Plant Molecular Biology **0:** 1–8, 2002.

The age of the *Arabidopsis thaliana* genome duplication

Maria D. Ermolaeva*, Martin Wu, Jonathan A. Eisen and Steven L. Salzberg
*The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA (*author for correspondence; e-mail mariae@tigr.org)*

Received .....; accepted in revised form 17 September 2002

Key words: Arabidopsis, bioinformatics, evolution, genome duplication, gene loss

Abstract

We estimate the timing of the *Arabidopsis thaliana* whole-genome duplication by means of phylogenetic and statistical analysis, and propose two possible scenarios for the duplication. The first one, based on the assumption that the duplicated segments diverged from an autotetraploid form, places the duplication at about 38 million years ago, after the *Arabidopsis* lineage diverged from that of soybean *(Glycine max*) and before it diverged from its sister genus, *Brassica*. The second scenario assumes that the ancestor was allotetraploid, and suggests that the duplication is younger than 38 million years and may have contributed to the *Arabidopsis-Brassica* divergence. In each case, our estimate places the age of the genome duplication as significantly younger than previously reported.

Introduction

Analysis of the genome of *Arabidopsis thaliana* showed that over 60% of the genome could be divided into about 24 large, megabase-scale duplicated segments (Arabidopsis Genome Initiative, 2000). In the paper reporting the complete genome sequence, it was also shown that none of these segments were found in three or more copies. This and other observations led the authors of that study to conclude that all the segments were duplicated simultaneously, in a single whole-genome duplication event. Because the sequence conservation remains strikingly clear at the DNA level, our original analysis (included in the *Arabidopsis* genome paper) suggested that the duplication might have occurred relatively recently. Prior to the completion of the genome, two studies attempted to date the duplication using the data available at the time. One such estimate, based on analysis of nonsynonymous substitutions, mapped the duplication to about 112 million years (Myr) ago (Ku *et al.*, 2000). Another analysis, based on synonymous substitution rates, gave a date of 65 Myr ago (Lynch and Conery, 2000).

The goal of this study is to estimate the age of the genome duplication on the basis of the complete *A. thaliana* genome sequence, phylogenetic methods based on duplicated gene families, and comparisons to preliminary genome sequence data from *Brassica oleracea*, a close relative of *A. thaliana*. As described in the *Arabidopsis* genome paper, the large-scale duplications between chromosomes occur in 24 discrete blocks. Within these blocks, most of the duplicated genes have been lost, with only about 1/3 remaining intact in two separate chromosomal locations. In many cases, one or both of these copies occur in a tandemly duplicated array of genes. These genes, which have been duplicated during both tandem and genome-scale events, provide a new type of molecular clock that allows us to assign a relative age to the genome duplication event.

Materials and methods

Phylogenetic analysis

Automated phylogenetic trees were inferred by (1) aligning proteins with ClustalW (Higgins *et al.*, 1996) and (2) building trees using the neighbor-joining algorithm of Phylip. Curated phylogenetic trees were inferred by (1) generating alignments with ClustalW; (2) refining the alignments by hand; and either (3a)

building trees using neighbor joining (NJ), for which trees were constructed based on pair-wise distances between amino acid sequences using the Protdist and Neighbor programs in PHYLIP (Felsenstein, 1989); or (3b) building trees by maximum likelihood (ML) with the Puzzle (Strimmer and von Haeseler, 1996) software, with 1000 puzzling steps and a variety of other parameters. The tree with greatest probability was chosen as the final tree, and the support values for each branch were used to evaluate their reliability. In the few cases in which the protein sequences were too conserved to be informative, the protein-coding DNA sequences were used instead and ML trees were made with Puzzle as in (3b) above. Bootstrap analyses from 100 replications were performed to evaluate the confidence for branches in the trees.

