



## METHODS

Phylogenetic dating involves a three-step process which culminates in rooted phylogenetic trees (Fig. 1). The starting material for the analysis is a file of duplicate protein pairs for a genome in which one wants to investigate the timing of duplication. Such a dataset is obtained through detailed analysis of patterns of gene duplication in the genome of interest. Software and algorithms have been described elsewhere for detecting duplicate pairs (Conant and Wagner, 2002; Calabrese *et al.*, 2003) and a number of *ad hoc* methods have also been used (Bowers *et al.*, 2003), so delineation of duplicate pairs is not a goal of this software package and will not be discussed here.

### Organism database creation

A preliminary step in running the dating analysis involves establishing sequence databases for a set of organisms. Our approach requires only partial sequence data and makes use of the taxonomy structure databases at the National Center for Biotechnology Information, so that organism-specific databases can be retrieved for any level of taxonomical organization.

Duplication events are evaluated relative to the divergence from a common ancestor of the organism in which we are analyzing duplication patterns (duplicate organism) and the organism represented in the database (comparison organism). Therefore, comparison databases should be chosen with varying evolutionary distances from the duplicate organism.

The organism databases can be subjected to clustering or other treatments to remove redundant or low-quality sequences.

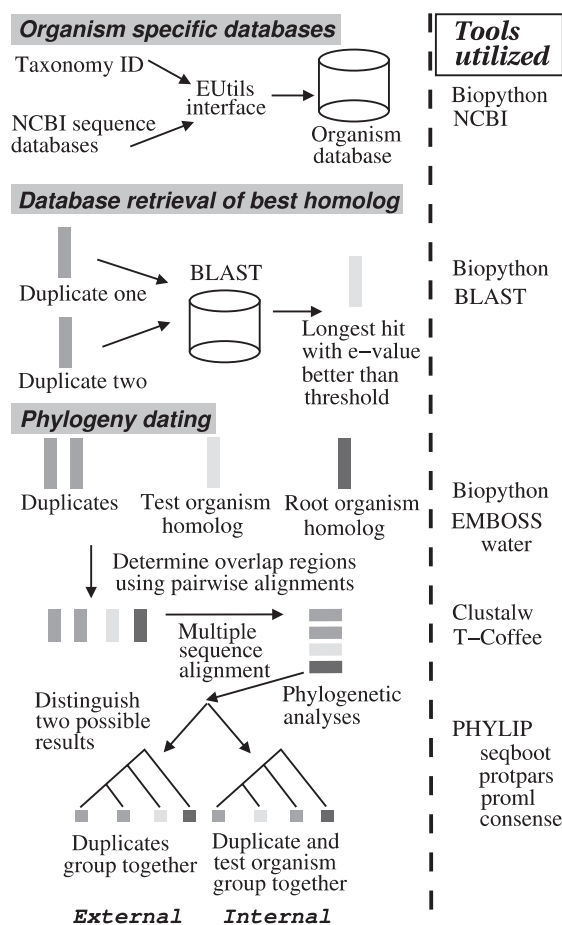
### Detection of homologs

With organism-specific databases in hand, the next step involves selection of a best detectable homolog for each duplicate protein pair. This homolog is defined as the BLAST (Altschul *et al.*, 1997) (*tblastn*) hit with the longest match region above a configurable significance (*E*-value) threshold (the default is an *E*-value of  $10^{-5}$ ).

In this way, we attempt to obtain the longest detectable region of sequence similarity. At this stage there are no restrictions on the lengths of alignments accepted as the best hit—length restrictions are applied during preparations for phylogenetic analysis. To try and obtain a useful sequence beyond the boundaries detected by BLAST, additional translated sequence to either side of the original hit location is included if available. Subsequent local alignment comparisons will remove non-informative added sequence.

### Generation of phylogenetic trees

The starting materials for phylogeny construction are: (a) the two protein sequences from a duplicate gene pair; (b) the best homolog from a comparison organism representing a particular taxonomic node; and (c) the best homolog from



**Fig. 1.** Three-step approach to performing dating of duplications. Step 1 involves download of a taxon-specific database using the EUtils interface at NCBI. Step 2 involves the identification of the best detectable homolog from a comparison organism and an outgroup for each duplicate protein pair using homology searches against the databases from step 1. Step 3 involves performing a four-sequence phylogeny, visualized as a tree. The freely available bioinformatics tools utilized at each stage are listed.

an outgroup organism known to be a very distant relative and thus used as the root in the final phylogeny. These four protein sequences are subjected to Smith–Waterman pairwise alignments [‘water’ in EMBOSS (Rice *et al.*, 2000)] to find the detectable regions conserved between all pairs.

