Journal of Structural and Functional Genomics **3:** 117–129, 2003.

Investigating ancient duplication events in the *Arabidopsis* genome

Jeroen Raes[†], Klaas Vandepoele[†], Cedric Simillion, Yvan Saeys & Yves Van de Peer*
Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology, Ghent University, K.L. Ledeganckstraat 35, B−9000 Gent, Belgium;
**Author for correspondence: E-mail yves.vandepeer@gengenp.rug.ac.be*
[†] *The two authors contributed equally to this work.*

Received 26.03.2002; accepted in final form 29.08.2002

Key words: large-scale gene duplications, plant genome evolution, polyploidy, synonymous substitution rate

Abstract

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.

Abbreviations: dA, fraction of amino acid substitutions; Kn, number of nonsynonymous substitutions per nonsynonymous site; Ks, number of synonymous substitutions per synonymous site; MYA, million years ago

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.*, this issue). 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 180×10^6 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 200×10^6 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 170×10^6 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 (200×10^6 years old), 10 from class E (170×10^6 years old) and 10 from class D (140×10^6 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 result 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 $< 1^{e-50}$ 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.*, 2002), 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 1.

Dating based on Ks

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 Ks by the method of Li (1993). We calculated Ks 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 $T = Ks/2\lambda$, 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 Ks value (average of the estimates obtained by the three

Table 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]	Anchors	dA	Age class	Age in MY	Anchors[b]	Ks[c]	Ks[d]	Ks[e]	Mean age[f]	StdDev
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] Ks values calculated according to Li (1993).
[d] Ks values calculated according to Nei and Gojobori (1986).
[e] Ks values calculated according to Yang and Nielsen (2000).
[f] Mean age of the block was derived from the mean Ks, excluding outliers (see Materials and Methods).

methods) for each block was derived for each duplicated pair. These values were then used to calculate the mean Ks 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 ClustalW (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

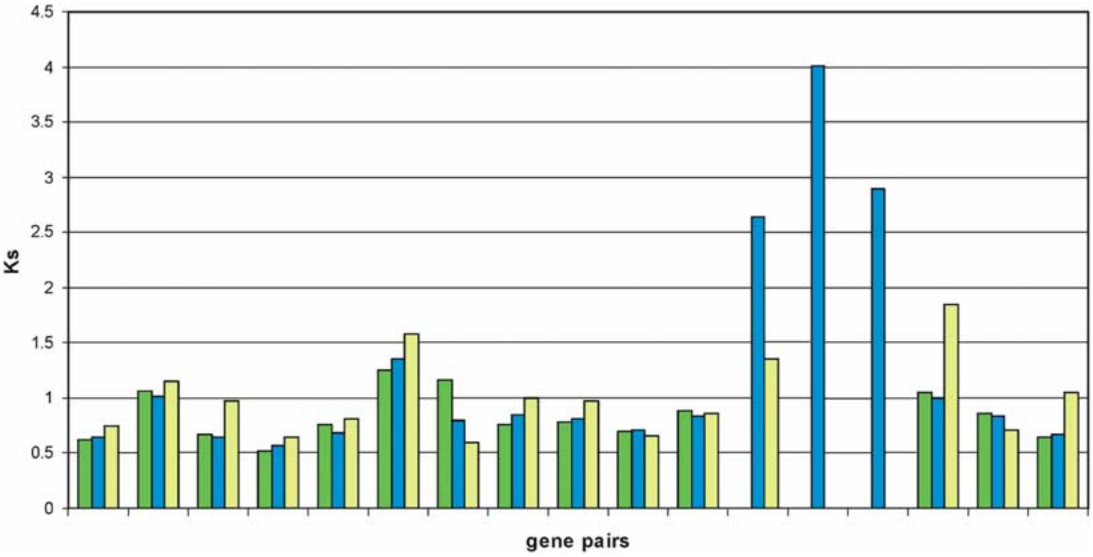

Figure 1. Distribution of Ks values for duplicated genes as found in block 37, and calculated with the methods of Li (green bars), Nei and Gojobori (blue bars) and Yang and Nielsen (yellow bars).

