Gene Order in Rosid Phylogeny, Inferred from Pairwise Syntenies among Extant Genomes

Chunfang Zheng[1] and David Sankoff[2]

[1] Département d'informatique et de recherche opérationnelle, Université de Montréal
[2] Department of Mathematics and Statistics, University of Ottawa

Abstract. Based on the gene order of four core eudicot genomes (cacao, castor bean, papaya and grapevine) that have escaped any recent whole genome duplication (WGD) events, and two others (poplar and cucumber) that descend from independent WGDs, we infer the ancestral gene order of the rosid clade and those of its main subgroups, the fabids and malvids. We use the gene order evidence to evaluate the hypothesis that the order Malpighiales belongs to the malvids rather than as traditionally assigned to the fabids. Our input data are pairwise synteny blocks derived from all 15 pairs of genomes. Our method involves the heuristic solutions of two hard combinatorial optimization problems, neither of which invokes any arbitrary thresholds, weights or other parameters. The first problem, based on the conflation of the pairwise syntenies, is the inference of disjoint sets of orthologous genes, at most one copy for each genome, and the second problem is the inference of the gene order at all ancestors simultaneously, minimizing the total number of genomic rearrangements over a given phylogeny.

1 Introduction

Despite a tradition of inferring common genomic structure among plants and despite plant biologists' interest in detecting synteny, e.g., [1,2], the automated ancestral genome reconstruction methods developed for animals [3,4,5,6] and yeasts [7,8,9,10,11] at the syntenic block or gene order levels, have yet to be applied to the recently sequenced plant genomes. Reasons for this include:

1. The relative recency of these data. Although almost twenty dicotyledon angiosperms have been sequenced and released, most of this has taken place in the last two years (at the time of writing) and the comparative genomics analysis has been reserved by the various sequencing consortia for their own first publication, often delayed for years following the initial data release.

2. Algorithms maximizing a well-defined objective function for reconstructing ancestors through the median constructions and other methods are computationally costly, increasing both with n, the number of genes orthologous across the genomes, and especially with $\frac{d}{n}$, where d is the number of rearrangements occurring along a branch of the tree.

3. Whole genome duplication (WGD), which is rife in the plant world, particularly among the angiosperms [12,13], sets up a comparability barrier between

J. Chen, J. Wang, and A. Zelikovsky (Eds.): ISBRA 2011, LNBI 6674, pp. 99–110, 2011.

those species descending from a WGD event and species in all other lineages originating before the event [2]. This is largely due to the process of duplicate gene reduction, eventually affecting most pairs of duplicate genes created by the WGD, which distributes the surviving members of duplicate pairs between two homeologous chromosomal segments in an unpredictable way [14,15,16], thus scrambling gene order and disrupting the phylogenetic signal. This difficulty is compounded by the residual duplicate gene pairs created by the WGD, complicating orthology identification essential for gene order comparison between species descended from the doubling event and those outside it.

4. Global reconstruction methods are initially designed to work under the assumption of identical gene complement across the genomes, but if we look at dicotyledons, for example, each time we increase the set of genomes being studied by one, the number of genes common to the whole set is reduced by approximately $\frac{1}{3}$. Even comparing six genomes, retaining only the genes common to all six, removes 85 % of the genes from each genome, almost completely spoiling the study as far as local syntenies are concerned.

Motivated in part by these issues, we have been developing an ancestral gene order reconstruction algorithm Pathgroups, capable of handling large plant genomes, including descendants of WGD events, as soon as they are released, using global optimization criteria, approached heuristically, but with well-understood performance properties [9,10]. The approach responds to the difficulties enumerated above as follows:

1. The software has been developed and tested with all the released and annotated dicotyledon genome sequences, even though "ethical" claims by sequencing consortia leaders discourage the publication of the results on the majority of them at this time. In this enterprise, we benefit from the up-to-date and well organized CoGe platform [1,17], with its database of thousands of genome sequences and its sophisticated, user-friendly SynMap facility for extraction of synteny blocks.

2. Pathgroups aims to rapidly reconstruct ancestral genomes according to a minimum total rearrangement count (using the DCJ metric [18]) along all the branches of a phylogenetic tree. Its speed is due to its heuristic approach (greedy search with look-ahead), which allows it to return a solution for values of $\frac{d}{n}$ where exact methods are no longer feasible. The implementation first produces a rapid initial solution of the "small phylogeny" problem (i.e., where the tree topology is given and the ancestral genomes are to be constructed), followed by an iterative improvement treating each ancestral node as a median problem (one unknown genome to be constructed on the basis of the three given adjacent genomes).

