) and one *Mutator* (Lisch et al., 2001) transposable element could be identified, together with other transposon-like remnants. In the homologous *Arabidopsis* genome segment, no retrotransposable elements were detected. Figure 1.2.2b shows another colinear region between rice chromosome 1 and *Arabidopsis* chromosome 3, characterized by eight anchor points. Removing dubiously predicted rice genes results in a gene density of one gene per 7.7 kb (or 42 genes on the stretch of 305.1kb rice genomic sequence). The probability of this colinear region to be generated by chance is <1%. Several rearrangements can be clearly observed; since the divergence of rice and *Arabidopsis*, two genes have undergone tandem duplications in *Arabidopsis*,

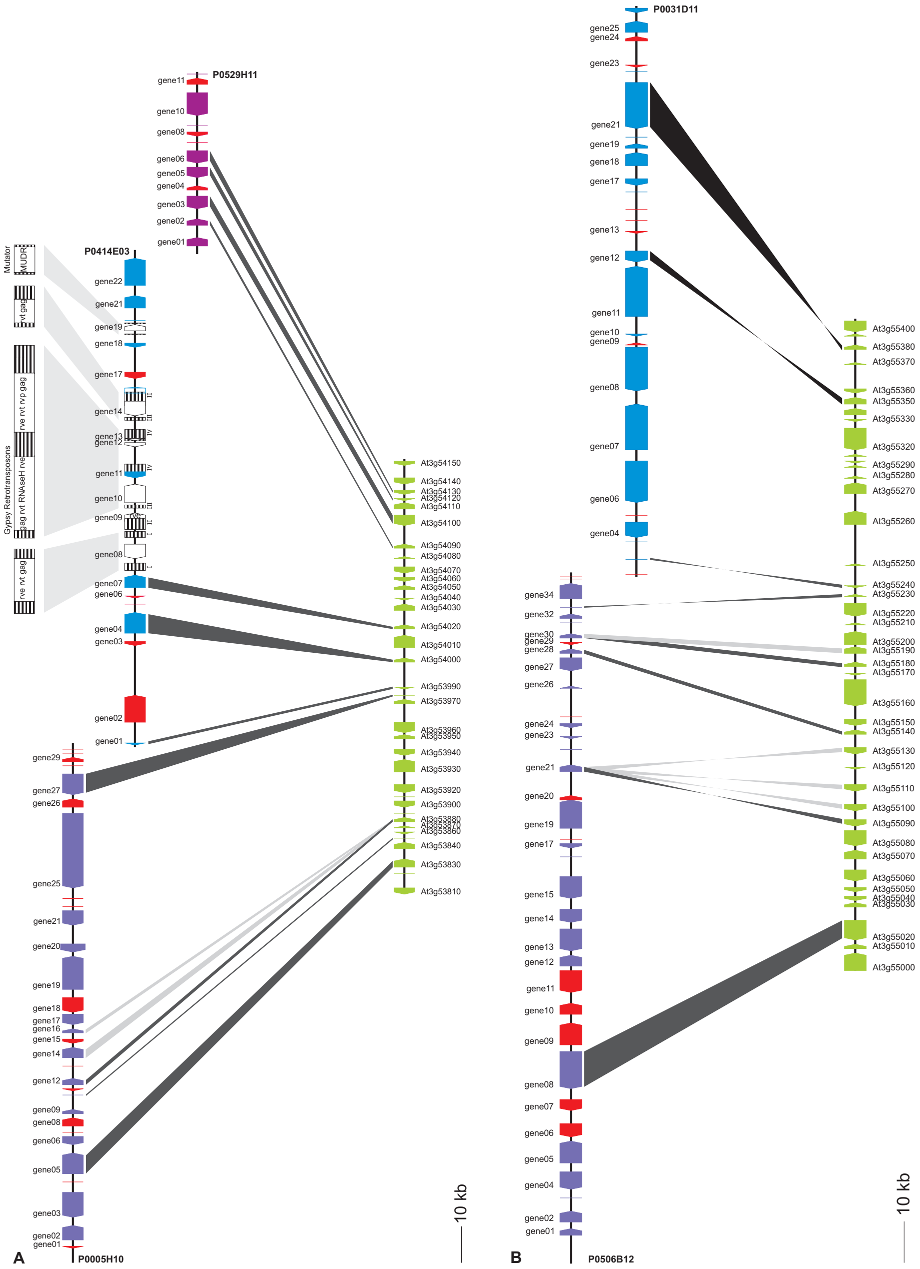


Figure 1.2.2 Examples of colinearity found between overlapping rice BACs and segments of the *Arabidopsis* genome. (a) Colinear segment between rice BACs (P0005H10, P0414E03, and P0529H11) and part of the *Arabidopsis* chromosome 3. Arrows indicate genes present on the genomic segment (black line), black bands connecting *Arabidopsis* and rice genes indicate anchor points (homologs), whereas gray bands indicate a tandem duplication. Genes probably erroneously predicted in rice are indicated in red (see text for details). LTRs are represented as hatched boxes. White boxes indicate gene products with similarity to proteins encoded by transposable elements. (gag) Retrotransposon gag protein; (rve) integrase core domain; (rvt) reverse transcriptase (RNA-dependent DNA polymerase); (rvp) retroviral aspartyl protease; (MUDR) MuDR family transposase. (b) Colinear segment between rice BACs (P0506B12 and P0031D11) and a segment of *Arabidopsis* chromosome 3.

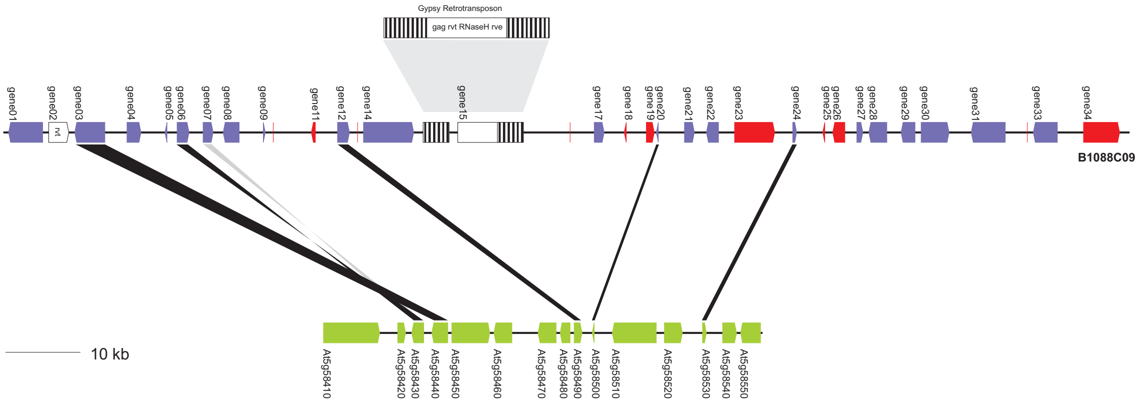


Figure 1.2.3 Colinearity between an individual rice BAC and a segment of the *Arabidopsis* chromosome 5. Interpretation is as in Figure 1.2.2.

whereas other genes have been inverted in *Arabidopsis* or in rice. A more drastic rearrangement event is shown in Figure 1.2.3. This colinear region between rice chromosome 1 and *Arabidopsis* chromosome 5 is characterized by five pairs of homologs (anchor points). Within the rice genomic fragment, a *gypsy*-like LTR-retrotransposon has been inserted, resulting in a much longer rice segment (96.8 kb) compared with the homologous *Arabidopsis* segment (39.8 kb). Next to the local gene inversions observed in a number of colinear regions, this example shows a more complex inversion event. Genes 03 and 06 located on rice BAC B1088C09 are part of a segment colinear with *Arabidopsis* chromosome 5, although their gene order and orientation are not conserved compared with the other anchor points. Therefore, a chromosomal segment encoding these two genes (or their *Arabidopsis* orthologs) seems to have been inverted after both species diverged from each other. However, reconstructing the history leading to this configuration requires an additional inversion event. Because for gene 06, in contrast to all other genes conserved within this homologous region, the orientation compared with the homologous *Arabidopsis* gene is different (see twisted black band in Figure 1.2.3), one extra gene inversion is required to explain the current gene organization between these two genomic fragments. Finally, gene 06 experienced a tandem duplication resulting in gene 07, or vice versa.

Discussion

It is estimated that rice and *Arabidopsis* have diverged ~200 million years ago (Yang et al., 1999; Wikström et al., 2001). Nevertheless, applying our newly developed tool to detect homologous regions between both plants revealed numerous examples of significant microcolinearity. On the other hand, of the total set of colinear regions present between rice and *Arabidopsis*, probably only a subset can be considered as genuine orthologous regions that originated from a common ancestral region. The major cause of this phenomenon is the fact that many genes are organized in multigene families, and consequently, the discrimination between paralogous and orthologous gene sequences is extremely difficult. Therefore, we incorporated a routine in the ADHoRe algorithm to determine whether a colinear region could have been generated by chance out of homologous gene couples. In other words, it was tested whether a particular colinear region is a homologous region or purely consists of homologous gene couples organized in a colinear way by chance. Analysis of a number of colinear regions characterized by a high probability to be generated by chance showed that low overall-similarity signals, such as similarities between DNA-binding sites, or badly conserved gene content and order were detected (data not shown).

Combining numerous rice BACs resulted in a set of long genomic rice stretches that could be investigated for colinearity with *Arabidopsis*. Although only a small fraction of the final rice genome sequence was used in this study (~38%, for which 62 MB was organized in overlapping BACs), already >20 regions between rice and *Arabidopsis* were found with biologically relevant colinearity, consisting of 4 up to 11 conserved genes. Because a large number of short colinear regions are found between individual rice BAC clones and an *Arabidopsis* genome segment, a major fraction of these regions were removed because they could represent colinear regions generated by chance. However, with more rice genomic sequence data becoming freely accessible very fast, we expect that concatenation of additional BACs will generate longer colinear stretches with *Arabidopsis*. Therefore, a number of colinear regions currently not retained in our final results could become statistically significant when analyzed over longer distances. Consequently, the real number of rice regions showing microcolinearity with *Arabidopsis* will most probably be higher than presented here. Preliminary results on the draft sequence of the rice genome show that larger colinear segments may exist between *Arabidopsis* and rice (Goff et al., 2002). However, as the annotation of the draft sequence is not yet publicly available, a comparison with the results described here remains difficult.

Detailed analysis of some colinear regions indicates that the quality of the rice annotation used in this comparison is not outstanding. Although the RiceGAAS system (Sakata et al., 2002) tries to benefit from combining a number of different gene prediction programs, a large number of errors still seem to be present. The crude quality assessment performed here to determine whether a predicted gene is a real gene (i.e., sensitivity) revealed that a major fraction of the protein-encoding genes were falsely predicted. Consequently, the initial gene density determined by the gene prediction system decreased drastically when removing unreliable predicted genes. In addition, a number of genes were split (one gene predicted as two separate genes) and some exons or complete genes were missing, which could be demonstrated by incorporating EST information. Especially the large number of ORFs predicted as genes poses a problem, because a small number of these ORFs actually are confirmed by EST information, but the major fraction was not. All of these annotation inaccuracies will definitely have their repercussions on the correct interpretation of the rice genome sequence, in a way similar to that faced in annotating the *Arabidopsis* genome sequence (Pavy et al., 1999). Therefore, further improvement and retraining of rice gene prediction programs, together with newly developed extrinsic gene prediction methods seems inevitable for fully exploiting the rice genome sequence (Rouzé et al., 1999; Bennetzen, 2002).

Next to the incorrectly predicted protein-encoding genes, a subset of these erroneously predicted genes seems to correspond with transposable elements. Although detailed analyses can unambiguously identify these elements, the presence of these elements annotated as protein-encoding genes is a major problem when performing genome-wide analyses such as described here. Although in the *Arabidopsis* genome 2,109 Class I transposable elements have been described already (Arabidopsis Genome Initiative, 2000), an additional screening reveals that within the *Arabidopsis* proteome nearly 600 predicted protein-encoding genes are present with high similarity to some retrotransposable elements (data not shown). Furthermore, it should be noted that the largest fraction of these genes resembling retrotransposable elements has been identified on chromosomes 1, 2, and 4. Because chromosomes 2 and 4 have been sequenced and analyzed first within the *Arabidopsis* sequencing project, an imperfect annotation protocol for transposons at that moment could be an explanation for this observation. For ~36% of these detected genes, an EST matches the structural annotation, which could explain why these genes have been allocated as protein-encoding genes in the automatic annotation protocols. Nevertheless, additional efforts seem most likely to increase the quality of the current annotation on a full-genome level toward transposable elements in both rice and *Arabidopsis* (Le et al., 2000).

Although transposable elements integrate and retrotransposons amplify within plant genomes, when correctly annotated, they should not interfere with the presented algorithm to detect homologous regions. Consequently, this level of complexity generated by transposable elements can be masked in our method, if all transposable elements are defined as such and not as protein-encoding genes in the genomic sequence. Analysis of multiple colinear regions showed that the number of retrotransposable elements in rice was considerably higher than in the homologous *Arabidopsis* segments, although the actual number of retrotransposable elements in *Arabidopsis* is probably higher than described so far (Arabidopsis Genome Initiative, 2000). Accumulation of retrotransposons in plant genomes clearly seems to be dependent of both the evolutionary lineage and the efficiency of mechanisms repressing this activity (Bennetzen and Kellog 1997; Fedoroff, 2000).

It is clear that all sorts of rearrangements have occurred since rice and *Arabidopsis* diverged from each other ~200 million years ago. Detailed analysis of colinearity between *Arabidopsis* and rice identified tandem duplications and gene loss, as well as gene and block inversions, although the frequency of these detectable events is rather low. In other words, it is not possible to trace all rearrangements that are responsible for the nonhomologous genes present in colinear regions. The main driving force responsible for degrading colinearity is seemingly a complex evolutionary mechanism, consisting of species-specific levels of large and small rearrangements (due to duplications, inversions, insertions, and deletions), transposon activity, and perhaps other unknown mechanisms. Ideally, the continuous improvement of data sets, methods, and additional genome sequences from intervening species will give us further insight into these mechanisms and their frequencies within different species.

Finally, the question remains whether, after detecting colinearity between genomes, the functions of the genes in one genome may be transferred to the homologous genes of the other genome. One major problem lies in the fact that a particular region of a chromosome can be duplicated in rice as well as in *Arabidopsis*. Even more drastically, complete genome duplication events may have occurred in both *Arabidopsis* (e.g., Arabidopsis Genome Initiative, 2000; Vision et al., 2000) and rice (e.g., Goff et al., 2002; Yu et al., 2002). Because after such a duplication event, all genes are present in duplicate, one copy may degenerate through loss-of-function mutations, or both duplicates may remain redundant, experience subfunctionalization, or diverge in function through positive Darwinian selection (e.g., Ohno, 1970; Force et al., 1999; Hughes, 1999; Van de Peer et al., 2001). This results in a situation in which one genomic segment of one species maps with two or more different segments in the other genome, or vice versa. Transferring functional annotations from one genome to the other genome, thus, has to be done with caution, as genes belonging to paralogous regions may have considerably diverged in function.

Materials and methods

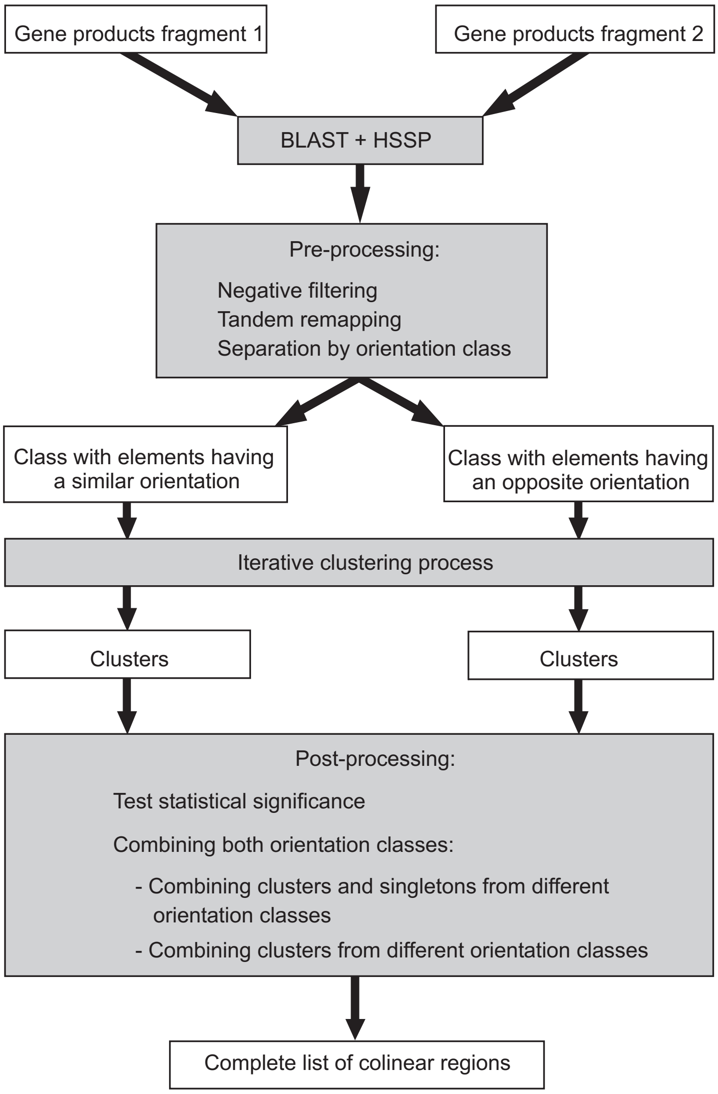

Gene products fragment 1
Gene products fragment 2
BLAST + HSSP
Pre-processing:
Negative filtering
Tandem remapping
Separation by orientation class
Class with elements having a similar orientation
Class with elements having an opposite orientation
Iterative clustering process
Clusters
Clusters
Post-processing:
Test statistical significance
Combining both orientation classes:
- Combining clusters and singletons from different orientation classes
- Combining clusters from different orientation classes
Complete list of colinear regions

Figure 1.2.4 Flowchart of the ADHoRe strategy used to define colinear regions between two genomic fragments. White boxes represent data items, gray boxes represent routines, and arrows indicate the dataflow.

The ADHoRe algorithm

Detection of homologous genes

To detect chromosomal locations of colinear genes, one has to look for regions that can be paired up because they contain sets of similar genes. Therefore, a data set containing all gene products, their absolute or relative position on a genomic sequence, and their orientation is required. The whole procedure is controlled by two parameters as follows: the gap size *G*, which describes the maximal number of intervening, non-homologous genes tolerated between two homologous genes within a colinear segment, and *Q*, the quality of the colinear regions (see below). Figure 1.2.4 presents a flowchart of the algorithm. For all gene products on two genomic fragments for which gene colinearity is to be detected, initially an all-against-all sequence similarity search is performed, using BLASTP (Altschul et al., 1990). In a second step, all of these results are converted into sequence identity scores (over a given alignable region) between query and hit sequences. Two protein sequences with >30% sequence identity over an alignable region of 150 amino acids are considered as being homologous. For matching sequences with an alignable region smaller than 150 amino acids, the

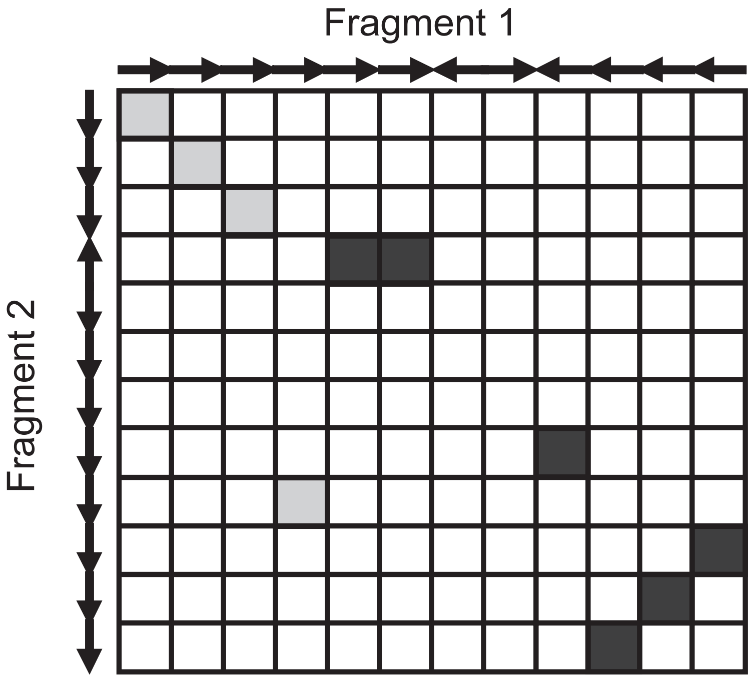


Figure 1.2.5 Matrix representation of homologous genes. Arrows indicate the orientation of the genes on the two genomic fragments compared. Homologous genes with the same orientation are colored in gray; homologous genes with an opposite orientation are in black.

Homology-derived Secondary Structure of Proteins (HSSP) identity cut-off curve is used to determine whether the two sequences are homologous (Rost, 1999). With this procedure, all pairs of homologous proteins between both genomic fragments are determined.

The information on homologous genes is then stored in a matrix of ($m \cdot n$) elements (m and n being the total number of genes on each genomic fragment), each non-zero element (x, y) being a pair of homologous genes (x and y denote the coordinates of these genes). Figure 1.2.5 shows such a small hypothetical matrix, in which gray elements indicate gene pairs having the same orientation, whereas black elements indicate homologous pairs of genes having an opposite orientation. In the matrix, colinear regions are represented as diagonal lines, whereas tandem duplications are manifested as purely horizontal or vertical lines; inversions can be detected by looking at the organization of the elements, and block duplications followed by gene loss form gaps in diagonal regions. To detect colinear regions, it is obvious that one has to find more or less diagonal series of elements in the matrix. This way of presenting the information reduces the problem to a clustering problem. When the matrix is constructed, it is subjected to a number of procedures that, in the end, returns all colinear regions present between both genomic fragments. In general, these procedures can be subdivided into three steps, pre-processing of the data, the actual clustering of homologous genes or blocks of genes, and post-processing.

Pre-processing of the data

As discussed above, during the pre-processing step, the two genomic fragments are compared, and homologous gene pairs are determined using BLAST and HSSP, after which, these are stored in a matrix. The orientation of the two genes determines the value in the matrix, whereas non-homologous pairs are represented as empty elements in the matrix.

The next step during the pre-processing is the removal of irrelevant data points, which we designate negative filtering. During this step, all elements that cannot belong to a cluster because they are too far away from other elements in the matrix, are removed. The last step in the pre-processing is to remap tandem duplicated blocks. Because we are looking for diagonal regions in the matrix, purely horizontal or vertical regions due to tandem duplications are remapped. This is done by collapsing all tandem duplications of a gene with the same orientation and within a distance *G*. This way, it is easier to detect diagonal regions, as they are no longer interrupted by horizontal or vertical elements. At the end of the pre-processing, the elements in the matrix are separated according to their orientation, yielding the two orientation classes (see Figures 1.2.4 and 1.2.5). This separation is made to facilitate the clustering and is based on the observation that colinear regions consist primarily of elements with the same orientation class. At the end of the process, both orientation classes are again combined, enabling the reconstruction of duplicated regions that have been subjected to small gene inversions.

Clustering of genes and blocks of genes

A colinear region is defined in the matrix representation as a number of points showing diagonal proximity. Therefore, a special distance function was used, yielding a shorter distance for points that are in diagonally closer proximity than

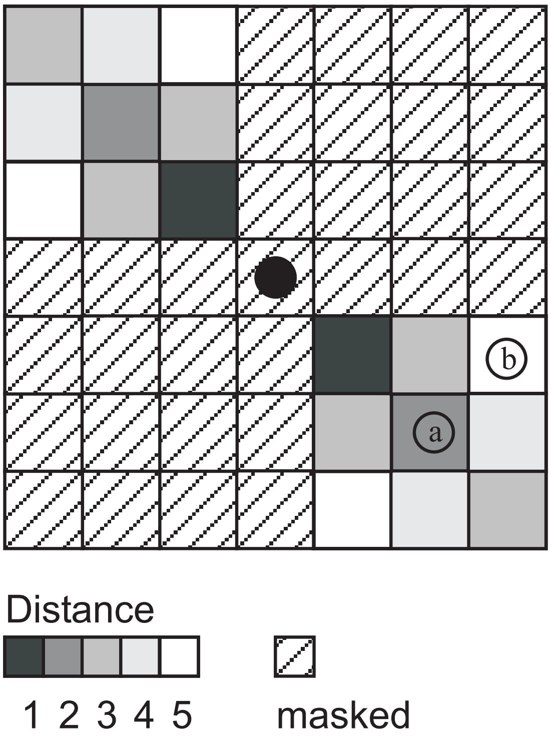


Figure 1.2.6 Graphical representation of the DPD function. Every rectangle represents a cell of the matrix. The central dot corresponds with an element of a cluster. Because the DPD distance to element *a* is 2 and the DPD distance to element *b* is 5, *a* is in closer proximity to the central dot under investigation than *b*. According to the orientation class, a specific region of the environment is masked (which corresponds to an infinite distance).

points that are in horizontal or vertical proximity. The formula for this function is:

$$d = 2\max(|y2 - y1|, |x2 - x1|) - \min(|y2 - y1|, |x2 - x1|)$$

Because the triangle inequality does not hold for this function, it cannot be regarded as a real distance function, but rather as a diagonal pseudo distance (DPD) function. Figure 1.2.6 shows the result of applying such a distance function on a hypothetical example. The actual clustering step is conceived as an iterative process, gradually increasing the gap size until the final gap size - one of the parameters of the algorithm - has been reached. During each iteration, the gap size represents the maximal distance between two points in a cluster. In each iteration, new clusters can be formed and existing clusters can be extended. The algorithm details of the clustering step are depicted in Figure 1.2.7. Starting with the elements of either one of the two orientation classes (a set of singletons, i.e., elements not yet clustered), the DPD function is used to cluster the elements according to the initial gap size. By default, the initial gap size is set to 3 and is then increased in 10 exponential steps until the final gap size G has been reached. This results in a set of clusters and a set of singletons.

Subsequently, the second parameter of the algorithm comes into play. This parameter determines to which extent the elements of a cluster fit on a diagonal line. This quality is estimated by calculating the coefficient of determination (r^2) by linear regression through the points in the clusters. Only clusters with a sufficiently high quality (higher than the cut-off Q, set by the second parameter) will be kept; the constituting elements of the other clusters are reassigned the status of singletons. Within each iteration, the remaining data set after applying the DPD clustering and the quality filtering is a collection of retained clusters and a collection of singletons (from the orientation class being analyzed) not yet clustered, or initially clustered, but rejected by the quality filtering (Figure 1.2.7). In the next step, which also uses the DPD function, it is tested whether the clusters can be enriched with singletons from the same orientation class without badly affecting the cluster’s diagonal properties. Therefore, three conditions must be fulfilled. First, the candidate singleton must be within a distance smaller than or equal to the current gap size in the iteration. Second, the candidate singleton must be positioned within the 99% confidence interval of the cluster. This confidence interval is computed by considering the best-fit line $y = ax + b$ through all of the points in the cluster using the least-squares fit method. Usually, the points in the cluster

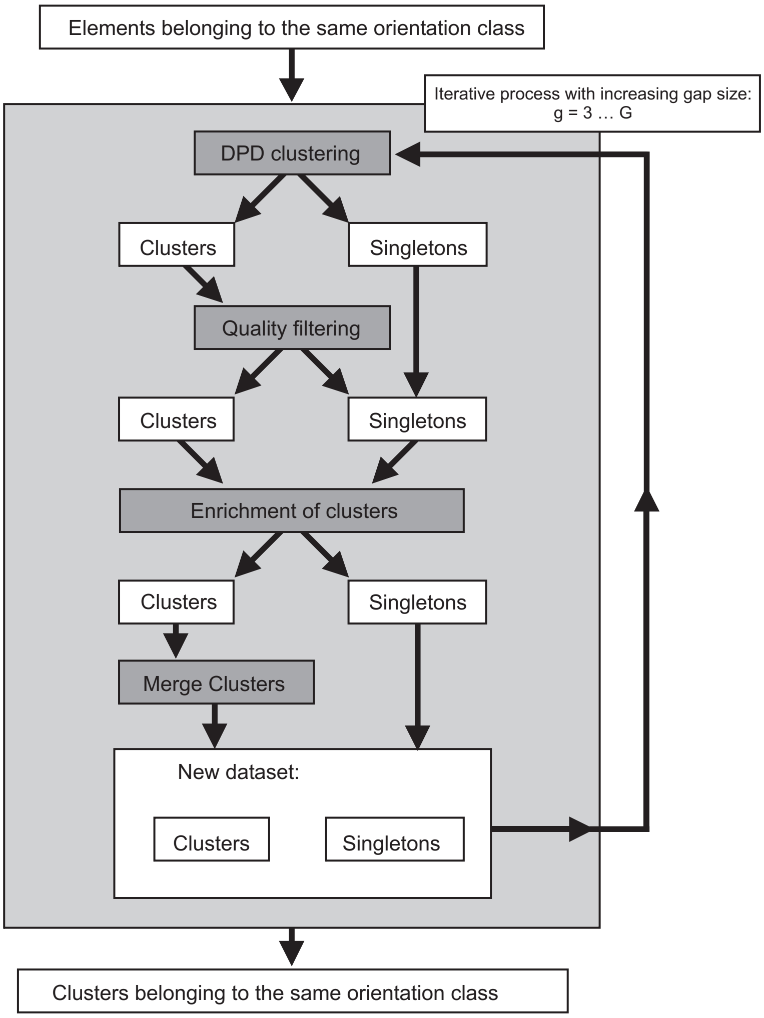


Figure 1.2.7 Flowchart of the ADHoRe core algorithm. Dark gray boxes represent the different steps in the clustering process, white boxes the data items, and the light gray boxes the actions performed during each iteration step. Arrows indicate the dataflow.

show a certain degree of deviation from this line. This deviation can be explained by two factors: (1) the error on the calculation of the constants *a* and *b* of the regression line, and (2) the error caused by the deviation of the point $x_{i,}$ y_i from this line. Assuming a normal distribution of this deviation, we can calculate a confidence interval that indicates the maximum deviation a candidate singleton can have from the best-fit line. Finally, if a singleton lies within these boundaries, it is also checked whether adding this singleton to the cluster will not decrease the r^2 value (see above) below the specified r^2 cut-off. If all criteria are met, the singleton is then added to the cluster. If not, the original configuration of both cluster and singleton is restored. The last step of the core algorithm aims at joining clusters. For each cluster within a distance smaller than *g* (*g* being the gap size in the current iteration) of another existing cluster, it is tested whether it can be merged with that cluster, again without badly affecting the cluster's diagonal properties. To determine whether two clusters can be joined, we first check whether the distance between the diagonal lines through the central points of both clusters is not larger than *g* (using DPD). Next, we check whether the distance between the endpoints of both clusters is small enough. If the clusters have overlapping *x* or *y* coordinates, we consider the distance between them to be 0. In this case, we have to check whether from both clusters at least one point lies in the confidence interval of the other or whether all points of one cluster lie in the interval of the other. This is to avoid grouping of closely, in-parallel-aligned clusters. Finally, we check whether the r^2 value of the resulting merged cluster does not drop below the specified r^2 cut-off. The resulting new data set again consists of a number of clusters and a number of singletons, which are used as input for the next iteration during the process (Figure 1.2.4). During the next iteration, the gap size is increased and new clusters are made or existing clusters extended, until the final gap size has been reached. The result is a set of clusters for each orientation class.

Post-processing

When all clusters have been compiled as described above, the fraction of colinear regions (clusters) that are not significant needs to be removed. The goal of this procedure is to determine the fraction of colinear regions that could have occurred purely by chance, and therefore are not biologically significant. This is implemented as a statistical test, sampling a large number of reshuffled data sets and calculating the probability that a colinear region, characterized by a number of conserved genes and an average gap size, can be found by chance. Using a default significance level of 99%, all regions with a probability to be generated by chance smaller than 1% are retained. The second step during post-processing is to combine the results for the two sets of clusters with different orientations. First, we try to enrich clusters from one orientation class with singletons from the other orientation class. This step is similar to the third step in the clustering algorithm, in which clusters are extended without badly affecting the quality. Second, it is tested whether clusters from the two different orientation classes can be merged. By combining the results of both orientation classes, it is possible to reconstruct larger colinear regions that might have been subjected to one or more inversion events.

The rice data set

For rice chromosomes 1, 2, 4, 6, 7, 8, and 10 (a set of chromosomes for which a large fraction of the chromosome was already sequenced), the public data of the different centers was collected (status January 14, 2002). All BAC sequences for which map position information was available and that were linked to one chromosome only were downloaded from the different consortia websites, for which an overview can be found at http://www.tigr.org/tdb/e2k1/osa1/BACmapping/description.shtml.

Concatenation of rice BACs

To obtain large stretches of genomic rice sequence to compare with *Arabidopsis*, we used a simple strategy to build rice contigs. Initially, for all BAC clones, the BAC extremities were compared with BAC ends of neighboring BACs using BLASTN (Altschul et al., 1990). These BAC ends were defined as the first and the last 20% of the genomic BAC sequence. For each BAC, the 25 closest neighboring BACs were scanned, given their putative map position. Two BACs were considered overlapping when an alignable region >300 bp showed >95% sequence identity. Next, all pairs of overlapping BACs were used to build larger stretches of adjacent overlapping BAC sequences (pair A-B and pair B-C producing stretch A-B-C, etc.). In the case in which one BAC overlapped with multiple other BACs, preferentially the BAC resulting in the longest stretch was selected. Note that these BAC stretches were not physically assembled into a contig sequence, but that this information was only used to locate and order the BACs relative to each other. This procedure divided the initial data set into two large fractions, a set of overlapping BACs (in total, 453 BACs, or 37% of the total size of the original data set) and a set of remaining individual BACs.

Annotation

For all rice BACs, gene annotation was performed using RiceGAAS (Sakata et al., 2002). This system combines a total of 14 analysis programs and automatically generates gene annotation for all rice BACs present in GenBank. For all BACs retained in the data set, the predicted coding sequence and corresponding protein sequences were retrieved from the RiceGAAS website (http:/

Table 1.2.2 Overview of the rice data set used[a]

		Total data set				Overlapping BACs		
Chromosome	Sequenced (%)	MB	BACs	Annotated genes	Gene density[b]	BACs	MB	Genes
1	100.0	50.68	370	10,300	4.92	266	34.97	6,237
2	44.7	18.40	154	3,692	4.98	2	0.30	38
4	92.4	18.90	143	3,766	5.02	75	10.76	2,064
6	63.0	21.57	159	4,410	4.89	6	0.94	163
7	75.5	20.33	164	4,149	4.90	7	0.86	168
8	46.2	16.85	139	3,398	4.96	24	3.55	615
10	95.6	19.28	145	3,806	5.07	73	10.83	1,892
Total		166.01	1,274	33,521	4.95	453	62.00	11,177

[a] Status on January 14, 2002; source TIGR.

[b] Genes/kb.

/ricegaas.dna.affrc.go.jp/). An overview of the number of BACs and proteins used can be found in Table 1.2.2. Finally, using the two sets of BAC clones (overlapping and individual BACs) and their corresponding gene annotation, gene lists were made and used as input for the ADHoRe algorithm. Parameters used for the ADHoRe algorithm were $G = 20$ for the maximum gap size and $Q = 0.8$ to denote the quality of the cluster. In total 1,000 reshuffled data sets were used to calculate the probability that a colinear region, characterized by a number of conserved genes and an average gap size, could have been generated by chance. For the genomic rice regions showing homology with an *Arabidopsis* genomic segment, which were analyzed in detail, the quality of the annotation retrieved from RiceGAAS was estimated. Therefore, for each predicted gene, we checked for the existence of a rice EST and for homology of the corresponding protein with any other protein present in the public protein databases. All predicted genes not confirmed by an EST and not showing similarity with any other protein were not considered as genes. Although these criteria are not biologically correct (i.e., these genes could be rice specific, not confirmed by ESTs and occur as a unique gene, not part of a multigene family in the rice genome), they were used here to determine rather crudely the quality of the annotation system. The same criteria applied to the total set of predicted genes in *Arabidopsis* shows that only 0.31% (79/25,439) genes are selected. Thus, on the basis of the ratios found in the *Arabidopsis* genome, we expect that from the complete set of rice genes we remove in this way, <0.3% might be real genes. For all analyzed rice segments, on average, 45% of the predicted genes were removed.

Annotation of transposable elements

Initially, the genomic BAC sequence was screened for repetitive elements using REPuter (Kurtz and Schleiermacher, 1999). In addition, predicted genes and ORFs were screened against a collection of protein families and domains using PFAM (Bateman et al., 2002) to determine similarities with proteins encoded in transposable elements. Artemis was used for sequence visualization and annotation (Rutherford et al., 2000).

Arabidopsis data set

Genomic sequences and gene annotation for the complete *Arabidopsis* genome was downloaded from the TIGR *Arabidopsis thaliana* Database (version August 2001, http://www.tigr.org/tdb/e2k1/ath1/) and processed with in-house Perl scripts.

1.3 Detecting the undetectable: uncovering duplicated segments in *Arabidopsis* by comparison with rice

Klaas Vandepoele, Cedric Simillion and Yves Van de Peer

Redrafted from: *Trends in Genetics* **18**, 606-608 (2002)

Genome analysis shows that large-scale gene duplications have occurred in fungi, animals and plants, creating genomic regions that show similarity in gene content and order. However, the high frequency of gene loss reduces colinearity resulting in duplicated regions that, in the extreme, no longer share homologous genes. Here, we show that by comparison with an appropriate second genome, such paralogous regions can still be identified.

Introduction

Genome sequencing projects reveal that genomes vary tremendously in size and organization, even among closely related organisms. This seems to be the result of a very dynamic process involving many different factors, such as recombinations, horizontal gene transfer, transposon activity, gene duplication and gene loss. In particular, duplications are being identified as important factors in the evolution of most genomes. Apart from small-scale tandem duplications, larger block duplications and even duplications of entire chromosomes or genomes are now postulated to have shaped the genomes of various animals, fungi and plants (Wolfe, 2001). From a population genetics point of view (Force et al., 1999), the frequency of gene preservation over a large evolutionary period after duplication is unexpectedly high and several models have recently been put forward to explain the retention of duplicates (Gibson and Spring, 1998; Lynch and Force, 2000; Wagner, 2002). However, the most likely fate of a gene duplicate is non-functionalization and consequent gene loss (Lynch and Conery, 2000).

This observation has consequences for the detection of duplicated regions in genomes. Identifying duplicated regions is usually based on a within-genome comparison that aims to define colinear regions (regions of conserved gene content and order) in different parts of the genome. In general, one tries to identify duplicated blocks of homologous genes that are statistically valid (i.e. that are probably not generated by chance). The statistics that determine colinearity usually depend on two factors, namely the number of pairs of genes that still can be identified as homologous (usually referred to as ‘anchor points’), and the distance over which these gene pairs are found, which usually depends on the number of ‘single’ genes that interrupt colinearity. When a putative colinear region has been detected, its statistical significance is usually evaluated by some sort of permutation test in which a large number of randomized datasets are sampled to calculate the probability that a cluster detected could have been generated by chance (Vision et al., 2000; Gaut, 2001; Friedman and Hughes, 2001; Vandepoele et al., 2002a). However, the high level of gene loss – together with phenomena such as translocations and chromosomal rearrangements – often renders it very difficult to find statistically significant homologous regions in the genome, particularly when the duplication events are ancient (Ku et al., 2000).

The search for traces of (ancient) large-scale gene duplications has received much attention lately, and hypotheses about the number and age of polyploidy events in eukaryotes are actively being discussed. Partly, this is because of the fact that the detection of homologous (paralogous) regions in genomes is not self-evident, for the reasons discussed above and, in consequence, the number of duplicated regions is likely to be underestimated. In plants, the systematic analysis of the *Arabidopsis thaliana* genome sequence has shown that this genome contains a large number of duplicated regions and that about 60% of the *Arabidopsis* genes occur in duplicated blocks (Blanc et al., 2000; The Arabidopsis Genome Initiative, 2000; Simillion et al., 2002). Here, we show that additional duplicated regions can be discovered in *Arabidopsis* when its genome is compared with that of rice.

Results and discussion

Recently, the draft genome sequences have been reported for two subspecies of rice (Goff et al., 2002; Yu et al., 2002), in addition to data being made available by the International Rice Gene Sequencing Project (Sasaki and Burr, 2000). We used the IRGSP data to compile a large set of BAC sequences for which the map position information is available and used these, where possible, to build longer rice contigs. This resulted in a dataset of 453 overlapping BACs, forming continuous genomic stretches of 62 Mb, and a remaining set of 821 individual BACs (representing 104 Mb). We compared these with the *Arabidopsis* genome to find statistically significant regions of colinearity between the genomes, using a new software tool called ADHoRe (for ‘Automatic Detection of Homologous Regions’) (Vandepoele et al., 2002a).

The comparison of rice, the major food source for billions of people and a model for larger cereal crop genomes (Shimamoto and Kyozuka, 2002) with *Arabidopsis*, a model plant organism for dicotyledons, revealed numerous examples of (short) genomic segments that shared conserved gene content and order, as reported previously (Mayer et al., 2001; Salse et al., 2002). In several cases, two (or more) regions of the *Arabidopsis* genome showed clear homology with a single region in rice. This is not surprising, because the *Arabidopsis* genome has undergone at least one (Lynch and Conery, 2000; The Arabidopsis Genome Initiative, 2000), and probably more (Vision et al., 2000; Simillion et al., 2002), polyploidizations. However, some of the duplicated regions escape detection in a within-genome comparison of *Arabidopsis*. More detailed analysis shows that each of these regions in *Arabidopsis* has lost a different set of genes (see Figure 1.3.1a).

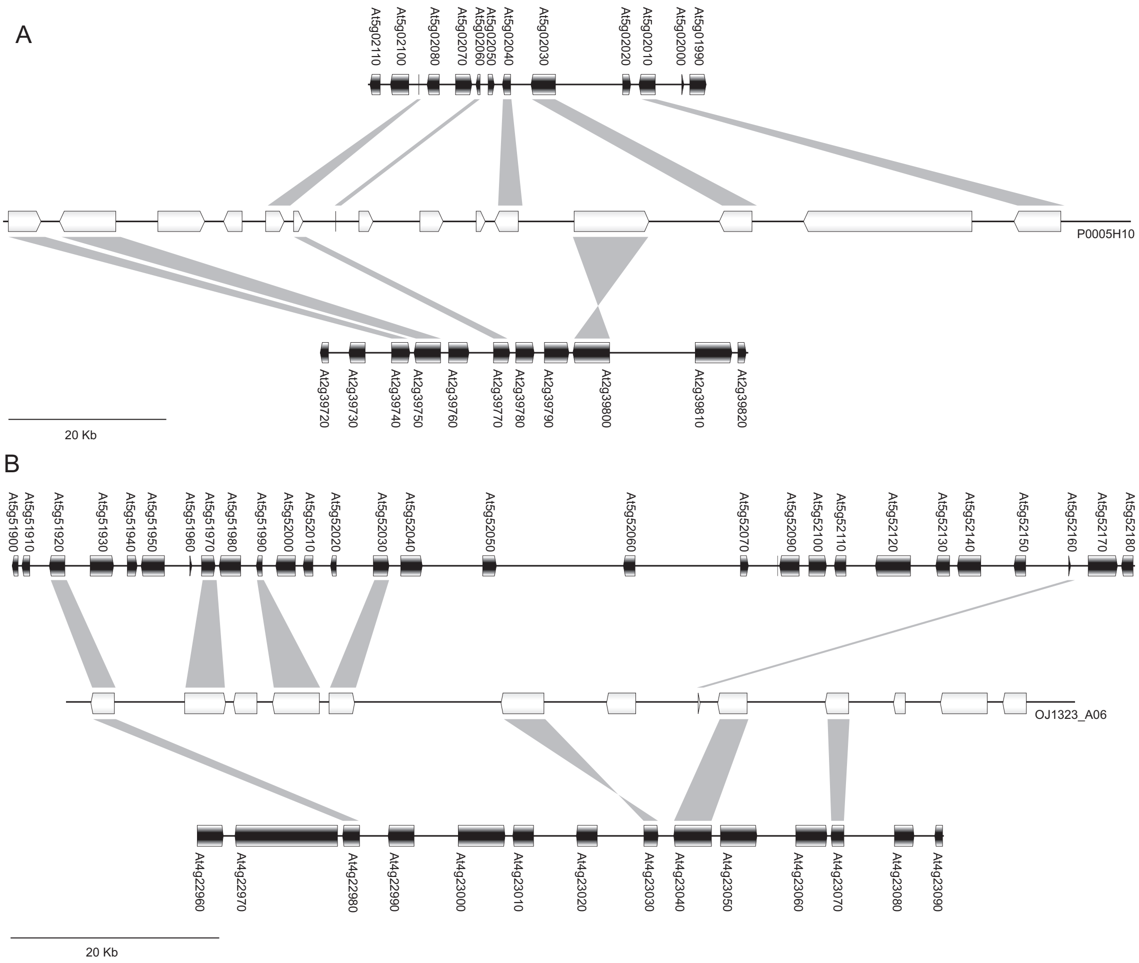


Figure 1.3.1 'Ghost' block duplications in the *Arabidopsis* genome. Homologous genes between *Arabidopsis* (black) and *Oryza sativa* (white) are indicated by grey bands. (a) Two genomic segments of *Arabidopsis*, on chromosomes 2 (top) and 5 (bottom), map to the same rice segment. Therefore, these segments are paralogous and result from a duplication event within the *Arabidopsis* genome. Because of differential gene loss, the duplicated *Arabidopsis* segments no longer have any paralogous genes in common. As a result, this duplication can not be detected anymore. (b) 'Ghost' block duplication between *Arabidopsis* chromosomes 4 (top) and 5 (bottom). One anchor point (i.e. the paralogous gene pair At5g51920 – At4g22980) is still present on both segments, but is insufficient to detect microcolinearity between the two segments.

This phenomenon, which we refer to as 'differential gene loss', turns the originally identical duplicated regions into two non-redundant sets of genes, divided over two distinct genome locations. Differential gene loss thus reduces the number of paralogs that can be identified by a within-genome comparison. For a few genes, both duplicates might have been retained, but in that case the number of anchor points is usually too small to detect significant colinearity when permutation tests are applied (Figure 1.3.1b). Therefore, the use of inter-genomic comparisons can help to recover block duplications that had seemingly disappeared.

By considering only a small amount of the rice genome sequence, we were able to detect several examples of such 'ghost' duplications in *Arabidopsis*. Once a completely assembled and well-annotated rice genome sequence is available, comparisons between rice and *Arabidopsis,* which diverged from one another ~200 million years ago (Wikström et al., 2001) will probably reveal many more of such regions. Furthermore, most probably, many other examples of such 'ghost' duplications are waiting to be discovered in other eukaryotic genomes as well.

- Part 2 -

Large-scale duplication events: key players in plant genome evolution

2.1 Investigating ancient duplication events in the *Arabidopsis* genome

Jeroen Raes[+], Klaas Vandepoele[+], Cedric Simillion, Yvan Saeys and
Yves Van de Peer

Redrafted from: *Journal of Structural and Functional Genomics* **3**, 117–129 (2003)

The complete genomic analysis of *Arabidopsis thaliana* has shown that a major fraction of the genome consists of paralogous genes that probably originated through one or more ancient large-scale gene or genome duplication events. However, the number and timing of these duplications still remains unclear, and several different hypotheses have been put forward recently. Here, we reanalyzed duplicated blocks found in the *Arabidopsis* genome described previously and determined their date of divergence based on silent substitution estimations between the paralogous genes and, where possible, by phylogenetic reconstruction. We show that methods based on averaging protein distances of heterogeneous classes of duplicated genes lead to unreliable conclusions and that a large fraction of blocks duplicated much more recently than assumed previously. We found clear evidence for one large-scale gene or even complete genome duplication event somewhere between 70 to 90 million years ago. Traces pointing to a much older (probably more than 200 million years) large-scale gene duplication event could be detected. However, for now it is impossible to conclude whether these old duplicates are the result of one or more large-scale gene duplication events.

[+] both authors contributed equally

Introduction

For over 30 years, geneticists, evolutionists and, more recently, developmental biologists have been debating on the number of genome duplications in the evolution of animal lineages and its impact on major evolutionary transitions and morphological novelties. Thanks to the recent progress made in gene mapping studies and large-scale genomic sequencing, the debate has been livelier than ever before. Indeed, huge amounts of sequence data have become available, amongst which the complete genome sequences of invertebrates, such as *Drosophila melanogaster*, *Caenorhabditis elegans*, and vertebrates, such as pufferfish and human, while others are being finalized. With these data at our disposition, we expect to address the ancient questions and hypotheses regarding genome duplications, as formulated by pioneers like J.B.S. Haldane (who already contemplated the benefits and evolutionary impact of polyploidy events in 1933) and S. Ohno. However, a great deal of controversy still exists on the prevalence of genome duplications in certain lineages. For example, the classic hypothesis of Ohno (1970) that at least one genome duplication occurred in the evolution of the vertebrates has not been evidenced yet. Several theories, which differ in the proposed number of duplications as well as in their timing, have been proposed, but without confirmation (Skrabanek and Wolfe, 1998; Hughes, 1999; Wolfe, 2001). More recently, a putatively ancient fish-specific genome duplication before the teleost radiation has been the subject of lively debate (Robinson-Rechavi et al., 2001; Taylor et al., 2001a, 2001b; Van de Peer et al., 2003). Given the already controversial nature of the occurrence and date of these genome duplications in vertebrates, their precise role in the evolution of new body plans (Holland, 1992) or in speciation (Lynch and Conery, 2001; Taylor et al., 2001c) remains even more speculative.

For plants, controversy about ancient genome duplications has long been nearly nonexisting. Polyploidy seems to have occurred frequently in plants. Up to 80% of angiosperms are estimated to be polyploid, with variation from tetraploidy (maize) and hexaploidy (wheat) to 80-ploidy (*Sedum suaveolens*) (for a review, see Leitch et al., 1997). Because of the complexity of many plant genomes and lack of sequence data, research on plant genome evolution was basically restricted to experimental techniques (Wendel, 2000) and, until very recently, few computational analyses had been performed to investigate the prevalence and timing of older large-scale duplications and their impact on plant evolution.

In 1996, the plant community decided to determine the complete genome sequence of *Arabidopsis thaliana*. This model plant was chosen because it has a small genome with a high gene density and seemed to be an “innocent” diploid. However, during and even before this huge enterprise, some indications were found that large-scale duplications had occurred (Kowalski et al., 1994; Paterson et al., 1996; Terryn et al, 1999; Lin et al., 1999; Mayer et al., 1999). After bacterial artificial chromosome sequences representing approximately 80% of the genome had been analyzed, almost 60% of the genome was found to contain duplicated genes and regions (Blanc et al., 2000). This phenomenon could only be explained by a complete genome duplication event, an opinion shared by the Arabidopsis Genome Initiative (2000). Previously, comparative studies of bacterial artificial chromosomes between *Arabidopsis* and soybean (Grant et al., 2000) and between *Arabidopsis* and tomato (Ku et al., 2000) had led to similar notions. In the latter study, two complete genome duplications were proposed: one 112 and another 180 million years ago (MYA). Vision et al. (2000) rejected the single genome duplication hypothesis by dating duplicated blocks through a molecular clock analysis. Several different age classes among the duplicated blocks were found, ranging from 50 to 220 MYA and at least four rounds of large-scale duplications were postulated. One of these classes, dated approximately 100 MYA, grouped nearly 50% of all the duplicated blocks, suggesting a complete genome duplication at that time (Vision et al., 2000). However, the dating methods used for these gene duplications were based on averaging evolutionary rates of different proteins, which was later criticized because of their high sensitivity to rate differences (Sankoff, 2001; Wolfe, 2001). Because the same methodology was also used by Ku *et al*. (2000), their results should also be considered with caution. On the other hand, Vision et al. (2000) discovered overlapping blocks, a phenomenon that can be explained only by multiple duplication events. Neither Blanc et al. (2000) nor the Arabidopsis Genome Initiative (2000) detected these overlapping blocks.

Using a different method of dating based on the substitution rate of silent substitutions, Lynch and Conery (2000) discovered that most *Arabidopsis* genes had duplicated approximately 65 MYA, which brings us back to a single polyploidy event. However, no duplicated blocks of genes, but only paralogous gene pairs were taken into account. Apparently, the evolutionary history of the first fully sequenced plant seems a lot more complex than originally expected. There is no clear answer on whether one single or multiple polyploidy events took place nor when they occurred. The results of the different analyses seem to be highly dependent of the methods used. For this reason, we reinvestigated the ancient large-scale gene duplications described by Vision et al. (2000) by applying two alternative dating methodologies on several of the more anciently duplicated blocks found in their study. Furthermore, we compared the results obtained to pinpoint the strengths and weaknesses of the methodology used in the two studies.

Materials and methods

Strategy

The original goal was to reinvestigate whether one or several ancient large-scale gene duplication(s) had occurred in the evolution of *Arabidopsis thaliana*. Furthermore, because Vision et al. (2000) dated one of the large-scale duplication events as approximately 200 million years old, we were curious to see whether this event pre- or postdated the monocot-dicot split, which is estimated to have occurred at about that time: 170-235 MYA (Yang et al., 1999) and 143-161 MYA (Wikström et al., 2001). We focused on the blocks that according to Vision et al. (2000), originated during this ancient round of duplication and consisted of six regions in the genome (class F). We mapped these regions to a more up-to-date data set (see below) and subjected them to two dating methodologies: dating based on synonymous substitution rates and molecular phylogeny. The former was done with three different approaches to estimate synonymous substitution rates, namely those of Li (1993), of Nei and Gojobori (1986) and of Yang and Nielsen (2000). Molecular phylogeny-based dating was performed through the construction of evolutionary trees by the Neighbor-joining method (Saitou and Nei, 1987). By using these different approaches, the possibility of drawing wrong conclusions caused by weaknesses of one particular method is minimized.

However, during the course of this study, it became clear that the most ancient blocks described by Vision et al. (2000) contained genes that had duplicated much more recently. Because the dating methodology of Vision et al. (2000) had been criticized before (Sankoff, 2001; Wolfe, 2001), we subsequently focused on two sets of 10 blocks of two younger age classes, D and E, estimated to be 140 and 170 million years old, respectively. These data sets were chosen in such a way that they represented a wide distribution in block size (number of anchor points) as well as amino acid substitution rate (dA) within each age class.

Data set of duplicated genes

From the complete set of segmentally duplicated blocks defined by Vision et al. (2000) that consisted of 103 regions with seven or more duplicated genes, we analyzed selected blocks covering the three oldest classes. This selection consisted of all six blocks from class F (200 million years old), 10 from class E (170 million years old) and 10 from class D (140 million years old). Because the original data set (i.e. the chromosomal DNA sequences) represented a preliminary version of the *Arabidopsis* genome sequence (incomplete and not always correctly assembled), the positions of these duplicated blocks were transferred to a data set that had been built recently. This new data set consisted of a genome-wide non-redundant collection of *Arabidopsis* protein-encoding genes, which were predicted with GeneMark.hmm (Lukashin and Borodvsky, 1998; genome version of January 18th, 2000 (v180101), downloaded from the Institute for Protein Sequences center Martiensried, Germany; ftp://ftpmips.gsf.de/cress/). In addition to the protein sequence, the position and orientation of the genes within the *Arabidopsis* genome were determined.

Within this protein set, all pairs of homologous gene products between two chromosomes were determined and the results stored in a matrix of (m, n) elements (m and n being the total number of genes on a certain chromosome). Two proteins were considered as homologous if they had an E-value < 1e-50 within a BLASTP (Altschul et al., 1997) sequence similarity search (Friedman and Hughes, 2001).

