Computational Biology and Chemistry 50 (2014) 75–81

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

Practical halving; the *Nelumbo nucifera* evidence on early eudicot evolution

Chunfang Zheng, David Sankoff*

Department of Mathematics and Statistics, University of Ottawa, 585 King Edward Avenue, Ottawa, Canada K1N 6N5

ARTICLE INFO

Article history:
Accepted 16 January 2014
Available online 26 January 2014

Keywords:
Sacred lotus
Eudicots
Flowering plants
Evolution
Genomics
Rearrangement
Polyploidization

ABSTRACT

We present a stepwise optimal genome halving algorithm designed for large eukaryote genomes with largely single-copy genes, taking advantage of a signature pattern of paralog distribution in ancient polyploids. This is applied to the genome of *Nelumbo nucifera*, the sacred lotus, which is the descendant of a duplicated basal eudicot genome. In concert with the reconstructed ancestor of the grape, we investigate early events in eudicot evolution and show that the chromosome number of the common ancestor of lotus and grape was likely between 5 and 7. We show that the duplication of the ancestor of lotus and the triplication of the ancestor of grape were not closely preceded by any additional such event before the divergence of their two lineages.

1. Introduction

The published genome sequences of flowering plants show that whole genome duplication or triplication events occurred in all lineages leading to modern species, and occurred many times in cases such as *Arabidopsis* (Blanc and Hokamp, 2003) and *Utricularia* (Ibarra-Laclette et al., 2013). These events at first create genomes consisting of two or more identical subgenomes. Duplicate genes are quickly lost, some of them from one copy of a chromosome and some from the other (interleaving loss, or *fractionation*), and the chromosomes are rearranged so that elements of one subgenome are interspersed with elements from the other. Analysis of gene order change through rearrangement is a well-studied avenue to the inference of evolutionary history, but this is seriously impeded by the presence of genome duplication and fractionation. Nevertheless, undertaking this task is essential to understanding the history of plant chromosomal structure.

Fortunately, even after extensive fractionation and rearrangement, genomes that have undergone duplication ($k=2$) or triplication ($k=3$) or higher multiplication ($k>3$), retain a signature pattern that can help in decoding the evolutionary history. This pattern involves the partition of all or most of the genome into a (usually large) number m of sets $\{S_1, \cdots, S_m\}$ of k mutually syntenic chromosomal fragments $S_i=\{f_{i1}, \cdots, f_{ik}\}$, sharing pairs, triples, . . ., or k-tuples of genes with only one copy per fragment. As illustrated in Fig. 1 for a small example, there may also be a large number or majority of single-copy genes in the fragments, but each pair of fragments $\{f_{ih}, f_{ig}\}$ within S_i is connected by a substantial number of these paralogs, and there are no, or very few, paralogs between fragments f_{ih} and f_{jg} in different sets S_i, and S_j of the partition. Explicit recognition of this pattern dates from the archetypical study of the *Vitis vinifera* (grapevine) genome (Jaillon et al., 2007), which contains the original discovery of the hexaploidization underlying the explosive radiation of the core eudicots.

Genome halving ($k=2$) (El-Mabrouk and Sankoff, 2003) and genome aliquoting ($k>2$) (Warren and Sankoff, 2009, 2011) are computational procedures for inferring the pre-polyploidization ancestor of a re-diploidized and rearranged tetraploid or polyploid where there are exactly k paralogous versions of each gene. In particular it finds the number of chromosomes in the ancestor. The analysis of halving has been generalized to allow single-copy genes as part of a number of packages for inferring the gene order of the common ancestor of a set of related genomes. Some of these, e.g. Savard et al. (2011), are impractical for large eukaryote genomes containing mostly single-copy genes, and others, e.g. Jones et al. (2012), handle duplicated regions as an exceptional case to a procedure for phylogenetic ancestral reconstruction of the ancestor of number of diploids. None place any special focus on respecting the signature pattern of paralogy following whole genome duplication, described above and in Section 2. In that section we propose a "practical halving" approach to reconstructing the pre-doubling ancestor, derived from the practical aliquoting procedure (Zheng and Sankoff, 2013), which prioritizes evidence for this signature pattern.

* Corresponding author.
E-mail address: sankoff@uottawa.ca (D. Sankoff).

1476-9271/$ – see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiolchem.2014.01.010

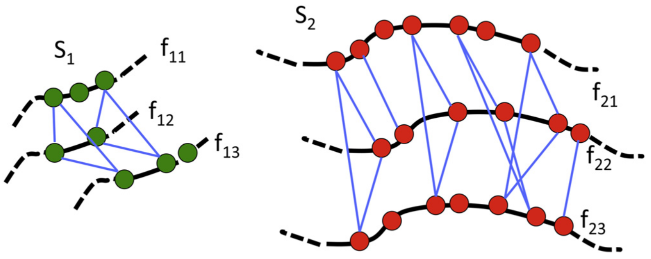

Fig. 1. Part of ancient polyploid where $k=3$ and $m=2$. Dots represent genes, line represent chromosome fragments. Note triples and pairs of paralogs as well as single copy genes, and rearranged gene order in fragment f_{23}.

