

Models for Similarity Distributions of Syntenic Homologs and Applications to Phylogenomics

David Sankoff¹, Chunfang Zheng, Yue Zhang, João Meidanis², Eric Lyons, and Haibao Tang³

Abstract—We outline an integrated approach to speciation and whole genome doubling (WGD) to resolve the occurrence of these events in phylogenetic analysis. We propose a more principled way of estimating the parameters of gene divergence and fractionation than the standard mixture of normals analysis. We formulate an algorithm for resolving data on local peaks in the distributions of duplicate gene similarities for a number of related genomes. We illustrate with a comprehensive analysis of WGD-origin duplicate gene data from the family Brassicaceae.

Index Terms—Gene tree, species tree, whole genome doubling, algorithms, mixture of distributions, Brassicaceae

1 INTRODUCTION

IN this paper, we model and analyze the set of $\binom{N}{2} + N$ distributions of similarity of homologous gene pairs within and across N species where whole genome doubling (or tripling, quadrupling, etc.—all subsumed under the abbreviation WGD) has affected one or more of these species. These distributions typically involve many thousands of genes, initially in colinear duplicate pairs along homeologous chromosomes. The remnant of this pattern after evolutionary chromosomal rearrangement and fractionation—the loss of large proportions of duplicate genes—is called syntenic paralogy. Although we do not formally discuss colinear gene order in this paper, syntenic paralogy is a unique source of unambiguous data on synchronously generated duplicate genes, in analogy to the syntenic orthology in the two genomes resulting from a speciation event. Although it is common practice to use methods such as EMMIX [1] to decompose distributions of homologous (paralogous or orthologous) gene pairs into sums of normal distributions, each component distribution indicative of a putative WGD or speciation event, there has been no previous attempt to account statistically for the parameters of these distributions aside from their location on a time axis. In particular, no work has been done on the quantitative effect of fractionation rates on the variance of, and area under, each normal curve, which are fundamental to our proposals. And other than the estimated mean of each component normal, the features of these distributions have

not been incorporated in phylogenomic methodology; the use of modal values, or “peaks”, which do not depend on any fixed subset of genes, is another contribution of the present work.

In Section 2.1, we first sketch out a model of gene similarity distribution under random sequence divergence, speciation and fractionation, providing the basis for a principled treatment of the statistical inference of divergence and fractionation rates and for speciation and WGD times.

In the rest of Section 2, we work out the detailed combinatorial and probability analyses of the case of two WGD in a genome (Section 2.2), the case of three WGD (Section 2.3), and the case of one whole genome tripling (WGT) followed by a WGD (Section 2.4). Although we have not yet developed a general software package implementing our methodology, we can nonetheless calculate particular cases using R routines. In Section 2.5, we illustrate with calculations for the *Populus trichocarpa* (poplar) genome under the WGT + WGD model.

Where the first parts of Section 2 deal only with paralogous gene pairs ($N = 1$), in Sections 2.6 and 2.7 we broaden the focus to include the orthologous gene pairs generated by speciation ($N \geq 2$). As an example of $N = 2$ in Section 2.6, we analyze the case of a WGD followed by a WGT followed by a speciation. The latter analysis allows us to dissect the similarity distribution between orthologous gene pairs in *Brassica rapa* and *Brassica oleracea*.

For $N \geq 3$, we enter the domain of phylogenomics. A combinatorial explosion of the number of possible gene trees for $N \geq 3$, already evident for the $N = 2$ case in Section 2.6, precludes the kind of exhaustive case-by-case analysis presented in Section 2 until some automated procedure can be developed. We can nonetheless proceed by simply identifying local modes—peaks—in all the similarity distributions, and translating these into phylogenetically related paralogous and orthologous event times. In Section 3, we propose an algorithm for inferring a rooted tree from these event times, where the WGD are located on specific lineages in a consistent way. This means that the peaks

- D. Sankoff, C. Zheng, Y. Zhang, and J. Meidanis are with the Department of Mathematics and Statistics, University of Ottawa, Ottawa, ON K1N 6N5, Canada. E-mail: {sankoff, yzhan481}@uottawa.ca, chunfang313@gmail.com, meidanis@ic.unicamp.br.
- E. Lyons and H. Tang are with the School of Plant Sciences, University of Arizona, Tucson, AZ 85721. E-mail: ericlyons@email.arizona.edu, tanghaibao@gmail.com.

Manuscript received 29 Oct. 2016; revised 2 May 2018; accepted 5 June 2018.
Date of publication 31 July 2018; date of current version 31 May 2019.
(Corresponding author: David Sankoff.)

For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2018.2849377

identified in a comparison of any two genomes correspond to all the WGD in the lineage of their common ancestor, plus a more recent peak corresponding to the speciation event from which the two genomes diverge, with no more recent WGD-generated peaks.

In Section 4, we pursue the study of the order Brassicales with the application of the new algorithm. Our data contains twelve genomes spanning several genera of the family Brassicaceae as well as one genome from the sister-family Cleomaceae. Our WGD-aware adaptation of the neighbour-joining algorithm corrects a serious error in a phylogeny produced without taking WGD into account.

2 DISTRIBUTIONS OF GENE SIMILARITY

There are two independent biological processes involved in the creation of syntenic homolog similarity distributions. The first is simply an event, either speciation or WGD that creates two copies of each chromosome in the genome and hence two copies of each gene, both copies in a single genome (WGD) or one copy in each of two new genomes (speciation). In the case of WGD, this process also involves fractionation, the gradual loss of duplicate genes over evolutionary time. The second process that determines the shape of homolog similarity distributions is the largely random mutational pattern that gradually degrades the similarity between duplicated genes over evolutionary time, with varying results from gene pair to gene pair. The initiative to integrate the processes of WGD, fractionation and mutational sequence divergence was taken in [2], and some of the formal material in this section is drawn from that source, although in corrected and simplified form.

In Section 2.1, we discuss the evolutionary process of sequence divergence, prior to investigating the consequences of some particular WGD histories in the rest of this section.

2.1 The Building Blocks

It is a common practice in genomics, when analyzing distributions of homolog similarity, to resort to numerical procedures embodied in software such as EMMIX [1] for resolving mixtures of normal distributions. These methods, however, powerful and flexible as they may be, are not tailored to the problem of detecting speciation and WGD in a *set* of related similarity distributions. For any mixture of normals, EMMIX will identify these components as long as there is enough data. But not every mixture of normals credibly reflects some sequence of genomic events. More important, among the $\binom{N}{2} + N$ gene similarity distributions within and across N species, there are many constraints that are not handled by software packages, such as requiring the time estimate to be the same for an event in all the distributions that are affected by it, or requiring the variance of the WGD in a lineage to be increasing over time.

Here we model gene pair divergence in terms of a probability p reflecting *similarity*—the proportion of nucleotide positions that are occupied by the same base in the two genes, orthologs or paralogs, although the same principles hold for *synonymous distance* K_s —the proportion of synonymous changes (not affecting translation to an amino

acid) over all eligible positions, or *fourfold degenerate synonymous distance* $4dTv$ —the transversion rate at fourfold degenerate third codon positions [3].