3. The comparability barrier erected by a WGD event is not completely impenetrable, even though gene order fractionation is further confounded by genome rearrangement events. The WGD-origin duplicate pairs remaining in the modern genome will contain much information about gene order in the ancestral diploid, immediately before WGD. The gene order information is retrievable through the method of *genome halving* [19], which is incorporated in a natural way into Pathgroups.

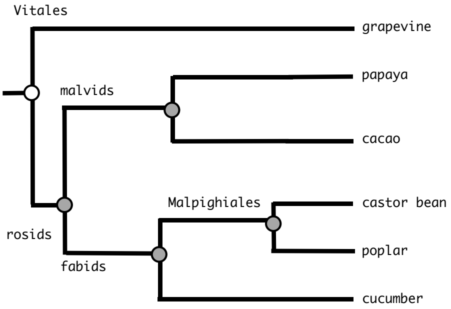

Fig. 1. Phylogenetic relationships among sequenced and non-embargoed eudicotyledon genomes (without regard for time scale). Poplar and cucumber each underwent WGD in their recent lineages. Shaded dots represent gene orders reconstructed here, including the rosid, fabid, malvid and Malpighiales ancestors.

4. One of the main technical contributions of this paper is the feature of PATHGROUPS that allows the genome complement of the input genomes to vary. Where the restriction to equal gene complement would lead to reconstructions involving only about 15 % of the genes, the new feature allows close to 100% of the genes with orthologs in at least two genomes to appear in the reconstructions. The other key innovation we introduce here is our "orthologs for multiple genomes" (OMG) method for combining the genes in the synteny block sets output by SYNMAP for pairs of genomes, into orthology sets containing at most one gene from every genome in the phylogeny.

Both the PATHGROUPS and the OMG procedures are parameter-free. There are no thresholds or other arbitrary settings. We argue that the the appropriate moment to tinker with such parameters is during the synteny block construction and not during the orthology set construction nor the ancestral genome reconstruction. A well-tuned synteny block method goes a long way to attenuate genome alignment problems due to paralogy. It is also the appropriate point to incorporate thresholds for declaring homology, since these depend on evolutionary divergence time, which is specific to pairs of genomes. Finally, the natural criteria for constructing pairwise syntenies do not extend in obvious ways to three or more genomes.

2 Six Eudicotyledon Sequences

There are presently almost twenty eudicotyledon genome sequences released. Removing all those that are embargoed by the sequencing consortia, all those who have undergone more than one WGD since the divergence of the eudicots from the other angiosperms, such as *Arabidopsis*, and some for which the gene annotations are not easily accessible leaves us the six depicted in Fig. 1, namely cacao [20], castor bean [21], cucumber [22], grapevine [23,24], papaya [25] and poplar [26]. Of the two main eudicot clades, asterids and rosids, only the latter is represented, as well as the order Vitales, considered the closest relative of the rosids [12,27]. Poplar and cucumber are the only two to have undergone ancestral WGD since the divergence of the grapevine.

3 Formal Background

A genome is a set of chromosomes, each chromosome consisting of a number of genes linearly ordered. The genes are all distinct and each has positive or negative polarity, indicating on which of the two DNA strands the gene is located.

Genomes can be rearranged through the accumulated operation of number of processes: inversion, reciprocal translocation, transposition, chromosome fusion and fission. These can all be subsumed under a single operation called double-cut-and-join which we do not describe here. For our purposes all we need is a formula due to Yancopoulos *et al.* [18], stated in Section 3.1 below, that gives the genomic distance, or length of a branch in a phylogeny, in terms of the minimum number of rearrangement operations needed to transform one genome into another.

3.1 Rearrangement Distance

The genomic distance $d(G_1, G_2)$ is a metric counting the number of rearrangement operations necessary to transform one multichromosomal gene order G_1 into another G_2, where both contain the same n genes. To calculate D efficiently, we use the breakpoint graph of G_1 and G_2, constructed as follows: For each genome, each gene g with a positive polarity is replaced by two vertices representing its two ends, i.e., by a "tail" vertex and a "head" vertex in the order g_t, g_h; for $-g$ we would put g_h, g_t. Each pair of successive genes in the gene order defines an adjacency, namely the pair of vertices that are adjacent in the vertex order thus induced.