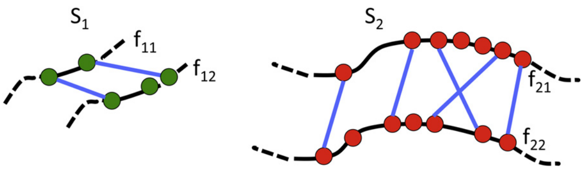

Fig. 3. Four chromosomes in an ancient tetraploid, arranged in two homeologous pairs. Dots represent genes, vertices in two kinds of graph: Black edges connect successive genes in linear orders, and blue edges indicate bipartite paralogy relationships. Note single-copy genes resulting from loss of paralog on the homeologous chromosome, and rearranged gene order in fragment f_{22}.(For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Nelumbo nucifera (sacred lotus) occupies a critical position in angiosperm phylogeny. By most accounts, and as illustrated in Fig. 2(i), it branched off from the rest of the eudicot lineage about 130 Mya, escaping the γ whole genome triplication 125 Mya responsible for the core eudicot radiation, but undergoing its own whole genome doubling, the "λ" event, some 65 Mya. Of particular interest is the number of chromosomes in the pre-λ ancestor of present-day *Nelumbo*, a quantity that we would like to compare to the seven pre-γ ancestral chromosomes of the core eudicots.

In Section 3 we will apply the practical halving algorithm to *Nelumbo* as a first step in estimating the number of ancestral chromosomes. In trying a large range of parameter settings, we find that the current version of the program tends to produce solutions that include one or two unrealistically large chromosomes and/or a large number of very small chromosomes. We opt for a more fragmented solution and invoke a comparison with the *V. vinifera* genome (Section 4) to filter out some of the smaller chromosomes, while still respecting the signature qualities of the reconstruction. This lead to a 5-chromosome solution, not very different from 7-chromosome core eudicot ancestor. It implies that 10 chromosomes existed in the post-duplication *Nelumbo* genome, and this has been reduced to 8 by chromosome fusion to produce the modern genome. In the process there has been considerable rearrangement, so this history could not have been discerned without carrying out the halving exercise.

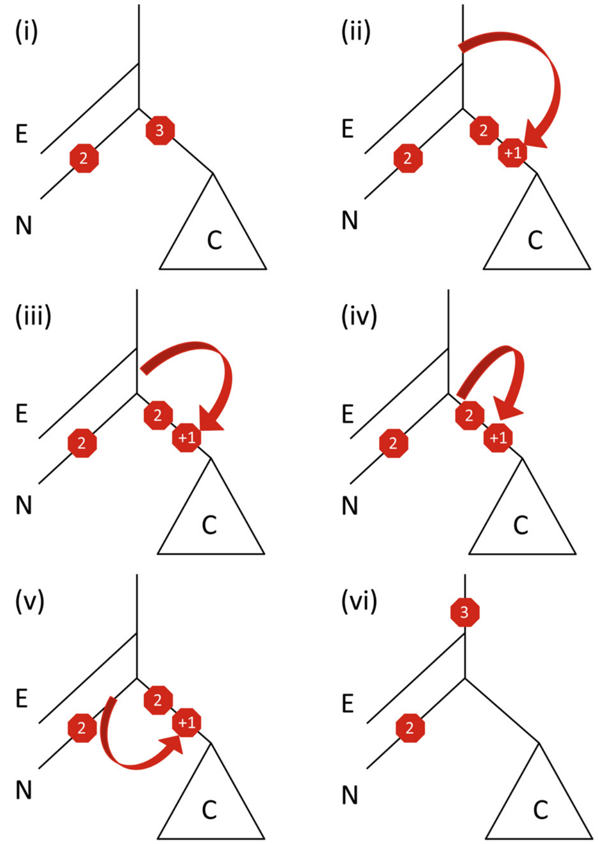

Fig. 2. Hypotheses about eudicot genome history. E = early basal eudicots, N = *Nelumbo*, C = core eudicots, including *Vitis*. A single icon containing "3", or the pair of neighboring icons containing "2" and "+1", pertains to γ, the hexaploidization preceding the core eudicot ancestor. The tail of each arrow indicates the lineage of origin of the third, dominant, subgenome that combines (+1) with an previously formed tetraploid genome to create the hexaploid. The icon containing "2" on the N lineage refers to λ, the whole genome duplication of the *Nelumbo* ancestor that we reconstruct.

The genome publication for *Nelumbo* (Ming et al., 2013) suggests that γ can be construed as two successive tetraploidizations $B+B'$ and $A+BB'$, the latter being an allopolyploidization with an earlier diverging sister genome A. The questions arise as to whether A diverged before or after the *Nelumbo*-core eudicot split. If it occurred before, was it early, before many the basal eudicots (Fig. 2(ii)) or later, after the other basal eudicots had diverged (Fig. 2(iii)), and if A originated after the split, whether it branched from the core eudicot ancestor (Fig. 2(iv)) or from the *Nelumbo* lineage itself (Fig. 2(v)).

However, the *Nelumbo* paper also cites "phylogenomic" data as being "...consistent with an earlier phylogenomic analysis using data from numerous plant genomes and basal eudicot transcriptomes suggesting that 18–28% of γ block duplications were eudicot-wide..., even though the signal is primarily observed in core eudicots". The wording in this interpretation by Jiao et al. (2012) suggests the possibility that γ occurred as in Fig. 2(vi), although Fig. 2(ii) is adduced as an explanation of these observations.

Each of the options (i)–(vi) in Fig. 2 makes predictions about the sequence divergence of the various subgenomes in the core-eudicots and in *Nelumbo*. We investigate these predictions systematically in Section 5 and find that only those in Fig. 2(i) are validated. Fig. 2(iv) could also be justified but only if the time intervals between the three events depicted, namely the doubling, the origination of the third subgenome and its incorporation with the other two, are very small on the evolutionary time scale.