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 Ks

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 (Ks) 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 (Ks $>$ 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 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 Ks 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 Ks values between 0.5 and 1 synonymous substitutions per synonymous site, and this for the three methods used (Fig. 1). For three pairs of genes within the duplicated block, the situation is less clear (Fig. 1). No results were obtained with the method of Li (1993), probably because the duplicated gene sequences are too divergent to calculate a Ks value using this method, whereas the two other methods gave extremely high or no Ks values. One possible explanation is a higher synonymous mutation rate specific for these genes, because fluctuations in Ks have been reported before (Li, 1997; Zeng *et al*., 1997). 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 alter-

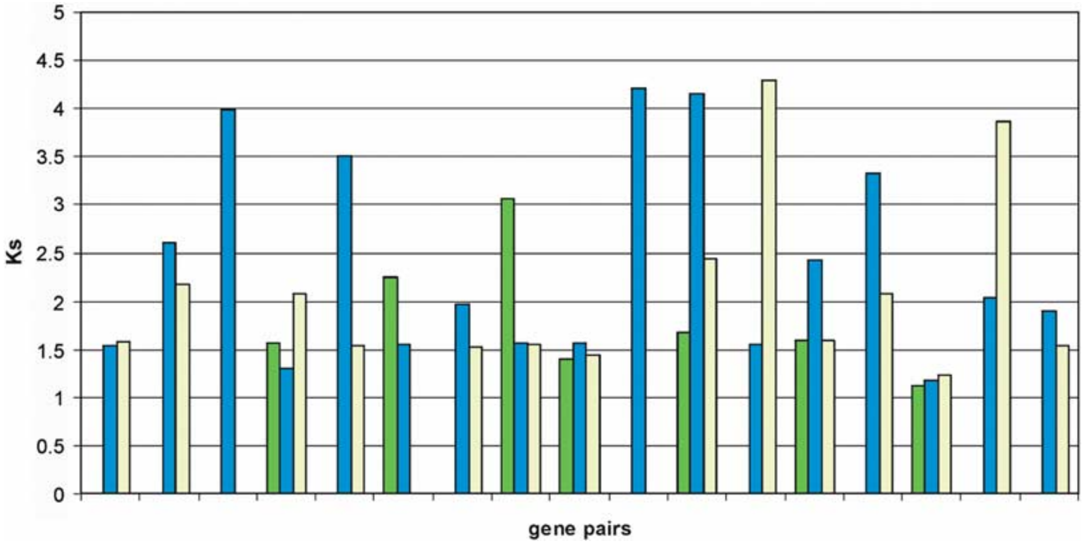

Figure 2. Distribution of Ks values for duplicated genes found in block 59, and calculated with the methods of Li (green bars), Nei and Gojobori (blue bars) and Yang and Nielsen (yellow bars).

nate 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).

However, most blocks of age class F had significantly higher Ks 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 1) and Ks 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; Fig. 2). The latter is probably due to saturation of synonymous substitutions, by which larger errors in Ks estimation are introduced, causing values of Ks > 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 Ks 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 Ks values within one block again gave similar results as above: most pairs had Ks 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

