Ancient angiosperm hexaploidy meets gene order reconstruction of the eudicot ancestor

Chunfang Zheng[1], Victor A. Albert[2], Eric Lyons[3], and David Sankoff[1]

[1] Department of Mathematics and Statistics, University of Ottawa
[2] Department of Biology, University at Buffalo
[3] iPlant, Department of Plant Sciences, University of Arizona

Abstract. We propose a protocol for the reconstruction and analysis of the post-polyploidization ancestor of a set of present-day descendant genomes. Our method, as applied to the post-hexaploidization ancestor of the core eudicot flowering plants, is based solely on the automated reconstruction of ancestral gene order, starting from all the orthologs obtained for each pair of data genomes, harmonized into complete sets of orthologs across multiple genomes.

We make use two independent approaches to infer the gene order at the root of the rosid phylogeny. One is of PATHGROUPS method for the small phylogeny problem, motivated by a chromosomal rearrangements (inversions, translocations, transpositions) model of evolution. The second is a maximum weighted matching technique based only on gene adjacencies in the data genomes in their phylogenetic context.

Aside from the confirmation of the triplication documented in the original grapevine and cacao genome sequence publications, we detect, in our reconstructed ancestor genome preceding the radiation of the core eudicots, the three regions corresponding to each of the seven original chromosomes hypothesized in these papers.

Keywords:

comparative genomics
genome rearrangement
plant genomes
phylogeny

Contact: sankoff@uottawa.ca

1 Introduction

The publication of the grapevine genome sequence in 2007 [1] showed the existence of an early hexaploidization event with clear traces in all the dicot genomes sequenced up to that time, and its absence from the monocot genome of rice. Since then, this event has been characterized in most detail with the publication of the cacao genome sequence [2]. These basis for this work was the observation of seven sets of three corresponding regions in each genome, each region (fragmented in some of the genomes) containing many genes homologous with genes in one, or occasionally both, of the other two corresponding regions. Meticulous work identified the boundaries of these regions and a credible history of chromosome fusions and other major rearrangements leading to the modern genomes from an ancestral $N = 3 \times 7 = 21$-chromosome hexaploid.

In this paper we propose a formal protocol for the detection, analysis and characterization of a post-hexaploidization ancestor of the core eudicots, before the radiation into the major clades, mainly rosids, malvids and the order Vitales. Our method, which should be applicable to other post-polyploidization genomes, is based solely on the automated reconstruction of ancestral gene order, starting from complete sets of orthologs inferred for pairs of genomes by the SynMap procedure [3–5], and harmonized into sets of orthologs across multiple genomes by the OMG! technique [6]. Note that this is not an attempt to reconstruct the ancestor at the moment of polyploidization, as is done with genome halving [7] or genome aliquoting algorithms [8], but to reconstruct the rediploidized, fractionated and rearranged descendant of that polyploid that is the most recent common ancestor of the data genomes.

There is a growing literature on gene order reconstruction in the phylogenetic context e.g., [9–14]; we will make use of our recently improved version of the Pathgroups [15] method to infer the gene order at the root of the rosid phylogeny. To provide a methodologically independent validation we also employ a maximum weighted matching approach to reconstruct this ancestor.

Before actually reconstructing the ancestral genome, involving procedures independent of the hexaploidization hypothesis, we quantitatively analyze the internal homologies of the six genomes, documenting the pervasive pattern of triples of syntenic blocks that are the signature of hexaploidization. We also develop a rough model for the fractionation process to account for the proportion of gene triples, gene pairs, and single copy genes within the genomes.

Once the reconstructions are obtained, we map the triplication evidence in the data genomes to the reconstructions, and then show a very clear pattern of triplicated regions consistent with each of the genomes. These regions corresponding to each of the seven hypothesized original chromosomes proposed in [1] and [2].

2 Gene triplicates resulting from the hexaploidization event

We analyzed five core eudicot genomes with no evidence of whole genome duplication (WGD) since the origin of this clade (∼110 Mya), namely *Vitis vinifera* (grapevine) [1, 16], *Carica papaya* (papaya) [17], *Ricinus communis* (castor bean) [18], *Theobroma cacao* (cacao) [2] and *Fragaria vesca* (strawberry) [19], as well as *Populus trichocarpa* (poplar) [20], which has undergone a relatively recent WGD (∼70 Mya).

