JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 16, Number 10, 2009

Pp. 1353–1367
DOI: 10.1089/cmb.2009.0103

Research Article

Towards Improved Reconstruction of Ancestral Gene Order in Angiosperm Phylogeny

DAVID SANKOFF,[1] CHUNFANG ZHENG,[2] P. KERR WALL,[3] CLAUDE DEPAMPHILIS,[3] JIM LEEBENS-MACK,[4] and VICTOR A. ALBERT[5]

ABSTRACT

Whole genome doubling (WGD), a frequent occurrence during the evolution of the angiopsperms, complicates ancestral gene order reconstruction due to the multiplicity of solutions to the genome halving process. Using the genome of a related species (the outgroup) to guide the halving of a WGD descendant attenuates this problem. We investigate a battery of techniques for further improvement, including an unbiased version of the guided genome halving algorithm, reference to two related genomes instead of only one to guide the reconstruction, use of draft genome sequences in contig form only, incorporation of incomplete sets of homology correspondences among the genomes, and addition of large numbers of "singleton" correspondences. We make use of genomic distance, breakpoint reuse rate, dispersion of sets of alternate solutions, and other means to evaluate these techniques, and employ the papaya (*Carica papaya*) and grapevine (*Vitis vinifera*) genomes to reconstruct the pre-WGD ancestor of poplar (*Populus trichocarpa*), as well as an early rosid ancestor. A significant result is that the papaya genome has rearranged at a greater rate from the rosid ancestor than phylogenetic relationships would predict.

Key words: gene order, genome halving, genome rearrangement, *Populus trichocarpa*, whole genome duplication.

1. INTRODUCTION

THE RECONSTRUCTION OF THE GENE ORDER in ancestral genomes requires that we make a number of choices, among the data on which to base the reconstruction, in the algorithm to use, and in how to evaluate the result. In this article, we illustrate an approach to making these choices in the reconstruction of the ancestor of the poplar *Populus trichocarpa* genome. This species has undergone whole genome duplication followed by extensive chromosomal rearrangement, and is one of four angiosperm genomes, along with those of *Carica papaya* (papaya), *Vitis vinifera* (grapevine), and *Arabidopsis thaliana*, whose sequences have been published to date (Fig. 1).

We have been developing methods to incorporate descendants of whole genome doubling into phylogenies of species that have been unaffected by the doubling event. The basic tool in analyzing

Departments of [1]Mathematics and [2]Biology, University of Ottawa, Ottawa, Canada.
[3]Biology Department, Penn State University, University Park, Pennsylvania.
[4]Department of Plant Biology, University of Georgia, Athens, Georgia.
[5]Department of Biological Sciences, SUNY Buffalo, Buffalo, New York.

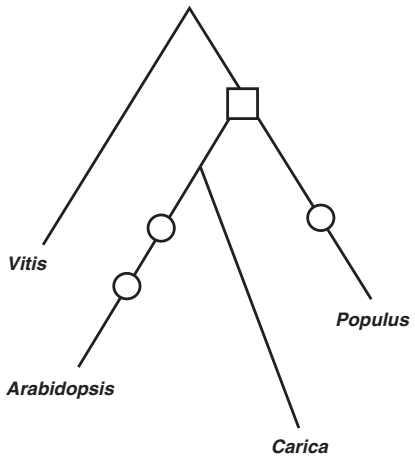

FIG. 1. Phylogenetic relationships among angiosperms with sequenced genomes. The circles indicate likely whole genome doubling events. The circle in the *Populus* lineage, representing the locus of the WGD event at the origin of the willow-poplar family, and the square, representing the ancestor of the rosid dicotyledons, indicate the target ancestors we reconstruct in this article.

descendants of whole genome doubling is the halving algorithm (El-Mabrouk and Sankoff, 2003). To overcome the propensity of the genome halving procedure to produce numerous, widely disparate solutions, we ''guide'' the execution of this procedure with information from genomes of related species (Zheng et al., 2006, 2008a,b, 2009; Sankoff et al., 2009), which we call outgroups. This, *ipso facto*, integrates the whole genome doubling descendant into the phylogeny of the related species.

Issues pertaining to data include the following:

Homology sets. Can we use defective sets of homologs, i.e., which have only one copy in the duplicated genome or are missing the ortholog completely in the guide genome?

Singletons. Should we purge singletons from the data, i.e., sets of homologous markers who have no homologous adjacent markers in common in the either the duplicated genome or the outgroup?

Contigs. Can we use guide genomes that are not fully assembled, but are available only as sets of hundreds or thousands of contigs?

Another choice to be made during reconstruction has to do with the guided halving algorithm itself. The original genome halving problem, with no reference to outgroup genomes, can be solved in time linear in the number of markers (El-Mabrouk and Sankoff, 2003). We can introduce information from an outgroup in order to guide this solution, without compromising the optimality of the result and without serious increase in computing time (Zheng et al., 2008a, 2009). We call this *constrained* guided halving. The true, *unconstrained*, guided halving problem, however, where the solution ancestor need not be a solution of the original halving problem, is likely to be NP-hard (Tannier et al., 2009). In the heuristics necessary for these two approaches, there is a trade-off between the speed of constrained halving versus the (theoretically) better solution obtainable by unconstrained halving.

Once we make our choices of data and algorithm, we may ask how to evaluate the results. As with most evolutionary reconstructions, this evaluation is necessarily completely internal, since there is no outside reference to check against, except simulations. There are many indices for evaluating a reconstruction:

Distance. Most important, there is the objective function; here our genomic distance definition attempts to recover the most economical explanation of the observed data, namely the minimum number of rearrangement events (reversals, reciprocal translocations, chromosome fusions/fissions, transpositions) required.

Reuse rate. Each rearrangement operation can create at most two breakpoints in the gene-by-gene alignment of a genome and its ancestor. When fewer than two are created, one or two pre-existing breakpoint(s) must be ''reused.'' Conversely, when rearranged genomes are optimally reconstructed, some breakpoints may be reused. In fact, breakpoint re-use is inferred far more frequently in reconstruction than it actually occurs in genome generation, and is actually a measure of the loss of evolutionary signal inherent in the gene order.

Dispersion. The motivation for guided halving is to resolve the ambiguities inherent in the large number of optimal halving solutions. One way to quantify the remaining non-uniqueness is to calculate the distances among a sample of

TABLE 1. GUIDED HALVING SOLUTIONS WITH AND WITHOUT SINGLETONS, CONSTRAINED AND UNCONSTRAINED HEURISTICS, *VITIS* OR *CARICA* AS OUTGROUP, AND ALL COMBINATIONS OF FULL AND DEFECTIVE HOMOLGY SETS