Analysis of gene loss

The coordinates of the segmental duplications and the list of the duplicated genes were defined as in the original *Arabidopsis* genome paper (Arabidopsis Genome Initiative, 2000). All *A. thaliana* genes (excluding genes in and around centromeres and telomeres) were divided into three groups: (1) duplicated genes, (2) genes that are located within duplicated regions but occur in just one copy, and (3) genes outside the segmental duplications. Each gene within a group was compared to all genes within the *A. thaliana* genome with BLASTP (Altschul *et al.*, 1990) and the average number of homologues per gene was calculated, with an E-value cutoff of 10–100. The number of homologues for each gene was summed, with self-hits and duplicates in the region of the putative segmented duplication not counted.

B. oleracea whole-genome shotgun sequences were obtained from the *Brassica oleracea*Genome Database at TIGR, http://www.tigr.org/tdb/e2k1/bog1

bog1. 10,000 *A. thaliana* sequences simulating random shotgun data were selected randomly from the whole genome. These sequences have the same average length and length variation as *Brassica* shotgun sequences. This set of sequences as well as the set of *Brassica* shotgun sequences was searched against the *A. thaliana* genome with BLASTX with an E-value cutoff of 10–10.

90% confidence intervals were calculated by dividing the 10 000 DNA sequences into groups, repeating the calculations for each group, and comparing the results in order to compute the mean and standard deviation.

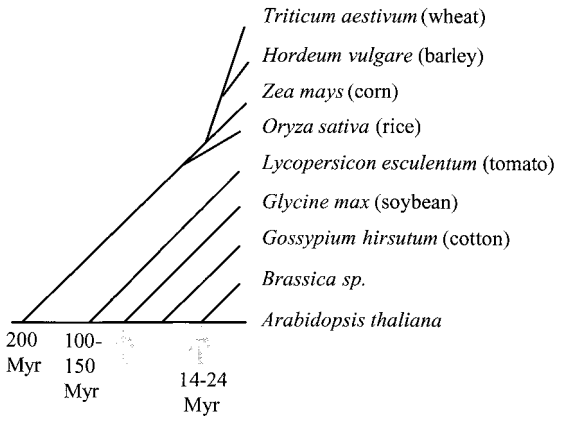

Figure 1. Evolutionary relationships between a few well-studied flowering plants (time scale is not preserved). Arrows show the two time points relative to which we date the *A. thaliana* genome duplication. The first time point is the *Arabidopsis*-soybean divergence and the second one corresponds to the *Arabidopsis-Brassica* divergence. The dates shown are estimates from the literature (Wolfe *et al.*, 1989; Gandolfo *et al.*, 1998; Yang *et al.*, 1999; Koch *et al.*, 2000, 2001).

Results and discussion

Phylogenetic analysis

Genome and chromosome duplication events result in the creation of large numbers of paralogous gene families within a species. The timing of such duplication events can be inferred by phylogenetic analysis of these gene families; in particular, one can determine the relative ordering of genome duplication events with respect to tandem duplications and to speciation events. For the analysis here, 5216 paired sets of *A. thaliana* proteins, with paralogues in both segments of a chromosomal duplication, were identified from searches of the *A. thaliana* proteome against itself. (Some genes are counted multiple times here due to tandem duplications.) Other plant species for which homologues of these duplicated sets are available were identified through searches of GenBank. Four species emerged as good candidates for analysis – tomato (*Lycopersicon esculentum*), soybean (*Glycine max*), cotton (*Gossypium hirsutum*), and *Brassica napus* – primarily because they diverged from the *Arabidopsis* lineage at time points near the lower and upper estimates for the date of the genome duplication. The

putative branching pattern of these species is shown in Figure 1.

We focused our analysis on soybean and *Brassica* after initial analyses indicated that the duplication likely occurred after the divergence of the tomato lineage, and after finding that too few homologous genes were available from cotton. To identify genes from these species that would be most useful for phylogenetic analysis, we compared all available genes from each (758 for *B. napus* and 1044 for *G. max)* to all *A. thaliana* gene pairs, and identified those with better matches to one of the members of a pair than to any other protein in the *A. thaliana* genome. For each of these genes, we constructed a phylogenetic tree containing the gene and all its *A. thaliana* homologues. These trees should provide a more reliable indication of evolutionary relationships than simple Blast (Altschul *et al*., 1990) searches. In order to root the trees, we identified additional homologues from other species by searching the plant proteins in GenBank and, if necessary, searching all of GenBank with a lower similarity cutoff. We generated and curated phylogenetic trees by both distance and maximum likelihood methods (Felsenstein, 1996).