With the regions of sequence similarity overlap for all four proteins, multiple alignments are performed using T-Coffee (Notredame *et al.*, 2000). Other multiple sequence alignment programs such as ClustalW (Chenna *et al.*, 2003) are also supported. Alignment quality is assessed using T-Coffee’s ability to produce reliability scores.

Multiple alignments are used as input into phylogenetic analyses to produce rooted trees comparing the duplicate pairs and best homolog. The PHYLIB set of programs (Felsenstein, 2003, <http://evolution.genetics.washington.edu/phylib.html>) is used for bootstrapped maximum likelihood

(‘proml’) and protein parsimony (‘protpars’) analyses. Bayesian approaches as implemented by MrBayes (Ronquist and Huelsenbeck, 2003) are also supported.

The results of phylogenetic analysis lead to only two possible rooted tree topologies. In one topology (Fig. 1, external tree), the members of the duplicated pair are more similar to one another than either is to the best homolog. In this case, the best homolog is referred to as external to the duplicate pair. In the alternative topology (Fig. 1, internal tree), the homolog is more similar to one member of the duplicate pair than is the other member of the duplicate pair. For this result, the best homolog is considered internal to the duplicate pair.

**ALGORITHM**

**Interpretation of individual trees**

The analyses described above produce the best possible tree of two duplicates and a homolog, rooted using an outgroup sequence. Interpretation of this tree provides the mechanism for dating the duplication event being examined. The goal of this analysis is to assign a relative order to two different events:

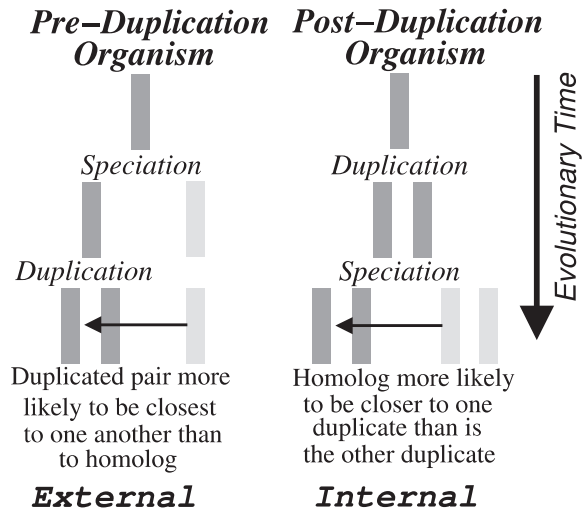
- The genome duplication being examined in our organism of interest.
- The divergence time, from a common ancestor, of two organisms—the study organism and the comparison organism.

Making the assumption that the length of evolutionary time since separation of genes from a common ancestral gene (whether by speciation or duplication) is reflected by their degree of phylogenetic similarity, we can thus evaluate which of the two scenarios pictured in Figure 2 correlates with our two possible phylogenetic tree topologies (Fig. 1). If speciation occurred prior to duplication, then we expect a tree with the duplicates that are closest (external) more often. In contrast, if the duplication event preceded speciation, then we expect the organism homolog to be closer to one of the duplicate pairs (internal) more often.

**Interpretation of collected results**

Large numbers of duplicate pairs are evaluated to minimize the impact of evolutionary events differing from our assumptions (tandem duplications and deletions, gene conversion). Practically, we also expect some failures to identify true homologs as a consequence of using sequence similarity searches with partial sequence databases (expressed sequence tag, genomic fragment) of necessarily limited size. The requirement of a divergent organism as an outgroup root results in highly conserved regions being most commonly represented in the analyses.

External trees are prone to include a disproportionately large share of these artifactual results, hence inferences rely heavily



**Fig. 2.** Distinguishing the timing of duplications from phylogenetic analysis. The primary assumption is that phylogenetic similarity is an indicator (albeit imperfect) of the length of time for which two or more genes have been evolving independently. External and internal classifications relate to the trees in Figure 1.

upon comparative ratios of internal trees to total trees evaluated. These ratios are examined across all comparisons to look for organisms which provide different amounts of evidence for one particular evolutionary history of duplication and speciation (Fig. 2). Methods such as correlated ANOVA analysis between duplication blocks (Bowers et al., 2003) can be applied to test the significance of differences in frequencies of internal trees between organisms.