Using SynMap to locate all synteny blocks of minimum size 5 in a self-comparison of each genome revealed thousands of pairs of genes, whose gene similarities are plotted in Fig. 1. All of the genomes show a clear peak at 70 % ± 3%. In addition, poplar showed a larger peak at 91 %, reflecting the WGD mentioned above.

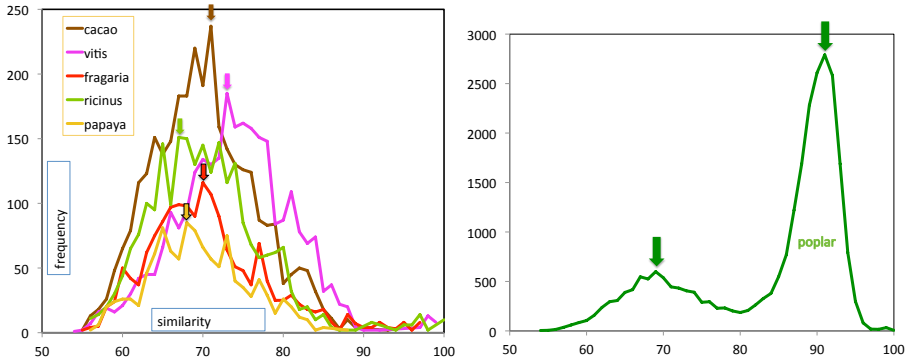

Fig. 1. Distribution of similarities in gene pairs in six genomes, showing triplication peaks between 67 % and 73%, with poplar duplication peak at 91%.

Dot-plots for the self-comparison of grapevine and of cacao clearly show a pattern of pairwise homologies falling largely into 21 groups according to the chromosomal location of the two genes [1, 2]. These groups are subdivided into 7 triples of regions, where there are numerous homologies among all three pairs of regions. For the other four genomes, the greater degree of rearrangement within and among the chromosomes means that we observe many more than 21 groups of synteny blocks in the SynMap output involving smaller regions of the genomes. But these regions can for the most part still be grouped into triples, where each of the three pairs of regions in a triple generally evidences at least a small number of homologous gene pairs.

As we will discuss in Section 3 below, there are very few *triples* of homologous genes within each genome, due to the rapid *fractionation* process. What is left

after this loss of iso-functional paralogs is an *interleaving* [21] pattern of pairwise homologies among three regions.

Adopting the seven-colour scheme in [2], if we colour each individual gene involved in one or more pairs in a genome by its colour in grapevine and/or cacao (there are virtually no conflicts between the two), we find that each of triple of regions corresponds to one of the seven triples in grapevine and cacao. We colour the block by the colour of its genes – there being no ambiguity in 99 % of the cases.

As we can see in Table 1, most of the duplicate genes in synteny blocks are coloured. The remaining genes and blocks represent either duplication pre- or post-hexaploidization, or originally triplicated genes not identified as such through our methodology. For virtually all of the coloured blocks, there is only one colour. Importantly, all of the gene pairs with two coloured genes involve two separate regions of the same colour. These observations are only consistent with a hexaploidization event as the origin of most of the gene pairs within these eudicot genomes.

	genome					
	cacao	Vitis	ricinus	Fragaria	papaya	populus
genes in synteny blocks	2878	2435	2147	1576	1098	5174
% genes coloured	78	84	58	52	60	52
in coloured blocks:						
% genes coloured	82	85	61	58	64	59
number of blocks	362	247	267	225	160	992
% one colour	95	97	93	82	93	76
% no colour	5	2	7	14	6	22
% different colour	0	2	1	4	1	2
number of gene pairs	3143	2709	2347	1638	1110	7242
% in coloured blocks	97	99	95	90	94	90
% same colour, different region	70	74	51	42	53	45
% same colour, same region	0	0	0	0	0	1
% one gene coloured	20	22	17	21	15	15
% neither coloured	10	4	32	37	31	38

Table 1. Distribution of colours among synteny blocks

3 Fractionation of duplicates and triplicates

The number of genes in the common ancestor of the six genomes studied here is unknown; in [2] it is implied that the pre-hexaploidization genome might have contained around 10,000 genes, so that the hexaploid itself would have had 30,000. Between the hexaploidization at time t_0 and the radiation of the core eudicots at time t_1, there would have been considerable loss of duplicates.