For the interpretation of these results, it is important to realize that the genomes of *B. napus* and *G. max* are not yet complete. This lack of completeness, and the fact that these species may have lost the closest homologues of the duplicated pair of genes, means that it is possible that the closest available homologue may actually be a distant paralogue of the gene of interest. Out of 20 gene families containing *Brassica* homologues, 3 families were discarded either because the sequence alignments were too short, or because the homology was too weak for phylogenetic analysis. Another family was removed because the curated phylogenetic tree shows that the *Brassica* gene is not the closest homologue of the pair of duplicated *A. thaliana* genes. All 16 remaining trees (all shown in Table 1) have the topologies shown in either Figure 2a or Figure 2b, with most branches having high ($>$70%) bootstrap support. These tree topologies are consistent with a genome duplication that pre-dated the divergence of *Arabidopsis* and *Brassica*.

Table 1 shows the ratios of branch lengths (t1/t2) for these 16 trees, where t1 is the age of the *Brassica-Arabidopsis* divergence and t2 is the age of the *Arabidopsis* genome duplication (Figure 3a). t1 and t2 were calculated with the Protdist program, and similar results were obtained with γ-corrected distances (Gu and Zhang, 1997).

The average ratio and 90% confidence interval are 0.48 ± 0.07. With estimates for the *Arabidopsis-Brassica* divergence ranging from 14 to 24 Myr old (Yang *et al.*, 1999; Koch *et al.*, 2000; Koch *et al.*, 2001), this t1/t2 ratio indicates that the *Arabidopsis* genome duplication occurred some 28–48 Myr ago. If the *Arabidopsis* ancestor was allotetraploid rather than autotetraploid, then the duplication event would have occurred more recently, because the two species (or ecotypes) that formed the allotetraploid would necessarily pre-date the duplication event.

To analyze the timing of the genome duplication in *Arabidopsis* relative to the species' divergence from soybean, 25 potentially suitable gene families were identified by the methods described above. Two of these were not included in further analysis: one because its sequence alignment was too short, and one because of complications due to potential lateral transfers from the chloroplast genome (Arabidopsis Genome Initiative, 2000). Phylogenetic trees were built and manually curated for the remaining 23 families (Table 2). These trees displayed a variety of distinct phylogenetic patterns, all illustrated in Figure 2. If a soybean gene and an *Arabidopsis* gene pair are separated by deeper branching species, as shown in Figure 2d, then these genes are probably distant paralogues and the closest soybean homologue is not available in the sequence databases.

Eight trees either had this topology or had low bootstrap support (below 70%) and thus were deemed inappropriate for further analysis. If the soybean lineage split from the *Arabidopsis* lineage before the genome duplication, then one would expect that homologous soybean genes would form a separate branch from a duplicated *Arabidopsis* gene pair. This is the pattern shown in Figure 2c. Of the 15 gene trees (Table 2), 12 have the pattern shown in Figure 2c based on both distance and likelihood methods. The remaining three trees are likely to include genes that are paralogues of the segmentally duplicated genes rather than the duplicated genes themselves, or possibly they contain soybean genes that are not true orthologues of the duplicated genes. For example, the gene At1g18080 from the tree (At1g18080, At1g48630) cannot be the original segmental duplicate of At1g48630 if At1g48630 and At3g18130 are also duplicates (Table 2, line 17). In the case of the (At2g28490, At1g07750) tree, the gene At2g28490 is likely to be a paralogue of the duplicated genes rather than the duplicate itself and, based on the tree structure, the best candidate for a true dupli-