**IMPLEMENTATION**

The analyses described above were implemented as a set of scripts and modules in the Python programming language. This code drives the bioinformatics tools listed in Figure 1 and is modularized for use in clustered and single-computer environments. A flexible Python-based configuration file allows the code to be utilized for a wide variety of duplicate organisms and comparison databases.

**RESULTS**

The implementation described above was used for dating of duplications in the *Arabidopsis* and *Oryza* (rice) genomes. Duplications have played an important role in shaping current plant genomes (Schmidt, 2002; Mitchell-Olds and Clauss, 2002). Understanding duplication events will have important consequences for detecting long-range synteny and applying information from sequenced plant genomes to important crop species with limited sequence data.

**Arabidopsis**

Since the discovery of large blocks of duplicated genes in the *Arabidopsis* genome (Arabidopsis Genome Initiative, 2000;

**Table 1.** *Arabidopsis* duplication dating results for three duplication events:  $\alpha$ ,  $\beta$  and  $\gamma$ , in order of increasing age (Bowers *et al.*, 2003)

Event	Pinaceae TaxId:3318	<i>Oryza</i> TaxId:4527	Solanaceae TaxId:4070	Medicago TaxId:3877	Malvaceae TaxId:3629	Citrus TaxId:2706	Brassica TaxId:3705
$\alpha$	0.034 (669)	0.055 (965)	0.070 (719)	0.090 (636)	0.070 (567)	0.073 (449)	<b>0.485 (526)</b>
$\alpha$ short	0.037 (54)	0.029 (68)	0.058 (52)	0.064 (47)	0.049 (41)	0.059 (34)	<b>0.488 (41)</b>
$\beta$	0.091 (243)	<b>0.214 (290)</b>	<b>0.313 (211)</b>	<b>0.328 (192)</b>	<b>0.328 (177)</b>	<b>0.292 (154)</b>	<b>0.688 (218)</b>
$\gamma$	0.324 (71)	0.532 (94)	0.655 (87)	0.675 (80)	0.545 (66)	0.596 (57)	0.808 (78)

Ratios indicate the number of internal trees to total trees examined; numbers in parentheses are the total number of trees analyzed for each event and organism. Grayed numbers indicate inferred organisms post-dating the duplication event based on eyeball comparisons of internal tree ratios.

Blanc *et al.*, 2000; Vision *et al.*, 2000; Paterson *et al.*, 2000; Kowalski *et al.*, 1994), much work has focused on understanding the nature of these duplicated blocks. The genome is made up of duplicated and rearranged blocks which are thought to have been generated by a whole genome duplication event that occurred relatively recently in evolutionary history, before the divergence of the legumes and the mallows from *Arabidopsis* (Ermolaeva *et al.*, 2003; Bowers *et al.*, 2003). There is also evidence of older duplication events. Different authors have suggested that from one to three or more additional partial duplication events may have occurred (Ziolkowski *et al.*, 2003; Blanc *et al.*, 2003; Bowers *et al.*, 2003; Raes *et al.*, 2003; Simillion *et al.*, 2002; Vision *et al.*, 2000).

Our laboratory has previously performed dating of *Arabidopsis* duplication events using a phylogenetic approach (Bowers *et al.*, 2003) similar to the one described in this paper. The results presented here (Table 1) use the new software tool described, incorporating more rigorous requirements for tree generation and interpretation. The analysis was performed using multiple sequence alignments of at least 35 amino acids. 100 bootstrap replicates were examined using protein parsimony, with PHYLIP's 'protpars' default implementation for scoring amino acid changes. Only trees with at least 80 (out of 100) bootstrap confidence support in the informative branch were used in the analysis.

The results of this dating analysis coincide with previous results (Bowers *et al.*, 2003), with a few notable exceptions. First, *Citrus* is included in the analysis, which provides another dating point along with the Malvaceae family to specify the most recent ( $\alpha$ ) duplication event. The  $\alpha$  event appears to have occurred after the divergence of these taxa, but before the divergence of Brassica and *Arabidopsis*. Dating results from short duplicate blocks generated by the  $\alpha$  event also confirm these results.