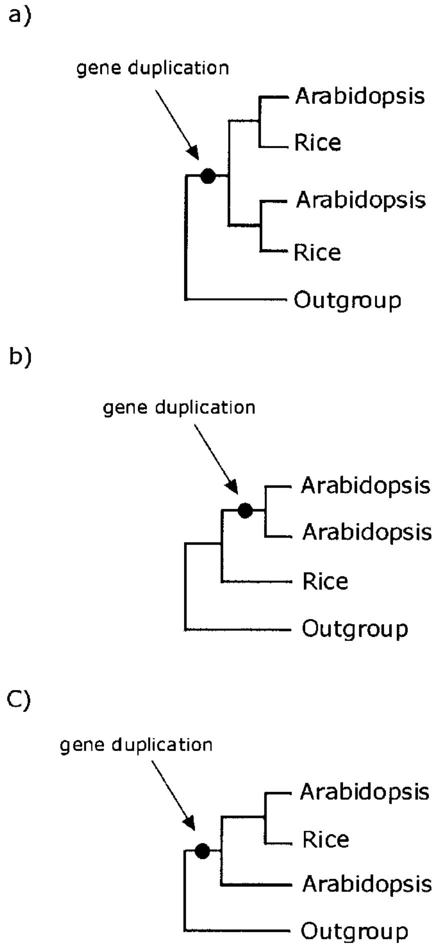

Figure 3. a) Expected tree topology for genes formed by a gene/genome duplication event prior to the split of monocots and dicots. b) Expected tree topology for genes formed by a gene/genome duplication event that occurred after the split of monocots and dicots and specific to *Arabidopsis*. c) Even if only one of the paralogues is known, due to gene loss or absence in the databases, the gene duplication can be inferred.

evolving blocks, or underestimating the age of blocks containing more slowly evolving proteins. The use of synonymous mutation rates is probably favorable 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 (Fig. 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 (Fig. 3b). Even if certain sequences are still missing from the databases (because of gene loss or nondetection), conclusions can be drawn. For example, the tree topology presented in Figure 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. 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 (Fig. 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.

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

Table 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 Fig. 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 Fig. 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.

apparently saturated for synonymous substitutions. However, all Ks values calculated for pairs in this block were above 2.2 synonymous substitutions per synonymous site (see Table 1), suggesting that this block is genuinely old. When we investigated the topology of the GATA family (Fig. 5), we observed a topology similar to that described in Figure 3c: although there is only one monocot sequence, this topology could be only explained if the duplication 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 given. Figure 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 Ks-based dating, we determined that this block had duplicated approximately 70 MYA, with approximately 80% of the Ks 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