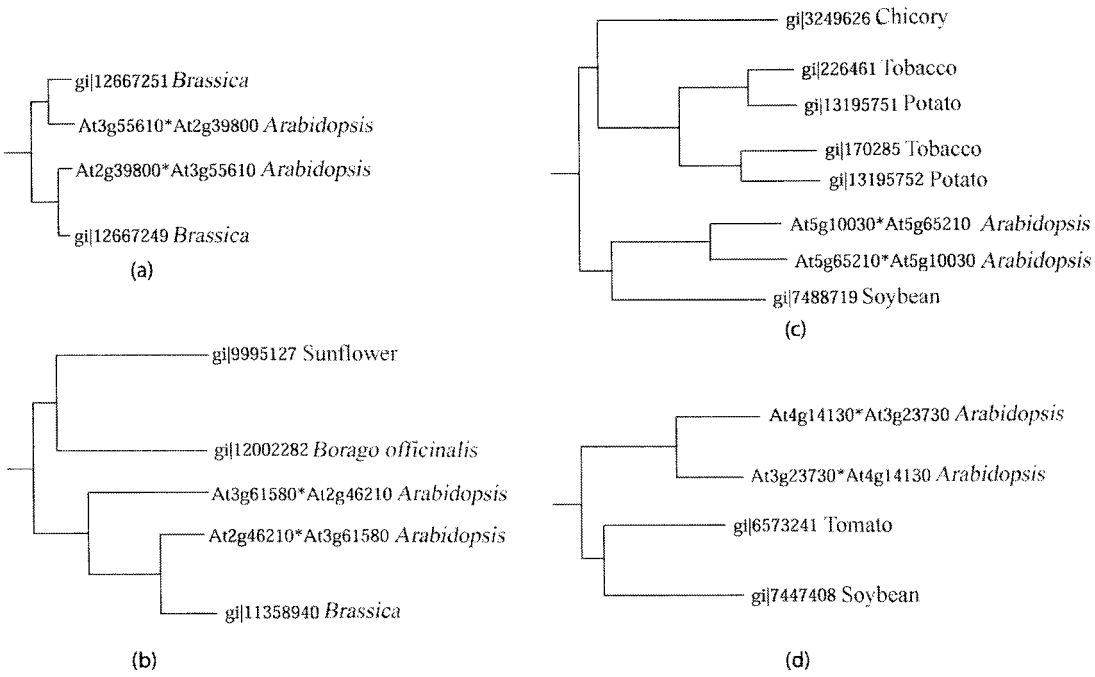

Figure 2. Examples of phylogenetic trees of gene families used to infer the timing of the *A. thaliana* genome duplication. Trees were generated as described in the text. *A. thaliana* genome IDs or GenBank gi numbers are listed. Branch lengths in the different trees are not to scale. The trees were rooted using outgroups from distantly related plants such as rice (not shown).

Table 1. Sixteen trees with segmentally duplicated *Arabidopsis* genes and their orthologues in *Brassica napus*. 'Method' is either a maximum-likelihood tree based on DNA sequence alignments (DNA ML) or a neighbor-joining tree built from protein sequence alignments (Protein NJ). The four possible branching patterns (a–d) are shown in Figure 2. *Arabidopsis* genes can be found in the *Arabidopsis* genome database at http://www.tigr.org/tdb/e2k1/ath1/

Arabidopsis gene	*Brassica* gene (GenBank gi identifier)	Method	Branching pattern	Bootstrap support, %	t1/t2 ratio
At1g15570 At1g80370	1076444	Protein NJ	Figure 2b	99	0.559078
At2g31680 At1g05810	1076457	Protein NJ	Figure 2b	98	0.801325
At5g06290 At3g11630	11119229	Protein NJ	Figure 2b	85	0.344086
At2g46210 At3g61580	11358940	Protein NJ	Figure 2b	100	0.359926
At1g12290 At1g62630	11761662	Protein NJ	Figure 2b	100	0.416413
At3g58780 At2g42830	12655901	Protein NJ	Figure 2b	93	0.617284
At2g39800 At3g55610	12667249 12667251	Protein NJ	Figure 2a	100	0.298755
At4g28510 At2g20530	12751303	Protein NJ	Figure 2b	88	0.4375
At3g17800 At1g48450	13561930	Protein NJ	Figure 2b	100	0.223404
At4g14080 At3g23770	322641	Protein NJ	Figure 2b	100	0.623853
At3g50820 At5g66570	5052366	Protein NJ	Figure 2b	100	0.297082
At4g13050 At3g25110	541913	Protein NJ	Figure 2b	93	0.416938
At3g49010 At5g23900	730449	DNA ML	Figure 2b	88	0.51938
At1g07370 At2g29570	7440029	Protein NJ	Figure 2b	73	0.689655
At4g14070 At3g23790	7451471	Protein NJ	Figure 2b	74	0.554307
At1g20130 At1g75910	99824	Protein NJ	Figure 2b	100	0.441982

