

The hidden duplication past of *Arabidopsis thaliana*

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

Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology, Ghent University, K.L. Ledeganckstraat 35, B-9000 Ghent, Belgium

Contributed by Marc C. E. Van Montagu, August 28, 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.

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 (1). After analysis of bacterial artificial chromosome sequences, representing $\approx 80\%$ of the genome, almost 60% was found to contain duplicated genes and regions (2), 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 (3), and by Lynch and Conery (4), 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 (5) and between *Arabidopsis* and tomato (6) 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 (3) did not reveal such segments. Vision *et al.* (7) 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 (7). However, the dating methods applied in their study have been criticized (8). A recent reanalysis of the duplicated blocks ascribed to different age classes, conducted by Raes *et al.* (9), indeed revealed that many of the ancient blocks described by Vision *et al.* (7) 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 (10–12), the most likely fate of a gene duplicate is nonfunctionalization and, consequently, gene loss (4). 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 (13–14). 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 (6, 13).

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 (15) 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 (16) against a set of known retrotransposable elements retrieved from SWISS-PROT (17). 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 $\geq 30\%$.

Detection of Block (Nonhidden) 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 (ref. 14 and www.psb.rug.ac.be/), we shall only briefly describe it here. The ADHoRe

Abbreviation: Mya, million years ago.

*To whom correspondence should be addressed. E-mail: yvdp@gengenp.rug.ac.be.

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 ($m \times n$) 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 (nonzero 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 (14). 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 nonhidden 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 (18). 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 (19). We also calculated K_s by using the dating methods of Nei and Gojobori (20) and Yang and Nielsen (21). The latter two methods are implemented in the YN00 program of the PAML phylogenetic analysis package (22). 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 (μ_{K_s}) and standard deviation (σ_{K_s}) for each block duplication, excluding outliers by using the Grubbs test with a 99% confidence interval (23, 24). 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 (4).

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

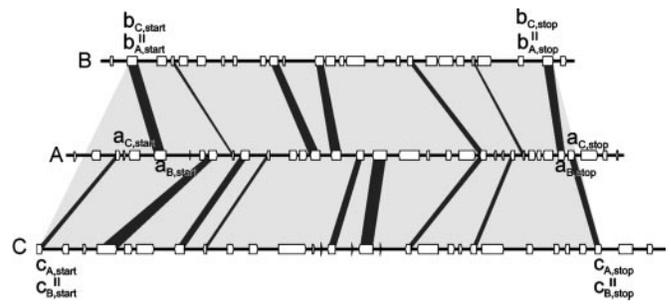


Fig. 1. Determination of the borders of a hypothetical hidden duplication. Gene coordinates increase from left to right. See *Materials and Methods* for details.

rejected by using a t test with a 99% confidence interval. When duplicated blocks can be grouped, the mean K_s (μ_{K_s}) and standard deviation (σ_{K_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 (cutoffs 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 nonhidden duplications (Fig. 1). If we consider three nonoverlapping chromosomal segments A, B, and C, for which it was shown that segments A and B form a nonhidden duplication, and segments A and C form an obvious nonhidden 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 Fig. 1). Note that $(a_{B,start}, b_{A,start})$ and $(a_{B,stop}, b_{A,stop})$ are consequently the coordinates of the outermost anchor points of the observed duplication. Let $(a_{C,start}, a_{C,stop})$ and $(c_{A,start}, c_{A,stop})$ denote the same for segments A and C. The positions $(b_{C,start}, b_{C,stop})$ and $(c_{B,start}, c_{B,stop})$ for the hidden duplication between segments B and C are then determined by considering the start positions of the nonhidden