And after the radiation, this loss process would be continued independently in each lineage. In Table 2, we see that the number of triplets of homologous genes remaining today, at time t_2 range from 34 – 180, averaging about 110. The number of pairs ranges from 500 – 1200, averaging about 820.

Let p represent the probability that a redundant (duplicate or triplicate) gene would be lost during the time span from t_0 to t_1, from hexaploidization to the radiation, had there been no functional constraints. However, we can assume that the event that all three copies were lost was prohibited. Adapting a derivation for a different scenario of compound polyploidization [22], the probability that

1. all three genes survived is $\frac{(1-p)^3}{1-p^3}$.
2. two of the three survived is $\frac{3p(1-p)^2}{1-p^3}$, and
3. only one survived is $\frac{3p^2(1-p)}{1-p^3}$.

Similarly, let q represent the probability that a redundant (duplicate or triplicate) gene would be lost during the time span from t_1 to t_2, from the radiation to the present, were there no functional constraints. Then the probability

4. a triplet would still survive is $\frac{(1-p)^3}{1-p^3}\frac{(1-q)^2}{1-q^2}$.
5. an original triplet would manifest as a pair is $\frac{(1-p)^3}{1-p^3}\frac{2q(1-q)}{1-q^2} + \frac{3p(1-p)^2}{1-p^3}\frac{(1-q)^2}{1-q^2}$, and
6. an original triplet would be reduced to a single copy is $\frac{(1-p)^3}{1-p^3}\frac{3q(1-q)^2}{1-q^3} + \frac{3p(1-p)^2}{1-p^3}\frac{2q}{1-q^2} + \frac{3p^2(1-p)}{1-p^3}$.

	frequencies of gene family sizes					
gene family size	cacao	Vitis	ricinus	Fragaria	papaya	populus
2	1159	965	885	672	496	813 (3-4)
3	180	151	123	63	34	116 (5-6)
≥ 4	5	12	2	10	1	5 (> 6)

Table 2. Reduction of triplicates and duplicates through fractionation

Solving the equations in 4. and 5. above, assuming 10,000 original triplets, 180 cacao triplets and 1159 cacao pairs, leads to values of $p = 0.48$ and $q = 0.80$. It is reassuring that p is smaller than than q, since it represents loss rates over a presumably much shorter interval $t_1 - t_0$ than the $t_2 - t_1$ associated with q. Vitis gives similar results, but the number of triples detected in the other genomes are too small to fit the model.

4 Gene order reconstruction

In the combinatorial optimization framework, gene order reconstruction attempts to minimize the overall cost of a given phylogenetic tree, with given genomes at

the leaves, namely by finding gene orders at all the ancestral nodes such that the sum of the rearrangement distance along the branches of the tree is as small as possible. Gene order phylogeny problems are difficult, with exact algorithms bogging down for realistic instances and heuristics risking seriously sub-optimal solutions. Nevertheless there are quite a number of methods available that give good results in specific contexts, as cited in Section 1.

In the present work, we use two reconstruction methods. One is the PATHGROUPS approach [11, 12, 15], which has been developed in the context of plant genomics, with its highly variable gene content and recurrent patterns of WGD. Since this has been detailed elsewhere, we will give only the briefest sketch here.

Since gene order is largely a matter of gene adjacency, and since gene order data with strandedness information can be represented as a graph matching where the vertices are gene ends and the edges are adjacencies, we also use a MAXIMUM WEIGHTED MATCHING (MWM) procedure as an alternative way of reconstructing the ancestral gene order.

In both methods, what is reconstructed is usually not whole chromosomal gene orders, but large fragments of these orders, containing anything from single isolated genes to half a chromosome or more. In the present study, the fifty or sixty largest fragments contain over 95 % of the genes.