		d(A, Vitis)			d(A + A, Populus)			
Data sets	*Genes in* A, *with singletons*	d	b	r	d	b	r	*Total* d
Solutions constrained to also be solutions of genome halving								
PPV	2104	638	751	1.70	454	690	1.32	1092
PPV, PP	2940	649	757	1.71	737	1090	1.35	1386
PPV, PV	5308	1180	1331	1.77	1083	1457	1.49	2263
PPV, PP, PV	6144	1208	1363	1.77	1337	1812	1.48	2545
Solutions unconstrained								
PPV	2104	593	734	1.62	512	733	1.40	1105
PPV, PP	2940	616	752	1.64	778	1119	1.39	1394
PPV, PV	5308	1121	1307	1.72	1147	1486	1.54	2268
PPV, PP, PV	6144	1129	1328	1.70	1437	1871	1.54	2566
		d(A, Carica)			d(A ⊕ A, Populus)			
Data sets	*Genes in* A, *with singletons*	d	b	r	d	b	r	*Total* d
Solutions constrained to also be solutions of genome halving								
PPC	2590	896	1075	1.67	565	823	1.37	1461
PPC, PP	3478	905	1085	1.67	884	1282	1.38	1789
PPC, PC	6334	1892	2224	1.70	1262	1700	1.48	3154
PPC, PP, PC	7222	1925	2241	1.72	1541	2065	1.49	3466
Solutions unconstrained								
PPC	2590	864	1043	1.66	628	870	1.44	1492
PPC, PP	3478	873	1039	1.68	951	1318	1.44	1824
PPC, PC	6334	1859	2172	1.71	1321	1742	1.52	3180
PPC, PP, PC	7222	1877	2211	1.70	1617	2126	1.52	3494
		d(A, Vitis)			d(A ⊕ A, Populus)			
Data sets	*Genes in* A, *without singletons*	d	b	r	d	b	r	*Total* d
Solutions constrained to also be solutions of genome halving								
PPV	2020	560	661	1.69	346	541	1.28	906
PPV, PP	2729	594	690	1.72	453	714	1.27	1047
PPV, PV	4203	573	686	1.67	751	1031	1.46	1324
PPV, PP, PV	4710	675	797	1.69	856	1211	1.41	1531
Solutions unconstrained								
PPV	2020	545	652	1.67	375	564	1.33	920
PPV, PP	2729	567	681	1.67	493	745	1.32	1060
PPV, PV	4203	544	674	1.61	782	1034	1.51	1326
PPV, PP, PV	4710	631	785	1.61	916	1250	1.47	1547
		d(A, Carica)			d(A ⊕ A, Populus)			
Data sets	*Genes in* A, *without singletons*	d	b	r	d	b	r	*Total* d
Solutions constrained to also be solutions of genome halving								
PPC	2464	772	934	1.65	412	607	1.36	1184
PPC, PP	3226	812	981	1.66	536	809	1.33	1348
PPC, PC	4651	779	926	1.68	774	1050	1.47	1554
PPC, PP, PC	5234	898	1088	1.65	892	1253	1.42	1790
Solutions unconstrained								
PPC	2464	758	921	1.65	454	639	1.42	1212
PPC, PP	3226	796	967	1.65	584	839	1.39	1380
PPC, PC	4651	764	911	1.68	804	1090	1.48	1568
PPC, PP, PC	5234	861	1058	1.63	952	1303	1.46	1813

A, pre-doubling ancestor of *Populus;* $A \oplus A$, doubled ancestor; PPV, PPC, full gene sets; PP, defective, missing grape or papaya ortholog; PV, PC, defective, missing one poplar paralog; *d*, genomic distance; *b*, number of breakpoints; $r = 2d/b$, reuse statistic.

solutions. Thus, we can compare the average distance between the alternate solutions in one method to the average in another, to see which is the less dispersed, or more compact. And we can compare these "within-group" distances to "between-group" distances, to assess statistically how much our methodological choices affect the results.

In this article, we will refer repeatedly to a main tabulation of results, Table 1, in which we discover the unexpected rapid evolution of the *Carica* gene order in comparison with that of *Vitis*. In Section 2, we report on the origin and processing of our gene-order data and the construction of the full and defective homology sets. In Section 3, we take up the discussion of our measures for assessing the quality of reconstructions. Then, in Section 4, we discuss the halving problems, and sketch a new algorithm for unconstrained guided halving. In Section 5, we evaluate the utility of singletons and of defective homology sets. Then, in Section 6, we assess the two guided halving algorithms on real and simulated data. Section 7 proposes a way to use unassembled genome sequence in contig form, as a input to the reconstruction algorithm, an approach that could potentially have wide use in gene order phylogeny. In Section 8, we demonstrate the phylogenetic validity of reconstructing the *Populus* ancestor using either *Vitis* or *Carica*, or both, as outgroups. Note that we have not included *Arabidopsis* in our analyses; as will be explained in Section 9, this was precluded by algorithmic complications due to two rounds of whole genome duplications and by a paucity of data in the appropriate configurations.

2. THE POPULUS, VITIS, AND CARICA DATA

Annotations for the *Populus*, *Vitis*, and *Carica* genomes were obtained from databases maintained by the U.S. Department of Energy's Joint Genome Institute (Tuskan et al., 2006), the French National Sequencing Center, Genoscope (Jaillon et al., 2007), and the University of Hawaii (Ming et al., 2008), respectively. An all-by-all BLASTP search was run on a data set that included all *Populus* and *Vitis* protein coding genes, and orthoMCL (Li et al., 2003) was used to construct 2104 full and 4040 defective gene sets, in the first case, denoted PPV, containing two poplar paralogs (genome P) and one grape ortholog (genome V), and in the second case, denoted PV or PP, missing a copy from either P or V. This was repeated with *Populus* and *Carica*, genomes P and C, respectively, to obtain 2590 full (PPC) and 4632 defective (PC or PP) sets. The location on chromosomes (or contigs in the case of *Carica*) and orientation of these paralogs and orthologs was used to construct our database of gene orders for these genomes. Contigs containing only a single gene were discarded from the *Carica* data.

3. EVALUATION OF SOLUTIONS

Developing methods for historical inference about genomes is an uncertain enterprise, since there is usually no way of checking the results against historical truth, the fossil record being extremely fragmentary, vast evolutionary time scales generally precluding laboratory experimentation and simulation being extremely dependent on simplifying assumptions.

Nevertheless, there are meaningful evaluation criteria. Parsimonious explanations are to be preferred to uneconomical ones, or if there is an accepted probability model, most likely explanations are better than unlikely ones. Low variance estimates are better than high variance ones. And methods that allow internal tests of significance, e.g., the boot-strap, are desirable. Here we will discuss the three sorts of evaluation we use in this study.

3.1. Genome distance and breakpoint graph

The distance measures we use are based on parsimony. As such they are likely to produce underestimates of the number of rearrangements historically intervening between two genomes, especially if this number is large. Nevertheless, lacking a credible probabilistic model for rearrangement processes, we can rely on the current measures, as long as we do not forget the inherent bias towards "shorter" solutions.

Genome comparison algorithms generally involve manipulations of the bicolored breakpoint graph (Bafna and Pevzner, 1996; Tesler, 2002) of two genomes, called the black and the gray genomes, on the same set of n genes, where two vertices are defined representing the two ends of each gene, and an edge of one color joins two vertices if the corresponding gene ends are adjacent in the appropriate

genome. Omitting the details pertaining to the genes at the ends of chromosomes, the genomic distance d, i.e., the minimum number of rearrangements necessary to transform one genome into the other, satisfies $d = n - c$, where c is the number of alternating color cycles making up the breakpoint graph (Yancopoulos, *et al*. 1995).

It is well-known (Mazowita et al. 2006; Sinha and Meller, 2008) that, in practice, genomic distance depends strongly on the degree of resolution of the genomic data. The smaller the threshold for conserved segment size and the greater the number of segments, the greater the distance. This is true for b and r as well. We will have to take account of the dependence of d on n when we investigate the effects of singletons, types of homology class, algorithm version and outgroup in Section 5.