Table 2. Twenty-three trees with segmentally duplicated *Arabidopsis* genes and their orthologues (or paralogues) in *Glycine max*. '(Figure 2d)' means that the bootstrap value for the branch that determines whether the soybean gene is an orthologue for the *Arabidopsis* duplicated genes is less than 70%. *Arabidopsis* genes can be found in the *Arabidopsis* genome database: http://www.tigr.org/tdb/e2k1/ath1/.

Arabidopsis genes	Soybean gene (GenBank gi identifier)	Method	Branching pattern	Bootstrap support, %
At3g60880 At2g45440	1084359	DNA ML	Figure 2c	100
At3g55840 At2g40000	12006354	Protein ML	Figure 2c	100
At3g14290 At1g53850	12229923	Protein ML	Figure 2c	82
At2g28490 At1g07750	12697782	Protein NJ	Opposite to Figure 2c	100
At1g62740 At1g12270	2129844	Protein NJ	Figure 2c	70
At4g12400 At4g22670	2129844	Protein NJ	Opposite to Figure 2c	70
At1g79470 At1g16350	4468193	Protein NJ	Figure 2c	100
At5g58950 At3g46930	478809	Protein NJ	Figure 2c	100
At3g48730 At5g63570	541940	Protein NJ	Figure 2d	–
At1g51590 At3g21160	6552504	Protein NJ	Figure 2c	96
At1g53240 At3g15020	7431174	Protein NJ	(Figure 2d)	–
At3g17240 At1g48030	7431870	Protein NJ	Figure 2c	100
At5g10450 At5g65430	7435033	Protein NJ	(Figure 2d)	–
At2g16500 At4g34710	7436498	Protein NJ	(Figure 2d)	–
At5g11300 At5g25380	7438491	Protein NJ	Figure 2c	100
At5g07340 At5g61790	7441504	Protein NJ	(Figure 2d)	–
At1g48630 At3g18130	7446123	Protein NJ	Figure 2c	100
At1g18080 At1g48630	7446123	Protein NJ	Opposite to Figure 2c	100
At3g23730 At4g14130	7447408	Protein NJ	Figure 2d	–
At5g65210 At5g10030	7488719	Protein NJ	Figure 2c	99
At2g34840 At1g30630	7670062	Protein NJ	Figure 2c	90
At3g53260 At2g37040	81807	Protein NJ	(Figure 2d)	–
At2g29570 At1g07370	99946	DNA NJ	Figure 2d	–

Table 3. Average number of homologues within different groups of genes in the *Arabidopsis* genome. The 90% confidence intervals for the average numbers of *Brassica* hits for duplicated and single genes within segmental duplications do not overlap, which means that there is not more than 0.0025 chance that the real value of R' for *Brassica* (i.e. the value of R' that we would get if the whole *Brassica* genome would be available) is more than 1.

Type of genes in the group	Number of genes in the group	Average number of homologues	Average number of hits with 10 000 random *A. thaliana* sequences (simulated shotgun)	Average number of hits with *Brassica* shotgun data (16383 sequences)
Duplicated genes within duplicated regions	6854	6.62	2.71 ± 0.3	1.58 ± 0.3
Single genes within duplicated regions	11427	4.50	2.04 ± 0.3	2.08 ± 0.2
Genes within single regions	4030	5.43	2.34	2.49