Our concluding remarks evaluate the relative accuracy of synteny-based and gene family-based estimates of evolutionary events.

2. Practical halving and the *N. nucifera* genome

Halving must take into account two independent characteristics of genome organization, synteny and paralogy. The first, *genome bipartition* (distinct from phylogenetic bipartition), has to do with homology among genes within a doubled genome, more particularly the pairs of paralogous genes created by a whole genome doubling.

The second, *double synteny*, involves gene positions on the chromosome. These two as aspects are illustrated in Fig. 3. After the

Table 1
Number of genes in first estimate of chromosome pairs.

Chromosome pair														
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Genes														
234	29	2717	124	1469	1839	462	195	1152	1134	2114	271	67	17	65

doubling of a genome originally containing n genes on C chromosomes, each of the $2n$ genes in the new $2C$-chromosome genome can be considered a vertex in a bipartite graph connected by an edge only to its paralog in the other part of the graph. This is bipartition. In addition each vertex is linearly ordered with respect to some subset χ of the other vertices – *with no edges (paralogies) among them* – representing one of the $2C$ chromosomes, and these subsets are disjoint. The orderings are reflected exactly within another chromosome, called its *homeolog*, containing a paralog of each of the genes in χ. The parallel orderings constitute perfect double synteny.

The paralogy graph and the homeology subsets representing an initially doubled genome evolve over time through chromosomal rearrangement and duplicate gene fractionation, introducing "defects" into both the bipartition and the double synteny. The rearrangements disrupt the linear order of the chromosomes, and may also involve the exchange of vertices between two subsets (chromosomes). Moreover, most of the vertices may simply be deleted from the graph, representing gene loss and paralogy loss, although at least one gene, "single-copy", in each pair of paralogs must remain.

The halving problem becomes: Given graph endowed with a bipartition of its vertices into $n > 0$ components, which are either single vertices or pairs of vertices connected by an edge, and given another partition of these vertices into a number of sets each of which is linearly ordered, to try to detect the "remains" of a doubled genome, by verifying whether it is bipartite, or almost so, and whether some regions of largely parallel linear ordering can be detected in two copies respecting the paralogy. To make this statement more precise requires specifying how deviations from strict bipartition are penalized relative to gaps between fragments in a region compared to the given linear ordering, as well as other considerations discussed in the next section.

3. The search for subgenomes

For the genome to be halved, the input to our procedure is its gene order along its chromosomes, together with a partition of all the genes into pairs of paralogs, plus single-copy genes. The latter are in fact ignored because they contain no information relevant to the choices made during halving. Each gene is

