Journal of Bioinformatics and Computational Biology
Vol. 7, No. 3 (2009) 499–520
© Imperial College Press

Imperial College Press
www.icpress.co.uk

GENE LOSS UNDER NEIGHBORHOOD SELECTION FOLLOWING WHOLE GENOME DUPLICATION AND THE RECONSTRUCTION OF THE ANCESTRAL POPULUS GENOME

CHUNFANG ZHENG

Department of Biology, University of Ottawa
Ottawa, Ontario K1N 6N5, Canada
czhen033@uottawa.ca

P. KERR WALL

Biology Department, Penn State University
University Park, PA 16802, USA
pkerrwall@gmail.com

JAMES LEEBENS-MACK

Department of Plant Biology, University of Georgia
Athens, GA 30602, USA
jleebensmack@plantbio.uga.edu

CLAUDE dePAMPHILIS

Biology Department, Penn State University
University Park, PA 16802, USA
cwd3@psu.edu

VICTOR A. ALBERT

Department of Biological Sciences, SUNY Buffalo
Buffalo, NY 14260, USA
vaalbert@buffalo.edu

DAVID SANKOFF*

Department of Mathematics and Statistics
University of Ottawa, Ottawa
Ontario K1N 6N5, Canada
sankoff@uottawa.ca

Received 18 July 2008
Revised 6 November 2008
Accepted 11 November 2008

*Corresponding author.

We develop criteria to detect neighborhood selection effects on gene loss following whole genome duplication, and apply them to the recently sequenced poplar (*Populus trichocarpa*) genome. We improve on guided genome halving algorithms so that several thousand gene sets, each containing two paralogs in the descendant T of the doubling event and their single ortholog from an undoubled reference genome R, can be analyzed to reconstruct the ancestor A of T at the time of doubling. At the same time, large numbers of defective gene sets, either missing one paralog from T or missing their ortholog in R, may be incorporated into the analysis in a consistent way. We apply this genomic rearrangement distance-based approach to the poplar and grapevine (*Vitis vinifera*) genomes, as T and R respectively. We conclude that, after chromosome doubling, the "choice" of which paralogous gene pairs will lose copies is random, but that the retention of strings of single-copy genes on one chromosome versus the other is decidedly non-random.

Keywords: Whole genome duplication; genome rearrangement; genome halving; *Populus trichocarpa*; *Vitis vinifera*.

1. Introduction

Following an episode of whole genome doubling (WGD), gene duplicates are lost at a high rate through processes such as pseudogenization and deletion of chromosomal segments containing one or more genes, while intra- and inter-chromosomal rearrangement mechanisms redistribute chromosomal segments, both large and small, across the genome. The genome of the present-day descendant can be largely decomposed into a set of duplicated DNA segments dispersed among the chromosomes, with all the duplicated pairs exhibiting a similar degree of sequence divergence, and with single-copy segments interspersed among them. In this paper, we introduce approaches to analyzing the evolution of doubled genomes, based entirely on gene order evidence, in order to explain aspects of the gene loss process and to reconstruct the rearrangement steps leading from the doubled ancestral genome to the present day descendant.

Though syntenic evidence, namely duplicated segments containing several genes in corresponding order, has long been used for WGD, studies of duplicate gene loss have focused on functional changes and divergence rates within individual gene families, to the exclusion of gene order considerations. Here we investigate how the fate of duplicate genes is correlated, or not, to the retention or loss of nearby genes on the same chromosomes.

As for reconstructing rearrangement history, a linear-time "genome halving" algorithm, based only on the ordering of duplicated chromosomal segments, can find an ancestral genome that minimizes the genomic distance to the present-day genome.[1,2] This does not suffice, however, as a solution to the reconstruction problem, since there may be a large number of very different, equally optimal solutions. Here we use a guided genome halving (GGH) strategy to overcome this non-uniqueness, guiding the reconstruction of the ancestor by one or more reference, or outgroup, genomes. This strategy does not sacrifice the optimality of the halving solution.

The flowering plants are well known for numerous historical events of genome doubling.[3] The recently sequenced poplar genome (*Populus trichocarpa*),[4] which shows very clear evidence of genome duplication 60 to 65 million years ago, and the grapevine genome (*Vitis vinifera*),[5,6] whose ancestor diverged before the aforementioned duplication, provide a pair of analytical incentives to the GGH strategy. On the one hand, the poplar data have an order of magnitude more duplicated elements than that previously analyzed, straining computational resources. On the other hand, the richness of the data allows us to assess neighborhood selection effects on duplicate gene loss and the implications of this loss of genes from thousands of duplicated pairs on the accuracy of ancestral genome reconstruction.

This paper thus contributes three advances on the methodological level: first, a way of analyzing chromosomal neighborhood selection effects on the retention or loss of duplicated genes; second, the scaling up, by more than an order of magnitude, of the amount of data amenable to our GGH analysis; and third, the incorporation into GGH of data from gene duplicate pairs that have lost one member, making use of chromosomal context in both the genome that can be traced to the doubling event and in the outgroup.

1.1. *Outline*

In Sec. 2, we describe the sources for our data and how we processed them to obtain the gene sets for the selection study and the GGH analysis. In Sec. 3, we present our method and results for detection of neighborhood selection effects. In Sec. 4, we sketch the necessary background about genomic rearrangement distance and the genome halving and GGH algorithms. In Sec. 5, we present the GGH algorithm incorporating both full and defective gene sets. We apply this method to the full gene sets in combination with one or both of two defective gene sets from *Populus* and *Vitis* in Sec. 6. We present the reconstructed undoubled *Populus* ancestor based on over 6000 gene sets and evaluate the evolutionary signal versus noise (a) in the ancestor-*Populus* and ancestor-*Vitis* comparisons, (b) in the full and defective gene sets, and (c) in genes with two or three common adjacencies in the data and those with weaker positional evidence.

2. The *Populus–Vitis* Comparison

Annotations for the *Populus* and *Vitis* genomes were obtained from databases maintained by the U.S. Department of Energy's Joint Genome Institute[4] and the French National Sequencing Center, Genoscope,[6] respectively. An all-by-all BLASTP search was run on a dataset including all *Populus* and *Vitis* protein coding genes, and orthoMCL[7] was used to construct 2104 full and 4040 defective gene sets, in the first case containing two poplar paralogs (genome T) and one grape ortholog (genome R), and in the second case missing a copy from either T or R. The chromosomal location and orientation of these paralogs and orthologs was used to construct our database of gene orders for these genomes, and the input to the GGH algorithm. In addition, 740 *Populus* single-copy genes with known chromosomal location but with no orthology detected in *Vitis* were added for the study of neighborhood selection effects.

3. Neighborhood Effects

The data include 3944 single-copy genes with or without orthologs in *Vitis*. We assume that the single-copy status of most of these genes is simply a consequence of the loss of their duplicates following WGD, either through pseudogenization, outright deletion or other process, though of course there may be a small proportion with other explanations.