The synchronization of our data set with the blocks detected by Vision et al. (2000) was done using their supplementary data (website: http://www.igd.cornell.edu/~tvision/arab/science_supplement.html). Initially, for a set of anchor points (i.e. pairs of duplicated genes), defining a duplicated block (Vision et al., 2000), the corresponding protein couples were detected in our data set and then these protein couples were localized in the matrix. To check whether these proteins were indeed part of a segmentally duplicated block, an automatic and manual detection was performed. The automatic detection was done with a new tool (Vandepoele et al., 2002a), primarily based on discovering clusters of diagonally organized elements (representing duplicated blocks) within the matrix of homologous gene products. Similar to the strategy of Vision et al. (2000), tandem repeats were remapped before defining a duplicated block. An overview of blocks analyzed in this study, together with the number of anchor points per block, is presented in Table 2.1.1.

Dating based on K_s

Blocks of duplicated genes were dated using the NTALIGN program in the NTDIFFS software package (Conery and Lynch, 2001). This package first aligns the DNA sequence of two mRNAs based on their corresponding protein alignment and then calculates K_s by the method of Li (1993). We calculated K_s also with two alternative dating methodologies (Nei and Gojobori, 1986; Yang and Nielsen, 2000) based on the same alignments. These two methods are implemented in the PAML phylogenetic analysis package (Yang, 1997). The time since duplication was calculated as $T=K_s/2\lambda$, with λ being the mean rate of synonymous substitution; in *Arabidopsis* the estimation is $\lambda=6.1$ synonymous subsitutions per 10^9 years (Lynch and Conery, 2000). The mean K_s value (average of the estimates obtained by the three methods) for each block was derived for each duplicated pair. These values were then used to calculate the mean K_s for each block, excluding outliers using the Grubbs test (Grubbs, 1969; Stefansky, 1972) with a 99% confidence interval.

Phylogenetic analysis

The public databases (PIR, GenBank/EMBL/DDBJ, Swiss-PROT) were scanned for homologues of the anchor points using BLASTP (Altschul et al., 1997). When homologues were found in other species next to the *Arabidopsis* paralogues, the gene family was selected for phylogenetic analysis. Protein sequences were subsequently aligned with CLUSTAL W (Thompson et al., 1994). Duplicates or sequences that were too short were removed from the data set. After manual optimization of the alignment and reformatting using BioEdit (Hall, 1999) and ForCon (Raes and Van de Peer, 1999), the more conserved positions of the alignment were subjected to phylogenetic analysis. Trees were constructed based on Poisson or Kimura distances using the Neighbor-joining algorithm as implemented in the TREECON package (Van de Peer and De Wachter, 1997).

Supplementary data such as sequences, accession numbers, alignments, and trees can be obtained from the authors upon request.

Results

Dating based on K_s

In contrast to mutations that result in amino acid changes (nonsynonymous substitutions), silent or synonymous substitutions do not affect the biochemical properties of the protein. As such they are generally believed not to be subjected to natural selection and, consequently, to evolve in a (nearly) neutral, clock-like way (Li, 1997). Absolute dating based on synonymous substitution rates (K_s) should be more accurate than dating based on the estimation of genetic distances between duplicated protein sequences. However, because of rapid saturation of synonymous sites, dates of older (K_s > 1) divergences/duplications will become unreliable (Li, 1997). We calculated Ks values with three different methods for all pairs of duplicated genes in 26 old blocks (classes D, E, and F, estimated to have originated between 140 and 200 MYA; Vision et al., 2000). From these values we calculated the duplication date of each block. The results of this analysis are given in Table 2.1.1.

Interestingly, several block duplications were dated to be much younger than what was found by Vision et al. (2000). For example, a duplication between chromosome 1 and 5, denoted as block 37 and based on 11 gene pairs (17 in our study; Table 1), was found to have occurred 72 MYA, and not 200 MYA. The distribution of the K_s values of the duplicated pairs in this block, calculated with the three different methods, confirmed our hypothesis that this is a younger block. With only a few exceptions, almost all duplicated pairs seemed to have K_s values between 0.5 and 1 synonymous substitutions per synonymous site, and this for the three methods used (Figure 2.1.1). For three pairs of genes within the duplicated block, the situation is less clear (Figure 2.1.1). No results were obtained with the method of Li (1993), probably because the duplicated gene sequences are too divergent to calculate a K_s value using this method, whereas the two other methods gave extremely high or no K_s values. One possible explanation is a higher synonymous mutation rate specific for these genes, because fluctuations in K_s have been reported before (Li, 1997; Zeng et al., 1998). Another possible explanation could be that these genes originated earlier than the other genes in that block and that the situation observed is due to differential deletions of alternate members of duplicated tandem pairs (Friedman and Hughes, 2001). For this reason, these gene pairs were not included in the calculation of the duplication date of the whole block (see Materials and methods).

Table 2.1.1 Re-analysis of the duplicated blocks as described by Vision *et al*. (2000)

Vision *et al*. (2000)							This study					
Block number	Chr1[a]	Chr2[a]	Anchor points [b]	dA	Age class	Age (MY)	Anchor points[b]	K_s [c]	K_s [d]	K_s [e]	Mean age[f]	Std Dev
15	1	3	7	0.8975	F	200	7	1.8641	2.5378	2.1679	213	92
25	1	5	7	0.8012	F	200	6	1.6757	1.7008	2.5515	160	27
37	1	5	11	0.8146	F	200	17	0.8386	0.8138	0.9698	72	19
39	1	3	8	0.8375	F	200	7	1.6053	1.9744	1.8768	170	62
57	2	3	7	0.8521	F	200	7	2.9251	3.2702	2.4395	269	64
59	2	5	15	0.8473	F	200	18	1.8078	2.3744	2.0642	191	70
34	1	5	23	0.7165	E	170	27	0.8723	0.8308	0.8900	71	18
71	3	5	31	0.6814	E	170	70	0.7933	0.8262	0.8312	67	19
100	4	5	20	0.6899	E	170	15	1.8656	1.9727	2.1682	170	45
78	3	5	26	0.701	E	170	35	0.7382	0.7551	0.8475	64	11
47	2	5	8	0.7397	E	170	8	1.8475	3.0169	2.1072	218	87
16	1	3	8	0.6562	E	170	7	0.8390	0.8536	1.0224	74	19
55	2	5	14	0.685	E	170	9	1.7585	2.0966	1.8341	162	32
9	1	3	24	0.6947	E	170	20	0.9098	0.9966	1.1350	83	20
87	3	4	11	0.7231	E	170	8	1.6049	1.8936	2.1889	164	67
48	2	3	11	0.7045	E	170	8	1.7175	1.9716	2.0465	162	56
6	1	5	30	0.6106	D	140	30	0.7754	0.8138	0.9228	69	17
30	1	3	92	0.5262	D	140	106	0.8047	0.8325	0.9668	71	20
95	4	5	88	0.5592	D	140	61	0.7337	0.7884	0.8707	65	10
17	1	1	153	0.5684	D	140	167	0.8110	0.8175	0.8983	69	18
92	4	5	97	0.6064	D	140	107	0.8741	0.8849	1.0507	77	25
33	1	4	18	0.5381	D	140	11	1.6283	1.6707	1.5669	133	26
5	1	4	13	0.5631	D	140	6	1.5232	1.5657	1.5324	126	16
73	3	5	26	0.5855	D	140	25	0.7965	0.8187	0.9105	69	15
93	4	5	42	0.6263	D	140	28	0.7719	0.8174	0.9010	68	16
26	1	4	35	0.5273	D	140	42	0.8719	0.8946	1.0867	78	23

[a] Chromosome numbers on which the two duplicated blocks are found.

[b] Number of anchor points in blocks detected in this study.

[c] K_s values calculated according to Li (1993).

[d] K_s values calculated according to Nei and Gojobori (1986).

[e] K_s values calculated according to Yang and Nielsen (2000).

[f] Mean age (in MY) of the block was derived from the mean K_s, excluding outliers (see Materials and Methods).

However, most blocks of age class F had significantly higher K_s values and consequently older divergence dates, which indeed points to a more ancient large-scale duplication event. This observation was strengthened by the fact that, with a few exceptions, duplicated blocks of this age class had less anchor points (Table 2.1.1) and K_s values seemed to fluctuate more between members of the same block (see, for example, the distribution of block 59, estimated to have duplicated approximately 190 MYA; Figure 2.1.2). The latter is probably due to saturation of

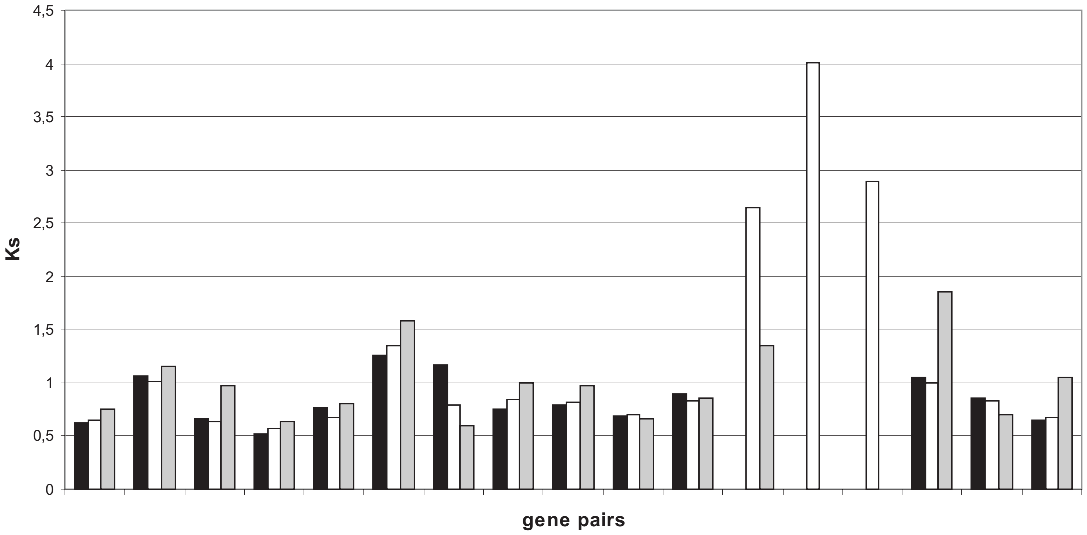


Figure 2.1.1 Distribution of K_s values for duplicated genes as found in block 37, and calculated with the methods of Li (black bars), Nei and Gojobori (white bars) and Yang and Nielsen (grey bars).

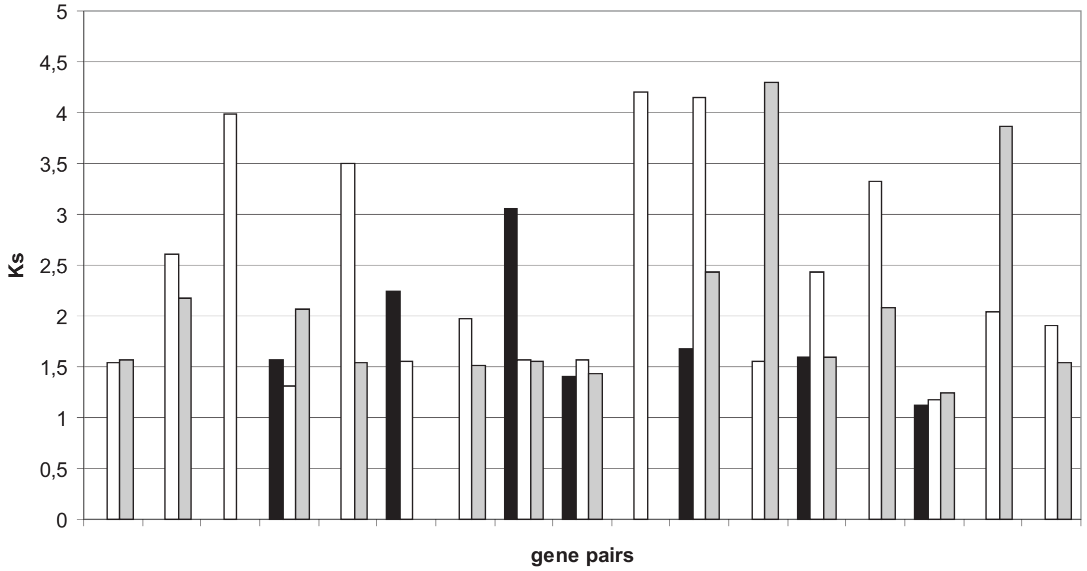


Figure 2.1.2 Distribution of K_s values for duplicated genes found in block 59, and calculated with the methods of Li (black bars), Nei and Gojobori (white bars) and Yang and Nielsen (grey bars).

synonymous substitutions, by which larger errors in K_s estimation are introduced, causing values of $K_s > 1$ to be unreliable.

In our evaluation of class E blocks (170 MYA; Vision et al., 2000), the situation is even more peculiar. From the 10 blocks we selected, a large part again seemed to be much younger than what was derived based on dA values. Five out of 10 blocks seemingly originated only approximately 70 MYA, less than half the age calculated by Vision et al. (2000). Here also, the distribution of K_s values clearly showed that a large majority of duplicated pairs in these blocks belonged to the same, much younger, age class, with only a few exceptions (data not shown). However, the other half of the 10 selected blocks seem to be older.

In the class D sample, dated 140 10^6 years old by Vision et al. (2000), 8 out of 10 blocks seemed to have duplicated approximately 70 MYA. The distribution of K_s values within one block again gave similar results as above: most pairs had K_s values between 0.5 and 1, with a minor fraction of exceptions (data not shown). Although only a subset of the complete set of duplicated blocks of age classes D and E were analyzed, many blocks appeared to be much younger than proposed by Vision and et al. (2000). Preliminary results of a more rigorous analysis seem to confirm our findings (unpublished results).

Dating by phylogenetic analysis

Absolute dating methods based on substitution numbers per site are very useful in high-throughput analyses, such as those by Lynch and Conery (2000) and Vision et al. (2000), but they have some serious drawbacks. Inferred divergence dates based on amino acid substitutions are not as quickly underestimated due to saturation, although saturation at the amino acid level has been demonstrated (Van de Peer et al., 2002). However, when using this technique, there is a serious risk of overestimating the age of more rapidly evolving blocks, or underestimating the age of blocks containing more slowly evolving proteins. The use of synonymous mutation rates is probably favourable because these positions evolve at nearly neutral rates and, so, give a more reliable estimate in the case of fast or slowly evolving genes. Unfortunately, these analyses are compromised for older duplications because of the rapid saturation of these sites.

To validate the results, an alternative technique was applied, namely relative dating using phylogenetic methods. If a duplication occurred before the monocot-dicot split, this could be proven by a tree topology (Figure 2.1.3a), in which the two dicot members of a gene family each group with a monocot sequence. If, however, the two *Arabidopsis* duplicates originated more recently, i.e. after the dicot-monocot split, the two dicot branches should be sister sequences, outgrouped by their monocot orthologue (Figure 2.1.3b). Even if certain sequences are still missing from the databases (because of gene loss or nondetection), conclusions

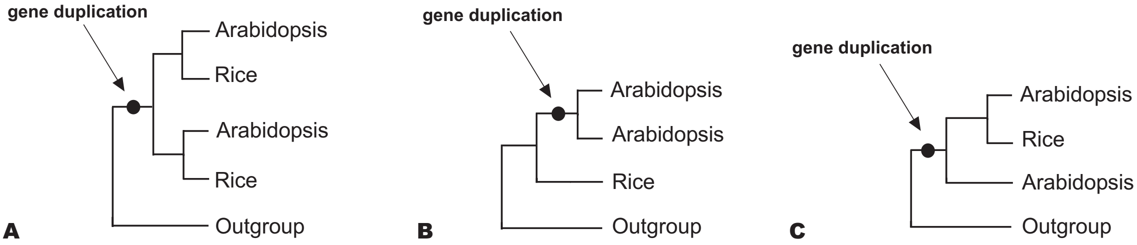


Figure 2.1.3 (a) Expected tree topology for genes formed by a gene/genome duplication event prior to the split of monocots and dicots. (b) Expected tree topology for genes formed by a gene/genome duplication event that occurred after the split of monocots and dicots and specific to *Arabidopsis*. (c) Even if only one of the paralogues is known, due to gene loss or absence in the databases, the gene duplication can be inferred.

can be drawn. For example, the tree topology presented in Figure 2.1.3c could only be explained by a duplication that occurred before the monocot-dicot split.

For all the anchor points of the oldest blocks (F), we searched the protein databases for homologues in other plant species to construct evolutionary trees. Unfortunately, it was impossible to construct trees for many of the duplicated genes, the main reason being the absence of homologues from plant species other than *Arabidopsis* in the databases. Furthermore, the sequences often contained too few conserved positions to get statistically significant results (i.e. high bootstrap values). An overview of constructed trees and conclusions is presented in Table 2.2.2. Gene families for which no homologues from other species than *Arabidopsis thaliana* could be found in the databases are not shown.

Although we could not draw conclusions on many of the genes/blocks, we would like to consider some of the constructed trees. A first interesting result was obtained from the analysis of the gluthatione synthase gene family; it has two members on chromosomes 1 and 5 that are part of block 37, which is a duplicated block of class F (200 MYA; Vision *et al*., 2000); but, according to our estimation, it had duplicated approximately 72 MYA. The tree topology (Figure 2.1.4) for this family clearly showed that the duplication that yielded the two duplicates occurred before the divergence of *Arabidopsis* and *Brassica*, but after the split between Asteridae and Rosidae. In consequence, the duplication between these two genes must have happened between 15-20 (Yang et al., 1999; Koch et al, 2001) and 135 MYA (the latter value being the mean of two estimations, 112-156 MYA [Yang et al., 1999]) and 114-125 MYA [Wikström et al., 2001]), which is in accordance with our findings for this block.

Table 2.1.2 Gene families selected for phylogenetic analysis for each paralogous block, belonging to age class F (Vision et al., 2000; 200 MYA)

Block[a]	Family[b]	Sites[c]	Conclusion	Reason
15	Unknown	279	None	No statistical support
25	-	None	No trees possible due to the absence of sequences from other species	
37	Calmodulin	105	None	No statistical support
	Calmodulin-like	112	Probably younger than the split between eurosids I and eurosids II	Genetic distance
	Glutamine synthase	314	Younger than the split with *asteridae* and older than the *Arabidopsis*-*Brassica* divergence (see Figure 2.1.3)	Topology with statistical support
39	Unknown	287	None	Too few monocot sequences for this family
57	DOF Zinc-finger	85	None	Highly inequal rates of evolution between duplicates
	GATA transcription factor	148	Older than the monocot-dicot split (see Figure 2.1.4)	Topology with statistical support
	Apetala 2	81	None	No statistical support
	Expansin	180	None	No statistical support
59	Protein phosphatase 2C	174	None	Too few monocot sequences available
	Putative Rab5 interacting protein	100	Probably younger than the monocot-dicot split	Genetic distance
	Cyclophilin	141	None	No statistical support
	Phosphoprotein phosphatase 1	305	None	No statistical support
	Apetala 2 (see also B57)	81	None	No statistical support

[a] Block number as defined by Vision et al. (2000).

[b] Name of the family analyzed, as far as could be deduced from the description line of the entries.

[c] Length of sequence alignment used for tree construction.

A second tree of interest is that of the GATA transcription factor family with a pair of duplicates on chromosomes 2 and 3 that belong to block 57, also of age class F. It was very hard to date this block with our dating methods, because the sequences were apparently saturated for synonymous substitutions. However, all K_s values calculated for pairs in this block were above 2.2 synonymous substitutions per synonymous site (see Table 2.1.1), suggesting that this block is genuinely old. When we investigated the topology of the GATA family (Figure 2.1.5), we observed a topology similar to that described in Figure 2.1.3c: although there is only one monocot sequence, this topology could be only explained if the duplication

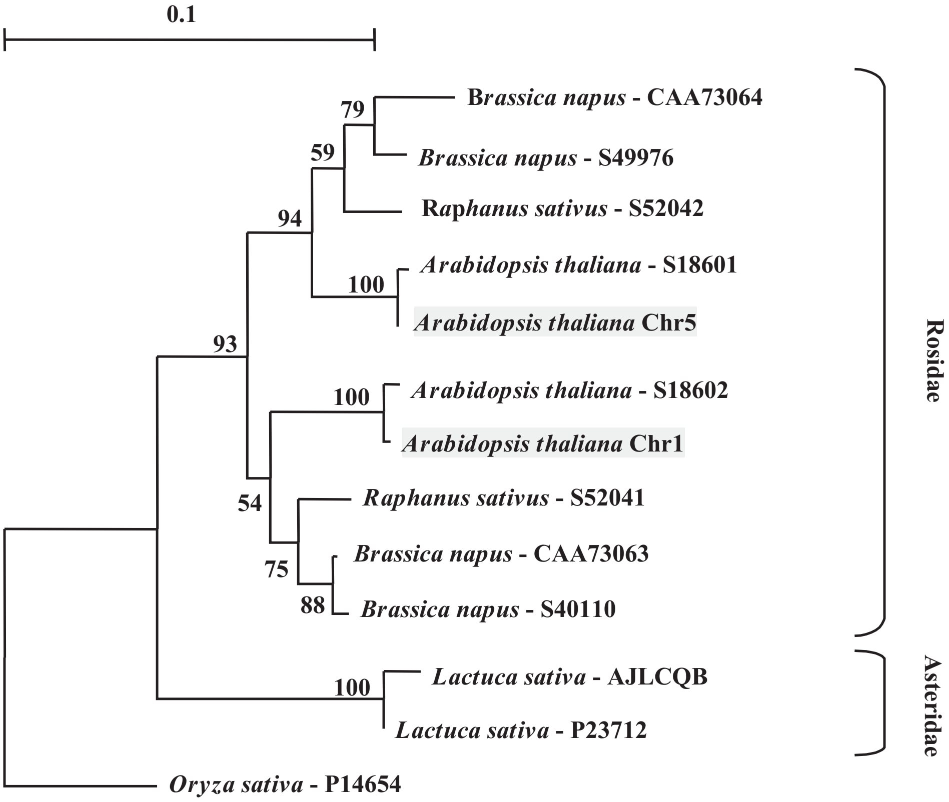

0.1
79
59
94
100
93
100
54
75
88
100
Brassica napus - CAA73064
Brassica napus - S49976
Raphanus sativus - S52042
Arabidopsis thaliana - S18601
Arabidopsis thaliana Chr5
Arabidopsis thaliana - S18602
Arabidopsis thaliana Chr1
Raphanus sativus - S52041
Brassica napus - CAA73063
Brassica napus - S40110
Lactuca sativa - AJLCQB
Lactuca sativa - P23712
Oryza sativa - P14654
Rosidae
Asteridae

Figure 2.1.4 Neighbor-joining tree of the glutamine synthase family, inferred from Poisson-corrected evolutionary distances. Sequences that belong to the analyzed duplicated blocks are indicated with their chromosome number. Bootstrap values (above 50%) are shown in percentages at the internodes. Scale = evolutionary distance in substitutions per amino acid.

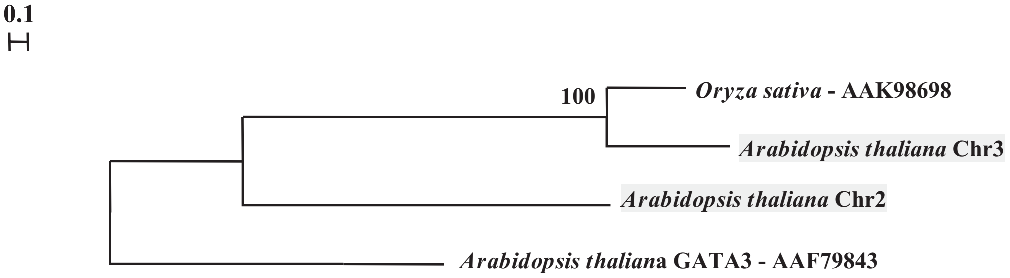

0.1
100
Oryza sativa - AAK98698
Arabidopsis thaliana Chr3
Arabidopsis thaliana Chr2
Arabidopsis thaliana GATA3 - AAF79843

Figure 2.1.5 Neighbor-joining tree of the GATA family of transcription factors, inferred from Poisson corrected evolutionary distances. Sequences that belong to the analyzed duplicated blocks are indicated by their chromosome number. Bootstrap values (above 50%) are shown in percentages at the internodes. Scale = evolutionary distance in substitutions per amino acid.

that gave rise to the two *Arabidopsis* genes occurred before the monocot-dicot split. This would mean that this block occurred at least 190 MYA (Yang et al., 1999; Wilkström et al., 2001).

In some cases, evolutionary distances can be informative of duplication dates. As illustration, an example from the age class D (140 MYA; Vision et al., 2000) is

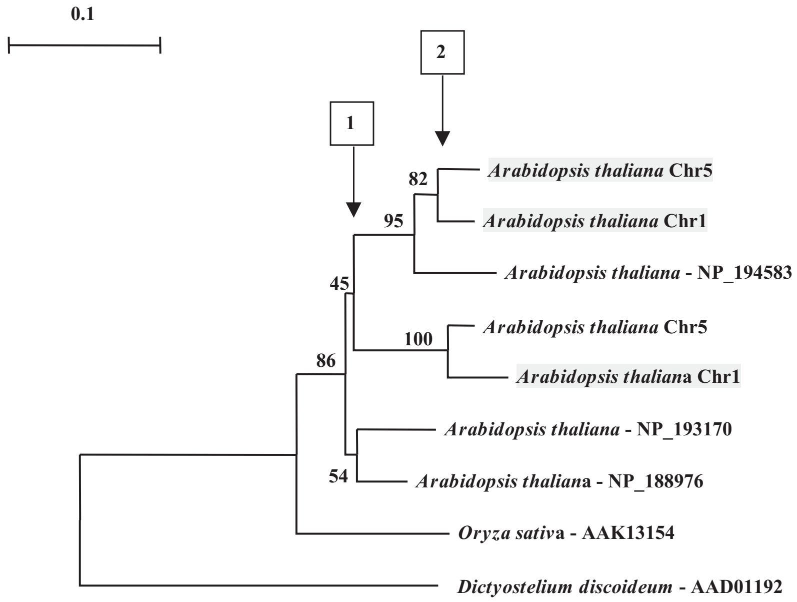

0.1
2
1
82
95
45
100
86
54
Arabidopsis thaliana Chr5
Arabidopsis thaliana Chr1
Arabidopsis thaliana - NP_194583
Arabidopsis thaliana Chr5
Arabidopsis thaliana Chr1
Arabidopsis thaliana - NP_193170
Arabidopsis thaliana - NP_188976
Oryza sativa - AAK13154
Dictyostelium discoideum - AAD01192

Figure 2.1.6 Neighbor-joining tree of the casein kinase family, using Poisson correction for evolutionary distance calculation. Sequences that belong to the analyzed duplicated blocks are indicated by their chromosome number. Arrows indicate (1) a tandem duplication and (2) the block duplication. Bootstrap values (above 50%) are shown in percentages at the internodes. Scale= evolutionary distance in substitutions per amino acid.

given. Figure 2.1.6 shows the topology of the casein kinase gene family that has two members on both chromosomes 1 and 5, all four of them belonging to the same duplicated block 6. Using K_s-based dating, we determined that this block had duplicated approximately 70 MYA, with approximately 80% of the K_s values in this block being smaller than 1. As can be seen from the tree topology, the two members of block 6 first originated (probably) through tandem duplication (arrow 1) and then through a larger-scale duplication including the other members of that block (arrow 2). Both these events happened after the monocot-dicot split, as can be derived from the fact that the group containing these four proteins is outgrouped by a rice sequence. The evolutionary distance from each of the duplicates to the block duplication point is approximately 0.025 amino acid substitutions per site, whereas the evolutionary distance between the genes originating by tandem duplication is approximately 0.158 amino acid substitutions per site. The average evolutionary distance between the sequences of rice and *Arabidopsis* is approximately 0.206 amino acid substitutions per site, meaning that, if a divergence date for monocots and dicots of 190 MYA (Yang et al., 1999; Wilkström et al., 2001) and a molecular clock-like evolution of this protein were assumed, the block duplication would have happened somewhere 46 MYA (with

λ=K/2T=0.206 substitutions per site/380 MY=5.42 10^{-4} substitutions per site/ MY). This value is much closer to our estimation based on Ks than that of 140 MYA obtained by Vision et al. (2000).

Discussion

Currently, three different methods to date gene duplication events are generally used: absolute dating based on synonymous substitution rates, absolute dating based on nonsynonymous substitution rates or protein-based distances, and relative dating through the construction of phylogenetic trees. Here, we provide some evidence that protein distances are not very reliable for large-scale dating of heterogeneous classes of proteins. For example, classes containing blocks of the same age based on mean protein distance (classes D, E, and F; Vision et al., 2000) seem to be very heterogeneous in age when dating is based on synonymous substitution rates. Protein-based distances are known to vary considerably among proteins (e.g. Easteal and Collet, 1994); therefore, duplicated blocks that contain a larger fraction of fast-evolving genes will have a relatively high mean protein distance between the paralogous regions and appear older than they actually are. In our opinion, the use of synonymous and, consequently, neutral substitutions for evolutionary distance calculations is more reliable. However, there is one important caveat: dating based on silent substitutions can only be applied when $K_s < 1$. A $K_s > 1$ points to saturation of synonymous sites and can no longer be used to draw any reliable conclusions regarding the origin of duplicated genes or blocks. In this case, a solution could be relative dating with phylogenetic means. Although the dating is rather crude, it offers a way of determining duplication dates relative to known divergences. The main problem here, however, is the availability of plant sequence data. Only a few duplicated pairs had enough orthologues in the public databases to allow any conclusions to be drawn. Furthermore, if orthologues would be found, the sequences may not be very suitable for phylogenetic analysis. Consequently, it seems that phylogenetic inference cannot yet be as widely applied to plant as to animal genomes (e.g., Wang and Gu, 2000; Friedman and Hughes, 2001; Van de Peer et al., 2001). However, as soon as more sequence data from key species such as mosses, ferns, and monocots, become available, this approach may become more useful.

From the three oldest age classes defined by Vision et al. (2000), only one (F) seems to contain many old duplicated blocks, whereas several blocks of the two other age classes have seemingly been duplicated approximately 70-90 MYA. In our opinion, the hypothesis of Vision et al. (2000) that at least four large-scale duplications have occurred is far from being proven. In contrast with the multimodal distribution of large-scale gene duplication, our results show that a major fraction of blocks has duplicated approximately at the same time and has probably originated by a complete genome duplication. On the other hand, a fraction of block duplications seems much older than the others. Unfortunately, because synonymous sites were saturated and trees were not reliable enough, these duplications could not be dated more accurately. Although these old duplicated blocks are scattered throughout the genome (Table 2.1.1), it is hard to prove that they are the result of a single duplication event.

The question whether large-scale gene duplications have occurred before the divergence of monocots and dicots still remains to be answered. Some of these events are probably anterior to the monocotyl-dicotyl split, as suggested by the GATA transcription factor topology (Figure 2.1.5). Large-scale gene duplication events prior to the monocot-dicot split may have led to the origin of flowering or even of seed plants: duplications of (sets of) developmentally important genes could have given the opportunity to develop new reproductive organs and strategies and consequently cause reproductive isolation, which may have resulted in speciation. The ongoing accumulation of sequence data delivered by several plant expressed sequence tags and genome sequencing projects will provide the means to answer the questions regarding the prevalence and timing of gen(om)e duplications in the evolution of plants and will hopefully help elucidating the role of these events in the diversification and evolution of plant species.

2.2 The hidden duplication past of *Arabidopsis thaliana*

Cedric Simillion, Klaas Vandepoele, Marc C. E. Van Montagu, Marc Zabeau, and Yves Van de Peer

Redrafted from: *Proc. Natl. Acad. Sci. U S A* **99,** 13627-32 (2002)

Analysis of the genome sequence of *Arabidopsis thaliana* shows that this genome, like that of many other eukaryotic organisms, has undergone large-scale gene duplications or even duplications of the entire genome. However, the high frequency of gene loss after duplication events reduces colinearity and therefore the chance of finding duplicated regions that, at the extreme, no longer share homologous genes. In this study we show that heavily degenerated block duplications that can no longer be recognized by directly comparing two segments because of differential gene loss, can still be detected through indirect comparison with other segments. When these so-called hidden duplications in *Arabidopsis* are taken into account, many homologous genomic regions can be found in five to eight copies. This finding strongly implies that *Arabidopsis* has undergone three, but probably no more, rounds of genome duplications. Therefore, adding such hidden blocks to the duplication landscape of *Arabidopsis* sheds light on the number of polyploidy events that this model plant genome has undergone in its evolutionary past.

Introduction

In 1996, when the research plant community decided to determine the genome sequence of the flowering plant *Arabidopsis thaliana*, few people suspected that this model plant organism is an ancient polyploid. Nevertheless, even before the completion of the genome sequence, it was clear that a large portion of its genome consists of duplicated segments (Terryn et al., 1999). After analysis of bacterial artificial chromosome sequences, representing ~80% of the genome, almost 60% was found to contain duplicated genes and regions (Blanc et al., 2000), which strongly suggested a large-scale gene or even entire genome duplication event in the evolutionary history of *Arabidopsis.* This opinion was later shared by the Arabidopsis Genome Initiative, based on the complete genome sequence (Arabidopsis Genome Initiative, 2000), and by Lynch and Conery (Lynch and Conery, 2000), who discovered that most *Arabidopsis* genes had duplicated approximately 65 million years ago (MYA), by using a dating method based on the rate of silent substitutions. Comparative studies between *Arabidopsis* and soybean (Grant et al., 2000) and between *Arabidopsis* and tomato (Ku et al., 2000) also suggested that one or more large-scale gene or genome duplications had occurred. For example, in the latter study, two complete genome duplications were proposed, namely one 112 MYA and another 180 MYA, based on the presence of chromosomal segments that seemed to have been duplicated multiple times. The analysis of duplicated regions by the Arabidopsis Genome Initiative (Arabidopsis Genome Initiative, 2000) did not reveal such segments. Vision *et al.* (2000) also rejected the single-genome duplication hypothesis and postulated at least four rounds of large-scale duplications, ranging from 50 to 220 MYA. One of the age classes of duplicated blocks they defined (~100 MYA) grouped nearly 50% of all of the duplicated blocks, strongly suggesting a complete genome duplication at that time (Vision et al., 2000). However, the dating methods applied in their study have been criticized (Wolfe, 2001). A recent reanalysis of the duplicated blocks ascribed to different age classes, conducted by Raes et al. (2002), indeed revealed that many of the ancient blocks described by Vision et al. (2000) had a much more recent origin than was initially postulated. It is clear that the discussion regarding the number and time of origin of large-scale duplications in *Arabidopsis* is far from settled, partly because obtaining a complete picture of all duplications (and their dating) that have occurred in the evolution of a genome is not self-evident. Although the frequency of gene preservation over a large evolutionary period after duplication is unexpectedly high, and several models have been recently put forward to explain the retention of duplicates (Gibson and Spring, 1998; Lynch and Force, 2000; Wagner, 2002), the most likely fate of a gene duplicate is nonfunctionalization and, consequently, gene loss (Lynch and Conery, 2000). This observation has great consequences for the detection of duplicated regions in genomes. Identifying duplicated chromosomal regions is usually based on a within-genome comparison that aims at delineating colinear regions (regions of conserved gene content and order) in different parts of the genome. In general, one tries to identify duplicated blocks of homologous genes that are statistically valid, i.e., that are shown not to have been generated by chance. The statistics that determine colinearity usually depend on two factors, namely the number of pairs of genes that still can be identified as homologous (usually referred to as anchor points), and the distance over which these gene pairs are found, which usually depends on the number of “single” genes that interrupt colinearity (Gaut, 2001; Vandepoele et al., 2002a). However, the high level of gene loss, together with phenomena such as translocations and chromosomal rearrangements, often renders it very difficult to find (statistically significant) paralogous regions in the genome, in particular when the duplication events are ancient (Ku et al., 2000: Gaut, 2001). In this study we show that heavily degenerated block duplications that cannot be observed by directly comparing the two segments because of extreme differential gene loss (Vandepoele et al., 2002b) can still be detected through the indirect comparison with other segments. We refer to this previously undescribed class of block duplications as hidden block duplications, as opposed to non-hidden block duplications. Adding these hidden block duplications to the global duplication landscape of *Arabidopsis thaliana* sheds more light on the number of large-scale gene duplications that this genome has undergone in its evolutionary past.

Materials and methods

Arabidopsis dataset

We retrieved the TIGR annotation of the *A. thaliana* genome (version of August 2001) and extracted the coding sequences (CDS), corresponding amino acid sequences, and the relative position and strand orientation for a total of 25,439 protein-encoding genes. For 50 genes, the translation of the annotated mRNA sequence did not correspond with the protein sequence because exons were removed from or added to the annotated mRNA sequence. In this case the mRNA sequence was corrected manually. Within this set of protein encoding genes, we identified genes that are likely to be retrotransposons by conducting a BLASTP search (Altschul et al., 1990) against a set of known retrotransposable elements retrieved from SWISSPROT (Bairoch and Apweiler, 2000). For each BLAST-hit we calculated the percent identity and removed all genes (i.e., 257 in total) from the dataset for which this was ≥30%.

Detection of block (non-hidden) duplications and tandem repeats

The detection of tandem and block duplications within the genome of *Arabidopsis* was done with ADHoRe. Because this tool is extensively discussed elsewhere (Vandepoele et al., 2002a), we shall only briefly describe it here. The ADHoRe algorithm performs a pairwise comparison of two genomic fragments (typically chromosomes) by comparing two lists of all protein-encoding genes (and their orientation) sorted in the order in which they are present on these fragments. By comparing all protein-coding genes of both fragments, the program identifies all homologous gene pairs. This information is then stored in a matrix of (*m* x *n*) elements (*m* and *n* being the length of the submitted gene lists) in which each nonzero element (*x*, *y*) is a pair of homologous genes, also called an anchor point (*x* and *y* denote the coordinates of both genes in their respective gene lists). We call this matrix the gene homology matrix. The value of a nonzero element is positive or negative, depending on whether the genes in every pair detected have the same strand orientation or do not, respectively. In this study, we performed pairwise comparisons between all five chromosomes of *Arabidopsis*, by using the annotation as described above. Once this matrix is compiled, block duplications can be easily identified as a diagonal series of anchor points (non-zero elements in the matrix), whereas tandem repeats can be identified as horizontal or vertical series of anchor points. First, the ADHoRe algorithm detects all tandem repeats and remaps them onto a single gene. For the determination of the actual number and size of tandem repeats within the *Arabidopsis* genome, only homologous genes with five or fewer unrelated intervening genes were taken into account. Next, all paralogous regions are identified as clusters of diagonal series of anchor points by using a maximum gap size (*G*) and a “quality” parameter (*Q*) that decides whether genes or gene clusters indeed form a diagonal (Vandepoele et al., 2002a).

These parameters were set to *G*=25 and *Q*=0.9. To test the statistical significance of identified block duplications, a permutation test was applied in which 1,000 randomized datasets were sampled. Based on the number of anchor points in a cluster and the average distance between anchor points in a cluster (reciprocal density), these datasets were then used to calculate the probability that a cluster detected in our real dataset could have been generated by chance. Only clusters that had a probability <1% were retained in our analysis.

Age estimation of block duplications

For all non-hidden duplicated blocks detected with the ADHoRe algorithm and shown to be statistically significant, each anchor point was dated by using the NTALIGN program in the NTDIFFS software package (Conery and Lynch, 2001). This program first aligns the RNA sequence of two mRNAs based on their corresponding protein alignment and then calculates the number of synonymous substitutions per synonymous sites (K_s) by the method of Li (1993). We also calculated K_s by using the dating methods of Nei and Gojobori (1986) and Yang and Nielsen (2000). The latter two methods are implemented in the YN00 program of the PAML phylogenetic analysis package (Yang, 1997). The mean K_s value (average of the estimates obtained by the three methods) was derived for each anchor point. These values were then used to calculate the mean K_s (μK_s) and standard deviation (σK_s) for each block duplication, excluding outliers by using the Grubbs test with a 99% confidence interval (Grubbs, 1969; Stefansky, 1972). For certain anchor points, the sequence divergence was too large to obtain an age estimate with any of the three methods. Such anchor points were also removed from the analysis. The time since duplication was calculated as $T = \mu K_s / 2\lambda$, with λ being the mean rate of synonymous substitutions, which was estimated in *Arabidopsis* to equal 6.1 synonymous substitutions per 10^9 years (Lynch and Conery, 2000).

Grouping duplicated blocks into age classes

Block duplications were grouped into age classes by comparing the mean K_s values of different blocks of duplicated genes. Two duplicated blocks are put into the same age class if the hypothesis that the mean K_s values of both duplications differ significantly could be rejected by using a t-test with a 99% confidence interval.

When duplicated blocks can be grouped, the mean K_s (μK_s) and standard deviation (σK_s) of the resulting total group are calculated, together with the coefficient of variance ($CV = \mu K_s / \sigma K_s$). For statistical significance we consider only duplications with five or more obvious anchor points. Age classes are generated by using the following procedure: a candidate age class is formed by taking a first duplication and adding to it the duplication that results in the age class with the lowest CV. This process continues until no further duplications can be added to the age class without exceeding a CV value of 0.3. Next, a second candidate age class is formed by starting with a second duplication and repeating the process. This process is then repeated for each duplication, such that there are as many candidate age classes as there are duplications. At this point, the largest age class is retained and the duplications that it contains are removed from further consideration. The previous steps are repeated for the remaining duplications until no more age classes can be defined containing five or more duplications. Determination of the different age classes by using the procedure described above has the advantage that duplicated blocks with a high variance on the estimated age will not be considered for defining the number of statistically significant age classes. The disadvantage is that a considerable fraction (sometimes up to 50%) of the dated block duplications is omitted from the analysis. However, it should be noted that the determination of age classes with different CVs (cut-offs are between 0.25 and 0.4) always yielded three age classes.

Detection of hidden duplications

Hidden duplications are detected by identifying chromosomal segments that are involved in different non-hidden duplications (Figure 2.2.1). If we consider three non-overlapping chromosomal segments A, B, and C, for which it was shown that segments A and B form a non-hidden duplication, and segments A and C form an obvious non-hidden duplication, it is then checked as to whether segments B and C show statistically significant colinearity, i.e., whether they share enough (or any) pairs of homologous genes. If this is not the case, it is concluded that segments B and C form a hidden block duplication. The exact coordinates in the gene homology matrix of this hidden block duplication are then determined as follows: Let ($a_{B,start}$, $a_{B,stop}$) and ($b_{A,start}$, $b_{A,stop}$) be the start and stop positions on segments A and B, respectively, of the duplication between these segments (see Figure 2.2.1). Note that ($a_{B,start}$, $b_{A,start}$) and ($a_{B,stop}$, $b_{A,stop}$) are consequently the coordinates of the

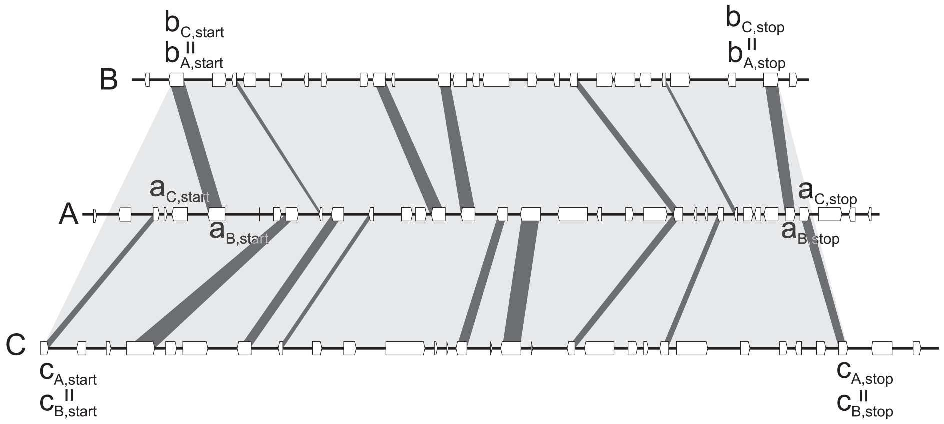


Figure 2.2.1 Determination of the borders of a hypothetical hidden duplication. Gene coordinates increase from left to right. See Materials and methods for details.

outermost anchor points of the observed duplication. Let ($a_{\mathrm{C,start}}$, $a_{\mathrm{C,stop}}$) and ($c_{\mathrm{A,start}}$, $c_{\mathrm{A,stop}}$) denote the same for segments A and C. The positions ($b_{\mathrm{C,start}}$, $b_{\mathrm{C,stop}}$) and ($c_{\mathrm{B,start}}$, $c_{\mathrm{B,stop}}$) for the hidden duplication between segments B and C are then determined by considering the start positions of the non-hidden duplications between A and C ($a_{\mathrm{C,start}}$) and A and B ($a_{\mathrm{B,start}}$). Suppose $a_{\mathrm{C,start}} \leq a_{\mathrm{B,start}}$. In this case the value of $c_{\mathrm{A,start}}$ is assigned to $c_{\mathrm{B,start}}$. The value of $b_{\mathrm{C,start}}$ is then determined by the coordinate b of the anchor point (a, b) in the duplication between segments A and B for which $a \geq a_{\mathrm{C,start}}$ and lies the closest to $a_{\mathrm{C,start}}$. The end positions ($b_{\mathrm{C,stop}}$, $c_{\mathrm{B,stop}}$) are determined in the same way. Thus, we infer the coordinates of the detected hidden duplication from the coordinates of the overlapping segments from the non-hidden duplications that lead to its detection. To rule out hidden duplications generated by statistical aberrances, we retain only those hidden duplications for which both non-hidden duplications have at least five anchor points on the common segment between them.

Results

Non-hidden block and tandem duplications

By using the ADHoRe algorithm (Vandepoele et al., 2002a), we identified a total of 304 non-hidden duplications (i.e., duplications that can be observed through direct comparison of chromosomal segments) in the *A. thaliana* genome (see Figure 2.2.2). These duplications contain a total of 3,571 anchor points. Eighty-two percent of all genes in the annotated genome and 80% of all sequenced nucleotide positions reside in duplicated segments (Table 2.2.1). This percentage is significantly higher than the 60% reported by the Arabidopsis Genome Initiative (2000). Nevertheless, it is clear that from the total set of genes located within duplicated segments, the major fraction of gene duplicates has been lost, whereas approximately 28% is retained. These findings are very similar to those reported by Vision et al. (2000). The smallest duplications consist of three anchor points with no intervening genes. The largest detected duplication concerned a 2.29-Mb segment containing 584 genes on chromosome 1 and a 2.00-Mb segment containing 479 genes also on chromosome 1, containing 172 anchor points. An example of a non-hidden duplication is shown in Figure 2.2.3a. Apart from these block duplications, 1,607 tandem repeats were detected, involving 4,193 individual genes. This result corresponds with 16.7% of all genes in our dataset. The largest tandem repeat contained 23 genes. These results are very similar to those reported (Arabidopsis Genome Initiative, 2000). A total of 137 non-hidden block duplications consisting of at least five paralogous gene pairs, and together containing 2,757 anchor points, were retained for dating duplication events. On the basis of these duplicated blocks of genes, three age classes could be defined (see Materials and methods) with mean K_s values of 0.91, 2.0, and 2.7, corresponding to duplication events 75 MYA, 163 MYA, and 221 MYA (see Table 2.2.2).

Hidden duplications and multiplication levels

In addition to the set of non-hidden duplications, we also identified 53 hidden duplications (see Materials and methods), with the smallest segments spanning 10 genes (51 kb) and the largest 218 genes (1.15 Mb). An example of such a hidden duplication can be found in Figure 2.2.3b. Detailed analysis of the hidden duplications reveals that in many cases some residual anchor points can still be identified (i.e., some degree of colinearity can still be observed). However, the

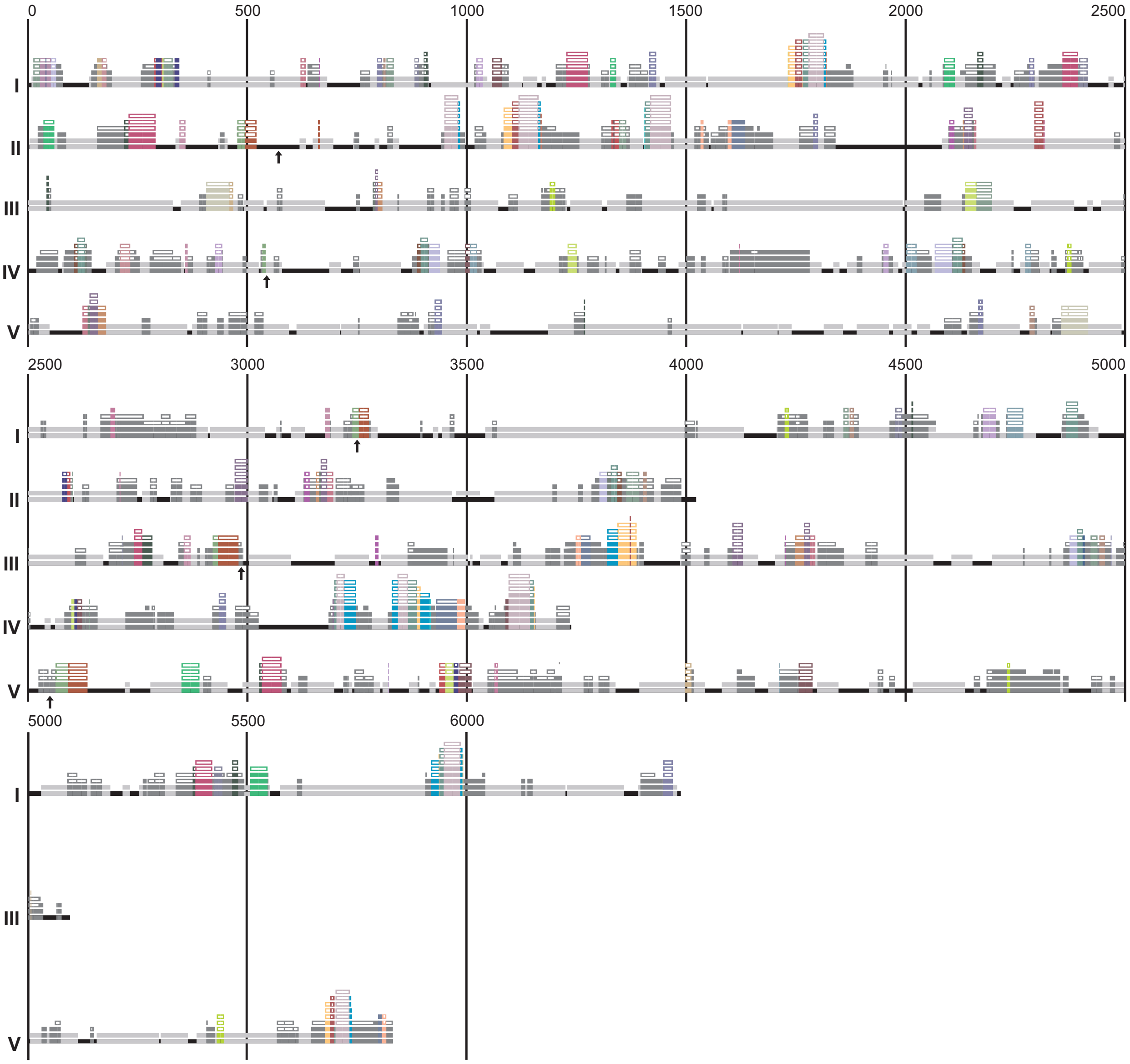


Figure 2.2.2 Overview of the chromosomal location of all multiplicons detected in the *Arabidopsis* genome. Baselines (black) represent all genes on the five chromosomes of *Arabidopsis*. Boxes on the baselines indicate segments that are part of a multiplicon (group of homologous segments). The number of boxes above the baselines indicates the number of additional segments that are homologous to the segment marked on the baseline. Filled boxes represent non-hidden duplications, whereas empty boxes denote hidden duplications, compared with the chromosome segment (see text for details). For all multiplicons with a multiplication level (the number of homologous segments in a multiplicon) greater than four (i.e., in agreement with three duplication events), a different color was used. Multiplicons with multiplication levels of three or four (in agreement with two rounds of duplication events) are marked in dark gray, whereas a multiplication level of two (a single duplication) is marked in light gray. Vertical black bars denote the number of genes, whereas arrows indicate the putative positions of the (collapsed) centromeres, which were removed from the initial dataset.

Table 2.2.1 Duplications in the *Arabidopsis* genome

Chr.	Number of genes in duplicated regions	Total number of genes	% of genes in duplicated regions	kb in duplicated regions	Total kb	% of kb in duplicated regions
1	5,532	6,488	85.27	24,846	29,640	83.83
2	3,163	4,023	78.62	14,129	19,643	71.93
3	4,335	5,096	85.07	19,582	23,333	83.92
4	3,027	3,738	80.98	13,723	17,549	78.20
5	4,637	5,832	79.51	20,451	26,269	77.85
Total	20,694	25,177	82.19	92,733	116,436	79.64

reason that these groups of anchor points are not recognized as non-hidden block duplications is that there are too few anchor points to be discriminated from random noise during the statistical filtering process of the ADHoRe algorithm (see Materials and methods). Furthermore, in some cases, not a single anchor point could be observed between two duplicated segments, indicating that, after being duplicated in *Arabidopsis*, these duplicated regions have lost a different, but complementary, set of genes (Vandepoele et al., 2002b). It should be noted that no duplications were found spanning the centromeric regions, which was also reported by Vision *et al.* (2000). Based on a complete analysis of all segmental duplications, we can identify a large number of chromosomal segments that have been involved in multiple duplications (Figure 2.2.2). We refer to such a group of homologous segments as a multiplicon. The multiplication level of a multiplicon is then defined as the number of chromosomal segments it contains. For example, if we consider only the 304 non-hidden duplications, the maximum multiplication level observed in the genome of *Arabidopsis* equals five (Figure 2.2.2). In other words, for certain genomic segments, another four homologous segments can be found elsewhere in the genome. However, when considering the set of 53 hidden duplications, the multiplication level increases significantly (Figures 2.2.2 and 2.2.4). The contribution of hidden duplications to the final multiplication level clearly shows the importance of considering such duplications. Although the major fraction of the set of multiplicons with a multiplication level greater than four has a maximum multiplication level of eight, one multiplicon was found with a level of nine (see below). Additional information describing hidden and non-hidden block duplications in greater detail can be obtained from our web site at http://bioinformatics.psb.ugent.be/.