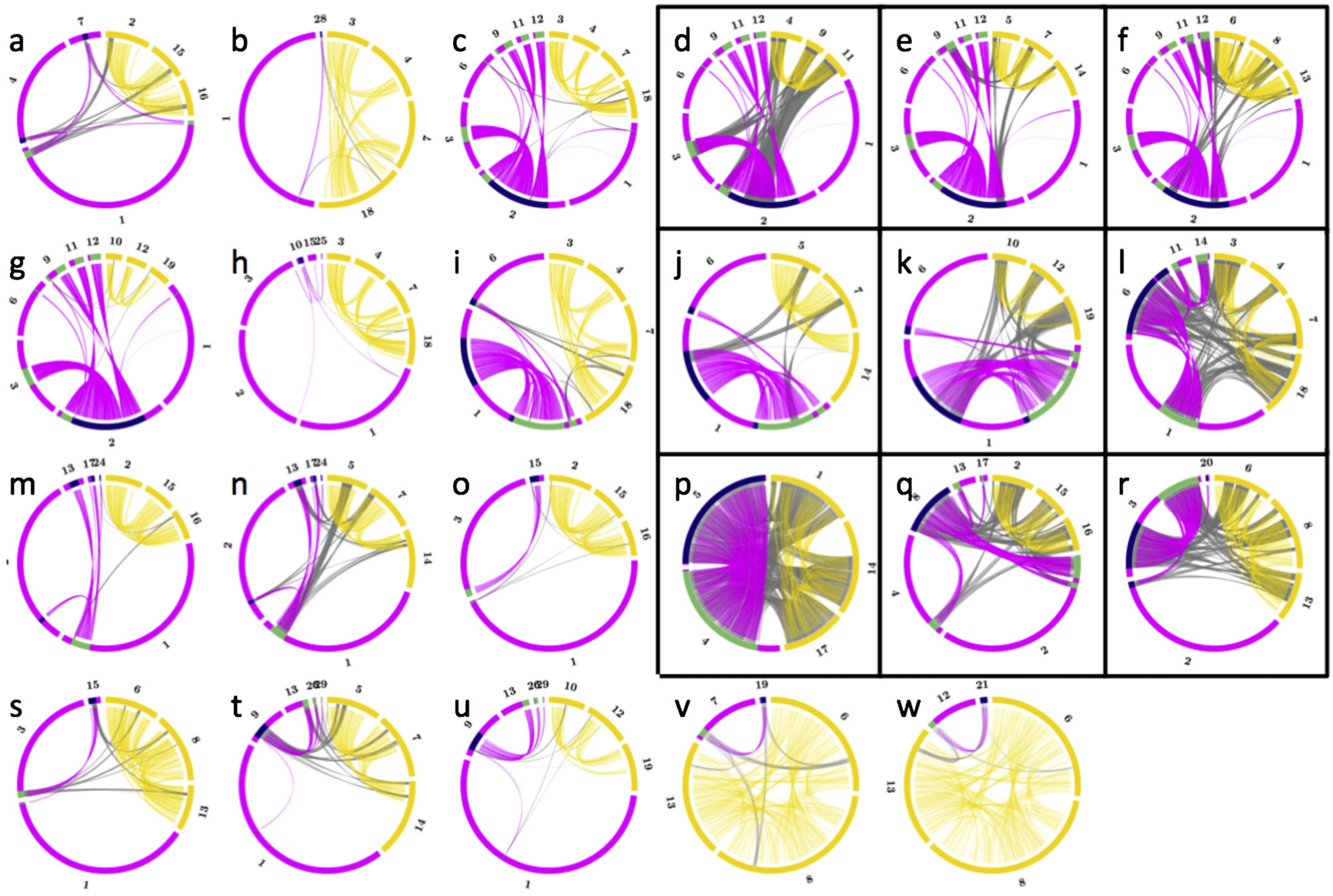

Fig. 4. Comparison of inferred pre-duplication chromosomes in *Nelumbo* ancestor and pre-triplication regions in the core eudicot ancestor. Yellow chromosomes and paralogy edges: *Vitis*. Violet chromosomes and paralogy edges: *Nelumbo*. Black edges: orthologies between three *Vitis* subgenomes and two *Nelumbo* subgenomes. Orthologies highly concentrated in comparisons d, e, f, j, k, l, p, q, r.

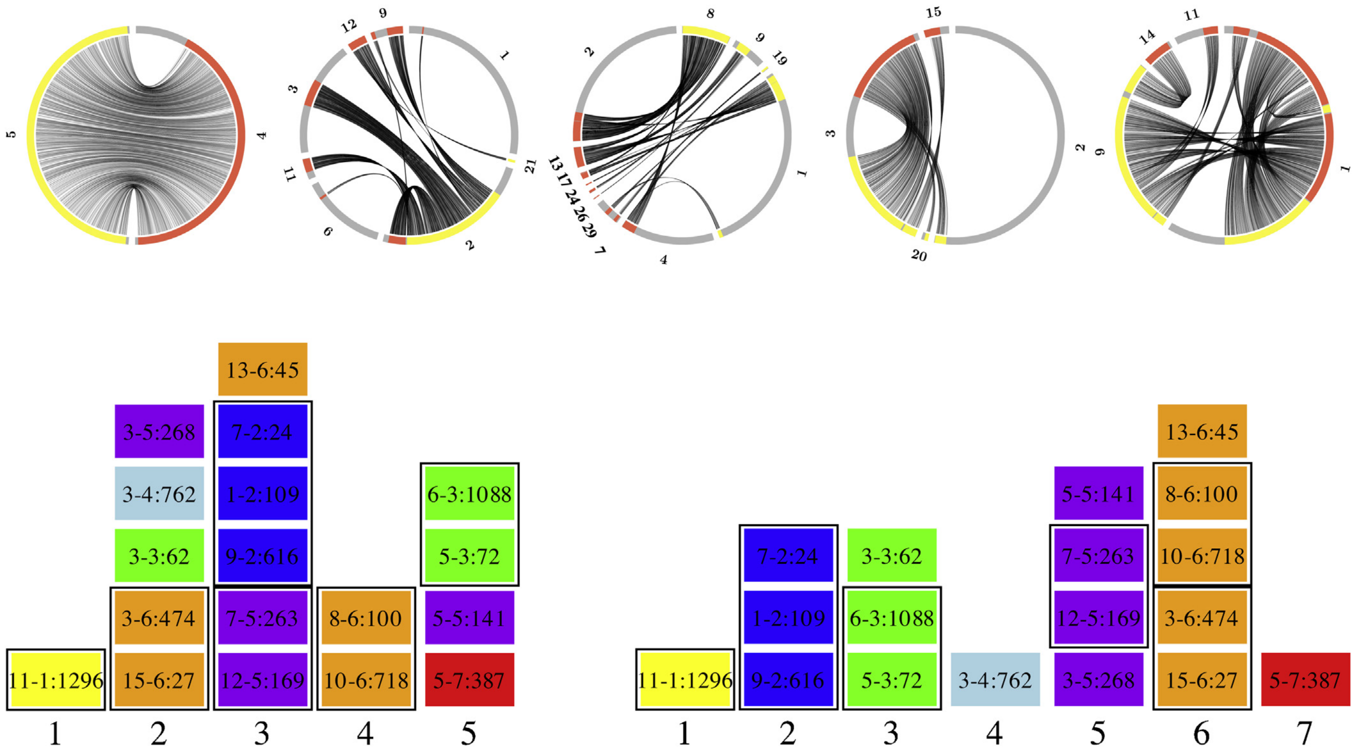

Fig. 5. Top: Ancestral *Nelumbo* duplicate chromosomes reflected in megascaffolds of current genome. Bottom: Reconstructed *Nelumbo* ancestral chromosomes and *Vitis* ancestral chromosomes (not to scale), showing common blocks and conserved block adjacencies. Each block contains the label of a *Nelumbo* chromosome pair in the original halving solution, followed by the *Vitis* chromosome triplet number, and the number of orthologs. Heavy outline surround blocks adjacent in both ancestral genomes, and undoubtedly in their common eudicot ancestor.

identified by a distinct label and its only two relevant properties are its position on a specific chromosome, and the identity of its paralog.