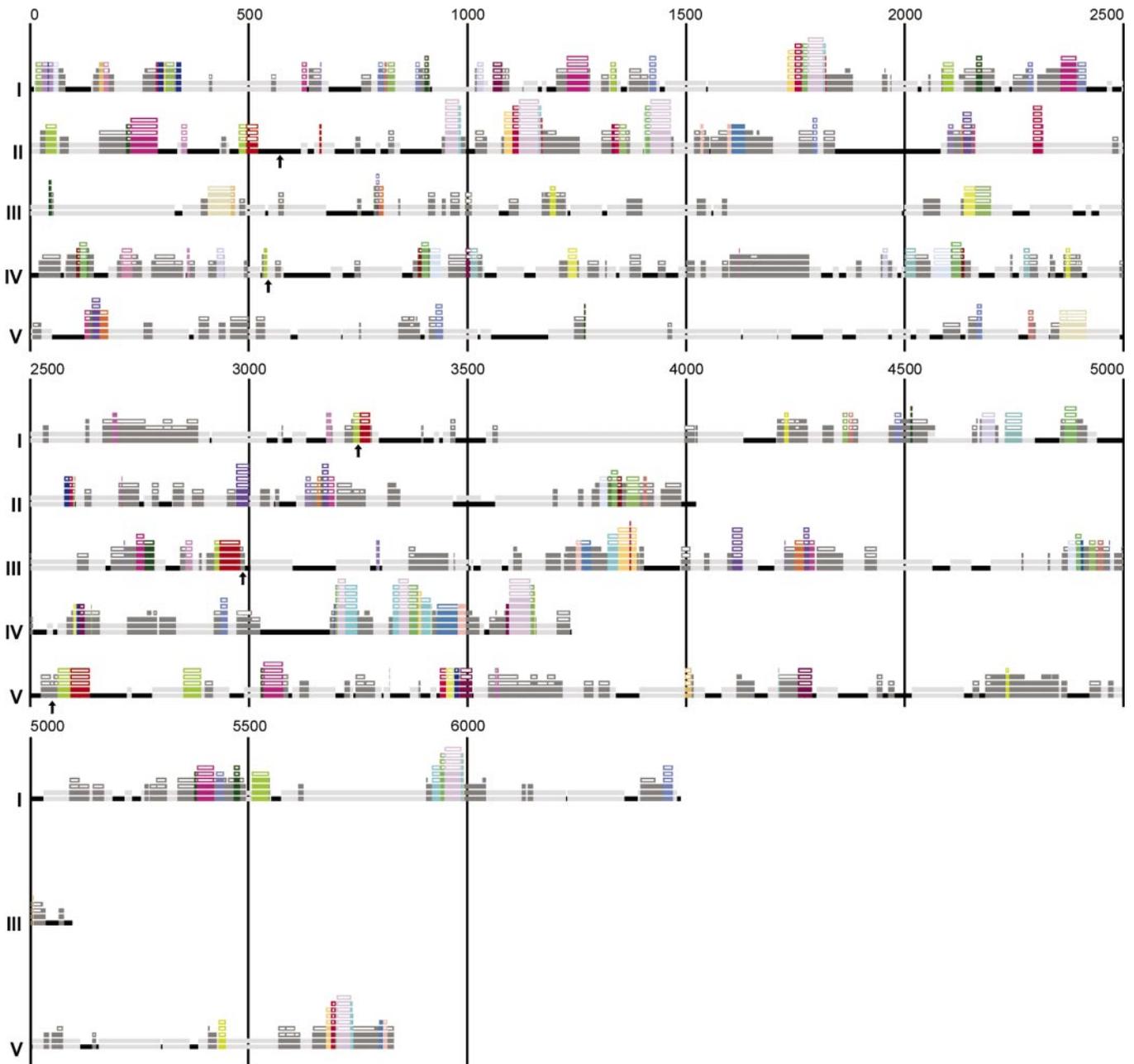


Fig. 2. Overview of the chromosomal location of all multiplicons detected in the *Arabidopsis* genome. Baselines (black) represent all genes on the five chromosomes of *Arabidopsis*. Boxes on the baselines indicate segments that are part of a multiplicon (group of homologous segments). The number of boxes above the baselines indicates the number of additional segments that are homologous to the segment marked on the baseline. Filled boxes represent nonhidden 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.

duplications between A and C ($a_{C,start}$) and A and B ($a_{B,start}$). Suppose $a_{C,start} \leq a_{B,start}$. In this case the value of $c_{A,start}$ is assigned to $c_{B,start}$. The value of $b_{C,start}$ is then determined by the coordinate b of the anchor point (a, b) in the duplication between segments A and B for which $a \geq a_{C,start}$ and lies the closest to $a_{C,start}$. The end positions ($b_{C,stop}, c_{B,stop}$) 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 nonhidden duplications that lead to its detection. To rule out hidden duplications generated by statistical aberrances,

we retain only those hidden duplications for which both nonhidden duplications have at least five anchor points on the common segment between them.