We represent by G the gene length, in terms of the number of nucleotides in the genes' coding region, setting aside for the moment that this varies greatly from gene to gene. We assume p follows the normal approximation to the sum of G binomial distributions, divided by G , and is related to the time $t \in [0, \infty)$ elapsed since the event that gave rise to the pair

$$\begin{aligned} \text{mean : } E[p] &= \frac{1}{4} + \frac{3}{4}e^{-\lambda t} \in [0, 1] \\ \text{variance : } E(p - E[p])^2 &= \frac{3}{16} \frac{(1 + 3e^{-\lambda t})(1 - e^{-\lambda t})}{G}, \end{aligned} \quad (1)$$

where $\lambda > 0$ is a divergence rate parameter.

In practice, p for duplicate gene pairs is generally much greater than 0.25, so we base our analysis on those pairs with similarity greater than, say, 0.5.

Fractionation, the loss of one gene—and only one—from a paralog pair, is represented by a parameter $u \in [0, 1]$, representing the probability, for a pair of duplicate genes, that neither gene is lost over a time interval of length t . The assumption that any gene pair has a constant probability (over time) of being fractionated entails

$$u = e^{-\rho t}, \quad (2)$$

where ρ is the fractionation parameter.

The two genes of a pair cannot both be lost, a fact that is motivated by both biological and modeling considerations. First, the loss of both WGD-generated copies of a functional gene would normally be lethal, so that a genome where this occurred would not be viable and such events would not be observed or inferred in genomic data. Second, the point of our model is to predict the distribution of gene pair similarities as a function of the time the pair was created from a single ancestral form, so that the parameters of the model can be inferred from counts and measurements of these pairs at the current time of observation. In effect, our model accounts for the lineages of observed pairs and single-copy (unpaired, singleton) genes, and does not generate incomplete (and thus unobservable) lineages caused by the loss of both duplicates of an ancestral gene. Thus there is no loss of generality in the condition that at least one gene in a paralog pair must survive until the next event, or until time of observation.

Thus, in the case of a single WGD, the mean of the distribution of duplicate gene pair similarities is an estimate of p (and also leads to an estimate of t), and the number of pairs compared to the number of unpaired genes provides an independent estimate of u (and of ρ).

2.2 Two WGD

Consider a genome that has undergone two successive WGD. We denote by “ t_1 -pairs” and “ t_2 -pairs” those duplicated gene pairs created at t_1 and t_2 respectively, with expected similarities p_1 and p_2 . For fixed ρ , u and v are functions of t_1 and t_2 only, representing the probabilities $e^{-\rho(t_1-t_2)}$ and $e^{-\rho(t_2-0)} = e^{-\rho t_2}$, respectively, for a pair of genes present at the start of the time interval, that neither

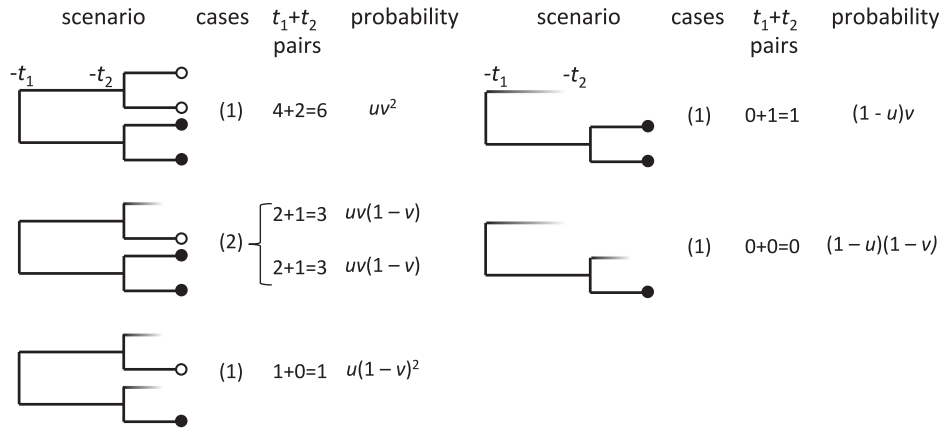


Fig. 1. Components of the number of surviving pairs created by WGDs at t_1 and t_2 . Fading lines illustrate gene loss, but play no role in the calculation of the number of cases.

gene is lost by the end of the interval. Note that in this and later models, we assume, for simplicity, that a fractionation regime from one WGD is supplanted by that set into operation by the next WGD. That is, fractionation involving older pairs is no longer operative.

In Fig. 1, let

$$\begin{aligned} A &= \mathbf{E}(t_1 \text{ pairs}) \\ &= 4uv^2 + 4uv(1-v) + u(1-v)^2 \\ &= u(1+v)^2 \end{aligned} \quad (3)$$

$$\begin{aligned} B &= \mathbf{E}(t_2 \text{ pairs}) \\ &= 2uv^2 + 2uv(1-v) + (1-u)v \\ &= v(1+u) \end{aligned} \quad (4)$$

$$\begin{aligned} C &= \mathbf{E}(\text{unpaired genes}) \\ &= (1-u)(1-v) \end{aligned} \quad (5)$$

$$\begin{aligned} P(A) &= \text{proportion of } t_1 \text{ pairs} \\ &= \frac{A}{A+B+C} \end{aligned} \quad (6)$$

$$\begin{aligned} P(B) &= \text{proportion of } t_2 \text{ pairs} \\ &= \frac{B}{A+B+C} \end{aligned} \quad (7)$$

$$\begin{aligned} P(C) &= \text{proportion unpaired} \\ &= \frac{C}{A+B+C}. \end{aligned} \quad (8)$$

For a fixed gene length G and λ , let $\mathbf{N}_p(s)$ be the density at point s of a normal distribution with mean p and variance $\frac{p(1-p)}{G}$. The probability that gene pair will be observed to be with similarity $s \in [0, 1]$ is

$$Q(s) = P(A)\mathbf{N}_{p_1}(s) + P(B)\mathbf{N}_{p_2}(s), \quad (9)$$

and the probability of an unpaired gene is

$$Q^* = P(C). \quad (10)$$

The likelihood of a data set with gene pairs at s_1, \dots, s_l and k unpaired genes is

$$\mathcal{L} = \prod_{i=1}^l Q(s_i) Q^{*k}. \quad (11)$$

The log likelihood $L = \log \mathcal{L}$ is

$$\begin{aligned} L &= \sum_{i=1}^l \log Q(s_i) + k \log Q^* \\ &= \sum_{i=1}^l [\log (P(A)\mathbf{N}_{p_1}(s_i) + P(B)\mathbf{N}_{p_2}(s_i))] + k \log Q^*. \end{aligned} \quad (12)$$

There is no closed form for the maximum likelihood of a mixture of normals, so in practice we use numerical means such as Newton-Raphson or an EM algorithm to derive the MLE.

2.3 Three WGD

Consider now three successive WGD affecting a genome (for example the τ, σ and ρ WGD that occurred in the common ancestor of the cereals [4]). The scenarios producing various numbers of gene pairs of various ages are depicted in Fig. 2, where u, v and w are the retention probabilities for pairs produced at t_1, t_2 and t_3 .

$$\begin{aligned} \mathbf{E}(t_1 \text{ pairs}) &= u(1+v)^2(1+w)^2 \\ \mathbf{E}(t_2 \text{ pairs}) &= (1+u)^2(1+v)w \\ \mathbf{E}(t_3 \text{ pairs}) &= w(1+u)(1+v) \\ \mathbf{E}(\text{unpaired}) &= (1-u)(1-v)(1-w). \end{aligned} \quad (13)$$

From this analysis, we can predict the number of pairs remaining from the each of the three events, and perform MLE calculations to determine the parameters.