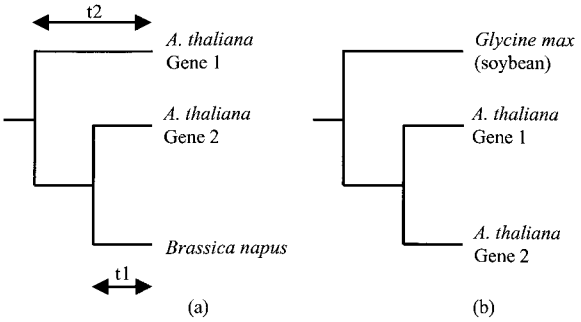

Figure 3. The most common branch order among the phylogenetic trees analyzed in this study.

cate is At2g28680. The third tree with an inconsistent branching pattern can be explained similarly. Thus the majority of trees indicate that soybean and *Arabidopsis* separated before the genome duplication. The most common branching pattern is schematically illustrated in Figure 3.

Analysis of gene loss

Genome duplication is frequently followed by extensive and rapid gene loss (Lynch and Conery, 2000). This suggests that if the *Arabidopsis* genome duplication occurred long before the split from the *Brassica* lineage, then most of the post-duplication gene losses would have occurred before the split. If the duplication happened shortly before the split, then much of the gene loss would have occurred separately in the two lineages. As we show below, most of the gene loss after the genome duplication appears to have occurred after the species diverged. Because no *Brassica* genome is yet complete, we developed a method to analyze the shotgun genome sequence data from *B. oleracea* to measure the extent of gene loss. This test is based on estimating the number of homologues within *A. thaliana* for genes both inside and outside duplicated chromosomal segments. We then make similar estimates for the partial genome data from *Brassica.*

The number of homologues within the *A. thaliana* genome was measured for three sets of *Arabidopsis* genes: (1) duplicated genes within duplicated segments (i.e., those genes that are found in both copies of a segmental duplication); (2) genes within duplicated segments that do not have homologues in the other duplicated segment, which likely represent gene loss events, and (3) genes occurring outside the segmentally duplicated regions, excluding genes in centromeric and telomeric regions. For each of these sets of genes, the average number of homologs per gene was measured (Table 3).

For example, the first set (duplicated genes within duplicated segments) includes 6854 genes. These 6854 genes were searched against all genes within *Arabidopsis* genome (some genes were not counted, as explained in Materials and methods) and the number of hits with an E-value below the cutoff (10–100) was 45 374, giving an average of 45 374/6854 = 6.62 homologues per gene. For segmentally duplicated genes, this average did not count the duplicated genes among each others' homologues. Table 3 shows that duplicated genes in duplicated segments have a significantly higher number of homologs than genes from the other groups.

One possible explanation for this result is that the duplicated genes in duplicated regions have more tandem duplicates than other genes. This is theoretically possible, because the extra copy of each gene in the duplicated segment might facilitate the creation of more tandem duplicates for that gene family. Based on the tree topologies from our phylogenetic analysis, only about 20% of the tandem duplications happened after the genome duplication. Previous reports showed that over 4000 genes, ca. 16% of the proteome, exist in tandemly repeated arrays ranging from 2 to 23 in length (Arabidopsis Genome Initiative, 2000). If every tandem duplication occurred after the genome duplication, then an average gene would have 0.16 tandem duplicates, and genes that were already duplicated would get an average of 0.32 new homologues through tandem duplication. Thus segmentally duplicated genes would have 0.16 'extra' homologues on average, for a ratio of 1.16:1. The observed ratio is 1.47:1, much greater than would be expected even if all tandem duplications occurred recently.

Another, more plausible explanation for this excess of homologues is that gene loss was a non-random process, directed by evolutionary pressure. Therefore, the probability that a gene will be lost is partially correlated with the probability that its homologues will be also lost. In other words, similar evolutionary pressure is likely to apply to all genes in a set of paralogues. This explanation gains plausibility from the observation that homologous genes, especially recent duplicates, are likely to be involved in similar functions. If a function is not critical for survival, then the duplicates with that function may be more likely to be lost. This phenomenon would cause the average number of homologues in the category 'duplicated region – single gene' to be less than for 'duplicated

region – duplicated gene'. For genes in the category 'single region – single gene', the average number of homologues should be intermediate, because the non-duplicated regions contain both functionally critical genes as well as functionally less important genes.