Results

Nonhidden Block and Tandem Duplications. By using the ADHoRe algorithm (14), we identified a total of 304 nonhidden duplications (i.e., duplications that can be observed through direct comparison of chromosomal segments) in the *A. thaliana* genome (see Fig. 2). These duplications contain a total of 3,571

Table 1. Duplications in the *Arabidopsis* genome

Chromosome no.	No. of genes in duplicated regions	Total no. 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

anchor points. Eighty-two percent of all genes in the annotated genome and 80% of all sequenced nucleotide positions reside in duplicated segments (Table 1). This percentage is significantly higher than the 60% reported by the *Arabidopsis* Genome Initiative (3). 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.* (7). 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 nonhidden duplication is shown in Fig. 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 (3).

A total of 137 nonhidden 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).

Hidden Duplications and Multiplication Levels. In addition to the set of nonhidden 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 Fig. 3B. Detailed analysis of the hidden duplications reveals that in many cases some residual anchor points can still be identified (i.e.,

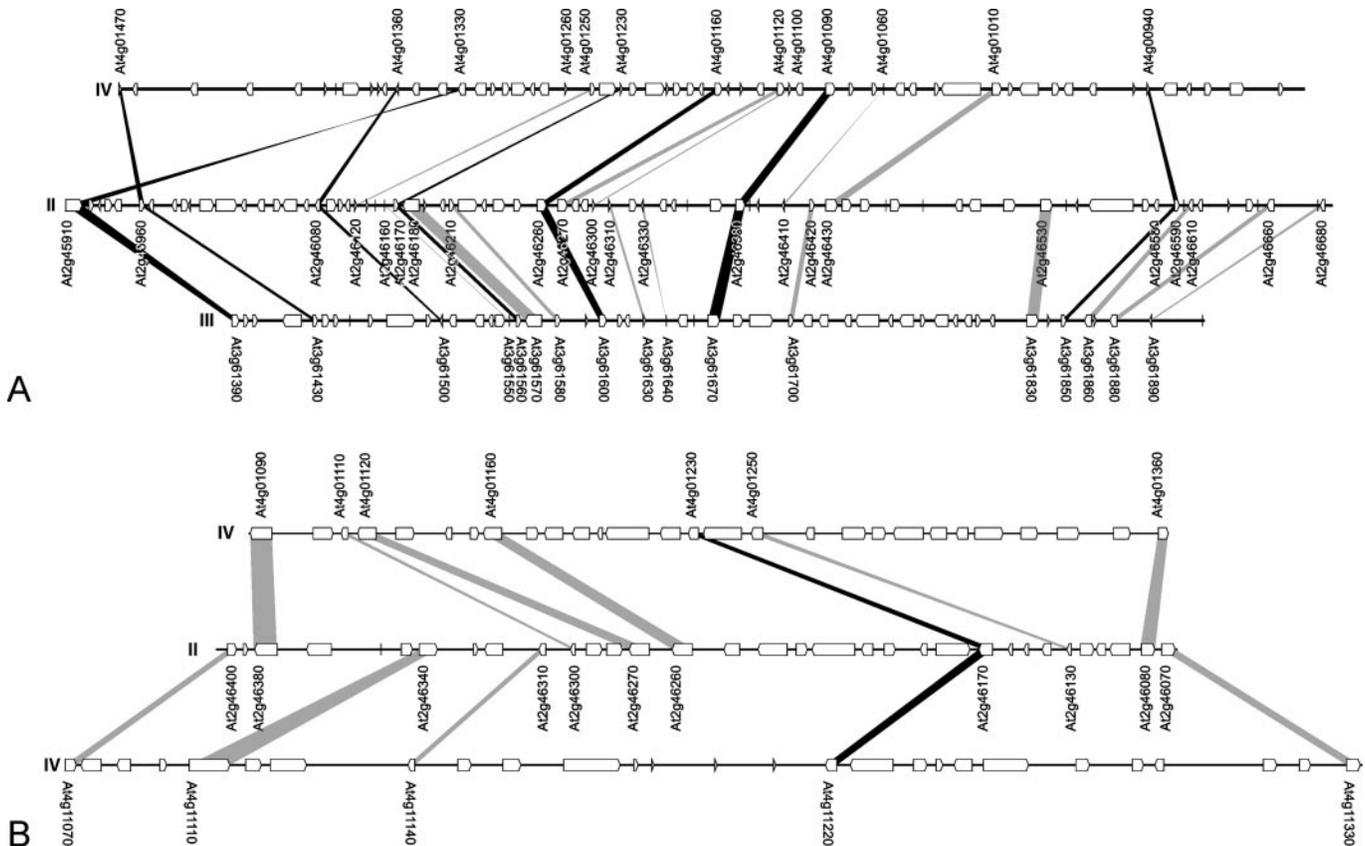


Fig. 3. Nonhidden and hidden duplicated blocks. (A) Example of a multiplicon in which nonhidden duplications can be observed between all three segments involved. Several genes can be distinguished that have homologs (indicated by black bands) on all segments. Light gray bands show homologs on two of three segments. (B) Example of a multiplicon in which no nonhidden 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.

Table 2. Detected age classes and age estimation

No. of blocks	No. of anchor points	Mean K_s (SD)	Age, 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)

My, million years.

some degree of colinearity can still be observed). However, the reason that these groups of anchor points are not recognized as nonhidden 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* and ref. 14). 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 (15). It should be noted that no duplications were found spanning the centromeric regions, which was also reported by Vision *et al.* (7).

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 (Fig. 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 nonhidden duplications, the maximum multiplication level observed in the genome of *Arabidopsis* equals five (Fig. 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 (Figs. 2 and 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 nonhidden block duplications in greater detail can be obtained from our web site at www.psb.rug.ac.be/.