Table 2.2.2 Detected age classes and age estimation

No. of blocks	No. of anchor points	Mean *K*s (SD)	Age in MY (SD)
21	311	0.91 (±0.27)	75 (±22)
33	266	2.0 (±0.60)	163 (±49)
7	50	2.7 (±0.82)	221 (±67)

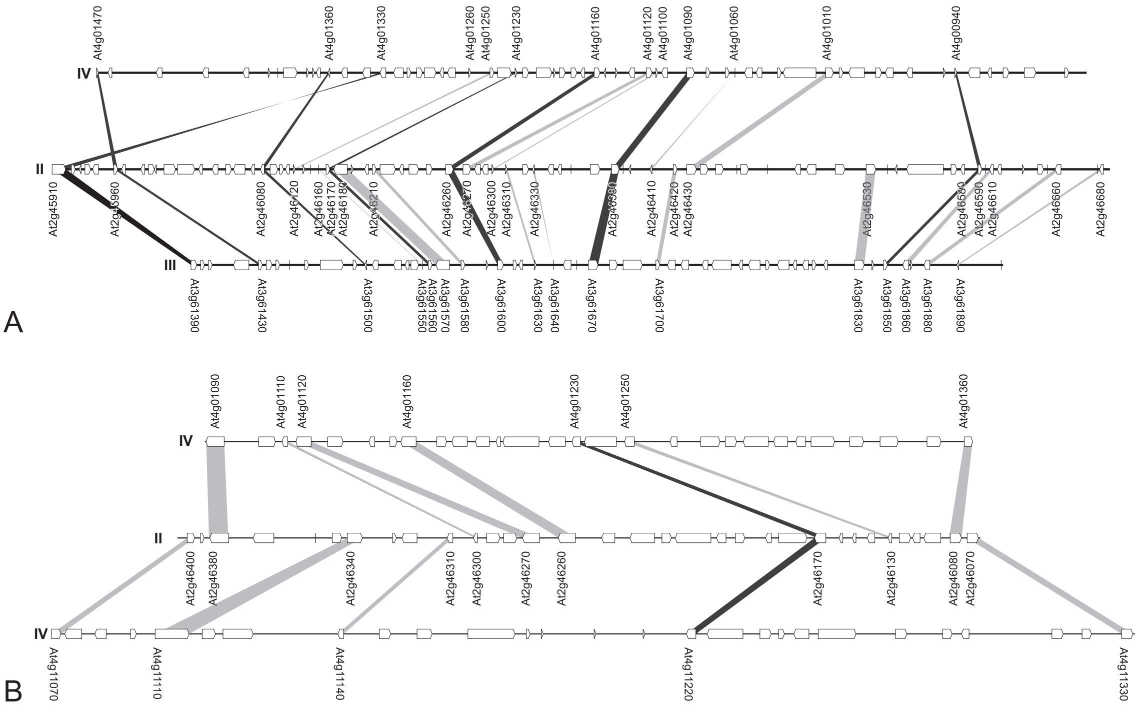


Figure 2.2.3 Non-hidden and hidden duplicated blocks. (a) Example of a multiplicon in which non-hidden duplications can be observed between all three segments involved. Several genes can be distinguished that have homologs (indicated by black bands) on all segments. Light gray bands show homologs on two of three segments. (b) Example of a multiplicon in which no non-hidden duplication can be observed between the two segments of chromosome IV. Both segments have only one homologous gene in common (dark gray band). However, both segments still share several, but different, homologous genes with a segment on chromosome II. Therefore, it can be concluded that both segments on chromosome IV form a hidden duplication.

Discussion

Careful analysis of duplicated regions shows that the majority of duplicated genes disappear during evolution. Nevertheless, in many cases, and with the right tools at hand, even after tens of millions of years of evolution, sufficient homologous gene pairs remain to detect many colinear, and thus duplicated, regions. Moreover, as shown in this study, even when the level of differential gene loss is too high to detect colinearity between two genomic segments, comparisons through a third segment can still reveal homology. Furthermore, when considering the set of 53 hidden duplications discovered in the genome of *Arabidopsis*, the multiplication level of many duplicated segments increases significantly. It is clear that, given the high multiplication levels observed in different multiplicons (see Figure 2.2.2), the genome of *Arabidopsis* must have undergone

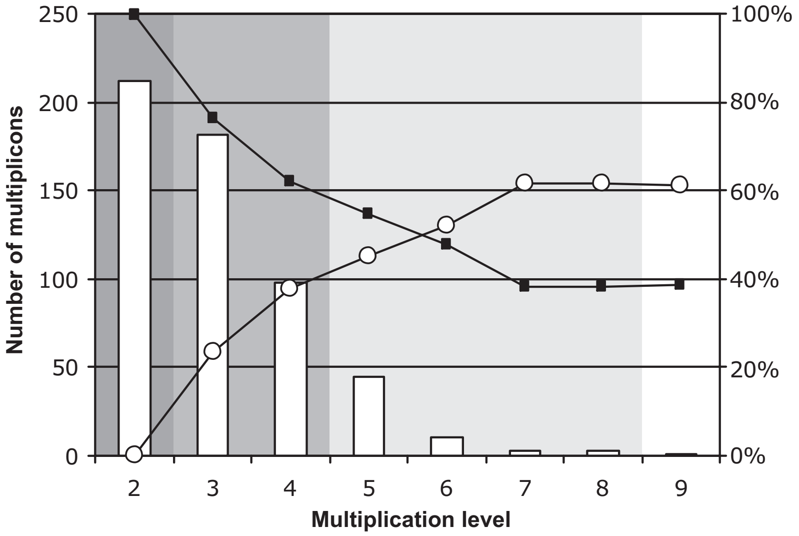


Figure 2.2.4 Multiplication levels and contribution of non-hidden and hidden duplications. Bars indicate the number of multiplicons (groups of homologous segments) for each multiplication level. The relative amount of non-hidden duplications within all multiplicons of a given multiplication level is represented as a black square, whereas white circles denote the contribution of hidden duplications. The multiplication levels supporting three rounds of duplication (mutiplication levels five to eight) are shaded in light gray, those supporting only two duplication events (multiplication levels three to four) are in gray, and the multiplication level of two (a single duplication) is marked in dark gray.

multiple rounds of large-scale gene or entire genome duplications. If, in a given genome a chromosomal segment appears in *n*-fold, then a lower bound for the number of duplications that have occurred is given by $d_{min} = \lceil \log_2(n) \rceil$ (take log2 of *n* and round up to the next integer), whereas the upper bound is given by $d_{max} = n$-1. Based on the parsimony principle, and assuming that all involved segments of the multiplicon have been detected, this lower bound number probably reflects the true number of large-scale gene duplication events that have occurred. In this study, we observe many multiplicons with multiplication levels between five and eight, which can be explained by assuming three rounds of duplications. However, the question remains whether the distribution of duplicated segments observed could be because of several smaller independent duplications rather than the observed multiplicity being the result of successive complete genome duplications followed by a large number of rearrangements and deletions. Although this cannot be completely ruled out, we agree with McLysaght *et al.* (2002) that this is probably the less plausible explanation. The hypothesis of several, small independent duplications requires a greater number of duplication events, whereas the hypothesis of successive genome duplications requires more deletion and rearrangement events. It has been shown that a polyploidization event is often followed by intense rearrangements and deletions, often involving large chromosomal segments or even entire chromosomes (Soltis and Soltis, 1993; Song et al., 1995). Thus, during these events large numbers of duplicated genes can be deleted simultaneously. This result, together with the fact that polyploidy is very often observed in land plants, probably favors the hypothesis of successive genome duplications. Furthermore, additional support for three rounds of genome duplications is provided by our dating analysis, although we are aware of the fact that dating must be interpreted cautiously. Dating was based on the inference of silent substitutions. Therefore, the obtained age estimates are unreliable for the two older age classes (dated 163 and 221 million years), because synonymous sites become quickly saturated and as a result, dates of older duplication events (with $K_s > 1$) become harder to estimate correctly (Li, 1997). Additionally, for older block duplications, the number of retained duplicated genes is usually low(er), and therefore fewer anchor points remain for the accurate dating of such blocks. The age of the youngest class (75 million years) is more reliable and is probably close to the true age of the most recent genome duplication in *Arabidopsis*. Other studies have suggested similar dates for the most recent polyploidization event of *Arabidopsis* (Arabidopsis Genome Initiative, 2000; Vision et al., 2000). However, one should keep in mind that the dating of duplication events was based on an estimated rate of 6.1 synonymous substitutions per 10^9 years (Li, 1997; Lynch, 1997). The use of other substitution rates (e.g., Bohle et al., 1996; Koch et al., 2000) might give quite different duplication dates. Nevertheless, to compare our study with recent studies that dealt with dating duplication events in *Arabidopsis* (Arabidopsis Genome Initiative, 2000; Vision et al., 2000) we have used the same substitution rate. Furthermore, although the absolute dating thus has to be considered cautiously, we believe that, whatever the exact synonymous substitution rate, dating based on synonymous substitutions will clearly reveal three significantly different age classes.

As stated previously, by using our method to determine the different age classes with different parameters always yielded a fixed number of three age classes, pointing to three large-scale gene duplication or polyploidization events in *Arabidopsis*. As can be observed in Figure 2.2.2, we detected one multiplicon with a multiplication level of nine. Although at first sight the detection of such a multiplicon seems to conflict with three genome duplications, detailed analysis revealed that the additional segment probably originated because of an additional

Table 2.2.3 Frequency of internal chromosomal duplications within the *Arabidopsis* genome

Chromosome No.	Hidden duplications	Nonhidden duplications	Anchor points in nonhidden duplications
1	4	24	478
2	2	8	25
3	3	2	8
4	3	10	113
5	1	13	152

duplication event on chromosome 1. One of the nine segments of the multiplicon indeed consists of an internal non-hidden duplication on chromosome 1, containing 172 anchor points. Overall, when comparing all internal duplications for each chromosome, we observe a significantly higher number of both non-hidden block duplications and anchor points involved in these internal duplications for chromosome 1 (see Table 2.2.3). When all internal chromosomal duplications in the *Arabidopsis* genome are excluded and the age classes are determined anew without these duplications, the same three age classes emerge. In other words, removing internal chromosomal duplications from the total dataset does not alter our view on the duplication history of *Arabidopsis*.

Our results clearly reject the single-genome duplication hypothesis as suggested (Arabidopsis Genome Initiative, 2000; Lynch and Conery, 2000). By plotting the frequency distribution of duplication dates inferred for duplicated blocks of genes based on amino acid sequence divergences, Vision et al. (2000) found a multimodal distribution, from which they concluded that at least four large-scale duplication events have occurred. However, as stated before, the dating methods applied in their study have been criticized. Although their method assumes that the overall distribution of amino acid substitution rates is the same throughout the genome, and therefore any contemporaneously duplicated block containing several homologous gene pairs provides an independent sample of that distribution (Vision et al., 2000; Todd Vision, personal communication), we have previously shown that many of their blocks have been dated erroneously (Raes et al., 2002). In our analysis, where we combined *K*s-based dating of non-hidden duplications with the multiple occurrences of homologous segments (i.e., multiplicons), we could not find any indication for a fourth polyploidy event in *Arabidopsis*. Although we agree that the more ancient duplication events are, the harder it is to detect them because of phenomena such as chromosomal rearrangements and translocations, we have shown here that at least the partial recovery of such ancient events should be possible. Therefore, we consider it unlikely that no traces could be detected of additional duplication events, if they have occurred.

2.3 Evidence that rice and other cereals are ancient aneuploids

Klaas Vandepoele, Cedric Simillion and Yves Van de Peer

Redrafted from: *Planct Cell* **15**, 2192-2202 (2003)

Detailed analyses of the genomes of several model organisms revealed that large-scale gene or even entire-genome duplications have played prominent roles in the evolutionary history of many eukaryotes. Recently, strong evidence has been presented that the genomic structure of the dicotyledonous model plant species *Arabidopsis* is the result of multiple rounds of entire-genome duplications. Here, we analyze the genome of the monocotyledonous model plant species rice, for which a draft of the genomic sequence was published recently. We show that a substantial fraction of all rice genes (~15%) are found in duplicated segments. Dating of these block duplications, their nonuniform distribution over the different rice chromosomes, and comparison with the duplication history of *Arabidopsis* suggest that rice is not an ancient polyploid, as suggested previously, but an ancient aneuploid that has experienced the duplication of one - or a large part of one - chromosome in its evolutionary past, ~70 million years ago. This date predates the divergence of most of the cereals, and relative dating by phylogenetic analysis shows that this duplication event is shared by most if not all of them.

Introduction

Large-scale duplication events have been considered important for the evolution of many organisms because they provide a way to considerably increase the genetic material on which evolution can work (Stephens, 1951; Ohno, 1970; Sidow, 1996; Holland, 2003). Because duplicated genes are redundant, one of the copies is, at least theoretically, freed from functional constraint and can evolve a new function (Van de Peer et al., 2001; Prince and Pickett, 2002). The search for traces of (ancient) large-scale gene duplications has received much attention of late, and hypotheses about the number and age of polyploidy events in eukaryotes are actively discussed (Wolfe, 2001; Durand, 2003). This is partly attributable to the fact that the detection of homologous (or paralogous) regions in genomes is not self-evident (Gaut, 2001; Vandepoele et al., 2002a).

Identifying duplicated regions at the gene level is based on a within-genome comparison that aims at delineating regions of conserved gene content and order (such regions are said to be colinear) in different parts of the genome. In general, one tries to identify a number of homologous gene pairs (usually referred to as anchor points) in relatively close proximity to each other between two different segments in the genome, either on the same chromosome or on different chromosomes. When such a candidate colinear region is detected, usually some sort of permutation test is performed in which a high number of randomized data sets are sampled to calculate the probability that the observed colinearity could have been generated by chance (Gaut, 2001). When it can be shown that the similarity between two genomic segments is unlikely to be the result of chance and therefore is statistically significant, the conclusion is reached that the duplicated genes are the result of a single segmental (block) duplication. The statistics that determine colinearity depend on two factors: the number of anchor points and the distance over which these are found, which usually depends on the number of “single” genes that interrupt colinearity. The high level of gene loss - together with phenomena such as translocations and chromosomal rearrangements - often renders it very difficult to find statistically significant homologous regions in the genome, particularly when the duplication events are ancient.

In plants, the systematic analysis of the *Arabidopsis* genome sequence has shown that this genome contains a large number of duplicated regions and that up to ~90% of the *Arabidopsis* genes occur in genomic segments that have been duplicated at one time or another (Vision et al., 2000; Simillion et al., 2002; Bowers et al., 2003). By applying novel techniques to detect heavily degenerated block duplications in *Arabidopsis*, we showed recently that the genome of this dicotyledonous model plant has been reshaped by not one but three large-scale gene, and probably even entire-genome, duplication events (Simillion et al., 2002). Apart from *Arabidopsis* (Arabidopsis Genome Initiative, 2000), rice is currently the only plant species for which draft sequences of the nuclear genome have been published (Goff et al., 2002; Yu et al., 2002). In addition, more complete versions of chromosomes 1, 4, and 10 have been published by the International Rice Genome Sequencing Project (Feng et al., 2002; Sasaki et al., 2002; Rice Chromosome 10 Sequencing Consortium, 2003). Rice is one of the most important cereal crops in the world and also is an excellent plant model system, as a result of its small genome size (430 Mb) and the high level of synteny with other cereals. Comparative mapping analyses of genomes of closely related grass species revealed a remarkably good conservation of markers within large chromosomal segments (for review, see Keller and Feuillet, 2000). Soon after the detection of colinearity based on genetic maps, detailed sequence analyses confirmed the existence of microcolinearity (i.e., conserved gene content and order at the gene level) between orthologous loci from closely related grass genomes, which varied extensively in size (Chen et al., 1997; Tikhonov et al., 1999; Paterson et al., 2000; Tarchini et al., 2000). Consequently, grasses can be studied as a single genetic system, allowing the transfer of biological information from a well-studied model grass genome, such as that of rice, to related plant species (Gale and Devos, 1998a). Although several studies that crossed the monocot-dicot boundary also identified numerous microcolinear segments between Arabidopsis and rice (Paterson et al., 1996; Liu et al., 2001; Mayer et al., 2001; Salse et al., 2002; Vandepoele et al., 2002a), the small size of these regions seems to seriously limit their value for comparative analysis of dicotyledonous and grass genomes.

In strong contrast to *Arabidopsis*, in which the initial sequencing of the genome sequence already revealed numerous duplicated segments (Terryn et al., 1999; Blanc et al., 2000; Paterson et al., 2000), very few studies have reported possible evidence for large-scale gene or complete-genome duplications in rice (Kishimoto et al., 1994; Nagamura et al., 1995), although a polyploid origin for rice has been suggested on several occasions (Goff et al., 2002; Levy and Feldman, 2002). Here, we report the detailed analysis of the rice genome, focusing on large-scale gene duplications. We show that large-scale gene duplication events did occur in the evolutionary past of rice but that the duplication history and magnitude are considerably different from those of its dicotyledonous counterpart *Arabidopsis*.

Results

Detection of non-hidden block duplications in the rice genome

Because one preferentially wants to use large genomic regions for the detection of duplicated segments in a genome, we built a data set of assembled rice genomic BAC sequences that were obtained from the International Rice Genome Sequencing Project (Sasaki and Burr, 2000). Where traditional sequence assembly programs are designed mainly to assemble large sets of individual sequence reads into larger contigs, the construction of large genomic scaffolds starting from already assembled genomic BAC clones is far from trivial. Because no publicly available assembly program was found that could handle and assemble genomic BAC clones, which range in size from 10 to 250 kb, we applied a newly developed assembly routine. The automatic sequence-to-genome assembly routine (ASGAR) is a conservative method that physically merges BAC clones with significant overlap (see Materials and methods).

After applying two rounds of assembly using ASGAR to the initial data set, the number of genomic sequences was reduced from 2897 BACs to 1025 genomic scaffolds (498 supercontigs and 527 singleton BACs). The total size of these scaffolds is 330.47 Mb, with an average size of 322 kb per scaffold. Gene annotation was retrieved from RiceGAAS (Sakata et al., 2002) and yielded 39,096 genes after filtering. This filtering step removed potential falsely predicted genes, based on the absence of homology for a predicted gene with a rice EST or any other protein present in the public protein databases (Vandepoele et al., 2002a). In addition, all predicted genes with similarity to transposable elements were removed. On average, 32 genes were present per genomic scaffold, which corresponds with an average gene density of one gene per 10 kb. An overview of the gene density and the length distribution of the scaffolds is shown in Figure 2.3.1.

By applying the ADHoRe (Automatic Detection of Homologous Regions) algorithm to an assembly covering ~70% of the annotated rice genome sequence, 193 statistically significant duplicated segments were identified ($P<0.001$), of which 150 contain three or four paralogous gene pairs (so-called anchor points) and 43 contain five or more gene duplicates. The complete set of block duplications,

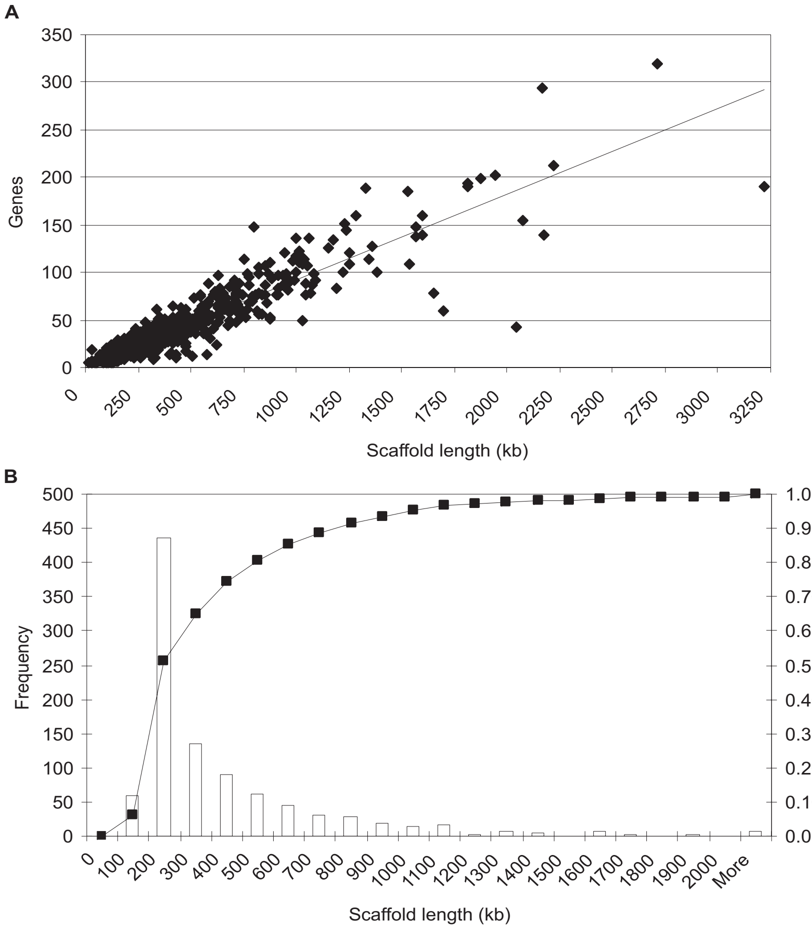


Figure 2.3.1 Overview of the genomic scaffolds generated by ASGAR. (a) Scatterplot showing the number of genes versus the scaffold length for all 966 genomic scaffolds that were used for the detection of duplicated blocks. The best-fit line, which shows a quite homogeneous gene density for the scaffolds (r^2 = 0.85), represents a gene density of 1 gene per 10 kb. (b) Length distribution of all genomic scaffolds that were subjected to block detection. The line indicates the relative (cumulative) contribution of the scaffolds assigned per bin (i.e., length segment) in the histogram.

omitting tandem duplications, contains 862 anchor points and includes nearly 15% of all rice proteins in our annotated nonredundant data set. Approximately two-thirds of the duplicated blocks (i.e., 129 of all detected duplicated blocks) are located at the beginning or the end of a genomic scaffold (i.e., the first or last five genes), which can be explained by the incomplete assembly of our data set. Regarding the 43 large block duplications (more than five anchor points), 34% of the total number of genes in these segments are retained duplicates. The largest block duplication in our assembled scaffold data set is formed by a 0.96-Mb segment with 107 genes on chromosome 1 and a 0.69-Mb segment with 62 genes on chromosome 5, governing 33 retained gene duplicates. Apart from the set of paralogous genes located in duplicate blocks, 1,609 tandem duplications were detected involving 4,308 individual genes. This number corresponds with 16.9% of all genes in our data set**,** which is very similar to what is found in Arabidopsis (Vision et al., 2000; Simillion et al., 2002). The largest tandem repeat was formed by 16 genes.

Hidden and ghost duplications

Apart from the large set of block duplications identifiable by direct comparisons of different genomic segments (so-called "non-hidden" duplications), an additional number of block duplications in the rice genome could be identified by indirect comparisons (so-called "hidden" and "ghost" duplications; see Materials and Methods) (Figure 2.3.2). Hidden duplications are heavily degenerated block duplications that cannot be observed by directly comparing the duplicated segments; rather, they are observed only through comparison with a third segment. Consequently, hidden duplications are important to consider for determining the actual number of duplication events that have occurred over time, as we demonstrated previously for *Arabidopsis* (Simillion et al., 2002). Reconstruction of multiplicons (i.e., sets of homologous segments; Simillion et al., 2002) for rice through the identification of hidden duplications revealed only two cases in which a chromosomal segment was involved in more than one duplication event.

Considering all 157 colinear regions detected between rice and *Arabidopsis*, another five ghost duplications were identified. The largest rice ghost duplication was found between genomic segments of chromosome 4 (46 genes spanning 477 kb) and chromosome 10 (64 genes spanning 761 kb), both colinear with chromosome 2 of *Arabidopsis*. More detailed analysis of these duplicated segments showed that each genomic segment has lost a different set of genes and that only a subset of the initial number of gene duplicates is retained (data not shown). Therefore, the combination of a limited number of gene duplicates with different types of rearrangements subsequent to the original duplication event does not allow the detection of this degenerated paralogous region using only the rice genome.

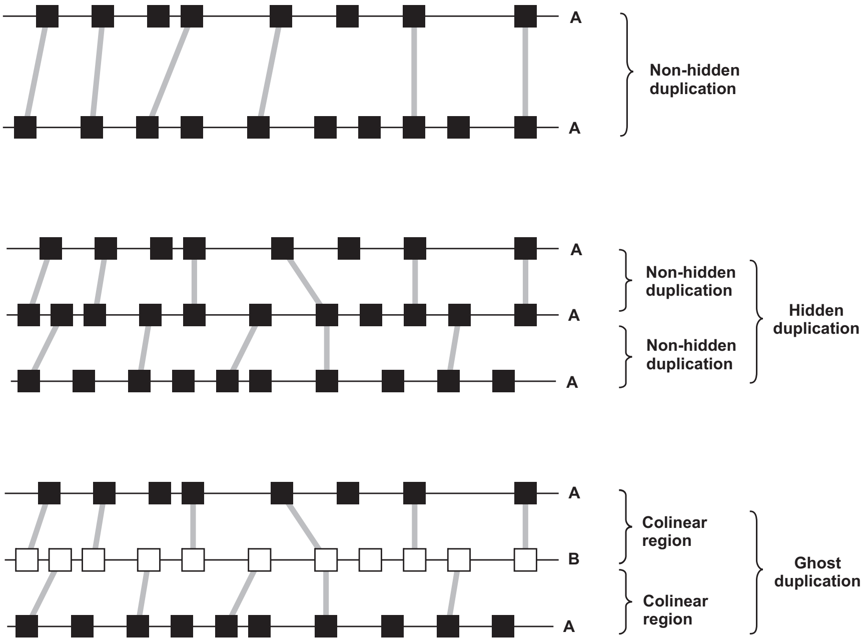


Figure 2.3.2 Scheme of non-hidden, hidden, and ghost duplications. Boxes represent the genes on chromosomal segments of genomes A and B, whereas connecting lines indicate the anchor points (i.e., homologous or duplicated genes). Hidden duplications are heavily degenerated block duplications that cannot be observed by directly comparing the duplicated segments; rather, they are observed only through comparison with a third segment from the same genome. Because non-hidden duplications are used to infer hidden duplications, no additional genomic segments are assigned to a duplication event, although the number of duplication events for a given segment increases. Ghost duplications are hidden block duplications that can be identified only through colinearity with the same segment in a different genome. In contrast to hidden duplications, the identification of ghost duplications increases the fraction of the genome involved in a duplication event.

Age estimation of duplicated blocks

For reasons of statistical significance (see Materials and Methods), only the set of block duplications with five or more anchor points (377 anchor points in total) was used to date the duplication events. Briefly, for a duplicated block, all anchor points were subjected to a dating method based on the number of synonymous substitutions per silent site (K_s), and all values obtained were used subsequently to calculate the mean K_s for each block duplication after removing outliers (Simillion et al., 2002). Although large variation in K_s estimates among contemporaneously duplicated genes in *Arabidopsis* has been reported (Zhang et al.,2002), removal of outliers greatly reduces the variation of the final K_s estimate for a duplicated block. Nearly half of all anchor points (i.e., 47%) have K_s values of between 0.6 and 1.1 (Figure 2.3.3), corresponding with duplication dates of 46

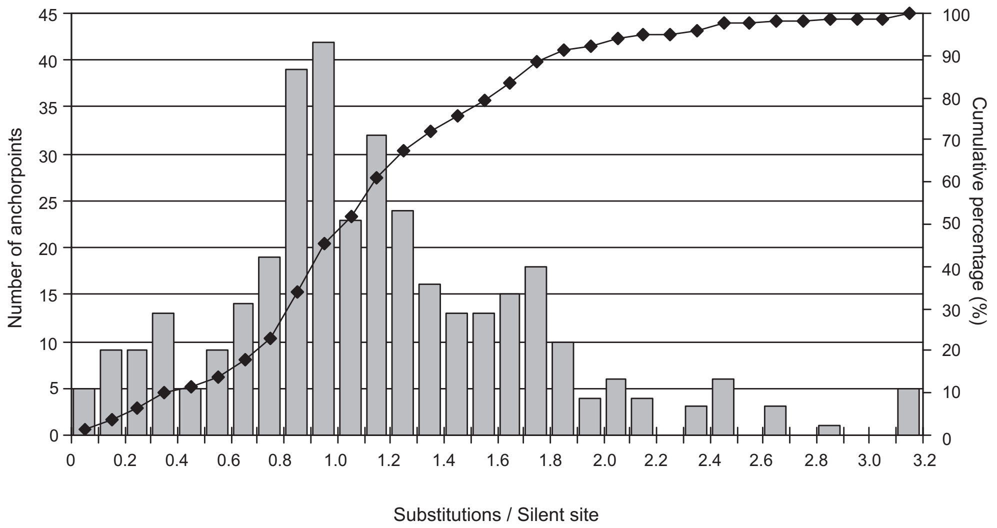


Figure 2.3.3 Absolute dating of block duplication events in the rice genome. Age distribution of all gene duplicates that are part of large (more than five anchor points) duplicated segments in the rice genome. The line indicates the relative (cumulative) contribution of the anchor points assigned per bin (i.e., age segment) in the histogram.

and 85 million years ago, respectively. The median, a K_s value of 0.87, corresponds with 67 million years ago.

Because absolute dating of duplication events has been criticized and may rely heavily on obtained K_s values and the estimated rate of synonymous substitutions for the organism of interest, which may not be very accurate (Li, 1997; Zeng et al., 1998; Blanc et al., 2003), we also applied relative dating by phylogenetic means (see Materials and Methods). In short, for a given pair of gene duplicates that is part of a duplicated block, homologous genes of related monocotyledonous plants were selected together with an appropriate outgroup sequence, and the evolutionary relationships between these different organisms were inferred based on the topology of the phylogenetic tree obtained. In total, 170 phylogenetic trees with bootstrap support were generated, representing a set of 99 block duplications (i.e., 1.7 trees per duplicated block on average). Fifty-four percent of these trees clearly supported the duplication event having occurred before the divergence of the cereals (Figure 2.3.4) (Kellogg, 2001).

Regarding the 18 large (more than five anchor points) block duplications with K_s values between 0.6 and 1.1, 74% of the topologies clearly supported duplication having occurred before the divergence of cereals. When more than one anchor point in the same block duplication could be used for tree construction (as was the case for 39 block duplications), 78% of the inferred trees within one duplicated block were congruent with one another. For all of the remaining tree

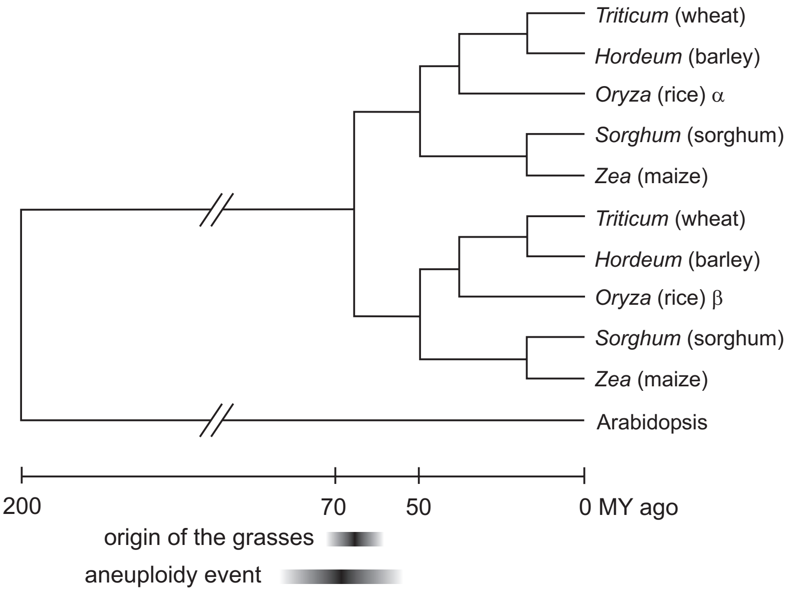


Figure 2.3.4 Dating of duplication events in the rice genome by phylogenetic means. Expected tree topology and date of origin for genes of the cereals wheat, barley, rice, maize, and sorghum if these genes have duplicated before the divergence of rice and other cereals. The large majority of tree topologies obtained in this study, including those of two copies of rice (i.e., the retained duplicates found in large duplicated segments) and at least one copy of another cereal, are congruent with this tree topology, in which one rice gene branches off before the divergence of rice and other cereals. Such topologies suggest a duplication before the divergence of rice, barley, wheat, maize, and sorghum, estimated at ~50 million years ago (Kellogg, 2001), and may have occurred just before the origin of the grasses, as suggested by the Ks-based dating (see text for more details).

topologies, no conclusions could be reached, for different reasons, such as the absence of real orthologs or sequences being too conserved. However, none of the trees was in clear conflict with a duplication event shared between rice and other cereals. Supplemental data on the block duplications detected, along with more detailed results from the dating analyses, are available at http://bioinformatics.psb.ugent.be/.

Discussion

The grass family has been the subject of many detailed comparisons of genome structure and gene order. Based on the presence of large colinear regions between different grass genomes, the creation of a grass consensus map clearly revealed the structural similarity between related grass genomes (Gale and Devos, 1998a). Although large chromosomal rearrangement events can be determined with the current resolution of these maps, information regarding large-scale duplication events within rice is scarce (Kishimoto et al., 1994; Nagamura et al., 1995).

Based on a BAC assembly covering >70% of the genome sequence of rice, we applied the ADHoRe algorithm to detect block duplications at the gene level. Subsequent to the detection of a large number of duplicated segments by direct comparison of all rice genomic scaffolds, a comparative approach using the genome sequence of *Arabidopsis* also yielded a set of ghost duplications, reflecting heavily degenerated duplicated segments. Regarding the 43 large (more than five anchor points) block duplications, 34% of the total number of genes in these segments are retained duplicates. This fraction of retained gene duplicates, when the estimated time of duplication is considered (see below), is very similar to what has been observed in Arabidopsis and yeast (28 and 25%, respectively) (Wolfe and Shields, 1997; Simillion et al., 2002), which seems to indicate similar rates of gene loss after duplication events.

When inferring the multiplication levels for all multiplicons (sets of homologous segments) present in the rice genome through non-hidden, hidden, and ghost duplications, ~1.3% of the genome resides in multiplicons with multiplication levels greater than two. This finding demonstrates that, given the quality of the current rice genomic data, a very small number of chromosomal regions seems to have been involved in multiple duplication events, in strong contrast to the findings in *Arabidopsis*, in which the majority of chromosomal regions have been involved in multiple duplication events (Vision et al., 2000; Simillion et al., 2002; Bowers et al., 2003).

It has been suggested that many polyploidy and/or aneuploidy events in the evolutionary history of the grasses are required to explain the current distribution of chromosome numbers among grass taxa (for review, see Gaut, 2002). Although an apparent whole-genome duplication, ~40 to 50 million years ago, was reported based on the rate of amino acid substitution of all possible paralogous protein pairs in the rice genome (Goff et al., 2002), there is good evidence that protein distances are not very reliable for the large-scale dating of heterogeneous classes of proteins (Li, 1997; Wolfe, 2001; Raes et al., 2003). To answer the question of whether rice is an ancient polyploid, we compared the duplication history of *Arabidopsis* and rice by plotting the total number of gene pairs in both species against their genetic distance inferred from the nucleotide substitutions at silent sites (Figure 2.3.5).

When all duplicated gene pairs in *Arabidopsis* and rice were plotted as a function of K_s, the shape and height of both curves were quite different. In *Arabidopsis*, the number of duplicates with K_s values between 0.6 and 0.9 increased dramatically, which corresponds with a genome duplication ~40 to 75 million years ago, as reported previously (Lynch and Conery, 2000; Simillion et al., 2002; Blanc et al., 2003; Bowers et al., 2003). Although overall, an exponential decay of the number of retained gene duplicates over time can be observed (Lynch and Conery, 2000), a small but significant increase also was observed for rice duplicates with K_s values between 0.6 and 1.1. However, because the increase in the number of

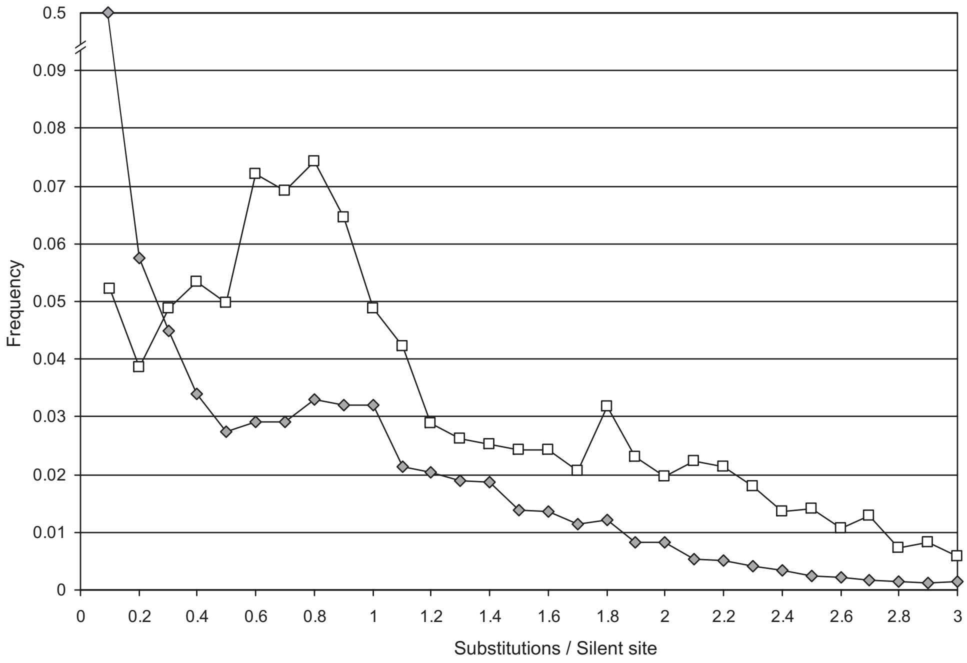


Figure 2.3.5 Frequency distribution of duplicated genes in *Arabidopsis* and rice as a function of the number of silent substitutions per silent site. All frequencies were corrected for the total number of dated gene duplicates per genome, which were 4928 for *Arabidopsis* (white squares) and 7698 for rice (gray diamonds). The fact that the total number of duplicated genes is higher in the rice than in the Arabidopsis gene family is attributable to the facts that the rice genome contains more predicted genes and that in Arabidopsis more gene families with >10 members have been omitted from the analysis.

duplicates, relative to the total number of duplicates, is much smaller in rice than in *Arabidopsis* (Figure 2.3.5), a complete genome duplication in rice seems highly unlikely.

In *Arabidopsis*, in which at least three rounds of large-scale gene duplication have been suggested (Vision et al., 2000; Simillion et al., 2002), 80% of the genome resides in duplicated blocks, 60% of which can be attributed to the most recent duplication event (data not shown). In the yeast *Saccharomyces cerevisiae* also supposedly an ancient tetraploid, ~50% of the genome is found in duplicated segments (Wolfe and Shields, 1997). Therefore, if similar rates of gene loss are assumed during diploidization (the process whereby a tetraploid species becomes a diploid) between different eukaryotes, the fact that only ~15% of the rice genome is found in duplicated blocks also disagrees with the notion of whole-genome duplication.

Mapping the locations of all block duplications on the different chromosomes provides an alternative way to estimate the distribution and overall impact of the duplicated blocks (Table 2.3.1). The physical size of all duplicated segments between two chromosomes was determined by comparing the fraction in our data set with the estimated chromosome sizes described by Chen et al. (2002). If a complete-genome duplication or polyploidization event had occurred in the evolutionary past of rice, it would be expected that all duplications would be spread uniformly over all chromosomes (i.e., null hypothesis). On the contrary, if only one chromosome, or a larger segment of a chromosome, had duplicated, such a uniform distribution would not be expected. Clearly, the observed distribution differs significantly from the null hypothesis (Table 2.3.1), which strongly suggests that the observed duplication landscape is not the result of an entire-genome duplication. Instead, a major fraction of the detected duplications involve chromosome 2, suggesting that this chromosome, or at least parts thereof, might have been involved in an aneuploidy event, followed by a number of chromosomal rearrangements. Both the dating based on synonymous substitutions and the dating by phylogenetic means support the notion that this event occurred before the divergence of cereals.

Because of the incomplete and fragmented nature of the current data set and the conservative approach used, our results only partially confirm previously reported duplicated segments based on marker analysis (Kishimoto et al., 1994; Nagamura et al., 1995). Therefore, we expect that the total number of duplications will be slightly higher once more complete chromosomal sequences become

Table 2.3.1 Coverage of block duplications between all 12 rice chromosomes

		0.003																						
1	0.003																							
	N.D.			0.011																				
2		N.D.	0.011																					
				0.003	N.D.																			
3			0.003			N.D.																		
		6E-04		**0.083**		0.003		**0.021**																
4	0.001		**0.096**		0.002		**0.021**																	
		0.108		0.002		0.005		0.001	N.D.															
5	**0.157**		0.002		0.003		5E-04			N.D.														
		0.002		**0.081**	N.D.			0.004		0.006		**0.019**												
6	0.001		**0.115**			N.D.	0.007		0.003		**0.019**													
		0.004		**0.018**		**0.032**		0.008		0.007		0.009		0.011										
7	0.008		**0.016**		**0.021**		0.013		0.001		0.017		0.011											
		0.006		0.006	N.D.			0.035		3E-04		**0.018**		0.012		0.006								
8	0.011		0.004			N.D.	0.011		2E-04		**0.017**		0.018		0.006									
	N.D.			7E-04		0.01		0.002	N.D.			0.004		0.003		0.007	N.D.							
9		N.D.	8E-04		0.023		0.007			N.D.	0.012		0.006		0.014			N.D.						
		0.003		**0.03**		0.009		**0.056**	N.D.			0.003		0.005		0.001	N.D.		N.D.					
10	0.012		**0.038**		0.007		**0.063**			N.D.	0.002		0.01		0.001			N.D.		N.D.				
	N.D.			5E-04	N.D.			0.002	N.D.			0.003		0.002	N.D.			0.007		0.002	N.D.			
11		N.D.	0.001			N.D.	0.006			N.D.	0.018		0.004			N.D.	0.005		0.005			N.D.		
		0.005		0.002				0.005	N.D.			0.011		7E-04		0.01		0.009	N.D.			0.008		0.009
12	0.006		0.004				0.012			N.D.	0.015		0.001		0.015		0.005			N.D.	0.002		0.009	
		1		2		3		4		5		6		7		8		9		10		11		12

Each pair of values in a particular cell describes the fraction covered by duplicated segments between both chromosomes (e.g., all duplicated segments between chromosome 1 and 5 cover 10.8% of chromosome 1 and 15.7% of chromosome 5). All values that differ significantly from the null hypothesis (i.e., that duplicated segments are distributed uniformly over all chromosomes) are indicated in boldface (P <0.001). N.D. (not determined) refers to the fact that no block duplications could be detected between the chromosomes.

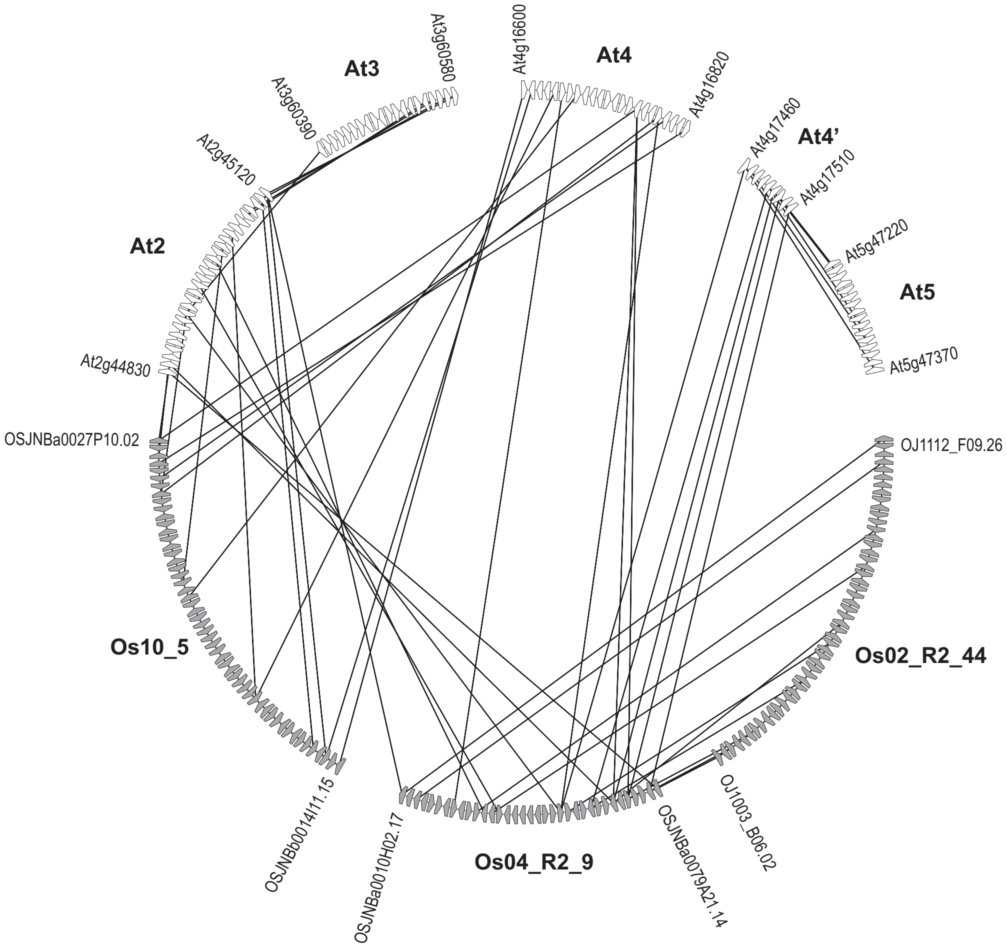


Figure 2.3.6 Set of homologous chromosomal segments (multiplicon) of *Arabidopsis* and rice. Arrows represent the genes on the chromosomal segments, and connecting lines indicate the anchor points (i.e., homologous or duplicated genes) that are part of a significant colinear relation determined by the ADHoRe algorithm. For each genomic segment, the names of the two genes delineating the segment are shown. Chromosomal segments of rice and *Arabidopsis* are shown in gray and white, respectively. By considering the colinearity between *Arabidopsis* and rice, a set of seemingly unrelated *Arabidopsis* segments can be joined into a multiplicon with a multiplication level of five, confirming the three duplication events in *Arabidopsis* described previously (Simillion et al., 2002). This colinearity also reveals that all three rice segments are linked with each other by two duplication events. Scaffold Os04_R2_9 includes BACs with accession numbers AL663006, AL662998, AL606459, AL607006, AL606728, AL606695, AL606587, AL606647, AL606633, AL663000, AL731613, AL606682, AL606687, AL606694, AL606628, AL607001, AL663003, and AL662954; scaffold Os10_5 includes BACs with accession numbers AC084763, AC079890, AC079874, AC069300, AC037426, and AC026758; and scaffold Os02_R2_44 includes BACs with accession numbers AP005108, AP004037, AP004883, AP005072, AP005289, AP005006, and AP004676.

available.

The presence of a small number of rice genomic regions that seem to have experienced multiple duplication events suggests that additional older block duplications occurred in the evolutionary past of rice. Indeed, analysis of mixed multiplicons (Figure 2.3.6), which represent all homologous relationships between genomic segments from *Arabidopsis* and rice, shows that additional information regarding genome evolution and duplication events within these plant model systems can be inferred. Careful investigation of colinear segments between *Arabidopsis* and rice shows that a number of very degenerated block duplications still can be recovered for both organisms, allowing a more realistic estimation of the number of duplication events that a homologous genomic segment in both species has undergone. Because for a number of mixed multiplicons the colinearity between homologous segments of rice and *Arabidopsis* still can be determined in a statistically significant way, which is not the case for paralogous segments within the genomes of rice and *Arabidopsis* (Figure 2.3.6), this pattern of conserved gene content and order could represent the remnants of a duplication event predating the monocot-dicot divergence, as was suggested recently (Bowers et al., 2003; Raes et al., 2003).

Materials and methods

Rice data set

A total of 2,897 rice (*Oryza sativa*) BAC sequences of the International Rice Genome Sequencing Project were retrieved from GenBank (September 2002). The total size of these genomic sequences amounts to 406.66 Mb, with an average size of 140 kb per BAC. Because both the sequence quality and the average length of genomic scaffolds from whole-genome shotgun approaches (fourfold to sixfold coverage and ~6 to 10 kb) (Goff et al., 2002; Yu et al., 2002) are inferior compared with BAC data, the former are less suited for the detection of block duplications. In addition, gene annotations for both whole-genome shotgun approaches are not publicly available.

Automatic Sequence-to-Genome Assembly Routine

A newly developed assembly routine for BAC sequences called the automatic sequence-to-genome assembly routine (ASGAR) was applied to merge significantly overlapping BAC sequences into larger contigs (so called supercontigs). For each genomic BAC sequence, ASGAR determines the BAC with the most significant overlap and creates a linked BAC pair. In the next step, either a new BAC pair is formed with no relation to the existing pair or a BAC pair that can be linked to an existing pair is formed. Afterwards, all overlapping BAC sequences that are linked and thus represent a tiling path are merged into supercontigs using the EMBOSS program megamerger (Rice et al., 2000). A significant overlap between two BAC sequences is defined by an overlap of at least 1,500 nucleotides with minimum 99% sequence identity. In addition, the overlap must be located at the end of one of the BAC sequences (i.e., the first or last 20% of the sequence). Sequence similarity searches were performed with BLASTN (Basic Local Alignment Search Tool; Altschul et al., 1997). Because both the input and output of ASGAR are a set of genomic sequences, multiple rounds of assembly can be performed until no more BAC sequences can be merged.

Detection of non-hidden block duplications, hidden block duplications, and ghost block duplications

All rice scaffolds covering five or more genes (966 scaffolds, or 286.01 Mb) were used for the detection of block duplications using ADHoRe, a recently developed tool for the automatic detection of homologous regions. Homologous gene pairs for the two genomic fragments compared were determined using BLAST and homology-derived secondary structure prediction (Rost, 1999). The ADHoRe parameters were set to Q=0.9 and G=25 (Vandepoele et al., 2002a). Only block duplications that had a probability of being generated by chance of <0.1% (or a significance level of 99.9%) were retained in our analysis. For the determination of the number of tandem duplications within the rice genome, only homologous genes with five or fewer unrelated intervening genes were considered.

Apart from block duplications that can be recognized clearly (so called obvious or non-hidden block duplications) and tandem duplications, we also discerned hidden and ghost duplications (Figure 2.3.2). Hidden duplications are heavily degenerated block duplications that cannot be observed by directly comparing the duplicated segments with each other; rather, they are observed only through comparison with a third segment (Simillion et al., 2002). Ghost duplications are defined as hidden duplications between different genomes. Thus, two genomic segments in the same genome form a ghost duplication when their homology can be inferred only through comparison with the genome of another species (Vandepoele et al., 2002b). To detect ghost duplications, initially, all colinear regions between rice and *Arabidopsis* were determined using ADHoRe (Q=0.9, G=25, and 99.9% significance level). Subsequently, all duplicated segments within *Arabidopsis* (Simillion et al., 2002) and all colinear regions between rice and *Arabidopsis* were mapped to infer networks of colinearity between both model plants and to detect ghost duplications in rice. Only non-hidden duplications and colinear regions with at least five anchor points were considered.

Age estimation of block duplications

For all non-hidden block duplications that were shown to be statistically significant, the time of duplication (age in million years) was determined using a dating method based on the fraction of synonymous substitutions per silent site (K_s), as described previously (Simillion et al., 2002). In short, the mean K_s value (average of the estimates obtained by three methods) was derived for each anchor point. These values then were used to calculate the mean K_s for each block duplication, excluding outliers. The mean rate of synonymous substitutions for rice was considered to be 6.5 synonymous substitutions per 10^9 years (Gaut et al., 1996; Li, 1997).

Age estimation of individual gene pairs

First, the complete set of rice and *Arabidopsis* genes was used to determine all gene families based on sequence similarity. In this procedure, an all-against-all sequence comparison is performed at the protein level for the complete set of genes in a genome. Subsequently, the alignable region and sequence identity between two similar proteins are validated to infer genuine paralogous relationships (Li et al., 2001). Finally, a simple-linkage clustering procedure is applied to assign individual genes to a gene family, given all paralogous relationships. For each gene family, the number of K_s was determined for all paralogous gene pairs by the method of Li (1993). Gene families with >10 members were excluded to reduce the number of gene family–specific pairwise comparisons.

Phylogenetic reconstruction

Phylogenetic trees were constructed with the neighbor-joining algorithm as implemented in LinTree (Takezaki et al., 1995), based on Poisson distances inferred from amino acid sequence alignments. Bootstrap analysis involving 1000 resamplings was performed to test the significance of the internodes. For each pair of duplicated rice genes, a sequence similarity search (BLASTP; Altschul et al., 1997) was performed to detect homologous monocotyledonous gene sequences and an appropriate dicotyledonous outgroup. The detection of a suitable outgroup was performed by selecting the best hit with an E value of <1e-50 for one of the gene duplicates among a set of dicotyledonous proteins (i.e., all *Arabidopsis* proteins from TIGR combined with all other dicotyledonous proteins present in SWISS-PROT [Boeckmann et al., 2003]). To detect homologous monocot gene sequences that contain sufficient information to reconstruct a reliable phylogenetic tree, two selection criteria were applied. First, all hits for both rice gene duplicates had to have an E value of <1e-10. Second, only sequences that had an alignable region of >150 amino acids with the query rice sequences were selected for the final phylogenetic analysis. The total set of monocotyledonous protein sequences contains 18,885 proteins, which were obtained by selecting SWISS-PROT proteins for *Triticum*, *Sorghum*, *Hordeum*, *Zea*, and *Avena*, translation of coding sequences from the National Center for Biotechnology Information (NCBI) Unigene collection for *Hordeum vulgare, Triticum aestivum*, and *Zea mays*, and the construction of open reading frames that show sequence similarity to rice proteins (BLASTX with E values of <1e-05) for all publicly available monocotyledonous ESTs and NCBI Unigenes lacking coding sequence information. Phylogenetic trees clearly in disagreement with the established grass phylogeny (Kellogg, 2001) or showing nonsignificant bootstrap values (<70%) were removed from further analysis.

2.4 Legume promoter sequences reveal reciprocal *cis*-regulatory divergence in *Arabidopsis* gene duplicates

Klaas Vandepoele, Cindy Martens, Cedric Simillion and Yves Van de Peer

Large-scale gene duplication events play an important role in the genome evolution of all major plant groups and many other eukaryotic organisms. Consequently, the substantial increase of raw genetic material associated with genome or chromosome duplications provides a potential source for evolutionary innovation. A set of duplicated regions derived from the youngest genome duplication in *Arabidopsis thaliana* was compared with homologous legume segments, which provide valuable information about the ancestral genome organization. A degraded network of microcolinearity was found between *Arabidopsis* segments and homologous Fabaceae genomic regions. Detailed comparative promoter analysis revealed that loss of *cis*-regulatory elements in duplicated genes is not completely random, but occurs according to a reciprocal pattern, with a complementary set of motifs partitioned over both paralogs as a result. For most gene duplicates with a high degree of reciprocal promoter divergence, a clearly dissimilar expression pattern was found, which is compatible with an evolutionary model predicting subfunction partitioning after gene duplication. Consequently, high levels of reciprocal promoter divergence, detectable through the comparison of gene duplicates with a suitable ortholog, are a good indicator for subfunctionalization after gene duplication. Finally, the biochemical function of a gene not only contributes to the survival rate of a gene duplicate, but also determines the *cis*-regulatory promoter complexity.

Introduction

The large number of duplicated genes and the discovery of ancient large-scale duplication events in a wide variety of eukaryotic model systems support the idea that gene and genome duplications are important drivers for biochemical innovation and evolution (Haldane, 1932; Stephens, 1951; Ohno, 1970; Stebbins, 1971). In plants, recent polyploidy events have been described in several species, such as *Triticum aestivum* (wheat), *Gossypium hirsutum* (cotton), and *Brassica* sp. (for an overview, see Wendel, 2000), whereas remnants of paleopolyploidy events have been uncovered in several monocotyledonous and dicotyledonous lineages (Gaut and Doebly, 1997; Bowers et al., 2003; Vandepoele et al., 2003; Paterson et al., 2004; Blanc and Wolfe, 2004; Sterck et al., 2005). Because paralogous genes often experience relaxed evolutionary constraints following duplication, most duplicates get lost through deleterious mutations (non-functionalization). Estimates in *Arabidopsis thaliana* and *Oryza sativa* (rice) indicate that only 21-34% of all gene duplicates have been retained after large-scale duplication events (Simillion et al., 2002; Blanc et al., 2003; Vandepoele et al., 2003; Paterson et al., 2004). Similarly, detailed comparison of homeologous regions in *Zea mays* (maize), which experienced a tetraploidy event approximately 11 million years ago (MYA), with orthologous segments of rice and *Sorghum bicolor* (sorghum) revealed 20-50% retention of gene duplicates (Ilic et al., 2003; Lai et al., 2004). Although excessive cases of fractionation through gene loss, leading to almost no retention at all, have been reported in maize (Langham et al., 2004), such extreme gene loss is probably exceptional. Intraspecies analysis of *Arabidopsis* gene duplicates formed by ancient polyploidy events also indicates that the process of gene loss is not random, but biased toward particular gene functions (Blanc and Wolfe, 2004; Seoighe and Gehring, 2004; Maere et al., 2005). In addition to gene loss, other structural modifications, such as tandem array expansions, recombination and local chromosomal rearrangements through inversions or translocations, are also responsible for the dynamic nature of plant chromosomes and explain why the chromosomal positions of genes can be quite different in closely related species (Ziolkowski et al., 2003; Lai et al., 2004).