Second, this analysis places the subsequent  $\beta$  event at a more ancient point than previously predicted (Bowers *et al.*, 2003). Specifically, we find a relatively larger number of internal trees in the *Oryza* analysis, indicating that the  $\beta$  event may have occurred near or prior to the divergence of the monocots from the dicots (represented by *Arabidopsis*). In addition to the difference in rigor of analysis, these results also include

**Table 2.** Rice duplication dating results

Block	Pinaceae TaxId:3318	<i>Sorghum</i> TaxId:4557	<i>Hordeum</i> TaxId:4512	<i>O.minuta</i> TaxId:63629
1	0.025 (121)	<b>0.318 (88)</b>	<b>0.309 (97)</b>	0.234 (47)
2	0.015 (65)	<b>0.286 (42)</b>	<b>0.419 (31)</b>	<b>0.414 (29)</b>
3	0.000 (61)	<b>0.381 (42)</b>	<b>0.333 (54)</b>	<b>0.429 (21)</b>
4	0.038 (53)	<b>0.333 (33)</b>	<b>0.300 (40)</b>	<b>0.471 (17)</b>
5	0.059 (34)	<b>0.375 (24)</b>	<b>0.267 (30)</b>	0.714 (14)
8	0.091 (33)	<b>0.433 (30)</b>	<b>0.481 (27)</b>	<b>0.333 (9)</b>
Total	0.033 (427)	<b>0.314 (309)</b>	<b>0.302 (338)</b>	<b>0.368 (155)</b>

Values, shading and taxonomy ids are as explained in Table 1. Results are shown for all duplicated blocks with 10 or more comparisons done, and for the overall event. Ten duplicated blocks were used in the analysis. The ratios indicate that the examined duplication event occurred after the divergence of Pinaceae from the rice lineage but prior to the divergence of *Sorghum* and *Hordeum*.

nearly twice as much rice sequence data as our previous work (Bowers *et al.*, 2003) due to the current emphasis on genome sequencing of *Oryza*. This highlights the importance of considering database size in interpreting dating results, and also the tendency toward external trees in incomplete datasets.

## Rice

Having gained some insights into the evolutionary history of dicotyledonous plants (dicots) from the analysis of *Arabidopsis*, a logical next step is to explore the history of the other major branch of the plant family tree, the monocots, as exemplified by the emerging sequence of the *Oryza* (rice) genome. Two pictures of rice duplications have already emerged. One analysis describes rice duplications as being generated from aneuploidy events, involving only a subset of the rice chromosomes (Vandepoele *et al.*, 2003). Our own early analysis suggests a duplication event involving most of the genome (Paterson *et al.*, 2003).

We performed an analysis of the rice duplication data (Paterson *et al.*, 2003) using the described phylogenetic dating implementation. Predicted rice proteins were ordered using bacterial artificial chromosomes (BACs) physically mapped along rice chromosomes (Paterson *et al.*, 2003). Intra-genome BLAST comparisons were then conducted and

**Table 3.** In-depth examination of the timing of the *Arabidopsis*  $\beta$  duplication event relative to monocot divergence times

$\beta$ Block	Pinaceae TaxId:3318	<i>Sorghum</i> TaxId:4557	<i>Hordeum</i> TaxId:4512	<i>Oryza</i> TaxId:4527	Solanaceae TaxId:4070
1	0.087 (23)	0.062 (16)	0.000 (11)	0.048 (21)	0.200 (20)
2	0.000 (16)	0.000 (12)	0.000 (13)	0.154 (13)	0.125 (8)
4	0.118 (34)	0.259 (27)	0.243 (37)	0.367 (49)	0.444 (36)
10	0.000 (13)	0.071 (14)	0.071 (14)	0.179 (28)	0.455 (11)
Total	0.091 (243) <sup>A</sup>	0.137 (161) <sup>AB</sup>	0.151 (199) <sup>AB</sup>	0.214 (290) <sup>B</sup>	0.313 (211) <sup>B</sup>

Correlated ANOVA analysis between duplicate blocks (Bowers *et al.*, 2003) indicates a duplication that occurred prior to the divergence of the Solanaceae and after the divergence of Pinaceae. However, interpretation of the timing relative to multiple monocot databases provides ambiguous placement of the event. Letters (A,B) indicate statistically different groups ( $\alpha = 0.05$ ) based on Tukey's analysis with 250 random gaussian bootstrapped blocks based on  $\mu$  and  $\sigma$  estimates from the *Arabidopsis* block data.

visualized to select detectable regions of duplication (blocks) (Bowers *et al.*, 2003; Paterson *et al.*, 2003).