2.4 WGT Followed by WGD

Whole genome tripling is rarer than doubling, but several important examples have had profound impacts on plant evolution. A whole genome tripling occurred in the eudicot lineage just before the emergence of the core eudicots, which today number in the hundreds of thousands of species, and a large number of these have undergone further WGD or WGT, as modelled in Fig. 3. A WGT characterizes the genus *Solanum*, which includes, tomato, pepper and eggplant; yet another WGT preceded the diversification of the mustard-cabbage-radish family Brassicaceae. Here

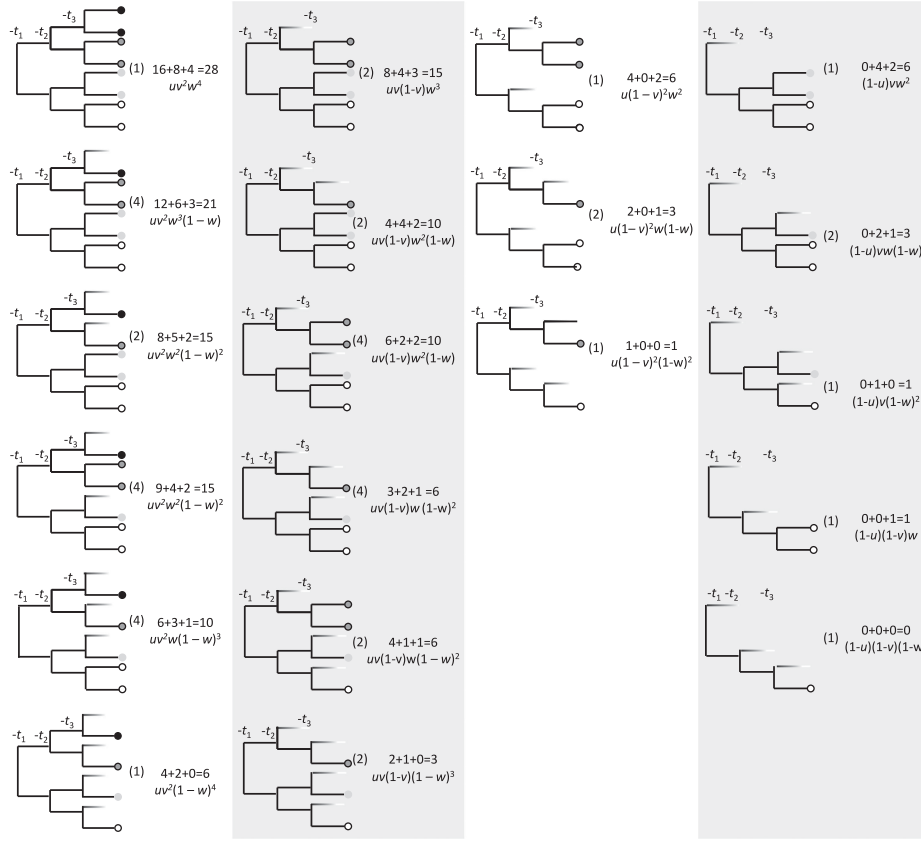


Fig. 2. Components of the number of surviving pairs created by WGDs at t_1 , t_2 , and t_3 . See Fig. 1 to interpret the content of the various columns.

$$\begin{aligned}
 E(t_1 \text{ pairs}) &= (u' + 3u''')(1+v)^2, \\
 E(t_2 \text{ pairs}) &= (1 + u' + 2u''')v, \\
 E(\text{unpaired}) &= (1 - u''' - u')(1-v).
 \end{aligned} \tag{14}$$

2.5 The Case of *Populus Trichocarpa*

Although the work we have presented consists of combinatorial models, and the inference procedures are not implemented in a user-friendly package, we did experiment with some data sets using functions on the R platform according to the model in the previous section and in subsequent sections. For example, we used coding sequence data on the poplar genome [5] stored on the CoGe platform [6], [7], and applied the SYNMAP program on that platform, comparing the genome to itself in order to locate all syntenic paralog pairs. From the output, we extracted the similarity of each pair, producing the distribution of similarities in Fig. 4. (Similarity, or identity, is just the percentage of identical base pairs in the coding regions.) Focusing on syntenic, or colinear, sets of duplicate genes tends to ensure that these pairs are all produced during the same WGD or speciation event. This excludes tandem pairs and isolated pairs out of syntenic context, namely pairs not created by a WGD.

In Fig. 4, the broadening of the calculated peak representing a recent WGD (at $p = 0.9$), as represented by the solid line, and the compression of the earlier peak, can be attributed to the strong link between the means and variances of the normal components of the distribution as expressed in Equation (1). There being no such constraint on EMMIX, the fit (of the dotted line) is of course

much better, but uninterpretable in terms of fractionation rates. According to our model, the small volume of the earlier peak is accurately reflected in very small

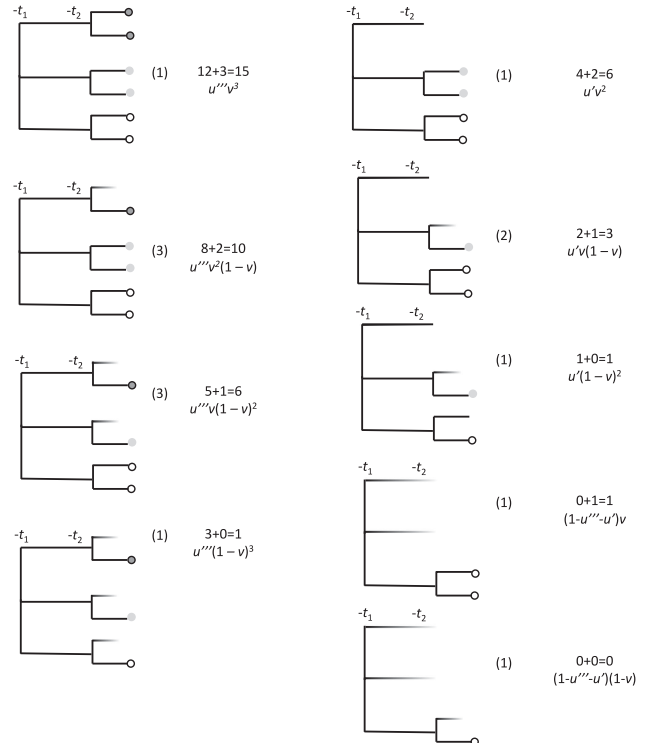


Fig. 3. Components of the number of surviving pairs created by a whole genome tripling at t_1 and a WGD at t_2 .

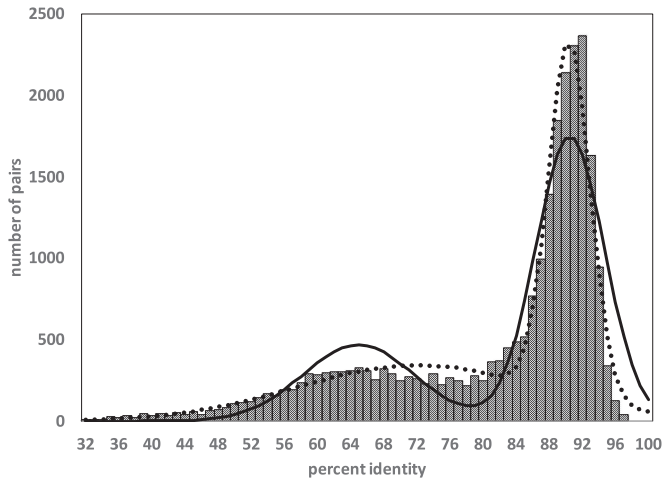


Fig. 4. The distribution of duplicate gene similarities in poplar. Some pairs with around 99 percent similarity were excluded, since they represent slightly divergent alleles of the same gene, not two distinct genes. Histogram: empirical data. Solid curve: fit of model with estimated fractionation rates. Dotted line curve: EMMIX fit.

estimates of u''' , less than 0.1, and $u' = 0.1$ compared to about 0.76 for v .