3.2. Breakpoint reuse

If d is the number of rearrangements and b the number of breakpoints, the reuse (Pevzner and Tesler, 2003) variable $r = 2d/b$ can take on values in $1 \leq r \leq 2$. Completely randomized genomes will have r close to 2, so that if an empirical comparison has $r \sim 2$, we cannot ascribe much significance to the details of the reconstruction (Sankoff, 2006). This is particularly likely to occur for genomes that are only very distantly related. In fact, studies of mammalian genomes (Sinha and Meller, 2008) have shown a very close correlation between r and d. This does not indicate an actual tendency towards breakpoint re-use throughout a phylogenetic domain, *since then r would be elevated even for closely related genomes*, but rather a loss of gene-order signal due to inadequate modeling of evolutionary processes and/or the reconstruction of homologous gene orders (Sankoff, 2006).

3.3. Dispersion

As we will see in Section 4, algorithms for reconstructing ancestral genomes generally allow two or more choices at many stages. It suffices to make this choice randomly to generate a sample of alternative solutions. The distances between these solutions are suggestive of the reliability of the method. A method that produces solutions within a few rearrangements of each other is preferable to one that generates a set of very heterogeneous solutions, as long as this improvement does not come with the cost of an increased bias.

4. GUIDED HALVING

The genome halving problem (El-Mabrouk and Sankoff, 2003) asks, given a genome T with two copies of each gene, distributed in any manner among the chromosomes, to find the "ancestral" genome, written $A \oplus A$, consisting of two identical halves, i.e., two identical sets of chromosomes with one copy of each gene in each half, such that the rearrangement distance $d(T, A \oplus A)$ between T and $A \oplus A$ is minimal. Note that part of this problem is to find an optimal labeling as "1" or "2" of the two genes in a pair of copies, so that all n copies labeled "1" are in one half of $A \oplus A$, and all those labeled "2" are in the other half. The genome A represents the ancestral genome at the moment immediately preceding the WGD event giving rise to $A \oplus A$.

The guided genome halving problem (Zheng et al., 2006) asks, given T as well as another genome R containing only one copy of each of the n genes, find A so that $d(T, A \oplus A) + d(A, R)$ is minimal. The solution A need not be a solution to the original halving problem.

In previous studies (Zheng et al., 2006, 2008a; Sankoff et al., 2007), we found that the solution of the guided halving problem is often a solution of the original halving problem as well, or within a few rearrangements of such a solution. This has led us to define a *constrained* version of the guided halving problem, namely to find A so that $A \oplus A$ is a solution to the original halving problem and $d(T, A \oplus A) + d(A, R)$ is minimal. This has the advantage that a good proportion of the computation, namely the halving aspect, is guaranteed to be rapid and exact, although the overall algorithm, which is essentially a search among all optimal A, remains heuristic. Without sketching out the details of the lengthy algorithm, the addition of gray edges representing genome A to the breakpoint graph, as in Figure 2, must favor configuration (b) over (c), even though there are as many cycles created by (c) as by (b). This is a consequence of the original halving theory in El-Mabrouk and Sankoff (2003). Otherwise $A \oplus A$ may not be a halving solution. This,

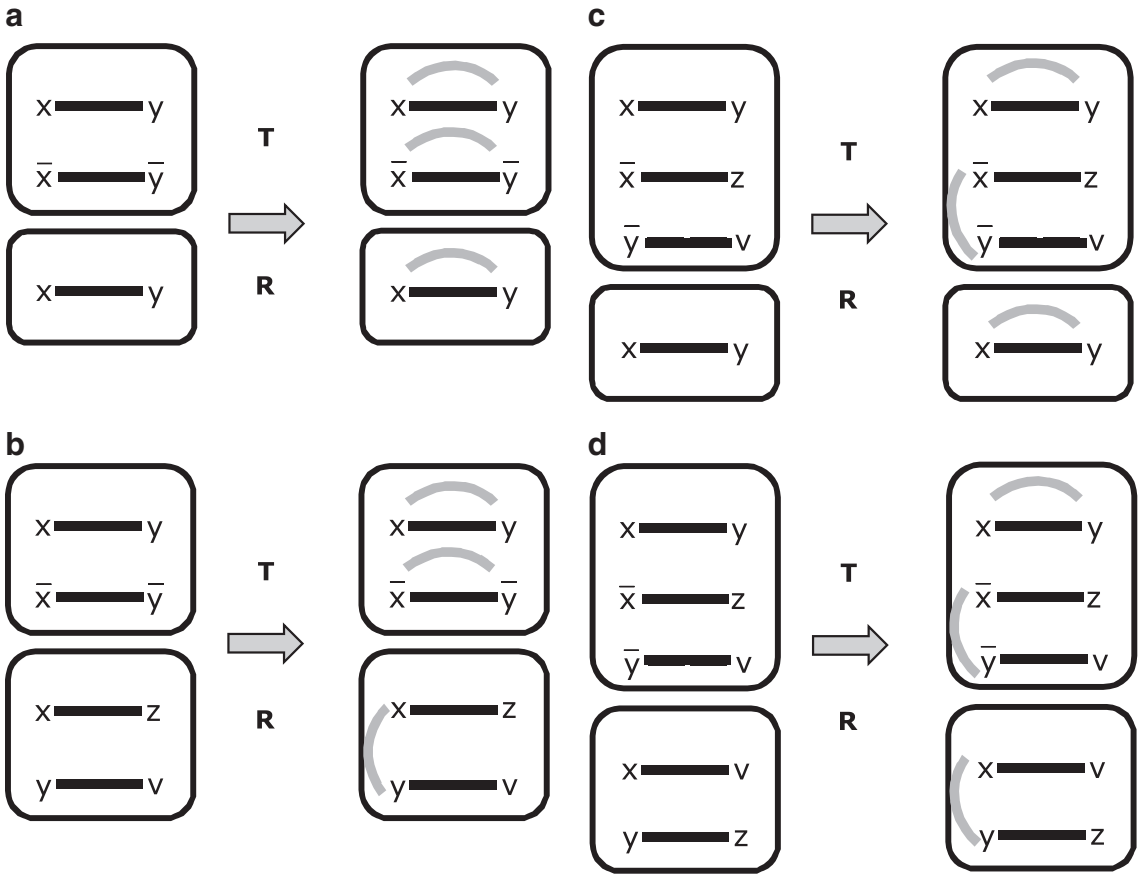

FIG. 2. Choice of gray edge to add at each stage of the reconstruction of A and $A \oplus \mathrm{A}$. Each black in the diagram represents either an adjacency in T or R or an alternating color path with a black edge at each endpoint. If vertex w is copy "1" in T, then $\bar{w}$ is copy "2," and vice versa. (**a**) Configuration requiring the creation of three cycles, two in the breakpoint graph of T and $A \oplus A$, and one in the breakpoint graph of A and R. (**b**) Configuration requiring the creation of two cycles in the breakpoint graph of T and $A \oplus A$, necessary for $A \oplus A$ to be a solution of the genome halving problem. (**c**) Alternative configuration in solution of guided halving $A \oplus A$ is not also required to be a solution of the halving problem. (**d**) Look-ahead when there are no configurations (a), (b), or (c). Here the addition of three gray edges creates a configuration (c).

however, may bias the reconstruction of A towards T and away from R. For the present work, we implemented a new version of the algorithm, as sketched in Section 4.1, treating configurations (b) and (c) equally in constructing A. The choice among two or more configurations of form (b) or (c) is based on a look-ahead calculation of what effect this choice will have on the remaining inventory of configurations of form (b) and (c). The new algorithm requires much more computation, but its objective function is better justified.