Conditions to prepare duplicates and homologs for phylogenetic analysis were as enumerated earlier for *Arabidopsis*. For phylogenetic analysis, 25 bootstrap replicates were examined with maximum likelihood, using the Jones–Taylor–Thornton amino acid substitution probabilities. Rate variability between sites was allowed, with a gamma model using one invariant site and two other sites with an  $\alpha$  value of 0.9. Only trees with at least 20 (out of 25) bootstrap confidence support were used in the analysis.

These results lead to two interesting insights (Table 2). First, the rice duplication event appears to be more ancient than its divergence from both *Sorghum* and *Hordeum* (barley). Because these represent very divergent members of the Poaceae (cereals), it appears likely that this duplication event predated the divergence of most or all of the Poaceae from a common ancestor.

Second, the results of a dataset for *Oryza minuta*, closely related to the organism in which we are analyzing duplicates (*Oryza sativa*), indicate the complications associated with the use of smaller databases. The presence of only 5286 available sequences leads to a subsequently larger amount of variability between blocks. Although the summed results agree with dating conclusions obtained from more distantly related species, we see two major deviations from our expectations—a low 0.234 ratio for block 1 and a high 0.714 ratio for block 5. These numbers are likely due to a combination of smaller numbers of comparisons done in each block and increased failure to select true homologs due to the limited representation of the database.

### Duplication events separating monocots and dicots

The dating results presented here are important to a new approach (Bowers *et al.*, 2003) for ameliorating previously encountered difficulties in monocot and dicot comparisons (Liu *et al.*, 2001). With the *Arabidopsis*  $\beta$  duplication event dating to near the monocot divergence, understanding the exact number of duplication events separating monocots and

dicots will be critical for attempting to detect homologous regions across this divide (Vandepoele *et al.*, 2002).

Dating of the *Arabidopsis*  $\beta$  event was explored in greater detail using additional monocot databases. *Oryza* (rice) results marginally support the hypothesis that the *Arabidopsis*  $\beta$  event predates monocot–dicot divergence (Table 3, block 4). However, results for *Sorghum* and *Hordeum* (barley) yield equivocal results (Table 3, blocks 2 and 10 and the statistical grouping of *Sorghum* and *Hordeum* with both *Oryza*, Solanaceae and Pinaceae). Accurate placement of the duplication event will require a more thorough understanding of the rice genome duplication event. More rigorous duplication analyses will be possible as work on *Oryza* genome finishing continues.

## DISCUSSION

Whole genome duplication is a major evolutionary force affecting the structure of many eukaryotic genomes. Understanding and characterizing these duplications is a complicated task due to rearrangements and deletions that obscure duplicated blocks and differential rates of sequence evolution that confound dating efforts.

A phylogenetic approach to dating duplication events has numerous advantages over alternative approaches to determine whether a genome-wide duplication event happened before or after divergence of different taxonomic lineages. The gene trees generated by the analysis are readily classifiable into two categories, and these categories correspond with two different models of speciation and duplication. The phylogenetic dating implementation described here provides a flexible tool for characterizing whole genome duplications.

*Arabidopsis* and *Oryza* duplications were examined using this implementation. The results showed the utility of the tool and raised critical questions about the timing of duplication events relative to the monocot–dicot divergence. Understanding the numerous duplication events which have occurred throughout the history of plant genome evolution will lead to better utilization of genomic comparisons across

species, facilitating the use of model species in understanding economically important crops.

Overall, this implementation provides a practical tool for researchers interested in elucidating the timing of whole genome duplication events.

## ACKNOWLEDGEMENTS

We thank J. Estill for valuable comments on the manuscript, and J. Kissinger and E. Kraemer for suggestions on the techniques. B.C. is supported by a Howard Hughes predoctoral fellowship.