In Fig. 4, if the early event is identified with the γ tripling some 100 Mya, then the more recent WGD must be dated older than 65 Mya, consistent with this event preceding the divergence of poplar and willow as argued in [5]. It is also consistent with a constant fractionation rate ρ over the whole time period covered in the analysis.

2.6 The Effect of Speciation

Up to now we have considered only WGD events, including tripling. In comparing two species, there are peaks at times corresponding to all their shared WGD, followed by a single peak dating from their speciation event, but no further peaks. Additional WGD after speciation increase the number of orthologs and paralogs in a uniform way across the board, but do not change the number of peaks.

2.7 Of Neep and Kraut

The broadening of effects of event age and of fractionation on the peaks in similarity distributions as time elapses are well illustrated by the two peaks in Fig. 4. Eventually, all events become indistinguishable from noise caused by random gene resemblances, widespread domain sharing, tandem and near-tandem duplications, gene-order rearrangements, gene conversion and other processes.

We keep this in mind as we compare *Brassica rapa* and *Brassica oleracea*. These species are thought to descend from a common WGT “ γ ” at the base of the eudicots (the same as the wide peak in Fig. 4), two WGD “ β ” and “ α ” (shared with *Arabidopsis* and other members of the order Brassicales) and another WGT specific to the family Brassicaceae [8]. It is known, however, that before the two most recent events there was a rapid speed-up of gene divergence processes, so that the earlier events are difficult to discern in comparative data. Data from CoGe on *B. rapa* and *B. oleracea* shows a clear speciation peak at 97 percent similarity and a clear WGT peak at 89 percent similarity, as in Fig. 5. There is a large shoulder on the curve, likely due to the large

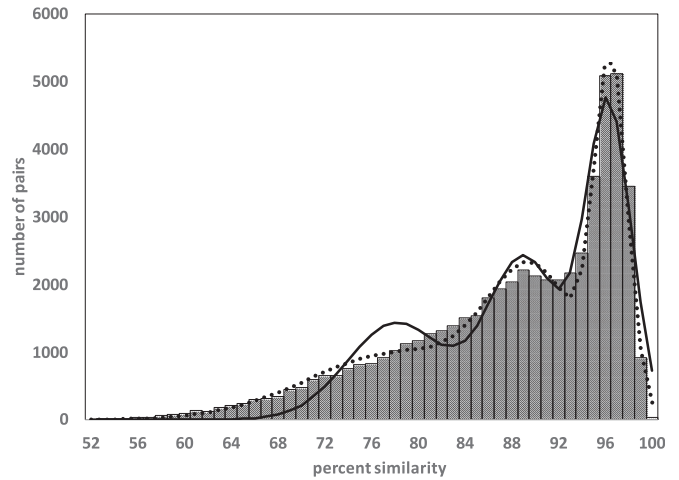


Fig. 5. Ortholog similarities in the *B. rapa* – *B. oleracea* comparison. Histogram: empirical data. Solid curve: fit of model with estimated fractionation rates. Dotted line curve: EMMIX fit.

overlap of the similarity distributions due to α, β and γ . Other evidence such as the distribution of K_s scores (not shown) and the comparisons within and between the *Arabidopsis* genomes (detailed below), suggest a hidden peak at 78 percent. This motivated us to set up a three-event model, with a WGD at 78 percent, a WGT at 89 percent and a speciation at 96 percent.

The model is depicted in Fig. 6. The probability that an α paralog pair (generated at time t_1 before the present) survives until time t_2 is u . The probability that all three pairs generated at the *Brassica* tripling time t_2 before the present survive until time t_3 is v and the probability that only one pair survives in v' . The probability that an ortholog pair generated at speciation time t_3 survives to the present is z .

For each kind of pair (i.e., according to its time of origin) adding up the number of these in each of the 39 configurations shown, multiplied by the number of versions (in parentheses), multiplied by the probability of the configuration as written, we obtain

$$\begin{aligned} E(t_1 \text{ pairs}) &= u(1 + 2v + v')(1 + z)^2, \\ E(t_2 \text{ pairs}) &= (1 + u)(2v + v')(1 + z)^2, \\ E(t_3 \text{ pairs}) &= 1 - u - v - v' - z - uv'z \\ &\quad + 3v'z + 4vz + uv' + uv + 3uz \end{aligned} \quad (15)$$

and

$$E(\text{unpaired}) = (1 - u)(1 - v' - v)(1 - z).$$

Among the t_1 pairs in this analysis half will be paralogous and will not show up in the predicted distribution of ortholog similarity. The same is true of the t_2 pairs, but the t_3 pairs are all orthologous. Based on these facts, we can calculate the density of ortholog pairs similarity for maximum likelihood or other estimates of u, v, v' and z . Fig. 5 compares this density, based on $u = 0.24, v = 0.15, v' = 0.13$ and $z = 0.5$, with the distribution of similarities in the data. It can be seen that the fit is reasonable, but not perfect, especially at the earliest times. The pairs at this time probably reflect the γ tripling. The combinatorial task of adding yet another