These proportions are precisely what we observe in Table 3. To further analyze the relationship between gene duplication and gene loss, we computed the average number of homologues in *B. oleracea* for each of the three sets of *A. thaliana* proteins to see if similar patterns were observed between species. Ideally, this analysis would include the complete *Brassica* genome, but because this is not available, we used partial shotgun sequencing data instead. We accomplished this by simulating a shotgun sequencing project at about the same depth of coverage for *A. thaliana*, and then asking whether the results were similar to those obtained from the complete genome.

A total of 10 000 sequences from the *A. thaliana* genome were generated from random locations on the genome. The number of Blast hits between the duplicated gene pairs and the set of the random *A. thaliana* DNA sequences was analyzed (Table 3, column 4). This analysis showed that the duplicated genes have significantly higher numbers of hits to random shotgun *A. thaliana* sequences than do single genes, although the ratio of the numbers of hits ($R' = 2.71/2.04 = 1.33$) is smaller than the corresponding ratio for the number of homologues in the whole genome ($R = 6.62/4.50 = 1.47$). This difference can be explained by the fact that the DNA sequences in the simulated shotgun are relatively short, necessitating a much higher Blast E-value threshold, which in turn may have added a significant amount of noise to the experiment. The simulation nonetheless shows that the number of hits to shotgun data reflects the number of homologues in the genome.

For the *B. oleracea* genome, we used a set of 16 383 random shotgun DNA sequences, with an average length of 685 nucleotides. Surprisingly, the average number of hits between the *A. thaliana* genes and the *B. oleracea* shotgun sequences is lower for duplicated genes than for single genes within the duplicated regions, with a corresponding ratio $R' = 1.58/2.08 = 0.76$. This ratio might be even lower if the random hits resulting from the higher E-value threshold were factored in. The average number of hits between non-duplicated regions of the *Arabidopsis* genome and *Brassica* (2.49) is also shown in Table 1. The results described above indicate that duplicated genes in *A. thaliana* have more homologues within that genome but have fewer homologues in *B. oleracea*.

Thus it appears that gene loss in *Arabidopsis* and *Brassica* are anti-correlated: genes that were lost in *Arabidopsis* were more likely to be preserved in *Brassica*, and vice versa. One plausible explanation of this anti-correlation is that most of the post-duplication gene loss happened after *Arabidopsis* and *Brassica* diverged. If instead the gene loss happened before the lineages diverged, then the sets of genes lost in *Arabidopsis* should significantly overlap those lost in *Brassica*, leading to a positive correlation, quite the opposite of what we observed. The value for the number of homologues of genes in single regions is difficult to interpret, in part because such genes may be contained within additional duplications in *Brassica* (Lan *et al.*, 2000; O'Neill and Bancroft, 2000; Parkin *et al.*, 2002).

According to our phylogenetic analysis, the whole-genome duplication happened after *A. thaliana* diverged from soybean but before it diverged from *Brassica*. The duplication is about twice as old as the *Arabidopsis-Brassica* divergence, which happened about 14–24 Myr ago (Yang *et al.*, 1999; Koch *et al.*, 2000, 2001); this puts the age of the duplication at 28–48 Myr ago. If the ancestor of *Arabidopsis* was allotetraploid, then the duplication occurred more recently. Our statistical analysis of lost genes suggests that most duplicated genes were lost after *Arabidopsis* and *Brassica* diverged, suggesting that the ago of the duplication is close to the lower end of that range, perhaps around 30–35 Myr ago. Also consistent with this estimate is the recent estimate that the half-life of duplicated genes is just 3–7 Myr (Lynch and Conery, 2000). Of course, there may be other plausible explanations for the uncorrelated gene loss in *Arabidopsis* and *Brassica*. Uncertainty about the precise age of the *Brassica-Arabidopsis* split, the possibility of an allotetraploid ancestor, and limitations on the accuracy of phylogenetic analysis prevent us from providing an exact date for the duplication; however, our results indicate that the duplication is much younger than 65 or 112 Myr as was previously suggested (Lynch and Conery, 2000; Ku *et al.*, 2000).