We may ask how the process of gene loss is distributed throughout the genome. In particular,

- is there any spatial non-randomness in the choice of duplicate pairs that will become single-copy? Are such pairs clustered together or randomly spaced in the genome?
- for any cluster of neighbouring gene pairs that becomes single-copy, do the remaining copies tend to reside on the same chromosome, or does each pair of duplicates "decide" independently from its neighbors which chromosome keeps the gene and which loses it?

3.1. *Choice of pairs to become single-copy*

One difficulty in addressing these problems is that we do not know which genes in the present-day genome are neighbors because they have always been neighbors, and which have become neighbors recently, for example through chromosomal rearrangement operations, and this is especially problematic for single-copy genes.

To circumvent this problem, we consider only sets of single-copy genes bounded at both ends by a pair of duplicate genes on the same two chromosomes, as in Fig. 1. Such a configuration, which we take as our basic unit of analysis, can arise only infrequently by rearrangement, especially if the genome contains more than a handful of different chromosomes. The unlikeliness of these units arising by chance rearrangement explains why it suffices to take only one flanking pair of duplicates at either end, and why we need not be more stringent and ask for two such consecutive pairs at either end, with the consequent loss of valuable data for our analysis. We

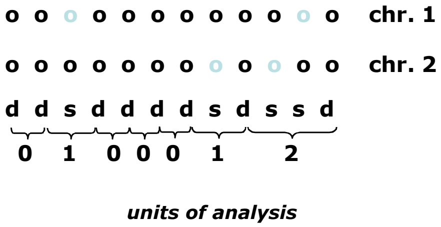

Fig. 1. Analytical units for the study of single-copy genes. Black circles represent existing genes, gray circles represent hypothesized lost genes. **d** = duplicate gene pair, **s** = single-copy remaining. Numbers indicate length of string in which one member of each pair had been lost.

assume that the single-copy status of all the intervening genes, on one or the other of the two chromosomes, arose through the loss of one copy from a corresponding position on the other chromosome. Note that these sets of single-copy genes are not "paralogons", since the genes are distributed on two chromosomes, although by hypothesis they are descended from two paralogons, which are duplicate regions on the two chromosomes.

Thus, the statistic we investigate is the frequency of the number of single-copy genes, on one chromosome or the other, between any two consecutive pairs of duplicate genes. Were the process of gene loss independent across two chromosomes of an artificial genome consisting of all the currently conserved duplicated genes (counting a pair only once if it is both at the end of one analytical unit and the beginning of another, e.g. the penultimate **d** in Fig. 1), plus two copies of the currently single-copy genes, we can predict this frequency distribution. Let n be the total number of duplicated pairs in this artificial construct. Then after S random selections of pairs from $\{1, \ldots, n\}$, with replacement but with only the first selection of any one pair resulting in a conversion to single-copy status, the probability that any particular duplicate pair A be conserved is $(1 - \frac{1}{n})^S$ and the probability that the run of m pairs following A have been converted to single-copy status, followed by a conserved duplicate pair B, is thus $(1 - \frac{1}{n})^{2S} \left(1 - (1 - \frac{1}{n})^S\right)^m$, as long as A is in the first $n - (m + 1)$ positions in the genome.

Then $\mathrm{E}[f(m)]$, the expected number of single-copy strings of length m, after S samples is

$$(n - m - 1)\left(1 - \frac{1}{n}\right)^{2S}\left(1 - \left(1 - \frac{1}{n}\right)^S\right)^m. \tag{1}$$

Requiring the predicted proportion of single-copy genes $1 - (1 - \frac{1}{n})^S$ to be equal to the observed proportion $\frac{r}{n}$, it leads to the predicted frequency:

$$(n - m - 1)\left(1 - \frac{r}{n}\right)^2\left(\frac{r}{n}\right)^m. \tag{2}$$

Comparing the predicted versus the observed frequency of single-copy unit of various sizes, as in Fig. 2, makes it clear that the null hypothesis of random choice of which duplicate pairs to convert to single-copy status would be difficult to dispute.

3.2. *Concentration of single-copies on one chromosome*

There remains, however, the second type of neighborhood selection we have mentioned, i.e. once a pair of duplicates is changed to single-copy status, is there an influence from neighboring genes on which copy is to be lost and which to be retained? Under the hypothesis of no neighbor effects, configurations like those on the right of Fig. 3, where both retained copies are on the same chromosome, should occur half of the time. Similarly for all m, we can calculate the probability that exactly q out of m genes occur on the same chromosome and $m - q$ on the other is $B(m, q) + B(m, m - q)$, for $q = 0, \ldots, \lfloor \frac{m}{2} \rfloor$, where B is the binomial distribution.

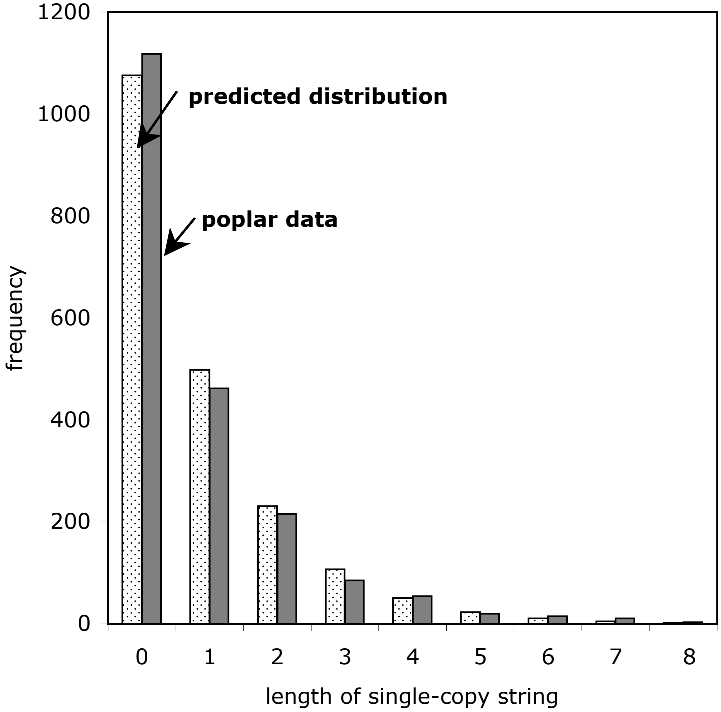

Fig. 2. Congruence of predicted versus observed lengths of strings of single-copy genes.

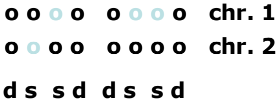

Fig. 3. Single-copy gene evidence for neighborhood selection. Black circles represent existing genes, gray circles represent hypothesized lost genes. **d** = duplicate gene pair, **s** = single-copy remaining. In the unit on the right, neighbors are conserved, on the left one gene is dropped from each chromosome.