If there are m genes on a chromosome, there are $2m$ vertices at this stage. The first and the last of these vertices are called telomeres. We convert all the telomeres in genome G_1 and G_2 into adjacencies with additional vertices all labelled T_1 or T_2, respectively. The breakpoint graph has a blue edge connecting the vertices in each adjacency in G_1 and a red edge for each adjacency in G_2. We make a cycle of any path ending in two T_1 or two T_2 vertices, connecting them by a red or blue edge, respectively, while for a path ending in a T_1 and a T_2, we collapse them to a single vertex denoted "T".

Each vertex is now incident to exactly one blue and one red edge. This bi-coloured graph decomposes uniquely into κ alternating cycles. If n' is the number of blue edges, then [18]:

$$d(G_1, G_2) = n' - \kappa. \tag{1}$$

3.2 The Median Problem and Small Phylogeny Problem

Let G_1, G_2 and G_3 be three genomes on the same set of n genes. *The rearrangement median problem is to find a genome M such that $d(G_1, M) + d(G_2, M) + d(G_3, M)$ is minimal.*

For a given unrooted binary tree T on N given genomes $G_1, G_2, \cdots, G_N$ (and thus with $N-2$ unknown ancestral genomes $M_1, M_2, \cdots, M_{N-2}$ and $2N-3$ branches), *the small phylogeny problem is to infer the ancestral genomes so that the total edge length of* T*, namely* $\sum_{XY \in E(T)} d(X,Y)$*, is minimal.*

The computational complexity of the median problem, which is just the small phylogeny problem with $N = 3$, is known to be NP-hard and hence so is that of the general small phylogeny problem.

4 The OMG Problem

4.1 Pairwise Orthologies

As justified in the Introduction, we construct sets of orthologous genes across the set of genomes by first identifying pairwise synteny blocks of genes. In our study, genomic data were obtained and homologies identified within synteny blocks, using the SynMap tool in CoGe [17,1]. This was applied to the six dicot genomes in CoGe shown in Fig. 1, i.e., to 15 pairs of genomes. We repeated all the analyses to be described here using the default parameters of SynMap, with minimum block size 1, 2, 3 and 5 genes.

4.2 Multi-genome Orthology Sets

The pairwise homologies SynMap provides for all 15 pairs of genomes constitute the set of edges E of the *homology graph* $H = (V, E)$, where V is the set of genes in any of the genomes participating in at least one homology relation.

The understanding of orthologous genes in two genomes as originating in a single gene in the most recent common ancestor of the two species, leads logically to transitivity as a necessary consequence. If gene x in genome X is orthologous both to gene y in genome Y and gene z in genome Z, then y and z must also be orthologous, even if SynMap does not detect any homology between y and z.

Ideally, then, all the genes in a connected component of H should be orthologous. Insofar as SynMap resolves all relations of paralogy, we should expect *at most* one gene from each genome in such an orthology set, or two for genomes that descend from a WGD event. We refer to such a set as *clean.*

In practice, gene x in genome X may be identified as homologous to both $y_1$and y_2 in genome Y. Or x in X is homologous both to gene y_1 in genome Y and gene z in genome Z, while z is also homologous to y_2. By transitivity, we again obtain that x is homologous to both $y_1$and y_2 in the same genome. While one gene being homologous to several paralogs in another genome is commonplace and meaningful, this should be relatively rare in the output from SynMap, where syntenic correspondence is a criterion for resolving paralogy. Aside from tandem duplicates, which do not interfere with gene order calculations, and duplicates stemming from WGD events, we consider duplicate homologs in the same genome, inferred directly by SynMap or indirectly by being members of the same connected component, as evidence of error or noise.

To "clean" a connected component with duplicate homologs in the same genome (or more than two in the case of a WGD descendant), we delete a number of edges, so that it decomposes into smaller connected components, each one of which is clean. To decide which edges to change, we define an objective function

$$F(X) = \sum_{i} \sum_{G, i \notin G} C_X(i, G), \tag{2}$$

where i ranges over all genes and G ranges over all genomes, and $C_X(i, G) = 1$ if there is exactly one one edge connecting i to any of the genes j in G (possibly two such edges if G descends from a WGD event), otherwise $C_X(i, G) = 0$.