4.1. The new algorithm

First we define paths, which represent intermediate stages in the construction of the breakpoint graph comparing T and $A \oplus A$ and the breakpoint graph comparing A and R. Then we define pathgroups, which focus on the three current paths leading from three "homologous" vertices in the graph, namely two copies in T and one in R. Note that each vertex represents one of the two ends of a gene.

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, as in Figure 2, is an ordered triple of paths, two in the partially constructed breakpoint graph involving T and $A \oplus A$, and one in the partially constructed breakpoint graph involving R and A, 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.

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 careful look-ahead. We can distinguish four 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. At each step, we add three gray edges.

- **Priority 1.** Adding the three gray edges would create two cycles in the breakpoint graph defined by T and $A \oplus A$, by closing two paths, and one cycle in the breakpoint graph comparison of A with the outgroup, as in Figure 2a.
- **Priority 2.** Adding three gray edges would create two cycles, one for T and one for the outgroup, or two for T and none for the outgroup, as in Figure 2b,c.
- **Priority 3.** Adding the gray edges would create only one cycle, either in the T versus $A \oplus A$ comparison, or in the R versus A comparison. In addition, it would create a higher priority pathgroup, as in as in Figure 2d.
- **Priority 4.** Adding the gray edges would create only one cycle, but would not create any higher priority pathgroup.

The algorithm simply completes the steps suggested by the highest priority pathgroup currently available, choosing among equal priority pathgroups according to a look-ahead to the configuration of priorities resulting from competing moves.

At each step, we must verify that a circular chromosome is not created, otherwise the move is blocked. As in El-Mabrouk and Sankoff (2003), this check requires a constant time. The algorithm terminates when no more pathgroups can be completed. Any remaining pathgroups define additional chromosomes in the ancestor A.

5. ON THE UTILITY OF SINGLETONS AND DEFECTIVE HOMOLOGY SETS

From the last column of Table 1, it is clear that d varies widely as a function of the four factors, inclusion/exclusion of singletons, inclusion/exclusion of defective homology sets, outgroup species, and heuristic. But it is also clear that d depends on n, in the first numerical column in the table (Mazowita et al., 2006; Sinha and Meller, 2008). Thus, we must control for the dependence of d on n in teasing out the relative contribution of each of these factors. In Figure 3, we group the 32 points in the plot of d versus n, taken from the 32 rows in Table 1, according to choice of outgroup, inclusion or not of singletons, combination of homology classes and algorithm version. We will return to the almost imperceptible differences between the constrained and unconstrained algorithms in Section 6, and to the choice of outgroup in Section 8, but we can observe here that the inclusion of singletons has a dramatic effect on the rate of increase of d on n. Though this effect can theoretically be generated by rearrangements, in practice it is better considered as noise in the analysis (Choi et al, 2007; Zheng et al., 2007). This is confirmed by the greater values of r, indicating degradation of evolutionary signal, almost everywhere in the upper half of Table 1 compared to the lower half.

We also note that the increase in d caused by adding defective homology sets to the analysis is really due to the disproportionate numbers of singletons in these sets. The trend lines for the four different combinations of homology sets are parallel and steeply sloped. This slope is largely due to the presence of singletons in the data for the two highest point on each line, and also to the higher rate of evolution of *Carica* for the highest and third highest points. The actual effect of homology class can be traced by comparing the lowest points on the four lines, the second lowest points, and so on. This shows a relatively gradual increase.

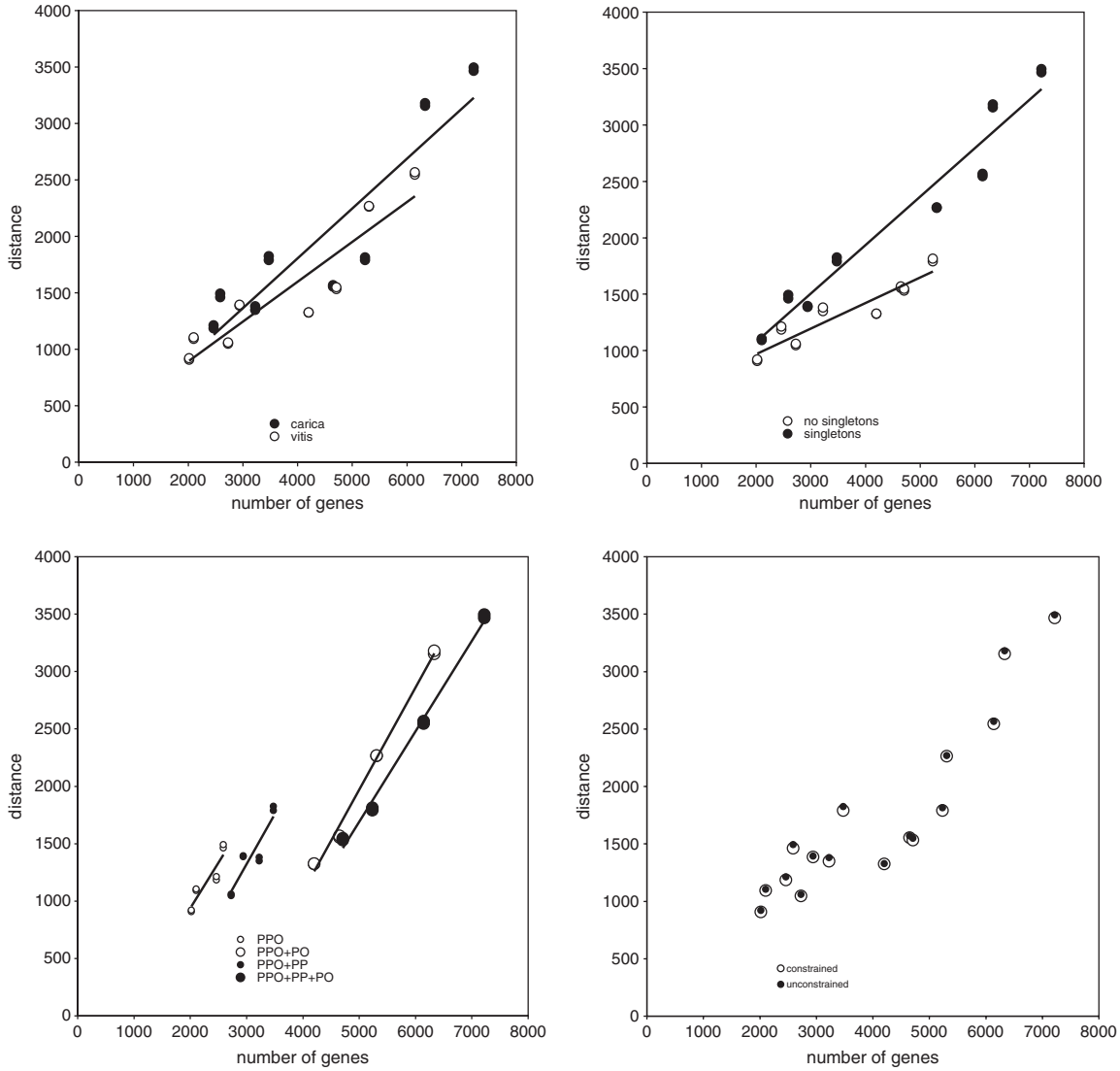

FIG. 3. Effect of controlling for the number of genes. (**Upper left**) *Carica* evolving faster than *Vitis.* (**Upper right**) Rapid increase in distance due to singletons. (**Lower left**) Effect of homology classes. (**Lower right**) Almost imperceptible effect of algorithm version.