The cumulative probability that q or fewer single-copy genes, out of m, will appear on the same chromosome is $\mathrm{Pr}_m(q) = \sum_{i=0}^{q} B(m,i) + B(m-q)$, for $q = 0, \ldots, \lfloor \frac{m}{2} \rfloor$. Plotting $\mathrm{Fr}_m(q)$, the observed cumulative relative frequency of q against this prediction $\mathrm{Pr}_m(q)$ should help us decide if the null hypothesis of randomness is justified. We use the cumulative distributions rather than the probability and relative frequency distributions themselves in order to combine the results for different values of m. Thus, in Fig. 4 we superimpose plots of $\mathrm{Fr}_m(q)$ against $\mathrm{Pr}_m(q)$, for $m = 2, \ldots, 8$ and for $q = 0, \ldots, \lfloor \frac{m}{2} \rfloor$. There should be 23 points on such a plot, but seven of them are necessarily coincidental at $(1, 1)$.

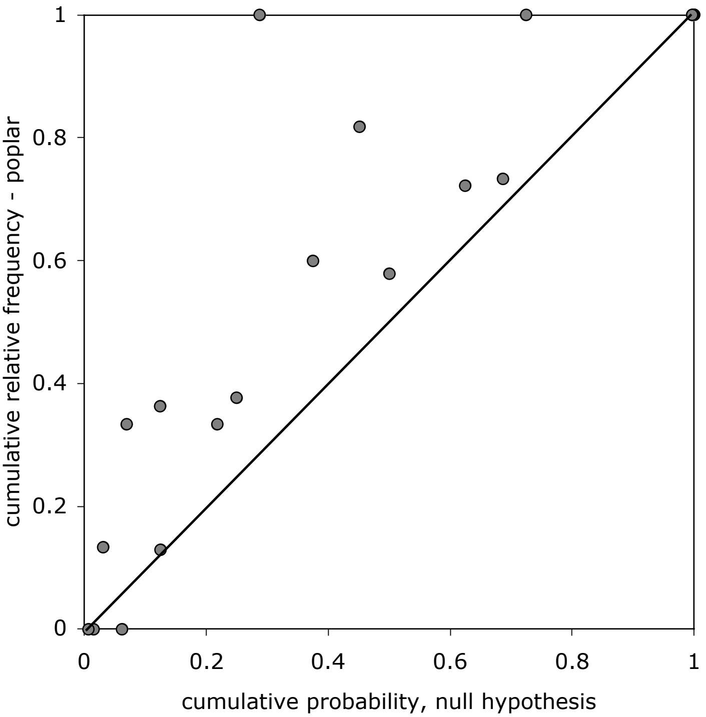

Fig. 4. Observed cumulative relative frequency against predicted values, for q or fewer single-copy genes, out of m, to appear on the same chromosome, for $m = 2, \ldots, 8$ and for $q = 0, \ldots, \lfloor \frac{m}{2} \rfloor$. There should be 23 points on such a plot, but seven of them are necessarily coincidental at $(1, 1)$.

The plausible alternative hypotheses are (a) a widespread neighborhood selection effect, possibly at the transcriptional level, perhaps involving co-regulation or common regulatory elements, or (b) deletion-based effect whereby gene loss involves a chromosomal segment, affecting more than one gene at a time. The latter is perhaps the simplest and most obvious choice, unless pseudogenization is the usual avenue to gene loss, and has been proposed in the context of gene loss after WGD in yeast.[8] Though selection-based explanations have been offered by other authors,[9] our own method (work to be reported elsewhere) yields even stronger support for selection than the poplar work reported here, rather than random large deletions. Under the null hypothesis, there is no reason for there to be more points above the diagonal than below it, and the particular concentration of points where $\mathrm{Fr}_m(q)$ is much greater than $\mathrm{Pr}_m(q)$ and $\mathrm{Pr}_m(q) < \frac{1}{3}$ suggests that the low values of q, where almost all the single-copy genes in each analytical unit are on the same chromosome, are greatly over-represented in these data.

With the poplar data, were random deletions involving segments larger than one gene, a quantitatively important explanation for the results in Fig. 4, this would necessarily also show up as deviation from the smooth geometric distribution in

Fig. 2, with a depressed value for $m = 1$ relative to increased values for at least $m = 2$ or higher m. This is emphatically not the case, and we can thus reject the large deletions theory in favour of a neighborhood selection effect.

3.3. *Long strings of single-copies*

Along with the data summarized in Figs. 3 and 4, we also found four analytical units with relatively long strings of single-copies. These occurred in analytical units on chromosome pair (6,18), with 12 genes on chromosome 18 and only 2 on chromosome 6, chromosome pair (2,14), with 12 genes on chromosome 14 and only 3 on chromosome 2, chromosome pair (13,19), with 15 genes on chromosome 13 and none on chromosome 19, and chromosome pair (8,10), with 16 genes on chromosome 10 and only 1 on chromosome 8.

These data provide additional striking confirmation of the tendency for single-copy genes to be largely on the same chromosome. The occasional instance of an active gene on the other chromosome is validation of our operational definition of an analytical unit: the surviving duplicate pairs delimiting either end of the single-copy segment did not get there through coincidental rearrangements, rather they define what remains of a genuine duplicated segment.

The extraordinary length of each of these four strings of single-copy genes remains unexplained, though functional selection effect is strongly suggested in several cases. For example, in the (2,14) pair and the (13,19) pair, about half of the genes also appear in a (single) cluster in the *Arabidopsis* genome. The genes in the (8,10) pair are almost evenly split between two such clusters in *Arabidopsis.* This suggests that the genes in question have a function-driven tendency to group together in a single-copy segment, a tendency which has survived separate WGD events in both *Populus* and *Arabidopsis.* This idea has recently been explored for yeast.[10]

Moreover, GO functional annotations show that four of the genes on chromosome 10 have nucleic acid or nucleotide binding capacity and three of the genes on chromosome 14 have hydrolase activity, in both cases more than could be expected by coincidence, and suggestive of some concerted functional connection favored by the close linkage of these genes.

3.4. *The units of analysis*

It is only in the context of WGD that we can set up the analytical units that enable our assessment of the presence of neighborhood selection in Secs. 3.1, 3.2 and 3.3 above, though these effects would exist of course whatever the source of duplicate gene pairs, and indeed in any process of genome shrinkage through gene loss.

4. Toward Efficient, Accurate Guided Genome Halving

Algorithms for guided genome halving (GGH), or reconstruction of the pre-doubling genome with the help of an outgroup, were first used for the ancestral doubled genome of the maize (*Zea mays*), with the rice (*Oryza sativa*) and sorghum (*Sorghum bicolor*) genomes as outgroups.[11] We generated all the 1.5×10^6 solutions to the genome halving problem for the maize genome, and then identified the subset, containing only a handful of relatively similar solutions that have a minimum rearrangement distance with the rice (or sorghum) genome.