The *Arabidopsis* genome duplication provided material for fast and efficient evolution, and could have helped contribute to the divergence of *Arabidopsis* and *Brassica* from each other by allowing each lineage to lose different genes or to evolve new functions. Speciation caused by large-scale genome duplication is consistent with the chromosomal model for speci-

ation (White, 1978). Similar large-scale duplications have been documented in the maize (Gaut and Doebley, 1997) and yeast genomes (Wolfe and Shields, 1997; Seoighe and Wolfe, 1999) and smaller-scale rearrangements (also followed by gene loss) have been described in yeast (Fischer *et al.*, 2000, 2001). Much more work, and many more genomes, will be necessary to determine the full extent of the role of genome duplication in the evolution of species.

Acknowledgements

We would like to thank Chris Town and Brian Haas for useful discussions and comments. M.D.E., M.W., and S.L.S. were supported in part by National Science Foundation grant KDI-9980088, S.L.S. was partly supported by NIH grant R01-LM06845, and M.D.E. was partly supported by NSF Cooperative Agreement DBI 98-13586.

References

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.

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

Felsenstein, J. 1989. PHYLIP – Phylogeny Inference Package (version 3.2). Cladistics 5: 164–166.

Felsenstein, J. 1996. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Meth. Enzymol. 266: 418–427.

Fischer, G., James, S.A., Roberts, I.N., Oliver, S.G. and Louis, E.J. 2000. Chromosomal evolution in *Saccharomyces*. Nature 405: 451–454.

Fischer, G., Neuveglise, C., Durrens, P., Gaillardin, C. and Dujon, B. 2001. Evolution of gene order in the genomes of two related yeast species. Genome Res. 11: 2009–2019.

Gandolfo, M.A., Nixon, K.C. and Crepet, W.L. 1998. A new fossil flower from the Turonian of New Jersey: *Dressiantha bicarpellata* gen. et sp. nov. (Capparales). Am. J. Bot. 85: 964–974.

Gaut, B.S. and Doebley, J.F. 1997. DNA sequence evidence for the segmental allotetraploid origin of maize. Proc. Natl. Acad. Sci. USA 94: 6809–6814.

Gu, X. and Zhang, J. 1997. A simple method for estimating the parameter of substitution rate variation among sites. Mol. Biol. Evol. 14: 1106–1113.

Higgins, D.G., Thompson, J.D. and Gibson, T.J. 1996. Using CLUSTAL for multiple sequence alignments. Meth. Enzymol. 266: 383–402.

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.A., Haubold, B. and Mitchell-Olds, T. 2001. Molecular systematics of the Brassiceae: evidence from coding plastidic *matK* and nuclear *Chs* sequences. Am. J. Bot. 88: 534–544.

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.

Lan, T.H., DelMonte, T.A., Reischmann, K.P., Hyman, J., Kowalski, S.P., McFerson, J., Kresovich, S. and Paterson, A.H. 2000. An EST-enriched comparative map of *Brassica oleracea* and *Arabidopsis thaliana*. Genome Res. 10: 776–788.

Lynch, M. and Conery, J.S. 2000. The evolutionary fate and consequences of duplicate genes. Science 290: 1151–1155.

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.

Parkin, I.A., Lydiate, D.J. and Trick, M. 2002. Assessing the level of collinearity between *Arabidopsis thaliana* and *Brassica napus* for *A. thaliana* chromosome 5. Genome 45: 356–366.

Seoighe, C. and Wolfe, K.H. 1999. Updated map of duplicated regions in the yeast genome. Gene 238: 253–261.

Strimmer, K. and von Haeseler, A. 1996. Quartet puzzling: A quartet maximum-likelihood method for reconstructing tree topologies. Mol. Biol. Evol. 13: 964–969.

White, M.J.D. 1978. Models of Speciation. Freeman, San Francisco, CA.

Wolfe, K.H., Gouy, M., Yang, Y.W., Sharp, P.M. and Li, W.H. 1989. Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. Proc. Natl. Acad. Sci. USA 86: 6201–6205.

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

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.