6. COMPARISON OF THE HEURISTICS

In Table 1, the constrained guided halving algorithm always does better than the unconstrained guided halving heuristic, as measured by the total distance in the last column. At the same time, the unconstrained heuristic had a clear effect in reducing the bias towards *Populus*, in each case decreasing the distance to the outgroup, compared to the constrained heuristic. This decrease was accompanied by a small decrease in r for the outgroup analysis.

In fact, the decrease in the bias was far greater than the increase in total cost, meaning that if bias reduction is important, then this heuristic is worthwhile, despite its inability to find a minimizing ancestor and its lengthy execution time.

To further investigate the behavior of the new algorithm, we simulated evolution by M inversions and translocations (in a 10:1 proportion) from a genome A to produce an outgroup genome R and 1000–M rearrangements from a WGD genome $A \oplus A$ to produce a descendant genome T. We then applied the constrained and the new algorithms, showing that the new one was superior when $M \leq 600$, but not for $M \geq 700$ (Fig. 4, right).

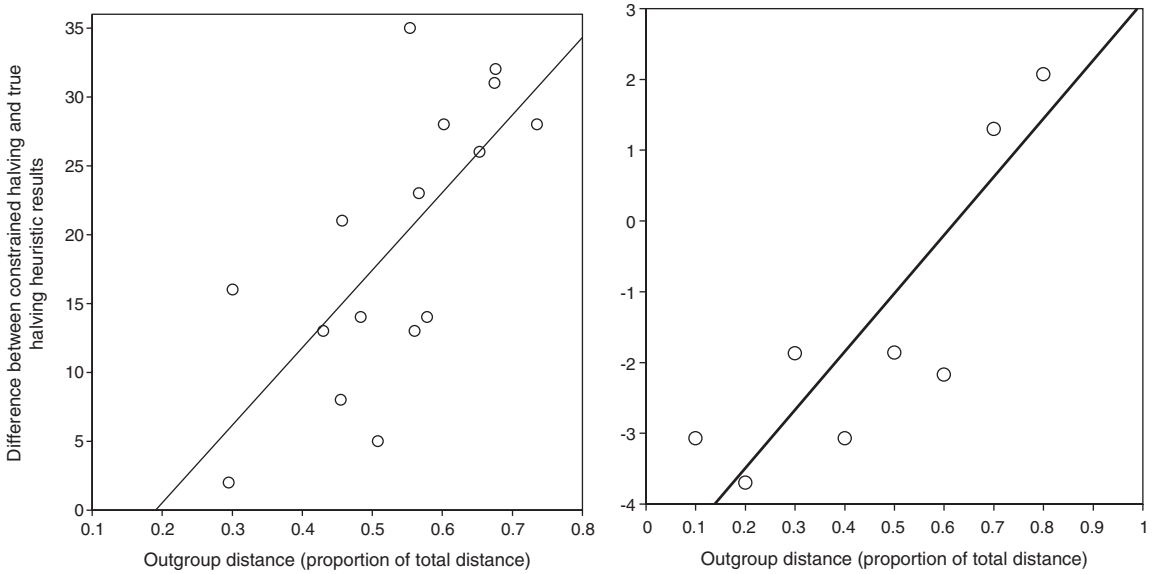

FIG. 4. Performance of the constrained and unconstrained heuristics as a function of the real (**left**) or simulated (**right**) distance of the outgroup from *A*. Note that, despite the similarity of the two curves, the simulated results indicate that the new (unconstrained) algorithm is better when the outgroup proportion of total distance is no larger than 0.6, whereas with the real data this is only predicted to happen when that proportion is below 0.2.

Considering the 16 comparisons in the real data between the constrained and the new algorithm, the change in the total distance also shows a distinct correlation ($\rho^2 = 0.5$) with the distance from the outgroup and *A*. We point this out even though the constrained algorithm, as we have seen, seems superior when the distance between *R* and *A* is more than 20% of the total distance. This is plotted in Figure 4 (left).

The difference between the simulations, where the new method is generally superior, and the real data, where the new method would seem to be superior only when the outgroup is very close to the ancestor, must be ascribed in large part to some way the model used for the simulations does not correspond to how the real data was generated. The ''failure'' of the new algorithm to do better with the real data cannot be ascribed to its inability to find good local optima, since it succeeds with simulated data. One clue is the relatively high reuse rate in the comparison between the outgroup and *A*, compared with that between *Populus* and $A \oplus A$.

7. REARRANGEMENTS OF PARTIALLY ASSEMBLED GENOMES

Our analyses involving *Carica* have incorporated an important correction. The genomic distance between *Carica* and *A* counts many chromosome fusion events that reduce the number of ''chromosomes'' in *Carica* from 223 to the 9. These are not a measure of the true rearrangement distance, but only of the current state of the *Carica* data. Since these may be considered to take place as a first step in the rearrangement scenario (Yancopoulos et al., 2005), we may simply subtract their number from *d* to estimate the true distance. At the same time, many of the breakpoints between *A* and *Carica* are removed by these same fusions, so these should be removed from the count of *b* as well. The calculations in Table 2 illustrate how the *d*(*A, Carica*) results in the bottom quarter of Table 1 were obtained.

Figure 5 (left) shows experimental results on how the increasing fragmentation of a genome into contigs, using a random fragmentation of *Vitis* grenome, decreases the estimated distance between *Vitis* and *A*. This is understandable, since the freedom of the contigs to fuse in any order without this counting as a rearrangment step, inevitably will reduce the distance by chance alone. But the linearity of the result suggests that this decrease is quite predictable, and that the estimates of the distance between *Carica* and *A* are actually underestimates by about 10%.

Figure 5 (right) shows that creating contigs by randomly breaking the *Vitis* genome does not create excessive variability among the solutions, only the same as the dispersion of alternate solutions for the original *Vitis* data, a few percentage points of the distance itself.

TABLE 2. CORRECTION FOR CONTIG DATA

		d(A, Carica)			*Correction*						
Data sets	*Genes in* A	d	b	r	c	d*	a	ct	ca	b*	r*
PPC	2464	986	1090	1.81	223	772	76	7	1371	934	1.65
PPC, PP	3226	1027	1132	1.81	224	812	74	6	2091	981	1.66
PPC, PC	4651	1084	1177	1.84	314	779	123	9	3470	926	1.68
PPC, PP, PC	5234	1214	1318	1.84	325	898	112	12	3910	1088	1.65

A, pre-doubling ancestor of *Populus;* PPC, full gene sets; PP, defective, missing papaya ortholog; PC, defective, missing one poplar paralog; *d*, genomic distance; *b*, number of breakpoints; $r = 2d/b$, the reuse statistic; *c*, number of contigs; *d**, distance corrected for excess of contigs over true number of chromosomes $= d - c + 9$; *a*, number of "obvious fusions"; *ca*, number of common adjacencies; *ct*, number of common telomeres; *b**, corrected number of breakpoints = number of genes $- ca - ct - 2a$; *r**, corrected reuse statistic $= 2d^*/b^*$. Data without singletons. Solutions obtained by constrained algorithm.