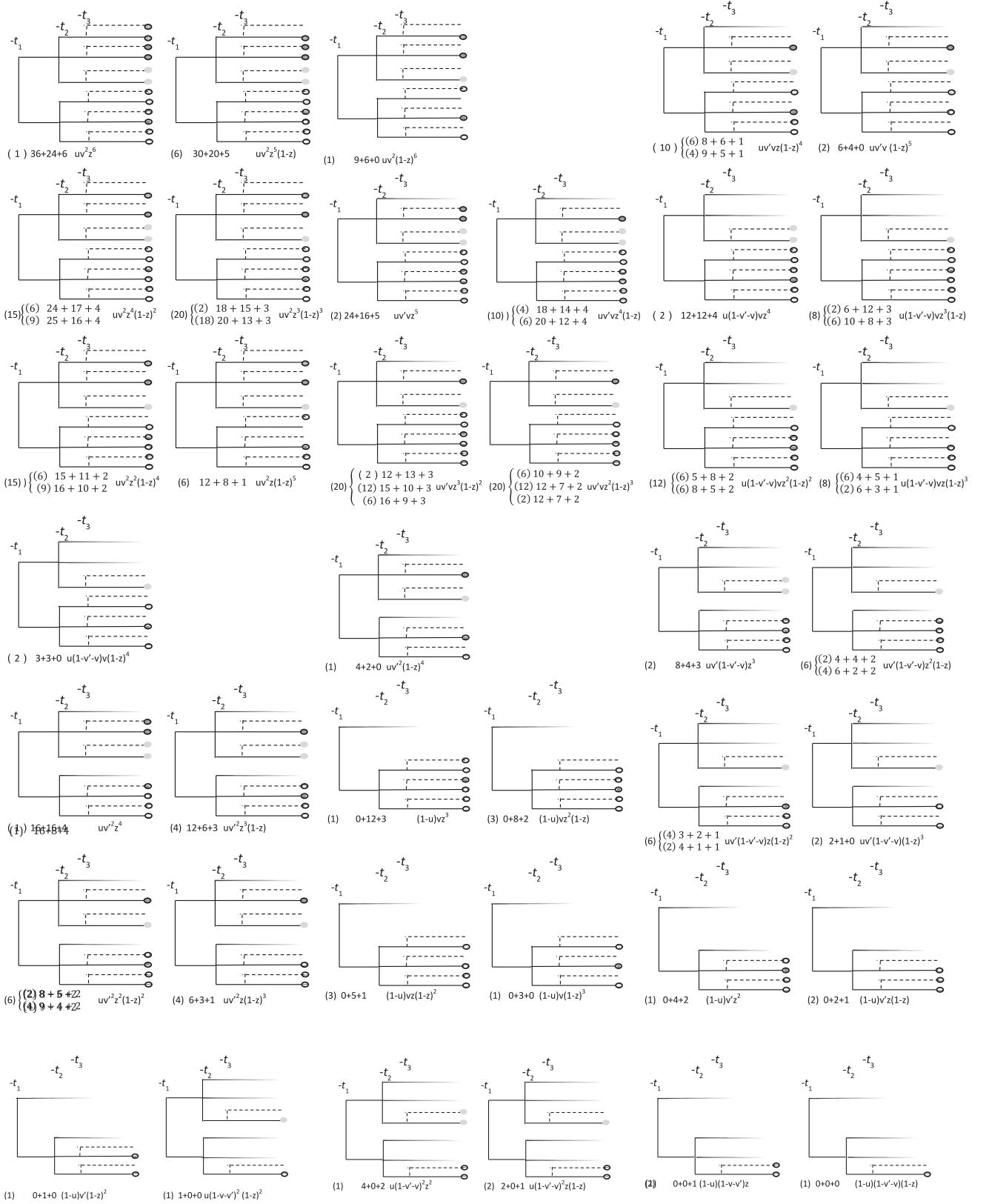


Fig. 6. Components of the set of surviving orthologous pairs in two species diverging at time t_3 after a (shared) WGD at t_1 , and a shared WGT at t_2 . Number of cases of same component with different labelling in parentheses. “ $w + x + y$ ” indicates w pairs dating from the shared WGD, x pairs created by the shared WGT, and y pairs from the speciation event. Parallel dotted and solid lines distinguish two species.

WGT event to the configurations in Fig. 6 is beyond our current scope of our manual methods.

Fitting three normals to these same data with EMMIX returns peaks at 0.79 and 0.90 and 0.96. In other words, our model

recovers essentially the same WGD and speciation times as EMMIX, but the lack of any constraint in EMMIX involving the mean and variance, such as Equation (1) in our model allows EMMIX to fit the data more closely, as was the case in Fig. 4.

3 A PHYLOGENOMIC ALGORITHM BASED ON SPECIATION AND WGD

It is commonplace that the divergence of many of the $\binom{N}{2}$ pairs of species in a N -species phylogeny may date from a single speciation event, and this should be represented by a common node in the inferred phylogenetic tree. In the same way, $N' \leq N$ species may be descendants of a given WGD, and this relationship should be represented by the position of this event on a branch of a directed tree, such that all N' of these species, and only these, are in the subtree descending from this event.

Because the computational problems with WGD-event detection illustrated in Section 2.7 preclude its application to large numbers of comparisons at this time, we will assume that we can infer p , and hence the age of an event, simply by identifying the mode, or “peak” of the similarity distribution, without recourse to other estimation procedures. This makes it difficult to pick out events visible as “shoulders” of other events on a similarity distribution derived from a pair of species, like we did for the *B. rapa*—*B. oleracea* comparison, though it allows for these events to emerge from the comparison of one or more other pairs of species.

3.1 The Algorithm

There are two principles underlying our method for reconstructing a phylogeny from a set of inter- and intra-genome syntenic homolog similarity distributions:

- each intra-genome distribution of similarities can only have peaks due to the WGD in its direct lineage, and
- each inter-genome distribution may contain several peaks due to WGDs, but only one peak due to speciation, the most recent one, i.e., at the date of the most recent common ancestor of the two species.

To incorporate these principles into a phylogenetic inference procedure, we adapt the neighbour-joining (NJ) algorithm [9], to produce a directed tree rather than the undirected tree for which it was first designed. To attempt to satisfy the two principles above, we need to specify not only on which tree branch a WGD is located, but whether it is the subtree emanating from one endpoint or the other of the branch that inherits the WGD, which implies specifying a direction to all branches leading from a WGD to a leaf.

We retain NJ’s transformation of a distance matrix that takes account of differing evolutionary rates on different branches of the phylogeny at each step of the agglomerative procedure. We add two elements to the procedure, namely an “age” calculation for each possible new node that could be produced, and a simple dynamic programming step to align and evaluate how similar are the ancestral WGD of two nodes, leaf or ancestral, being aligned. This step potentially adds a penalty to the distance between two nodes if they do not share the same WGD, especially a recent WGD. With these added elements, we no longer have the NJ algorithm, strictly speaking, since the new elements cannot be incorporated into the distance matrix at the outset, but must be calculated during execution.

The dynamic programming step compares two series of WGD times (positive real values) in ascending order. As in standard sequence comparison, this includes insertion/

deletion costs (for entire WGD events), and in the place of substitution penalty, a cost depending on the difference between WGD times. The minimum cost between the two sequences of WGD times is used as the penalty to add to the distance measure between the two nodes being linked to an immediate common ancestor. To compute a consensus sequence of WGD times, we use the traceback of the dynamic programming to align all or some of the WGD, ignore the remaining WGD, and assign new WGD times midway between the times of each aligned pair. A straightforward implementation of this step has complexity $O(k^2)$, where k is the maximum length of any WGD series; in practice will seldom more than 4, so for all intents and purposes we may consider that the dynamic programming step requires constant time. Algorithm 2, which examines all pairs of unconnected nodes will then be $O(n^2)$ and the main Algorithm 1, whose main loop is traversed n times, is $O(n^3)$.

Algorithm 1. WGD-Modified Neighbour Joining (NJ)

Input: list of n leaf genomes, with WGD times symmetric
 $n \times n$ matrix of speciation times
Output: tree with WGD events

- 1 Initially there are n disconnected nodes (leaf genomes), each node with its own WGDs.
- 2 **while** $n \geq 3$ **do**
- 3 **for each pair of unconnected nodes do**
- 4 get a candidate new node (parent) by joining them (two child nodes)
- 5 calculate the distances from this candidate to the other nodes as in standard NJ
- 6 use **Algorithm 2** to get minimum cost candidate
 / * update the tree with this minimum cost candidate */
- 7 replace two old nodes by a new node and set
- 8 the WGDs and branch direction above this node
- 9 get $n - 1$ unconnected nodes and distances among these
 $n - 1$ nodes for the next iteration
- 10 $n \leftarrow n - 1$
- 11 set the root node anywhere that respects WGD-determined directions
- 12 **return** the resulting tree

The input to Algorithm 1 includes a standard distance matrix, say of divergence times calculated from the p at a speciation peak in the similarity distribution of each pair of genomes, or K_s values at the corresponding peak, plus a list of WGD times ancestral to each genome, obtained from the similarity or K_s distribution of syntenic paralogs.