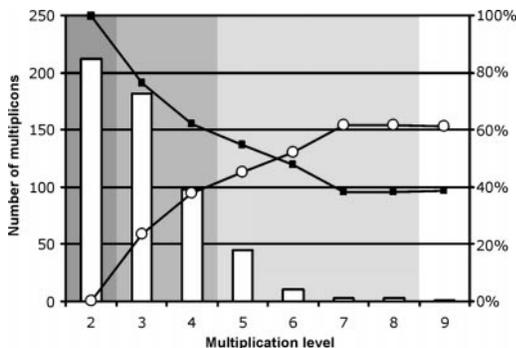


Fig. 4. Multiplication levels and contribution of nonhidden and hidden duplications. Bars indicate the number of multiplicons (groups of homologous segments) for each multiplication level. The relative amount of nonhidden duplications within all multiplicons of a given multiplication level is represented as a black square, whereas white circles denote the contribution of hidden duplications. The multiplication levels supporting three rounds of duplication (multiplication 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.

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 Fig. 2), the genome of *Arabidopsis* must have undergone 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 \log_2 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.* (25) 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 (26–27). 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 (28). 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* (4, 7). 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 (28, 29). The use of other substitution rates (e.g., those in refs. 30 and 31) might give quite different duplication dates. Nevertheless, to compare our study with recent studies that dealt with dating duplication events in *Arabidopsis* (4, 7) we have used the same substitution rate. Furthermore, although the absolute dating thus has to be

Table 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

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 Fig. 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 duplication event on chromosome 1. One of the nine segments of the multiplicon indeed consists of an internal nonhidden 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 nonhidden block duplications and anchor points involved in these internal duplications for chromosome 1 (see Table 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 (3, 4). By plotting the frequency distribution of duplication dates inferred for duplicated blocks of genes based on amino acid sequence divergences, Vision *et al.* (7) 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 (ref. 7; Todd Vision, personal communication), we have previously shown that many of their blocks have been dated erroneously (9). In our analysis, where we combined K_s -based dating of nonhidden 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.