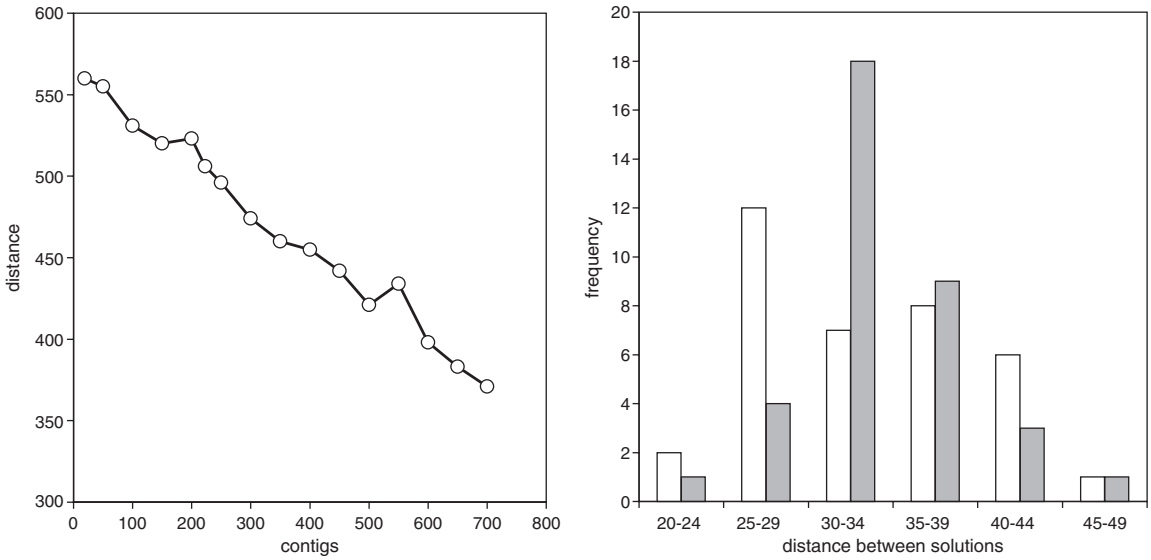

FIG. 5. (**Left**) Effect of increasing fragmentation of *Vitis* into "contigs" on the distance between the reconstructed *A* and *Vitis*. (**Right**) Distributions of distances among solutions for *A* based on *Vitis* data (white bars) and among solutions for *Vitis* fragmented into contigs in different random ways (gray bars).

8. A COMPARISON OF THE OUTGROUPS

Perhaps the most surprising result in Table 1 is that the *Vitis* gene order is decidedly closer to *Populus* and its ancestor *A* than *Carica* is. Both the Tree of Life and the NCBI Taxonomy Browser currently exclude the Viticae family from the rosids, though some older taxonomies do not make this distinction.

Before interpreting this result, we mention two sources of error in the comparison of *Vitis* and *Carica*. The first is that the *Carica* distances are based on a larger gene set; without singletons and defective homology sets PPC is 22% larger than PPV. As a rule of thumb, we can expect distances to be approximately proportional to the number of genes. However, as we have seen in Figure 3, *Carica* evolves faster even if we control for gene number.

The other source of error is due to the contig data, and this results in an *underestimate* of the *Carica*-ancestor distance. From Figure 5, we can estimate that the *Carica* distances are underestimated by about 10% because of the 223 contigs in the *Carica* data. Thus, the discrepancy between the two outgroups is actually larger than it appears to be.

We may conclude that this difference is genuine and substantial. Then, assuming that *Populus* and *Carica* have a closer phylogenetic relationship, or even a sister relationship, our results can only be explained by a faster rate of gene order evolution in *Carica* than in *Vitis*.

8.1. Dispersion

As described in Section 3.3, we generated 100 different solutions with the constrained halving and unguided halving algorithms using each outgroup, and 54 for unconstrained halving with *Vitis* as the outgroup and 15 with *Carica*. The genomic distances were normalized by number of genes in common in the two genomes being compared before input to the analysis. This number was 2464 genes for *Carica* comparisons, 2020 for *Vitis* comparisons, and 1514 for *Carica-Vitis* comparison.

For each of the six outgroup/method combinations, we calculated the average normalized distance between all of its solutions to each of the other combinations, leading to the 6×6 bottom right submatrix in Table 3. We input this into a two-dimensional principal coordinates analysis in the R package, producing the pattern of six black dots in Figure 6. It can be seen that the first dimension of the figure represents the right-to-left movement from unguided halving towards increasing influence of the outgroup. The second dimension distinguishes the *Carica* and *Vitis* analyses.

We then calculated the average distance between all the alternate solutions *within* each outgroup/method combination and divided this by the original input distances to the vertical and horizontal neighbors of the corresponding point on Figure 6; these two factors were multiplied by the corresponding distances on the principal coordinates graph in order to obtain the axes of ellipses. The ellipses, shaded in the figure, represent the degree of dispersion of the solutions around each of the six points. In the case of the constrained solution based on *Vitis*, a quadrilateral shape was employed because of the asymmetry of the horizontal comparisons involving the unconstrained guided halving and the unguided halving solutions.

Figure 7 situates the ancestral reconstructions in a principal coordinates analysis including the *Carica* and *Vitis* genomes. The *Populus* genome is added after the analysis on the basis of halving distances; it was not included in the principal coordinates analysis because of orthology assignment inconsistencies arising in the calculations of the distances between several unduplicated genomes and the descendant of a WGD.

This figure shows that the ancestral reconstructions all occupy a relatively small area of solution space. It also represents the movement already studied in Figure 6 from unguided to constrained to unconstrained analysis in the direction of the outgroup.

8.2. Using both outgroups

There are 1734 complete homologous gene sets, including two *Populus* copies and one copy in each of *Carica* and *Vitis*. (Some of these, constituting 1-gene contigs in *Carica*, were not used for the analyses in Table 1; here we have 332 *Carica* contigs, instead of the 223 in the previous analysis.) In the same way as the unconstrained algorithm in Section 4 is based on a modification of the guided halving algorithm for one outgroup in reference (Zheng et al., 2008a), we could define an unconstrained version of the two-outgroup guided halving algorithm implemented in that earlier work. For convenience, however, we use the constrained version of two-outgroup guided halving from reference (Zheng et al., 2008a) to find the ancestor (small circle) genome in Figure 8a as a first step, then compute the ''median'' genome based on this ancestor, *Carica* and *Vitis*. The median problem here is to find the genome, the sum of whose distances from ancestor *A*, *Carica* and *Vitis* is minimal. This problem is NP-hard (Tannier et al., 2009), and solving it is barely feasible with the 1734 genes in our data, requiring some 300 hours of MacBook computing time.

Table 3. Matrix of Average Distances between Analyses, Normalized ×1000

	Carica	Vitis	*ConC*	*ConV*	*UncC*	*UncV*	*UgC*	*UgPV*
Carica	0.0	458.4	400.2	432.6	394.5	432.6	448.9	463.7
Vitis		0.0	289.3	278.7	305.2	270.3	313.7	329.7
ConC			14.2	52.2	40.6	57.5	68.6	81.9
ConV				16.8	73.3	35.1	76.6	68.3
UnvC					16.6	69.4	94.6	107.0
UncV						20.3	89.2	87.1
UgC							45.0	63.4
UgV								47.0

Matrix of average distances between analyses, normalized ×1000. Diagonal contains average within-group distances (not input into principal components analysis). *Carica* and *Vitis* data used in Figure 7 but not in Figure 6. Con, constrained; Unc, unconstrained; Ug, unguided; C, *Carica*; V, *Vitis*.