A global optimum, maximizing F, the exact solution of the "orthology for multiple genomes" (OMG) problem, would be hard to compute; instead we chose edges in H one at a time, to delete, so that the increase in F is maximized. We continue in this greedy way until we obtain a graph H^* with all clean connected components . Note that F is designed to penalize the decrease of $C_X(i, G) = 1$ to $C_X(i, G) = 0$ by the removal of the only homology relation between gene i and some gene in genome G, thus avoiding the trivial solution where $H*$ contains no edges.

Note that it is neither practical nor necessary to deal with H in its entirety, with its hundred thousand or so edges. It suffices to do the cleaning on each connected component independently. Typically, this will contain only a few genes and very rarely more than 100. The output of the cleaning is generally a decomposition of the homology set into two or more smaller, clean, sets. These we consider our orthology sets to input into the gene order reconstruction step.

5 Pathgroups

Once we have our solution to the OMG problem on the set of pairwise syntenies, we can proceed to reconstruct the ancestral genomes. First, we briefly review the Pathgroups approach (previously detailed in [9,10]) as it applies to the median problem with three given genomes and one ancestor to be reconstructed, *all having the same gene complement.* The same principles apply to the simultaneous reconstruction of all the ancestors in the small phylogeny problem, and to the incorporation of genomes having previously undergone WGD.

We redefine a path to be any connected subgraph of a breakpoint graph, namely any connected part of a cycle. Initially, each blue edge in the given genomes is a path. A *fragment* is any set of genes connected by red edges in a linear order. The set of fragments represents the current state of the reconstruction procedure. Initially the set of fragments contains all the genes, but no red edges, so each gene is a fragment by itself.

The objective function for the small phylogeny problem consists of the sum of a number of genomic distances, one distance for every branch in the phylogeny. Each of these distances corresponds to a breakpoint graph. A given genome determines blue edges in one breakpoint graph, while the red edges correspond to the ancestral genome being constructed. For each such ancestor, *the red edges are identical in all the breakpoint graphs corresponding to distances to that ancestor.*

A pathgroup is a set of three paths, all beginning with the same vertex, one path from each partial breakpoint graph currently being constructed. Initially, there is one pathgroup for each vertex.

Our main algorithm aims to construct three breakpoint graphs with a maximum aggregate number of cycles. At each step it adds an identical red edge to each path in the pathgroup, altering all three breakpoint graphs. It is always possible to create one cycle, at least, by adding a red edge between the two ends of any one of the paths. The strategy is to create as many cycles as possible. If alternate choices of steps create the same number of cycles, we choose one that sets up the best configuration for the next step. In the simplest formulation, the pathgroups are prioritized, 1. by the maximum number of cycles that can be created within the group, without giving rise to circular chromosomes, and
2. for those pathgroups allowing equal numbers of cycles, by considering the maximum number of cycles that could be created in the next iteration of step 1, in any one pathgroup affected by the current choice.

By maintaining a list of pathgroups for each priority level, and a list of fragment endpoint pairs (initial and final), together with appropriate pointers, the algorithm requires $O(n)$ running time.

In the current implementation of Pathgroups, much greater accuracy, with little additional computational cost, is achieved by designing a refined set of 163 priorities, based on a two-step look-ahead greedy algorithm.

5.1 Inferring the Gene Content of Ancestral Genomes

The assumption of equal gene content simplifies the mathematics of Pathgroups and allows for rapid computation. Unfortunately it also drastically reduces the number of genes available for ancestral reconstruction, so that the method loses its utility when more than a few genomes are involved.

Allowing unequal gene complements in the data genomes, we have to decide how to construct the gene complement of the ancestors.

Using dynamic programming on unrooted trees, our assignment of genes to ancestors simply assures that if a gene is in at least two of the three adjacent nodes of an ancestral genome, it will be in that ancestor. If it is in less than two of the adjacent nodes, it will be absent from the ancestor.