This approach was feasible with the small number (34) of doubled blocks identified in maize that were also present in one copy in each outgroup, but in a subsequent analysis,[12] when we attempted to reconstruct the ancient doubled yeast genome from which *Saccharomyces cerevisiae* is descended, guided simultaneously by both of the undoubled outgroup genomes *Ashbya gossypii* and *Kluyveromyces waltii*, the number of doubled genes we could use as evidence was an order of magnitude greater than the number of blocks in the cereals data, and the number of solutions to the halving problem is astronomical. It was no longer feasible to exhaustively search the halving solutions to find those that are closest to the outgroups. Instead we took a random sample of several thousand solutions in the hope that the best one might be optimal, or close to it. It was not clear, however, how large the sample should be, or how to validate the results, since the local optima found in that study remained fairly far apart, as measured by genomic rearrangement distance.

In our current use of GGH, on yeast[13] and on the flowering plants studied in the present article, we seek to replace the brute force approach of generating all (or a random sample of) halving solutions first, i.e. before taking into consideration the outgroup genome. Instead, we inject all pertinent information derivable from the outgroup into the halving algorithm, influencing hitherto arbitrary choices in that algorithm so that the halving solution is guided towards the outgroup.

4.1. *Definitions: Genomes, rearrangement operations and genomic distance*

A genome G is represented by a set of strings (called *chromosomes*) of form $\{g_{11} \cdots g_{1n_1}, \ldots, g_{\chi 1} \cdots g_{\chi n_\chi}\}$, where $n = n_1 + \cdots + n_\chi$ and $\{|g_{..}|\} = \{1, \ldots, n\}$; i.e. each integer $i \in \{1, \ldots, n\}$ appears exactly once in the genome and may have either positive or negative polarity. The biologically-motivated operations of reversal or inversion, reciprocal translocation, chromosome fission or fusion, and transposition, can all be represented by an operation (called double-cut and join, or DCJ) of cutting the genome twice, each time between two elements on one of the chromosomes and rejoining the four resulting cut ends differently.[14,15] Whether the two cuts are on the same chromosome or not, and how the endpoints are rejoined, determine which rearrangement operation pertains.

The genome rearrangement distance $d(G, H)$ is defined to be the minimum number of DCJ operations required to convert one of the genomes, G, into the other, H.

Rearrangement algorithms[14,16,17] can be formulated in terms of the bi-colored "breakpoint graph", where each end (either $5'$ or $3'$) of a gene in genome G is represented by a vertex joined by a black edge to the vertex for adjoining end of the adjacent gene, and these same ends, represented by the same $2n$ vertices in the graph, are joined by gray edges determined by the adjacencies in genome H. In addition, each vertex representing a first or last term of some chromosome in G or in H is connected by an edge of the appropriate color to an individual "cap" vertex, and there are specific rules for adding caps to the genome with fewer chromosomes and for joining the caps among themselves. If G has χ chromosomes and H has no more than χ, there are $2n + 4\chi$ vertices in all. The breakpoint graphs necessarily consist of disjoint alternating color cycles, and it can be shown that, in the DCJ formulation, $d(G, H) = n + \chi - c$, where c is the number of cycles in the breakpoint graph. Calculation of the distance can be done in time that is linear in n.

4.2. *Genome halving*

Let T be a genome consisting of ψ chromosomes and $2n$ genes $a_1^{(1)}, \ldots, a_n^{(1)}$; $a_1^{(2)}, \ldots, a_n^{(2)}$, dispersed in any order on the chromosomes. For each i, we call $a_i^{(1)}$ and $a_i^{(2)}$ "duplicates", but there is no particular property distinguishing all elements of the set of $a_i^{(1)}$ in common from all those in the set of $a_i^{(2)}$. A potential "doubled ancestor" of T is written as $A' \oplus A''$, and consists of 2χ chromosomes, where some half (χ) of the chromosomes, symbolized by the A', contains exactly one of $a_i^{(1)}$ or $a_i^{(2)}$ for each $i = 1, \ldots, n$. The remaining χ chromosomes, symbolized by the A'', are each identical to one in the first half, in that where $a_i^{(1)}$ appears on a chromosome in the A', $a_i^{(2)}$ appears on the corresponding chromosome in A'', and where $a_i^{(2)}$ appears in A', $a_i^{(1)}$ appears in A''. We define A to be either of the two halves of $A' \oplus A''$, where the superscript (1) or (2) is suppressed from each $a_i^{(1)}$ or $a_i^{(2)}$. *The genome halving problem for T is to find an A for which some $d(A' \oplus A'', T)$ is minimal.*

In the rearrangement distance algorithm, construction of the breakpoint graph is an easy step. The genome halving algorithms[2] also make use of the breakpoint graph, but the problem here is the more difficult one of building the breakpoint graph where one of the genomes (the doubled ancestor $A' \oplus A''$) is unknown. This is done by segregating the vertices of the graph in a natural way into subsets, such that all the vertices of each cycles must fall within a single subset, and then constructing these cycles in an optimal way within each subset so that the black edges correspond to the structure of the known genome T and the gray edges define the adjacencies of $A' \oplus A''$.

As a first step, each gene a in a doubled descendant is replaced by a pair of vertices (a_t, a_h) or (a_h, a_t) depending if the DNA is read from left to right or right to left. The duplicate of gene $a = (a_t, a_h)$ is written as $\bar{a} = (\bar{a_t}, \bar{a_h})$.

Following this, for each pair of neighboring genes, say (a_t, a_h) and (b_h, b_t), the two adjacent vertices a_h and b_h are linked by a black edge, denoted by $\{a_h, b_h\}$ in the notation of Ref. 15. For a vertex at the end of a chromosome, say b_t, it generates a virtual edge of form $\{b_t, \mathrm{end}\}$. Note that the use of "end" instead of "cap" reflects a somewhat different bookkeeping for the beginnings and ends of chromosome in the halving algorithm compared to the distance algorithm in Sec. 4.1.

The edges thus constructed are then partitioned into *natural graphs* according to the following principle: If an edge $\{x, y\}$ belongs to a natural graph, then so does some edge of form $\{\bar{x}, z\}$ and some edge of form $\{\bar{y}, w\}$. If a natural graph has an even number of edges, it can be shown that in all optimal ancestral doubled genomes, the edges colored gray, say, representing adjacent vertices in the ancestor, and incident to one of the vertices in this natural graph, necessarily have as their other endpoint another vertex within the same natural graph.

For all other natural graphs, there are one or more ways of grouping them pairwise into *supernatural graphs* so that an optimal doubled ancestor exists such that the edges colored gray incident to any of the vertices in a supernatural graph have as their other endpoint another vertex within the same supernatural graph. Thus the supernatural graph may be completed one at a time.

An important detail in this construction is that before a gray edge is added during the completion of a supernatural graph, it must be checked to see that it would not inadvertently result in a circular chromosome. The key to the linear worst-case complexity of the halving algorithm is that this check may be made in constant time.

Along with the multiplicity of solutions caused by different possible constructions of supernatural graphs, within such graphs and within the natural graphs, there may be many ways of drawing the gray edges. Without repeating here the lengthy details of the halving algorithm, it suffices to note that these alternate ways can be generated by choosing one of the vertices within each supernatural graph as a starting point.