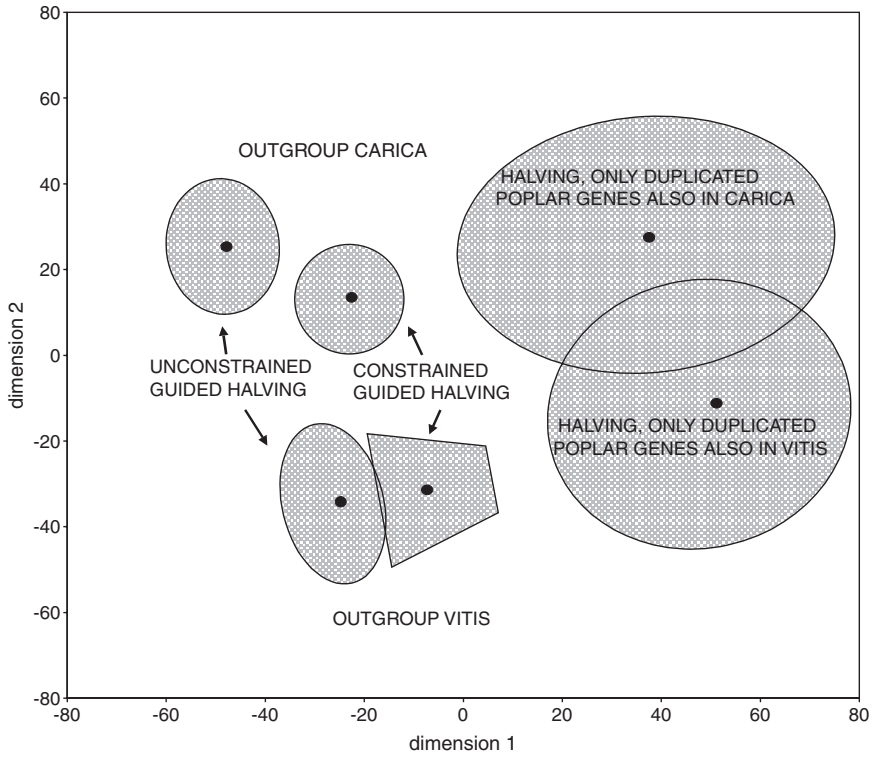

FIG. 6. Principal coordinates analysis of average distances between reconstructed ancestral genomes. Black points represent ''average'' genomes. Shaded areas around each point represent the dispersion of alternate solutions of the same halving problem.

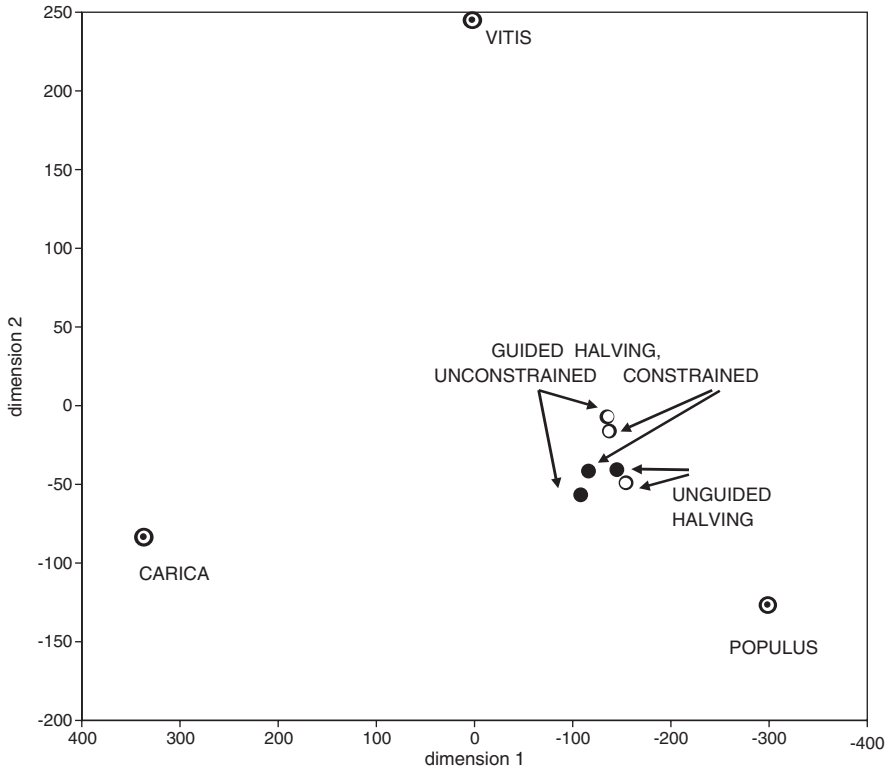

FIG. 7. Principal coordinates analysis of distances between reconstructed and present-day genomes. Filled and open dots indicate *Carica* and *Vitis* used as outgroup, respectively.

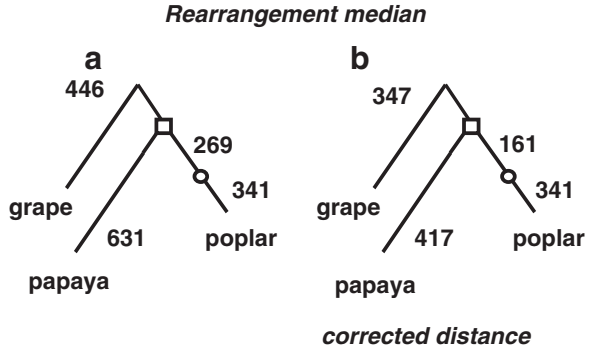

FIG. 8. Branch lengths in angiosperm phylogeny, using two estimates of the median, and applying the contig correction. (**a**) Before correcting for contig fusions. (**b**) After correction.

This initial result unfortunately inherits the same defect as the *Carica* data, i.e., it is composed of contigs rather than true chromosomes. In this case, the median genome contains 118 ''contig-chromosomes.'' We correct the distances to the median by subtracting the difference in the number of chromosomes/contigs between the three genomes and the median. This corresponds to disregarding the fusions counted in the original distances that are essentially carrying out an optimal assembly, modeling an analytical process, not a biological one. This produces the corrected values in Figure 8b.

Let us compare the distance from *Vitis* and from *Carica* to ancestor *A*, passing through the median, in Figure 8 (508 and 578, respectively), with the minimum distances[1] in Table 1, and proportionately adjusted for the reduced number of genes ($560 \times \frac{1734}{2020} = 481$ and $772 \times \frac{1734}{2464} = 543$, respectively. Passing through the median modestly augments (by 27 and by 35, respectively) both trajectories. But using the median diminishes the total cost of the phylogeny, i.e., in comparison with a phylogeny where there is no common evolutionary divergence of the outgroups from *Populus* from $481 + 543 = 1024$ to $341 + 417 + 161 = 919$.

Figure 8b confirms that the papaya genome has evolved more rapidly than the grapevine one.

8.3. Molecular evolutionary correlates of rearrangement rates

With obvious sources of error in our papaya/poplar and three-way comparisons (such as the incomplete assembly of the papaya genome and potentially error-prone ortholog/paralog determination) being insufficient explanations for papaya's enhanced rearrangement rate relative to *Vitis* or the diploid poplar ancestor, *A*, we have sought a biological interpretation.

Papaya, grapevine, and poplar all share the ancient γ WGD. ''Paleologous'' (paralogous) gene pairs identified as γ descendants, as mined from the three genomes, show different rates (K_s) of synonymous substitutional change (Tang et al., 2008). Median K_s for *Vitis* γ pairs (1.22) is substantially lower than that for poplar (1.54) or papaya (1.76).

Synonymous substitutional rates can be interpreted as placeholders for divergence times (Cui et al., 2006), but they have also been correlated with different life strategies in plants, e.g., the woody perennial versus the annual habit, and as such, generation time (Gaut and Morgan, 1996). Recent evidence from large-scale phylogenetic studies incorporating many taxa and many genes has backed the latter inference (Smith and Donoghue, 2008).