5.2 Median and Small Phylogeny Problems with Unequal Genomes

To generalize our construction of the three breakpoint graphs for the median problem to the case of three unequal genomes, we set up the pathgroups much as before, and we use the same priority structure. Each pathgroup, however, may have three paths, as before, or only two paths, if the initial vertex comes from a gene absent from one of the leaves. Moreover, when one or two cycles are completed by drawing a red edge, this edge must be left out of the third breakpoint graph if the corresponding gene is missing from the third genome.

The consequence of this strategy is that some of the paths in the breakpoint graph will never be completed into cycles, impeding the search for optimality.

We could continue to search for cycles, but this would be computationally costly, spoiling the linear run time property of the algorithm.

The small phylogeny problem can be formulated and solved using the same principles as the median problem, as with the case of equal genomes. The solution, however, only serves as an initialization. As in [10], the solution can be improved by applying the median algorithm to each ancestral node in turn, based on the three neighbour nodes, and iterating until convergence. The new median is accepted if the sum of the three branch lengths is no greater than the existing one. This strategy is effective in avoiding local minima.

6 Results on Rosid Evolution

In the process of reconstructing the ancestors, we can also graphically demonstrate the great spread in genome rearrangement rates among the species studied, in particular the well-known conservatism of the grapevine genome, as illustrated by the branch lengths in Fig. 2.

It has been suggested recently that the order Malpighiales should be assigned to the malvids rather than the fabids [28]. In our results, the tree supporting this suggestion is indeed more parsimonious than the more traditional one. However, based on the limited number of genomes at our disposal, this is not conclusive.

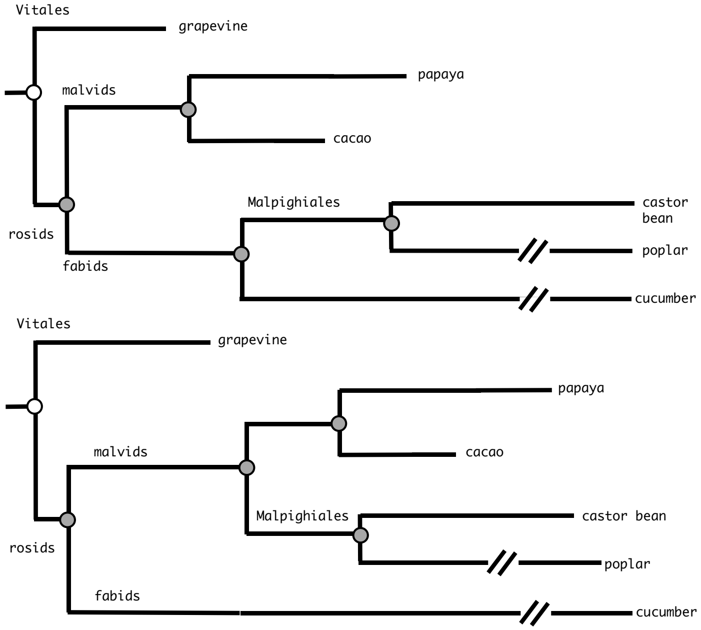

Fig. 2. Competing hypotheses for the phylogenetic assignment of the Malpighiales, with branch lengths proportional to genomic distances, following the reconstruction of the ancestral genomes with PATHGROUPS

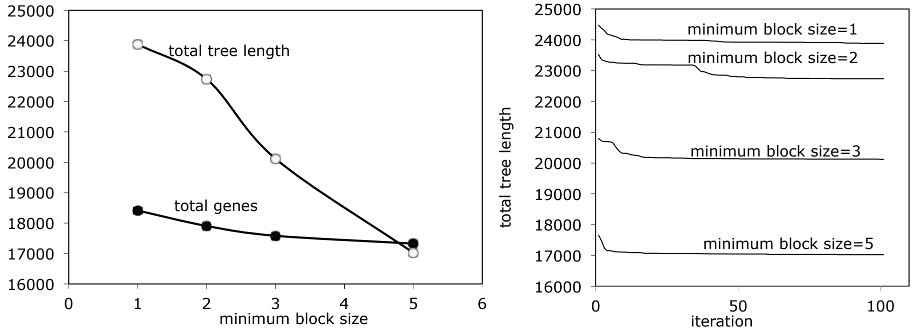

Fig. 3. Left: Effect of minimum block size on number of orthology sets and total tree length. Right: Convergence behaviour as a function of minimum block size.