4.1 Orthology for multiple genomes

The input to the reconstruction methods is the complete gene order for each of the data genomes, although some missing data and unassembled contigs can be tolerated with little difficulty. The main problem is to construct s complete as possible orthology sets containing at most one gene per genome (two for ancient tetraploids). We obtain orthologs for pairs of genomes from SYNMAP; we can have a high degree of confidence in these inferences. Conceptually, the orthology relation should be transitive; if two genes in genome A and genome B originate in a single gene in their common ancestor, and if the same gene in Genome B is orthologous to one in Genome C, then logically, the genes in Genome A and Genome C should also be orthologous. In practice, fluctuation in divergence rates may result in the latter orthology slipping below he threshold of detection; worse a low degree of false orthology inference is quickly compounded as the number of genomes being compared grows, so that simply conflating pairwise orthology sets to produce a master graph H, where the vertices represent genes in specific genomes, and the edges represent orthology inferences from SYNMAP, leads to an increasing number of conflicts, i.e., connected components of H, meant to be coherent orthology sets, but containing false paralogs. (There should be *no* paralogs except WGD "ohnologs".)

We use the OMG procedure [6] to judiciously eliminate misleading edges from H to resolve these conflicts. Suppose $G = (V_G, E_G)$ is a connected component of H with (false) paralogs in one or more genomes. We delete a subset of edges $E' \subset E_G$, so that the remaining graph Q decomposes into smaller connected components, $Q = Q_1 \cup \cdots \cup Q_t$, where each $Q_i = (V_i, E_i)$ is free from paralogy. To decide which edges to delete, we define an objective function to be the total

number of edges in the transitive closure of Q, i.e., in all the cliques generated by the components Q_i. In other words, we seek to maximize $\sum_1^t \binom{|E_i|}{2}$, where $Q_i = (V_i, E_i)$. The details of our heuristic solution are given in [6].

4.2 PATHGROUPS

The adjacencies between two gene ends define a "matching" in the graph where the vertices are the ends. For two genomes (red and blue, say) sharing the same genes (i.e., orthologous genes labelled by the same name or identifier) the two matchings define a graph (the *breakpoint* graph) consisting of a number of alternating colour cycles. (It is necessary to complete some alternating color paths from a telomeric gene to another telomeric gene by adding an edge or superimposing one telomere on the other.) Then the number of chromosomal rearrangements necessary to convert one of the genomes into the other is just $b - c$ where b and c are the number of blue edges and the number of cycles, respectively [23]. This number serves as a measure of the distance between the red and blue gene orders.

PATHGROUPS is essentially a dynamic data structure that constructs all the breakpoint graphs for all the branches in the tree simultaneously. It allows a rapid greedy algorithm with look-ahead that remains accurate as long as neighboring genomes on the tree are not too rearranged with respect to each other [12].

4.3 MWM

Where the PATHGROUPS approach incorporates adjacency information via the cycles in the breakpoint graph, and where it reconstructs all the non-leaf nodes in the phylogeny, we may take a shortcut to ancestral gene order reconstruction by simply combining all the matches in all the data genomes in some principled way to produce the adjacencies (the matching) of the ancestral gene order.

The most obvious approach is to weight the adjacencies according to how many data genomes they appear in, and then find a maximum weight matching. The best known polynomial-time exact algorithm for this is Edmond's "path, trees and flower" method, but we use a simpler algorithm for which code is easily available [24].

There is a bias built in to this direct approach to weighting due to how the genomes are selected; if many closely related genomes are included in one lineage while another lineage is sparsely represented, counting all adjacencies will favor the former over the latter. Our solution to this is to consider the subtrees defined by the three branches incident to the root – in this study, the malvid and fabid subclades of the rosids, plus Vitis. If an adjacency occurs anywhere in the subtree, it adds one to its weight. Thus adjacencies are weighted 0,1,2 or 3.

Another problems with MWM is that it can produce circular chromosomes. Indeed, typically a quarter to a half of the reconstructed chromosomal gene orders are circular. This is not a major problem, however, since deleting a single weight 1 edge can be deleted to linearize a chromosome, something which displaces the reconstruction a minuscule distance from the optimum.

5 Triplicate regions in the ancestor

Our reconstruction of the ancestral genomes was carried out independently of our study of triplicated regions. What is the result of mapping the "coloured" regions on the reconstructed ancestor. Fig. 2 answers this questions for the PATHGROUPS. Each row in a coloured cell of the table lists the chromosomal fragments that contain genes in that colour. Basically, there is a near perfect correspondence across all the genomes. This is partly circular since all the genomes except cacao and Vitis were assigned colors and "subcolours" with the help of these two relatively unrearranged genomes. But it does not explain that *all* the rows contain almost all the same fragments. And of course the cacao and Vitis genomes were themselves coloured independently.