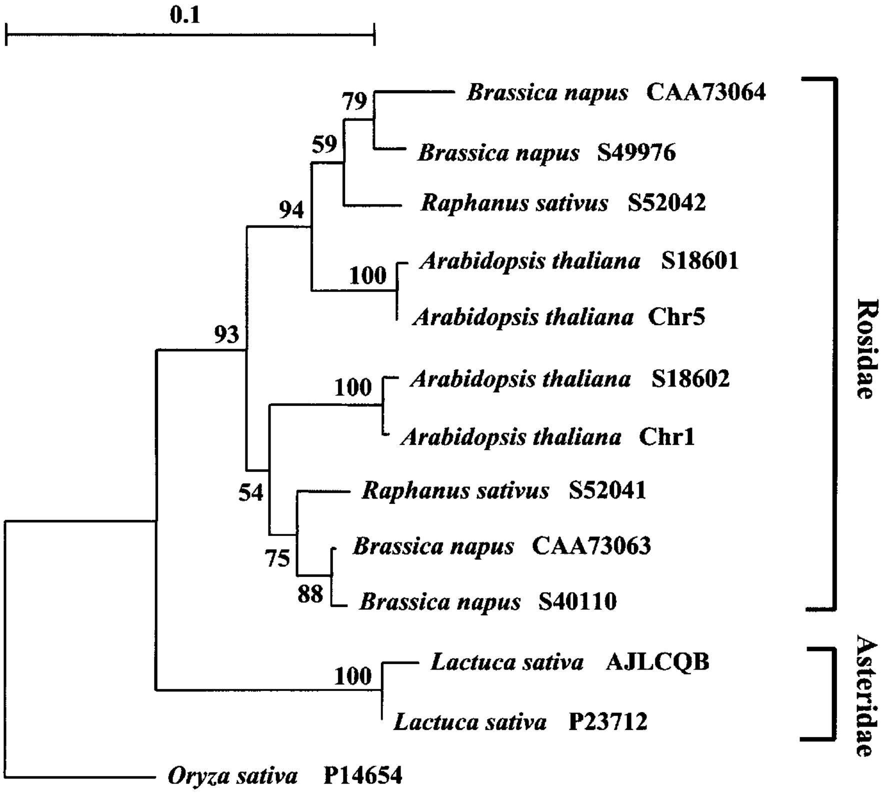

Figure 4. Neighbor-joining tree of the glutamine synthase family, inferred from Poisson-corrected evolutionary distances. Sequences that belong to the analyzed duplicated blocks are indicated with their chromosome number. Bootstrap values (above 50%) are shown in percentages at the internodes. Scale = evolutionary distance in substitutions per amino acid.

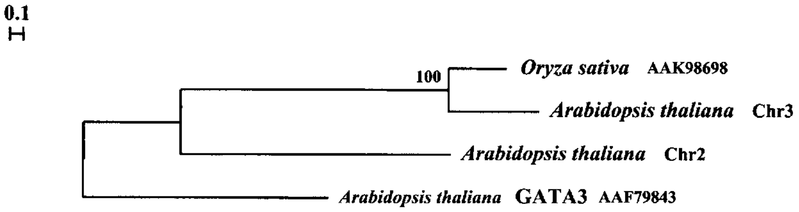
0.1
100
Oryza sativa AAK98698
Arabidopsis thaliana Chr3
Arabidopsis thaliana Chr2
Arabidopsis thaliana GATA3 AAF79843

Figure 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.

46 MYA (with $\lambda = K/2T = 0.206$ 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 dis-

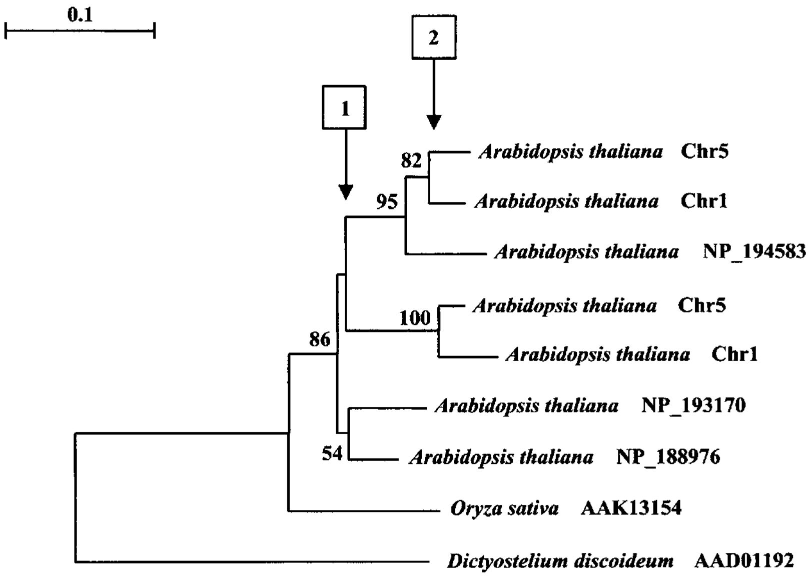
0.1
1
2
82
95
86
100
54
Arabidopsis thaliana Chr5
Arabidopsis thaliana Chr1
Arabidopsis thaliana NP_194583
Arabidopsis thaliana Chr5
Arabidopsis thaliana Chr1
Arabidopsis thaliana NP_193170
Arabidopsis thaliana NP_188976
Oryza sativa AAK13154
Dictyostelium discoideum AAD01192

Figure 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.

tances 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 Ks < 1. A Ks > 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 1), it is hard to prove that they are the result of a single duplication event.

The question of 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 monocotyldicotyl split, as suggested by the GATA transcription factor topology (Fig. 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.