4.3. *Genome halving with outgroups*

Let T be a genome consisting of ψ chromosomes and $2n$ genes $a_1^{(1)}, \ldots, a_n^{(1)}; a_1^{(2)}, \ldots, a_n^{(2)}$, dispersed in any order on the chromosomes, where for each i, genes $a_i^{(1)}$ and $a_i^{(2)}$ are duplicates. Any genome R is a reference or outgroup genome for T if it contains the n genes $a_1, \ldots, a_n$.

Let R be a reference genome for T. The GGH problem with one outgroup is to find a potential ancestral genome A such that some $d(R, A) + d(A' \oplus A'', T)$ is minimal. In practice, A is either one of the solutions to the unconstrained halving problem, or it is close to such a solution,[18] so little is lost in restricting our search to the set of solutions of the genome halving problem for T.

One strategy, suitable for small datasets, as in Ref. 11, is to generate the entire set $\mathbf{S}$ of genome halving solutions of T, then to evaluate each $A \in \mathbf{S}$ to find the one that minimizes $d(R, A)$.

When $\mathbf{S}$ is so large that it is not feasible to generate all of $\mathbf{S}$ in order to find the best A, we may resort to sampling $\mathbf{S}$, as in Ref. 12. In defining the gray edges in the supernatural graphs of Sec. 4.2, we generally have several choices at some of the steps. By randomizing this choice, we are effectively choosing a random sample of $X \in \mathbf{S}$.

5. The GGH Algorithm

The key idea in our improvement over brute force algorithms is to incorporate information from R during the halving process. It is important to take advantage of the common structure in T and R as early as possible, before it can be destroyed in the course of construction. To this end, we drop the practice of completing all the gray edges in one supernatural graph before starting another. We simply look for elements of common structure and add gray edges accordingly, always making sure that no circular chromosomes are inadvertently created.

Missing homologs. The halving algorithm requires full gene sets at several steps in reconstructing the ancestor, so we algorithmically restore the missing homologs to the most appropriate positions in T and R at the outset. The criterion for restoring a gene to a position in a genome is the net decrease in the number of disrupted adjacencies in the three-way comparison of the augmented genomes versus the situation before the gene was restored. Note that the fictional genes thus included do not count in the main GGH algorithm when it comes to choosing among steps of equal priority.

Paths. We define a path to be any connected fragment of a breakpoint graph, namely any connected fragment of a cycle. We represent each path by an unordered pair $(u, v) = (v, u)$ consisting of its current endpoints, though we keep track of all its vertices and edges. Initially, each black edge in T is a path, and each black edge in R is a path.

Pathgroups. A pathgroup Γ is an ordered triple of paths, two in T and one in R, where one endpoint of one of the paths in T is the duplicate of one endpoint of the other path in T, and both are orthologous to one of the endpoints of the path in R. The other endpoints may be duplicates or orthologs to each other, or not.

5.1. *The algorithms*

In adding pairs of gray edges to connect duplicate pairs of terms in the breakpoint graph of T versus $A' \oplus A''$ (which is being constructed), our approach is basically greedy, but with a sophisticated look-ahead. We can distinguish five different levels of desirability, or priority, among potential gray edges, i.e. potential adjacencies in the ancestor.

Recall that in constructing the ancestor A to be close to the outgroup R, such that $A' \oplus A''$ is simultaneously close to T, we must create as many cycles as possible in the breakpoint graphs between A and R and in the breakpoint graph of $A' \oplus A''$ versus T.

(1) Adding two gray edges would create two cycles in the breakpoint graph defined by T and $A' \oplus A''$, by closing two paths. When this possibility exists, it must

be realized, since it is an obligatory choice in any genome halving algorithm. It may or may not also create cycles in the breakpoint graph comparison of X with the outgroup, but this does not affect its priority.

(2) Adding two gray edges would create two cycles, one for T and one for the outgroup.

(3) Adding two gray edges would create a cycle in the T versus $A' \oplus A''$ comparison, but none for the outgroup. It would, however, create a higher priority pathgroup.

(4) Adding two gray edges would create a cycle in the T versus $A' \oplus A''$ comparison, but none for the outgroup, nor would it create any higher priority pathgroup.

(5) Each remaining path terminates in duplicate terms, which cannot be connected to form a cycle, since in $A' \oplus A''$ these must be on different (and identical) chromosomes. In supernatural graphs containing such paths, there is always another path and adding two gray edges between the endpoints of the two paths can create a cycle.

In not completing each supernatural graph before moving on to another, we lose the advantage in Ref. 2 of a constant time check against creating circular chromosomes. The worst case becomes a linear time check. This is a small liability, because the worst case scenario is seldom realized, the check almost always requiring only one or two steps.

6. GGH Results and Discussion

Our data consisted of 6144 gene sets, of which only 2104 were full sets. There were only 836 defective sets by virtue of a missing ortholog in R, while 3204 genes lacked one paralog in T.

Table 1 shows the results of the analysis on the full gene sets only, on combinations of the full sets with one kind of defective sets, and all three sets. For each case we study not only the reconstructed ancestor but also a "projected" version where genes from the defective sets are simply erased, in order to assess the changes in gene order due to the defective gene sets. Whereas the distance between each T and its reconstructed ancestor A is given by GGH, the distance between projected ancestor and T required a heuristic, explained in detail in Refs. 19 and 20, for attributing each paralog in T to one of the two copies of the ancestral genome. Note that we choose only one optimal ancestor A for each analysis; this does not affect d for the comparison between $A \oplus A$ and *Populus*, but it may have a very small effect on b, and r and on all three quantities for the comparison between A and *Vitis*.

Figure 5 depicts the result of analyzing all the 6144 gene sets with GGH, although the 836 genes with no grape orthologs are not visible. This is just one