## REFERENCES

- Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
- Blanc,G., Barakat,A., Guyot,R., Cooke,R. and Delseny,M. (2000) Extensive duplication and reshuffling in the Arabidopsis genome. *Plant Cell*, **12**, 1093–1101.
- Blanc,G., Hokamp,K. and Wolfe,K.H. (2003) A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res.*, **13**, 137–144.
- Bowers,J.E., Chapman,B.A., Rong,J. and Paterson,A.H. (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, **422**, 433–438.
- Calabrese,P.P., Chakravarty,S. and Vision,T.J. (2003) Fast identification and statistical evaluation of segmental homologies in comparative maps. *Bioinformatics*, **19**(Suppl. 1), I74–I80.
- Chenna,R., Sugawara,H., Koike,T., Lopez,R., Gibson,T.J., Higgins,D.G. and Thompson,J.D. (2003) Multiple sequence alignment with the clustal series of programs. *Nucleic Acids Res.*, **31**, 3497–3500.
- Conant,G.C. and Wagner,A. (2002) GenomeHistory: a software tool and its application to fully sequenced genomes. *Nucleic Acids Res.*, **30**, 3378–3386.
- Ermolaeva,M.D., Wu,M., Eisen,J.A. and Salzberg,S.L. (2003) The age of the *Arabidopsis thaliana* genome duplication. *Plant Mol. Biol.*, **51**, 859–866.
- Felsenstein,J. (2003). PHYLIP (Phylogeny Inference Package) version 3.6a3.
- Gu,Z., Nicolae,D., Lu,H.H. and Li,W.H. (2002) Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet.*, **18**, 609–613.
- 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.
- Langkjaer,R.B., Cliften,P.F., Johnston,M. and Piskur,J. (2003) Yeast genome duplication was followed by asynchronous differentiation of duplicated genes. *Nature*, **421**, 848–852.
- Liu,H., Sachidanandam,R. and Stein,L. (2001) Comparative genomics between rice and *Arabidopsis* shows scant collinearity in gene order. *Genome Res.*, **11**, 2020–2026.
- Lynch,M. and Conery,J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151–1155.
- Makova,K.D. and Li,W.H. (2003) Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome Res.*, **13**, 1638–1645.
- McLysaght,A., Hokamp,K. and Wolfe,K.H. (2002) Extensive genomic duplication during early chordate evolution. *Nat. Genet.*, **31**, 200–204.
- Mitchell-Olds,T. and Clauss,M.J. (2002) Plant evolutionary genomics. *Curr. Opin. Plant Biol.*, **5**, 74–79.
- Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
- Paterson,A., Bowers,J., Peterson,D., Estill,J. and Chapman,B. (2003) Structure and evolution of cereal genomes. *Curr. Opin. Genet. Devel.*, **13**, 644–650.
- Paterson,A.H., Bowers,J.E., Burow,M.D., Draye,X., Elsik,C.G., Jiang,C.X., Katsar,C.S., Lan,T.H., Lin,Y.R., Ming,R. and Wright,R.J. (2000) Comparative genomics of plant chromosomes. *Plant Cell*, **12**, 1523–1540.
- Raes,J., Vandepoele,K., Simillion,C., Saeys,Y. and Van de Peer,Y. (2003) Investigating ancient duplication events in the *Arabidopsis* genome. *J. Struct. Funct. Genomics*, **3**, 117–129.
- Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European molecular biology open software suite. *Trends Genet.*, **16**, 276–277.
- Ronquist,F. and Huelsenbeck,J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.
- Schmidt,R. (2002) Plant genome evolution: lessons from comparative genomics at the dna level. *Plant Mol. Biol.*, **48**, 21–37.
- Simillion,C., Vandepoele,K., Van Montagu,M.C., Zabeau,M. and Van de Peer,Y. (2002) The hidden duplication past of *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA*, **99**, 13627–13632.
- Vandepoele,K., Saeys,Y., Simillion,C., Raes,J. and Van De Peer,Y. (2002) The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between *Arabidopsis* and rice. *Genome Res.*, **12**, 1792–1801.
- Vandepoele,K., Simillion,C. and Van De Peer,Y. (2003) Evidence that rice and other cereals are ancient aneuploids. *Plant Cell*, **15**, 2192–2202.
- Vision,T.J., Brown,D.G. and Tanksley,S.D. (2000) The origins of genomic duplications in *Arabidopsis*. *Science*, **290**, 2114–2117.
- Wolfe,K.H. and Shields,D.C. (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, **387**, 708–713.
- Zhang,L., Vision,T.J. and Gaut,B.S. (2002) Patterns of nucleotide substitution among simultaneously duplicated gene pairs in *Arabidopsis thaliana*. *Mol. Biol. Evol.*, **19**, 1464–1473.
- Zhu,H., Kim,D.J., Baek,J.M., Choi,H.K., Ellis,L.C., Kuester,H., McCombie,W.R., Peng,H.M. and Cook,D.R. (2003) Syntenic relationships between *Medicago truncatula* and *Arabidopsis* reveal extensive divergence of genome organization. *Plant. Physiol.*, **131**, 1018–1026.
- Ziolkowski,P.A., Blanc,G. and Sadowski,J. (2003) Structural divergence of chromosomal segments that arose from successive duplication events in the *Arabidopsis* genome. *Nucleic Acids Res.*, **31**, 1339–1350.