Ohno's classical model (1973) predicts that, besides nonfunctionalization, a duplicated gene might acquire a new function through the accumulation of a series of non-deleterious mutations (neofunctionalization). More recently, Force et al. (1999) proposed the duplication-degeneration-complementation (DDC) model that predicts that degenerative mutations preserve gene duplicates by changing their functions (Force et al., 1999). The DDC model predicts that the likelihood of preservation is correlated with the number of “subfunctions” that can be ascribed to a gene. The model starts from the assumption that a gene can perform several different functions; for instance, genes expressed in different tissues and at different times during development may be controlled by different DNA regulatory elements. When duplicated genes lose different regulatory subfunctions, each affecting different spatial and/or temporal expression patterns, they must complement each other by jointly retaining the full set of subfunctions that were present in the ancestral gene. Therefore, degenerative mutations facilitate the retention of duplicated genes, in which both duplicates now perform different, but necessary, subfunctions. As predicted by the DDC model, the sum of the subfunctions associated with each of the retained duplicates must be the same as the total number of subfunctions performed by the ancestral gene. Gene duplication then allows each daughter gene to specialize for one (or more) of the functions of the ancestral genes.

The DDC model is attractive because it suggests a mechanism through which both duplicates can be preserved in the genome and seems to fit well with the large number of duplicated genes present in most eukaryotic genomes (Force et al., 1999; Lynch and Conery, 2000). Consequently, degenerative nucleotide substitutions can promote functional divergence after gene duplication affecting gene expression or protein function (Mena et al., 1996; Prince and Pickett 2002; Adams et al., 2003; Van de Peer et al., 2003). The observation that in particular genes involved in signal transduction and transcription are preferentially retained in Arabidopsis (Blanc and Wolfe, 2004; Seoighe and Gehring, 2004; Maere et al., 2005) is in accordance with the view that the *cis*-regulatory evolution of transcriptional regulators provides a predominant mechanism for generating novel phenotypes and genetic diversity (Doebley and Lukens, 1998). In *Arabidopsis*, Blanc and Wolfe (2004) found that more than 50% of all duplicates that originated from the youngest genome duplication have diverged in expression, while a significant rate of divergence in the duplicated protein sequence was observed in 21% of all analyzed paralogs. In contrast, Zhang et al. (2002) found little evidence for rate differences in paralogs and no evidence for positive selection. Despite these observations, the evolutionary mechanisms responsible for gene expression changes in duplicated plant genes remain unclear, mainly because intraspecies comparisons offer only a limited resolution (Haberer et al., 2004).

In order to unravel the dynamics of both structural and functional divergence in duplicated regions, a set of duplicated segments (paralogons) of the *Arabidopsis* genome that arose through whole-genome duplication was compared with homologous genomic regions of three legume species. Because the leguminous plant species have diverged from Arabidopsis prior to its youngest genome duplication (Blanc et al., 2003; Bowers et al., 2003; Ermolaeva et al., 2003), the genomic organization of *Fabaceae* can provide novel insights into the genome evolution of the ancient polyploid *Arabidopsis*. Furthermore, based on a comparative promoter analysis, we investigated the evolution of promoters in gene duplicates of Arabidopsis and verified whether subfunctionalization at the regulatory level is likely to affect the evolution of duplicates in plants.

Results

Microcolinearity between Arabidopsis and Fabaceae species

There is compelling evidence that the duplicated chromosomal segments resulting from the youngest genome duplication in *Arabidopsis* were formed some 20-60 MYA, before the split of the *Arabidopsis* and *Brassica* lineages, but after the divergence of the *Brassicacea* and *Malvaceae* (Simillion et al., 2002; Blanc et al., 2003; Bowers et al., 2003; Ermolaeva et al., 2003). Because of the absence of sequencing projects for species of the *Malvaceae* family, genomic data of the earlier diverging *Fabaceae* can provide valuable information about the ancestral genome organization before the youngest polyploidization event in *Arabidopsis*. Therefore, we selected 126 bacterial artificial clone (BAC) sequences from *Medicago truncatula*, 115 of *Lotus corniculatus* var. japonicus (lotus), and 20 of *Glycine max* (soybean) to determine the degree of conservation in gene content and order (colinearity) with duplicated *Arabidopsis* segments. For all 261 BAC sequences (approximately 26.7 MB), gene models were determined by using EuGène, a software tool that combines extrinsic homology-based structural annotation with ab-initio gene prediction (Schiex et al., 2000; see Materials and methods). In total, 5,269 genes were annotated, corresponding with gene densities of 1 gene per 5.9 kb, 4.4 kb, and 5.1 kb, in *M. truncatula*, lotus, and soybean, respectively. Approximately 81% of all predicted proteins have significant similarity (BLAST E-value <1e-10) with one or more publicly available protein or expressed sequence tags (ESTs). In order to identify *Fabaceae* genomic segments that could provide additional information on the preduplicate genome organization in *Arabidopsis*, we identified all colinear segments between the *Arabidopsis* genome and all *Fabaceae* BAC sequences using i-ADHoRe, a novel method that creates profiles by combining gene content and order information of multiple homologous genomic segments to detect colinearity in one or more genomes (Simillion et al., 2004). We found significant colinearity between Arabidopsis and 14% of all legume BACs, with the highest number of colinear BACs in soybean (4/20) and the lowest in lotus (14/115). In total, we identified 19 sets of homologous segments (multiplicons) grouping two duplicated Arabidopsis segments with a homologous *Fabaceae* segment, covering 8.5% of all initially selected leguminous genomic sequences.

Structural divergence in duplicated segments

For all valid multiplicons that were clearly colinear between one single legume BAC and two duplicated chromosomal segments in *Arabidopsis*, all homologous gene strings, which represent the gene organization of the different homologous segments, were aligned with i-ADHoRe (Figure 2.4.1; Simillion et al., 2004). For each multiplicon, phylogenetic trees were constructed for all retained gene duplicates in *Arabidopsis* with a conserved homolog on the *Fabaceae* segment to verify if the homologous *Fabaceae* segment was indeed an outgroup to both duplicated *Arabidopsis* segments. For 36 out of 47 sets of homologous genes conserved on all segments of a particular multiplicon, a suitable outgroup sequence was found for the construction of a neighbor-joining phylogenetic tree. Three multiplicons contained each one gene leading to an unexpected tree topology, whereas all other genes (92%) had the expected tree topology, supporting a duplication event in the *Brassicaceae*, after its divergence from the *Fabaceae*. Consequently, for all 19 multiplicons, the *Fabaceae* segment can provide valuable information about the ancestral gene organization of the duplicated *Arabidopsis* segments. Additional information about the gene annotation, multiplicons and phylogenetic trees can be found at http://bioinformatics.psb.ugent.be/.

By using the genomic profiles and after discarding all genes involved in local tandem duplications, we identified 47 retained *Arabidopsis* gene duplicates (designated *Ath* duplicates) and 59 genes for which one of the duplicates had been lost (designated single-copy (SC) genes) (Figure 2.4.1; Supplemental Table 1). For most duplicated blocks, the number of SC gene duplicates on both segments

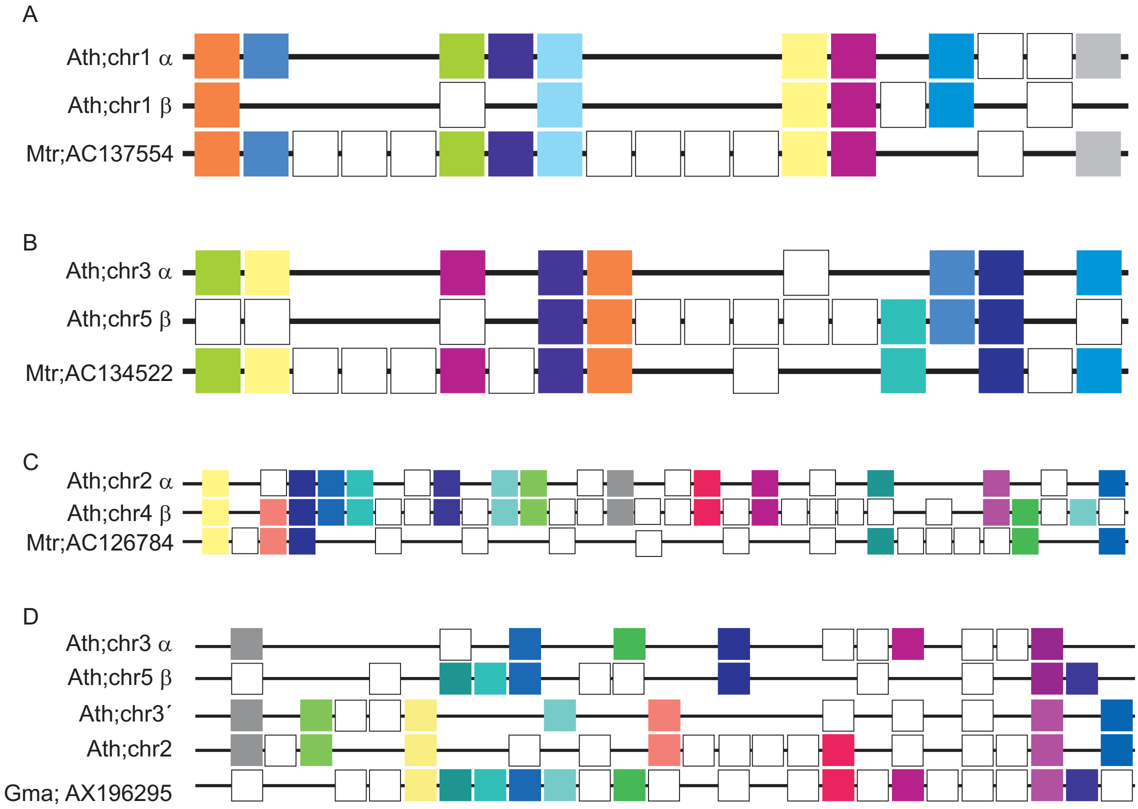


Figure 2.4.1 Microcolinearity between two paralogous *Arabidopsis* segments and one homologous legume segment. The genes on the chromosomal segments (black lines) are represented by boxes that are colored for different gene families and white when no homolog on one of the other segments could be found. Note that tandem duplications are remapped to a single gene. Gene coordinates (a) segment alpha start: At1g72700.1; stop: At1g72810.1; segment beta start: At1g72810.1; stop: At1g17560.1 (b) segment alpha start: At3g05530.1; stop: At3g05600.1; segment beta start: At5g27730.1; stop: At5g27860.1 (c) segment alpha start: At2g23940.1; stop: At2g24130.1; segment beta start: At4g30500.1; stop: At4g30820.1 (d) segment alpha start: At3g11050.11; stop: At3g11170.1; segment beta start: At5g05480.1; stop: At5g05590.1; segment chr3' start: At3g56090.1; stop: At3g55980.1; segment chr2 start: At2g40300.1; stop: At2g40140.1

was very similar, indicating that the amount of gene loss on both segments was balanced, as also previously observed in paleopolyploid genomes (Langkjaer et al., 2003; Dietrich et al., 2004; Kellis et al., 2004). However, there were some indications that, besides individual gene losses or insertions, also larger segmental insertion or deletion events have occurred as well (e.g. segment Ath; chr5 β of multiplicon B in Figure 2.4.1). The overall retention rate of gene duplicates on all 19 pairs of duplicated segments is 38%, whereas the fraction of gene duplicates that returned to single-copy state accounts for 13% of all genes. The evolutionary history of the remaining genes in these duplicated segments is unclear. On average, each legume BAC in these 19 multiplicons shares 6.4 homologous genes with the duplicated *Arabidopsis* segments, corresponding to 37% conservation of colinearity between legumes and *Arabidopsis*.

For some duplicated blocks, a large fraction of retained *Arabidopsis* duplicates without homologs on the corresponding homologous *Fabaceae* segment could be found (Figure 2.4.1c), suggesting that additional, yet unrevealed, homologous segments might exist. Consequently, it is important to realize that the homologous legume segment within a multiplicon does not necessarily represent the complete ancestral genome organization, because the ancestral gene content might be spread over multiple chromosomal regions due to translocations or legume-specific duplication events (Zhu et al., 2003).

Identification of cis-acting regulatory elements through phylogenetic footprinting

Apart from analyzing differential loss of duplicated genetic material (also known as fractionation) at the gene level (Vandepoele et al., 2002b), detected colinearity between legumes and *Arabidopsis* paralogons can also be used to investigate whether fractionation acts on *cis*-acting units of a gene (Langham et al., 2004; Lockton and Gaut, 2005). Consequently, analysis of intragenic loss of *cis*-acting functions in duplicated genes through comparison with legume outgroups might reveal subfunctionalization, in which a complementary set of regulatory elements has been degenerated. To investigate this hypothesis, phylogenetic footprints, i.e. non-coding sequences that are unusually well conserved because of some functional constraint, were identified in 1000 bp promoter sequences with the shared motif method (SMM; Castillo-Davis et al., 2004). Briefly, this method quantifies conserved motifs in upstream regions of homologous genes by using a recursive local alignment algorithm, without respect to their order, orientation, or spacing. Subsequently, the fraction of shared motifs between both promoters is calculated and the promoter divergence (d_{SM}) is defined as 1 minus the shared fraction (see Figure 2.4.2). Although this method reports overlapping motifs, it is important to note that this does not affect the way the promoter divergence is calculated (see Materials and methods; Castillo-Davis et al., 2004). A minimum alignment score corresponding with at least 14 matches was applied for the detection of conserved elements. Despite the existence of smaller *cis*-regulatory elements in plants, the very high false-positive rate associated with the computational identification of such small elements makes it impossible to consider them in systematic promoter analyses (see Materials and methods for details).

Initially, the promoter sequences of all SC and *Ath* duplicates were compared with those from the homologous legume promoters. The average number of detected motifs between Arabidopsis and legume promoters was 6.3, whereas the average d_{SM} was 0.88 (± 0.12). No significant differences in d_{SM} were observed when SC or *Ath* duplicate promoter sequences were compared with the legume promoters (data not shown). When comparing the promoter sequences of both *Ath* duplicates, on average 6.5 conserved motifs were found, whereas the average paralog promoter divergence (PPd_{SM}) was 0.80 (± 0.13). For a fraction of duplicates, the d_{SM} values differed strongly between both *Arabidopsis* promoters, when compared to the legume promoter (data not shown). In addition, although both duplicates might share a similar number of conserved promoter elements with the homologous legume promoter, the motifs might be different (Figures 2.4.2 and 2.4.3). Therefore, we defined a new measure, referred to as d_R, to quantify the amount of reciprocal sequence divergence between two paralogous promoter sequences. When an identical set of motifs is identified between both duplicates and their homologous legume promoter, then d_R will be zero, even when, theoretically, d_{SM} can be quite large. In contrast, when the (ancestral) legume promoter contains, for example, five promoter elements (a, b, c, d, and e), and gene duplicates alpha and beta have three conserved elements ([c, d, and e] and [a, b, and c], respectively), then the fraction of shared elements will be 1/5 (namely element c) and the reciprocal divergence d_R will be 0.8 (1-1/5 = 4/5, assuming identical motif sizes; Figure 2.4.2). Thus, whereas d_{SM} is a general measure for promoter sequence divergence, d_R is a measure describing the differential pattern of motif loss in the promoter sequences of gene duplicates (Figures 2.4.2 and 2.4.3).

Fifty percent of all retained *Ath* duplicates, for which both copies still have some conserved elements shared with the homologous legume promoter, had a d_R > 0.8 (Table 2.4.1), implying that many duplicates derived from the youngest genome duplication have lost a different, but complementary, set of *cis*-regulatory elements in their promoter. The *Ath* duplicates with the lowest reciprocal promoter divergence encodes a galactosyltransferase (At1g53290/At3g14960; d_R=0.187), whereas 13 duplicates with complete reciprocal divergence (d_R=1) were identified, coding for a variety of gene products (such as protein kinase, ABC transporter, GTP-binding protein, glycosyl hydrolase, and polynucleotide adenylyltransferase).

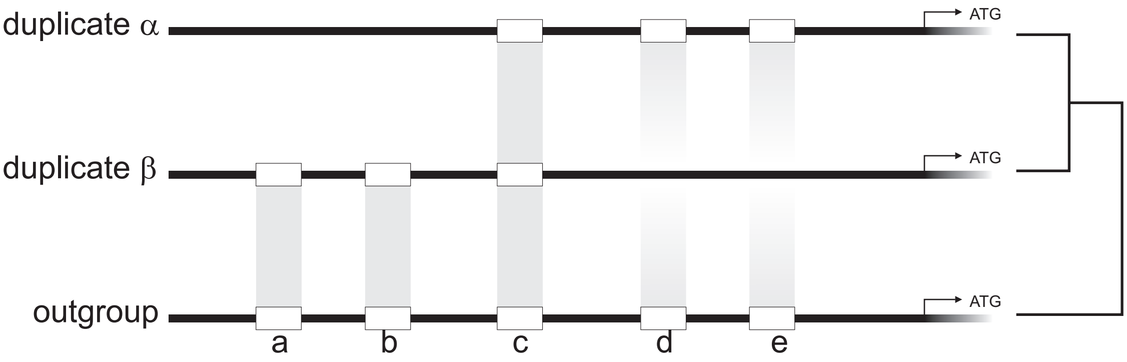


Figure 2.4.2 Investigating *cis*-regulatory evolution in gene duplicates with an outgroup promoter sequence. The three black bars indicate the promoter sequences (all 500 nt long) and white boxes the conserved *cis*-regulatory elements. In this hypothetical example, in which the actual size of each shared motif is considered identical (20 nucleotides), one ancestral motif (c) is shared between both paralogs and results in a paralog promoter divergence PPd_{SM} = 0.96 (1-[1 motif x 20 nt per motif/500 nt]). Although both duplicated genes have three motifs shared with the outgroup (d_{SM}=0.88; 1-[3 motifs x 20 nt per motif/500 nt]), which reflects the ancestral pre-duplicated state, four out of five *cis*-regulatory elements (i.e. a, b, d, and e) have been complementary partitioned over both paralogs, leading to an reciprocal promoter divergence d_R = 0.8.

When excluding four pairs of *Ath* duplicates for which no conserved motifs could be identified, a significant positive correlation was found between the paralog promoter divergence PPd_{SM} and d_R (Spearman rank correlation N=43, rho=0.53, P<0.001; Pearson correlation R=0.52, P<0.001; Figure 2.4.4a). This indicates that a reduced number of shared motifs between both *Ath* paralogs is correlated with an increasing amount of reciprocal divergence. Although this pattern might seem logical and expectable, it is interesting to note that gene duplicates of similar ages (73% of all *Ath* paralogous genes have a K_s value between 0.6 and 1.1; see Table 1) had very different degrees of complementary motif loss (Figure 2.4.4).

In order to investigate whether promoter divergence is correlated with the age of duplicates, a second data set was created of gene duplicates including many younger genes (see Supplemental table 2). Based on the observation that orthologous gene pairs between *Arabidopsis* and the *Fabaceae* have an average K_s of 1.83 (data not shown; Blanc et al., 2003), all gene families were identified (Methods) with only two members in the *Arabidopsis* genome that have a K_s<1.8, plus a legume homolog. In total, we selected 133 gene families grouping two *Arabidopsis* duplicates and a legume homolog, spanning a K_s range between 0.01 and 1.74. As described above for the set of *Ath* duplicates derived from the 19 multiplicons, the shared motif method was again applied to determine the

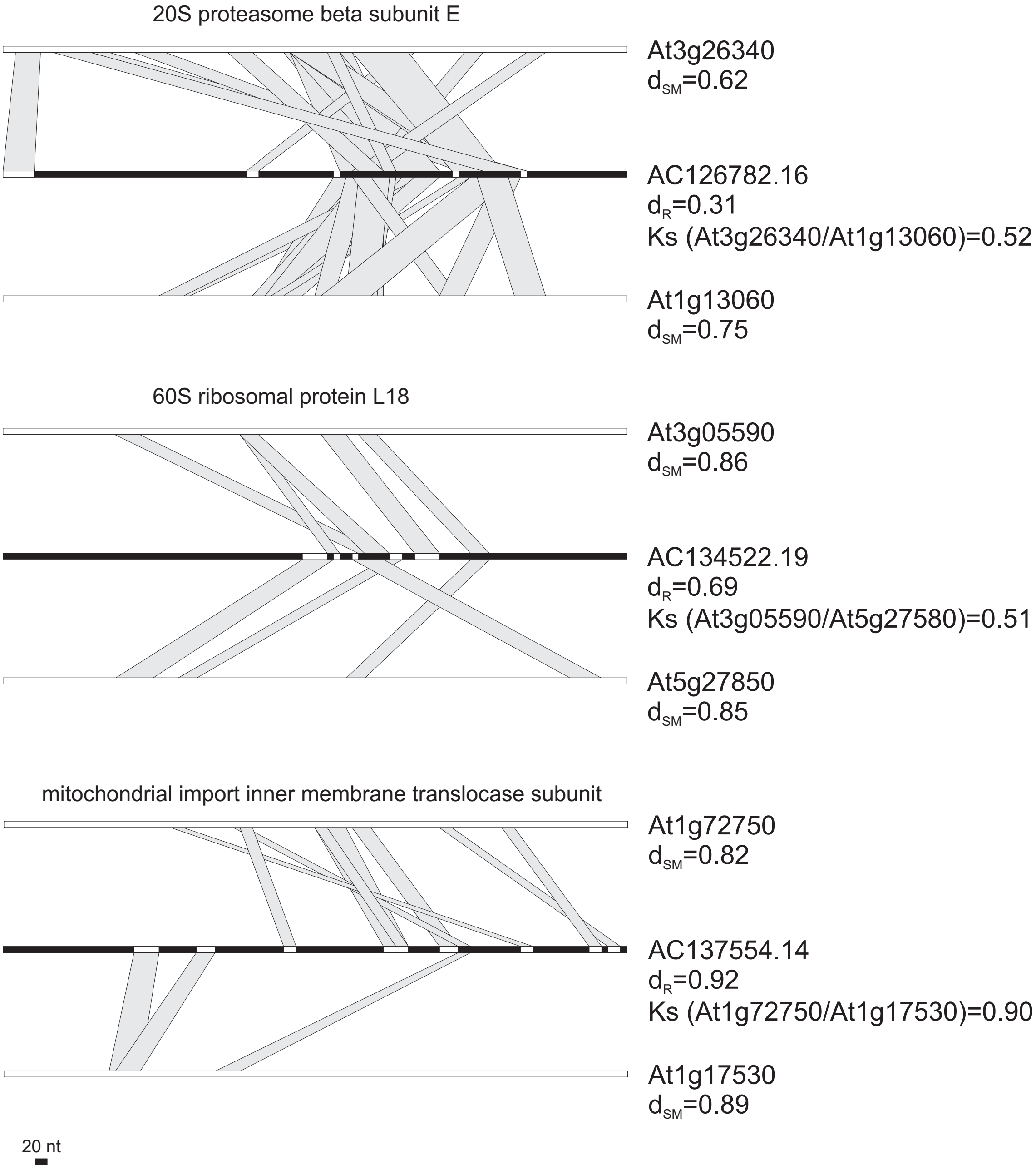


Figure 2.4.3 Examples of reciprocal promoter divergence in *Arabidopsis* gene duplicates. The horizontal white and black bars represent the promoters of the duplicated genes and of the legume homolog, respectively. Significant phylogenetic footprints are shown as grey bands and the set of motifs that diverged reciprocally is indicated in white on the legume promoter. d_{SM} gives the promoter divergence between the legume and the *Arabidopsis* paralog and the K_s values the age of a gene duplicate.

Table 2.4.1 Overview of paralog promoter divergence (PPd$_{SM}$) and reciprocal promoter divergence (d$_R$) for 45 retained gene duplicates in *Arabidopsis* for which at least one paralog contains shared motifs with the homologous legume promoter.

Legume locus	Duplicate 1	Duplicate 2	PPdSM	Ks	Gene description	d$_R$	GO label	GO Description	R[a]	E(Duplicate 1)[b]	E(Duplicate 2)[b]
AC125477.10	At4g30170	At2g18980	0.71	0.8	peroxidase, putative	0.716	GO:0016209 / GO:0003824	antioxidant activity / catalytic activity	0.75		
AC125477.16	At4g30130	At2g19090	0.72	0.8	expressed protein	0.718	GO:0005554	molecular_function unknown	n.a.		
AC125477.6	At4g30190	At2g18960	0.62	0.4	ATPase 2, plasma membrane-type, putative	0.856	GO:0005215 / GO:0016787	transporter activity / hydrolase activity	0.21		
AC126784.15	At2g23950	At4g30520	0.58	0.7	leucine-rich repeat family protein / protein kinase family protein	0.661	GO:0016301	kinase activity	0.17		
AC126784.18	At2g23940	At4g30500	0.96	0.7	expressed protein	0.831	GO:0005554	molecular_function unknown	0.04		
AC134522.15	At3g05550	At5g27760	1	0.6	hypoxia-responsive family protein	1	GO:0005554	molecular_function unknown	n.a.		
AC134522.19	At3g05590	At5g27850	0.83	0.5	60S ribosomal protein L18 (RPL18B)	0.686	GO:0005198	structural molecule activity	0.58		
AC135319.6	At3g05280	At5g27490	0.71	0.6	integral membrane Yip1 family protein	1	GO:0005554	molecular_function unknown	0.19	459 (0.97)	427 (0.90)
AC135565.11	At3g04910	At5g28080	0.74	0.9	protein kinase family protein	0.625	GO:0004672	protein kinase activity	0		
AC135565.19	At3g04860	At5g28150	0.77	0.5	expressed protein	0.892	GO:0005554	molecular_function unknown	0		
AC135565.5	At3g04920	At5g28060	0.9	0.5	40S ribosomal protein S24 (RPS24A)	0.74	GO:0005198	structural molecule activity	0.33		
AC136973.4	At2g47000	At3g62150	0.91	0.9	multidrug resistant (MDR) ABC transporter, putative	1	GO:0003674	molecular function	0.25	302 (0.64)	351 (0.74)
AC136973.9	At2g47060	At3g62220	0.9	0.6	serine/threonine protein kinase, putative	1	GO:0016301	kinase activity	0	414 (0.88)	182 (0.38)
AC137554.1	At1g72700	At1g17500	0.58	0.6	haloacid dehalogenase-like hydrolase family protein	n.d.	GO:0005215 / GO:0016787	transporter activity / hydrolase activity	n.a.		
AC137554.14	At1g72750	At1g17530	0.84	0.9	mitochondrial import inner membrane translocase subunit	0.919	GO:0005215	transporter activity	0	458 (0.97)	425 (0.90)
AC137554.15	At1g72760	At1g17540	0.82	0.7	protein kinase family protein	1	GO:0016301	kinase activity	0.12	128 (0.27)	467 (0.99)
AC137554.8	At1g72740	At1g17520	0.84	0.8	DNA-binding family protein / histone H1/H5 family protein	0.997	GO:0003700	transcription factor activity	0.03	430 (0.91)	119 (0.25)
AC146681.1	At2g20290	At4g28710	0.86	0.4	myosin, putative	n.d.	GO:0003774 / GO:0005515	motor activity / protein binding	n.a.		

AC146681.6	At2g20260	At4g28750	0.92	0.7	photosystem I reaction center subunit IV, chloroplast	1	GO:0005554	molecular_function unknown	0.87	449 (0.95)	433 (0.92)
AC146681.7	At2g20240	At4g28760	0.93	0.7	expressed protein	0.825	GO:0005554	molecular_function unknown	0.06		
AC146852.20	At3g22750	At4g14780	0.72	0.8	protein kinase, putative	0.615	GO:0016301	kinase activity	0.1		
AC146852.5	At3g22530	At4g14830	0.66	0.9	expressed protein	0.387	GO:0005554	molecular_function unknown	n.a.		
AC148293.10	At2g21230	At4g38900	0.73	1	bZIP family transcription factor	0.883	GO:0003677 / GO:0003700	DNA binding / transcription factor activity	0.12		
AP004512.12	At1g73620	At1g18250	0.86	0.7	thaumatin-like protein, putative / pathogenesis-related protein	n.d.	GO:0005554	molecular_function unknown	n.a.		
AP004512.13	At1g73630	At1g18210	0.8	1.1	calcium-binding protein, putative	0.909	GO:0005509	calcium ion binding	0.04	404 (0.85)	461 (0.97)
AP004512.15	At1g73640	At1g18200	0.98	0.6	Ras-related GTP-binding family protein	1	GO:0000166	nucleotide binding	0.13	411 (0.87)	416 (0.88)
AP004512.17	At1g73650	At1g18180	1	0.6	expressed protein	1	GO:0005554	molecular_function unknown	0	457 (0.97)	97 (0.21)
AP004512.18	At1g73655	At1g18170	0.8	0.9	immunophilin / FKBP-type peptidyl-prolyl cis-trans isomerase family	0.493	GO:0003824 / GO:0005488	catalytic activity / binding	0.54		
AP004512.19	At1g73660	At1g18160	0.81	0.5	protein kinase family protein	1	GO:0016301	kinase activity	0.02	417 (0.88)	454 (0.96)
AP006392.16	At3g48830	At5g23690	0.78	0.5	polynucleotide adenylyltransferase family protein / RNA recognition motif (RRM)-containing protein	1	GO:0003723 / GO:0016740	RNA binding / transferase activity	0.04	359 (0.76)	318 (0.67)
AP006392.9	At3g48800	At5g23680	0.9	0.7	sterile alpha motif (SAM) domain-containing protein	1	GO:0005554	molecular_function unknown	n.a.		
AP006629.6	At1g23490	At1g70490	0.66	0.6	ADP-ribosylation factor	0.808	GO:0003674	molecular function	0.51		
AP006629.9	At1g23460	At1g70500	0.74	0.6	polygalacturonase, putative / pectinase, putative	n.d.	GO:0016787	hydrolase activity	n.a.		
AP006654.16	At1g53290	At3g14960	0.68	0.7	galactosyltransferase family protein	0.187	GO:0016740	transferase activity	0.08		
AP006654.18	At1g53300	At3g14950	0.54	0.7	thioredoxin family protein	0.457	GO:0005554	molecular_function unknown	0.027		
AP006654.23	At1g53310	At3g14940	0.78	0.6	phosphoenolpyruvate carboxylase, putative / PEP carboxylase, putative (PPC1)	0.544	GO:0016301	kinase activity	0.07		

AX196295.12	At2g40370	At5g05390	0.85	1.7	laccase, putative / diphenol oxidase, putative	0.672	GO:0003824 / GO:0005488	catalytic activity / binding	0.12		
AX196295.16	At2g40330	At5g05440	0.94	1.8	Bet v I allergen family protein	0.922	GO:0005554	molecular_function unknown	0.53	171 (0.36)	396 (0.84)
AX196295.23	At3g11040	At5g05460	0.82	0.6	glycosyl hydrolase family 85 protein	1	GO:0016787	hydrolase activity	0.1	359 (0.76)	303 (0.64)
AX196295.30	At2g40290	At5g05470	1	1.3	eukaryotic translation initiation factor 2 subunit 1, putative	0.738	GO:0003723	RNA binding	0.39		
AX196295.31	At2g40280	At3g56080	1	1.9	dehydration-responsive family protein	0.935	GO:0005554	molecular_function unknown	0.04	455 (0.96)	293 (0.62)
AX196295.35	At2g40270	At3g56050	0.8	0.7	protein kinase family protein	0.507	GO:0016301	kinase activity	0.56		
AX196295.38	At3g11070	At5g05520	0.94	0.7	outer membrane OMP85 family protein	1	GO:0005554	molecular_function unknown	n.a.		
AX196295.54	At3g11180	At5g05600	0.88	0.8	oxidoreductase, 2OG-Fe(II) oxygenase family protein	0.869	GO:0003824	catalytic activity	0		
AX196297.16	At4g14130	At3g23730	0.57	0.8	Xyloglucan:xyloglucosyl transferase, putative	0.708	GO:0016787	hydrolase activity	0.42		

n.d. No shared motifs could be found for both duplicates between the *Arabidopsis* and the outgroup *Fabaceae* promoter sequence.

n.a. No expression for both paralogs available.

[a] Pearson correlation coefficients are calculated based on 473 expression values retrieved from NASC microarrays using Genevestigator (Zimmernann et al., 2004).

[b] The number of experiments the gene was found expressed when querying 473 NASC microarrays using Genevestigator (Zimmernann et al., 2004). The percentage is indicated in parenthesis.

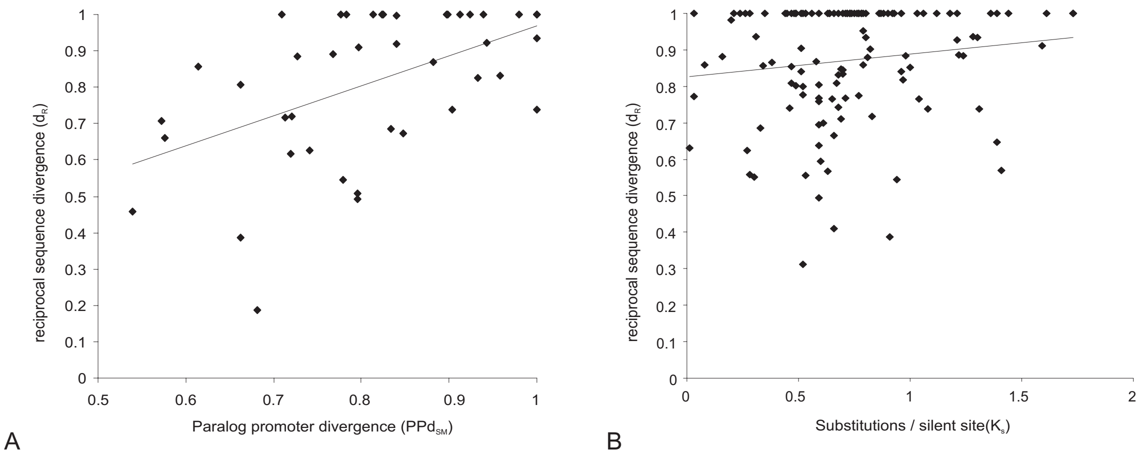


Figure 2.4.4 Correlation between paralog promoter divergence, age, and reciprocal promoter divergence of Arabidopsis gene duplicates. The black line indicates a linear fit of the data. (a) The Paralog Promoter divergence (PPd_{SM}) measured between paralogous promoters derived from the 19 multiplicons plotted against the reciprocal promoter divergence (d_R). Spearman rank correlation N=43, rho=0.53, P<0.001; Pearson correlation R=0.52, P<0.001. (b) Age (K_s) of 125 gene duplicates plotted against the reciprocal promoter divergence. Spearman rank correlation N=125, rho=0.14, P=0.12; Pearson correlation R=0.14, P=0.13.

number of motifs conserved between *Arabidopsis* and legume promoters and to calculate the reciprocal motif divergence for all 133 paralogous gene pairs (designated *Ath'* duplicates). Including more recently duplicated genes, no correlation was found between the age of a gene duplicate and the fraction of motifs shared between each of the promoters of the *Ath'* duplicates and the legume homolog, defined as 1-d_{SM} (Spearman rank correlation N=170, rho= -0.014, P=0.86; Pearson correlation R=0.002, P=0.98; Supplemental Figure 1). In a second step, we tested the relationship between the age of *Ath'* duplicates and the reciprocal promoter divergence. Based on 125 gene duplicates containing shared motifs with the homologous legume sequence, no significant correlation between K_s and d_R (Spearman rank correlation N=125, rho=0.14, P=0.12; Pearson correlation R=0.14, P=0.13; Figure 2.4.4b) was observed. These results reveal that reciprocal promoter divergence and the age of duplicates are not strongly related.

We also investigated whether the number of identified *cis*-regulatory elements, again defined as conserved motifs with ≥14 identical nucleotides present in the upstream regions, differs for genes with different biochemical functions. To this end, we identified 422 pairs of single-copy *Arabidopsis* and legume homologs (Supplemental Table 3), and compared the promoter sequences as described before. When plotting the distribution of the number of shared motifs for all functional gene ontology (GO) classes with more than three genes, a difference

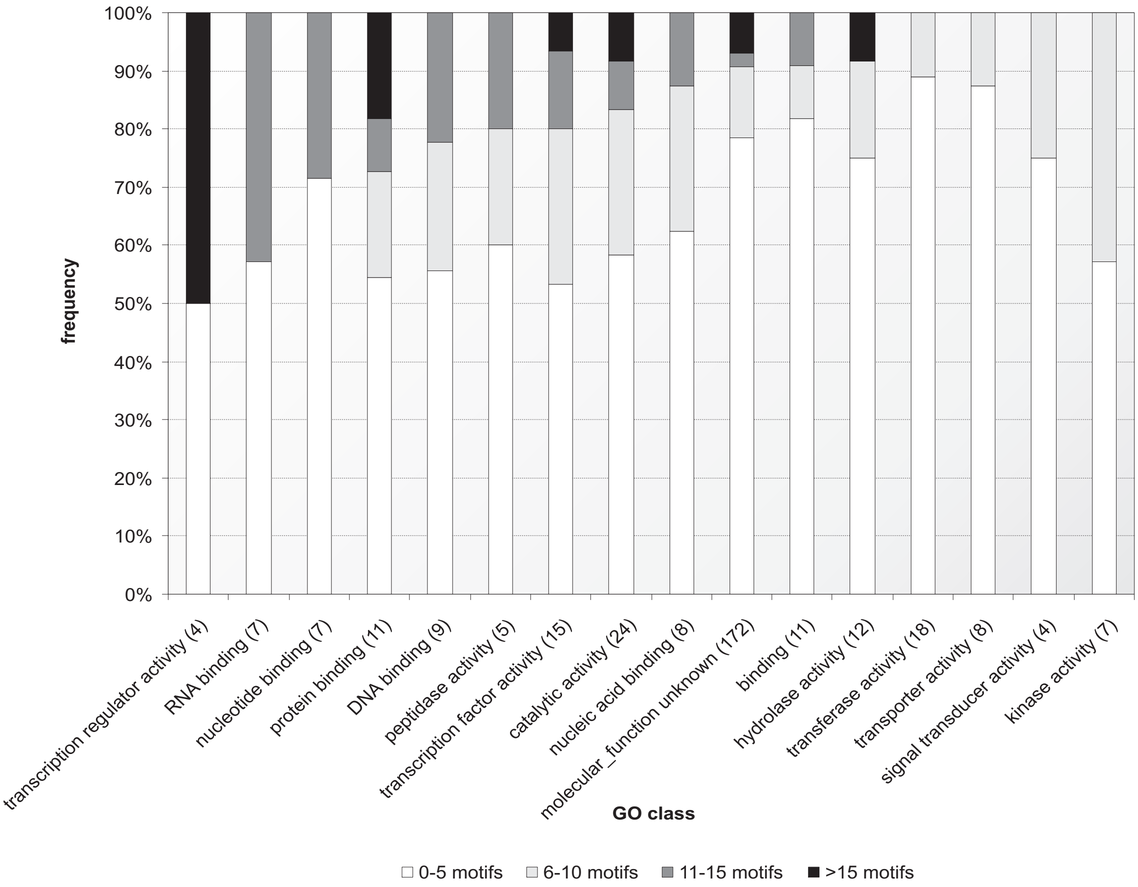


Figure 2.4.5 Distribution of the number of shared regulatory motifs identified in 422 orthologs sorted by functional GO class. For each functional GO class containing more than three genes, the motif composition is shown grouped by the number of motifs, with the number of orthologous gene pairs assigned to a particular functional GO class in parentheses.

between the relative abundance of motifs per functional category can indeed be observed (Figure 2.4.5). Most genes showing hydrolase, transferase, or transport activity contained only a limited number of regulatory elements (typically between 0 and 5 motifs with a length of 14 nucleotides or more), whereas other classes of proteins showing peptidase activity or catalytic activity possessed a wide variety of simple and complex promoters (with a small or large number of elements, respectively). Finally, a substantial number of genes (43-50%) involved in protein binding, DNA and RNA binding, or showing transcription regulator activity or transcription factor activity, seemed to contain a large number of conserved regulatory elements (>5-10 motifs), suggesting a more complex transcriptional control (Inada et al., 2003).

Does promoter divergence reflect divergence of expression?

Comparison of the degree of reciprocal promoter divergence with the expression correlation of both *Ath* paralogs using genome-wide expression data (Zimmermann et al., 2004) showed a weak negative, but nonsignificant correlation (Spearman rank correlation N=36, rho=-0.23, P=0.18; Pearson correlation R=-0.16, P=0.36; Table 2.4.1 and Supplemental Figure 2). Detailed analysis of expression levels for several gene duplicates was hindered because both paralogs were often not present in these genome-wide expression repositories. However, the Pearson correlation coefficients between the expression profiles for all *Ath* duplicates with d_R>0.9 in the different multiplicons were very low (≤0.25) in 13 out of 15 paralogous gene pairs (Table 2.4.1). Based on the distribution of correlation coefficients for a large number of functionally unrelated gene pairs, which revealed that 95% of these random pairs have a correlation coefficient <0.52 (Blanc and Wolfe, 2004), this indicates that these duplicates are no longer significantly co-regulated. Although this observation seems to suggest that the divergence in expression pattern is caused by the reciprocal promoter divergence, these results actually provide little information about the nature of this functional divergence. Therefore, the expression levels for these 13 gene pairs were retrieved for 473 NASC microarrays (Digital Northern tool; Zimmermann et al., 2004) and the occurrence of expression of both paralogs over the different experiments (i.e. expression breadth) was compared. For one third of these gene pairs, showing high reciprocal promoter divergence and low expression similarities, the expression breadth of one paralog was significantly underrepresented, which indicates that the low Pearson expression correlation coefficient is mainly the result of the reduced expression of one duplicate in the different experiments (Table 2.4.1). For the remaining pairs (9/13), we observed a similar breadth of expression throughout the 473 experiments, revealing that the low Pearson expression correlation coefficient of these gene pairs, which have lost a complementary set of regulatory motifs, is caused by a clearly differential, but quantitatively similar, expression of both paralogs (Table 2.4.1).

Discussion

Degraded microcolinearity between Arabidopsis and legume species

Comparative sequence analysis at the sub-megabase level indicates that microcolinearity is abundant between closely related plant species. Conserved gene content and order has been demonstrated between tomato and *Arabidopsis*, which diverged approximately 112 MYA (Ku et al., 2000), between *Arabidopsis* and soybean (Grant et al., 2000), and between *Arabidopsis*, tomato, and *Capsella* (Rossberg et al., 2001). However, exceptions of interspecies colinearity caused by gene loss, duplications or rearrangements, such as inversions, deletions, and insertions, reveal the dynamic nature of this microstructure (Bennetzen, 2002; O'Neill and Bancroft, 2000; Ziolkowski et al., 2003; Vandepoele et al., 2004b).

Comparisons between related diploid genomes and duplicated segments that arose by large-scale duplication events provide novel insights into the intraspecies colinearity and the evolution of ancient polypolyploid species (Vandepoele et al., 2002b; Langkjaer et al., 2003; Dietrich et al., 2004; Kellis et al., 2004; Lai et al., 2004). This study provides a detailed comparative analysis of duplicated segments in the paleopolyploid genome of *Arabidopsis* and homologous genomic sequences of different legume species. After annotating 26.7 MB of genomic BAC sequences from *M. truncatula*, lotus, and soybean, only 8.8% of all legume BAC sequences were significantly colinear with an individual *Arabidopsis* segment. Interestingly, when genomic profiles are used that combine gene content and order information of multiple homologous genomic segments for the detection of genomic homology (Simillion et al., 2004), the number of legume BACs that are colinear with one or more *Arabidopsis* segments increased to 14%. Taking into account that nearly half of the legume BACs have fewer than 20 annotated genes, which greatly reduces the chance of finding significant interspecies colinearity, this finding confirms the existence of highly degenerated networks of microcolinearity between Arabidopsis and species belonging to the *Fabaceae* (Figure 2.4.1d; Yan et al., 2003; Zhu et al., 2003). By comparing 19 duplicated *Arabidopsis* segments with a homologous legume genomic region, an overall gene retention level of 38% has been found for gene duplicates that originated through the youngest genome duplication. This percentage is very similar to that previously obtained by comparing all duplicated segments created by the youngest polyploidy event in *Arabidopsis* (36.5%; Simillion et al., 2002).

The gene order and content in the legume BAC sequences reflects, at least to some extent, the genome organization prior to the youngest polyploidization event in *Arabidopsis*. Based on the number of colinear genes shared between at least two segments, the overall amount of non-colinear genes on the legume segments is 57%, whereas that on the *Arabidopsis* segments is 45%. Assuming that the divergence time between the *Brassicaceae* and the *Fabaceae* is 110 MYA (Wikström et al., 2001; Chaw et al., 2004; Davies et al., 2004) and that the youngest genome duplication in *Arabidopsis* occurred some 50 MYA (Simillion et al., 2002; Blanc et al., 2003; Bowers et al., 2003; Ermolaeva et al., 2003), the current data illustrate that the gene organization in the duplicated segments degrades faster than that of the interspecies colinearity (45% *Arabidopsis* intraspecies non-colinearity/50 MYA = 0.9% degradation per MYA and 57% *Arabidopsis*-legume non-colinearity/110 MYA = 0.52% degradation per MYA, respectively). This result confirms observations between related grass species, such as rice, sorghum, and maize, which have diverged approximately 50-70 MYA and where the interruption of intraspecies colinearity in the paleotetraploid maize is much greater than that in interspecies comparisons (Ilic et al., 2003; Lai et al., 2004). Also, the rate of gene loss associated with diploidization, the evolutionary process whereby a polyploidy species becomes a diploid again, seems to be higher in maize than in *Arabidopsis* (15% maize intraspecies non-colinearity/11 MYA = 1.4% degradation per MYA; Lai et al., 2004).

cis-regulatory sequence divergence

Through the comparative sequence analysis of promoter sequences with legume homologs, which, to some extent, resemble a pre-duplicate state, we analyzed the *cis*-regulatory evolution in *Arabidopsis* gene duplicates. We applied a conservative approach for the identification of phylogenetic footprints and observed, on average, 12% conservation between the promoter sequences of *Arabidopsis* and legume homologs. This percentage is in good agreement with the observed 15.2% noncoding sequence conservation of maize and rice sequences, which diverged 50-70 MYA (Guo and Moose, 2003). However, it should be noted that, given the rudimentary knowledge of plant *cis*-acting binding sites and the difficulties associated with distinguishing small functional elements from noise, only a subset of all *cis*-regulatory elements is currently covered (Guo and Moose, 2003; Inada et al., 2003; Lockton and Gaut, 2005).

For half of all gene duplicates created during the youngest genome duplication in *Arabidopsis*, a highly complementary set of motifs ($d_R > 0.8$) is conserved compared to that of the homologous legume promoter. Moreover, a decreasing number of motifs shared between *Ath* duplicates seems to correlate with an increasing reciprocal promoter divergence (rho=0.53, P<0.001). This relationship suggests that loss of *cis*-regulatory elements in duplicated genes is not completely random, but occurs according to a well-defined pattern, with a complementary set of motifs partitioned over both paralogs as a result. By analyzing a data set containing both recent and older duplicates (N=133; K_s values between 0.01 and 1.74), we found no significant correlation between the age of a duplicated gene pair and the number of motifs shared between the promoters of the *Ath'* duplicates and the legume promoter. In addition, we found no evidence that older gene duplicates experienced a higher degree of reciprocal promoter divergence (Figure 2.4.4b). This observation is in agreement with studies in other organisms, where it was shown that promoter divergence can occur very rapidly after duplication, although other duplicates can maintain a high degree of co-expression and promoter similarity for a long period of time (Figure 2.4.4b; Pickett and Meeks-Wagner, 1995; Makova and Li, 2003; Papp et al., 2003; Haberer et al., 2004). Recently, Haberer and co-workers (2004) suggested a time-dependent increase of expression divergence for *Arabidopsis* duplicates, based on a rather weak correlation observed between promoter similarities and expression correlations for tandem duplicates. However, the lack of a significant correlation between the age of a duplicate and the expression similarities identified for a genome-wide data set of *Arabidopsis* duplicates (Haberer et al., 2004), together with our observations, suggest that such a time-dependent divergence mechanism, if existing, is most probably not acting on all *Arabidopsis* gene duplicates. If this were true, then the significant negative correlations between divergence time K_s and expression similarity observed in yeast and human (Gu et al., 2002; Makova and Li, 2003) might represent an oversimplified picture of the actual evolution acting on the full set of gene duplicates.

In agreement with studies on cereal genes, we observed that the number of *cis*-regulatory elements is rather small for genes with basic enzymatic functions, whereas genes involved in signal transduction pathways, such as transcription factors or genes involved in RNA binding contain a larger number of regulatory motifs (Inada et al., 2003; Harbison et al., 2004). Consequently, the *cis*-regulatory complexity embedded in genes governing different molecular or biochemical functions will probably also determine the degree of promoter divergence tolerated in duplicated genes (Gu et al., 2002).

Does reciprocal promoter divergence provides evidence for subfunctionalization?

Based on our promoter analysis of *Ath* duplicates derived from the youngest genome duplication and *Ath'* duplicates with varying age, it is clear that increased promoter divergence in gene duplicates yields highly reciprocal patterns of *cis*-regulatory conservation. Consequently, one expects that this difference in the content of *cis*-regulatory elements between duplicated genes will be responsible for dissimilar spatio-temporal expression profiles. A low expression correlation (≤0.25) between both expression profiles was indeed observed for more than 85% of all analyzed paralogs having a high degree of reciprocal promoter divergence (d_R>0.9), indicating that both genes are no longer significantly co-regulated. Therefore, we believe that the reciprocal promoter divergence identified through the comparison of a homologous legume promoter reveals subfunctionalization after gene duplication. Our results are in good agreement with those of Blanc and Wolfe (2004), who estimated, based on expression data that 57% of all gene duplicates from the youngest polyploidy event in *Arabidopsis* have functionally diverged. Based on a very small data set of 36 gene duplicates, no strong linear relationship between the degree of reciprocal promoter divergence and expression correlation was observed (rho=-0.23, P=0.18). Zhang and co-workers suggested that the lack of correlation between *cis*-regulatory motif content and expression similarity might be caused by the absence or presence of additional *trans*-acting factors, which, in turn, could contribute substantially to the final expression pattern (Zhang et al., 2004). This hypothesis seems valid, because, in addition to *cis*-regulatory, also *trans*-regulatory changes contribute substantially to divergent expression patterns (Romano and Wray, 2003; Wittkopp et al., 2004). Nevertheless, analysis on a much larger set of gene duplicates, for which both expression data and outgroup promoter sequences are available, is required to fully understand the role of *cis*- and *trans*-regulatory changes in the evolution of gene duplicates.

For the set of 13 gene duplicates with reciprocal promoter divergence and highly dissimilar expression profiles, we studied the nature of the divergence of gene expression in more detail. For most (70%) of the analyzed gene pairs, the expression breadth of both duplicates (measured as the number of experiments in which the gene was significantly expressed over the total number of experiments), was very similar. This result suggests that these genes acquired a clearly different expression profile after duplication, but maintained a similar level of overall activity. For the remaining *Ath* duplicates with reciprocal promoter divergence and very low expression correlations, the expression breadth of one duplicate was significantly reduced compared to that measured for the other paralog. This pattern has also been observed for human duplicates, in which gene pairs that rapidly diverged in their expression either altered their expression pattern in terms of presence or absence in different tissues or in terms of absolute amounts of mRNA transcripts (Makova and Li, 2003). Although different, both patterns seem to be in agreement with the subfunctionalization model describing the complementary loss of regulatory motifs (Force et al., 1999). Because it is very unlikely that each *cis*-regulatory element has the same quantitative contribution to the overall gene expression pattern (Wray et al., 2003; Harbison et al., 2004), a complementary pattern of motif loss might lead indeed to a severely reduced expression level of one of the duplicates.

Finally, although the reciprocal divergence patterns of promoter sequences and expression profiles observed here seem to be in agreement with the subfunctionalization model, it is difficult to conclude whether the degenerative complementary mutations themselves are responsible for the preservation of gene duplicates, as predicted in the DDC model (Force et al., 1999). If the DDC model is responsible for the preservation of all gene duplicates through this mechanism, we should clearly find evidence of reciprocal promoter divergence for all *Ath* duplicates. Despite the difficulty to investigate the initial evolutionary stages in gene duplicates, the presence of a number of ancient gene duplicates in eukaryotic genomes for which no divergence could be observed (Table 2.4.1; Makova and Li, 2003; Papp et al., 2003; Haberer et al., 2004) indicates that also other mechanisms determine which gene duplicates are retained (Pickett and Meeks-Wagner, 1995; Wendel 2000). However, given the limited knowledge on promoter architecture in plants on the one hand, and the conservative parameters applied here, on the other hand, these paralogs might be preserved because of complementary subfunction partitioning, either at the regulatory or protein levels. In addition, the limited amount of functional data and the absence of compatible interspecies expression data (e.g. Huminiecki and Wolfe, 2004) for plants makes it very difficult to characterize the modular structure of promoters in great detail or to determine the precise function of individual *cis*-acting regulatory elements.

Materials and methods

Genomic data and gene prediction

The annotation of the *Arabidopsis thaliana* genome was retrieved from The Institute for Genome Research (TIGR; release 5; Wortman et al., 2003). We extracted the coding sequences, the corresponding amino acid sequences, and the relative position and strand orientation for a total of 26,192 protein-encoding genes. From the European Molecular Biology Laboratory database (Kulikova et al., 2004), 261 genomic BAC sequences of the *Fabaceae* family were obtained showing sequence similarity with multiple nonhomologous *Arabidopsis* protein loci. The goal of this selection criterion was to discard genomic sequences encoding only a single gene or a cluster of tandemly duplicated genes. The gene predictor EuGène was used to define gene models with the intrinsic *Arabidopsis* parameters (i.e. determined through training on validated *Arabidopsis* gene models) and taking into account sequence similarity to publicly available plant ESTs and proteins (Schiex et al., 2000). For the 126 *Medicago truncatula*, 115 *Lotus corniculatus* var. japonicus*,* and 20 *Glycine max* BACs, 2,912, 2,413, and 354 genes were predicted, respectively. The complete annotation of all legume genomic BAC clones can be found at http://bioinformatics.psb.ugent.be/.

Identification of homologous genomic segments

All protein sequences of *Arabidopsis* and the *Fabaceae* BACs were compared with each other using BLASTP (Altschul et al., 1997) and significant homologous gene pairs (Li et al., 2001) were retained. With this method two proteins are considered homologous only when they share a substantially conserved region on both molecules with a minimum amount of sequence identity. In this manner, homology based on the partial overlap of single protein domains between two multidomain proteins, which occasionally leads to significant E-values in BLAST, is not retained. Colinear genomic segments between *Arabidopsis* and *Medicago*, lotus*,* and soybean were identified using i-ADHoRe (with gap_size 20, cluster_gap 20, q_value 0.7, anchor_points 5 and prob_cutoff 0.01; Simillion et al., 2004). For all retained gene duplicates in valid level 3 multiplicons (sets of three homologous segments), containing two duplicated *Arabidopsis* segments and one homologous *Fabaceae* segment, phylogenetic trees were constructed. Both paralogs and their *Fabaceae* homolog plus an outgroup sequence were used for phylogenetic inference. Outgroup sequences from *Pinus*, *Physcomitrella* and/or *Chlamydomonas* were retrieved from the Sequence Platform for the Phylogenetic analysis of Plant Genes (SPPG) database (Vandepoele and Van de Peer, 2005) and aligned with T-coffee (Notredame et al., 2000). Neighbor-Joining phylogenetic trees were constructed with PHYLIP (Felsenstein, 1993) using the Dayhoff PAM matrix and 100 bootstrap samples. Only tree topologies with an overall 70% bootstrap support were considered as significant.