6.1 Properties of the Solution as a Function of Synteny Block Size

To construct the trees in Fig. 2, from the 15 pairwise comparisons of the gene orders of the six dicot genomes, we identified some 18,000 sets of orthologs using SynMap and the OMG procedure. This varied surprisingly little as the minimum size for a synteny block was set to 1, 2, 3 or 5, as in Fig. 3. On the other hand, the total tree length was quite sensitive to minimum synteny block size. This can be interpreted in terms of risky orthology identifications for small block sizes.

Of the 18,000 orthology sets, the number of genes considered on each branch ranged from 12,000 to 15,000. When the minimum block size is 5, the typical branch length over the 11 branches of the tree (including one branch from each WGD descendant to its perfectly doubled ancestor plus one from that ancestor to a speciation node) is about 1600, so that $\frac{d}{n}$ is around 0.12, a low value for which simulations have shown Pathgroups to be rather accurate [10].

Fig. 3 shows the convergence behaviour as the set of medians algorithms is repeated at each ancestral node. Each iteration required about 8 minutes on a MacBook.

6.2 Block Validation

To what extent do the synteny blocks output by SynMap for a pair of genomes appear in the reconstructed ancestors on the path between these two genomes in the phylogeny? Answering this could validate the notion of syntenic conservation implicit in the block construction. Since our reconstructed ancestral genomes are not in the curated CoGe database (and are lacking the DNA sequence version required of items in the database), we cannot use SynMap to construct synteny blocks between modern and ancestor genomes. We can only see if the genes in the original pairwise syntenies tend to be colinear as well in the ancestor.

On the path connecting grapevine to cacao in the phylogeny in Fig. 1, there are two ancestors, the malvid ancestor and the rosid ancestor. There are 308 syntenic blocks containing at least 5 genes in the output of SynMap. A total of 11,229 genes are involved, of which 10,872 and 10,848 (97 %) are inferred to be in the malvid and rosid ancestor respectively.

Table 1. Integrity of cacao-grapevine syntenic blocks

	malvid ancestor		rosid ancestor	
synteny breaks	number	intra-block movement ≤ 1.0)	number	intra-block movement ≤ 1.0
0	140 (45%)	126 (90%)	153 (50%)	146 (95%)
1	66 (21%)	62 (94%)	64(21%)	58 (91%)
2	42 (14%)	39 (93%)	47(15%)	37 (79%)
> 2	60 (19%)	58 (97%)	44(14%)	38 (86%)

Table 1 shows that in each ancestor, roughly half of the blocks appear intact. This is indicated by the fact there are zero syntenic breaks in these blocks (no rearrangement breakpoints) and the average amount of relative movement of adjacent genes within these blocks is less than one gene to the left or right of its original position almost all of the time. Most of the other blocks are affected by one or two breaks, largely because the ancestors can be reconstructed with confidence by Pathgroups only in terms of a few hundred chromosomal fragments rather than intact chromosomes, for reasons given in Section 6.1. And it can be seen that the average shuffling of genes within these split blocks is little different from in the intact blocks.

7 Discussion and Future Work

We have developed a methodology for reconstructing ancestral gene orders in a phylogenetic tree, minimizing the number of genome rearrangements they imply over the entire tree. The input is the set of synteny blocks produced by SynMap for all pairs of genomes. The two steps in this method, OMG and Pathgroups, are parameter-free. Our method rapidly and accurately handles large data sets (tens of thousands of genes per genome, and potentially dozens of genomes). There is no requirement of equal gene complement.

For larger numbers of genomes, a problem would become the quadratic increase in the number of pairs of genomes, but this can be handled by SynMap only to pairs that are relatively close phylogenetically.

Future work will concentrate first on ways to complete cycles in the breakpoint graph which are currently left as paths, without substantially increasing computational complexity. This will increase the accuracy (optimality) of the results. Second, to increase the biological utility of the results, a post-processing component will be added to differentiate regions of confidence in the reconstructed genomes from regions of ambiguity.

Acknowledgments

Thanks to Victor A. Albert for advice, Eric Lyons for much help and Nadia El-Mabrouk for encouragement in this work. Research supported by a postdoctoral fellowship to CZ from the NSERC, and a Discovery grant to DS from the same agency. DS holds the Canada Research Chair in Mathematical Genomics.