The generation times of papaya, poplar, and grapevine show a pattern entirely (negatively) consistent with median K_s values for γ paralogs. Papaya can reproduce in 9–15 months (Ming et al., 2008), poplar in 4–6 years (Tuskan et al., 2006), and grape has reproduced sexually in approximately 80-year intervals since domestication (Arroyo-García et al., 2006). In turn, these generation times and K_s values correlate well with the genomic rearrangement rates calculated here (Table 1; Fig. 8). As such, we hypothesize a common cause argument, short generation time, to explain the aberrant-seeming rearrangement history for papaya relative to its phylogenetic relationships, which would otherwise have suggested this taxon to be closer than grapevine to the diploid poplar ancestor, *A*. Future median genome/guided halving analyses incorporating

[1]Constrained analyses, no singleton or defective homology sets.

weedy species such as *Arabidopsis* and *Mimulus* might help bolster or refute this hypothesis depending on their rearrangement rates relative to other plant genomes.

9. CONCLUSION

The main contributions of this article are as follows:

- the discovery of the rapid rate of gene order evolution in *Carica* compared to *Vitis*,
- a systematic way of controlling for the dependence of rearrangement distance on the number of genes,
- a way of visualizing the reduction of dispersion of the solutions to a problem when comparing methods to solve it,
- a way to use incompletely assembled contigs in genome rearrangement studies,
- a new unbiased algorithm for guided genome halving, and
- the systematic use of reuse rates to show that the inclusion of singletons are not helpful in ancestral genome reconstruction.

In this article, we have not considered the *Arabidopsis* genome. One reason is simply the paucity of full homology sets containing exactly four *Arabidopsis* copies, with or without copies from one or more outgroups. This means we would have to depend on defective homology sets to a greater degree, with the concomitant problem of deciding which if any of the genes in the sets are paralogs dating from the most recent WGD. A more fundamental problem is that the logic of guided genome halving is to use the gene order of the outgroup to influence the placement of the two paralogs in the ancestral tetraploid. But, as in the case of *Arabidopsis*, if the pair of paralogs is itself the result of an earlier WGD, the influence of the outgroup is self-contradictory, trying to position each pair of genes in the same two positions. We plan to elaborate a more comprehensive analysis avoiding both these difficulties, but this is beyond the scope of this article.

ACKNOWLEDGMENTS

Research was supported in part by a Discovery grant to D.S. and a doctoral fellowship to C.Z. from the Natural Sciences and Engineering Research Council of Canada (NSERC). D.S. holds the Canada Research Chair in Mathematical Genomics.

DISCLOSURE STATEMENT

No conflicting financial interests exist.

REFERENCES

Arroyo-García, R., Ruiz-García, L., Bolling, L., et al. 2006. Multiple origins of cultivated grapevine (*Vitis vinifera* L. ssp. sativa) based on chloroplast DNA polymorphisms. *Mol. Ecol.* 15, 3707–3714.

Bafna, V., and Pevzner, P. 1996. Genome rearrangements and sorting by reversals. *SIAM J. Comput.* 25, 272–289.

Bergeron, A., Mixtacki, J., and Stoye, J. 2006. A unifying view of genome rearrangements. *Lec. Notes Comput. Sci.* 4175, 163–173.

Choi, V., Zheng, C., Zhu, Q., et al. 2007. Algorithms for the extraction of synteny blocks from comparative maps. *Lect. Notes Bioinform.* 4645, 277–288.

Cui, L., Wall, P.K., Leebens-Mack, J.H., et al. 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Res.* 16, 738–749.

El-Mabrouk, N., and Sankoff, D. 2003. The reconstruction of doubled genomes. *SIAM J. Comput.* 32, 754–792.

Gaut, B.S., Morton, B.R., McCaig, B.C., et al. 1996. Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. *Proc. Natl. Acad. Sci. USA* 93, 10274–10279.

Jaillon, O., Aury, J.M., Noel, B., et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449, 463–467. Available at: www.genoscope.cns.fr/externe/English/Pro-jets/Projet_ML/data/annotation/pt. Accessed July 29, 2009.

Pevzner, P.A., and Tesler, G. 2003. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc. Natl. Acad. Sci. USA* 100, 7672–7677.

Li, L., Stoeckert, C.J. Jr, and Roos, D.S. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189.

Mazowita, M., Haque, L.,and Sankoff, D. 2006. Stability of rearrangement measures in the comparison of genome sequences. *J. Comput. Biol.* 13, 554–566.

Ming, R., Hou, S., Feng, Y., et al. 2008. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* 452, 991–996. Available at: http://asgpb.mhpcc.hawaii.edu. Accessed July 29, 2009.

Sankoff, D. 2006. The signal in the genomes. *PLoS Comput. Biol.* 2, e35.

Sankoff, D., Zheng, C., and Zhu, Q. 2007. Polyploids, genome halving and phylogeny. *Bioinformatics* 23, i433–i439.

Sinha, A.U., and Meller, J. 2008. Sensitivity analysis for reversal distance and breakpoint reuse in genome rearrangements. *Pac. Symp. Biocomput.* 13, 37–48.

Smith, S.A., and Donoghue, M.J. 2008. Rates of molecular evolution are linked to life history in flowering plants. *Science* 322, 86–89.

Soltis, D.E., Albert, V.A., Leebens-Mack, J., et al. 2009. Polyploidy and angiosperm diversification. *Am. J. Bot.* 96, 336–348.

Tang, H., Wang, X., Bowers, J.E., et al. 2008. Unraveling ancient hexaploidy through multiply aligned angiosperm gene maps. *Genome Res.* 18, 1944–1954.

Tannier, E., Zheng, C., and Sankoff, D. 2009. Multichromosomal median and halving problems under different genomic distances. *BMC Bioinform.* 10, 120.

Tesler, G. 2002. Efficient algorithms for multichromosomal genome rearrangements. *J. Comput. Syst. Sci.* 65, 587–609.

Tuskan, G.A., Difazio, S., Jansson, S., et al. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313, 1596–1604. Available at: http://genome.jgi-psf.org/Poptr1/Poptr1.download.html. Accessed July 29, 2009.

Velasco, R., Zharkikh, A., Troggio, M., et al. 2007. A high-quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS ONE* 2, e1326.

Yancopoulos, S., Attie, O., and Friedberg, R. 2005. Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics* 21, 3340–3346.

Zheng, C., Zhu, Q., Adam, Z., et al. 2008a. Guided genome halving: hardness, heuristics and the history of the Hemiascomycetes. *Bioinformatics* 24, i96–i104.

Zheng, C., Zhu, Q., and Sankoff, D. 2006. Genome halving with an outgroup. *Evol. Bioinform.* 2, 319–326.

Zheng, C., Zhu, Q., and Sankoff, D. 2007. Removing noise and ambiguities from comparative maps in rearrangement analysis. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 4, 515–522.

Zheng, C., Zhu, Q., and Sankoff, D. 2008b. Descendants of whole genome duplication within gene order phylogeny. *J. Comput. Biol.* 15, 947–964.

Zheng, C., Wall, P.K., Leebens-Mack, J., et al. 2009. Gene loss under neighborhood selection following whole genome duplication and the reconstruction of the ancestral *Populus* genome. *J. Bioinform. Comput. Biol*. 7, 499–520.

Address correspondence to:
Dr. David Sankoff
Department of Mathematics
University of Ottawa
585 King Edward Avenue
Ottawa K1N 6N5, Canada

E-mail: sankoff@uottawa.ca