	Vitis	cacao	Ricinus	poplar	Fragaria	papaya
b1(chr1)	0 1 44 69 265	0 1 69	0 1 44 69 265	0 1 29 44 69 265	0 265	0 1 265
b2(chr14b)	44 102 133	44 102 133	44 102 133	44 102 133	44 102 133	44 102 133
b3(chr17)	1 102 133	102 133	1 102 133	102 133 262	102 133	1 102 133
b1(chr2)	204 245	204 245	204 245	204	204	204
b2(chr15)	69	69	69	69	69	69
b3(chr16)	102	102	102	102	102	102
b1(chr3)	16 81	16 81	16 81	16 81 159	81	81
b2(chr4b)	16 73	16	16 73	16 73	16	16
b3(chr7b)	16 39 73	16 39 73	16 39	16 39 73	16 39 73	16 39 73
b4(chr18)	66 79 81 154	66 79 81 154	66 79 81 154	66 79 81 154	66 79 81	66 79 81 154
b1(chr4a)	129	129	129	129	129	129
b2(chr9)	44 229	44 229	44 229	44 229	44 229	229
b3(chr11)	129	129	129	129	129	129
b1(chr5)	69 105	69 105	69 105	69 105	69 105	69 105
b2(chr7a)	69	69	69	69	69	69
b3(chr14a)	79 102 115 133	79 102 115	79 102 115	79 102 133	79 133	102
b1(chr6)	102 117	102 117	102	102 117	102 117	102
b2(chr8)	69 97 102 105	69 105	105	69 97 102 105 217	69 105	105
b3(chr13)	33 53 54 123	33 54 123	33	33 54 123	33 54	33 54
b1(chr10)	51 171	51 171	51 171	51 171	51 171	51 171
b2(chr12)	20 138 250	20 250	20 250	20 250	20 138 250	20 250
b3(chr19)	137 177	137	137	137	137	137

Fig. 2. Triplicated fragments of pathgroups ancestor, from evidence in six eudicots.

Repeating the same analysis on the reconstruction by MWM gives the entirely parallel results in Fig. 3, although the chromosomal fragments in this reconstruction are different from those in the PATHGROUPS reconstruction.

6 Coherence of alternative constructions of the ancestral gene order

We have seen that the PATHGROUPS and MWM approaches produce ancestral orders that are largely mutually confirmatory about the content of the triplicated regions. In this section, we discuss how similar the two sets of results are in terms of large-scale order.

	Vitis	cacao	Ricinus	poplar	Fragaria	papaya
b1(chr1)	0 57	0 57	0 57	0 57 65	0 57	0 57
b2(chr14b)	65	65	65	7 65	65	65
b3(chr17)	7	7	7	7	7	7
b1(chr2)	10 75	10 75	10 75	10	10	10
b2(chr15)	117	117	117	117	117	117
b3(chr16)	5 7 48	5 7 48	5 48	5	5	5
b1(chr3)	12 23 70	12 23 70	12 70	12 23 70	12 23	12 23
b2(chr4b)	10 12	10	10 12	10 12	10	10
b3(chr7b)	10 12 35 84	10 12 35	10	10 12 35 84	10 12	10 12 35 84
b4(chr18)	2 106	2 106	2 106	2 106	2 106	2 106
b1(chr4a)	70	70	70	70	70	70
b2(chr9)	26 27 60 69	27 60 69	27 60 69	27 60 69	26 27 60 69	27 60 69
b3(chr11)	70	70	70	70	70	70
b1(chr5)	0 44 82	44 82	0 44	0 44 82	0 44 82	44 82
b2(chr7a)	56 117	56 117	56 117	56 117	56 117	56 117
b3(chr14a)	49 65 66	49 66	49 66	49 65 66	65 66	49
b1(chr6)	4 8 61 68 76 81	4 8 61 68 76 81	68 76	4 8 61 68 76 81	4 8 61 68 76 81	4 8 68 76 81
b2(chr8)	8 9 49 75 76	8 9 75 76	9 76	8 9 49 75 76	8 9 76	9 76
b3(chr13)	33 41 68	33 68	33	33 68	33	33
b1(chr10)	8	8	8	8	8	8
b2(chr12)	13 56 116	13 56	13 116	13 56 116	13 56 116	13 56 116
b3(chr19)	8 17 51	17 51	17	13 17 51	17 51	17 51

Fig. 3. Triplicated fragments of MWM ancestor, from evidence in six eudicots.