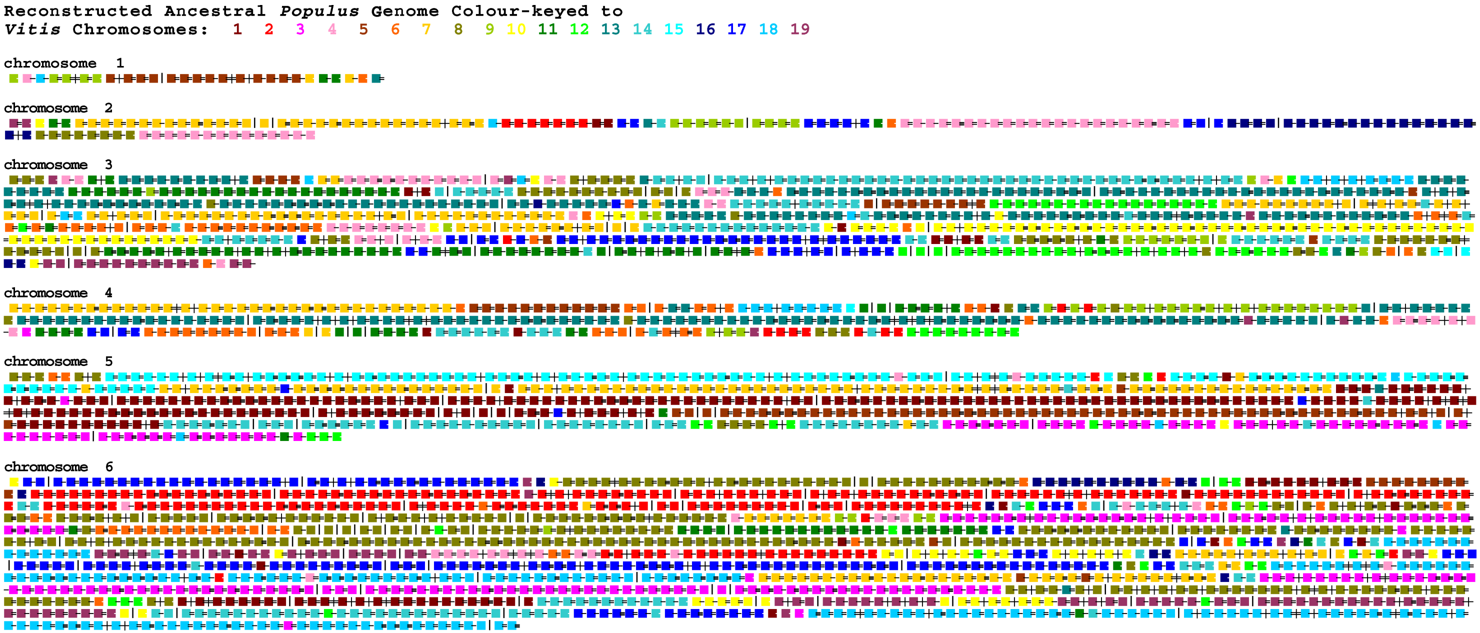

Fig. 5. Ten chromosomes (wrapped) of ancestral poplar genome reconstructed by GGH algorithm from 6144 full and defective gene sets. Only the genes with grape orthologs are indicated. Adjacencies present three times, i.e. twice in poplar and once in grape, are indicated by ≡, those present twice by = and those present once by −. Intrachromosomal breakpoints within segments indicated by |.

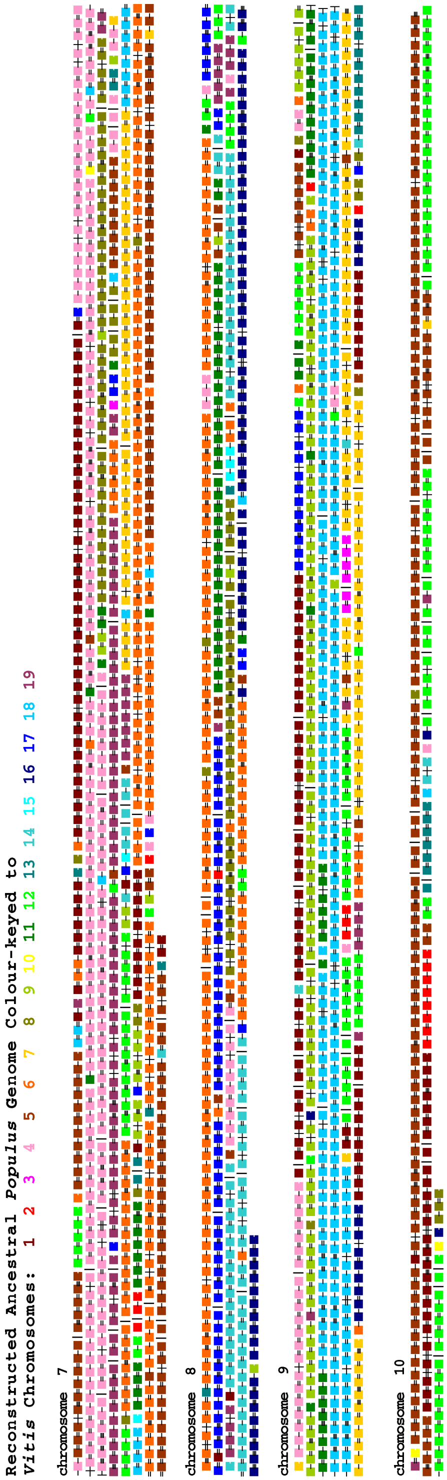

Fig. 5. (*Continued*)

Table 1. Comparisons of the reconstructed immediate pre-doubling ancestor A with the *Vitis* genome and of the immediate doubled ancestor $A \oplus A$ with *Populus*.

	Genes	$d\ (A, Vitis)$			$d\ (A \oplus A, Populus)$		
Data sets	in A	d	b	r	d	b	r
PPV	2104	638	751	1.70	454	690	1.32
PPV, PP	2940	649	757	1.71	737	1090	1.35
projected	2104	649	757	1.71	581	823	1.41
PPV, PV	5308	1180	1331	1.77	1083	1457	1.49
projected	2104	663	758	1.75	670	833	1.61
PPV, PP, PV	6144	1208	1363	1.77	1337	1812	1.48
projected	2104	664	757	1.75	750	926	1.62
without singletons							
PPV	2020	560	661	1.69	346	541	1.28
PPV, PP	2729	594	690	1.72	453	714	1.27
projected	2006	571	664	1.72	416	628	1.32
PPV, PV	4203	573	686	1.67	751	1031	1.46
projected	1955	489	580	1.69	490	644	1.52
PPV, PP, PV	4710	675	797	1.69	856	1211	1.41
projected	1986	528	622	1.70	558	744	1.50

PPV: full gene sets, PP: defective, missing grape ortholog, PV: defective, missing one poplar paralog. Projected: genes not in PPV ancestor deleted from solution A, d: genomic distance, b: number of breakpoints, $r = 2d/b$: the re-use statistic.

of many equally parsimonious solutions of the GGH problem, differing largely in how they concatenate chromosomal segments where there are two different possibilities suggested by the *Populus* genome and a third by the *Vitis* genome. The reconstruction is given as an example, and the details, including the telomeric positions determining the size of the chromosomes, are not definitive.

Despite this ambiguity in the reconstruction, the numerical results on d, b and r are quite robust, and can be used for the comparison of genomes and for evaluating methods.

The large number of singleton genes disrupting otherwise homogeneous synteny-blocks suggests that “noise” due to uncertainties inherent in homology identification and especially orthology identification may be artifactually inflating genomic distance d and the number of breakpoints b. Since the rigorous noise elimination techniques of Refs. 21 and 22, which are the gene-order equivalent of synteny block construction methods for genome sequences, have not yet been extended in the context of genome doubling, we simply identified singletons as gene sets lacking two real (i.e. not inferred from **insertMH**; see algorithm below) common adjacencies out of six possible cases in the original genomes, and ran all the analyses again without these genes.

In each case, we counted the breakpoints and calculated the appropriate genomic distance d, i.e. from the doubled ancestor to *Populus* and from the undoubled version of the same ancestor to *Vitis*.