Acknowledgments

The authors would like to thank Eric Bonnet, Sven Degroeve, and John S. Taylor for helpful discussions and Martine De Cock for help with the manuscript. K.V. and C.S. are indebted to the Vlaams Instituut voor de Bevordering van het Wetenschappelijk-Technologisch Onderzoek in de Industrie for a predoctoral fellowship. Y.V.d.P. was a Research Fellow of the Fund for Scientific Research (Flanders).

Note added in proof

Since acceptance of this paper, novel tools to identify heavily degenerated block duplications allowed us to find evidence for the recent genome duplication described in this study. The occurrence of two additional, but probably no more, ancient genome duplicatons in *Arabidopsis* was also demonstrated [Simillian, C., Vandepoele, K., Van Montagu, M.C.E., Zabeau, M., and Van de Peer, Y. (2002). The hidden-duplication past of *Arabidopsis thaliana. Proc. Natl. Acad. Sci. USA* **99**, 13627–13632].

References

Altschul, S.F., Madden, T.L., Schäffer, 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.

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

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.

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

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.

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

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. USA*, **97,** 4168–4173.

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

Haldane, J.B.S. (1933) The part played by recurrent mutation in evolution. *Am. Nat.*, **67,** 5–19.

Hall, T.A. (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.*, **41,** 95–98.

Holland, P. (1992) Homeobox genes in vertebrate evolution. *BioEssays*, **14,** 267–273.

Hughes, A.L. (1999) Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. *J. Mol. Evol.*, **48,** 565–576.

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

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.

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. USA*, **97,** 9121–9126.

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

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*, Sinauer Associates, Sunderland, MA.

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., Adams, M.D., Carrera, A.J., Creasy, T.H., Goodman, H.M., Somerville, C.R., Copenhaver, G.P., Preuss, D., Nierman, W.C., White, O., Eisen, J.A., Salzberg, S.L., Fraser, C.M., and Venter, J.C.

(1999) Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature*, **402,** 761–768.

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.