We use the SynMap procedure in CoGe (Lyons and Freeling, 2008; Lyons et al., 2008) to extract these data via a self-comparison of the genome. We assume this information is completely accurate, or very nearly so, both with respect to gene order and paralogy assignment.

While the paralogy relations among genes can be assumed to have been constant since the polyploidization event, the gene positions have been subject to rearrangement and we can only hope to identify relatively long multiply copied regions in the two subgenomes.

Our procedure is essentially an agglomerative clustering algorithm producing clusters that each have two internal orderings, called *regions* representing parts of the original subgenomes. At the outset each paralogy set is considered a cluster containing one item, namely the set itself.

We use three parameters to control the agglomeration step in the algorithm, a "short gap" reward $r > 0$, a chromosome "jump" penalty $j < 0$ and an "halving defect" penalty h. A fourth parameter, threshold $t > 0$, is applied in post-processing to modify very short regions.

Some terminological distinctions: A *fragment* is a contiguous set of genes on a chromosome of the input genome. (This ignores any single-copy genes, which have already been removed from consideration.) A *region* is an ordered set of fragments, with successive fragments being separated by a *gap* of one or more genes on a chromosome, or by a *chromosome jump*, i.e., the two fragments are on different chromosomes. In a pair of regions, ideally all the paralogs of all the genes are between the regions and none are within a single region. Pairs of paralogs that are exceptions to this rule are called *halving defects*.

The key step in the algorithm sketched below is the iterative clustering together of two existing clusters, which are pairs of regions, to make a larger region. The best pair to merge is determined by a score calculated by comparing the two original clusters with the potential new one. When two regions are merged, some gaps may be filled in, completely or in part, and some gaps may be created, such as between the end of one region and the beginning of the other. If the merger were to reduce the total number of gapped genes, it is assigned score r. If it does not reduce the total number of gapped genes, the score component due to gaps is $\max(0, r-x)$ where x is the change in total number of gapped genes in the new region. In addition there is a penalty j if the number of chromosomes of the input genome in the two regions being merged is less than the number in the output. Finally, if the number of halving defects in the merged regions is d greater than that in both of the original regions, a penalty of hd is assessed. The score $S(i_1, i_2)$ associated with the candidate merger of regions i_1 and i_2 is thus the gap component plus the chromosome component, summed across two paralogous regions, plus a halving defect component:

$$S(i_1,i_2)=\sum_{\text{pairofregions}}[\max(0,r-x)-j\chi(\text{jump})]-hd\chi(d>0), \quad (1)$$

where $x = 0$ if the number of gapped genes does not increase, and $\chi(\text{jump})$ and $\chi(d > 0)$ are indicator functions of increased jumps and increased aliquoting defects, respectively.

Algorithm practical halving

- **Parameters:** short gap reward $r > 0$, jump $j > 0$, halving defect penalty $h > 0$, threshold $t \geq 0$.
- **Input:** $n > 0$ paralogy sets, each containing two genes. Genes distributed and ordered on C' chromosomes.
- **Output:** A number $C'' \geq 1$ of pairs of regions

Table 2
Similarities between *Nelumbo* and *Vitis* orthologs and between *Vitis* paralogs. Numbers in blue indicate the larger of two similarities. The is no tendency for the non-dominant genomes to more similar, and there are almost no statistically significant difference (*t*-test) in any case. The more significant comparison is listed in red.

	similarity of genes in *Vitis* subgenomes to *Nelumbo* orthologs		
	two subgenomes being compared		
	small-medium	medium-large	small-large
triple "colour"			
yellow	720 795 75.47 75.63 p<.60	795 1086 75.47 **75.62** **p<.57**	720 1086 **75.63** 75.62 p<.97
dark blue	393 525 75.43 74.83 p<.11	525 569 74.83 **75.86** **p<.0023**	393 569 75.43 **75.86** p<.22
green	561 563 75.46 75.35 p<.75	563 1218 75.35 **75.45** **p<.72**	561 1218 **75.46** 75.45 p<.99
light blue	482 502 75.38 75.41 p<.93	502 617 75.41 **75.93** p<.11	482 617 75.38 **75.93** **p<.11**
purple	315 596 75.90 75.48 **p<.29**	596 714 75.48 **75.75** p<.38	315 714 **75.90** 75.75 p<.70
orange	730 967 75.48 75.69 **p<.45**	967 1091 **75.69** 75.56 p<.59	730 1091 75.48 **75.56** p<.77
red	197 233 74.88 75.70 **p<.15**	233 354 **75.70** 75.39 p<.51	197 354 74.88 **75.39** p<.31

similarity of genes in pairs of *Vitis* subgenomes		
two subgenomes being compared		
small-medium	medium-large	small-large
156 172 73.45 74.2 p<.32	172 186 74.2 **74.28** p<.91	156 186 73.45 **74.28** **p<.27**
66 94 72 73.13 p<.21	94 132 73.13 **73.64** p<.53	66 132 72 **73.64** **p<.09**
74 114 76.11 73.82 p<.0156	114 190 **73.82** 73.75 p<.92	74 190 **76.11** 73.75 **p<.002**
74 122 73.97 76.12 **p<.0317**	122 128 **76.12** 75.03 p<.16	74 128 73.97 **75.03** p<.29
20 30 74.2 75.2 p<.66	30 161 **75.2** 74.16 **p<.40**	20 161 **74.2** 74.16 p<.98
174 197 75.20 75.27 p<.92	197 258 75.27 **75.55** p<.63	174 258 75.20 **75.55** **p<.58**
50 53 74.08 73.91 p<.90	53 94 73.91 **75.34** **p<.15**	50 94 74.08 **75.34** p<.26