Algorithm: GGH

Guided Genome Halving with Full and Defective Gene Sets

```
Input. Two genomes: duplication descendant $T'$, outgroup genome $R'$, where
    each gene has three homologs (full set) or two homologs (defective set), in the
    patterns TTR, TT or TR.
Output. Augmented genomes $T$, and $R$, where all gene sets are full, and
    Genome $A$, a halving solution of $T$, minimizing $d(A' \oplus A'', T) + d(A, R)$.
insertMH
Initialize paths (black edges) in $T$ and $R$.
Construct supernatural graphs.
Construct two pathgroups for each gene $g$ in $R$, one based on $g_t$, the other
    on $g_h$.
If number of chromosomes in $T$ is odd,
   add pathgroup with two paths of form $(\mathrm{end}, \mathrm{end})$.
While there remains at least one pathgroup
   For each pathgroup $((x, y), (\bar{x}, z), (x, m))$
   classify it by case and priority, and find a pathgroup $\Gamma$ that has the highest
   priority. To choose among Priority 2 pathgroups, find one that maximizes the
   number of "real" black edges, i.e. edges in $T'$ and $R'$, not just edges created by
   insertMH. Similarly for Priority 3 pathgroups.
   Case 1: $\bar{x} \neq y$, and adding $xy$ and $\bar{x}\bar{y}$ would not create a circular chromosome.
      Priority 1: $z = \bar{y}$.
      Priority 2: $y = m$.
      Priority 3: adding $xy$ and $\bar{x}\bar{y}$ would create a pathgroup with priority 2.
      Priority 4: None of 1, 2 or 3.
   Case 2: $\bar{x} \neq y$, and adding $x\bar{z}$ and $\bar{x}z$ would not create a circular chromosome.
      Priority 2: $z = m$.
      Priority 3: adding $x\bar{z}$ and $\bar{x}z$ would create a pathgroup with priority 2.
      Priority 4: Neither of 2 or 3.
   Case 3:$\bar{x} = y$.
      Priority 5:
   If $\Gamma$ is Case 1, addGrayEdge$(xy, \bar{x}\bar{y})$.
   If $\Gamma$ is Case 2, addGrayEdge$(x\bar{z}, \bar{x}z)$.
   If $\Gamma$ is Case 3, find some
      $W = ((w, \bar{w}), (\bar{w}, w), (w, s))$ in the same supernatural graph and
      addGrayEdge$(xw, \bar{x}\bar{w})$.
```

Algorithm: addGrayEdge$(rt, \bar{r}\bar{t})$

```
Add gray edges $rt, \bar{r}\bar{t}$ to partially completed genome $X$” $\oplus X''$.
Add gray edge $rt$ to partially completed genome $X$.
Update paths in pathgroups that are affected by the new gray edges.
Remove pathgroups that start with $r$ and $t$.
```

Algorithm: insertMH

Insert Missing Homologs in Chromosomes

```
Input. Two genomes: duplication descendant $T'$, outgroup $R'$, where each gene
    has two or three homologs, in the patterns TTR, TT, TR.
Output. Augmented genomes $T$ and $R$ containing exactly three homologs for
    each gene, in the pattern TTR, maximizing the number of common edges of
    form $\{a_1, b_1\}$, $\{a_2, b_2\}$ in $T$ and $\{a, b\}$ in $R$.
    (Or $\{a_1, b_2\}$, $\{a_2, b_1\}$ in $T$ and $\{a, b\}$ in $R$.)
While there are genes that have only two copies, count edgeDiff for each
    such, which simultaneously finds the BestPosition.
    Insert the gene with the minimum edgeDiff value into the BestPosition of this
        gene.
```

Algorithm: count edgeDiff

```
If a gene $g$ just has one copy ($g_1$) in $T'$ and one copy ($g$) in $R'$, then we must
    insert another copy ($g_2$) into $T'$.
If a gene $g$ just has two copies ($g_1, g_2$) in $T'$, then we must insert $g$ into $R'$.

(The details are omitted here. This is essentially a greedy heuristic to add
adjacencies reflecting, as if possible, adjacencies already existing in $R'$ and $T'$.)
```

This enabled us to calculate the "breakpoint re-use" statistic $r = 2d/b$, which is a measure of how much signal about conserved order (among segments, not within segments) remains in the comparison of two genomes after a period of evolutionary rearrangements. When $r = 1$, we can have high confidence in the rearrangement distance and history. When r approaches two, the segment order in the two genomes being compared are essentially random with respect to each other, i.e. calculating r for random genomes gives a value approaching 2[a]. In Table 1, we see both from changes in d and changes in r that

- most of the signal contained in the order among conserved chromosomal segments has been lost between the ancestor and *Vitis*, but is retained to a great degree

[a] If breakpoints are frequently re-used during evolution, then r will also be close to 2; unfortunately there is no internal way of testing the breakpoint re-use hypothesis against the null hypothesis of complete loss of signal about segment order.[23]

between the ancestor and *Populus*, probably reflecting the difference in divergence time but also possible biases towards T in the GGH algorithm,[19]

- the addition of the defective PV gene sets degrades the analysis, more than the addition of PP sets, though this may be due to the four times greater number of gene sets in the former,
- the elimination of singletons improves all the analyses, but where PV is present, this comes about largely by discarding most of the sets, which turn out to be singletons.

The analysis with 6144 gene sets required almost 48 h on a MacBook, but this was anomalously large, since those with 4000 or 5000 required less than five hours and those with 2000 about one hour. Much of the running time is due to the check on the number of real edges in a pathgroup to choose among Priority 2 or among Priority 3 options. This could be reduced by optimizing data structures in our software.

7. Conclusions

We have formalized a new way of assessing neighborhood selection constraints on duplicate gene loss, taking into account the particular genomic structure of descendants of WGD. Dividing possible selective effects into those affecting which gene pairs will lose one member, and those affecting which member of the pair is lost, we found no effect of the first type, but a clear effect of the second.

With the application of our GGH method to more than 6000 gene sets, we have shown that any realistic case of genome doubling should be amenable, even if all the gene paralogs remain in the sequenced descendant.

The reconstruction of long conserved segments in Fig. 5 attests to the coherence of orthoMCL homolog sorting and GGH.

The inclusion of defective PV gene sets would appear to add little more than noise to the analysis, but the PP sets would seem to add significant information, especially to the ancestor–*Populus* comparison.

The elimination of singletons proves to be a meaningful way of drastically decreasing the number of segments (as measured by b) and the genomic distance to credible levels, though this still does not result in a detectible signal in the ancestor–*Vitis* comparison.

The recently sequenced *Carica papaya* genome,[24] which is phylogenetically more closely related to *Populus*, but like *Vitis* is diverged before the *Populus* doubling event, can also play the outgroup role in our analysis,[18] but accuracy is diminished since genome assembly has not been completed. In general, as sequences become more polished and further complete homology sets can be more accurately detected, our methods should become more accurate.

Acknowledgments

The research was supported in part by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC). DS holds the Canada Research Chair in Mathematical Genomics.