In this first formalization of our distance-based algorithm taking into account of WGDs and the entire set of gene duplicates generated by WGD or speciation events, we deliberately leave some elements unspecified, to emphasize the generality of the two innovative concepts. For the matrix of genome distance between pairs of genomes, we illustrate with two examples: the speciation peak (most recent peak) in the similarity distribution of homologous gene pairs, and the same peak in the distribution of $\log K_s$ values for these gene pairs, though many other types of distance matrix are possible. For the other key innovation, the penalty due to discordant WGD history between two genomes, we have yet to optimize the parameters of the dynamic programming step,

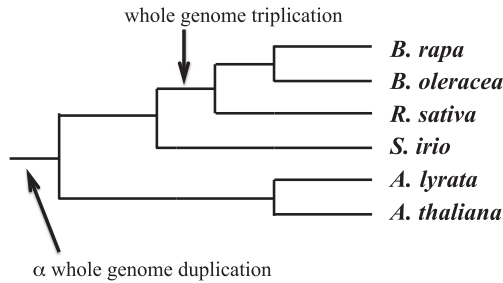


Fig. 7. Phylogenetic relationship of six species in the family Brassicaceae, showing lineages affected by WGD and WGT events.

and how they relate to the measure of distance between genomes.

Algorithm 2. Count Cost

Input: n unconnected nodes,
 $n(n-1)/2$ candidate nodes,
distances from candidate nodes to unconnected nodes

Output: minimum cost candidate, WGD placements

- 1 **for each candidate node do**
- 2 **for each genome (child node) of this candidate do**
- 3 parental age \leftarrow age of this child node + branch length
between this child and its parent, the candidate
- 4 separate the WGDs into two groups:
- 5 group1: WGDs younger than parental age
- 6 group2: WGDs older than parental age
- 7 keep the group1 WGDs on branch between child and
parent
- 8 apply dynamic programming on the group2 WGDs to
determine placement on branch above parent
- 9 age \leftarrow minimum of two parental ages
- 10 cost \leftarrow age
- 11 **if the final penalty of the dynamic programming exceeds a
threshold then**
- 12 this penalty is added to the cost of the candidate
- 13 **return candidate with minimum cost, WGD placement**

4 THE BRASSICACEAE

To illustrate our discussion, we draw on twelve published genomes in the Brassicaceae family, two in the genus *Brassica*: *B. rapa* (turnip, Chinese cabbage) [10] and *B. oleracea* (cabbage, cauliflower) [11], two in *Raphanus*: *Raphanus sativus* (radish) [12] and *R. raphanistrum* (wild radish) [13], two in the genus *Arabidopsis*: *A. lyrata* (rock cress) [14] and *A. thaliana* (thale cress, mouse-ear cress) [15] and one each in the genera *Sisymbrium*: *S. irio* (London rocket) [16], *Schrenkiella*: *S. parvula* (dwarf spikeweed) [17], *Eutrema*: *E. salsugineum* (saltwater cress) [18], *Capsella*: *C. rubella* [19] (red shepherds purse), *Leavenworthia*: *L. alabamica* (Alabama glaucous) [16] and *Aethionema*: *A. arabicum* [16]. In addition, we included one species from the sister family Cleomaceae, genus *Cleome*: *Tarenaya hassleriana* (spider flower) [20].

Before reporting on the analysis of this full data set, we illustrate with the details of a subset of six of the species (i.e., only $\binom{6}{2} = 15$ of the $\binom{13}{2} = 78$ of the pairwise comparisons and 6 self-comparisons), namely *B. rapa*, *B. oleracea*, *R. sativus*, *A. lyrata*, *A. thaliana* and *S. irio*. Fig. 7 shows the phylogenetic relationship among the six species (cf. [16]).

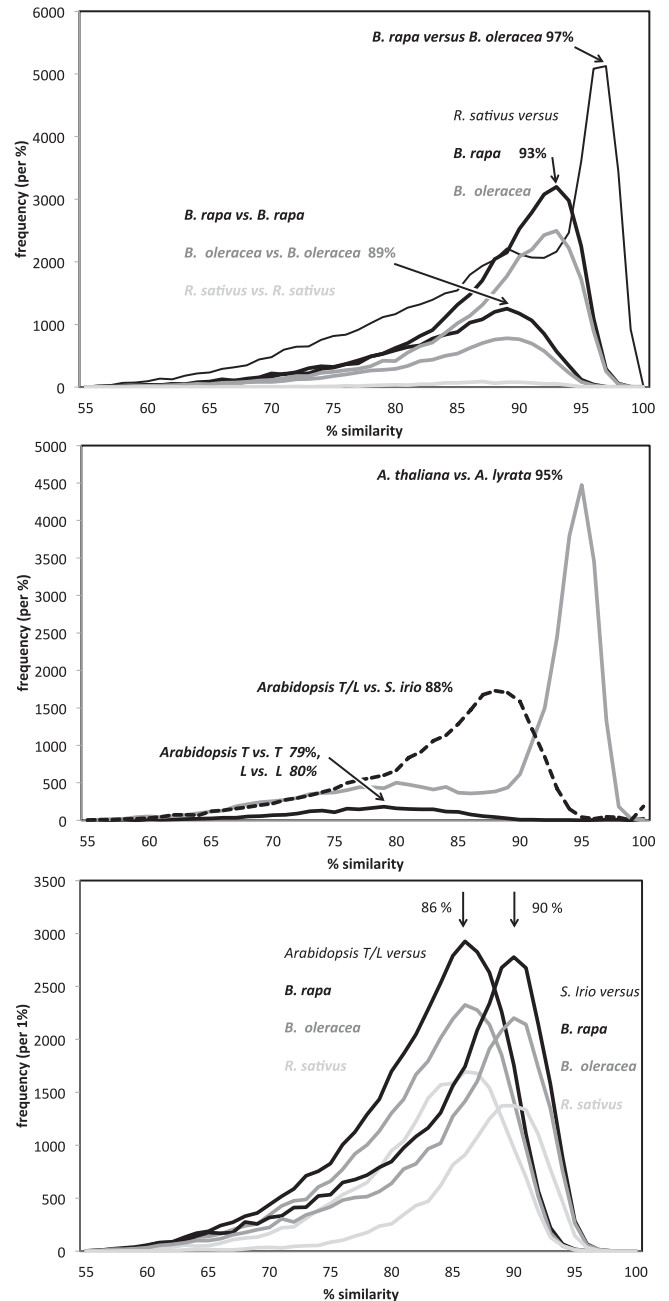


Fig. 8. Gene similarity distribution between 15 pairs of genomes in the Brassicaceae and five self comparisons. Local modes ("peaks") are indicated. Only one of each comparison is shown for *Arabidopsis*, the other is superimposed and indistinguishable.

We extracted genomic data from these species using the database in CoGe [6], [7] as previously discussed in Section 2.5. We then used the SynMap routine (with default parameters) on this platform to compare the gene orders of each of the $\binom{6}{2} = 15$ pairs of genomes. This procedure identifies orthologs produced by speciation by detecting colinear arrays of several duplicate pairs in two species with approximately the same divergence: "syntenic blocks". Similarly, we did a self-comparison of five of the six genomes; the sixth one, the *Sisymbrium* genome, did not have enough closely spaced duplicate pairs for SynMap to produce paralogous syntenic blocks. The distributions of similarities calculated are shown in Fig. 8. The peaks found in each genome are tabulated in Table 1.