In each reconstruction, we have fifty to sixty chromosomal fragments containing multiple genes. Many of these contain genes from different triplicated regions. Within each reconstruction there is no inherent order among the fragments, but we may ask if the commonalities and differences between the two reconstructions can help order the fragments.

Let be the PATHGROUPS fragments be numbered arbitrarily $i = 1, 2, \ldots$ and initially ordered by $pos1(i) = i$ and the MWM fragments numbered $j = 1, 2, \ldots$ and initially ordered by $pos1(j) = j$. Let $m(i, j)$ be the number of genes in common between fragments i and j in the two reconstructions. We define the crossing number

$$X = \sum_{i,j} m(i_1, j_1)m(i_2, j_2)\chi(i_1, i_2, j_1, j_2) \tag{1}$$

where $\chi(i_1, i_2, j_1, j_2) = 1$ if $pos1(i_1 < i_2)$ and $pos2(j_1 \geq j_2)$ and $\chi(i_1, i_2, j_1, j_2) = 0$, otherwise, to be the degree of discordance in the two orders. By minimizing X, we optimally order the two reconstructions with respect to each other. This can be carried out in a heuristic manner by iteratively moving a chromosome in either ancestor to a position in the order that decreases the crossing number. Repeating the procedure with different initializations produces slightly different orders, and the best one is selected for further analysis.

Fig. 4 shows that despite the fact that triplication information was used neither in the PATHGROUPS nor in the MWM procedures, most of the 21 triplicated regions or located solely or largely on a single fragment in one of the reconstructions and, tellingly, on two or more fragments that are ordered close together in the other reconstruction. Thus each reconstruction is serving to "assemble" the fragments of the other into whole chromosomes or at least larger chromosomal fragments, and this assembly is coherent with a pattern of each triplicated region

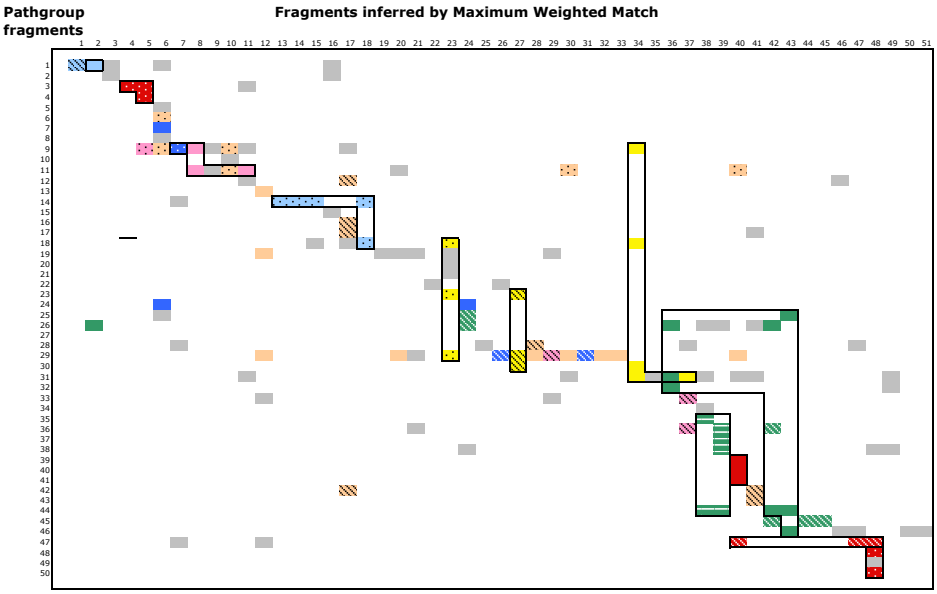

Fig. 4. Genes in common of in PATHGROUPS fragments (listed vertically) and MWM fragments (horizontal axis), coloured by inferred triplication region.

being conserved on a single chromosome, or rearranged into a small number of chromosomes.

7 Conclusions

As a preliminary to ancestral genome reconstruction, we have quantitatively documented the details of the triplicated regions in the six eudicots under study, and suggested a model for fractionation in two steps, one pre-radiation and the other post-radiation.