- **Initialization:**
  - Each set of paralogs defines a pair of regions, each region consisting of at most one fragment made up of one gene.
  - For any "pairs of pairs" of regions, calculate their clustering score *S*.
- **while** there remain pairs of pairs of regions with $S > 0$,
  - merge the pair of pairs of regions with max *S*,
  - delete merged pairs of regions and add the resulting larger pair of regions,
  - calculate the clustering score *S* of the new pair of regions with all other pairs of regions
- **Post-processing** If the gaps between two consecutive fragments in any region is smaller than threshold *t*, move the missing genes from their current location to fill in the gap as long as any resulting halving defects in the bipartition are not excessive. It is preferable to set *t* to as low a value as possible if this does not cause a proliferation of very small regions.

The initialization of the coefficients requires quadratic time, but may they be stored to allow rapid search; the update step proceeds in linear time since only the coefficients involving the two clusters being combined are affected. The iteration stops when no further amalgamation has positive score, after a number of steps less than *n*, so that the total running time requirement is quadratic.

The post-processing step involves some subjective judgment about how many aliquoting defects and how many small regions

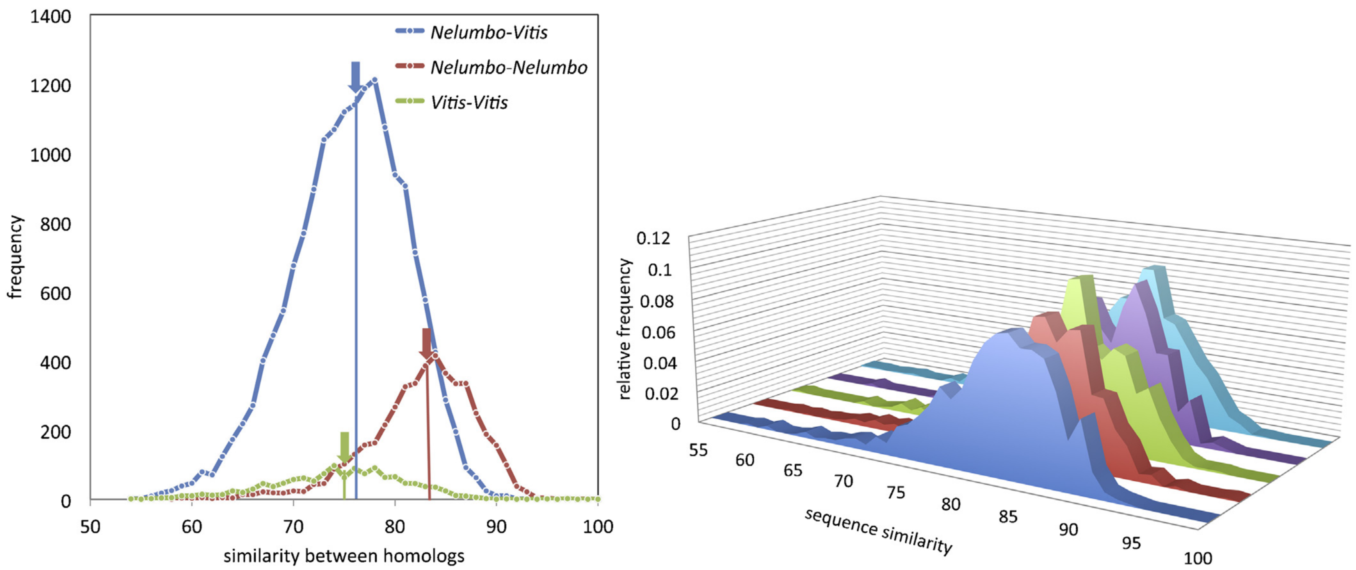

Fig. 6. Left: Similarity between *Nelumbo* and *Vitis* orthologs, between *Nelumbo* paralogs and between *Vitis* paralogs. Right:Distribution of similarities between *Nelumbo* paralogs separately for each ancestral chromosome.

are tolerable. This can of course be formalized, but it will always be dependent on the specific problem instance and to what purposes the solution will be applied.

4. The comparison of *Nelumbo* and *Vitis* ancestors

We used the *N. nucifera* genome data accessible in the CoGe database. In comparing this genome with itself, using the SynMap program to find synteny blocks, out of a total of 26,473 genes, there were 9262 paralogs in 4631 pairs in syntenic contexts. As mentioned in the introduction we set the parameters to achieve a halving result of 15 pairs of ancestral chromosomes: these contained over 99% of the genes. The goal was to avoid concentrating too many genes in unrealistically large chromosomes, while at the same time not identifying too many small fragments as ancestral chromosomes. As listed in Table 1 the 15 pairs of chromosomes included six with over a thousand genes and four with less than a hundred. The remaining five all had less than 500.

Only one pair of chromosomes reflected two contiguous homeologous blocks in the extant *Nelumbo* genome. The remaining 14 pairs were fragmented across a total of 62 blocks, so that a single chromosome ancestral chromosome could project to several blocks at scattered locations on a *Nelumbo* chromosome and/or to blocks on several *Nelumbo* chromosomes.