References

1. El-Mabrouk N, Bryant D, Sankoff D, Reconstructing the pre-doubling genome, in Istrail S, Pevzner P, Waterman M (eds.), *Third Annual International Conference on Computational Molecular Biology (RECOMB 99)*, ACM Press, New York, pp. 154–163, 1999.
2. El-Mabrouk N, Sankoff D, The reconstruction of doubled genomes, *SIAM Journal on Computing* **32**:754–792, 2003.
3. Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A, Albert VA, Ma H, dePamphilis CW, Widespread genome duplications throughout the history of flowering plants, *Genome Res* **16**:738–749, 2006.
4. Tuskan GA, Difazio S, Jansson S *et al.*, The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray), *Science* **313**:1596–1604, 2006. http://genome.jgi-psf.org/Poptr1/Poptr1.download.html
5. Velasco R, Zharkikh A, Troggio M *et al.*, A high quality draft consensus sequence of the genome of a heterozygous grapevine variety, *PLoS ONE* **2**:e1326, 2007.
6. Jaillon O, Aury JM, Noel B *et al.*, French-Italian Public Consortium for Grapevine Genome Characterization, The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla, *Nature* **449**:463–467, 2007. http://www.genoscope.cns.fr/externe/English/Projets/Projet_ML/data/annotation/
7. Li L, Stoeckert CJ Jr, Roos DS, OrthoMCL: Identification of ortholog groups for eukaryotic genomes, *Genome Res* **13**:2178–2189, 2003.
8. van Hoek MJ, Hogeweg P, The role of mutational dynamics in genome shrinkage, *Mol Biol Evol* **24**:2485–2494, 2007.
9. Byrnes JK, Morris GP, Li WH, Reorganization of adjacent gene relationships in yeast genomes by whole-genome duplication and gene deletion, *Mol Biol Evol* **23**:1136–1143, 2006.
10. Poyatos JF, Hurst LD, The determinants of gene order conservation in yeasts, *Genome Biol* **8**:R233, 2007.
11. Zheng C, Zhu Q, Sankoff D, Genome halving with an outgroup, *Evol Bioinform* **2**:319–326, 2006.
12. Sankoff D, Zheng C, Zhu Q, Polyploids, genome halving and phylogeny, *Bioinformatics* **23**:i433–i439, 2007.
13. Zheng C, Zhu Q, Adam Z, Sankoff D, Guided genome halving: Hardness, heuristics and the history of the Hemiascomycetes, *Bioinformatics* **24**(13):96–104, 2008.
14. Yancopoulos S, Attie O, Friedberg R, Efficient sorting of genomic permutations by translocation, inversion and block interchange, *Bioinformatics* **21**:3340–3346, 2005.
15. Bergeron A, Mixtacki J, Stoye J, A unifying view of genome rearrangements, in Bücher P, Moret BME (eds.), *Workshop on Algorithms in Bioinformatics (WABI 2006)*, *Lecture Notes in Computer Science 4175*, pp. 163–173, 2006.
16. Bafna V, Pevzner P, Genome rearrangements and sorting by reversals, *SIAM Journal of Computing* **25**:272–289, 1996.
17. Tesler G, Efficient algorithms for multichromosomal genome rearrangements, *J Comput Syst Sci* **65**: 587–609, 2002.
18. Sankoff D, Zheng C, Wall PK, dePamphilis C, Leebens-Mack J, Albert VA, Internal validation of ancestral gene order reconstruction in angiosperm phylogeny, in Nelson C, Viallette S (eds.), *Proceedings of the RECOMB Satellite Conference on Comparative Genomics (RECOMB CG 2008)*, *Lecture Notes in Bioinformatics 5267*, pp. 252–264, 2008.

19. Zheng C, Zhu Q, Sankoff D, Descendants of whole genome duplication within gene order phylogeny, *J Computational Biology* **15**:947–964, 2008.
20. Tannier E, Zheng C, Sankoff D, Multichromosomal median and halving problems under different genomic distances. In: Crandall KA, Lagergren J (eds.), *Workshop on Algorithms in Bioinformatics (WABI 2008), Lecture Notes in Bioinformatics 5251*, pp. 1–13, 2008.
21. Zheng C, Zhu Q, Sankoff D, Removing noise and ambiguities from comparative maps in rearrangement analysis, *Trans. Comput. Biol. Bioinform.* **4**:515–522, 2007.
22. Choi V, Zheng C, Zhu Q, Sankoff D, Algorithms for the extraction of synteny blocks from comparative maps, in Giancarlo R, Hannenhalli S (eds.), *Workshop on Algorithms in Bioinformatics (WABI 2007), Lecture Notes in Bioinformatics 4645*, pp. 277–288, 2007.
23. Sankoff D, The signal in the genomes, *PLoS Comput. Biol* **2**:e35, 2006.
24. Ming R, Hou S, Feng Y *et al.*, The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya Linnaeus*), *Nature* **452**:991–9966, 2008. http://asgpb.mhpcc.hawaii.edu

Chunfang Zheng received her Bachelor's degree in Biology from Beijing Sports University, China, in 1989, her Bachelor's degree in Computer Science and her Master's degree from the University of Ottawa, Canada, in 2001 and 2005, respectively. She completed her Ph.D. in 2009 in the Biology Department at the University of Ottawa. She has published several computational biology papers on partially ordered genomes, the genome halving problem, and removing noise from comparative maps.

P. Kerr Wall received his Ph.D. in Bioinformatics from Penn State University, USA, under the direction of Claude dePamphilis and Webb Miller. He has been the lead bioinformatics programmer for the Floral Genome Project, and continued with the Ancestral Angiosperm Genome Project until recently moving to a new position in industry with BASF Plant Biology.

James Leebens-Mack received his Ph.D. in Botany from the University of Texas, USA. He currently holds a position in the Plant Biology at the University of Georgia, USA. His research employs computational and phylogenomic approaches to explore the ecological, genetic and developmental processes that contribute to phenotypic diversification and speciation in flowering plants.

Claude dePamphilis received his Ph.D. in Botany from the University of Georgia, USA, under the direction of Robert Wyatt. He is currently Professor of Biology at Penn State University, USA, where he is the Principal Investigator of the Floral Genome Project and the Ancestral Angiosperm Genome Project. His research interests are in genome evolution, ancestral reconstruction, and bioinformatics, with focus on early angiosperms and parasitic plants.

Victor A. Albert is Empire Innovation Professor of Biological Sciences at the University at Buffalo (SUNY), USA. He is also a Faculty Investigator at the New York State Center of Excellence in Bioinformatics and Life Sciences. His current research interests are in developmental and genomic evolution in plants.

David Sankoff received his Ph.D. degree in Mathematics from McGill University, Canada, and has been a member of the Centre de recherche mathmatiques in Montreal for many years. He currently holds the Canada Research Chair in Mathematical Genomics in the Mathematics and Statistics Department at the University of Ottawa, and is cross-appointed to the Biology Department and the School of Information Technology and Engineering. His research interests include comparative genomics, particularly probability models, statistics, and algorithms for genome rearrangements.