TABLE 1
Peak Similarity Level, by Genome

peak number	description	genome					
		BR	BO	RS	SI	AL	AT
1	alpha doubling [21]	np	np	np	np	80	80,79
2	divergence of genus <i>Arabidopsis</i>	86	86	86,87	88	88-86	88-86
3	whole genome tripling	89	89	87	np	np	np
4	divergence of genus <i>Sisymbrium</i>	90	90	90	-	-	-
5	divergence of genus <i>Raphanus</i>	93	93	93	np	-	-
6	speciation of <i>Arabidopsis</i> T & L	-	-	-	-	95	95
7	speciation of <i>B. rapa</i> & <i>B. oleracea</i>	97	97	-	-	-	-

np: no peak, but one could be found by mixtures of distribution methods. - : no peak expected. Note peak 3 occurring before peak 4 due to slow evolutionary rate (λ) of *Sisymbrium*.

From Fig. 8 and Table 1, we note that the earliest doubling, detected at 79-80 percent in the *Arabidopsis* self-comparisons, shows no peaks in the other self-comparisons—there is a shoulder or heavy tail in the appropriate place in the *Brassica* self-comparisons, but this is swamped by the later tripling event. The tripling itself is visible in all three *Brassica* self-comparisons and in the comparison of *B. oleracea* and *B. rapa*, but not in the weaker signals involving *Raphanus*.

More interesting is that the peaks at 90 percent reflecting the *Sisymbrium* speciation, known to occur before the *Brassica* tripling, suggest that speciation is more recent, since the tripling peak is at 89 percent. This apparent conflict is clearly ascribable to a slower rate of evolution (lower λ), since the divergence of *Arabidopsis* from *Sisymbrium* also seems to occur more recently (88 percent) than the divergence of *Arabidopsis* from the *Sisymbrium* sister genus *Brassica* (86 percent). Note that the small differences between peak similarities are not insignificant, given the many thousands of gene pairs involved in these comparisons.

Applying our algorithm to the data derived from these six genomes reflects this rate anomaly, with *S. irio* branching after the tripling instead of preceding it, unless the dynamic programming penalty on discordant WGD evidence is increased, in which case *S. irio* branches with the *Arabidopsis-Capsella* group, an equally poor result.

Fig. 9 depicts a phylogeny of our full set of Brassicaceae genomes (all of the sequenced ones we have been able to access through CoGe) plus one genome from the sister family Cleomaceae. Though most of this tree is uncontroversial, the taxonomic positions of *S. irio*, *E. salsugineum* and *S. parvula* has been changed several times in recent years [22]. The particular configuration shown in the figure is drawn from [16].

We computed all 78 pairwise comparisons and 12 of the self-comparisons and picked out the visible peaks in each. These were used as input to our algorithm. With any of range of reasonable values for the dynamic programming penalty, the output tree was as shown in Fig. 10.

The only difference with Fig. 9 is that the taxonomically volatile group *S. irio*, *E. salsugineum* and *S. parvula* appear as a monophyletic clade, sister to the *Brassica-Raphanus* group, which is not implausible, and is certainly preferable than branching *S. irio* within the latter, after the *Brassica* tripling. More important, without the dynamic programming constraint taking account of the placement of the WGD, *L. alabamica* branches before the *Arabidopsis-Brassica* split, whereas in Fig. 10 it is appropriately grouped as a sister taxon to the *Arabidopsis-Capsella* clade.

5 CONCLUSION

We have introduced a concerted approach to plant phylogenomics that gives a central role to the whole genome

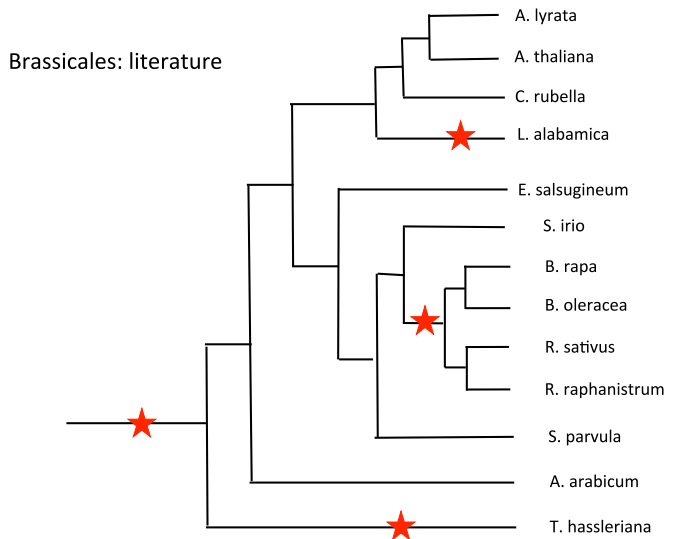


Fig. 9. Brassicales phylogeny derived from the literature.

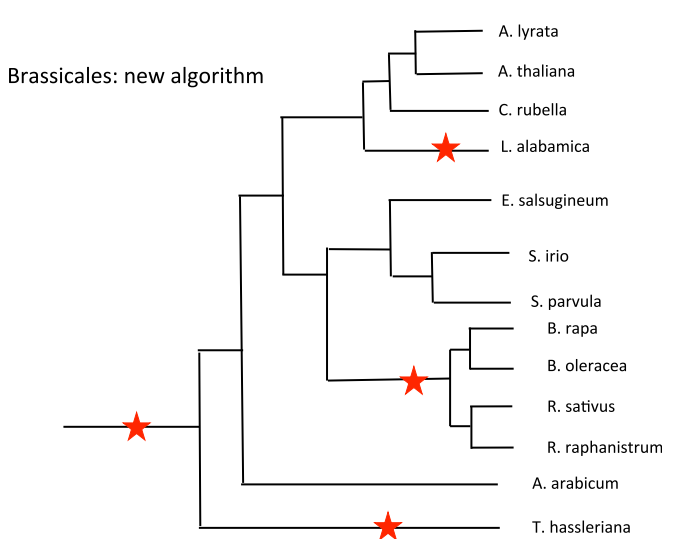


Fig. 10. Phylogeny calculated by proposed algorithm.

doubling (WGD) and tripling evidence from similarity distributions of syntenic (colinear) homologs. The first component of this methodology is a combined combinatorial and probabilistic analysis of orthologous and paralogous gene pairs in pairwise genome comparisons and self-comparisons, biologically more interpretable than statistical mixture-of-distributions analyses. The second half uses the results from this analysis in a phylogenetic algorithm inspired by neighbour-joining, but that produces rooted trees to accommodate the inherent directionality of WGD. This algorithm incorporates a dynamic programming subroutine to align the ancestral WGD events of two observed or inferred ancestral genomes. For both the first part and the second part, we have implemented proof-of-principle software and applied these to genomes from the order Brassicales, in particular to the comparison of *B. rapa* and *B. oleracea* for the detailed analysis of a similarity distribution, and to 12 members of the family Brassicaceae, plus one outgroup, for the phylogenomics.

Our theoretical considerations pertain to the simple model assumed at the beginning of this section. In practice, various other processes affect the distribution of similarities so that the number of gene homologs between and within genomes may be severely reduced from those expected from the model. We have seen that the divergence rate λ may vary somewhat for individual lineages, and ρ is certainly even more variable. The parameter G should be allowed to increase in time to account for the greater than predicted increase in the variance of older components. Genomic processes such as chromosomal rearrangements disrupt gene order and degrade the recovery of syntenic blocks and duplicate gene pairs. These issues should all be addressed in future work. DNA substitution models with more parameters and rate variation among sites could also be incorporated.