To improve these results, i.e., to either amalgamate or discard tiny chromosomes and possibly decompose the largest ones into two more reasonably sized ones, we introduced a comparison with the *V. vinifera* genome, whose ancestral chromosomal history is well-known (Zheng et al., 2013). Since there are many *Nelumbo–Vitis* orthologs that have only one copy in *Nelumbo*, we first filled in each of the 64 *Nelumbo* blocks with the single-copy genes falling into the range between the 5′-most paralog and the 3′-most paralog in that block.

Comparing *Nelumbo* to *Vitis* with SynMap, we found a total of 12,610 *Nelumbo* genes with *Vitis* orthologs in syntenic blocks. Of these 10,162 were in both our *Nelumbo* blocks and in known *Vitis* triplicated regions. We then compared each of the 15 tentative ancestral *Nelumbo* ancestral chromosome pairs to each of the 7 known *Vitis* ancestral triples. Only 23 of the 105 combinations shared more than one or two ortholog pairs, as depicted in Fig. 4. Moreover, most orthologs are concentrated in only nine of the 23 combinations, which reflect the six largest chromosome pairs in Table 1.

The data in Fig. 4 prompted us to concentrate on 52 blocks of *Nelumbo–Vitis* orthologs showing more than 20 contiguous genes in each chromosome. We used these as building blocks of our final estimate of five *Nelumbo* ancestral chromosomes, as depicted in Fig. 5 largely according to the following criteria:

- two blocks adjacent in both *Nelumbo* and a *Vitis* ancestral chromosome,
- two blocks adjacent at least twice in *Nelumbo*, but clearly not due to a recent reversal.

5. Gene divergence evidence

The scenarios in Fig. 2 predict the following about the similarities of *Nelumbo–Vitis* orthologs and of *Vitis* paralogs:

–2(i) The similarities of *Nelumbo–Vitis* orthologs should all be equal, regardless of subgenome. The similarities between paralogs in *Vitis* should all be equal, regardless of subgenomes.

–2(ii) and (iii) The similarities of *Nelumbo–Vitis* orthologs should be less for the dominant *Vitis* subgenome. The similarities between paralogs in *Vitis* should be greater for the two non-dominant subgenomes.

–2(iv) The similarities between paralogs in *Vitis* should be greater for the two non-dominant subgenomes.

–2(v) The similarities of *Nelumbo–Vitis* orthologs should be greater for the dominant *Vitis* subgenome. The similarities between paralogs in *Vitis* should be greater for the two non-dominant subgenomes.

–2(vi) The similarities between paralogs in *Nelumbo* should have a bimodal distribution.

None of the predicted differences between the dominant and non-dominant subgenomes in Fig. 2(ii)–(iv) or in (v) hold. Table 2 shows that there are almost no significant differences between pairs of subgenomes. And Fig. 6 shows no evidence of the bimodality predicted by Fig. 2(vi).

6. Conclusions

We have used the practical aliquoting algorithm to gain insight into the pre- and post-WGD structure of the *N. nucifera* genome. The results were not as clear as was hoped, but the addition of the *Vitis* genome to the analysis enabled us to reconstruct the pre-doubling ancestor of *Nelumbo*.

The disproportionately large number of genes in one of the subgenomes in the triplicated core eudicot genome has been cited as evidence for one of scenarios (ii), (iii) or (iv) in Fig. 2. The dominant subgenome would be the one added in to the genome some time after the initial tetraploidization, and would thus have had less time to lose genes through fractionation. However, all these scenarios predict that the other two subgenomes should be less divergent in their gene sequences from each other than they are from the dominant one, since their duplication event is relatively recent. But there is absolutely no evidence that such a prediction is validated. The dominant subgenome is no more divergent than the other two, as is evident in comprehensive statistical testing. More compelling explanations of subgenome dominance are to be found in epigenetic mechanisms that establish patterns of preferential gene expression between homeologous chromosomes, perhaps as a side-effect of transposon repression mechanism (Freeling et al., 2012; Schnable et al., 2011).

Nor does the *Nelumbo* genome contain evidence of γ in the basal eudicots. The similarity of all the paralog pairs are distributed around a high value, indicative of a relatively recent event, even taking into account *Nelumbo*'s slow rate of sequence evolution. There is no syntenic evidence, such as a bump in the histogram of similarities, that "18–20 %" of the pairs (Ming et al., 2013) originated at an earlier date.

What of the "phylogenomic" evidence for an early basal eudicot origin for such a large portion of duplicate pairs observed in the "γ blocks", especially considering the rigorous methodology utilized in the original paper (Jiao et al., 2012)? The answer lies simply is the inherent uncertainty in the data input to the phylogenetic programs. Compared to the whole-genome synteny methods we have applied to one or two genomes at a time, phylogenomic methods have the advantage of phylogenetic scope, the recruitment of data from a wider range of genomes. However, they cannot attain the accuracy afforded by the thousands of similarity measures bearing on a single event.

Individual genes are do not contain enough variable nucleotide position to estimate divergence times and branching orders accurately, and inferred trees are easily distorted by a few statistical outliers. Add to this rate variation among lineages, long branch effects, sparse taxon sampling for some genes, missing paralogs, incorrectly identified paralogs, i.e., general lack of syntenic control, except in *Vitis*, and multiple trees of approximately equal credibility. Some but not all of these are controlled for, but only partially. Despite the care taken over the data collection, the use of the best phylogenetic software and interpretive procedures, nothing can overcome the highly variable nature of the outcome. High bootstrap values do nothing to correct biases due inherently highly variable data.