Mayer, K., Schüller, C., Wambutt, R., Murphy, G., Volckaert, G., Pohl, T., Düsterhöft, 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., Müller, M., Kreis, M., Delseny, M., Puigdomenech, P., Watson, M., Schmidtheini, T., Reichert, B., Portatelle, D., Perez-Alonso,, M., Boutry, M., Bancroft, I., Vos, P., Hoheisel, J., Zimmermann, W., Wedler, H., Ridley, P., Langham, S.-A., McCullagh, B., Bilham, L., Robben, J., Van der Schueren, J., Grymonprez, B., Chuang, Y.-J., Vandenbussche, F., Braeken, M., Weltjens, I., Voet, M., Bastiaens, I., Aert, R., Defoor, E., Weitzenegger, T., Bothe, G., Ramsperger, U., Hilbert, H., Braun, M., Holzer, E., Brandt, A., Peters, S., van Staveren, M., Dirkse, W., Mooijman, P., Klein Lankhorst, R., Rose, M., Hauf, J., Kötter, P., Berneiser, S., Hempel, S., Feldpausch, M., Lamberth, S., Van den Daele, H., De Keyser, A., Buysschaert, C., Gielen, J., Villarroel, R., De Clercq, R., Van Montagu, M., Rogers, J., Cronin, A., Quail, M., Bray-Allen, S., Clark, L., Foggett, J., Hall, S., Kay, M., Lennard, N., McLay, K., Mayes, R., Pettett, A., Rajandream, M.-A., Lyne, M., Benes, V., Rechmann, S., Borkova, D., Blöcker, H., Scharfe, M., Grimm, M., Löhnert, T.-H., Dose, S., de Haan, M., Maarse, A., Schäfer, M., Müller-Auer, S., Gabel, C., Fuchs, M., Fartmann, B., Granderath, K., Dauner, D., Herzl, A., Neumann, S., Argiriou, A., Vitale, D., Liguori, R., Piravandi, E., Massenet, O., Quigley, F., Clabauld, G., Mündlein, A., Felber, R., Schnabl, S., Hiller, R., Schmidt, W., Lecharny, A., Aubourg, S., Chefdor, F., Cooke, R., Berger, C., Montfort, A., Casacuberta, E., Gibbons, T., Weber, N., Vandenbol, M., Bargues, M., Terol, J., Torres, A., Perez-Perez, A., Purnelle, B., Bent, E., Johnson, S., Tacon, D., Jesse, T., Heijnen, L., Schwarz, S., Scholler, P., Heber, S., Francs, P., Bielke, C., Frishman, D., Haase, D., Lemcke, K., Mewes, H.W., Stocker, S., Zaccaria, P., Bevan, M., Wilson, R.K., de la Bastide, M., Habermann, K., Parnell, L., Dedhia, N., Gnoj, L., Schutz, K., Huang, E., Spiegel, L., Sehkon, M., Murray, J., Sheet, P., Cordes, M., Abu-Threideh, J., Stoneking, T., Kalicki, J., Graves, T., Harmon, G., Edwards, J., Latreille, P., Courtney, L., Cloud, J., Abbott, A., Scott, K., Johnson, D., Minx, P., Bentley, D., Fulton, B., Miller, N., Greco, T., Kemp, K., Kramer, J., Fulton, L., Mardis, E., Dante, M., Pepin, K., Hillier, L., Nelson, J., Spieth, J., Ryan, E., Andrews, S., Geisel, C., Layman, D., Du, H., Ali, J., Berghoff, A., Jones, K., Drone, K., Cotton, M., Joshu, C., Antonoiu, B., Zidanic, M., Strong, C., Sun, H., Lamar, B., Yordan, C., Ma, P., Zhong, J., Preston, R., Vil, D., Shekher, M., Matero, A., Shah, R., Swaby, I'K., O'Shaughnessy, A., Rodriguez, M., Hoffman, J., Till, S., Granat, S., Shohdy, N., Hasegawa, A., Hameed, A., Lodhi, M., Johnson, A., Chen, E., Marra, M., Martienssen, R., and McCombie, W.R. (1999) Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. *Nature*, **402,** 769–777.

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

Ohno, S. (1970) *Evolution by Gene Duplication*, Springer-Verlag, Berlin.

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. *Nature Genet*., **14,** 380–382.

Raes, J., and Van de Peer, Y. (1999) ForCon: a software tool for the conversion of sequence alignments. *EMBnet.news*, 6 (http://www.ebi.ac.uk/embnet.news/vol6_1).

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-R459.

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

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

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

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

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. *Phil. Trans. R. Soc. Lond*., **B 356,** 1661–1679.

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

Taylor, J.S., Van de Peer, Y., and Meyer, A. (2001c) Revisiting recent challenges to the ancient fish-specific genome duplication hypothesis. *Curr. Biol*., **11,** R1005-R1007.

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., Déhais, P., Rombauts, S., Van Montagu, M., Rouzé, P., and Vos, P. (1999) Evidence for an ancient chromosomal duplication in *Arabidopsis thaliana* by sequencing and analyzing a 400-kb contig of the *APETALA2* locus on chromosome 4. *FEBS Lett*., **445,** 237–245.

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.

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.S., 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.

Vandepoele, K., Saeys, Y., Simillion, C., Raes, J., and Van de Peer, Y. (2002) A new tool for the Automatic Detection of Homologous Regons (ADHoRe) and its application to microcolinearity between *Arabidopsis* and rice. *Genome Res.*, in press.

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

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

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

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

Wolfe, K.H. (2001) Yesterday's polyploids and the mystery of diploidization. *Nature Rev. Genet.*, **2,** 333–341.

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. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13,** 555–556.

Yang, Z., and Nielsen, R. (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.*, **17,** 32–43.

Zeng, 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*, **103,** 369–382.