Regulatory sequence analysis

Promoter sequences 1000 bp upstream from the translation start site were isolated for all genes of *Arabidopsis*, *Medicago,* lotus, and soybean. For some genes, shorter sequences had to be extracted because the upstream gene was located less than 1000 bp upstream. Subsequently, interspersed and simple sequence repeats were masked by RepeatMasker (Smit, AFA & Green, P; http://www.repeatmasker.org/). Conserved motifs in promoter sequences were identified using the shared motif method (SMM), described by Castillo-Davis et al. (2004). Because few general characteristics of *cis*-regulatory elements in plants are known and to reduce the false-positive rate associated with the detection of shared motifs, we empirically inferred the minimum alignment score by analyzing the distribution of d_{SM} values for random sequence pairs with a nucleotide composition similar to that of the data examined. Analysis of 1000 random sequence pairs of 1000 bp using a shared motif minimum alignment score of 56 (i.e. a combination of ≥14 matches, mismatches, and gaps that sum to a final score ≥56; Castillo-Davis et al., 2004) showed that >95% of all sequence pairs exhibited a d_{SM} > 0.90. Although we are aware that smaller *cis*-regulatory elements, i.e., fewer than 14 nucleotides, do exist in plants, the noise generated by such small elements is extremely high (see also Castillo-Davis et al., 2004). This is illustrated by d_{SM} values between 0 and 0.1 for random promoters when tolerating small shared motifs with only 5 or 6 nucleotide matches, (Supplemental Figure 3), suggesting that the *cis*-regulatory content between both promoters is (nearly) identical.

Besides quantifying the promoter divergence through d_{SM}, we defined a measure (d_R) to describe the degree of reciprocal motif divergence in promoters of duplicated genes (Figure 2.4.2). First, for each promoter sequence of a paralogous *Arabidopsis* gene pair, the number of shared motifs with a homologous *Fabaceae* promoter is counted. Subsequently, for each paralog, the number of shared motifs is compared with the total number of shared motifs found between both paralogs and the *Fabaceae* promoter sequence. When in both pairwise *Arabidopsis-Fabaceae* promoter comparisons all shared motifs are the same, then d_R will be zero, whereas, when no similar motifs are found between the *Fabaceae* promoter and both paralogous promoter sequences, then d_R=1. The reciprocal sequence divergence was not calculated based on the number of motifs, but with the exact positions and sizes of the detected motifs shared between the legume promoter and the *Arabidopsis* duplicated genes (Figure 2.4.2).

GO functional annotation

GO associations for *Arabidopsis* proteins were retrieved from TIGR (ftp.tigr.org/pub/data/a_thaliana/ath1/DATA_RELEASE_SUPPLEMENT/) and remapped to the generic GO Slim classification scheme (ftp.geneontology.org/pub/go/GO_slims/goslim_generic.go) with the Perl script map2slim.pl (available at http://www.geneontology.org).

Dating paralogous gene pairs

Pairwise alignments of the paralogous nucleotide sequences were created with CLUSTAL W (Thompson et al., 1994) using the corresponding protein sequences as an alignment guide. Alignment columns were removed when they had gaps 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 eliminated until the residues were conserved in all columns of the sequence alignment: for every pair of residues in the column, the BLOSUM62 value was retrieved and the median value for all these values was calculated. If the median was ≥0, the column was considered as containing homologous amino acids. K_s estimates were obtained with the CODEML (Goldman and Yang, 1994) program of the PAML package (Yang, 1997). Codon frequencies were calculated from the average nucleotide frequencies at the three codon positions (F3x4), whereas the gamma shape parameter α and the transition-transversion ratio κ were estimated for every pairwise comparison.

Gene family delineation

An all-against-all sequence comparison of all *Arabidopsis* and legume proteins (as generated by gene prediction; see above) was performed with BLASTP (Altschul et al., 1997) and relevant hits were retained (Li et al., 2001). All valid homologous protein pairs (e.g. protein A is homologous to protein B, protein B is homologous to protein C) were subject to a simple-linkage clustering routine to delineate protein gene families (for example, family with proteins A, B, and C).

- Part 3 -

The evolution of plant gene families

3.1 Genome-wide analysis of core cell cycle genes in *Arabidopsis*

Klaas Vandepoele, Jeroen Raes, Lieven De Veylder, Pierre Rouzé, Stephane Rombauts and Dirk Inzé

Redrafted from: *Plant Cell* **14**, 903-916 (2002)

Cyclin-dependent kinases and cyclins master together with the help of different interacting proteins the progression through the eukaryotic cell cycle. A high-quality, homology-based annotation protocol was applied to determine all core cell cycle genes in the recently completed *Arabidopsis* genome sequence. In total, 61 genes were identified belonging to seven selected families of cell cycle regulators, for which 30 are new or corrections of the existing annotation. A new class of putative cell cycle regulators was found that probably are competitors of E2F/DP transcription factors, which mediate the G1-to-S progression. In addition, the existing nomenclature for cell cycle genes of *Arabidopsis* was updated and physical positions of all genes were compared with segmentally duplicated blocks in the genome, showing that 22 core cell cycle genes emerged through block duplications. This genome-wide analysis illustrates the complexity of the plant cell cycle machinery and provides a tool for elucidating the function of new family members in the future.

Introduction

Cell proliferation is controlled by a universally conserved molecular machinery, in which the core key players are serine/threonine kinases, known as cyclin-dependent kinases (CDKs). CDK activity is regulated in a complex manner, including phosphorylation/dephosphorylation by specific kinases/phosphatases and the association with regulatory proteins. Although many cell cycle genes of plants have been identified in the last decade (for review, see Stals and Inzé, 2001), the correct number of CDKs, cyclins, and interacting proteins with a role in the cell cycle control is still unknown. Now that the complete sequence of the nuclear genome of *Arabidopsis* is available (Arabidopsis Genome Initiative, 2000), it is possible to scan an entire plant genome for all these core cell cycle genes and determine their number, position on the chromosomes, and phylogenetic relationship. From an evolutionary point of view, this core cell cycle gene catalogue would be extremely interesting because it allows us to determine which processes are plant specific and which are conserved among all eukaryotes. Furthermore, there is a unique opportunity to unravel in future experiments the function and interactions of newly found family members of primary cell cycle regulators, thus expanding our knowledge on how cell cycle is regulated in plants.

Nevertheless, a genome-wide inventory of all core cell cycle genes is only possible when the available raw sequence data are correctly annotated. Although the genome-wide annotation of organisms sequenced by large consortia produced a huge amount of information, which, no doubt, benefits the scientific community, one has to realize that this automated high-throughput annotation is far from optimal (Devos and Valencia, 2001). For this reason, it is often not trivial to extract clear biological information out of these public databases. When high-quality annotation is needed, a supervised semi-automatic annotation may be a good compromise between quality and speed. Annotation is generally performed in two steps: first, a structural annotation that aims at finding and characterizing biologically relevant elements within the raw sequence (such as exons and translation starts), and secondly, functional annotation, in which biological information is attributed to the gene or its elements. Unfortunately, there are some problems inherent to both. When structural annotation is performed, the first problem occurs whenever no cDNA or expressed sequence tag (EST) information is available, which is the case for 60% of all *Arabidopsis* genes (Arabidopsis Genome Initiative, 2000). Then, one has to resort to intrinsic gene prediction software, which remains limited, although a lot of improvement has been made over the last few years. Errors range from wrongly determined splice sites or start codons, over so-called spliced (one gene predicted as two) or fused (two genes predicted as one) genes, up to completely missed or nonexisting predicted genes (Rouzé et al., 1999). In addition, no general and well-defined prediction protocol is used by the different annotation centers with the generation of redundant, non-uniform, structural annotation as a result. Furthermore, clear information is lacking on methods and programs used as well as the motivation for applying a special protocol, making it impossible to trace the annotation grounds.

The problem with functional annotation is related to the difficulty to couple biological knowledge to a gene. Such a link is made generally on the basis of sequence similarity that is derived either from full-length sequence comparisons or by means of multiple alignments, patterns, and domain searches. Of major concern is the origin of the assigned function, because transfer of low-quality or bad functional annotation propagates wrong annotations in the public databases. Even correct annotations can be erroneously disseminated: one can easily imagine the wrong transfer of a good functional assignment from a multidomain protein to a protein that only has one of the domains. This problem can be avoided by using only experimentally derived information to predict unambiguously a gene's structure and function.

Here, we applied a homology-based annotation by using experimental references to build a full catalogue with 61 core cell cycle genes of *Arabidopsis*. In total, 30 genes are new or are genes for which the previous annotation was incorrect. Based on phylogenetic analysis we updated and rationalized their nomenclature. Furthermore, relations between gene family members were correlated with large segmental duplications.

Materials and methods

Annotation of Arabidopsis cell cycle genes

The genome version of January 18, 2001 (v180101) was downloaded from the ftp site (ftp://ftpmips.gsf.de/cress/) of the Martiensried Institute for Protein Sequences (MIPS) center (Martiensried, Germany). Regions of interest on the chromosomes were localized by the BLAST software (Altschul et al., 1997) with experimental representatives as query sequence. For the regions returned by

BLAST, chromosome sequences were extracted with 15 kb upstream and downstream from the hit to prevent unreliable prediction due to border effects.

Gene prediction was done with Eugene (Schiex et al., 2001), in combination with GeneMark.hmm (Lukashin and Borodovsky, 1998), because the latter had been reported previously to give the best scores in *Arabidopsis* (Pavy et al., 1999). New analysis (C. Mathé, personal communication), however, showed that Eugene has become the best gene prediction tool for *Arabidopsis*. The Eugene program combines NetGene2 (Tolstrup et al., 1997) and SplicePredictor (Brendel and Kleffe, 1998) for splice site prediction, NetStart (Pedersen and Nielsen, 1997) for translation initiation prediction, Interpolated Markov model-based content sensors, and information from protein, EST, and cDNA matches to predict the final gene model.

The predicted candidate gene products were aligned with the experimental representatives by using CLUSTAL W (Thompson et al., 1994). On the final alignments, HMMer was used to generate profiles for each specific gene family with hidden Markov models. These profiles were then used to search for new family members (Eddy, 1998). The genome-wide non-redundant collection of *Arabidopsis* protein-encoding genes was predicted with GeneMark.hmm. Based on these predictions, we built a database of virtual transcripts (and corresponding protein database) that we designated genome-predicted transcripts (GPTs). Manual annotation was done with Artemis (Rutherford et al., 2000).

Phylogeny and nomenclature

Phylogenetic analysis was performed on more conserved positions of the alignment. Editing of the alignment and reformatting was done with BioEdit (Hall, 1999) and ForCon (Raes and Van de Peer, 1999). Similarity between proteins was based on a BLOSUM62 matrix (Henikoff and Henikoff, 1993). Trees were constructed with various distance and parsimony methods. Distance matrices were calculated based on Poisson, Kimura, or PAM correction and trees were constructed with the Neighbor-joining algorithm by means of the software packages TREECON (Van de Peer and De Wachter, 1994) and PHYLIP (Felsenstein, 1993). The latter was also used for the parsimony analysis. Bootstrap analysis with 500 replicates was performed to test the significance of nodes.

Protein structure analysis

Protein secondary structure prediction was done with PSIpred v2.0 (Jones, 1999).

Segmental duplications in the Arabidopsis genome

For the detection of large segmental duplications, duplicated blocks were identified by a method similar to that by Vision et al. (2000). Initially, protein-coded genes predicted by GeneMark.hmm (in total 26,352 present in our GPT database) were ordered according to the location on the corresponding chromosome. BLASTP was used to identify genes with high sequence similarity and all BLASTP scores were stored in a matrix to be analyzed. Initially, filtering was performed to reduce low-similarity hits (E-value < 1e-50; Friedman and Hughes, 2001), followed by a procedure to define duplicated blocks in the scoring matrix. Finally, by post-processing only blocks of appropriate size (i.e. blocks containing more than seven genes) were selected.

Results

Strategy

In order to correctly annotate all core cell cycle genes, a strategy was defined that uses as much reliable information as possible, combining experimentally derived data with the best prediction tools available for *Arabidopsis* (see Materials and methods). First, experimental representatives for each family were used as bait to locate regions of interest on the different chromosomes. For these selected regions, genes were predicted and candidate genes were validated; the presence of mandatory domains in their gene products was determined by aligning them with the experimental representatives and, if necessary, the predicted gene structure was modified by using the family-related characteristics or ESTs. Still, in some cases, this approach did not allow us to conclude whether a region of interest really coded for a potential gene or whether a candidate gene was a core cell cycle gene. To clarify such situations, a more integrated analysis was performed. First, the members of every family were used to build a profile for that specific family.

By taking into account the new predicted genes for creating the profile, a more “flexible” (i.e. all diversity within a class/subclass being represented) and plant-specific profile could be established. With this new profile, novel family members were sought within a collection of genome-wide predicted *Arabidopsis* proteins. Subsequently, the predicted gene products were again validated or modified by comparing them with those of other family members in a multiple alignment. With this additional approach, we could determine clearly whether the predicted genes were similar to a certain class of cell cycle genes.

To characterize subclasses within the gene families, phylogenetic trees were generated that included reference cell cycle genes from other plants and known genes from *Arabidopsis*; by different methods and statistical analysis of nodes the significance of the derived classification was tested. Based on the position in the tree and the presence of class-specific signatures, genes were named according to the proposed nomenclature rules for cell cycle genes (Renaudin et al., 1996; Joubès et al., 2000). A complete list of core cell cycle genes in *Arabidopsis* in presented in Table 3.1.1. Additional data regarding nomenclature and gene models can be found at http://bioinformatics.psb.ugent.be/.

Annotation and nomenclature

CDK

In yeasts one CDK is sufficient to drive cells through all cell cycle phases, whereas multicellular organisms evolved to use a family of related CDKs, all with specific functions. In plants, two major classes of CDKs have been studied so far, known as A-type and B-type CDKs. The A-type CDKs regulate both the G1-to-S and G2-to-M transitions and the B-type CDKs seem to control the G2-to-M checkpoint only (Hemerly et al., 1995; Magyar et al., 1997; Porceddu et al., 2001). In addition, the presence of C-type CDKs and CDK-activating kinases (CAKs) have been reported (Magyar et al., 1997; Umeda et al., 1998; Joubès et al., 2001). Whereas the latter were shown to regulate the activity of the A-type CDKs, the function of the C-type CDKs remains unknown. Until now, one A-type and four B-type CDKs have been described for *Arabidopsis* (Joubès et al., 2000; Boudolf et al., 2001). Furthermore, C-type CDKs and one CAK have been reported as well (Umeda et al., 1998; Lessard et al., 1999). In alfalfa, one E-type CDK has been identifed, but no counterparts had been found previously in *Arabidopsis* (Magyar

Table 3.1.1 Characteristics of all 61 core cell cycle genes in *Arabidopsis*

Gene	Chr.	Start[a]	Stop[b]	Strand	Status[c]	Features[d]	ORF Name
Arath;*CDKA;1*	3	18,368,303	18,370,279	+	EXP	PSTAIRE	AT3g48750
Arath;*CDKB1;1*	3	20,355,861	20,357,226	+	EXP	PPTALRE	AT3g54180
Arath;*CDKB1;2*	2	16,301,446	16,302,758	+	EXP	PPTALRE	AT2g38620
Arath;*CDKB2;1*	1	28,430,923	28,429,129	-	EXP	PSTTLRE	AT1g76540
Arath;*CDKB2;2*	1	7,294,679	7,292,770	-	EXP	PPTTLRE	AT1g20930
Arath;*CDKC;1*	5	3,224,679	3,221,723	-	AI993037	PITAIRE	AT5g10270
Arath;*CDKC;2*	5	25,955,460	25,958,387	+	AV439592	PITAIRE	AT5g64960
Arath;*CDKD;1*	1	27,423,792	27,425,694	+	PRED	NVTALRE	AT1g73690
Arath;*CDKD;2*	1	24,603,461	24,605,698	+	AV554642	NFTALRE	AT1g66750
Arath;*CDKD;3*	1	6,206,888	6,209,316	-	AF344314	NITALRE	AT1g18040
Arath;*CDKE;1*	5	25,465,021	25,463,612	-	BG459367	SPTAIRE	AT5g63610
Arath;*CDKF;1*	4	13,494,330	13,495,958	+	EXP	None	AT4g28980
Arath;*CYCA1;1*	1	16,354,762	16,352,618	-	AV556475	LVEVxEEY	AT1g44110
Arath;*CYCA1;2*	1	28,792,710	28,790,480	-	PRED	LVEVxEEY	AT1g77390
Arath;*CYCA2;1*	5	8,885,657	8,887,990	+	EXP	LVEVxEEY	AT5g25380
Arath;*CYCA2;2*	5	3,604,472	3,601,820	-	EXP	LVEVxDDY	AT5g11300
Arath;*CYCA2;3*	1	5,363,054	5,365,235	+	EXP[e]	LVEVxEEY	AT1g15570
Arath;*CYCA2;4*	1	29,923,266	29,925,430	+	AV558333	LVEVxEEY	AT1g80370
Arath;*CYCA3;1*	5	17,293,193	17,294,681	+	PRED	LVEVxEEY	AT5g43080
Arath;*CYCA3;2*	1	17,022,212	17,023,757	+	AT50514	LVEVxEEY	AT1g47210
Arath;*CYCA3;3*	1	17,024,852	17,026,370	+	PRED	LVEVxEEY	AT1g47220
Arath;*CYCA3;4*	1	17,027,927	17,029,762	+	PRED	LVEVxEEY	AT1g47230
Arath;*CYCB1;1*	4	16,830,051	16,827,976	-	EXP	HxRF	AT4g37490
Arath;*CYCB1;2*	5	1,861,577	1,859,551	-	EXP	HxKF	AT5g06150
Arath;*CYCB1;3*	3	3,627,150	3,625,489	-	EXP[f]	HxKF	AT3g11520
Arath;*CYCB1;4*	2	11,548,850	11,552,088	+	PRED	HxKF	AT2g26760
Arath;*CYCB2;1*	2	7,813,050	7,815,144	+	EXP	HxKF	AT2g17620
Arath;*CYCB2;2*	4	16,107,598	16,109,617	+	EXP	HxKF	AT4g35620
Arath;*CYCB2;3*	1	7,137,288	7,135,091	-	PRED	HxKF	AT1g20610
Arath;*CYCB2;4*	1	28,338,772	28,336,622	-	PRED	HxKF	AT1g76310
Arath;*CYCB3;1*	1	5,584,476	5,582,409	-	PRED	HxKF	AT1g16330
Arath;*CYCD1;1*	1	26,148,702	26,150,664	+	EXP	LxCxE	AT1g70210
Arath;*CYCD2;1*	2	9,704,757	9,703,043	-	EXP	LxCxE	AT2g22490
Arath;*CYCD3;1*	4	15,563,758	15,565,156	+	EXP	LxCxE	AT4g34160
Arath;*CYCD3;2*	5	26,836,277	26,837,626	+	AI995751	LxCxE	AT5g67260
Arath;*CYCD3;3*	3	18,862,632	18,861,289	-	AV527915	LxCxE	AT3g50070
Arath;*CYCD4;1*	5	26,143,713	26,141,558	-	EXP	LxCxE	AT5g65420
Arath;*CYCD4;2*	5	3,282,347	3,280,801	+	PRED	no LxCxE	AT5g10440
Arath;*CYCD5;1*	4	16,885,341	16,886,338	+	AI998509	LFLCxE	AT4g37630
Arath;*CYCD6;1*	4	1,432,497	1,431,184	-	PRED	no LxCxE	AT4g03270
Arath;*CYCD7;1*	5	417,084	418,547	+	PRED	LxCxE	AT5g02110
Arath;*CYCH;1*	5	9,813,161	9,816,075	+	AV560893	None	AT5g27620
Arath;*CKS1*	2	12,060,430	12,059,793	-	EXP	None	AT2g27960
Arath;*CKS2*	2	12,061,999	12,061,350	-	AV553882	None	AT2g27970
Arath;*DEL1*	3	18,079,607	18,081,809	+	EXP	None	AT3g48160
Arath;*DEL2*	5	4,858,640	4,861,044	+	PRED	None	AT5g14960
Arath;*DEL3*	3	126,812	124,606	-	EXP	None	AT3g01330
Arath;*DPa*	5	544,155	844,977	-	EXP	None	AT5g02470
Arath;*DPb*	5	842,841	845,196	+	EXP	None	AT5g03410
Arath;*E2Fa*	2	15,268,582	15,271,784	+	EXP	None	AT2g36010
Arath;*E2Fb*	5	7,431,826	7,434,541	+	EXP	None	AT5g22220
Arath;*E2Fc*	1	17,356,113	17,358,730	+	EXP	None	AT1g47870
Arath;*KRP1*	2	10,126,806	10,125,908	-	EXP	None	AT2g23430
Arath;*KRP2*	3	19,096,470	19,097,325	+	EXP	None	AT3g50630
Arath;*KRP3*	5	19,794,310	19,792,575	-	EXP	None	AT5g48820
Arath;*KRP4*	2	14,022,387	14,024,238	+	EXP	None	AT2g32710
Arath;*KRP5*	3	9,060,905	9,061,654	+	EXP	None	AT3g24810
Arath;*KRP6*	3	6,617,597	6,616,567	-	EXP	None	AT3g19150
Arath;*KRP7*	1	18,087,625	18,086,761	-	EXP	None	AT1g49620
Arath;*Rb*	3	3,919,344	3,913,685	-	AF245395	None	AT3g12280
Arath;*WEE1*	1	673,409	676,125	+	EXP[g]	None	AT1g02970

[a] Position of start codon on the chromosome.

[b] Position of stop codon on the chromosome.

[c] Expression status of the gene. EXP, experimentally characterized; PRED, prediction. Numbers are EST accession numbers.

[d] Family-specific protein signatures.

[e] EST BE528080 found for the first exon completes the structural annotation.

[f] Gene structure was determined using partial mRNA L27224 and AV546264.

[g] Gene structure was determined using two cDNA sequences, confirming the manual annotation.

et al., 1997). By the homology-based annotation method used here, we identified in total eight CDKs (one A-type, four B-type, two C-type, one E-type) and four CAKs (three D-type and one F-type).

The previously described CAK homolog of *Arabidopsis* (*cak1At*) differs substantially from the known rice CAK, R2 (Umeda et al., 1998; Yamaguchi et al., 1998). R2 has been suggested to be specific for monocots (Yamaguchi et al., 1998). However, with the rice sequence as experimental reference, three related sequences were identified in *Arabidopsis*, designated CDKD;1, CDKD;2 and CDKD;3 with 75%, 68% and 79% sequence similarity with R2 from rice, respectively. These genes are only distantly related to *cak1At*, indicating that *Arabidopsis* has two functional classes of CAK. To stress this functional difference and to have a more uniform nomenclature, *cak1At* was renamed as *CDKF;1*. The phylogenetic relationship among CDKs of *Arabidopsis* are shown in Figure 3.1.1.

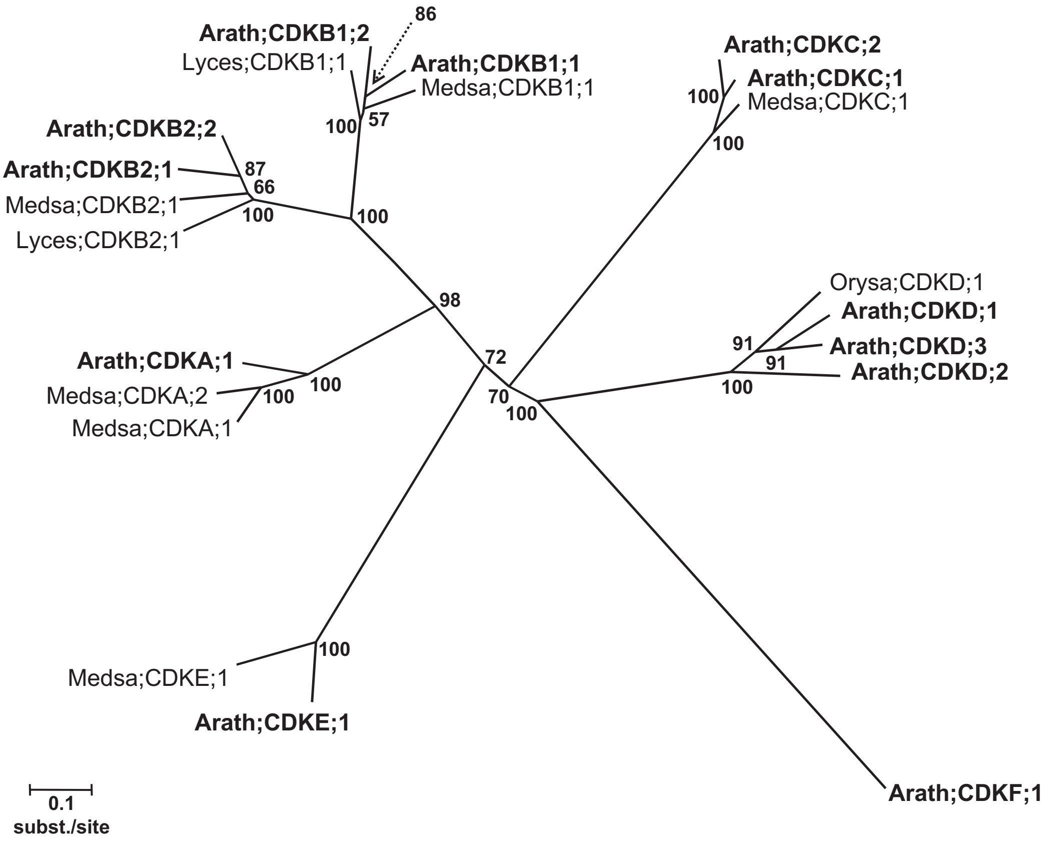


Figure 3.1.1 Unrooted Neighbor-Joining tree of the A, B, C, D, E, and F classes of CDKs with the Poisson correction for evolutionary distance calculation. Bootstrap values of 500 bootstrap iterations are shown. Numbers indicate evolutionary distance. Arath, *Arabidopsis*; Lyces, tomato (*Lycopersicon esculentum*); Medsa, alfalfa (*Medicago sativa*); Orysa, rice (*Oryza sativa*). Reference genes are Medsa;CDKC;1, Orysa;CDKD;1, Medsa;CDKE;1, Medsa;CDKA;1, Medsa;CDKA;2, Medsa;CDKB1;1, Lyces;CDKB1;1, Lyces;CDKB2;1, and Medsa;CDKB2;1.

Cyclins

Monomeric CDKs have no kinase activity and have to associate with regulatory proteins called cyclins to be activated. Because the cyclin protein levels fluctuate in the cell cycle, cyclins are the major factors that determine the timing of CDK activation. Cyclins can be grouped into mitotic cyclins (designated A- and B-type cyclins in higher eukaryotes and CLBs in budding yeast) and G1-specific cyclins (D-type cyclins in mammals and CLNs in budding yeast). H-type cyclins regulate the activity of the CAKs. All four types of cyclins known in plants were identified mostly by analogy to their human counterparts. For *Arabidopsis*, currently four A-type, five B-type, five D-type, but no H-type, cyclins have been described (Soni et al., 1995; Renaudin et al., 1996; De Veylder et al., 1999; Swaminathan et al., 2000). By using the known plant cyclin sequences as probes, a total of 30 cyclins could be detected in the *Arabidopsis* genome. For 19 cyclins, an EST could be found (Table 3.1.1).

Three different subclasses of plant A-type cyclins (A1, A2, and A3) have been described previously (Renaudin et al., 1996) and were all found in *Arabidopsis*, comprising 10 cyclins. Two members of A1-type members (*CYCA1;1* and *CYCA1;2*), four A2-type (*CYCA2;1*, *CYCA2;2*, *CYCA2;3*, and *CYCA2;4*), and four A3-type genes were detected (*CYCA3;1*, *CYCA3;2*, *CYCA3;3*, and *CYCA3;4*). B-type cyclins are subdivided into two subclasses, B1 and B2. In total, *Arabidopsis* contains nine B-type cyclins, of which four belong to the B1 class (*CYCB1;1*; *CYB1;2*, *CYCB1;3*, and *CYCB1;4*) and four to the B2 class (*CYCB2;1*, *CYCB2;1*, *CYCB2;3*, and *CYCB2;4*). One gene could be attributed neither the B1 nor the B2 classes, although it clearly contained a B-type-like cyclin box in combination with the B-type specific HxKF signature. On the other hand, no B1- nor B2-like destruction box could be detected. The phylogenetic position of this gene within the B cluster depended on the number of positions used for the analysis. Because cyclin sequences are known to be saturated with substitutions (Renaudin et al., 1996), a technique was applied to construct trees on unsaturated positions only (Van de Peer et al., 2002). No support was found to designate this gene to one of the two classes of B-type cyclins (data not shown). On this basis, it seems justified to create a new subclass of cyclins, the B3-type (Figure 3.1.2).

In addition to the five D-type cyclins already described (*CYCD1;1*, *CYCD2;1*, *CYCD3;1*, *CYCD3;2*, and *CYCD4;1*), five new D-type genes were detected. Based on their phylogenetic position, two were attributed to the D3 (*CYCD3;3* and *CYCD3;4*) and one to the D4 (*CYCD4;2*) classes. The remaining new D-type

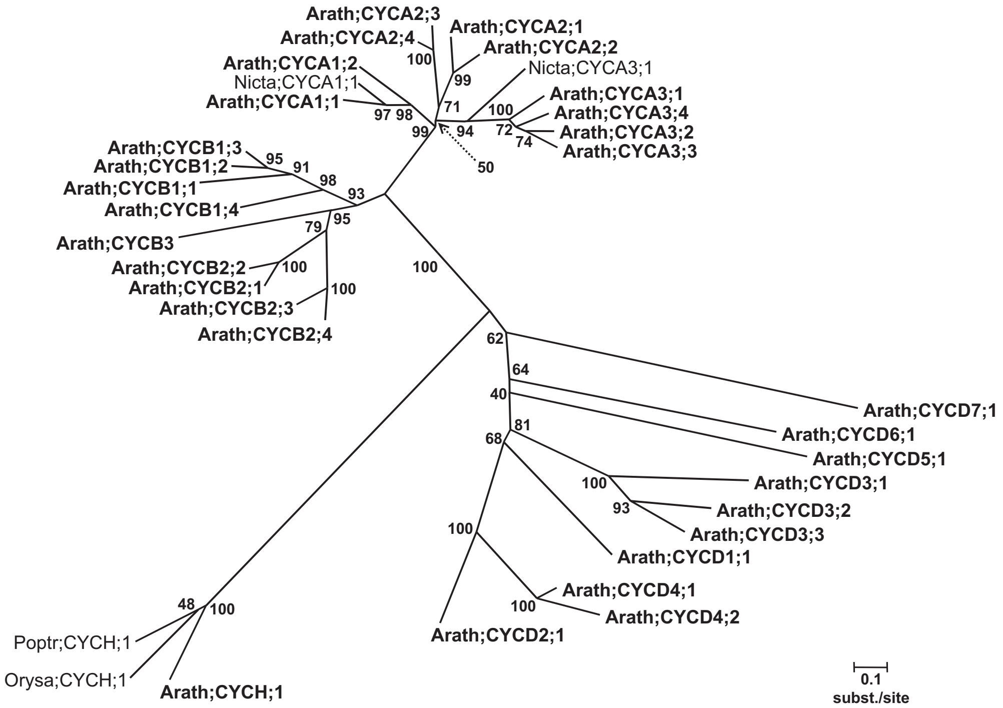


Figure 3.1.2 Unrooted Neighbor-Joining tree of the A, B, D, and H subgroups of the cyclin family with Poisson correction for evolutionary distance calculation. Bootstrap values of 500 bootstrap iterations are shown. Scales indicate evolutionary distance. Arath, *Arabidopsis*; Nicta, tobacco (*Nicotiana tabacum*); Orysa, rice; Poptr, poplar (*Populus tremula – Populus tremuloides*). Reference genes are Nicta;CYCA1;1, Nicta;CYCA3;1, Poptr;CYCH, and Orysa;CYCH.

cyclins were further subdivided into classes CYCD5, CYCD6, and CYCD7 according to their phylogenetic positions. It is remarkable that *CYCD4;2* and *CYCD6;1* do not contain the LxCxE retinoblastoma (Rb)-binding motif, whereas *CYCD5;1* contains a divergent Rb-binding motif (FxCxE), located at the N-terminus. The biological function of cyclins lacking the conserved Rb-binding motif remains unclear. One *Arabidopsis* gene was found with high sequence similarity to cyclin H of poplar (71%) and rice (66%).

Aligning all cyclins allowed us to identify the cyclin and destruction box consensus sequences for A-, B-, D-, and H-type cyclins (Table 3.2.2). Although A- and B-type cyclin boxes are very similar, these two types of cyclins can be discriminated by their destruction boxes. For two genes within the A- and B-type cyclins (*CYCA3;1* and *CYCB3;1*), no destruction box could be detected. In addition, these genes have a highly diverged cyclin box compared with their subclass consensus. The low overall sequence similarity within D-type cyclins is also reflected in their cyclin box. In addition to the cyclins described above, two presumed pseudogenes were predicted, which were very similar to B-type cyclins.

Table 3.1.2 Consensus sequences for cyclin and the destruction Box in *Arabidopsis* cyclins

Subclass	Cyclin Box Signature	Destruction Box
Cyclin A1	MR-(I/V)L(I/V)DW	RAPL(G/S)(D/N)ITN
Cyclin A2	MR-(I/V)L(I/V)DW	RAVL(K/G)(D/E)(I/V)(T/S)N
Cyclin A3[a]	MR-(I/V)L(I/V)DW	RVVLGEL(P/L)N
Cyclin B1	MR-IL(I/V/F)DW	R-(A/V)LGDIGN
Cyclin B2	MR-IL(I/V/F)DW	RR(A/V)L–IN
Cyclin B3	TRGILINW	N.D.[b]
Cyclin D1	REDSVAW	N.D.
Cyclin D2	RNQALDW	N.D.
Cyclin D3	R(E/K)(E/K)A(L/V)(D/G)W	N.D.
Cyclin D4	R(R/I)(D/Q)AL(N/G)W	N.D.
Cyclin D5	RLIAIDW	N.D.
Cyclin D6	RNQAISS	N.D.
Cyclin D7	RFHAFQW	N.D.
Cyclin H[c]	MRAFYEAK	N.D.

[a] In CYCA3;1, cyclin box KRGVLVDW was not included in the consensus; no destruction box was detected.

[b] N.D., not detected.

[c] Plant cyclin H consensus for cyclin box MR(A/V)(F/Y)YE-K (based on the sequence of Arath;*CYCH*, Orysa;*CYCH*, and cyclin H of poplar).

The precise number of pseudogenes for the seven selected families remains unclear, because the detection of pseudogenes depends on the degree of conservation still present in their gene structure and of detection by prediction tools of these degenerated structures.

CDK/cyclin interactors and regulatory proteins

CKS proteins act as docking factors that mediate the interaction of CDKs with putative substrates and regulatory proteins. Besides the already described CDK subunit gene in *Arabidopsis* (*Arath;CKS1*; De Veylder et al., 1997), a second CKS gene was found (*Arath;CKS2*) with sequence (83% identical and 90% similar amino acids) and gene structure (number and size of exons and introns) very similar to those of *Arath;CKS1* (Figure 3.1.3a). The two CKS gene products miss both the N- and C-terminal extension when compared with the yeast Suc1p/Cks1p homologs (De Veylder et al., 1997). Upon the occurrence of stress or the perception of antiproliferation agents, the CDK/cyclin complexes are repressed by the CDK inhibitor (CKI) proteins. In mammals, two different classes of CKIs exist (the INK4 and the Kip/Cip families), each with their own CDK-binding specificity and protein structure. Seven *CKI* genes, belonging to the group of Kip/Cip CKIs, have been

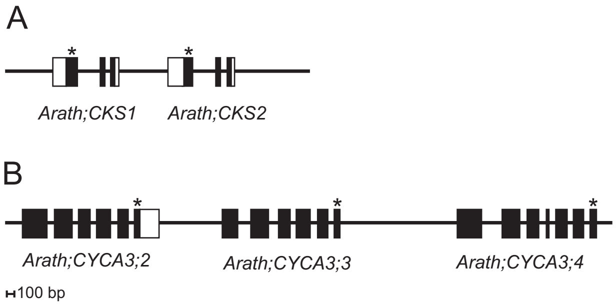


Figure 3.1.3 Gene tandem duplication of CKS and A3-type cyclin genes. Black rectangles represent protein-encoding exons, and white rectangles represent untranslated regions based on hits with ESTs or mRNA. Asterisks denote the exon with the stop codon. (a) Gene structure of *CKS1* and *CKS2* on chromosome 2. The indicated chromosome region spans from 12,059 to 12,063 kb. (b) Gene structure of *CYCA3;2*, *CYCA3;3*, and *CYCA3;4* on chromosome 1. The indicated region spans from 17,022 to 17,030 kb. ESTs AT50714, AT50514, and AT37419 hit with *CYCA3;2* (data not shown).

described previously for *Arabidopsis*, designated *KRP1* to *KRP7* (De Veylder et al., 2001). No extra KRPs could be detected in the complete genome and no plant counterparts of the INK4 family were found as well.

CDK/cyclin activity is negatively regulated by phosphorylation of the CDK subunit by the WEE1 kinase and positively when the inhibitory phosphate groups are removed by the CDC25 phosphatase. A single *WEE1* gene was identified on chromosome 1. The WEE1 kinase was annotated by using two cDNA sequences that were at our disposal (L. De Veylder, unpublished results) and has the highest homology to the WEE1 kinase of maize, showing 56% similarity with the gene product of a partial mRNA (Sun et al., 1999). No CDC25 phosphatase could be identified.

Rb and E2F/DP

Rb and the E2F/DP proteins are key regulators that control the entry of DNA replication. When the E2F/DP transcription factors are bound to Rb, they are inactive, but they become active when Rb is phosphorylated by G1-specific CDK/cyclin complexes, stimulating transcription of genes needed for G1-to-S and S phase progression. Only one Rb could be identified in the *Arabidopsis* genome that was located on chromosome 3. *E2F* genes are known for tobacco, carrot, and wheat (Ramírez-Parra et al., 1999; Sekine et al., 1999; Albani et al., 2000; Magyar et al., 2000), but no *Arabidopsis* family members have been described until now, whereas two *Arabidopsis* DP genes (*DPa* and *DPb*) have been reported.

The *E2F* and *DP* genes were analyzed in a combined approach, because the sequence of both types of proteins are partially similar (22% overall similarity). In total, eight genes were detected in *Arabidopsis*. Although the sequence similarity between these eight members of the *E2F/DP* family is rather low (20% overall mean similarity), three groups had emerged based on prior experimental information (Magyar et al., 2000) and phylogenetic analysis (Figure 3.1.4). The first group comprises the E2F transcription factors that are most similar to the mammalian *E2F* factors and were designated *E2Fa*, E2Fb, and E2Fc (46% overall similarity). The second group consists of the two already known DP factors.

The third group contains three new genes with an internal similarity of 59% and a sequence similarity with both *E2F* (21%) and *DP* genes (18%), initially indicating some kind of relation with the *E2F/DP* genes. When the boxes present in the E2F genes (DNA-binding, dimerization, Marked and Rb-binding box) and DP genes (DNA-binding and dimerization box) were compared with these three new genes, only a DNA-binding domain was found, but in duplex (Figure 3.1.5a). Both DNA-binding domains are highly similar to the E2F DNA-binding domain. Because of their phylogenetic position, they form a distinct class, which we designated as DP-E2F-like (DEL). The DNA-binding domain of the E2F and DP genes have a limited across-family homology (Figure 3.1.5b), including the RRxYD DNA recognition motif (in their α3 helices), which interacts with half of the palindromic promoter-binding site (CGCGCG and CGCGCG). Within all three

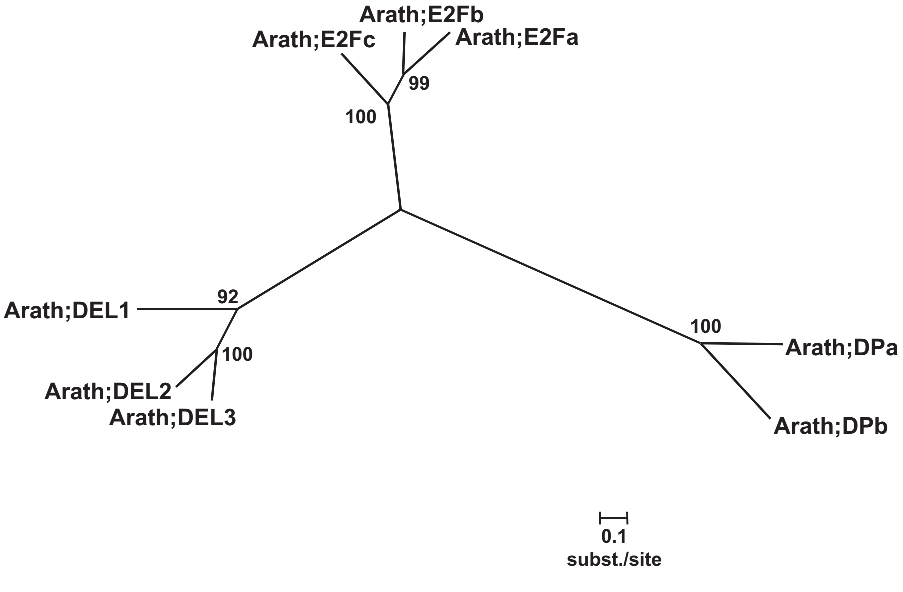


Figure 3.1.4 Unrooted Neighbor-Joining tree of the E2F, DP, and DEL Families with Poisson Correction for evolutionary distance calculation. Bootstrap values of 500 bootstrap iterations are shown. Scales indicate evolutionary distance. Arath, *Arabidopsis*.

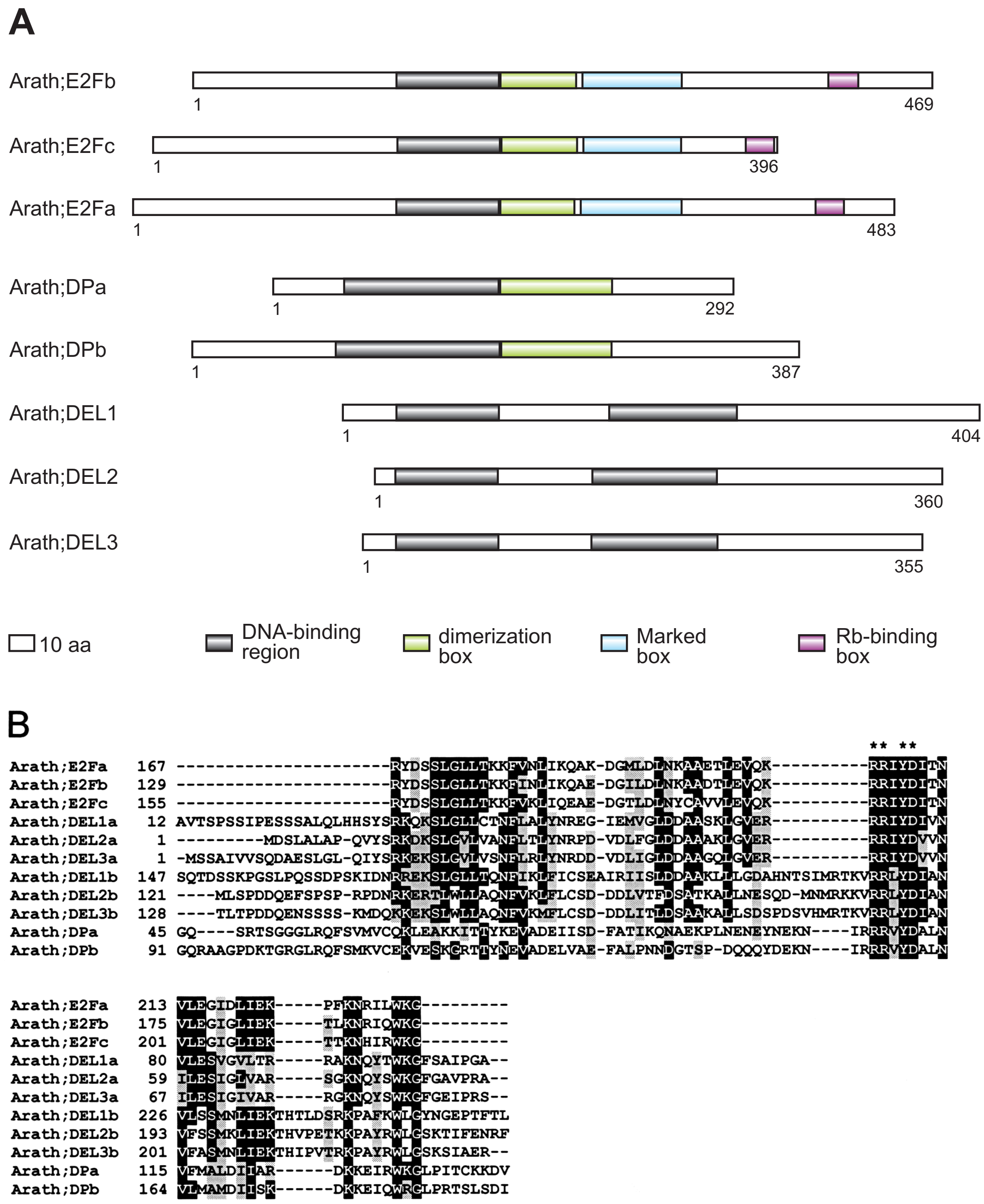

A
Arath;E2Fb
1
469
Arath;E2Fc
1
396
Arath;E2Fa
1
483
Arath;DPa
1
292
Arath;DPb
1
387
Arath;DEL1
1
404
Arath;DEL2
1
360
Arath;DEL3
1
355
10 aa
DNA-binding region
dimerization box
Marked box
Rb-binding box
B
** **
Arath;E2Fa 167 ----------------------RYDSSLGLLTKKFVNLIKQAK-DGMLDLNKAAETLEVQK----------RRIYDITN
Arath;E2Fb 129 ----------------------RYDSSLGLLTKKFINLIKQAE-DGILDLNKAADTLEVQK----------RRIYDITN
Arath;E2Fc 155 ----------------------RYDSSLGLLTKKFVKLIQEAE-DGTLDLNYCAVVLEVQK----------RRIYDITN
Arath;DEL1a 12 AVTSPSSIPESSSALQLHHSYSRKQKSLGLLCTNFLALYNREG-IEMVGLDDAASKLGVER----------RRIYDIVN
Arath;DEL2a 1 ---------MDSLALAP-QVYSRKDKSLGVLVANFLTLYNRPD-VDLFGLDDAAAKLGVER----------RRIYDVVN
Arath;DEL3a 1 -MSSAIVVSQDAESLGL-QIYSRKEKSLGVLVSNFLRLYNRDD-VDLIGLDDAAGQLGVER----------RRIYDVVN
Arath;DEL1b 147 SQTDSSKPGSLPQSSDPSKIDNRREKSLGLLTQNFIKLFICSEAIRIISLDDAAKLLLGDAHNTSIMRTKVRRLYDIAN
Arath;DEL2b 121 ----MLSPDDQEFSPSP-RPDNRKERTLWLLAQNFVKLFLCSD-DDLVTFDSATKALLNESQD-MNMRKKVRRLYDIAN
Arath;DEL3b 128 ----TLTPDDQENSSSS-KMDQKKEKSLWLLAQNFVKMFLCSD-DDLITLDSAAKALLSDSPDSVHMRTKVRRLYDIAN
Arath;DPa 45 GQ----SRTSGGGLRQFSVMVCQKLEAKKITTYKEVADEIISD-FATIKQNAEKPLNENEYNEKN----IRRRVYDALN
Arath;DPb 91 GQRAAGPDKTGRGLRQFSMKVCEKVESKGRTTYNEVADELVAE-FALPNNDGTSP-DQQQYDEKN----IRRRVYDALN
Arath;E2Fa 213 VLEGIDLIEK-----PFKNRILWKG---------
Arath;E2Fb 175 VLEGIGLIEK-----TLKNRIQWKG---------
Arath;E2Fc 201 VLEGIGLIEK-----TTKNHIRWKG---------
Arath;DEL1a 80 VLESVGVLTR-----RAKNQYTWKGFSAIPGA--
Arath;DEL2a 59 ILESIGLVAR-----SGKNQYSWKGFGAVPRA--
Arath;DEL3a 67 ILESIGIVAR-----RGKNQYSWKGFGEIPRS--
Arath;DEL1b 226 VLSSMNLIEKTHTLDSRKPAFKWLGYNGEPTFTL
Arath;DEL2b 193 VFSSMKLIEKTHVPETKKPAYRWLGSKTIFENRF
Arath;DEL3b 201 VFASMNLIEKTHIPVTRKPAYRWLGSKSIAER--
Arath;DPa 115 VFMALDIIAR------DKKEIRWKGLPITCKKDV
Arath;DPb 164 VLMAMDIISK------DKKEIQWRGLPRTSLSDI

Figure 3.1.5 Structural organization of the E2F, DP, and DEL families at the protein level. (a) Scheme of the DNA binding, dimerization, Marked, and Rb binding boxes in *E2F*, *DP*, and *DEL* genes of *Arabidopsis*. (b) Alignment of putative DNA binding domains of E2F, DP, and DEL proteins. All DEL proteins were split in two (parts a and b) to compare both DNA binding motifs with those of E2F and DP. The RRxYD DNA binding motif is indicated by asterisks. Numbers indicate protein length in amino acids (aa).

DEL genes, the conserved DNA recognition motif RRxYD is also present in two copies. The E2F/DP heterodimer binds and recognizes the palindromic sequence of the binding site in an essentially symmetric arrangement (Zheng et al., 1999). Protein secondary structure prediction for the *DEL* genes showed that the winged-helix DNA-binding motif, a fold found in the cell cycle transcription factors E2F/DP (three α helices and a ß sheet), is present in duplex in all these DEL genes. The first and second DEL DNA-binding domain have an overall similarity of 61% and 47% with the E2F DNA-binding domain, respectively. Currently, no experimental data are available about the putative function and role of the *DEL* genes in cell cycle regulation.

Gene/Genome organization

In order to find out whether the segmental or genomic duplications and the acquisition of new cell cycle regulation mechanisms are linked, we mapped all cell cycle genes on the five different chromosomes (Figure 3.1.6). Subsequently, all duplicated regions in the *Arabidopsis* genome were defined and the position of every cell cycle gene was compared with the coordinates of each duplicated block.

Comparison of the position of A2 cyclin genes with the position of duplicated blocks in the *Arabidopsis* genome revealed that all four members are located in duplicated blocks: one internal duplication on chromosome 1 (*CYCA2;3* linked with *CYCA2;4*) and one on chromosome 5 (*CYCA2;2* linked with *CYCA2;1*). The three CYCA3 genes were organized in tandem (*CYCA3;2*, *CYCA3;3*, and *CYCA3;4* spanning a region of less than 8 kb) and have a highly similar gene structure (number and size of exons and introns), as well as highly similar protein sequences (74.3% overall similarity). Only *CYCA3;2* had one significant EST hit, whereas *CYCA3;4* had an additional small predicted exon (33 nucleotides) when compared with the other CYCA3 genes that occur in the same tandem (Figure 3.1.3b). Similar to the A2-type cyclins, all four B2-type cyclins were located within duplicated blocks: one duplicated block between chromosomes 2 and 4 (linking *CYCB2;1* and *CYCB2;2*) and one internal duplication on chromosome 1 (linking *CYCB2;3* and *CYCB2;4*).

Although in total 10 D-type cyclins were detected, only few of them were located in duplicated blocks. *CYCD3;2* and *CYCD3;3* are members of an inverted block between chromosome 5 and 3, whereas *CYCD4;1* and *CYCD4;2* are located within an internal block of chromosome 5. The two CKS genes were located in a

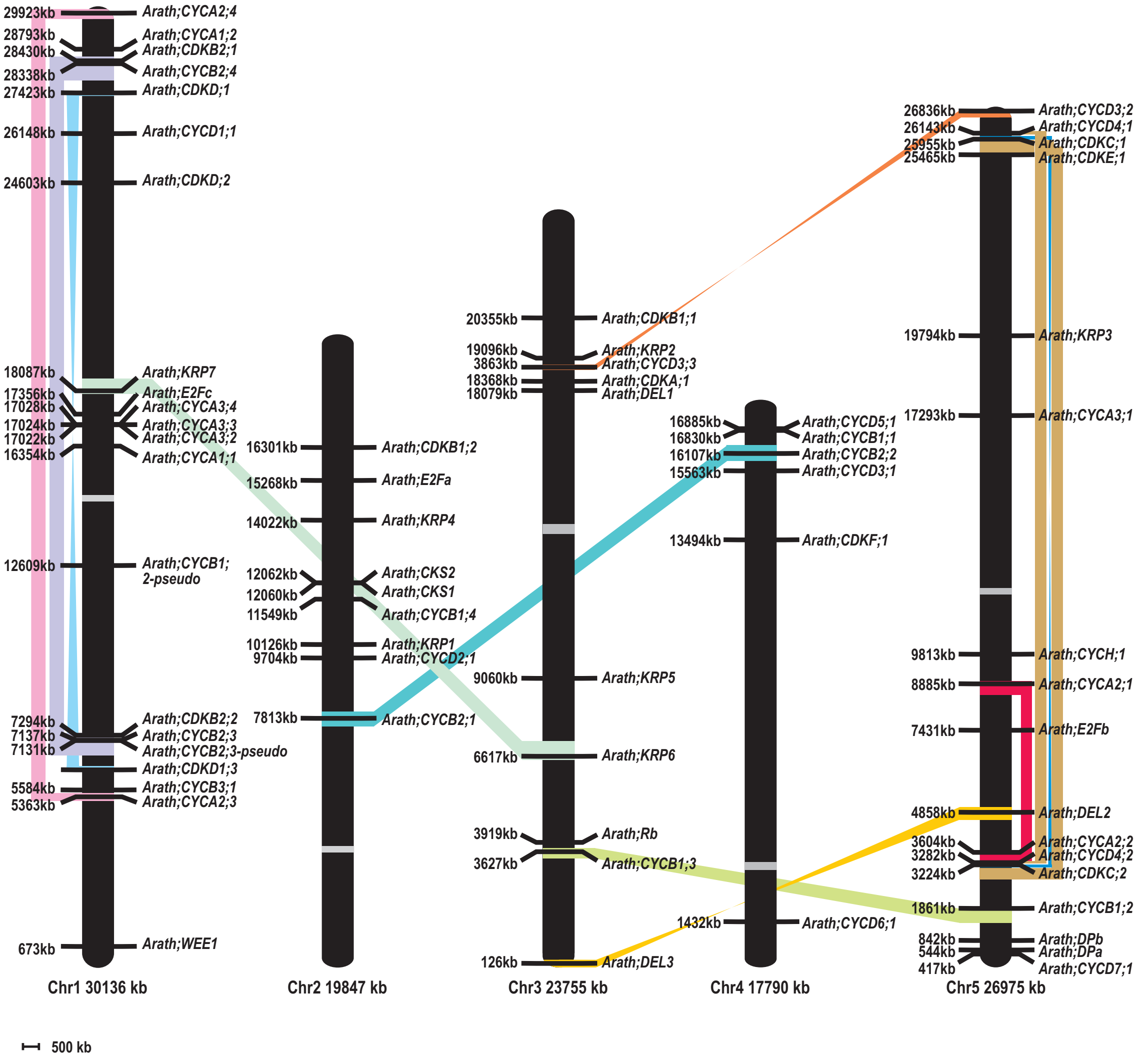


Figure 3.1.6 Physical positions of core cell cycle genes on the *Arabidopsis* genome. Segmental duplicated regions are shown only when a cell cycle gene is present in a duplication event. Colored bands connect corresponding duplicated blocks. Duplicated blocks in reverse orientation are connected with twisted colored bands. Centromeres are represented as gray boxes. Chr1 to Chr5, chromosomes 1 to 5.

gene tandem duplication, where the stop codon of *CKS2* was separated by only 916 bp from the start codon of *CKS1* (Figure 3.1.3a).