References

1. Lyons, E., et al.: Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar and grape: CoGe with rosids. Plant Phys. 148, 1772–1781 (2008)
2. Tang, H., et al.: Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. Genome Res. 18, 1944–1954 (2008)
3. Murphy, W.J., et al.: Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. Science 309, 613–617 (2005)
4. Ma, J., et al.: Reconstructing contiguous regions of an ancestral genome. Genome Res. 16, 1557–1565 (2006)
5. Adam, Z., Sankoff, D.: The ABCs of MGR with DCJ. Evol. Bioinform. 4, 69–74 (2008)
6. Ouangraoua, A., Boyer, F., McPherson, A., Tannier, É., Chauve, C.: Prediction of Contiguous Regions in the Amniote Ancestral Genome. In: Salzberg, S.L., Warnow, T. (eds.) ISBRA 2009. LNCS, vol. 5542, pp. 173–185. Springer, Heidelberg (2009)
7. Gordon, J.L., Byrne, K.P., Wolfe, K.H.: Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern *Saccharomyces* cerevisiae genome. PLoS Genet. 5, 1000485 (2009)
8. Tannier, E.: Yeast ancestral genome reconstructions: The possibilities of computational methods. In: Ciccarelli, F.D., Miklós, I. (eds.) RECOMB-CG 2009. LNCS, vol. 5817, pp. 1–12. Springer, Heidelberg (2009)
9. Zheng, C.: Pathgroups, a dynamic data structure for genome reconstruction problems. Bioinformatics 26, 1587–1594 (2010)
10. Zheng, C., Sankoff, D.: On the Pathgroups approach to rapid small phylogeny. BMC Bioinformatics 12(Suppl 1), S4 (2011)
11. Bertrand, D., et al.: Reconstruction of ancestral genome subject to whole genome duplication, speciation, rearrangement and loss. In: Moulton, V., Singh, M. (eds.) WABI 2010. LNCS, vol. 6293, pp. 78–89. Springer, Heidelberg (2010)
12. Soltis, D.E., et al.: Polyploidy and angiosperm diversification. Am. J. Bot. 96, 336–348 (2009)
13. Burleigh, J.G., et al.: Locating large-scale gene duplication events through reconciled trees: implications for identifying ancient polyploidy events in plants. J. Comp. Biol. 16, 1071–1083 (2009)
14. Langham, R.A., et al.: Genomic duplication, fractionation and the origin of regulatory novelty. Genetics 166, 935–945 (2004)
15. Thomas, B.C., Pedersen, B., Freeling, M.: Following tetraploidy in an *Arabidopsis* ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. Genome Res. 16, 934–946 (2006)
16. Sankoff, D., Zheng, C., Zhu, Q.: The collapse of gene complement following whole genome duplication. BMC Genomics 11, 313 (2010)
17. Lyons, E., Freeling, M.: How to usefully compare homologous plant genes and chromosomes as DNA sequences. Plant J. 53, 661–673 (2008)
18. Yancopoulos, S., Attie, O., Friedberg, R.: Efficient sorting of genomic permutations by translocation, inversion, and block interchange. Bioinformatics 21, 3340–3346 (2005)

19. El-Mabrouk, N., Sankoff, D.: The reconstruction of doubled genomes. SIAM J. Comput. 32, 754–792 (2003)
20. Argout, X., et al.: The genome of *Theobroma cacao*. Nat. Genet. 43, 101–108 (2011)
21. Chan, A.P., et al.: Draft genome sequence of the oilseed species *Ricinus communis*. Nat. Biotechnol. 28, 951–956 (2010)
22. Haung, S., et al.: The genome of the cucumber, *Cucumis sativus* L. Nat. Genet. 41, 1275–1281 (2010)
23. Jaillon, O., et al.: The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449, 463–467 (2007)
24. Velasco, R., et al.: A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PLoS ONE 2, e1326 (2007)
25. Ming, R., et al.: The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). Nature 452, 991–996 (2008)
26. Tuskan, G.A., et al.: The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science 313, 1596–1604 (2006)
27. Forest, F., Chase, M.W.: Eudicots. In: Hedges, S.B., Kumar, S. (eds.) The Timetree of Life, pp. 169–176. Oxford University Press, Oxford (2009)
28. Shulaev, V., et al.: The genome of woodland strawberry (*Fragaria vesca*). Nat. Genet. 43, 109–116 (2011)