We found labeling of the reconstructed genomes by the "colors" of the 21 triplicate regions to be remarkably coherent and complete across all the data genomes.

The minimization of crossing number provides a way to partially assemble each fragmented reconstructed ancestral gene order, and also orders the fragments in a way rather consistent with the hexaploidization model.

References

1. Jaillon O *et al.* (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**: 463–7.
2. Argout X *et al.* (2011) The genome of *Theobroma cacao*. *Nat Genet* **43**: 101–8.
3. Lyons E, Freeling M (2008) How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J* **53**: 661–73.
4. Lyons E *et al.* (2008) Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar and grape: CoGe with rosids. *Plant Phys* **148**:1772–81.
5. Tang H *et al.* (2011) Screening synteny blocks in pairwise genome comparisons through integer programming. *BMC Bioinformatics* **12**:102.
6. Zheng C, Swenson KM, Lyons E, Sankoff D (2011) OMG! Orthologs in multiple genomes – competing graph-theoretical formulations. In Przytycka TM, Sagot M-F (eds), *Algorithms in Bioinformatics (WABI). Eleventh International Workshop, Lect Notes Comput Sc*, **6833**: 364-375.
7. El-Mabrouk N, Sankoff D (2003) The reconstruction of doubled genomes. *SIAM Journal on Computing* **32**: 754–792.
8. Warren R, Sankoff D (2011) Genome aliquoting revisited. *Journal of Computational Biology* **18**:1065–75.
9. Adam Z, Sankoff D (2008) The ABCs of MGR with DCJ. *Evol Bioinform*, **4**: 69–74.
10. Tannier E (2009) Yeast ancestral genome reconstructions: the possibilities of computational methods. In Ciccarelli FD, Miklós I (eds), *Comparative Genomics (RECOMB CG). Seventh Annual Workshop, Lect Notes Comput Sc*, **5817**: 1–12.
11. Zheng C (2010) PATHGROUPS, a dynamic data structure for genome reconstruction problems. *Bioinformatics* **26**: 1587–94.
12. Zheng C, Sankoff D (2011) On the PATHGROUPS approach to rapid small phylogeny. *BMC Bioinformatics* **12** (Suppl 1): S4.
13. Xu, AW (2010) On exploring genome rearrangement phylogenetic patterns. In Tannier E (ed), *Comparative Genomics (RECOMB CG). Eighth Annual Workshop, Lect Notes Comput Sc*, **6398**: 121-136.
14. Kovác J, Brejová B, Vinar T (2011) A practical algorithm for ancestral rearrangement reconstruction. In Przytycka TM, Sagot M-F (eds), *Algorithms in Bioinformatics (WABI). Eleventh International Workshop, Lect Notes Comput Sc*, **6833**: 163–174.
15. Zheng C, Sankoff D (2011) Gene order in Rosid phylogeny, inferred from pairwise syntenies among extant genomes. In Chen J, Wang J, Zelikovsky A (eds), *Bioinformatics Research and Applications, Seventh International Symposium, Lect Notes Comput Sc*, **6674**: 99–110.
16. Velasco R *et al.* (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS ONE* 2, e1326.
17. Ming R *et al.* (2008) The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). *Nature* 452, 991–9966. http://asgpb.mhpcc.hawaii.edu
18. Chan AP *et al.* (2010) Draft genome sequence of the oilseed species Ricinus communis. *Nat Biotechnol* doi:10.1038/nbt.1674.
19. Shulaev V *et al.* (2011) The genome of woodland strawberry (*Fragaria vesca*). *Nat Genetics* **43**: 109–16.

20. Tuskan GA *et al.* (2006) The genome of black cottonwood, Populus trichocarpa (Torr. & Gray) *Science* 313,1596–1604.
21. Kellis M, Birren BW, Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* **428**:617–24.
22. Nadeau JH, Sankoff D (1997) Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. *Genetics* **147**: 1259–1266.
23. Yancopoulos S, Attie O, Friedberg R (2005) Efficient sorting of genomic permutations by translocation, inversion, and block interchange. *Bioinformatics* **21**: 3340–6.
24. Galil Z (1986) Efficient algorithms for finding maximum matching in graphs. *ACM Computing Surveys* **18**: 23-38.