Special attention is required for two duplication events. On chromosome 1, a large internal duplication occurred (spanning an area of approximately 4,890 kb or 16% of chromosome 1) that was followed by several inversions (data not shown), leading to the formation of multiple smaller blocks, one of which contained two pairs of cell cycle genes: *CDKB2;2* linked with *CDKB2;1* and *CYCB2;3* linked with *CYCB2;4*. The *CYCB2;3* gene was present in tandem (interspersed by one gene) and the second copy was designated Arath;CYCB2;3_pseudo, because its gene structure was degraded and imperfect with respect to *CYCB2;3*. We conclude that this tandem duplication occurred after the segmental duplication event, because in the region linked to the duplicated block, no trace of another extra B2-like cyclin was found. Another special, internally duplicated event was found on chromosome 5. Two duplicated blocks (Figure 3.1.6, brown blocks) were detected that connected both extremities of the chromosome. Although these blocks could be regarded as one, we clearly distinguished an invertedly duplicated block in between (Figure 3.1.6, blue block). *CYCD4;1* and *CYCD4;2* both fit nicely into the first block. *CDKC;1* and *CDKC;2* mapped in this region as well, located in the small invertedly duplicated block. It is remarkable that, although both couples of linked genes were located in duplicated blocks with different orientations, their relative positions were the same (i.e. at the bottom and the top of chromosome 5, a C-type CDK was followed by a D4-type cyclin). This configuration suggests that initially one large duplication event occurred (Figure 3.1.6; the region spanning brown and blue blocks) that was later reshuffled by inversions (and perhaps some deletions), resulting in adjacent, duplicated blocks with different orientations and sizes.

Discussion

The members of the *Arabidopsis* genome sequencing consortia use different tools to perform automated genome annotations together with similarities to ESTs and known protein sequence to refine gene models. This procedure has generated a large quantity of information on the *Arabidopsis* gene repertoire. However, the extraction of clear biological information for a particular process from these public databases is not always that trivial (for instance, the word ‘cyclin’ as query in the MIPS database returned 37 hits with 23 putative cyclin or cyclin-like hits). To solve this problem, we designed a protocol, mainly focused on high-quality homology-based annotation.

We used a combination of two selected high-quality *Arabidopsis* prediction tools (Pavy et al., 1999; Schiex et al., 2001; C. Mathé and P. Rouzé, personal communication), together with pure experimental information as reference material. A first advantage of this method is that the chance of finding new and rarely expressed genes is maximized because it is structurally characterized by tools with higher specificity and sensitivity than those used by the different consortia for generating genome annotation (Gopal et al., 2001). Secondly, focus on families with available experimental references allows comparisons with functionally well-characterized genes and diminishes the risk of propagation of wrong annotation. In addition, the use of hidden Markov profiles, which represent the complete diversity within a family, is clearly more powerful than that of a single sequence for remote-homolog detection (Karplus et al., 1998).

With this strategy, we have built a catalogue of 61 core cell cycle genes, belonging to seven selected families. From these, 30 had not been described before and for 22 of them the gene prediction provided by the *Arabidopsis* Genome Initiative was incorrect. Corrected gene models have been submitted to TAIR and can also be found at the web site http://bioinformatics.psb.ugent.be/. These results highlight the complexity of the cell cycle regulation in *Arabidopsis*, indicating a larger variety of genes than what was currently known experimentally.

Like in mammals, plants evolved to use different classes of CDKs to regulate their cell cycle. In *Arabidopsis*, a total of six different CDK classes can be identified, designated from A through F. Although some of these CDKs have been proven to be active during specific phases of the cell cycle (Magyar et al., 1997; Porceddu et al., 2001; Sorrell et al., 2001), no functional correlation can be made with CDKs of other eukaryotes on the basis of protein sequences. For example, no clear ortologs can be identified for the mammalian G1/S-specific *CDK4* and *CDK6*, suggesting that plants developed independently additional CDKs for more specialized functions in the cell cycle control. This hypothesis is in agreement with the observation that the cyclin-binding motifs found in the plant B-type CDKs cannot be found in any CDK of other eukaryotes.

Within the CDK family, we identified three new CAK members, being close homologs of the rice R2 gene (Hata, 1991). These CAKs (*CDKD;1*, *CDKD;2* and *CDKD;3*) differ structurally from the previously isolated *Arabidopsis cak1At*, renamed *CDKF;1*. The high sequence diversity (35% overall sequence similarity between D- and F-type CDKs) suggests that plants utilize two distinct classes of CAKs. When the *Arabidopsis CDKF;1* is compared with the rice R2, both classes are functionally different: they both can complement yeast CAK mutant strains, but show a different substrate specificity; the rice R2 phosphorylates both CDKs and the carboxyl-terminal domain of the largest subunit of RNA polymerase II, whereas *CDKF;1* phosphorylates CDKs only (Umeda et al., 1998; Yamaguchi et al., 1998).

The complexity of the cyclin gene family appears to be higher in plants than in mammals. Compared to human, *Arabidopsis* has approximately 14 more A- and

B-type cyclins, and seven more D-type cyclins. A major part of the A-cyclins originated through large segmental duplications. For the 10 A-type cyclins, all four members of the A2-type subclass are part of duplicated blocks and three genes out of the four A3-type cyclins are organized in tandem. Several analyses of the *Arabidopsis* genome sequence had already concluded that genes had duplicated extensively in the history of the model plant. More than 50% of the genes in *Arabidopsis* belong to a gene family with three or more members. After analyzing regions of chromosomes 2, 4 and 5, Blanc et al. (2000) estimated that more than 60% of the genome consisted of duplicated regions and suggested the possibility that *Arabidopsis* was an ancient tetraploid. In a later analysis, Vision et al. (2000) concluded that in fact several large independent duplications of chromosome segments had happened at different time points in the plants' evolution. This view was blurred by extensive deletion, inversion and translocation of genes and chromosome segments, as well as smaller and tandem gene duplications (The Arabidopsis Genome Initiative, 2000; Vision et al., 2000). In our analysis, we detected that 22 core cell cycle genes are part of a segmental duplication in the *Arabidopsis* genome. Whether there is functional redundancy within A- and B-type cyclins, or whether some cyclin subclasses are differently regulated (and expressed) will have to be analyzed.

In contrast to the A- and B-type cyclins, D-type cyclins lack high sequence similarity among each other, which is reflected within the phylogenetic analysis resulting in seven D-type subclasses. When compared with A- and B-type cyclins, of which some complete subclasses (A2 and B2) are located within segmentally duplicated blocks, no large duplications can be found for the D-type cyclins. Only the D3 and D4 subclasses have different members. Redundancy of the D3-type cyclins has been proposed previously as an explanation of the failure to observe mutant phenotypes, when knocking out a single D3-type cyclin (Swaminathan et al., 2000). Our analysis clearly confirms this hypothesis: the fact that two D3-type cyclins are linked via a recent segmental duplication strengthens our belief that these D3 cyclins are functionally redundant. A similar hypothesis could hold for D4-type cyclins, because two out of three are located in a duplicated block.

The much larger divergence seen for D-type cyclins when compared to A- and B-type cyclins might reflect the presumed role of D-type cyclins in integrating developmental signals and environmental cues into the cell cycle. For example, D3-type cyclins have been shown to respond to plant hormones, such as cytokinins and brassinosteroids, whereas *CYCD2* and *CYCD4* are activated earlier in G1 and react to sugar availability (for review, see Stals and Inzé, 2001). Because of the large number of various D-type cyclins with different response to developmental and environmental signals, cell division and growth in sessile plants might be more flexible than what is observed in other eukaryotes.

Whereas plants clearly share all elements needed for G1/S entry with other higher eukaryotes, they lack the typical class of E-type cyclins, known to be essential regulators of DNA replication (Duronio et al., 1996). Presumably some of the A- or D-type cyclins take over the role of the E-type cyclins. Also the lack of a consensus Rb-binding motif in some D-type cyclins suggests that some cyclins might have gained other novel functions during evolution. Alternatively, some of the core cell cycle genes might have undergone such dramatic changes during evolution that they cannot be recognized anymore as functional homologs of animal and yeast counterparts, of which the *CDC25* gene is the most likely example. Both the presence of the antagonistic *WEE1* kinase and accumulating biochemical evidence point to the existence of a *CDC25* phosphatase in plants (Zhang et al., 1996; Sun et al., 1999), although it could not be identified as such in the *Arabidopsis* genome.

It is surprising that mammals and plants have approximately the same number of core cell cycle genes, with the exception of the above described difference in cyclin number. Complex, multicellular organisms may need many more cell cycle genes to coordinate cell cycle progression with the diverse developmental pathways. Therefore, the pool of mammalian cell cycle genes is probably larger than expected because of the frequent occurrence of alternative splicing. For example, spliced variants of cyclin E are known, with an expression profile and substrate specificity different from that of cyclin E itself (Mumberg et al., 1997; Porter and Keyomarsi, 2000). At least five distinct DP-2 mRNAs are synthesized in a tissue-specific fashion (Rogers et al., 1996). Depending on the splice variant, the DP family members lack a nuclear localization signal and, when associated with E2F, these different DP molecules have opposing effects on the E2F/DP activity (De la Luna et al., 1996). Furthermore, alternative splicing in humans is known for CDKs, CDC25, and CKIs (Wegener et al., 2000; Hirano et al., 2001; Herrmann and Mancini, 2001). For cell cycle genes of plants, only one case of alternative splicing has been reported (Sun et al., 1997).

E2F/DP transcription factors are characterized by the presence of both a DNA-binding and transcription activation domain. Binding of these transcription factors to the E2F/DP palindromic binding site is mediated by a small DNA recognition motif (RRxYD). By scanning the genome for E2F/DP-related proteins, a putatively novel class of cell cycle-regulating genes was identified, designated DEL. The DEL proteins have two E2F-like DNA-binding boxes, each including the RRxYD motif, but have no activation domain. By competing for the same DNA binding sites, monomeric DEL proteins could act as competitors of the E2F/DP proteins and, because they lack an activation domain, they would act as a repressor of E2F/DP-regulated genes. This mechanism would avoid G1-to-S transition, in cases where conditions are not appropriate for entry in the S phase (such as DNA damage and stress). This new class of putative cell cycle regulators seems not to be plant specific, because one homolog was found in *Caenorabditis elegans* (data not shown). In conclusion, our genome-wide analysis demonstrated an unexpected complexity of the core cell cycle machinery in plants that is comparable with that seen in mammals. The major challenge for the future is to understand the specific role of all these individual genes in regulating cell division during plant development.

Note added in proof

The postulated function of the DEL proteins has recently been confirmed (Mariconti, L., Pellegrini, B., Cantoni, R., Stevens, R., Bergounioux, C., Cella, R., and Albani, D. [January 10, 2002] J. Biol. Chem. 10.1074/jbc.M110616200), but the gene prediction for one DEL family member (E2Ff~DEL3) differs from the one we present here. The gene structure we propose has been validated experimentally in our laboratory.

3.2 Exploring the plant transcriptome through phylogenetic profiling

Klaas Vandepoele and Yves Van de Peer

Redrafted from: *Plant Physiology* **137**, 31-24 (2005)

Publicly available protein sequences represent only a small fraction of the full catalogue of genes encoded by the genomes of different plants, such as green algae, mosses, gymnosperms and angiosperms. In contrast, an enormous amount of expressed sequence tags exists for a wide variety of plant species, representing a substantial part of all transcribed plant genes. Integrating protein and EST sequences in comparative and evolutionary analyses is not straightforward because of the heterogeneous nature of both types of sequence data. By combining information from publicly available EST and protein sequences for 32 different plant species, we identified more than 250,000 plant proteins organized in over 12,000 gene families. Approximately 60% of the proteins are absent from current sequence databases, but provide important new information about plant gene families. Analysis of the distribution of gene families over different plant species through phylogenetic profiling reveals interesting insights into plant gene evolution, and identifies species and lineage-specific gene families, orphan genes, and conserved core genes across the green plant lineage. We counted a similar number of approximately 9,500 gene families in monocotyledonous and eudicotyledonous plants and found strong evidence for the existence of at least 33,700 genes in rice. Interestingly, the larger number of genes in rice compared to *Arabidopsis* can partially be explained by a larger amount of species-specific single copy genes and species-specific gene families. In addition, a majority of large gene families, typically containing more than 50 genes, is bigger in rice than *Arabidopsis*, whereas the opposite seems true for small gene families.

Introduction

Comparative genomics provides a powerful means to study gene structure and the evolution of gene function and regulation. Analysis of genes or pathways in a broad phylogenetic context allows scientists to better understand how complex biological processes are regulated and evolve (Soltis and Soltis, 2003; Koonin et al., 2004). Although phylogenetic studies can provide important insights into gene and genome evolution (for examples, see Ermolaeva et al., 2003; Griffiths et al., 2003; Vandepoele et al., 2003), a dense taxonomical sampling is necessary to obtain a complete and accurate view of the evolutionary history of a biological process and its underlying genes. Similarly, to draw biologically relevant conclusions, the inference of orthology and paralogy between homologous genes requires a good phylogenetic sampling (for review, see Doyle and Gaut, 2000). Moreover, a coherent classification of homologous genes is essential for the high-throughput extraction of functional and evolutionary information from gene phylogenies. In this respect, the availability of numerous large-scale sequencing projects offers the opportunity to study homologous genes, typically gene families, from an evolutionary point of view. The construction of phylogenetic profiles, which reflect the presence or absence of a particular gene family in a biological species, is an effective method for the detection of conserved core genes, species-specific single copy genes, species or lineage-specific gene family expansions, gene loss and genes that have been transferred between nuclear and organellar genomes. Furthermore, analysis of the phylogenetic profiles of protein families and of domain fusion events helps to predict functional interactions and to deduce specific functions for numerous proteins (Kriventseva et al., 2001).

Perhaps the best known example of an integrated sequence-based system applying phylogenetic profiles is the COG database, which is a comprehensive repository of functionally annotated clusters of bacterial and eukaryotic orthologous genes (Tatusov et al., 2003). Although in Bacteria, Fungi and animals various sequencing projects constantly enlarge the gene space (for an overview, see http://www.ncbi.nlm.nih.gov/genomes/static/EG_T.html), the situation is different for plants (Pryer et al., 2002). Apart from *Arabidopsis thaliana* and *Oryza sativa* (rice), where genome sequencing projects present a first overview of the eudicotyledonous and monocotyledonous gene repertoire, respectively (Arabidopsis Genome Initiative, 2000; Feng et al., 2002; Goff et al., 2002: Sasaki et al., 2002; Yu et al., 2002; Rice Chromosome 10 Sequencing Consortium, 2003), the majority of all other *Viridiplantae*, ranging from early land plants such as mosses and ferns, to highly developed flowering plants lack a comprehensive overview of the proteins encoded by their genomes. On the other hand, an enormous amount of plant expressed sequence tags (ESTs) - single-pass sequence reads from reverse transcribed mRNAs - is publicly available and provides a substantial representation of the plant transcriptome (Rudd 2003). Because the overall number of plant ESTs is by far larger than that of plant proteins currently stored in public sequence repositories, the phylogenetic analysis of plant genes based on complete protein sequences is difficult and inefficient (Raes et al., 2003) and offers only a very limited view on the total amount of plant sequence information currently available.

Here, we present an integrated sequence repository (available as – Sequence platform for the Phylogenetic analysis of Plant Genes [SPPG] - in the section Databases at http://bioinformatics.psb.ugent.be/) that combines EST sequence data with protein information, providing an excellent starting point for plant comparative and evolutionary genomics. This is illustrated by the examination of several thousands of gene families distributed over a large number of different plant species, which reveals unique features about the evolution of plant gene families.

Results and discussion

EST assembly, ORF detection, protein clustering and functional annotation

Initially, 106,174 proteins and 2,884,000 EST sequences from 32 different plant species were retrieved from EMBL and TIGR to construct a non-redundant and high-quality data set of plant proteins. After the assembly of the EST sequences, annotation of open-reading frames (ORFs) on EST clusters, and processing all currently available proteins for the plant species selected here (see Materials and methods for technical details), a total of 86,077 non-redundant plant proteins from EMBL and TIGR were obtained, together with 253,857 EST clusters derived from more than 1.8 million clustered EST sequences (Table 3.2.1; Figure 3.2.1). Fifty-seven percent of all initial EST sequences could be assembled into an EST cluster comprising on average 6.16 ESTs. These results are very comparable with similar plant EST assembly initiatives (TIGR Plant Gene Indices; Quackenbush et al., 2001 and PlantGDB; Dong et al., 2004). Nevertheless, because we applied more stringent assembly criteria to reduce the creation of chimeras and other artificial cDNAs (see Materials and methods), the overall number of EST assemblies per species is slightly smaller than that in PlantGDB and TIGR Gene Indices. For two-thirds of all EST clusters an ORF longer than 50 codons could be determined, resulting in 166,306 protein sequences (Figure 3.2.1). Thus, in total 252,383 non-redundant plant proteins were assigned to the final data set. Approximately 82% of all proteins (which corresponds to 207,023 proteins) could be assigned to 14,369 gene families, here defined as a set of two or more homologous gene sequences. Overall, a good correlation between the initial number of ESTs and the final number of clustered plant proteins was observed (r^2=0.88), which indicates that there is no significant bias in the EST assembly and ORF annotation routines applied for these different *Viridiplantae* species (see Materials and methods for details). Whereas a minority of gene families (i.e. 4,275) contains only proteins derived from EST clusters, the majority (i.e. 10,094) consists of proteins from EMBL, TIGR or both. In addition, 46% (6,664) of all gene families contain proteins derived from both EST clusters and EMBL or TIGR. Consequently, this subset corresponds to gene families with a dense sampling over the different plant species included in the data set, with an average total of 27 proteins per family from 9.7 different plant species. In contrast, the overall sampling density for all 14,369 gene families is 14.4 genes sampled over 5.5 different plants per family. Despite the fact that only 25% of all proteins derived from EST clusters are truly full-length (i.e. the protein begins with a start codon and ends with a stop codon), the majority (86%) of all these proteins has significant homology with other proteins, offering additional information for the phylogenetic profiles (see below). Approximately 45,000 protein sequences were not clustered into gene families. Although 30% of these unclustered proteins represent single-copy species-specific genes (or orphan genes, see below), the majority corresponds to partial proteins, derived from incomplete ORFs annotated on non full-length EST clusters, with sometimes only partial homology to other plant proteins. Indeed, one might expect that a number of gene families only comprising proteins derived from EST clusters will represent partial proteins. These proteins will not be clustered with the corresponding full-length proteins because they do not fulfill the global homology criterion required for being added to such a group of related proteins. We estimate that approximately 11% of all gene families form a group of related partial proteins, derived from EST clusters, for which a related full-length gene family exists (see Materials and methods).

Table 3.2.1 Number of EST and protein sequences combined in EST assembly, CDS annotation and protein clustering

Species name	FrameD IMM[a]	ESTs	EST-clusters	ESTs in cluster (%)	ESTs per cluster	ORF from EST-clusters[b]	ePROT[b,c]	Total proteins	Proteins in gene family	Orphan proteins[d]
Arabidopsis thaliana	-	-	-	-	-	-	26,294	26,294	22,412	1,050 (253)
Beta vulgaris	Caryo	19,039	2,334	0.33	2.72	1,785	76	1,855	1,607	38
Brassica napus	Arabidopsis	37,896	4,913	0.58	4.49	3,676	813	4,350	3,996	95
Capsicum annuum	Aster	23,361	2,420	0.53	5.09	1,835	218	2,013	1,885	16
Chlamydomonas reinhardtii	Chlamy	141,100	9,528	0.72	10.62	2,480	738	3,015	1,889	131
Glycine max	Arabidopsis	342,149	28,726	0.78	9.32	19,189	1,332	20,180	17,726	584
Gossypium arboreum	Arabidopsis	38,967	3,577	0.43	4.67	2,613	38	2,648	2,397	57
Gossypium hirsutum	Arabidopsis	14,307	1,150	0.32	4.03	742	609	1,114	1,040	21
Helianthus annuus	Aster	60,785	5,024	0.60	7.30	2,589	306	2,827	2,496	70
Hordeum vulgare	Oryza	202,705	14,925	0.74	10.03	10,060	1,150	10,894	9,752	201
Lactuca sativa	Aster	69,319	6,698	0.66	6.87	4,679	83	4,755	4,150	90
Lotus corniculatus var. Japonicus	Arabidopsis	24,896	2,695	0.74	6.85	1,780	132	1,893	1,568	127
Lycopersicon esculentum	Aster	151,147	14,428	0.82	8.62	10,478	1,442	11,546	10,292	155
Medicago truncatula	Arabidopsis	188,367	16,867	0.83	9.22	12,567	233	12,772	11,143	191
Mesembryanthemum crystallinum	Caryo	26,563	2,471	0.70	7.50	1,673	207	1,857	1,674	34
Nicotiana tabacum	Aster	11,276	453	0.11	2.70	69	1,697	1,435	1,289	47
Oryza sativa	-	-	-	-	-	-	47,475	47,475	30,993	7,882 (704)
Physcomitrella patens	Physco	104,161	12,924	0.82	6.65	10,217	355	10,298	6,319	2,053
Pinus pinaster	Pinus	9,059	1,222	0.51	3.79	835	66	895	822	9
Pinus taeda	Pinus	73,349	5,325	0.40	5.53	2,895	154	3,004	2,466	97
Populus balsamifera subsp. Trichocarpa	Arabidopsis	24,579	2,320	0.58	6.18	1,725	49	1,758	1,621	28
Populus tremula	Arabidopsis	14,081	1,120	0.42	5.23	833	33	839	791	16
Populus tremula x Populus tremuloides	Arabidopsis	56,048	4,757	0.55	6.50	3,401	68	3,467	3,197	81
Populus x canescens	Arabidopsis	10,499	754	0.26	3.60	526	20	546	513	13
Prunus persica	Arabidopsis	10,939	1,071	0.54	5.52	829	127	940	886	12
Solanum tuberosum	Aster	95,632	16,151	0.78	4.63	11,634	985	12,341	10,755	161
Sorghum bicolor	Oryza	134,740	15,303	0.80	7.06	10,925	426	11,278	9,771	219
Sorghum propinquum	Oryza	21,390	3,148	0.65	4.44	2,349	6	2,355	2,130	28
Triticum aestivum	Oryza	508,406	35,271	0.45	6.51	22,615	1,432	23,788	21,115	535
Vitis vinifera	Arabidopsis	111,849	11,163	0.82	8.17	6,674	233	6,861	5,952	260
Zea mays	Oryza	362,796	26,807	0.65	8.81	14,704	3,819	16,919	14,256	397
Zinnia elegans	Aster	9,836	312	0.07	2.15	119	57	171	120	4
Total / ***Average***		2,899,241	253,857	*0.57*	*6.16*	166,496	96,219	252,383	207,023	14,457

[a] The Interpolated Markov Model used to determine the ORF of an EST cluster (see Material and Methods for more details). [b] The number of sequences before removing redundant entries. [c] The number of protein sequences available in EMBL/TIGR after removing transposon-like genes in *Arabidopsis* and rice. [d] The number in parenthesis give the number of predicted orphan genes supported by EST/cDNA (match with >95% identity across >100 bp) for *A. thaliana* and *O. sativa* .

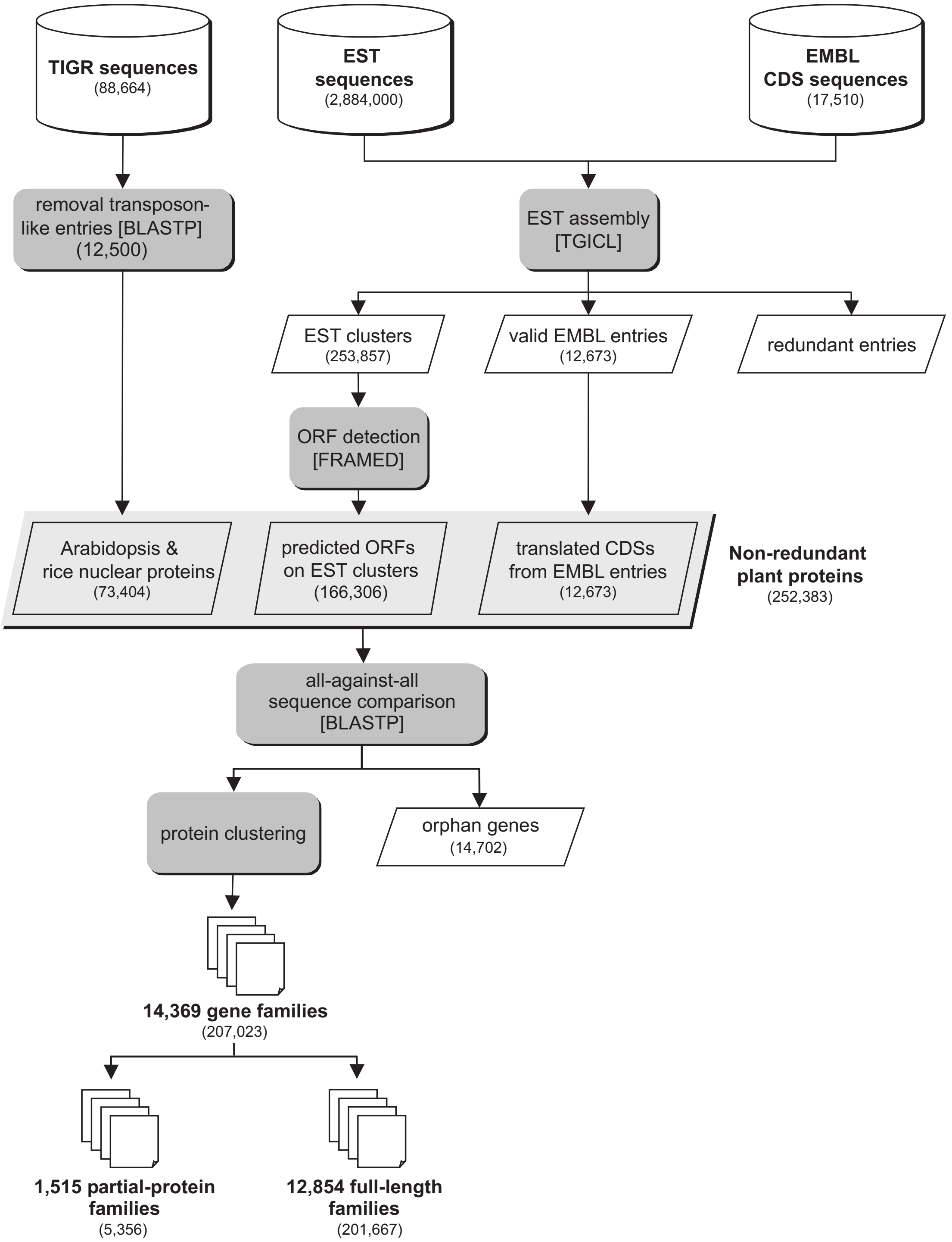


Figure 3.2.1 Schematic overview of the construction of the data set. The white barrels represent the initial sequence data retrieved from TIGR and EMBL, the dark grey boxes routines applied to manipulate and organize the data, whereas the light grey box describes the final amount of sequence data derived from the different sources (see text for details). Except for *Arabidopsis* and rice, whose nuclear protein-encoding genes were retrieved from TIGR, all other sequence data for the 32 species was obtained through EMBL. The numbers of sequences are indicated in brackets.

Gene families and individual genes have been functionally annotated based on the available gene descriptions and Gene Ontology (GO) annotations of protein sequences derived from EMBL and TIGR. Approximately 58,000 gene descriptions could be mapped on 11,938 different gene families and 22,395 functional GO labels of *Arabidopsis* could be assigned to 4,099 gene families. When gene descriptions are transferred between different members of the same gene family, more than 80% of plant sequences can be labeled with functional information.

Gene content in chloroplast and mitochondrial genomes

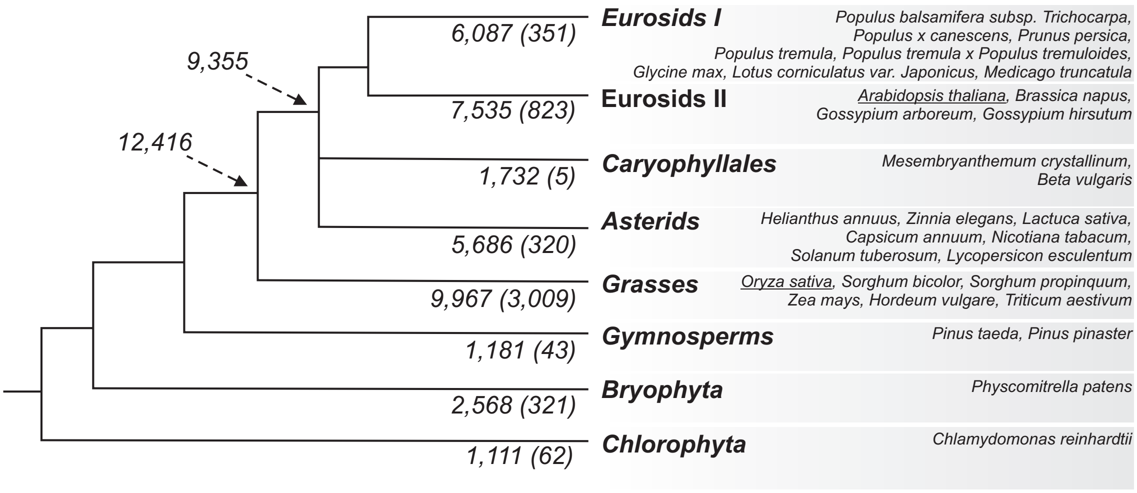
In addition to assigning general gene descriptions to families or individual proteins, information about the nuclear or organellar origin of genes has also been integrated, which allows us to determine the amount of chloroplast and mitochondrial DNA sequences that have been inserted into or transferred to the nucleus. In total, 704 chloroplast and 275 mitochondrial gene products were identified that could be clustered into 202 distinct gene families. Interestingly, in numerous gene families genes from different origins were grouped. Sixty-six and 24 gene families were found uniquely for chloroplast or mitochondrial genomes, respectively, whereas 110 organelle families were identified for which homologs were also detected in the nuclear plant genome of *Arabidopsis* or rice. Two gene families were identified encoded by the chloroplast and mitochondrial genome (NADH dehydrogenase subunits 1 and 4). Gene families in both mitochondrial and nuclear genomes encode for cytochrome c subunits, ribosomal proteins and transfer RNAs, whereas a wide variety of genes, covering 66 different gene families, was found in both chloroplast and nuclear genomes (for full list, see Supplemental Table I available at www.plantphysiol.org). In addition, ten families were identified in the mitochondrial, chloroplast and nuclear genomes of different species encoding ribosomal proteins, NADH dehydrogenase subunits, Fe-superoxide dismutase, ATP synthase subunit 1 and an asparagine transfer RNA. This confirms previous findings that genes frequently are transferred from the chloroplast or mitochondrial genome to the nucleus, where they acquire new expression control and targeting signals for the correct expression, translation and re-import into the organelle (Martin, 2003).

Strikingly, whereas 19% (15 out of 76 gene families) of all chloroplast gene functions in *Arabidopsis* are also present in the nuclear genome, in rice 37% (30 out of 81 gene families) of all chloroplast gene functions are found in the nuclear genome. This difference confirms previous findings that the rice nuclear genome is significantly more enriched with plastid genome sequences than that of *Arabidopsis* (Shahmuradov et al., 2003). Although recent gene transfers from the chloroplast to the nuclear genome might be associated with chloroplast genome reduction due to subsequent gene loss, the overall number of distinct gene functions in the rice chloroplast genome is not significantly different from that of the *Arabidopsis* chloroplast genome (81 and 76 gene families, respectively). Therefore, it is currently unclear whether this current redundancy represents the first step of the transfer of chloroplast gene functions to the rice nucleus and has any evolutionary consequences (Timmis et al., 2004).

Application of phylogenetic profiles for the evolutionary classification of plant genes

An overview of the number of proteins ascribed to gene families is shown in Table 3.2.1. As expected, the largest numbers of proteins that can be assigned to gene families are derived from *Arabidopsis* and rice (22,412 and 30,993 genes, respectively), for which nearly complete nuclear genome sequences have been determined. Monocotyledonous plants, such as *Triticum aestivum*, *Zea mays*, *Sorghum bicolor* and *Hordeum vulgare*, are also well represented, as well as the eudicotyledonous plants *Glycine max*, *Medicago truncatula*, *Solanum tuberosum*, *Lycopersicon esculentum* and *Vitis vinifera*. For the moss *Physcomitrella patens*, more than 6,300 proteins are clustered into gene families, which can be explained by the exhaustive EST-sequencing efforts lately (Nishiyama et al., 2003). In contrast, for other plants only a limited number of protein sequences are available.

In addition to defining *sensu stricto* phylogenetic profiles at the species level, we also determined the overall presence of each gene family over distinct taxa of the *Viridiplantae*. The different taxa scored were, at lower taxonomic levels, *Chlorophyta*, *Bryophyta*, gymnosperms and angiosperms, the latter being further subdivided in monocots and eudicots. At a higher taxonomic level, *eurosids I*, *eurosods II*, *rosids*, *asterids* and *Caryophyllales* were discerned. Given the still very incomplete nature of most available plant gene sequences, these high-level phylogenetic profiles offer an alternative representation of the distribution of gene families within the green lineage (Figure 3.2.2). Moreover, these alternative profiles provide a valuable tool for the extraction of information about the evolution of gene functions.


9,355
12,416
6,087 (351)
Eurosids I
Populus balsamifera subsp. Trichocarpa, Populus x canescens, Prunus persica, Populus tremula, Populus tremula x Populus tremuloides, Glycine max, Lotus corniculatus var. Japonicus, Medicago truncatula
7,535 (823)
Eurosids II
Arabidopsis thaliana, Brassica napus, Gossypium arboreum, Gossypium hirsutum
1,732 (5)
Caryophyllales
Mesembryanthemum crystallinum, Beta vulgaris
5,686 (320)
Asterids
Helianthus annuus, Zinnia elegans, Lactuca sativa, Capsicum annuum, Nicotiana tabacum, Solanum tuberosum, Lycopersicon esculentum
9,967 (3,009)
Grasses
Oryza sativa, Sorghum bicolor, Sorghum propinquum, Zea mays, Hordeum vulgare, Triticum aestivum
1,181 (43)
Gymnosperms
Pinus taeda, Pinus pinaster
2,568 (321)
Bryophyta
Physcomitrella patens
1,111 (62)
Chlorophyta
Chlamydomonas reinhardtii

Not shown: *Vitis vinifera*

Figure 3.2.2. Phylogenetic distribution of all gene families over different taxa of the *Viridiplantae*. The number of gene families in one or more species belonging to a particular taxon is shown beneath the branches. The number of gene families exclusively found for a particular taxon is shown between parentheses. Families grouping partial protein sequences were discarded (see text for details). Arrows indicate the number of gene families found in the eudicots (9,355) and angiosperms (12,416).

Core plant genes, species- and lineage specific gene families, and orphans

Examination of the high-level phylogenetic profiles revealed that a total of 397 gene families covering 53,796 proteins were present in *chlorophytes*, *bryophytes*, gymnosperms and angiosperms. These conserved gene families thus represent a set of core genes found in all major divisions of the *Viridiplantae*. As expected, the functional classification of these gene families shows that they encode basic components of the plant cell machinery, such as genes involved in translation, ribosomal structure, post-translational modifications, energy production, secretion, amino acid transport and metabolism (see Supplemental Figure 1). The number of core proteins in *Arabidopsis* identified here (4,177) is larger than the 1,152 *Arabidopsis* proteins conserved in all eukaryotes (Guttierez et al., 2004), which can be explained by the presence of a large number of gene functions specific to the green lineage but absent from other eukaryotic kingdoms. Indeed, we find that only 10% of the *Arabidopsis* plant core genes is part of the eukaryotic core as defined by Guttierez (2004), suggesting a large number of plant-specific core gene functions. As expected, a large number of these plant-specific core genes are involved in photosynthesis. Surprisingly, when combining the set of 3,848 plant-specific *Arabidopsis* proteins identified by Guttierez et al. (2004) with the phylogenetic profiles computed here, only 3% of these proteins belong to the set of core gene families. This indicates that a large fraction of these putative plant-specific genes are part of species- or lineage specific gene families and do not belong to the set of plant core genes, as it is defined now by including more plant species. It should be noted that in our data set only eight gene families were found in all 32 plant species, which is very illustrative for the current poor status of gene sampling in plants. Surprisingly, 26 core gene families (i.e. 7% of all core families) correspond to genes with unknown function, which suggests that they represent essential, albeit unexplored, gene functions in plants. This number is significantly higher than the 2% of uncharacterized core gene families in pan-eukaryotic KOGs (18/860, http://www.ncbi.nlm.nih.gov/COG/new/ ; Koonin et al., 2004). Genes typically used for reconstructing the phylogenetic relationships between different plant phyla were found in a majority, if not all, species (e.g. tubulin, actin, Rubisco subunits, heat shock protein hsp70 and elongation factor 1 alpha).

In contrast to the set of core genes, a large number of gene families are specific to one particular plant. Initially, 3,337 species-specific gene families (SSGFs) were identified when querying the profiles of all gene families. Because the general gene family delineation was performed with rather conservative criteria, less stringent protein clustering parameters were applied in order to determine the real number of SSGFs, lineage-specific families (LSGFs) and orphan genes (see Materials and methods). In total, 1,116 SSGFs containing 5,180 proteins were detected, with the largest number in rice, *Arabidopsis* and *Physcomitrella*, covering 637 (~4,258 proteins), 187 (~1,241 proteins) and 164 (~408 proteins) gene families, respectively. The availability of a complete genome sequence for *Arabidopsis* and rice may be the reason for the larger number of SSGF proteins, whereas for *Physcomitrella* the absence of sequence data from closely related species in combination with the large number of available EST/cDNA sequences explains the high amount of SSGF proteins. Approximately 82% of all SSGF proteins lack a functional annotation, which indicates that they play a role in unknown or poorly characterized biological processes. Although one might expect that LSGFs will be hard to detect in an incomplete and fragmented plant data set (Jabbari et al., 2004), several examples were obtained by querying the phylogenetic profiles. An overview of some SSGFs and LSGFs for which functional information

Table 3.2.2 Examples of species-specific and lineage-specific gene families

Phylogenetic profile	Taxon[a]	Family ID	Number of homologous proteins	Function	Comments
SSGF	*Arabidopsis thaliana*	10207	102	Ulp1 protease family	
SSGF	*Arabidopsis thaliana*	1607	72	F-box protein family	confirmed by EST/cDNA
SSGF	*Arabidopsis thaliana*	2397	21	cytochrome P-450 aromatase-related	
SSGF	*Chlamydomonas reinhardtii*	12880	10	matrix metalloprotease	homologs found in *Volvox* and vertebrates
SSGF	*Chlamydomonas reinhardtii*	6240	2	hydroxyproline-rich glycoprotein	
SSGF	*Chlamydomonas reinhardtii*	7610	4	sulfur deprivation response regulator	homologs found in Bacteria
SSGF	*Chlamydomonas reinhardtii*	4173	4	nitrite transporter NAR1	homologs found in Bacteria
SSGF	*Chlamydomonas reinhardtii*	287	6	perphorin	homologs found in *Volvox*
SSGF	*Glycine max*	6927	6	nodulin 22	homologs found in *Phaseolus*
SSGF	*Medicago truncatula*	5884	3	nodule-specific glycine-rich protein	
SSGF	*Nicotiana tabacum*	4926	5	putative translation transactivator	
SSGF	*Pinus taeda*	4061	6	nonspecific lipid transfer protein	
SSGF	*Vitis vinifera*	11153	7	putative proline-rich cell wall protein	
SSGF	*Zea mays*	10123	7	basal layer antifungal peptide	
SSGF	*Zea mays*	12935	6	MURA-like protein	
SSGF	*Mesembryanthemum crystallinum*	5565	3	antimicrobial peptide 1 precursor	homologs found in related *Caryophyllales*
LSGF	Gymnosperms (2 species)	7133	3	glycine-rich protein	
LSGF	Asterids (4 species)	9108	22	proteinase inhibitor II protein	
LSGF	Asterids (3 species)	1844	12	gamma-thionin 1 precursor	
LSGF	Asterids (2 species)	5981	10	probable metallocarboxypeptidase inhibitor	
LSGF	Asterids (3 species)	1996	8	cysteine-rich extensin-like protein	
LSGF	Asterids (4 species)	7470	5	hypothetical protein SENU1, senescence up-regulated	
LSGF	Eurosids I (2 *Fabaceae* species)	3404	18	albumin 1	
LSGF	Eurosids I (2 *Fabaceae* species)	2428	2	nodulin 6	
LSGF	monocots (3 species)	4462	20	gamma-gliadin	
LSGF	monocots (3 species)	5153	80	zein-alpha precursor	
LSGF	monocots (5 species)	6364	39	pollen allergen	
LSGF	monocots (4 species)	10095	39	alpha-amylase inhibitor	
LSGF	monocots (5 species)	9761	12	protein synthesis inhibitor	

[a] The number of species covered by the gene family for specific taxa is indicated in parenthesis.

is available is given in Table 3.2.2. The largest SSGF was found in *Arabidopsis* and codes for Ulp1 proteases, a eukaryotic class of cysteine proteases. Examples of genes driving unique taxa-specific biological processes are matrix metalloproteases, lytic enzymes digesting the cell walls of mating-type gametes during mating in *Chlamydomonas reinhardtii* (Kinoshita et al., 1992), specific nodulin genes participating in nodule formation and function in legume plants (Kevei et al., 2002; Mergaert et al., 2003), and zeins, a class of seed storage proteins typically found in panicoid cereals (Shewry and Halford, 2002).

In order to estimate the real number of orphan genes for a particular organism, we compared these proteins with the total data set by using less strict sequence similarity criteria than those used for the construction of the gene families. Still more than 14,000 orphan genes were detected, the largest number being found in rice and the lowest in *Zinnia elegans* (Table 3.2.1). Interestingly, the number of expressed orphan genes is only 6,482, because almost half of all putative orphans are predicted genes of *O. sativa* and *Arabidopsis* lacking proof of expression (no EST- or cDNA supported gene model). *P. patens* seems to be the organism with the highest number of expressed orphan genes (2,053) in the full data set, which can be explained by its unique taxonomic position and current EST/cDNA sequencing status. Indeed, *P. patens* is the only moss representative in the data set and has a high number of ESTs yielding more than 10,000 different moss proteins. Overall, disregarding *P. patens*, the observed correlation between the number of initial EST sequences and the final number of orphan genes for all plant species is linear (r^2=0,83; y=0,0011x + 25.482). Hence, within these plant species, the chance of detecting new orphan genes only increases with one new orphan per ~900 additional ESTs. In this respect, the 131 orphan genes for *C. reinhardtii*, which also lacks closely related species in this data set and has a high number of ESTs (>140,000), seems unexpectedly low. Most probably, the fact that only 26% of all *C. reinhardtii* EST clusters yielded a protein sequence of more than 50 amino acids compared to 79% for *P. patens*, for which overall longer cDNA sequences could be obtained, reduces the number of detectable *Chlamydomonas* orphan genes. The current sequencing and gene annotation of the *Chlamydomonas* genome will probably reveal additional information about the amount of *Chlorophyta*-specific and orphan genes (Grossman et al., 2003).

Gene loss in Arabidopsis and rice

In order to determine specific gene loss events in *Arabidopsis* and rice, we searched the phylogenetic profiles for conserved gene functions present in numerous eudicots and grasses but absent in *Arabidopsis* and rice, respectively. Subsequently, we used less stringent sequence similarity criteria (see Materials and methods) to validate whether a particular gene family indeed was absent in the full proteome of *Arabidopsis* or rice. We identified seven gene families that were present in five or more plant species, including related *Eurosid II* species, but were absent from *Arabidopsis*. A detailed search with protein sequences of related plants for the missing genes against the raw genomic *Arabidopsis* BAC sequences yielded three loci with significant similarity (see Table 3.2.3 and Supplemental Table II). This indicates that these loci may represent active genes missed by the current gene annotation efforts, whereas the absence of the other four gene families could point to gene loss in *Arabidopsis*. An alternative explanation is that these four gene functions do exist in *Arabidopsis* but are located in currently unsequenced chromosomal regions, such as centromeres (Yamada et al., 2003; Nagaki et al., 2004). In rice, 62 gene loss events were detected for gene families with homologs in five or more other species, including other cereals. For more than 70% (45/62) of the missing gene families a homologous rice locus could be identified on the raw BAC sequences. Although this higher number might reflect a similar degree of gene loss in rice than *Arabidopsis*, this observation is most probably biased due to the current incomplete status of the rice sequencing project. The gene families that are currently untraceable in *Arabidopsis* and rice are shown in Table 3.2.3.

Despite the high number of publicly available protein and EST sequences for monocots that are extremely valuable for extrinsic gene prediction approaches (Mathé et al., 2002; Allen et al., 2004), these observations indicate that the current gene annotation in rice still suffers from a number of missed genes. In addition, the high number of unclustered rice genes (~8,600 genes) and putative orphans currently lacking any evidence of expression (~7,000 genes) indicate that further improvement and retraining of gene prediction programs, together with newly developed extrinsic gene prediction methods seems inevitable for fully exploiting the rice genome sequence (Rouzé et al., 1999; Bennetzen et al., 2004). When compiling all results, our data provides strong evidence for the existence of 33,708 rice genes (30,993 genes organized in gene families + 704 expressed orphan genes + 2,011 unclustered genes with EST/cDNA support) when excluding 12,398

Table 3.2.3 Potential gene loss events in *Arabidopsis* and rice

Loss in	Family ID	Species[a]	Family size[b]	Function	Probe[c]	Evolutionary conservation[d]
Arabidopsis thaliana	152	6	11	unknown	29729_1996.1	*Gossypium arboreum, grasses*
Arabidopsis thaliana	7748	8	13	unknown	3635_1091.1	*Gossypium hirsutum, Glycine max, Solanum tuberosum*
Arabidopsis thaliana	10777	11	13	unknown	29729_1339.1	*Gossypium arboreum, Glycine max, Medicago truncatula*
Arabidopsis thaliana	11343	7	8	unknown	29729_2160.1	*Gossypium arboreum, Glycine max, Medicago truncatula*
Oryza sativa	2118	6	7	unknown	4513_12561.1	*Hordeum vulgare, Lycopersicon esculentum*
Oryza sativa	2431	8	18	unknown	4513_2513.1	*Triticum aestivum, Hordeum vulgare*
Oryza sativa	3448	5	5	oxidoreductase, 2OG-Fe(II) oxygenase family	4577_13672.1	*Zea mays, Arabidopsis thaliana, Beta vulgaris*
Oryza sativa	3563	6	6	unknown	132711_1331.1	*Hordeum vulgare, Sorghum propinquum*
Oryza sativa	3705	7	8	unknown	4565_16492.1	*Triticum aestivum, Zea mays*
Oryza sativa	5304	6	6	unknown	4577_22894.1	*Zea mays, Lycopersicon esculentum, Glycine max*
Oryza sativa	5357	6	7	ubiquitin family protein	4513_2676.1	*Hordeum vulgare, Arabidopsis thaliana, Medicago truncatula*
Oryza sativa	6190	7	8	unknown	4577_1043.1	*Zea mays, Arabidopsis thaliana, Glycine max*
Oryza sativa	6248	7	7	unknown	4513_12850.1	*Triticum aestivum, Hordeum vulgare, Zea mays*
Oryza sativa	6840	5	6	brix domain-containing protein	4513_1319.1	*Triticum aestivum, Hordeum vulgare, Zea mays*
Oryza sativa	8373	11	14	quinolinate phosphoribosyltransferase	4558_2586.1	*Triticum aestivum, Zea mays, Sorghum bicolor*
Oryza sativa	9577	5	5	MA3 domain-containing protein	4513_5042.1	*Triticum aestivum, Zea mays, Hordeum vulgare*
Oryza sativa	9601	8	13	unknown	4577_2849.1	*Zea mays, Arabidopsis thaliana, Glycine max*
Oryza sativa	10255	5	6	temperature sensing protein-related	4558_5791.1	*Triticum aestivum, Sorghum bicolor, Arabidopsis thaliana*
Oryza sativa	10829	11	12	unknown	4513_7620.1	*Triticum aestivum, Hordeum vulgare, Zea mays*
Oryza sativa	10975	5	5	unknown	4513_5641.1	*Triticum aestivum, Hordeum vulgare, Zea mays*
Oryza sativa	13242	11	12	unknown	4513_2178.1	*Triticum aestivum, Hordeum vulgare, Zea mays*

[a] The number of species in which the gene family was found.

[b] The total number of proteins assigned to this gene family.

[c] The protein ID of the probe that was used to search against the raw genomic BAC sequences (see text for details).

[d] A subset of taxa where the gene family was found.

proteins resembling transposable elements (see Materials and methods). Note that this is a very conservative estimation, since it has been shown that a considerable amount, up to 37% in *Arabidopsis*, of genes lacking EST/cDNA support do represent active genes (Yamada et al., 2003). When taking into account the large number of unclustered rice proteins that are partially homologous with other plant proteins (6,252 proteins matched other rice or plant proteins with a BLASTP E-value< 1e-05), the estimated number of rice genes increases to 39,960. Whether this set of proteins corresponds to genuine genes or pseudogenes, as observed in other eukaryotic genomes (Mounsey et al., 2002; Torrents et al., 2003), remains to be determined.

A closer look at Arabidopsis and rice

Comparing all conserved gene families between *Arabidopsis* and rice makes it possible to verify whether the larger number of genes in rice, as suggested in the past (Goff et al., 2002; Yu et al., 2002) and partially confirmed here, can be the consequence of gene amplification in specific families. A detailed comparison of all 5,910 gene families containing 18,461 and 22,149 genes in *Arabidopsis* and rice, respectively, is given in Figure 3.2.3. We found that 51% of these gene families have the same copy number in both model plants, whereas 10% of all gene families have a more than two-fold size difference. Interestingly, the best-fit line shows that in general large gene families, containing more than 50 genes, are larger in rice than in *Arabidopsis*, whereas the opposite, slightly counterbalancing pattern is observed for small gene families containing less than 5 genes (Figure 3.2.3). Moreover, 76% of all gene families with a >5 fold size difference are bigger in rice compared to *Arabidopsis*. Examples of gene families that strongly vary in copy number are coding for TIR and non-TIR NBS-LRR disease resistance genes (Zhou et al., 2004), Kelch repeat-containing F-box proteins, BTB/POZ domain-containing proteins, glycosyl hydrolases and F-box family proteins (Figure 3.2.3). Phylogenetic analysis on a subset of gene families with a higher copy number in rice than in *Arabidopsis* indicates that they have expanded significantly in rice, after the divergence of monocots and eudicots from their last common ancestor (Supplemental Figure 2). The expansion of the chalcone synthase (CHS) family in rice, a catalyst in the first steps of flavonoid biosynthesis, might reflect an adaptive strategy in its evolution, because previous analyses have reported the extensive differentiation in gene expression among duplicate copies of CHS genes (Durbin

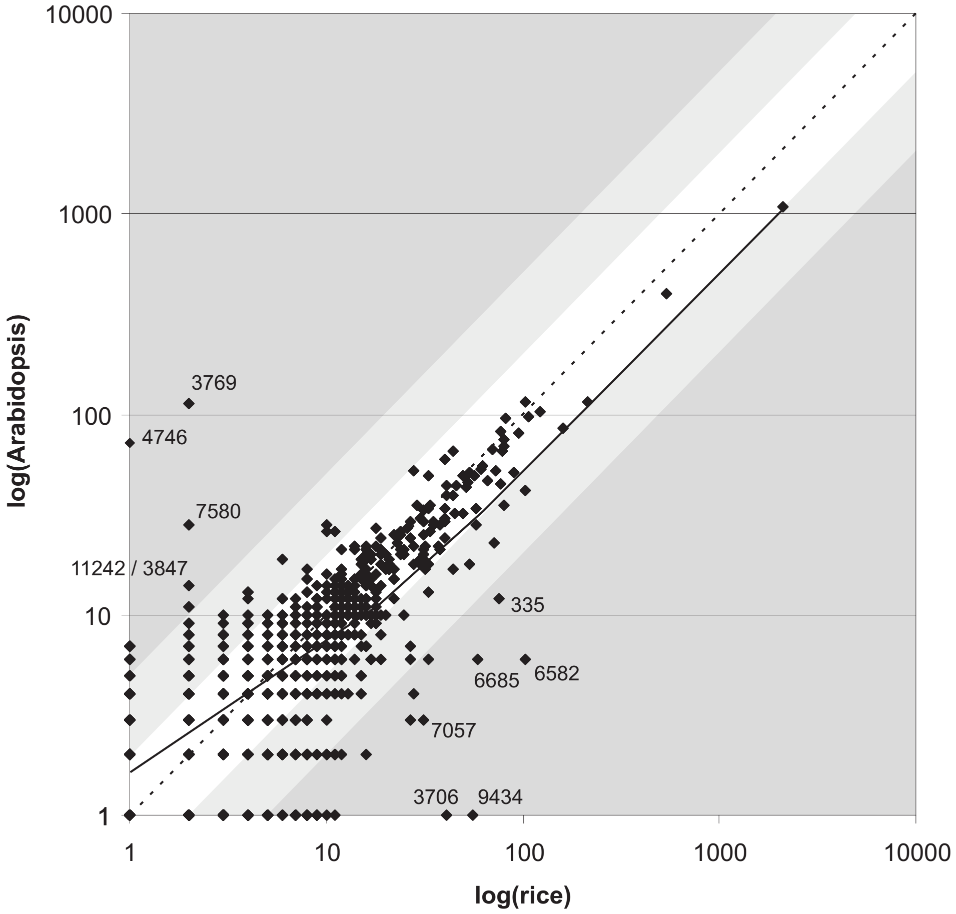


Figure 3.2.3 Size variations of all 5,910 gene families shared by *Arabidopsis* and rice. The position of each dot representing a gene family describes the number of genes identified in *Arabidopsis* and rice (abscissa and ordinate, respectively). The dotted line shows the 1:1 ratio and the black line the best-fit line (y=0.5399x + 1.1002; r^2=0.95). The dark grey and light grey areas indicate a >5 fold and >2 fold size difference, respectively, whereas the white area indicates a <2 fold size difference. The gene families indicated by their family ID are: 335, F-box domain containing protein; 3706, NB-ARC domain / disease resistance protein (CC-NBS-LRR class); 3769, disease resistance protein (TIR-NBS-LRR class); 3847, EXS family protein; 4746, kelch repeat-containing F-box family protein; 5858, chalcone synthase; 6582, putative speckle-type protein / BTB/POZ domain; 6685, unknown; 7057, glycosyl hydrolase family 18; 7580, pentatricopeptide (PPR) repeat-containing protein; 9434, disease resistance protein (NBS-LRR class); 11242, F-box family protein.

et al., 2000). Likewise, the expansion of receptor-like kinases involved in defense and disease control in rice, for which we observe a >1.9 size difference, offers advanced sensing towards diverse extracellular signals (Shiu et al., 2004). Similar patterns of gene family expansion were also observed in gene families that are larger in *Arabidopsis* than in rice, which suggests that the extension of gene families through gene duplication is a more common phenomenon in higher plants than massive reduction through gene loss. The presence of a number of large gene families with similar copy numbers in both plant model systems, such as gene families covering transcription factors, transporter proteins, cytochrome P450s and phosphatases, corresponds with previously reported findings (Goff et al., 2002).