We thank Jeroen Raes for helpful discussions and Bill Martin and Todd Vision for critical comments on the manuscript. C.S. and K.V. 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. is a Research Fellow of the National Fund for Scientific Research-Flanders.

1. Terryn, N., Heijnen, L., De Keyser, A., Van Asseldonck, M., De Clercq, R., Verbakel, H., Gielen, J., Zabeau, M., Villarroel, R., Jesse, T., *et al.* (1999) *FEBS Lett.* **445**, 237–245.
2. Blanc, G., Barakat, A., Guyot, R., Cooke, R. & Delseny, M. (2000) *Plant Cell* **12**, 1093–1101.
3. The *Arabidopsis* Genome Initiative (2000) *Nature* **408**, 796–815.
4. Lynch, M. & Conery, J. S. (2000) *Science* **290**, 1151–1155.
5. Grant, D., Cregan, P. & Shoemaker, R. C. (2000) *Proc. Natl. Acad. Sci. USA* **97**, 4168–4173.
6. Ku, H. M., Vision, T., Liu, J. & Tanksley, S. D. (2000) *Proc. Natl. Acad. Sci. USA* **97**, 9121–9126.
7. Vision, T., Brown, D. G. & Tanksley, S. D. (2000) *Science* **290**, 2114–2117.
8. Wolfe, K. H. (2001) *Nat. Rev. Genet.* **2**, 333–341.
9. Raes, J., Vandepoele, K., Simillion, C., Saeys, Y. & Van de Peer, Y. (2002) in *Genome Evolution*, eds Meyer, A. & Van de Peer, Y. (Kluwer, Dordrecht, The Netherlands), in press.
10. Gibson, T. J. & Spring, J. (1998) *Trends Genet.* **14**, 46–49.
11. Lynch, M. & Force, A. (2000) *Genetics* **154**, 459–473.
12. Wagner, A. (2002) *Genome Biol.* **3**, reviews 1012.1–1012.3.
13. Gaut, B. S. (2001) *Genome Res.* **11**, 55–66.
14. Vandepoele, K., Saeys, Y., Simillion, C., Raes, J. & Van de Peer, Y. (2002) *Genome Res.*, in press.
15. Vandepoele, K., Simillion, C. & Van de Peer, Y. (2002) *Trends Genet.*, in press.
16. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215**, 403–410.
17. Bairoch, A. & Apweiler, R. (2000) *Nucleic Acids Res.* **28**, 45–48.
18. Conery, J. S. & Lynch, M. (2001) in *Pacific Symposium on Biocomputing*, eds Altman, R. B., Dunker, A. K., Hunter, L., Lauderdale, K. & Klein, T. E. (World Scientific, Singapore), pp. 167–178.
19. Li, W. H. (1993) *J. Mol. Evol.* **36**, 96–99.
20. Nei, M. & Gojobori, T. (1986) *Mol. Biol. Evol.* **3**, 418–426.
21. Yang, Z. & Nielsen, R. (2000) *Mol. Biol. Evol.* **17**, 32–43.
22. Yang, Z. (1997) *Comput. Appl. Biosci.* **13**, 555–556.
23. Grubbs, F. (1969) *Technometrics* **11**, 1–21.
24. Stefansky, W. (1972) *Technometrics* **14**, 469–479.
25. McLysaght, A., Hokamp, K. & Wolfe, K. H. (2002) *Nat. Genet.* **31**, 200–204.
26. Soltis, D. E. & Soltis, P. S. (1993) *Crit. Rev. Plant Sci.* **12**, 243–273.
27. Song, K., Lu, P., Tang, K. & Osborn, T. C. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 7719–7723.
28. Li, W. H. (1997) *Molecular Evolution* (Sinauer, Sunderland, MA)
29. Lynch, M. (1997) *Mol. Biol. Evol.* **14**, 914–925.
30. Böhle, U. R., Hilger, H. H. & Martin, W. F. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 11740–11745.
31. Koch, M. A., Haubold, B. & Mitchell-Olds, T. (2000) *Mol. Biol. Evol.* **17**, 1483–1498.