It is thus not surprising that only three-quarters of the gene trees in Jiao et al. (2012) produce time estimates for γ in the expected range. We conclude that only the scenario in Fig. 2(i) is valid.

References

Blanc, G., Hokamp, K., H, W.K., 2003. A recent polyploidy superimposed on older large-scale duplications in the arabidopsis genome. Genome Research 13, 137–144.

El-Mabrouk, N., Sankoff, D., 2003. The reconstruction of doubled genomes. SIAM Journal on Computing 32, 754–792.

Freeling, M., Woodhouse, M.R., Subramaniam, S., Turco, G., Lisch, D., Schnable, J.C., 2012. Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. Current Opinion in Plant Biology 15, 131–139.

Ibarra-Laclette, E., Lyons, E., Hernández-Guzmán, G., Pérez-Torres, C.A., Carretero-Paulet, L., Chang, T.-H., Lan, T., Welch, A.J., Juárez, M.J.A., Simpson, J., Fernández-Cortés, A., Arteaga-Vázquez, M., Góngora-Castillo, E., Acevedo-Hernández, G., Schuster, S.C., Himmelbauer, H., Minoche, A.E., Xu, S., Lynch, M., Oropeza-Aburto, A., Cervantes-Pérez, S.A., de Jes&rsquo;us Ortega-Estrada, M., Cervantes-Luevano, J.I., Michael, T.P., Mockler, T., Bryant, D., Herrera-Estrella, A., Albert, V.A., Herrera-Estrella, L., 2013. Architecture and evolution of a minute plant genome. Nature 498, 94–98.

Jaillon, O., Aury, J.M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., Vezzi, A., Legeai, F., Hugueney, P., Dasilva, C., Horner, D., Mica, E., Jublot, D., Poulain, J., Bruyère, C., Billault, A., Segurens, B., Gouyvenoux, M., Ugarte, E., Cattonaro, F., Anthouard, V., Vico, V., Del Fabbro, C., Alaux, M., Di Gaspero, G., Dumas, V., Felice, N., Paillard, S., Juman, I., Moroldo, M., Scalabrin, S., Canaguier, A., Le Clainche, I., Malacrida, G., Durand, E., Pesole, G., Laucou, V., Chatelet, P., Merdinoglu, D., Delledonne, M., Pezzotti, M., Lecharny, A., Scarpelli, C., Artiguenave, F., Pè, M.E., Valle, G., Morgante, M., Caboche, M., Adam-Blondon, A.F., Weissenbach, J., Quétier, F., Wincker, P., French-Italian Public Consortium for Grapevine Genome Characterization, 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449, 463–467.

Jiao, Y., Leebens-Mack, J., Ayyampalayam, S., Bowers, J.E., McKain, M.R., McNeal, J., Rolf, M., Ruzicka, D.R., Wafula, E., Wickett, N.J., Wu, X., Zhang, Y., Wang, J., Zhang, Y., Carpenter, E.J., Deyholos, M.K., Kutchan, T.M., Chanderbali, A.S., Soltis, P.S., Stevenson, D.W., McCombie, R., Pires, J.C., Wong, G.K., Soltis, D.E., dePamphilis, C.W., 2012. A genome triplication associated with early diversification of the core eudicots. Genome Biology 13, R3.

Jones, B.R., Rajaraman, A., Tannier, E., Chauve, C., 2012. Anges: reconstructing ancestral genomes maps. Bioinformatics 28, 2388–2390.

Lyons, E., Freeling, M., 2008. How to usefully compare homologous plant genes and chromosomes as DNA sequences. The Plant Journal 53, 661–673.

Lyons, E., Pedersen, B., Kane, J., Alam, M., Ming, R., Tang, H., Wang, X., Bowers, J., Paterson, A., Lisch, D., Freeling, M., 2008. Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar and grape: CoGe with rosids. Plant Physiology 148, 1772–1781.

Ming, VanBuren, R., Liu, Y., Yang, M., Han, Y., Li, L.-T., Zhang, Q., Kim, M.-J., Schatz, M., Campbell, M., Li, J., Bowers, J., Tang, H., Lyons, E., Ferguson, A., Narzisi, G., Nelson, D., Blaby-Haas, C., Gschwend, A., Jiao, Y., Der, J., Zeng, F., Han, J., Min, X.J., Hudson, K., Singh, R., Grennan, A., Karpowicz, S., Watling, J., Ito, K., 2013. Genome of the long-living sacred lotus (*Nelumbo nucifera* Gaertn.). Genome Biology 14, R41.

Savard, O.T., Gagnon, Y., Bertrand, D., El-Mabrouk, N., 2011. Genome halving and double distance with losses. Journal of Computational Biology 18, 1185–1199.

Schnable, J.C., Springer, N.M., Freeling, M., 2011. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proceedings of the National Academy of Sciences 108, 4069–4074.

Warren, R., Sankoff, D., 2009. Genome aliquoting with double cut and join. BMC Bioinformatics 10 (Suppl. 1), S2.

Warren, R., Sankoff, D., 2011. Genome aliquoting revisited. Journal of Computational Biology 18, 1065–1075.

Zheng, C., Sankoff, D., 2013. Practical aliquoting of flowering plant genomes. BMC Bioinformatics 14 (Suppl. 15), S8.

Zheng, C., Chen, E., Albert, V.A., Lyons, E., Sankoff, D., 2013. Ancient eudicot hexaploidy meets ancestral eurosid gene order. BMC Genomics 13, S7:S5.