Apart from analyzing the conserved gene families between *Arabidopsis* and rice, we also examined the distribution of gene families containing *Arabidopsis* or rice genes over a wider range of plant species using the high-level phylogenetic profiles (see above). Although 69% of the gene families in grasses are also present in eudicots, 3,006 gene families are unique to the grasses, of which 42% represent grass-specific families found in multiple cereals. These results correspond with previous estimates of putative monocot-specific genes using sugarcane ESTs (Vincentz et al., 2004). In addition, we found that 11% of all families present in the grasses with homologs in eudicots were absent in *Arabidopsis*, which confirms our findings that gene loss in specific lineages or species is common. This number is considerably higher than the 2% of sugarcane sequences that matched homologous non-*Arabidopsis* eudicot sequences and is most probably caused by the higher number of eudicotyledonous species used here, compared to the analysis of Vincentz et al. (2004). The reverse query indicates that also 11% of all families conserved between monocots and eudicots are absent in rice, suggesting that the amount species-specific gene loss in monocots and eudicots is very similar. Although the overall evolutionary distributions of gene families is very similar for *Arabidopsis* and rice (see Figure 3.2.3 and Supplemental Figure 3), the number of rice-specific gene families (and genes) is approximately two to three fold larger for rice than for *Arabidopsis* (see above). Thus, this set of genes, together with the set of orphan genes, also accounts for the larger number of genes currently found in rice than in *Arabidopsis*. Finally, the fact that 914 gene families are detected solely in the fully sequenced genomes of *Arabidopsis* and rice indicates that a fraction of plant gene functions is currently not covered by gene discovery efforts through EST sequencing.

Conclusion

Recent estimates show that approximately 43,000 plant protein sequences are known, which can be classified into approximately 4,053 gene families (Mohseni-Zadeh et al., 2004). Although an enormous amount of ESTs are publicly available for a variety of plant species, these sequences only represent partial information about transcribed genes and lack annotated coding sequence information. Consequently, phylogenetic analysis of plant genes and gene families based on protein information combined with manual addition of homologous plant ESTs is very time consuming and has an overall low success rate. Analysis of the data set described here suggests that approximately 19,300 different gene functions (i.e. 12,854 full-length gene families + 6,482 expressed orphans) exist in the green plant lineage. When all gene families covering partial proteins are discarded, 9,355 gene families are found in the eudicots, of which 89% are found in multiple species, with an additional 2,353 expressed orphan genes. Similarly, 9,967 gene families have been detected in the grasses, of which 82% are found in multiple species, together with 2,084 expressed orphan genes for specific cereals. These numbers suggest that the total number of gene functions in monocots and eudicots is comparable and seems to indicate that a substantial portion of the recently described rice genes are anomalous sequences representing incorrect gene predictions or pseudogenes (Goff et al., 2002; Yu et al., 2002; Jabbari et al., 2004; Bennetzen et al., 2004). Nevertheless, a significant difference in copy number between *Arabidopsis* and rice was uncovered for a subset of large gene families, SSGFs and orphan genes, confirming the larger number of genes in rice compared to *Arabidopsis*. Clearly, the large number of expressed orphans, together with numerous examples of SSGFs and LSGFs, complemented with the observations of gene loss in *Arabidopsis* and rice, illustrate the high plasticity of plant genomes.

Materials and methods

Construction of the data set

The data set consists of two subsets, one including publicly available plant proteins, the other containing EST sequences. The protein data set covers data extracted from EMBL (Kulikova et al., 2004) for 30 different plant species, whereas the EST set contains data of more than 2.8 million ESTs for these plant species. Sequence information for *Arabidopsis thaliana* and *Oryza sativa*, for which a nuclear genome sequence is available, was obtained from TIGR (*Arabdidopsis* release 5 from January 2004; Wortman et al., 2003; rice release April 2004; Yuan et al. 2003). If multiple protein sequences were available for the same locus, the protein of the first gene model was retained. Hundred two *Arabidopsis* proteins with similarity to known plant transposable elements (BLASTP E-value <1e-05 with Swiss-Prot transposable elements) were not retained for further analysis. For rice, all 12,398 proteins with gene description “transposon” or “retrotransposon” were discarded.

EST sequences were transformed into EST clusters (also called unigene or tentative consensus) and a set of singleton ESTs with the EST clustering software developed by TIGR (Pertea et al., 2003). ESTs were clustered and assembled in such a way that paralogous gene sequences should be maintained as such and not merged into a single chimeric EST cluster. To this end, conservative parameters were applied (minimum percent identity 99%, minimum length of overlap 50 bp and a maximum mismatched overhang of 20 bp), which are more stringent than those of TIGR Gene Indices or NCBI Unigenes (for a detailed comparison of these and others EST assembly efforts, see Parkinson et al., 2002). All mRNA sequences of all genes in the protein data set were also incorporated during the EST clustering. As a consequence, all ESTs perfectly matching an existing plant mRNA were remapped to one gene sequence, avoiding inclusion of redundancy in the data set. Similarly, redundant genes in the protein layer were removed, because identical mRNAs were merged into a single gene sequence.

Next, putative ORFs were delineated for all EST clusters. For these EST clusters containing experimentally derived mRNAs, the corresponding coding sequence (CDS) information was retained. For all other sequences, the coding frame and putative CDS were determined with the FrameD software tool (Schiex et al., 2003). When validating the FrameD software against a subset of mRNAs from the protein set from different species, its overall sensitivity was good (85% for mRNAs with an EMBL CDS annotation using the *Arabidopsis* Interpolated Markov Model [IMM]), but rather low for species without a specific IMM, such as *Chlamydomonas* and *Pinus* (2% and 63%, respectively). Because different plants have different codon usages and only a limited number of plant IMMs is available in FrameD, additional IMMs were required to have a good overall ORF detection sensitivity not biased towards particular plant species. Therefore, we first created training sets for each plant species for which no IMM was available, based on the annotated CDS of mRNAs present in the EMBL database. After a careful evaluation of the available FrameD IMMs on the training sequences from different plants and a detailed comparison of the codon usage in the 30 plant species under investigation (data not shown), we constructed five new IMMs (one for

Chlamydomonas, *Physcomitrella*, the *Pinaceae*, the *Asterids* and the *Caryophyllales*). Note that not all new models are species specific because some models were built with sequences from several closely related plant species (see Supplemental Table III). Finally, for each plant species, ORFs were determined on the EST clusters with FrameD using a specific IMM (Table 3.2.1; additional parameters –E for eukaryotic EST analysis and –C for correcting frameshifts; Schiex et al., 2003). Only putative ORFs with a minimal length of 50 codons were retained.

All translated coding sequences of the EST clusters and all sequences from the protein data set were used to construct gene families by applying sequence-based protein clustering (Li et al., 2001). First, an all-against-all sequence comparison was performed using BLASTP (Altschul et al., 1997) and relevant hits were retained (Li et al., 2001). Briefly, this method considers two proteins as being homologous only when they share a substantially conserved region on both molecules with a minimum amount of sequence identity. In this manner, homology based on the partial overlap of single protein domains between two multi-domain proteins, which occasionally leads to significant E-values in BLAST, is not retained. The proportion of identical amino acids in the aligned region between the query and target sequence is recalculated to $I' = I \times \mathrm{Min}(n_1/L_1, n_2/L_2)$, where L_i is the length of sequence i and n_i is the number of amino acids in the aligned region of sequence i. This value I' is then used in the empirical formula for protein clustering proposed by Rost (1999). These additional criteria prevent that partial ORFs derived from two EST clusters, which in reality originated from the same gene, were counted as two distinct family members. Finally, all valid homologous protein pairs (e.g. protein A is homologous to protein B, protein B is homologous to protein C) were subject to a simple-linkage clustering routine to delineate protein gene families (for example, family with proteins A, B and C). In total, more than 39 million blast hits were evaluated and >6.4 million valid homologous protein pairs were used for delineating the gene families. An evaluation of Li’s method (2001) applied on yeast sequences showed that it behaves equally well compared to other automatic protein clustering algorithms (Yang et al., 2003). Although one might argue that by using this method partial proteins will be split from their complete homologous counterparts (see below), we prefer this conservative clustering approach because a less stringent protein clustering would lead to the creation of superfamilies, obscuring every pattern of evolutionary conservation for a specific gene function. Additional information about different protein clustering strategies that were evaluated can be found on our website.

GO functional annotation

Gene Ontology gene associations for *Arabidopsis* proteins were retrieved from TIGR (ftp.tigr.org/pub/data/a_thaliana/ath1/ DATA_RELEASE_SUPPLEMENT/) and remapped to the generic GO Slim classification scheme (ftp.geneontology.org/pub/go/GO_slims/goslim_generic.go) with the Perl script map2slim.pl (available at www.geneontology.org).

Analysis of gene families consisting of partial proteins

Throughout this analysis, we assumed that *Arabidopsis* and rice genes derived from the genome sequencing projects represented full-length proteins. Given the fact that the family delineation algorithm does not create family relationships between homologous proteins that vary extremely in length (i.e. that lack global homology), we believe that gene families including *Arabidopsis* and rice proteins will generally not contain clustered partial proteins. These full-length families represent the majority of all gene families (i.e. 68% of all 14,639 gene families). We obtained 4,341 gene families without *Arabidopsis* and/or rice homologs that might contain partial proteins (designated partial protein families, PPF). For each of the 14,369 gene families, a random gene representative was selected and compared with all other gene representatives. Subsequently, all significant similarities (BLASTP E-value < 1e-15) between genes representing full-length families and PPFs were scored. Finally, we identified these PPFs that were significantly shorter than the homologous full-length family. We found 1,415 and 1,515 PPFs that were more than 50% and more than 30% shorter than the homologous full-length family, respectively. In order to reduce the chance of overpredicting the final number of gene families, we selected the 1,515 gene families that were at least 30% shorter than their full-length counterpart as gene families consisting of partial proteins. These families were discarded when the number of gene families in the different lineages is discussed (Figure 3.2.2). Applying other E-value similarity and length difference cut-offs yielded similar results (data not shown).

Analysis of orphans, SSGFs and LSGFs

All orphan proteins or proteins of gene families specific for one plant species or lineage were compared against the full set of proteins using less stringent criteria (BLASTP E-value <1e-05) compared to the criteria applied by the protein clustering algorithm for delineating gene families (see above). These proteins without non-self BLAST hits (i.e. only hitting themselves) were designated orphans, whereas only those genes uniquely matching proteins of the same species or lineage were retained as species or lineage specific, respectively.

Section III

Conclusions & future perspectives

Concluding remarks

During the last decade of the twentieth century, plant geneticists discovered that plants often use homologous genes for very similar functions and that they exhibit extensive conservation of genome structure, despite major differences in genome size. Apart from different constraints acting on the structural organization of plant genomes, the complete genome sequence of *Arabidopsis thaliana* revealed that plant genomes exhibit a much higher degree of apparent redundancy (65% of all genes belong to multi-gene families), compared to other multicellular eukaryotic organisms (e.g. less than 20% of all human genes belong to multi-gene families; Arabidopsis Genome Initiative, 2000). Moreover, initial comparative analysis of *Arabidopsis* genes illustrated the dynamic nature of plant genomes, with the identification of plant-specific gene functions, genes of bacterial origin whose functions are now integrated in eukaryotic processes, and independent evolution of several families of transcription factors (Arabidopsis Genome Initiative, 2000). As demonstrated in chapter 3.2, the wide-spread occurrence of orphan genes, together with species-specific and lineage-specific gene families indicate that this dynamic state of gene content is a universal feature within the green plant lineage, and contributes to the developmental and genetic diversity seen in different plant species today.

Filling the bioinformatics and evolutionary analysis toolbox…

It is clear that large-scale genome sequencing and advanced comparative sequence analysis offer a powerful combination to study the complex evolutionary forces that shape the gene content and structure of plant genomes. Therefore, dedicated tools are required for the gene annotation of raw genomic DNA sequences, together with various computational methods that enable direct and detailed comparisons at different levels of resolution. As demonstrated in chapter 3.1, the annotation and delineation of a limited number of gene families involved in cell cycle regulation required a substantial amount of manual modifications on the first gene models, predicted in 2001. Now, four years later, the situation has improved significantly, because a large number of EST and full-length cDNA sequences have been generated, which enhance the quality of gene prediction

(Haas et al., 2002; Kikuchi et al., 2003; Castelli et al., 2004). In addition, the application of advanced extrinsic gene prediction strategies, which take full advantage of the large set of protein and EST sequences stored in public sequence repositories, also contributes to the overall quality of current gene annotation efforts in different species (Chapter 2.4; Foissac et al., 2003; Brent and Guigo, 2004). Nevertheless, based on the results of a comparative analysis on predicted genes from *Arabidopsis* and rice, combined with EST data from thirty other plant species, it seems that several deficiencies are still present in the currently applied annotation protocols (Chapter 3.2; Bennetzen et al., 2004).

When comparing large genomic segments in order to identify paralogous segments within a genome or homologous/orthologous segments between different genomes, objective analytical and statistically supported approaches are required (Bennetzen, 2000b; King, 2002). As illustrated in chapter 1.2, the application of a newly developed tool for the detection of genomic homology (ADHoRe), based on Monte-Carlo simulations for assessing the significance of colinearity, confirmed the existence of small but significant microcolinear segments conserved between *Arabidopsis* and rice (Chapter 1.2; Bennetzen, 2000b). Therefore, the availability of accurate tools for assessing genomic homology (Chapter 1.1) not only allows scientists to properly and objectively delineate macro- and microcolinearity between closely and distantly related species (e.g. Chapter 2.4), but also illustrates that well established examples of colinearity (e.g. “the unified grass genome”) determined using loosely defined criteria might require reassessment (Gaut, 2002).

Detection and dating of large-scale duplication events in plants

Detailed intraspecies comparisons provide evidence for recent and ancient duplication events in the nuclear genomes of *Arabidopsis* and rice (Chapters 2.2 and 2.3). Application of the map-based approach for the detection of duplicated segments combined with several dating strategies (see chapter 2.1) offers a successful approach for investigating the duplication past of plants and other eukaryotic organisms (e.g. yeast, Wolfe and Shields, 1997; human, McLysaght et al., 2002; pufferfish, Vandepoele et al., 2004a). As illustrated in chapters 1.3 and 2.3, the use of interspecies comparisons can help to recover block duplications that are seemingly undetectable when applying an intraspecies comparison. Although in this analysis only a limited number of degraded duplicated segments in the *Arabidopsis* genome were discovered through comparison with sequence data of rice, this methodology was later successfully applied for the identification of cryptic cycles of polyploidy in plants (see chapter 2.3; Simillion et al., 2004) and yeast (Langkjaer et al., 2003; Dietrich et al., 2004; Kellis et al., 2004). Consequently, the process of differential gene loss, which turns originally identical duplicated regions into two non-redundant sets of genes, divided over two distinct genome locations, appears not to be plant specific. The continuous development of new approaches (e.g. genomic profiles, Simillion et al., 2004), which try to cope with the destructive nature of gene loss and different types of rearrangements, therefore will provide a better view on the occurrence of ancient duplication events. Nevertheless, since many of these methods still require that both gene content and order is conserved, the detection of shared gene content in the absence of shared order between chromosomes (i.e. synteny), provides a valid alternative for the identification of more degenerated genomic homology, as illustrated in vertebrate genomes (McLysaght et al., 2002; Vandepoele et al., unpublished results).

In order to determine the timing of large-scale duplication events, it seems that dating based on the construction of phylogenetic trees offers several advantages compared to dating based on synonymous substitutions rates. The two main disadvantages of the latter method, which offers only a crude age estimation, are that determining the fraction of synonymous substitutions per silent site (K_s) becomes very difficult when saturation occurs (K_s>1), and that calibrating the rate of synonymous substitutions can be problematic, because rate differences in specific lineages have been observed (Gaut et al., 1996; Li, 1997; Koch et al., 2000; Blanc et al., 2003). Although dating through phylogenetic inference is labour-intensive and not always very efficient, especially when sequence data of related species is absent (see chapter 2.1), it provides an alternative and attractive solution for dating evolutionary events more precisely. In addition, this method is less sensitive to saturation, which allows accurate relative dating of events that occurred tens or even hundreds of million years ago (Chapter 2.3; Bowers et al., 2003; Ermolaeva et al., 2003; Vandepoele et al., 2004a).

Even though the accessibility to a large set of sophisticated computational methods considerably enhances the speed and quality of large-scale evolutionary analysis, it is also clear that the quality of the initial sequence data and associated gene predictions can have serious implications on the final results and conclusions. This is partially illustrated in chapter 2.3, where a detailed analysis on a large number of rice bacterial artificial chromosome sequences was performed. Although the existence of a substantial amount of duplicated segments in the rice genome was confirmed, the obtained results at that time did not provide evidence for a genome duplication event in the evolutionary history of rice. Recent investigations on a nearly full assembly of the rice genome however, indicate that the fraction of the rice genome in duplicated segments is considerably higher than estimated using the incomplete and highly fragmented data set described in chapter 2.3 (Paterson et al., 2004). Consequently, this finding might suggest that rice and other grasses are ancient tetraploids (Paterson et al., 2004). Although the timing of this large-scale duplication event seems firm, detailed dating of all duplicated blocks in the rice genome and comparisons between the paranomes (i.e. the complete set of paralogous genes in a genome) of rice and *Arabidopsis* does not provide conclusive support for this "whole-genome duplication" hypothesis (Paterson et al., 2005). Therefore, more sequence data of related grass species together with better tools are required to fully uncover the contribution and timing of tandem and other types of (large-scale) duplications in these cereal genomes.

The consequences of genome evolution on gene function and regulation

Although comparative mapping and detailed sequence comparisons start to provide information about the patterns of genome organization and the mechanisms altering the structure of nuclear plant genomes, still little is known about the consequences of these processes at the gene level. Some examples of transposon-mediated alterations of gene structure and regulation have been described (Martienssen et al., 1998; Jiang et al., 2004), but the effect of chromosomal modifications such as translocations, (retro-)transpositions, insertions, sequence deletions and duplication events on a gene's function and regulation remains poorly understood. Based on the results described in this thesis and other studies, it is clear that the majority of genes that originated through large-scale duplication events get lost rapidly, most probably during the process of diploidization subsequent to polyploidy. In chapter 2.4, where the evolution of paralogous genes arisen through the youngest genome duplication in *Arabidopsis* was investigated, detailed comparative promoter analysis using homologous legume sequences revealed that half of all retained duplicates have lost a different set of *cis*-regulatory elements, suggesting subfunctionalization. For a limited number of genes, no evidence for divergence, either at the expression level or in the promoter sequence, was found, which seems to confirm the role of redundancy after duplication, apart from gene loss and functional divergence (Pickett and Meeks-Wagner, 1995; Wendel, 2000). Interestingly, for a large fraction of gene duplicates with a high degree of reciprocal promoter divergence between both paralogs, a clearly dissimilar expression pattern was found, which seems to be compatible with an evolutionary model predicting subfunction partitioning after gene duplication (Force et al., 1999). Certainly, future studies, which will have access to larger amounts and more diverse types of experimental data (e.g. spatio and temporal expression data, protein localization data, yeast-two-hybrid data), will make it possible to fully unravel the evolution of gene function after duplication.

Future perspectives

The accessibility to a diverse set of tools makes it possible to study the degree of conservation between closely and more distantly related species, and allows scientists to dig into the evolutionary history of fully sequenced genomes. Nevertheless, more advanced computational analysis tools will be required to fully characterize and date older events, such as cryptic cycles of polyploidy, which are frequently observed in plants (Wendel, 2000; Bowers et al., 2003). As described in chapters 2.2 and 2.3, traces of very ancient duplication events, perhaps more than 200 million years old, have been found in the genomes of both *Arabiopsis* and rice. Nevertheless, whether these duplicated segments are remnants of an ancient large-scale or genome duplication event, predating the split between monocotyledonous and dicotyledonous plants, is not clear. Similarly, it is currently unknown to what extent the older large-scale duplication events identified in *Arabidopsis* are shared with other dicot plants. Moreover, whether these ancient events are responsible for the radiation of angiosperm or dicotyledonous plants, is currently far from proven. Therefore, a comprehensive evolutionary analysis based on a phylogenetic tree-based approach could provide valuable information about conserved features, similarities and differences in the evolutionary history of different plant species (e.g. *Arabidopsis*, *Populus*, *Solanum* and *Medicago*).

A second currently unanswered question is which genetic-evolutionary processes are active following large-scale duplication events. Moreover, the consequences at the gene level of these events accompanying polyploidy are largely unknown. Although it is clear that novel forms of gene expression, altered regulatory interactions and rapid epigenetic changes go together with the acquisition of new phenotypes, increased genetic complexity and the success of polyploids in nature, the mechanisms driving these changes are poorly understood (Wendel, 2000; Osborn et al., 2003). Therefore, detailed sequence analysis of the short and long-term evolutionary changes in gene duplicates, together with experiments monitoring the altered gene expression in several model polyploid systems should enlarge our knowledge about these biological processes. The huge amount of sequence data currently generated for several plant species therefore will be instrumental for studying the changes, either at the regulatory or at the protein level, in new-born and ancient gene duplicates. These future experiments will offer a better understanding of the effects caused by the different evolutionary actors driving gene and genome evolution in plants. Finally, comparing these features between plants and other eukaryotic taxa (e.g. yeasts, animals, protists) will be the next best thing.

Section IV

Summary

Plants come in a wide range of forms and colours, and their genomes exhibit a large degree of variation, even between different species from the same family. Apart from the diversity present in the construction and organization of DNA sequences in different species, molecular and evolutionary processes are continuously shaping nuclear genome structures. Although it has become clear that major genome size differences can be explained by differences in ploidy levels and dissimilar amounts of mobile and tandem repetitive elements, the mechanisms driving gene and genome evolution in higher plants, together with their implications on a gene's function and regulation, are largely unknown. We developed a set of tools for advanced comparative sequence analysis and applied these to study the evolutionary forces that shape the gene content and structure of plant genomes. Apart from the detailed characterization of large-scale duplication events - key players in plant genome evolution -, we also focussed on the consequences of these events on the evolution of individual gene families. Finally, a pilot study was initiated to verify whether duplication indeed is responsible for the acquisition of novel gene functions or altered, more complex, regulatory control.

Application of a newly developed tool for the detection of genomic homology (ADHoRe) revealed evidence for colinearity (conserved content and order of genes) between both closely and distantly related species. Comparative restriction fragment length polymorphism mapping studies identified numerous examples of macrosynteny between related species in the past (Bonierbale et al., 1988; Helentjaris et al., 1988; Chao et al., 1989; Ahn et al., 1993). However, it is clear that the availability of fully sequenced genomes, grouping thousands of genes over different chromosomes, requires objective and statistically supported criteria, implemented in flexible computational tools, for studying genome evolution. Determining intraspecies colinearity in *Arabidopsis* and rice using ADHoRe revealed the presence of a large number of duplicated segments. In *Arabidopsis*, evidence for 3 large-scale duplication events, together with an additional duplication event on chromosome 1, was found. Dating of these duplicated blocks using synonymous substitutions and phylogenetic trees shows that the youngest genome duplication occurred 40-70 MYA and is shared with other *Brassicacea* species. Also in rice, traces of a large-scale duplication event, predating the divergence of the grasses 50-70 MYA, were found. In both plants genomes, remnants of older duplication events were also identified, although the age of these events, together with their significance for angiosperm evolution, is currently unclear.

Apart from investigating gene and genome evolution in different plant species, comparisons grouping data from several taxa can provide additional insights on gene and genome evolution. An interspecies sequence comparison grouping genomic data of *Arabidopsis* and rice identified numerous duplicated blocks that are seemingly undetectable (“ghost duplications”) in both species when applying an intraspecies comparison. Similarly, the analysis of approximately 250,000 genes organized in more than 12,000 gene families over a wide variety of species within the green plant lineage allowed us to identify genes driving the core machinery in plants, together with orphans and species- and lineage specific gene families. Interestingly, the methodology for the identification of ghost duplications was later successfully applied in yeast, providing conclusive evidence for an ancient polyploidy event in yeast (Langkjaer et al., 2003; Dietrich et al., 2004; Kellis et al., 2004).

Although detailed sequence comparisons start to provide information about the patterns of genome organization and the mechanisms altering the structure of nuclear plant genomes, still little is known about the consequences of these processes at the gene level. Based on genomic sequence data from *Medicago* and other legumes, which diverged from *Arabidopsis* before it’s youngest genome duplication, we were able to study the *cis*-regulatory evolution for a small set of gene duplicates. For nearly half of all analyzed gene duplicates, traces of reciprocal promoter divergence were found using phylogenetic footprinting. Through the identification of conserved non-coding sequences, we observed that for a large number of genes the *cis*-regulatroy elements present in the legume outgroup promoter were complementary partitioned over both retained duplicated genes. Interestingly, for a majority of these genes, a high degree of expression divergence was observed when analyzing expression levels over several hundreds of microarray experiments. This confirms that subfunctionalization is an important mechanism responsible for creating genetic novelty and introducing altered transcriptional regulatory control. Nevertheless, more genomic and functional data from a variety of plant species is required to fully unravel the consequences of large-scale duplication events and other actors driving gene and genome evolution in plants.

Nederlandse samenvatting

In de natuur komen planten voor in een brede waaier van vormen en kleuren. Ook in de grootte van hun genomen komt veel variatie voor, zelfs tussen verschillende species van eenzelfde familie. Naast de diversiteit die terug te vinden is in de opbouw en organisatie van het DNA in diverse species, spelen verschillende moleculaire en evolutionaire processen tevens een belangrijke rol bij het vormgeven van nucleaire genomen. Alhoewel het duidelijk is dat grote verschillen in genoom grootte te wijten zijn aan verschillen in *ploidy* niveaus en ongelijke hoeveelheden van mobiele en tandem repetitieve elementen, is tot op heden weinig gekend over de verschillende mechanismen die gen- en genoomevolutie in planten sturen. Tijdens dit doctoraatswerk werden nieuwe methoden ontwikkeld om op een gedetailleerde manier vergelijkende sequentie analyses uit te voeren, en werden deze gebruikt om de evolutionaire processen te bestuderen die de inhoud en structuur van plant genomen modelleren. Naast het bestuderen van grootschalige duplicatie gebeurtenissen – hoofdrolspelers in plant genoom evolutie –, werden ook de gevolgen van deze duplicaties voor individuele genen en genfamilies in detail onderzocht. Tevens werd een pilootstudie uitgevoerd om na te gaan of duplicatie inderdaad verantwoordelijk is voor het ontstaan van nieuwe genfuncties of gewijzigde, complexere vormen van regulatorische controle.

Het gebruik van een recent ontwikkelde methode voor de detectie van genomische homologie (ADHoRe) toonde duidelijk aan dat colineariteit (conservering van geninhoud en volgorde) voorkomt tussen nauw- en ververwante species. Alhoewel vergelijkende mapping experimenten in het verleden alreeds voorbeelden van *macrosynteny* tussen verschillende planten hebben aangetoond (Bonierbale et al., 1988; Helentjaris et al., 1988; Chao et al., 1989; Ahn et al., 1993), wordt het meer en meer duidelijk dat de beschikbaarheid van volledig gesequeneerde genomen, die duizenden genen groeperen verdeeld over meerdere chromosomen, objectieve methoden vereisen om colineariteit te bepalen en genoomevolutie te bestuderen. Tevens is het essentieel dat deze nieuwe methoden statistisch onderbouwd zijn. De detectie van intraspecies colineariteit door middel van ADHoRe toonde duidelijk aan dat zowel in *Arabidopsis* als rijst een groot aantal gedupliceerde genomische segmenten voorkomen. In *Arabidopsis* werden bewijzen gevonden voor 3 grootschalige duplicaties plus een extra duplicatie binnen chromosoom 1. Datering van deze gedupliceerde segmenten door middel van synonieme substituties en boomconstructie toonde aan dat de jongste genoomduplicatie ongeveer 40-70 miljoen jaar geleden gebeurde en gedeeld is met andere planten van de *Brassicaceae*. Ook in rijst werden sporen van een grootschalige duplicatie gebeurtenis teruggevonden, die dateert van voor de divergentie van de grassen, 50-70 miljoen jaar geleden. In beide genomen werden tevens sporen van oudere duplicaties teruggevonden, maar het is tot op heden onduidelijk wanneer deze duplicaties zich hebben voorgedaan en wat hun rol of bijdrage in de evolutie van angiosperme planten is.

Naast het bestuderen van gen- en genoomevolutie in verschillende planten, kunnen vergelijkende studies, waarbij data van verschillende planten wordt gecombineerd, nieuwe inzichten bieden in de evolutie van genomen. Een inter-species vergelijking tussen *Arabidopsis* en rijst, waarbij genomische sequenties van beide planten samen werden onderzocht, maakte het mogelijk om een aantal schijnbare ondetecteerbare gedupliceerde blokken (*ghost duplications*) op te sporen. Analoog maakte een vergelijkende analyse van ongeveer 250,000 genen, gegroepeerd in meer dan 12,000 genfamilies, afkomstig van een brede waaier van planten, het mogelijk om een set van genen aan te duiden die algemeen geconserveerde processen binnen planten sturen, evenals *orphan* genen en species- en *lineage* specifieke genfamilies. Het is interessant om op te merken dat de methode voor het opsporen van *ghost duplications* later ook succesvol op gist genomen werd toegepast, waar ze een ontegensprekelijk bewijs vormen voor een oude genoomduplicatie (Langkjaer et al., 2003; Dietrich et al., 2004; Kellis et al., 2004).

Alhoewel het duidelijk is dat vergelijkende sequentie analyses ons meer kennis verschaffen omtrent de patronen van genoomorganisatie en de mechanismen die de structuur van nucleaire plant genomen beïnvloeden, is tot op heden weinig gekend omtrent de gevolgen van deze processen op het genniveau. Aan de hand van genomische sequenties van *Medicago* en andere *Fabaceae* species, die gedivergeerd zijn van *Arabidopsis* voor diens jongste genoomduplicatie, was het mogelijk op de *cis*-regulatorische evolutie van een set van gedupliceerde genen te bestuderen. Voor ongeveer de helft van alle duplicaten konden we door middel van *phylogenetic footprinting* sporen van reciproque promoter divergentie vaststellen, waarbij een complementaire set van regulatorische elementen verloren is gegaan in vergelijking met de voorouderlijke *Fabaceae* promoter. Interessant hierbij was dat onderzoek van honderden microarray expressie experimenten aantoonde dat de overgrote meerderheid van deze genduplicaten niet langer co-gereguleerd zijn. Dit bevestigt dat functionele divergentie na duplicatie een belangrijk mechanisme is voor het ontstaan van nieuwe genetische interacties en het introduceren van gewijzigde transcriptionele controle. Echter, grotere hoeveelheden genomische en functionele data zijn noodzakelijk om in detail de implicaties van grootschalige genduplicaties en andere evolutionaire processen betrokken bij genoomevolutie in planten te onderzoeken.

Section V

Bibliography

Acarkan A., Rossberg M., Koch M. and Schmidt R. (2000) Comparative genome analysis reveals extensive conservation of genome organisation for Arabidopsis thaliana and Capsella rubella. Plant J **23:** 55-62.

Adams K. L., Cronn R., Percifield R. and Wendel J. F. (2003) Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. Proc Natl Acad Sci U S A **100:** 4649-4654

Ahn S., Anderson J. A., Sorrells M. E. and Tanksley S. D. (1993) Homoeologous relationships of rice, wheat and maize chromosomes. Mol Genet Genomics **241:** 483-490

Albani D., Mariconti L., Ricagno S., Pitto L., Moroni C., Helin K. and Cella R. (2000) DcE2F, a functional plant E2F-like transcriptional activator from Daucus carota. J Biol Chem **275:** 19258-19267

Alexandersson M., Cawley S. and Pachter L. (2003) SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model. Genome Res **13:** 496-502.

Allen J. E., Pertea M. and Salzberg S. L. (2004) Computational gene prediction using multiple sources of evidence. Genome Res **14:** 142-148

Altschul S. F., Gish W., Miller W., Myers E. W. and Lipman D. J. (1990) Basic local alignment search tool. J Mol Biol **215:** 403-410.

Altschul S. F., Madden T. L., Schaffer A. A., Zhang J., Zhang Z., Miller W. and Lipman D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:** 3389-3402.

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature **408:** 796-815.

Avramova Z., Tikhonov A., Chen M. and Bennetzen J. L. (1998) Matrix attachment regions and structural colinearity in the genomes of two grass species. Nucleic Acids Res **26:** 761-767

Bairoch A. and Apweiler R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res **28:** 45-48

Barakat A., Carels N. and Bernardi G. (1997) The distribution of genes in the genomes of Gramineae. Proc Natl Acad Sci U S A **94:** 6857-6861

Bateman A., Birney E., Cerruti L., Durbin R., Etwiller L., Eddy S. R., Griffiths-Jones S., Howe K. L., Marshall M. and Sonnhammer E. L. (2002) The Pfam protein families database. Nucleic Acids Res **30:** 276-280.

Bennetzen J. L. and Freeling M. (1997) The unified grass genome: synergy in synteny. Genome Res **7:** 301-306.

Bennetzen J. L. and Kellog E. (1997) Do plants have a one-way ticket to genomic obesity? Plant Cell **7:** 1509-1514

Bennetzen J. L. (1998) The structure and evolution of angiosperm nuclear genomes. Curr Opin Plant Biol **1:** 103-108

Bennetzen J. L. (2000a) Transposable element contributions to plant gene and genome evolution. Plant Mol Biol **42:** 251-269

Bennetzen J. L. (2000b) Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions. Plant Cell **12:** 1021-1029.

Bennetzen J. L. and Ramakrishna W. (2002) Numerous small rearrangements of gene content, order and orientation differentiate grass genomes. Plant Mol Biol **48:** 821-827

Bennetzen J. L., Coleman C., Liu R., Ma J. and Ramakrishna W. (2004) Consistent over-estimation of gene number in complex plant genomes. Curr Opin Plant Biol **7:** 732-736

Blanc G., Barakat A., Guyot R., Cooke R. and Delseny M. (2000) Extensive duplication and reshuffling in the Arabidopsis genome. Plant Cell **12:** 1093-1101.

Blanc G., Hokamp K. and Wolfe K. H. (2003) A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. Genome Res **13:** 137-144.

Blanc G. and Wolfe K. H. (2004) Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell **16:** 1679-1691

Boeckmann B., Bairoch A., Apweiler R., Blatter M. C., Estreicher A., Gasteiger E., Martin M. J., Michoud K., O'Donovan C., Phan I., Pilbout S. and Schneider M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res **31:** 365-370.

Bohle U. R., Hilger H. H. and Martin W. F. (1996) Island colonization and evolution of the insular woody habit in Echium L. (Boraginaceae). Proc Natl Acad Sci U S A **93:** 11740-11745

Bonierbale M. W., Plaisted R. L. and Tanksley S. D. (1988) RFLP maps based on a common set clones reveal modes of chromosomal evolution in potato and tomato. Genetics **120:** 1095-1103

Boudolf V., Rombauts S., Naudts M., Inze D. and De Veylder L. (2001) Identification of novel cyclin-dependent kinases interacting with the CKS1 protein of Arabidopsis. J Exp Bot **52:** 1381-1382

Bowers J. E., Chapman B. A., Rong J. and Paterson A. H. (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature **422:** 433-438.

Bray N., Dubchak I. and Pachter L. (2003) AVID: A global alignment program. Genome Res **13:** 97-102.

Brendel V. and Kleffe J. (1998) Prediction of locally optimal splice sites in plant pre-mRNA with applications to gene identification in Arabidopsis thaliana genomic DNA. Nucleic Acids Res **26:** 4748-4757

Brent M. R. and Guigo R. (2004) Recent advances in gene structure prediction. Curr Opin Struct Biol **14:** 264-272

Brubaker C. L., Paterson A. H. and Wendel J. F. (1999) Comparative genetic mapping of allotetraploid cotton and its diploid progenitors. Genome **42:** 184-203

Brudno M., Do C. B., Cooper G. M., Kim M. F., Davydov E., Green E. D., Sidow A. and Batzoglou S. (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res **13:** 721-731.

Calabrese P. P., Chakravarty S. and Vision T. J. (2003) Fast identification and statistical evaluation of segmental homologies in comparative maps. Bioinformatics **19:** I74-I80.

Callan H. G. (1972) Replication of DNA in the chromosomes of eukaryotes. Proc R Soc Lond B Biol Sci **181:** 19-41

Cannon S. B. and Young N. D. (2003) OrthoParaMap: distinguishing orthologs from paralogs by integrating comparative genome data and gene phylogenies. BMC Bioinformatics **4:** 35

Carels N., Hatey P., Jabbari K. and Bernardi G. (1998) Compositional properties of homologous coding sequences from plants. J Mol Evol **46:** 45-53

Castelli V., Aury J. M., Jaillon O., Wincker P., Clepet C., Menard M., Cruaud C., Quetier F., Scarpelli C., Schachter V., Temple G., Caboche M., Weissenbach J. and Salanoubat M. (2004) Whole genome sequence comparisons and “full-length” cDNA sequences: a combined approach to evaluate and improve Arabidopsis genome annotation. Genome Res **14:** 406-413

Castillo-Davis C. I., Hartl D. L. and Achaz G. (2004) cis-Regulatory and protein evolution in orthologous and duplicate genes. Genome Res **14:** 1530-1536

Cavalcanti A. O., Ferreira R. O., Gu Z. O. and Li W. H. (2003) Patterns of Gene Duplication in Saccharomyces cerevisiae and Caenorhabditis elegans. J Mol Evol **56:** 28-37.

Cavalier-Smith T. (1985) Selfish DNA and the origin of introns. Nature **315:** 283-284

Chao S., Sharp P. J., Worland A. J., Warham E. J., Koebner R. M. D. and Gale M. D. (1989) RFLP-based genetic maps of wheat homoeologous group 7 chromosomes. Theor. Appl. Genet. **78:** 495-504

Chaw S. M., Chang C. C., Chen H. L. and Li W. H. (2004) Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. J Mol Evol **58:** 424-441

Chen M. and Bennetzen J. L. (1996) Sequence composition and organization in the Sh2/A1-homologous region of rice. Plant Mol Biol **32:** 999-1001.

Chen M., SanMiguel P., de Oliveira A. C., Woo S. S., Zhang H., Wing R. A. and Bennetzen J. L. (1997) Microcolinearity in sh2-homologous regions of the maize, rice, and sorghum genomes. Proc Natl Acad Sci U S A **94:** 3431-3435.

Chen M., SanMiguel P. and Bennetzen J. L. (1998) Sequence organization and conservation in sh2/a1-homologous regions of sorghum and rice. Genetics **148:** 435-443

Clamp M., Andrews D., Barker D., Bevan P., Cameron G., Chen Y., Clark L., Cox T., Cuff J., Curwen V., Down T., Durbin R., Eyras E., Gilbert J., Hammond M., Hubbard T., Kasprzyk A., Keefe D., Lehvaslaiho H., Iyer V., Melsopp C., Mongin E., Pettett R., Potter S., Rust A., Schmidt E., Searle S., Slater G., Smith J., Spooner W., Stabenau A., Stalker J., Stupka E., Ureta-Vidal A., Vastrik I. and Birney E. (2003) Ensembl 2002: accommodating comparative genomics. Nucleic Acids Res **31:** 38-42.

Cliften P. F., Hillier L. W., Fulton L., Graves T., Miner T., Gish W. R., Waterston R. H. and

Johnston M. (2001) Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis. Genome Res **11:** 1175-1186.

Coghlan A. and Wolfe K. H. (2002) Fourfold faster rate of genome rearrangement in nematodes than in Drosophila. Genome Res **12:** 857-867.

Collins J. E., Goward M. E., Cole C. G., Smink L. J., Huckle E. J., Knowles S., Bye J. M., Beare D. M. and Dunham I. (2003) Reevaluating human gene annotation: a second-generation analysis of chromosome 22. Genome Res **13:** 27-36.

Comai L., Tyagi A. P., Winter K., Holmes-Davis R., Reynolds S. H., Stevens Y. and Byers B. (2000) Phenotypic instability and rapid gene silencing in newly formed arabidopsis allotetraploids. Plant Cell **12:** 1551-1568

Conery J. S. and Lynch M. 2001. Nucleotide substitutions and the evolution of duplicate genes. In *In Pacific Symposium on Biocomputing* (eds. R.B. Altman A.K. Dunker L. Hunter K. Lauderdale, and T.E. Klein), pp. 167-178. World Scientific, Singapore.

Davies T. J., Barraclough T. G., Chase M. W., Soltis P. S., Soltis D. E. and Savolainen V. (2004) Darwin's abominable mystery: Insights from a supertree of the angiosperms. Proc Natl Acad Sci U S A **101:** 1904-1909

de la Luna S., Burden M. J., Lee C. W. and La Thangue N. B. (1996) Nuclear accumulation of the E2F heterodimer regulated by subunit composition and alternative splicing of a nuclear localization signal. J Cell Sci **109 (Pt 10):** 2443-2452

De Veylder L., Segers G., Glab N., Casteels P., Van Montagu M. and Inze D. (1997) The Arabidopsis Cks1At protein binds the cyclin-dependent kinases Cdc2aAt and Cdc2bAt. FEBS Lett **412:** 446-452

De Veylder L., de Almeida Engler J., Burssens S., Manevski A., Lescure B., Van Montagu M., Engler G. and Inze D. (1999) A new D-type cyclin of Arabidopsis thaliana expressed during lateral root primordia formation. Planta **208:** 453-462

Delcher A. L., Kasif S., Fleischmann R. D., Peterson J., White O. and Salzberg S. L. (1999) Alignment of whole genomes. Nucleic Acids Res **27:** 2369-2376.

Dermitzakis E. T., Reymond A., Lyle R., Scamuffa N., Ucla C., Deutsch S., Stevenson B. J., Flegel V., Bucher P., Jongeneel C. V. and Antonarakis S. E. (2002) Numerous potentially functional but non-genic conserved sequences on human chromosome 21. Nature **420:** 578-582.

Devos K. M., Atkinson M. D., Chinoy C. N., Harcourt R. L., Koebner R. M. D., Liu C. J., Masojc P., Xie D. X. and Gale M. D. (1993) Chromosomal rearrangements in the rye genome relative to that of wheat. Theor. Appl. Genet. **85:** 673-680

Devos K. M. and Gale M. D. (1997) Comparative genetics in the grasses. Plant Mol Biol **35:** 3-15.

Devos K. M., Beales J., Nagamura Y. and Sasaki T. (1999) Arabidopsis-rice: will colinearity allow gene prediction across the eudicot-monocot divide? Genome Res **9:** 825-829.

Devos D. and Valencia A. (2001) Intrinsic errors in genome annotation. Trends Genet **17:** 429-431.

Devos K. M., Brown J. K. and Bennetzen J. L. (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. Genome Res **12:** 1075-1079

Dieterich C., Cusack B., Wang H., Rateitschak K., Krause A. and Vingron M. (2002) Annotating regulatory DNA based on man-mouse genomic comparison. Bioinformatics **18:** S84-90.

Dietrich F. S., Voegeli S., Brachat S., Lerch A., Gates K., Steiner S., Mohr C., Pohlmann R., Luedi P., Choi S., Wing R. A., Flavier A., Gaffney T. D. and Philippsen P. (2004) The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome. Science **304:** 304-307

Doebley J. and Lukens L. (1998) Transcriptional regulators and the evolution of plant form. Plant Cell **10:** 1075-1082

Dong Q., Schlueter S. D. and Brendel V. (2004) PlantGDB, plant genome database and analysis tools. Nucleic Acids Res **32 Database issue:** D354-359

Doolittle W. F. and Sapienza C. (1980) Selfish genes, the phenotype paradigm and genome evolution. Nature **284:** 601-603

Doring H. P. and Starlinger P. (1986) Molecular genetics of transposable elements in plants. Annu Rev Genet **20:** 175-200

Doyle J. J. and Gaut B. S. (2000) Evolution of genes and taxa: a primer. Plant Mol Biol **42:** 1-23

Durand D. (2003) Vertebrate evolution: doubling and shuffling with a full deck. Trends Genet **19:** 2-5.

Durbin M. L., McCaig B. and Clegg M. T. (2000) Molecular evolution of the chalcone synthase multigene family in the morning glory genome. Plant Mol Biol **42:** 79-92

Duronio R. J., Brook A., Dyson N. and O’Farrell P. H. (1996) E2F-induced S phase requires cyclin E. Genes Dev **10:** 2505-2513

Easteal S. and Collet C. (1994) Consistent variation in amino-acid substitution rate, despite uniformity of mutation rate: protein evolution in mammals is not neutral. Mol Biol Evol **11:** 643-647

Eddy S. R. (1998) Profile hidden Markov models. Bioinformatics **14:** 755-763

Eichler E. E. and Sankoff D. (2003) Structural dynamics of eukaryotic chromosome evolution. Science **301:** 793-797.

Enright A. J., Van Dongen S. and Ouzounis C. A. (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res **30:** 1575-1584.

Ermolaeva M. D., Wu M., Eisen J. A. and Salzberg S. L. (2003) The age of the Arabidopsis thaliana genome duplication. Plant Mol Biol **51:** 859-866

Fedoroff N. (2000) Transposons and genome evolution in plants. Proc Natl Acad Sci U S A **97:** 7002-7007.

Feldman M., Lupton F. G. H. and Miller T. E. 1995. Wheats. In *Evolution of Crop Plants* (eds.

J. Smartt and N.W. Simmonds), pp. 184-192. Longman Scientific, London.

Felsenstein J. (1993) PHYLIP (Phylogeny Inference Package), Version 3.5c. (Seattle, WA: Department of Genetics, University of Washington).

Feng Q., Zhang Y., Hao P., Wang S., Fu G., Huang Y., Li Y., Zhu J., Liu Y., Hu X., Jia P., Zhao Q., Ying K., Yu S., Tang Y., Weng Q., Zhang L., Lu Y., Mu J., Zhang L. S., Yu Z., Fan D., Liu X., Lu T., Li C., Wu Y., Sun T., Lei H., Li T., Hu H., Guan J., Wu M., Zhang R., Zhou B., Chen Z., Chen L., Jin Z., Wang R., Yin H., Cai Z., Ren S., Lv G., Gu W., Zhu G., Tu Y., Jia J., Chen J., Kang H., Chen X., Shao C., Sun Y., Hu Q., Zhang X., Zhang W., Wang L., Ding C., Sheng H., Gu J., Chen S., Ni L., Zhu F., Chen W., Lan L., Lai Y., Cheng Z., Gu M., Jiang J., Li J., Hong G., Xue Y. and Han B. (2002) Sequence and analysis of rice chromosome 4. Nature **420:** 316-320.

Ferris S. D. and Whitt G. S. (1977) Loss of duplicate gene expression after polyploidization. Nature **265:** 258-260

Feuillet C. and Keller B. (1999) High gene density is conserved at syntenic loci of small and large grass genomes. Proc Natl Acad Sci U S A **96:** 8265-8270.

Finnegan E. J., Genger R. K., Peacock W. J. and Dennis E. S. (1998) DNA Methylation in Plants. Annu Rev Plant Physiol Plant Mol Biol **49:** 223-247

Flicek P., Keibler E., Hu P., Korf I. and Brent M. R. (2003) Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map. Genome Res **13:** 46-54.

Foissac S., Bardou P., Moisan A., Cros M. J. and Schiex T. (2003) EUGENE'HOM: A generic similarity-based gene finder using multiple homologous sequences. Nucleic Acids Res **31:** 3742-3745

Force A., Lynch M., Pickett F. B., Amores A., Yan Y. L. and Postlethwait J. (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151:** 1531-1545.

Friedman R. and Hughes A. L. (2001) Pattern and timing of gene duplication in animal genomes. Genome Res **11:** 1842-1847.

Gale M. D. and Devos K. M. (1998a) Comparative genetics in the grasses. Proc Natl Acad Sci U S A **95:** 1971-1974.

Gale M. D. and Devos K. M. (1998b) Plant comparative genetics after 10 years. Science **282:** 656-659.

Gaut B. S., Morton B. R., McCaig B. C. and Clegg M. T. (1996) Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. Proc Natl Acad Sci U S A **93:** 10274-10279.

Gaut B. S. and Doebley J. F. (1997) DNA sequence evidence for the segmental allotetraploid origin of maize. Proc Natl Acad Sci U S A **94:** 6809-6814

Gaut B. S. (2001) Patterns of chromosomal duplication in maize and their implications for comparative maps of the grasses. Genome Res **11:** 55-66.

Gaut B. S. (2002) Evolutionary dynamics of grass genomes. New Phytologist **154:** 15-28

Gibson T. J. and Spring J. (1998) Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. Trends Genet **14:** 46-49; discussion 49-50.

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

Goldman N. and Yang Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol **11:** 725-736

Gopal S., Schroeder M., Pieper U., Sczyrba A., Aytekin-Kurban G., Bekiranov S., Fajardo J. E., Eswar N., Sanchez R., Sali A. and Gaasterland T. (2001) Homology-based annotation yields 1,042 new candidate genes in the Drosophila melanogaster genome. Nat Genet **27:** 337-340

Grandbastien M. A. (1992) Retroelements in higher plants. Trends Genet **8:** 103-108

Grant D., Cregan P. and Shoemaker R. C. (2000) Genome organization in dicots: genome duplication in Arabidopsis and synteny between soybean and Arabidopsis. Proc Natl Acad Sci U S A **97:** 4168-4173.

Griffiths S., Dunford R. P., Coupland G. and Laurie D. A. (2003) The evolution of CONSTANS-like gene families in barley, rice, and Arabidopsis. Plant Physiol **131:** 1855-1867

Grossman A. R., Harris E. E., Hauser C., Lefebvre P. A., Martinez D., Rokhsar D., Shrager J., Silflow C. D., Stern D., Vallon O. and Zhang Z. (2003) Chlamydomonas reinhardtii at the crossroads of genomics. Eukaryot Cell **2:** 1137-1150

Grubbs F. (1969) Procedures for detecting outlying observations in samples. Technometrics **11:** 1-21

Gu Z., Nicolae D., Lu H. H. and Li W. H. (2002) Rapid divergence in expression between duplicate genes inferred from microarray data. Trends Genet **18:** 609-613

Guo H. and Moose S. P. (2003) Conserved noncoding sequences among cultivated cereal genomes identify candidate regulatory sequence elements and patterns of promoter evolution. Plant Cell **15:** 1143-1158

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

Haas B. J., Volfovsky N., Town C. D., Troukhan M., Alexandrov N., Feldmann K. A., Flavell R. B., White O. and Salzberg S. L. (2002) Full-length messenger RNA sequences greatly improve genome annotation. Genome Biol **3:** RESEARCH0029

Haberer G., Hindemitt T., Meyers B. C. and Mayer K. F. (2004) Transcriptional similarities, dissimilarities, and conservation of cis-elements in duplicated genes of Arabidopsis. Plant Physiol

136: 3009-3022

Haldane J. B. S. (1932). The Causes of Evolution. (Ithaca, NY: Cornell Univ. Press.).

Hall T. A. (1999) BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser**:** 95-98

Hampson S., McLysaght A., Gaut B. and Baldi P. (2003) LineUp: statistical detection of chromosomal homology with application to plant comparative genomics. Genome Res **13:** 999-1010.

Harbison C. T., Gordon D. B., Lee T. I., Rinaldi N. J., Macisaac K. D., Danford T. W., Hannett N. M., Tagne J. B., Reynolds D. B., Yoo J., Jennings E. G., Zeitlinger J., Pokholok D. K., Kellis M., Rolfe P. A., Takusagawa K. T., Lander E. S., Gifford D. K., Fraenkel E. and Young R. A. (2004) Transcriptional regulatory code of a eukaryotic genome. Nature **431:** 99-104

Hardison R. C., Oeltjen J. and Miller W. (1997) Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. Genome Res **7:** 959-966.

Hata S. (1991) cDNA cloning of a novel cdc2+/CDC28-related protein kinase from rice. FEBS Lett **279:** 149-152

Heckman D. S., Geiser D. M., Eidell B. R., Stauffer R. L., Kardos N. L. and Hedges S. B. (2001) Molecular evidence for the early colonization of land by fungi and plants. Science **293:** 1129-1133

Helentjaris T., Weber D. and Wright S. (1988) Identification of the genomic locations of duplicate nucleotide sequences in maize by analysis of restriction fragment length polymorphisms. Genetics **118:** 353-363

Hemerly A., Engler Jde A., Bergounioux C., Van Montagu M., Engler G., Inze D. and Ferreira P. (1995) Dominant negative mutants of the Cdc2 kinase uncouple cell division from iterative plant development. Embo J **14:** 3925-3936

Henikoff S. and Henikoff J. G. (1993) Performance evaluation of amino acid substitution matrices. Proteins **17:** 49-61

Herrmann C. H. and Mancini M. A. (2001) The Cdk9 and cyclin T subunits of TAK/P-TEFb localize to splicing factor-rich nuclear speckle regions. J Cell Sci **114:** 1491-1503

Hirano K., Hirano M., Zeng Y., Nishimura J., Hara K., Muta K., Nawata H. and Kanaide H. (2001) Cloning and functional expression of a degradation-resistant novel isoform of p27Kip1. Biochem J **353:** 51-57

Holland P. W. H. (2003) More genes in vertebrates? Journal of Structural and Functional Genomics **3:** 75-84

Hong R. L., Hamaguchi L., Busch M. A. and Weigel D. (2003) Regulatory elements of the floral homeotic gene AGAMOUS identified by phylogenetic footprinting and shadowing. Plant Cell **15:** 1296-1309

Hughes M. K. and Hughes A. L. (1993) Evolution of duplicate genes in a tetraploid animal, Xenopus laevis. Mol Biol Evol **10:** 1360-1369

Hughes A. L. (1999) Adaptive evolution of genes and genomes. Oxford University Press, New York

Huminiecki L. and Wolfe K. H. (2004) Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. Genome Res **14:** 1870-1879

Ilic K., SanMiguel P. J. and Bennetzen J. L. (2003) A complex history of rearrangement in an orthologous region of the maize, sorghum, and rice genomes. Proc Natl Acad Sci U S A **100:** 12265-12270

Inada D. C., Bashir A., Lee C., Thomas B. C., Ko C., Goff S. A. and Freeling M. (2003) Conserved noncoding sequences in the grasses. Genome Res **13:** 2030-2041

Jabbari K., Cruveiller S., Clay O., Le Saux J. and Bernardi G. (2004) The new genes of rice: a closer look. Trends Plant Sci **9:** 281-285

Jensen R. A. (2001) Orthologs and paralogs - we need to get it right. Genome Biol **2:** INTERACTIONS1002.

Jiang J., Nasuda S., Dong F., Scherrer C. W., Woo S. S., Wing R. A., Gill B. S. and Ward D. C. (1996) A conserved repetitive DNA element located in the centromeres of cereal chromosomes. Proc Natl Acad Sci U S A **93:** 14210-14213

Jiang N., Bao Z., Zhang X., Eddy S. R. and Wessler S. R. (2004) Pack-MULE transposable elements mediate gene evolution in plants. Nature **431:** 569-573

Jones D. T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol **292:** 195-202

Joubes J., Chevalier C., Dudits D., Heberle-Bors E., Inze D., Umeda M. and Renaudi J. P. (2000) CDK-related protein kinases in plants. Plant Mol Biol **43:** 607-620

Joubes J., Lemaire-Chamley M., Delmas F., Walter J., Hernould M., Mouras A., Raymond P. and Chevalier C. (2001) A new C-type cyclin-dependent kinase from tomato expressed in dividing tissues does not interact with mitotic and G1 cyclins. Plant Physiol **126:** 1403-1415

Kaplinsky N. J., Braun D. M., Penterman J., Goff S. A. and Freeling M. (2002) Utility and distribution of conserved noncoding sequences in the grasses. Proc Natl Acad Sci U S A **99:** 6147-6151

Karplus K., Barrett C. and Hughey R. (1998) Hidden Markov models for detecting remote protein homologies. Bioinformatics **14:** 846-856

Keller B. and Feuillet C. (2000) Colinearity and gene density in grass genomes. Trends Plant Sci **5:** 246-251.

Kellis M., Patterson N., Endrizzi M., Birren B. and Lander E. S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature **423:** 241-254.

Kellis M., Birren B. W. and Lander E. S. (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature **428:** 617-624

Kellogg E. A. (2001) Evolutionary history of the grasses. Plant Physiol **125:** 1198-1205.

Kent W. J. (2002) BLAT—the BLAST-like alignment tool. Genome Res **12:** 656-664.

Kevei Z., Vinardell J. M., Kiss G. B., Kondorosi A. and Kondorosi E. (2002) Glycine-rich proteins encoded by a nodule-specific gene family are implicated in different stages of symbiotic nodule development in Medicago spp. Mol Plant Microbe Interact **15:** 922-931

Kikuchi S., Satoh K., Nagata T., Kawagashira N., Doi K., Kishimoto N., Yazaki J., Ishikawa M., Yamada H., Ooka H., Hotta I., Kojima K., Namiki T., Ohneda E., Yahagi W., Suzuki K., Li C. J., Ohtsuki K., Shishiki T., Otomo Y., Murakami K., Iida Y., Sugano S., Fujimura T., Suzuki Y., Tsunoda Y., Kurosaki T., Kodama T., Masuda H., Kobayashi M., Xie Q., Lu M., Narikawa R., Sugiyama A., Mizuno K., Yokomizo S., Niikura J., Ikeda R., Ishibiki J., Kawamata M., Yoshimura A., Miura J., Kusumegi T., Oka M., Ryu R., Ueda M., Matsubara K., Kawai J., Carninci P., Adachi J., Aizawa K., Arakawa T., Fukuda S., Hara A., Hashizume W., Hayatsu N., Imotani K., Ishii Y., Itoh M., Kagawa I., Kondo S., Konno H., Miyazaki A., Osato N., Ota Y., Saito R., Sasaki D., Sato K., Shibata K., Shinagawa A., Shiraki T., Yoshino M., Hayashizaki Y. and Yasunishi A. (2003) Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. Science **301:** 376-379

Kilian A., Chen J., Han F., Steffenson B. and Kleinhofs A. (1997) Towards map-based cloning of the barley stem rust resistance genes Rpg1 and rpg4 using rice as an intergenomic cloning vehicle. Plant Mol Biol **35:** 187-195.

King G. J. (2002) Through a genome, darkly: comparative analysis of plant chromosomal DNA. Plant Mol Biol **48:** 5-20

Kinoshita T., Fukuzawa H., Shimada T., Saito T. and Matsuda Y. (1992) Primary structure and expression of a gamete lytic enzyme in Chlamydomonas reinhardtii: similarity of functional domains to matrix metalloproteases. Proc Natl Acad Sci U S A **89:** 4693-4697

Kirik A., Salomon S. and Puchta H. (2000) Species-specific double-strand break repair and genome evolution in plants. Embo J **19:** 5562-5566

Kishimoto N., Higo H., Abe K., Arai S., Saito A. and Higo K. (1994) Identification of the duplicated segments in rice chromosomes 1 and 5 by linkage analysis of cDNA markers of known functions. Theor. Appl. Genet. **88:** 722-726

Koch M. A., Haubold B. and Mitchell-Olds T. (2000) Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in Arabidopsis, Arabis, and related genera (Brassicaceae). Mol Biol Evol **17:** 1483-1498

Koch M., Haubold B. and Mitchell-Olds T. (2001) Molecular systematics of the Brassicaceae: evidence from coding plastidic matK and nuclear Chs sequences. Am J Bot **88:** 534-544

Koonin E. V., Fedorova N. D., Jackson J. D., Jacobs A. R., Krylov D. M., Makarova K. S., Mazumder R., Mekhedov S. L., Nikolskaya A. N., Rao B. S., Rogozin I. B., Smirnov S., Sorokin A. V., Sverdlov A. V., Vasudevan S., Wolf Y. I., Yin J. J. and Natale D. A. (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol **5:** R7

Kowalski S. P., Lan T. H., Feldmann K. A. and Paterson A. H. (1994) Comparative mapping of Arabidopsis thaliana and Brassica oleracea chromosomes reveals islands of conserved organization. Genetics **138:** 499-510

Kriventseva E. V., Biswas M. and Apweiler R. (2001) Clustering and analysis of protein families. Curr Opin Struct Biol **11:** 334-339

Ku H. M., Vision T., Liu J. and Tanksley S. D. (2000) Comparing sequenced segments of the tomato and arabidopsis genomes: large-scale duplication followed by selective gene loss creates a network of synteny. Proc Natl Acad Sci U S A **97:** 9121-9126

Kulikova T., Aldebert P., Althorpe N., Baker W., Bates K., Browne P., van den Broek A., Cochrane G., Duggan K., Eberhardt R., Faruque N., Garcia-Pastor M., Harte N., Kanz C., Leinonen R., Lin Q., Lombard V., Lopez R., Mancuso R., McHale M., Nardone F., Silventoinen V., Stoehr P., Stoesser G., Tuli M. A., Tzouvara K., Vaughan R., Wu D., Zhu W. and Apweiler R. (2004) The EMBL Nucleotide Sequence Database. Nucleic Acids Res **32 Database issue:** D27-30

Kumar A. and Bennetzen J. L. (1999) Plant retrotransposons. Annu Rev Genet **33:** 479-532

Kurtz S. and Schleiermacher C. (1999) REPuter: fast computation of maximal repeats in complete genomes. Bioinformatics **15:** 426-427.

Lagercrantz U. (1998) Comparative mapping between Arabidopsis thaliana and Brassica nigra indicates that Brassica genomes have evolved through extensive genome replication accompanied by chromosome fusions and frequent rearrangements. Genetics **150:** 1217-1228.

Lagos-Quintana M., Rauhut R., Meyer J., Borkhardt A. and Tuschl T. (2003) New microRNAs from mouse and human. Rna **9:** 175-179.

Lai J., Ma J., Swigonova Z., Ramakrishna W., Linton E., Llaca V., Tanyolac B., Park Y. J., Jeong O. Y., Bennetzen J. L. and Messing J. (2004) Gene loss and movement in the maize genome. Genome Res **14:** 1924-1931

Langham R. J., Walsh J., Dunn M., Ko C., Goff S. A. and Freeling M. (2004) Genomic duplication, fractionation and the origin of regulatory novelty. Genetics **166:** 935-945

Langkjaer R. B., Cliften P. F., Johnston M. and Piskur J. (2003) Yeast genome duplication was followed by asynchronous differentiation of duplicated genes. Nature **421:** 848-852.

Le Q. H., Wright S., Yu Z. and Bureau T. (2000) Transposon diversity in Arabidopsis thaliana. Proc Natl Acad Sci U S A **97:** 7376-7381.

Leitch I. J. and Bennett M. D. (1997) Polyploidy in angiosperms. Trends Plant Sci **2:** 470-476

Lessard P., Bouly J. P., Jouannic S., Kreis M. and Thomas M. (1999) Identification of cdc2cAt: a new cyclin-dependent kinase expressed in Arabidopsis thaliana flowers. Biochim Biophys Acta **1445:** 351-358

Levin D. A. (1983) Polyploidy and novelty in flowering plants. American Naturalist **122:** 1-25

Levy S., Hannenhalli S. and Workman C. (2001) Enrichment of regulatory signals in conserved non-coding genomic sequence. Bioinformatics **17:** 871-877.

Levy A. A. and Feldman M. (2002) The impact of polyploidy on grass genome evolution. Plant Physiol **130:** 1587-1593.

Li W. H. (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution.

J Mol Evol **36:** 96-99.

Li W. H. (1997). Molecular Evolution. (Sunderland, MA: Sinauer Associates).

Li W. H., Gu Z., Wang H. and Nekrutenko A. (2001) Evolutionary analyses of the human genome. Nature **409:** 847-849.

Li W. H., Gu Z., Cavalcanti A. R. and Nekrutenko A. (2003) Detection of gene duplications and block duplications in eukaryotic genomes. J Struct Funct Genomics **3:** 27-34.

Lin X., Kaul S., Rounsley S., Shea T. P., Benito M. I., Town C. D., Fujii C. Y., Mason T., Bowman C. L., Barnstead M., Feldblyum T. V., Buell C. R., Ketchum K. A., Lee J., Ronning C. M., Koo H. L., Moffat K. S., Cronin L. A., Shen M., Pai G., Van Aken S., Umayam L., Tallon L. J., Gill J. E., Venter J. C. and et al. (1999) Sequence and analysis of chromosome 2 of the plant Arabidopsis thaliana. Nature **402:** 761-768

Lisch D. R., Freeling M., Langham R. J. and Choy M. Y. (2001) Mutator transposase is widespread in the grasses. Plant Physiol **125:** 1293-1303.

Liu B., Vega J. M. and Feldman M. (1998) Rapid genomic changes in newly synthesized amphiploids of Triticum and Aegilops. II. Changes in low-copy coding DNA sequences. Genome **41:** 535-542

Liu H., Sachidanandam R. and Stein L. (2001) Comparative genomics between rice and Arabidopsis shows scant collinearity in gene order. Genome Res **11:** 2020-2026.

Lockton S. and Gaut B. S. (2005) Plant conserved non-coding sequences and paralogue evolution. Trends Genet**:** in press

Lukashin A. V. and Borodovsky M. (1998) GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res **26:** 1107-1115.

Lynch M. and Conery J. S. (2000) The evolutionary fate and consequences of duplicate genes. Science **290:** 1151-1155.

Lynch M. and Force A. (2000) The probability of duplicate gene preservation by subfunctionalization. Genetics **154:** 459-473.

Maere S., De Bodt S., Raes J., Casneuf T. V., Kuiper M. and Van de Peer Y. (2005) The importance of ancient genome duplications for plant evolution. submitted

Magyar Z., Meszaros T., Miskolczi P., Deak M., Feher A., Brown S., Kondorosi E., Athanasiadis A., Pongor S., Bilgin M., Bako L., Koncz C. and Dudits D. (1997) Cell cycle phase specificity of putative cyclin-dependent kinase variants in synchronized alfalfa cells. Plant Cell **9:** 223-235

Magyar Z., Atanassova A., De Veylder L., Rombauts S. and Inze D. (2000) Characterization of two distinct DP-related genes from Arabidopsis thaliana. FEBS Lett **486:** 79-87

Makova K. D. and Li W. H. (2003) Divergence in the spatial pattern of gene expression between human duplicate genes. Genome Res **13:** 1638-1645

Mariconti L., Pellegrini B., Cantoni R., Stevens R., Bergounioux C., Cella R. and Albani D. (2002) The E2F family of transcription factors from Arabidopsis thaliana. Novel and conserved components of the retinoblastoma/E2F pathway in plants. J Biol Chem **277:** 9911-9919

Martienssen R. (1998) Transposons, DNA methylation and gene control. Trends Genet **14:** 263-264

Martin W. (2003) Gene transfer from organelles to the nucleus: frequent and in big chunks. Proc Natl Acad Sci U S A **100:** 8612-8614

Mathe C., Sagot M. F., Schiex T. and Rouze P. (2002) Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res **30:** 4103-4117

Mayer K., Schuller C., Wambutt R., Murphy G., Volckaert G., Pohl T., Dusterhoft A., Stiekema W., Entian K. D., Terryn N., Harris B., Ansorge W., Brandt P., Grivell L., Rieger M., Weichselgartner M., de Simone V., Obermaier B., Mache R., Muller M., Kreis M., Delseny M., Puigdomenech P., Watson M., McCombie W. R. and et al. (1999) Sequence and analysis of chromosome 4 of the plant Arabidopsis thaliana. Nature **402:** 769-777

Mayer K., Murphy G., Tarchini R., Wambutt R., Volckaert G., Pohl T., Dusterhoft A., Stiekema W., Entian K. D., Terryn N., Lemcke K., Haase D., Hall C. R., van Dodeweerd A. M., Tingey S. V., Mewes H. W., Bevan M. W. and Bancroft I. (2001) Conservation of microstructure between a sequenced region of the genome of rice and multiple segments of the genome of Arabidopsis thaliana. Genome Res **11:** 1167-1174.

McClintock B. (1984) The significance of responses of the genome to challenge. Science **226:** 792-801

McCutcheon J. P. and Eddy S. R. (2003) Computational identification of non-coding RNAs in Saccharomyces cerevisiae by comparative genomics. Nucleic Acids Res **31:** 4119-4128.

McLysaght A., Hokamp K. and Wolfe K. H. (2002) Extensive genomic duplication during early chordate evolution. Nat Genet **31:** 200-204.

Mena M., Ambrose B. A., Meeley R. B., Briggs S. P., Yanofsky M. F. and Schmidt R. J. (1996) Diversification of C-function activity in maize flower development. Science **274:** 1537-1540

Mergaert P., Nikovics K., Kelemen Z., Maunoury N., Vaubert D., Kondorosi A. and Kondorosi E. (2003) A novel family in Medicago truncatula consisting of more than 300 nodule-specific genes coding for small, secreted polypeptides with conserved cysteine motifs. Plant Physiol **132:** 161-173

Messing J. and Llaca V. (1998) Importance of anchor genomes for any plant genome project. Proc Natl Acad Sci U S A **95:** 2017-2020

Meyers B. C., Tingey S. V. and Morgante M. (2001) Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res **11:** 1660-1676

Mohseni-Zadeh S., Louis A., Brezellec P. and Risler J. L. (2004) PHYTOPROT: a database of clusters of plant proteins. Nucleic Acids Res **32 Database issue:** D351-353

Moore G., Devos K. M., Wang Z. and Gale M. D. (1995) Cereal genome evolution. Grasses, line up and form a circle. Curr Biol **5:** 737-739

Mounsey A., Bauer P. and Hope I. A. (2002) Evidence suggesting that a fifth of annotated

Caenorhabditis elegans genes may be pseudogenes. Genome Res **12:** 770-775

Mumberg D., Wick M., Burger C., Haas K., Funk M. and Muller R. (1997) Cyclin ET, a new splice variant of human cyclin E with a unique expression pattern during cell cycle progression and differentiation. Nucleic Acids Res **25:** 2098-2105

Nadeau J. H. and Taylor B. A. (1984) Lengths of chromosomal segments conserved since divergence of man and mouse. Proc Natl Acad Sci U S A **81:** 814-818.

Nagaki K., Cheng Z., Ouyang S., Talbert P. B., Kim M., Jones K. M., Henikoff S., Buell C. R. and Jiang J. (2004) Sequencing of a rice centromere uncovers active genes. Nat Genet **36:** 138-145

Nagamura Y., Inoue T., Antonio B. A., Shimano T., Kajiya H., Shomura A., Lin S. Y., Kuboki Y., Harushima Y., Kurata N., Minobe Y., Yano M. and Sasaki T. (1995) Conservation of duplicated segments between rice chromosome-11 and chromosome-12. Breeding Science **45:** 373-376

Nei M. and Gojobori T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol **3:** 418-426

Ning Z., Cox A. J. and Mullikin J. C. (2001) SSAHA: a fast search method for large DNA databases. Genome Res **11:** 1725-1729.

Nishiyama T., Fujita T., Shin I. T., Seki M., Nishide H., Uchiyama I., Kamiya A., Carninci P., Hayashizaki Y., Shinozaki K., Kohara Y. and Hasebe M. (2003) Comparative genomics of Physcomitrella patens gametophytic transcriptome and Arabidopsis thaliana: implication for land plant evolution. Proc Natl Acad Sci U S A **100:** 8007-8012

Notredame C., Higgins D. G. and Heringa J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol **302:** 205-217

Ohno S. (1970) Evolution by gene duplication. New York. Springer Verlag

Ohno S. (1973) Ancient linkage groups and frozen accidents. Nature **244:** 259-262

O'Neill C. M. and Bancroft I. (2000) Comparative physical mapping of segments of the genome of Brassica oleracea var. alboglabra that are homoeologous to sequenced regions of chromosomes 4 and 5 of Arabidopsis thaliana. Plant J **23:** 233-243.

Orgel L. E. and Crick F. H. (1980) Selfish DNA: the ultimate parasite. Nature **284:** 604-607

Osborn T. C., Pires J. C., Birchler J. A., Auger D. L., Chen Z. J., Lee H. S., Comai L., Madlung A., Doerge R. W., Colot V. and Martienssen R. A. (2003) Understanding mechanisms of novel gene expression in polyploids. Trends Genet **19:** 141-147

Panstruga R., Buschges R., Piffanelli P. and Schulze-Lefert P. (1998) A contiguous 60 kb genomic stretch from barley reveals molecular evidence for gene islands in a monocot genome. Nucleic Acids Res **26:** 1056-1062

Papp B., Pal C. and Hurst L. D. (2003) Evolution of cis-regulatory elements in duplicated genes of yeast. Trends Genet **19:** 417-422

Parkinson J., Guiliano D. B. and Blaxter M. (2002) Making sense of EST sequences by CLOBBing them. BMC Bioinformatics **3:** 31

Paterson A. H., Lan T. H., Reischmann K. P., Chang C., Lin Y. R., Liu S. C., Burow M. D., Kowalski S. P., Katsar C. S., DelMonte T. A., Feldmann K. A., Schertz K. F. and Wendel J. F. (1996) Toward a unified genetic map of higher plants, transcending the monocot- dicot divergence. Nat Genet **14:** 380-382.

Paterson A. H., Bowers J. E., Burow M. D., Draye X., Elsik C. G., Jiang C. X., Katsar C. S., Lan T. H., Lin Y. R., Ming R. and Wright R. J. (2000) Comparative genomics of plant chromosomes. Plant Cell **12:** 1523-1540.

Paterson A. H., Bowers J. E. and Chapman B. A. (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proc Natl Acad Sci U S A **101:** 9903-9908

Paterson A. H., Bowers J. E., Van de Peer Y. and Vandepoele K. (2005) Ancient duplication of cereal genomes. New Phytologist**:** in press

Pavy N., Rombauts S., Dehais P., Mathe C., Ramana D. V., Leroy P. and Rouze P. (1999) Evaluation of gene prediction software using a genomic data set: application to Arabidopsis thaliana sequences. Bioinformatics **15:** 887-899.

Pearson W. R. and Lipman D. J. (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A **85:** 2444-2448.

Pedersen A. G. and Nielsen H. 1997. Neural network prediction of translation initiation sites in eukaryotes: Perspectives for EST and genome analysis. In *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology* (eds. T. Gaasterland P. Karp K. Karplus C. Ouzounis C. Sander, and A. Valencia), pp. 226–233. American Association for Artificial Intelligence Press, Menlo Park, CA.

Pedersen J. S. and Hein J. (2003) Gene finding with a hidden Markov model of genome structure and evolution. Bioinformatics **19:** 219-227.

Pertea G., Huang X., Liang F., Antonescu V., Sultana R., Karamycheva S., Lee Y., White J., Cheung F., Parvizi B., Tsai J. and Quackenbush J. (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics **19:** 651-652

Petrov D. A., Lozovskaya E. R. and Hartl D. L. (1996) High intrinsic rate of DNA loss in Drosophila. Nature **384:** 346-349

Petrov D. (1997) Slow but Steady: Reduction of Genome Size through Biased Mutation. Plant Cell **9:** 1900-1901

Pevzner P. and Tesler G. (2003) Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. Genome Res **13:** 37-45.

Pickett F. B. and Meeks-Wagner D. R. (1995) Seeing double: appreciating genetic redundancy. Plant Cell **7:** 1347-1356

Porceddu A., Stals H., Reichheld J. P., Segers G., De Veylder L., Barroco R. P., Casteels P., Van Montagu M., Inze D. and Mironov V. (2001) A plant-specific cyclin-dependent kinase is involved in the control of G2/M progression in plants. J Biol Chem **276:** 36354-36360

Porter D. C. and Keyomarsi K. (2000) Novel splice variants of cyclin E with altered substrate specificity. Nucleic Acids Res **28:** E101

Prince V. E. and Pickett F. B. (2002) Splitting pairs: the diverging fates of duplicated genes. Nat Rev Genet **3:** 827-837.

Pryer K. M., Schneider H., Zimmer E. A. and Ann Banks J. (2002) Deciding among green plants for whole genome studies. Trends Plant Sci **7:** 550-554

Quackenbush J., Cho J., Lee D., Liang F., Holt I., Karamycheva S., Parvizi B., Pertea G., Sultana R. and White J. (2001) The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. Nucleic Acids Res **29:** 159-164

Raes J. and Van de Peer Y. (1999) ForCon: A software tool for the conversion of sequence alignments. http://www.ebi.ac.uk/embnet.news/vol6_1/ForCon/body_forcon.html.

Raes J., Vandepoele K., Simillion C., Saeys Y. and Van de Peer Y. (2002) Investigating ancient duplication events in the Arabidopsis genome. Genome evolution. Kluwer Academic Publishers, Dortrecht / Boston / London

Raes J., Vandepoele K., Simillion C., Saeys Y. and Van de Peer Y. (2003) Investigating ancient duplication events in the Arabidopsis genome. Journal of structural and functional genomics **3:** 117-129

Ramirez-Parra E., Xie Q., Boniotti M. B. and Gutierrez C. (1999) The cloning of plant E2F, a retinoblastoma-binding protein, reveals unique and conserved features with animal G(1)/S regulators. Nucleic Acids Res **27:** 3527-3533

Ranz J. M., Casals F. and Ruiz A. (2001) How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus Drosophila. Genome Res **11:** 230-239.

Remm M., Storm C. E. and Sonnhammer E. L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol **314:** 1041-1052.

Renaudin J. P., Doonan J. H., Freeman D., Hashimoto J., Hirt H., Inze D., Jacobs T., Kouchi H., Rouze P., Sauter M., Savoure A., Sorrell D. A., Sundaresan V. and Murray J. A. (1996) Plant cyclins: a unified nomenclature for plant A-, B- and D-type cyclins based on sequence organization. Plant Mol Biol **32:** 1003-1018

Rice Chromosome 10 Consortium. (2003) In-depth view of structure, activity, and evolution of rice chromosome 10. Science **300:** 1566-1569

Rice P., Longden I. and Bleasby A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet **16:** 276-277.

Robinson-Rechavi M., Marchand O., Escriva H. and Laudet V. (2001) An ancestral whole-genome duplication may not have been responsible for the abundance of duplicated fish genes. Curr Biol **11:** R458-459

Rogers K. T., Higgins P. D., Milla M. M., Phillips R. S. and Horowitz J. M. (1996) DP-2, a heterodimeric partner of E2F: identification and characterization of DP-2 proteins expressed in vivo. Proc Natl Acad Sci U S A **93:** 7594-7599

Romano L. A. and Wray G. A. (2003) Conservation of Endo16 expression in sea urchins despite evolutionary divergence in both cis and trans-acting components of transcriptional regulation. Development **130:** 4187-4199

Rossberg M., Theres K., Acarkan A., Herrero R., Schmitt T., Schumacher K., Schmitz G. and Schmidt R. (2001) Comparative sequence analysis reveals extensive microcolinearity in the lateral suppressor regions of the tomato, Arabidopsis, and Capsella genomes. Plant Cell **13:** 979-988.

Rost B. (1999) Twilight zone of protein sequence alignments. Protein Eng **12:** 85-94.

Rouzé P., Pavy N. and Rombauts S. (1999) Genome annotation: which tools do we have for it? Curr Opin Plant Biol **2:** 90-95.

Rudd S. (2003) Expressed sequence tags: alternative or complement to whole genome sequences? Trends Plant Sci **8:** 321-329

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

Saitou N. and Nei M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol **4:** 406-425

Sakata K., Nagamura Y., Numa H., Antonio B. A., Nagasaki H., Idonuma A., Watanabe W., Shimizu Y., Horiuchi I., Matsumoto T., Sasaki T. and Higo K. (2002) RiceGAAS: an automated annotation system and database for rice genome sequence. Nucleic Acids Res **30:** 98-102.

Salinas J., Matassi G., Montero L. M. and Bernardi G. (1988) Compositional compartmentalization and compositional patterns in the nuclear genomes of plants. Nucleic Acids Res **16:** 4269-4285

Salse J., Piegu B., Cooke R. and Delseny M. (2002) Synteny between Arabidopsis thaliana and rice at the genome level: a tool to identify conservation in the ongoing rice genome sequencing project. Nucleic Acids Res **30:** 2316-2328.

Sankoff D. (2001) Gene and genome duplication. Curr Opin Genet Dev **11:** 681-684

SanMiguel P., Tikhonov A., Jin Y. K., Motchoulskaia N., Zakharov D., Melake-Berhan A., Springer P. S., Edwards K. J., Lee M., Avramova Z. and Bennetzen J. L. (1996) Nested retrotransposons in the intergenic regions of the maize genome. Science **274:** 765-768

SanMiguel P. and Bennetzen J. L. (1998) Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. Ann Bot **82:** 37-44

Sasaki T. and Burr B. (2000) International Rice Genome Sequencing Project: the effort to completely sequence the rice genome. Curr Opin Plant Biol **3:** 138-141.

Sasaki T., Matsumoto T., Yamamoto K., Sakata K., Baba T., Katayose Y., Wu J., Niimura Y., Cheng Z., Nagamura Y., Antonio B. A., Kanamori H., Hosokawa S., Masukawa M., Arikawa K., Chiden Y., Hayashi M., Okamoto M., Ando T., Aoki H., Arita K., Hamada M., Harada C., Hijishita S., Honda M., Ichikawa Y., Idonuma A., Iijima M., Ikeda M., Ikeno M., Ito S., Ito T., Ito Y., Iwabuchi A., Kamiya K., Karasawa W., Katagiri S., Kikuta A., Kobayashi N., Kono I., Machita K., Maehara T., Mizuno H., Mizubayashi T., Mukai Y., Nagasaki H., Nakashima M., Nakama Y., Nakamichi Y., Nakamura M., Namiki N., Negishi M., Ohta I., Ono N., Saji S.,

Sakai K., Shibata M., Shimokawa T., Shomura A., Song J., Takazaki Y., Terasawa K., Tsuji K., Waki K., Yamagata H., Yamane H., Yoshiki S., Yoshihara R., Yukawa K., Zhong H., Iwama H., Endo T., Ito H., Hahn J. H., Kim H. I., Eun M. Y., Yano M., Jiang J. and Gojobori T. (2002) The genome sequence and structure of rice chromosome 1. Nature **420:** 312-316.

Schiex T., Moisan A. and Rouzé P. 2001. EuGène: An eukaryotic gene finder that combines several sources of evidence. In *Computational Biology: Selected Papers (Lecture Notes in Computer Science)* (eds. O. Gascuel and M.-F. Sagot), pp. 111-125. Springer-Verlag, Berlin.

Schiex T., Gouzy J., Moisan A. and de Oliveira Y. (2003) FrameD: A flexible program for quality check and gene prediction in prokaryotic genomes and noisy matured eukaryotic sequences. Nucleic Acids Res **31:** 3738-3741

Schmidt R. (2000) Synteny: recent advances and future prospects. Curr Opin Plant Biol **3:** 97-102.

Schmidt R. (2002) Plant genome evolution: lessons from comparative genomics at the DNA level. Plant Mol Biol **48:** 21-37

Schwartz S., Zhang Z., Frazer K. A., Smit A., Riemer C., Bouck J., Gibbs R., Hardison R. and Miller W. (2000) PipMaker—a web server for aligning two genomic DNA sequences. Genome Res **10:** 577-586.

Schwartz S., Kent W. J., Smit A., Zhang Z., Baertsch R., Hardison R. C., Haussler D. and Miller W. (2003a) Human-mouse alignments with BLASTZ. Genome Res **13:** 103-107.

Schwartz S., Elnitski L., Li M., Weirauch M., Riemer C., Smit A., Green E. D., Hardison R. C. and Miller W. (2003b) MultiPipMaker and supporting tools: alignments and analysis of multiple genomic DNA sequences. Nucleic Acids Res **31:** 3518-3524.

Sekine M., Ito M., Uemukai K., Maeda Y., Nakagami H. and Shinmyo A. (1999) Isolation and characterization of the E2F-like gene in plants. FEBS Lett **460:** 117-122

Seoighe C., Federspiel N., Jones T., Hansen N., Bivolarovic V., Surzycki R., Tamse R., Komp C., Huizar L., Davis R. W., Scherer S., Tait E., Shaw D. J., Harris D., Murphy L., Oliver K., Taylor K., Rajandream M. A., Barrell B. G. and Wolfe K. H. (2000) Prevalence of small inversions in yeast gene order evolution. Proc Natl Acad Sci U S A **97:** 14433-14437.

Seoighe C. and Gehring C. (2004) Genome duplication led to highly selective expansion of the Arabidopsis thaliana proteome. Trends Genet **20:** 461-464

Shahmuradov I. A., Akbarova Y. Y., Solovyev V. V. and Aliyev J. A. (2003) Abundance of plastid DNA insertions in nuclear genomes of rice and Arabidopsis. Plant Mol Biol **52:** 923-934

Shewry P. R. and Halford N. G. (2002) Cereal seed storage proteins: structures, properties and role in grain utilization. J Exp Bot **53:** 947-958

Shirasu K., Schulman A. H., Lahaye T. and Schulze-Lefert P. (2000) A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. Genome Res **10:** 908-915

Shiu S. H., Karlowski W. M., Pan R., Tzeng Y. H., Mayer K. F. and Li W. H. (2004) Comparative analysis of the receptor-like kinase family in Arabidopsis and rice. Plant Cell **16:** 1220-1234

Shoemaker R. C., Polzin K., Labate J., Specht J., Brummer E. C., Olson T., Young N., Concibido V., Wilcox J., Tamulonis J. P., Kochert G. and Boerma H. R. (1996) Genome duplication in soybean (Glycine subgenus soja). Genetics **144:** 329-338

Sidow A. (1996) Gen(om)e duplications in the evolution of early vertebrates. Curr Opin Genet Dev **6:** 715-722.

Simillion C., Vandepoele K., Van Montagu M. C., Zabeau M. and Van de Peer Y. (2002) The hidden duplication past of Arabidopsis thaliana. Proc Natl Acad Sci U S A **99:** 13627-13632.

Simillion C., Vandepoele K., Saeys Y. and Van de Peer Y. (2004) Building genomic profiles for uncovering segmental homology in the twilight zone. Genome Res **14:** 1095-1106

Skrabanek L. and Wolfe K. H. (1998) Eukaryote genome duplication - where's the evidence? Curr Opin Genet Dev **8:** 694-700

Smith T. F. and Waterman M. S. (1981) Identification of common molecular subsequences. J Mol Biol **147:** 195-197.

Soltis D. E. and Soltis P. S. (1993) Molecular data and the dynamic nature of polyploidy. Crit. Rev. Plant Sci. **12:** 243.273

Soltis D. E. and Soltis P. S. (1995) The dynamic nature of polyploid genomes. Proc Natl Acad Sci U S A **92:** 8089-8091.

Soltis D. E. and Soltis P. S. (2003) The role of phylogenetics in comparative genetics. Plant Physiol **132:** 1790-1800

Song K., Lu P., Tang K. and Osborn T. C. (1995) Rapid genome change in synthetic polyploids of Brassica and its implications for polyploid evolution. Proc Natl Acad Sci U S A **92:** 7719-7723.

Soni R., Carmichael J. P., Shah Z. H. and Murray J. A. (1995) A family of cyclin D homologs from plants differentially controlled by growth regulators and containing the conserved retinoblastoma protein interaction motif. Plant Cell **7:** 85-103

Sonnhammer E. L. and Durbin R. (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene **167:** GC1-10.

Sorrell D. A., Menges M., Healy J. M., Deveaux Y., Amano C., Su Y., Nakagami H., Shinmyo A., Doonan J. H., Sekine M. and Murray J. A. (2001) Cell cycle regulation of cyclin-dependent kinases in tobacco cultivar Bright Yellow-2 cells. Plant Physiol **126:** 1214-1223

Stals H. and Inze D. (2001) When plant cells decide to divide. Trends Plant Sci **6:** 359-364.

Stebbins G. L. (1971). Chromosomal evolution in higher plants. (London: Edward Arnold).

Stefansky W. (1972) Rejecting outliers in factorial designs. Technometrics **14:** 469-479

Stephens S. G. (1951) Possible significance of duplication in evolution. Advances in Genetics **4:** 247-265

Sterck L., Rombauts S., Jansson S., Sterky F., Rouzé P. and Van de Peer Y. (2005) EST data suggest that poplar is an ancient polyploid. New Phytologist**:** in press

Sun Y., Dilkes B. P., Zhang C., Dante R. A., Carneiro N. P., Lowe K. S., Jung R., Gordon-Kamm W. J. and Larkins B. A. (1999) Characterization of maize (Zea mays L.) Wee1 and its activity in developing endosperm. Proc Natl Acad Sci U S A **96:** 4180-4185

Swaminathan K., Yang Y., Grotz N., Campisi L. and Jack T. (2000) An enhancer trap line associated with a D-class cyclin gene in Arabidopsis. Plant Physiol **124:** 1658-1667

Takezaki N., Rzhetsky A. and Nei M. (1995) Phylogenetic test of the molecular clock and linearized trees. Mol Biol Evol **12:** 823-833.

Tarchini R., Biddle P., Wineland R., Tingey S. and Rafalski A. (2000) The complete sequence of 340 kb of DNA around the rice Adh1-adh2 region reveals interrupted colinearity with maize chromosome 4. Plant Cell **12:** 381-391.

Tatusov R. L., Galperin M. Y., Natale D. A. and Koonin E. V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res **28:** 33-36.

Tatusov R. L., Fedorova N. D., Jackson J. D., Jacobs A. R., Kiryutin B., Koonin E. V., Krylov D. M., Mazumder R., Mekhedov S. L., Nikolskaya A. N., Rao B. S., Smirnov S., Sverdlov A. V., Vasudevan S., Wolf Y. I., Yin J. J. and Natale D. A. (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics **4:** 41

Taylor J. S., Van de Peer Y., Braasch I. and Meyer A. (2001a) Comparative genomics provides evidence for an ancient genome duplication event in fish. Philos Trans R Soc Lond B Biol Sci **356:** 1661-1679

Taylor J. S., Van de Peer Y. and Meyer A. (2001b) Revisiting recent challenges to the ancient fish-specific genome duplication hypothesis. Curr Biol **11:** R1005-1008

Taylor J. S., Van de Peer Y. and Meyer A. (2001c) Genome duplication, divergent resolution and speciation. Trends Genet **17:** 299-301

Terryn N., Heijnen L., De Keyser A., Van Asseldonck M., De Clercq R., Verbakel H., Gielen J., Zabeau M., Villarroel R., Jesse T., Neyt P., Hogers R., Van Den Daele H., Ardiles W., Schueller C., Mayer K., Dehais P., Rombauts S., Van Montagu M., Rouze P. and Vos P. (1999) Evidence for an ancient chromosomal duplication in Arabidopsis thaliana by sequencing and analyzing a 400-kb contig at the APETALA2 locus on chromosome 4. FEBS Lett **445:** 237-245

Thacker C., Marra M. A., Jones A., Baillie D. L. and Rose A. M. (1999) Functional genomics in Caenorhabditis elegans: An approach involving comparisons of sequences from related nematodes. Genome Res **9:** 348-359.

Thompson J. D., Higgins D. G. and Gibson T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res **22:** 4673-4680.

Tikhonov A. P., SanMiguel P. J., Nakajima Y., Gorenstein N. M., Bennetzen J. L. and Avramova Z. (1999) Colinearity and its exceptions in orthologous adh regions of maize and sorghum. Proc Natl Acad Sci U S A **96:** 7409-7414.

Timmis J. N., Ayliffe M. A., Huang C. Y. and Martin W. (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet **5:** 123-135

Tolstrup N., Rouze P. and Brunak S. (1997) A branch point consensus from Arabidopsis found by non-circular analysis allows for better prediction of acceptor sites. Nucleic Acids Res **25:** 3159-3163

Torrents D., Suyama M., Zdobnov E. and Bork P. (2003) A genome-wide survey of human pseudogenes. Genome Res **13:** 2559-2567

U N. (1935) Genome analysis in Brassica with special reference to the experimental formation of B. napus and peculiar mode of fertilization. Japanese Journal of Botany **7:** 389-452

Umeda M., Bhalerao R. P., Schell J., Uchimiya H. and Koncz C. (1998) A distinct cyclin-dependent kinase-activating kinase of Arabidopsis thaliana. Proc Natl Acad Sci U S A **95:** 5021-5026

Van de Peer Y. and De Wachter R. (1994) TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. Comput Appl Biosci **10:** 569-570.

Van de Peer Y., Taylor J. S., Braasch I. and Meyer A. (2001) The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. J Mol Evol **53:** 436-446.

Van de Peer Y., Frickey T., Taylor J. and Meyer A. (2002) Dealing with saturation at the amino acid level: a case study based on anciently duplicated zebrafish genes. Gene **295:** 205-211

Van de Peer Y., Taylor J. S. and Meyer A. (2003) Are all fishes ancient polyploids? J Struct Funct Genomics **3:** 65-73

van Dodeweerd A. M., Hall C. R., Bent E. G., Johnson S. J., Bevan M. W. and Bancroft I. (1999) Identification and analysis of homoeologous segments of the genomes of rice and Arabidopsis thaliana. Genome **42:** 887-892.

Vandepoele K., Saeys Y., Simillion C., Raes J. and Van de Peer Y. (2002a) The Automatic Detection of Homologous Regions (ADHoRe) and Its Application to Microcolinearity Between Arabidopsis and Rice. Genome Res **12:** 1792-1801.

Vandepoele K., Simillion C. and Van de Peer Y. (2002b) Detecting the undetectable: uncovering duplicated segments in Arabidopsis by comparison with rice. Trends Genet **18:** 606-608.

Vandepoele K., Raes J., De Veylder L., Rouze P., Rombauts S. and Inze D. (2002c) Genome-wide analysis of core cell cycle genes in Arabidopsis. Plant Cell **14:** 903-916

Vandepoele K., Simillion C. and Van de Peer Y. (2003) Evidence that rice and other cereals are ancient aneuploids. Plant Cell **15:** 2192-2202

Vandepoele K., De Vos W., Taylor J. S., Meyer A. and Van de Peer Y. (2004a) 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:** 1638-1643

Vandepoele K., Simillion C. and Van de Peer Y. (2004b) The quest for genomic homology. Curr Genomics **5:** 299-308

Vandepoele K. and Van de Peer Y. (2005) Exploring the plant transcriptome through phylogenetic profiling. Plant Physiol **137:** 31-42

Vicient C. M., Jaaskelainen M. J., Kalendar R. and Schulman A. H. (2001) Active retrotransposons are a common feature of grass genomes. Plant Physiol **125:** 1283-1292.

Vincentz M., Cara F. A., Okura V. K., da Silva F. R., Pedrosa G. L., Hemerly A. S., Capella A. N., Marins M., Ferreira P. C., Franca S. C., Grivet L., Vettore A. L., Kemper E. L., Burnquist W. L., Targon M. L., Siqueira W. J., Kuramae E. E., Marino C. L., Camargo L. E., Carrer H., Coutinho L. L., Furlan L. R., Lemos M. V., Nunes L. R., Gomes S. L., Santelli R. V., Goldman M. H., Bacci M., Jr., Giglioti E. A., Thiemann O. H., Silva F. H., Van Sluys M. A., Nobrega F. G., Arruda P. and Menck C. F. (2004) Evaluation of monocot and eudicot divergence using the sugarcane transcriptome. Plant Physiol **134:** 951-959

Vision T. J., Brown D. G. and Tanksley S. D. (2000) The origins of genomic duplications in Arabidopsis. Science **290:** 2114-2117.

Wagner A. (1998) The fate of duplicated genes: loss or new function? Bioessays **20:** 785-788

Wagner A. (2002) Selection and gene duplication: a view from the genome. Genome Biol **3**

Wang Y. and Gu X. (2000) Evolutionary patterns of gene families generated in the early stage of vertebrates. J Mol Evol **51:** 88-96

Waterston R. H., Lindblad-Toh K., Birney E., Rogers J., Abril J. F., Agarwal P., Agarwala R., Ainscough R., Alexandersson M., An P., Antonarakis S. E., Attwood J., Baertsch R., Bailey J., Barlow K., Beck S., Berry E., Birren B., Bloom T., Bork P., Botcherby M., Bray N., Brent M. R., Brown D. G., Brown S. D., Bult C., Burton J., Butler J., Campbell R. D., Carninci P., Cawley S., Chiaromonte F., Chinwalla A. T., Church D. M., Clamp M., Clee C., Collins F. S., Cook L. L., Copley R. R., Coulson A., Couronne O., Cuff J., Curwen V., Cutts T., Daly M., David R., Davies J., Delehaunty K. D., Deri J., Dermitzakis E. T., Dewey C., Dickens N. J., Diekhans M., Dodge S., Dubchak I., Dunn D. M., Eddy S. R., Elnitski L., Emes R. D., Eswara P., Eyras E., Felsenfeld A., Fewell G. A., Flicek P., Foley K., Frankel W. N., Fulton L. A., Fulton R. S., Furey T. S., Gage D., Gibbs R. A., Glusman G., Gnerre S., Goldman N., Goodstadt L., Grafham D., Graves T. A., Green E. D., Gregory S., Guigo R., Guyer M., Hardison R. C., Haussler D., Hayashizaki Y., Hillier L. W., Hinrichs A., Hlavina W., Holzer T., Hsu F., Hua A., Hubbard T., Hunt A., Jackson I., Jaffe D. B., Johnson L. S., Jones M., Jones T. A., Joy A., Kamal M., Karlsson E. K., Karolchik D., Kasprzyk A., Kawai J., Keibler E., Kells C., Kent W. J., Kirby A., Kolbe D. L., Korf I., Kucherlapati R. S., Kulbokas E. J., Kulp D., Landers T., Leger J. P., Leonard S., Letunic I., Levine R., Li J., Li M., Lloyd C., Lucas S., Ma B., Maglott D. R., Mardis E. R., Matthews L., Mauceli E., Mayer J. H., McCarthy M., McCombie W. R., McLaren S., McLay K., McPherson J. D., Meldrim J., Meredith B., Mesirov J. P., Miller W., Miner T. L., Mongin E., Montgomery K. T., Morgan M., Mott R., Mullikin J. C., Muzny D. M., Nash W. E., Nelson J. O., Nhan M. N., Nicol R., Ning Z., Nusbaum C., O’Connor M. J., Okazaki Y., Oliver K., Overton-Larty E., Pachter L., Parra G., Pepin K. H., Peterson J., Pevzner P., Plumb R., Pohl C. S., Poliakov A., Ponce T. C., Ponting C. P., Potter S., Quail M., Reymond A., Roe B. A., Roskin K. M., Rubin E. M., Rust A. G., Santos R., Sapojnikov V., Schultz B., Schultz J., Schwartz M. S., Schwartz S., Scott C., Seaman S., Searle S., Sharpe T., Sheridan A., Shownkeen R., Sims S., Singer J. B., Slater G., Smit A., Smith D. R., Spencer B., Stabenau A., Stange-Thomann N., Sugnet C., Suyama M., Tesler G., Thompson J., Torrents D., Trevaskis E., Tromp J., Ucla C., Ureta-Vidal A., Vinson J. P., Von Niederhausern A. C., Wade C. M., Wall M., Weber R. J., Weiss R. B., Wendl M. C., West A. P., Wetterstrand K., Wheeler R., Whelan S., Wierzbowski J., Willey D., Williams S., Wilson R. K., Winter E., Worley K. C., Wyman D., Yang S., Yang S. P., Zdobnov E. M., Zody M. C. and Lander E. S. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature **420:** 520-562.

Wegener S., Hampe W., Herrmann D. and Schaller H. C. (2000) Alternative splicing in the regulatory region of the human phosphatases CDC25A and CDC25C. Eur J Cell Biol **79:** 810-815

Wendel J. F. (2000) Genome evolution in polyploids. Plant Mol Biol **42:** 225-249

Wendel J. F., Cronn R. C., Alvarez I., Liu B., Small R. L. and Senchina D. S. (2002) Intron size and genome size in plants. Mol Biol Evol **19:** 2346-2352

Wessler S. R., Bureau T. E. and White S. E. (1995) LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. Curr Opin Genet Dev **5:** 814-821

White O., Soderlund C., Shanmugan P. and Fields C. (1992) Information contents and dinucleotide compositions of plant intron sequences vary with evolutionary origin. Plant Mol Biol **19:** 1057-1064

White S. E., Habera L. F. and Wessler S. R. (1994) Retrotransposons in the flanking regions of normal plant genes: a role for copia-like elements in the evolution of gene structure and expression. Proc Natl Acad Sci U S A **91:** 11792-11796

Whitelaw C. A., Barbazuk W. B., Pertea G., Chan A. P., Cheung F., Lee Y., Zheng L., van Heeringen S., Karamycheva S., Bennetzen J. L., SanMiguel P., Lakey N., Bedell J., Yuan Y., Budiman M. A., Resnick A., Van Aken S., Utterback T., Riedmuller S., Williams M., Feldblyum T., Schubert K., Beachy R., Fraser C. M. and Quackenbush J. (2003) Enrichment of gene-coding sequences in maize by genome filtration. Science **302:** 2118-2120

Wikström N., Savolainen V. and Chase M. W. (2001) Evolution of the angiosperms: calibrating the family tree. Proc R Soc Lond B Biol Sci **268:** 2211-2220.

Wittkopp P. J., Haerum B. K. and Clark A. G. (2004) Evolutionary changes in cis and trans gene regulation. Nature **430:** 85-88

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

Wolfe K. H. (2001) Yesterday’s polyploids and the mystery of diploidization. Nat Rev Genet **2:** 333-341.

Wong S., Butler G. and Wolfe K. H. (2002) Gene order evolution and paleopolyploidy in hemiascomycete yeasts. Proc Natl Acad Sci U S A **99:** 9272-9277.

Wortman J. R., Haas B. J., Hannick L. I., Smith R. K., Jr., Maiti R., Ronning C. M., Chan A. P., Yu C., Ayele M., Whitelaw C. A., White O. R. and Town C. D. (2003) Annotation of the Arabidopsis genome. Plant Physiol **132:** 461-468.

Wray G. A. (2003) Transcriptional regulation and the evolution of development. Int J Dev Biol **47:** 675-684

Yamada K., Lim J., Dale J. M., Chen H., Shinn P., Palm C. J., Southwick A. M., Wu H. C., Kim C., Nguyen M., Pham P., Cheuk R., Karlin-Newmann G., Liu S. X., Lam B., Sakano H., Wu T., Yu G., Miranda M., Quach H. L., Tripp M., Chang C. H., Lee J. M., Toriumi M., Chan M. M., Tang C. C., Onodera C. S., Deng J. M., Akiyama K., Ansari Y., Arakawa T., Banh J., Banno F., Bowser L., Brooks S., Carninci P., Chao Q., Choy N., Enju A., Goldsmith A. D., Gurjal M., Hansen N. F., Hayashizaki Y., Johnson-Hopson C., Hsuan V. W., Iida K., Karnes M., Khan S., Koesema E., Ishida J., Jiang P. X., Jones T., Kawai J., Kamiya A., Meyers C.,

Nakajima M., Narusaka M., Seki M., Sakurai T., Satou M., Tamse R., Vaysberg M., Wallender E. K., Wong C., Yamamura Y., Yuan S., Shinozaki K., Davis R. W., Theologis A. and Ecker J. R. (2003) Empirical analysis of transcriptional activity in the Arabidopsis genome. Science **302:** 842-846

Yamaguchi M., Umeda M. and Uchimiya H. (1998) A rice homolog of Cdk7/MO15 phosphorylates both cyclin-dependent protein kinases and the carboxy-terminal domain of RNA polymerase II. Plant J **16:** 613-619

Yan H. H., Mudge J., Kim D. J., Larsen D., Shoemaker R. C., Cook D. R. and Young N. D. (2003) Estimates of conserved microsynteny among the genomes of Glycine max, Medicago truncatula and Arabidopsis thaliana. Theor Appl Genet **106:** 1256-1265

Yang Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci **13:** 555-556

Yang Y. W., Lai K. N., Tai P. Y. and Li W. H. (1999) Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between Brassica and other angiosperm lineages. J Mol Evol **48:** 597-604.

Yang Z. and Nielsen R. (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol **17:** 32-43

Yang J., Lusk R. and Li W. H. (2003) Organismal complexity, protein complexity, and gene duplicability. Proc Natl Acad Sci U S A **100:** 15661-15665

Yoder J. A., Walsh C. P. and Bestor T. H. (1997) Cytosine methylation and the ecology of intragenomic parasites. Trends Genet **13:** 335-340

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

Yuan Q., Ouyang S., Liu J., Suh B., Cheung F., Sultana R., Lee D., Quackenbush J. and Buell C. R. (2003) The TIGR rice genome annotation resource: annotating the rice genome and creating resources for plant biologists. Nucleic Acids Res **31:** 229-233

Z**eng L. W., Comeron J. M., Chen B. and Kreitman M.** (1998) The molecular clock revisited: the rate of synonymous vs. replacement change in Drosophila. Genetica **102-103:** 369-382

Zhang K., Letham D. S. and John P. C. (1996) Cytokinin controls the cell cycle at mitosis by stimulating the tyrosine dephosphorylation and activation of p34cdc2-like H1 histone kinase. Planta **200:** 2-12

Zhang L., Vision T. J. and Gaut B. S. (2002) Patterns of nucleotide substitution among simultaneously duplicated gene pairs in Arabidopsis thaliana. Mol Biol Evol **19:** 1464-1473

Zhang Z., Gu J. and Gu X. (2004) How much expression divergence after yeast gene duplication could be explained by regulatory motif evolution? Trends Genet **20:** 403-407

Zheng N., Fraenkel E., Pabo C. O. and Pavletich N. P. (1999) Structural basis of DNA recognition by the heterodimeric cell cycle transcription factor E2F-DP. Genes Dev **13:** 666-674

Zhou T., Wang Y., Chen J. Q., Araki H., Jing Z., Jiang K., Shen J. and Tian D. (2004) Genome-wide identification of NBS genes in japonica rice reveals significant expansion of divergent non-TIR NBS-LRR genes. Mol Genet Genomics **271:** 402-415

Zhu H., Kim D. J., Baek J. M., Choi H. K., Ellis L. C., Kuester H., McCombie W. R., Peng H. M. and Cook D. R. (2003) Syntenic relationships between Medicago truncatula and Arabidopsis reveal extensive divergence of genome organization. Plant Physiol **131:** 1018-1026

Zimmermann P., Hirsch-Hoffmann M., Hennig L. and Gruissem W. (2004) GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. Plant Physiol **136:** 2621-2632

Ziolkowski P. A., Blanc G. and Sadowski J. (2003) Structural divergence of chromosomal segments that arose from successive duplication events in the Arabidopsis genome. Nucleic Acids Res **31:** 1339-1350

Zohary D. and Feldman M. (1962) Hybridization between amphidiploids and the evolution of polyploids in the wheat (Aegilops-Triticum) group. Evolution **16:** 44-61

Section VI

Appendix

CURRICULUM VITAE

Personalia

Vandepoele Klaas
Born on May 21st 1978
Oostende, Belgium

Educational background

1990-1996
KTA Koekelare, Latijn-Wiskunde
KA Veurne, Wetenschappen-Wiskunde

1996-2000
Ghent University, Bachelor in Chemistry
Ghent University, Master in Biotechnology

Dissertation: Characterization of *Arabidopsis* sequences for the construction of prediction-free gene structure database

2000- present
Ghent University, Ph. D. student, Bioinformatics and Evolutionary Genomics division, Department of Molecular Genetics (Plant Systems Biology – VIB)
Specialization grant IWT (Flemish government institution)

Dissertation: Mode and tempo of gene and genome evolution in plants

Scientific activity

Oral presentations

“Evidence that rice and other cereals are ancient aneuploids” - Belgian Bioinformatics Conference 2003 (May 13th 2003)

“Major events in the genome evolution of vertebrates: Paranome age and size differ considerably between ray-finned fishes and land vertebrates” – Belgian Bioinformatics Conference 2004 (April 23rd 2004)

Poster presentations

“In silico construction of the physicial map of cell cycle genes in the Arabidopsis genome” – Belgian Bioinformatics Conference 2001 (April 6th 2001)

“Detecting microcolinearity between Arabidopsis and Rice” - 6th Gatersleben Research Conference: Plant Genetic Resources in the Genomic Era: Genetic Diversity, Genome Evolution and New Applications (March 7-11 2002, Germany)

“Building networks of colinearity to study genome evolution” – CSHL conference on Comparative Plant Genomics (December 12-15 2002, Cold Spring Harbor, NY USA)

“A closer look into large-scale genome duplication in plants” - Keystone symposium on Comparative Genomics of Plants (March 4-9 2004, Taos, NM USA)

Publications

Breyne P., Dreesen R., Vandepoele K., De Veylder L., Van Breusegem F., Callewaert L., Rombauts S., Raes J., Cannoot B., Engler G., Inze D. and Zabeau M. (2002) Transcriptome analysis during cell division in plants. Proc. Natl. Acad. Sci. U S A **99**: 14825-14830

Simillion C., Vandepoele K., Van Montagu M. C., Zabeau M. and Van de Peer Y. (2002) The hidden duplication past of Arabidopsis thaliana. Proc. Natl. Acad. Sci. U S A **99**: 13627-13632

Vandepoele K., Raes J., De Veylder L., Rouze P., Rombauts S. and Inze D. (2002) Genome-wide analysis of core cell cycle genes in Arabidopsis. Plant Cell **14**: 903-916

Vandepoele K., Saeys Y., Simillion C., Raes J. and Van de Peer Y. (2002) The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between Arabidopsis and rice. Genome Res. **12**: 1792-1801

Vandepoele K., Simillion C. and Van de Peer Y. (2002) Detecting the undetectable: uncovering duplicated segments in Arabidopsis by comparison with rice. Trends Genet. **18**: 606-608

Breyne P., Dreesen R., Cannoot B., Rombaut D., Vandepoele K., Rombauts S., Vanderhaeghen R., Inze D. and Zabeau M. (2003) Quantitative cDNA-AFLP analysis for genome-wide expression studies. Mol. Genet. Genomics **269**: 173-179

Raes J., Vandepoele K., Simillion C., Saeys Y. and Van de Peer Y. (2003) Investigating ancient duplication events in the Arabidopsis genome. J Struct. Funct. Genomics **3**: 117-129

Vandepoele K., Simillion C. and Van de Peer Y. (2003) Evidence that rice and other cereals are ancient aneuploids. Plant Cell **15**: 2192-2202

Gevers D., Vandepoele K., Simillon C. and Van de Peer Y. (2004) Gene duplication and biased functional retention of paralogs in bacterial genomes. Trends Microbiol. **12**: 148-154

Landrieu I., da Costa M., De Veylder L., Dewitte F., Vandepoele K., Hassan S., Wieruszeski J. M., Faure J. D., Van Montagu M., Inze D. and Lippens G. (2004) A small CDC25 dual-specificity tyrosine-phosphatase isoform in Arabidopsis thaliana. Proc. Natl. Acad. Sci. U S A **101**:13380-13385

Simillion C., Vandepoele K., Saeys Y. and Van de Peer Y. (2004) Building genomic profiles for uncovering segmental homology in the twilight zone. Genome Res. **14**: 1095-1106

Simillion C., Vandepoele K. and Van de Peer Y. (2004) Recent developments in computational approaches for uncovering genomic homology. Bioessays **26**: 1225-1235

Vandepoele K., De Vos W., Taylor J. S., Meyer A. and Van de Peer Y. (2004) 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**: 1638-1643

Vandepoele K., Simillion C. and Van de Peer Y. (2004) The quest for genomic homology. Current Genomics **5**: 299-308

Vercammen D., van de Cotte B., De Jaeger G., Eeckhout D., Casteels P., Vandepoele K., Vandenberghe I., Van Beeumen J., Inze D. and Van Breusegem F. (2004) Type II metacaspases Atmc4 and Atmc9 of Arabidopsis thaliana cleave substrates after arginine and lysine. J Biol. Chem. **279**: 45329-45336

Vandepoele K. and Van de Peer Y. (2005) Exploring the plant transcriptome through phylogenetic profiling. Plant Physiology **137**: 31-42

Paterson, A. H., Bowers, J. E., Van de Peer, Y and Vandepoele, K. (2005) Ancient duplication of cereal genomes. New Phytologist **165**: 658-61