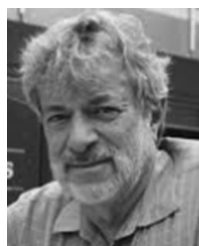
ACKNOWLEDGMENTS

This work was partially funded by a NSERC Discovery grant, NSF grant IOS-1339156 and Fapesp grant 2016/01511-7. J. Meidanis was with the the Institute of Computing, University of Campinas, while this work was done.

REFERENCES

- [1] G. J. McLachlan, D. Peel, K. E. Basford, and P. Adams, "The Emmix software for the fitting of mixtures of normal and t-components," *J. Statistical Softw.*, vol. 4, no. 2, pp. 1–14, 1999.
- [2] Y. Zhang, C. Zheng, and D. Sankoff, "Evolutionary model for the statistical divergence of paralogous and orthologous gene pairs generated by whole genome duplication and speciation," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 5, pp. 1579–1584, Sep./Oct. 2018.
- [3] S. Kumar and S. Subramanian, "Mutation rates in mammalian genomes," *Proc. Nat. Acad. Sci. United States America*, vol. 99, pp. 803–808, 2002.
- [4] M. R. McKain, H. Tang, J. R. McNeal, S. Ayyampalayam, J. I. Davis, C. W. dePamphilis, T. J. Givnish, J. C. Pires, D. W. Stevenson, and J. H. Leebens-Mack, "A phylogenomic assessment of ancient polyploidy and genome evolution across the Poales," *Genome Biol. Evol.*, vol. 8, pp. 1150–1164, 2016.
- [5] G. A. Tuskan, S. Difazio, J. Jansson, J. Bohlmann, I. Grigoriev, U. Hellsten, N. Putnam, S. Ralph, S. Rombauts, A. Salamov et al., "The genome of black cottonwood, *Populus trichocarpa* (torr. & gray)," *Sci.*, vol. 313, pp. 1596–1604, 2006.
- [6] E. Lyons and M. Freeling, "How to usefully compare homologous plant genes and chromosomes as DNA sequences," *Plant J.*, vol. 53, pp. 661–673, 2008.
- [7] E. Lyons, B. Pedersen, J. Kane, and M. Freeling, "The value of non-model genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates rosids," *Tropical Plant Biol.*, vol. 1, pp. 181–190, 2008.
- [8] M. Barker, H. Vogel, and M. Schranz, "Paleopolyploidy in the Brassicales: Analyses of the cleome transcriptome elucidate the history of genome duplications in Arabidopsis and other brassicales," *Genome Biol. Evol.*, vol. 1, pp. 391–399, 2009.
- [9] N. Saitou and M. Nei, "The neighbor-joining method: A new method for reconstructing phylogenetic trees," *Mol. Biol. Evol.*, vol. 4, pp. 406–425, 1987.
- [10] X. Wang, H. Wang, J. Wang, R. Sun, J. Wu, S. Liu, Y. Bai, J.-H. Mun, I. Bancroft, F. Cheng, S. Huang, X. Li, W. Hua, J. Wang, X. Wang, M. Freeling, J. C. Pires, A. H. Paterson, B. Chalhoub, B. Wang, A. Hayward, A. G. Sharpe, B.-S. Park, B. Weisshaar, B. Liu, B. Li, B. Liu, C. Tong, C. Song, C. Duran, C. Peng, C. Geng, C. Koh, C. Lin, D. Edwards, D. Mu, D. Shen, E. Soumpourou, F. Li, F. Fraser, G. Conant, G. Lassalle, G. J. King, G. Bonnema, H. Tang, H. Wang, H. Belcram, H. Zhou, H. Hirakawa, H. Abe, H. Guo, H. Wang, H. Jin, I. A. P. Parkin, J. Batley, J.-S. Kim, J. Just, J. Li, J. Xu, J. Deng, J. A. Kim, J. Li, J. Yu, J. Meng, J. Wang, J. Min, J. Poulain, K. Hatakeyama, K. Wu, L. Wang, L. Fang, M. Trick, M. G. Links, M. Zhao, M. Jin, N. Ramchiary, N. Drou, P. J. Berkman, Q. Cai, Q. Huang, R. Li, S. Tabata, S. Cheng, S. Zhang, S. Zhang, S. Huang, S. Sato, S. Sun, S.-J. Kwon, S.-R. Choi, T.-H. Lee, W. Fan, X. Zhao, X. Tan, X. Xu, Y. Wang, Y. Qiu, Y. Yin, Y. Li, Y. Du, Y. Liao, Y. Lim, and Y. Narusaka, "The genome of the mesopolyploid crop species *Brassica rapa*," *Nature Genetics*, vol. 43, pp. 1035–1039, 2011.
- [11] S. Liu, Y. Liu, X. Yang, C. Tong, D. Edwards, I. Parkin, M. Zhao, J. Ma, J. Yu, S. Huang, X. Wang, J. Wang, K. Lu, Z. Fang, I. Bancroft, T.-J. Yang, Q. Hu, X. Wang, Z. Yue, H. Li, L. Yang, J. Wu, Q. Zhou, W. Wang, G. J. King, J. C. Pires, C. Lu, Z. Wu, P. Sampath, Z. Wang, H. Guo, S. Pan, L. Yang, J. Min, D. Zhang, D. Jin, W. Li, H. Belcram, J. Tu, M. Guan, C. Qi, D. Du, J. Li, L. Jiang, J. Batley, A. G. Sharpe, B.-S. Park, P. Ruperao, F. Cheng, N. E. Waminal, Y. Huang, C. Dong, L. Wang, J. Li, Z. Hu, M. Zhuang, Y. Huang, J. Huang, J. Shi, D. Mei, J. Liu, T.-H. Lee, J. Wang, H. Jin, Z. Li, X. Li, J. Zhang, L. Xiao, Y. Zhou, Z. Liu, X. Liu, R. Qin, X. Tang, W. Liu, Y. Wang, Y. Zhang, J. Lee, H. H. Kim, F. Denoeud, X. Xu, X. Liang, W. Hua, X. Wang, J. Wang, B. Chalhoub, and A. H. Paterson, "The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes," *Nature Commun.*, vol. 5, 2014, Art. no. 3930.
- [12] H. Kitashiba, F. Li, H. Hirakawa, T. Kawanabe, Z. Zou, Y. Hasegawa, K. Tonosaki, S. Shirasawa, A. Fukushima, S. Yokoi, Y. Takahata, T. Kakizaki, M. Ishida, S. Okamoto, K. Sakamoto, K. Shirasawa, S. Tabata, and T. Nishio, "Draft sequences of the radish (*Raphanus sativus* L.) genome," *DNA Res.*, vol. 21, no. 5, pp. 481–490, 2014.
- [13] G. D. Moghe, D. E. Hufnagel, H. Tang, Y. Xiao, I. Dworkin, C. D. Town, J. K. Conner, and S.-H. Shiua, "Consequences of whole-genome triplication as revealed by comparative genomic analyses of the wild radish *Raphanus raphanistrum* and three other Brassicaceae species," *The Plant Cell*, vol. 26, pp. 1925–1937, 2014.
- [14] T. T. Hu, P. Pattyn, E. G. Bakker, J. Cao, J.-F. Cheng, R. M. Clark, N. Fahlgren, J. A. Fawcett, J. Grimwood, H. Gundlach, G. Haberer, J. D. Hollister, S. Ossowski, R. P. Otillar, A. A. Salamov, K. Schneeberger, M. Spannagl, X. Wang, L. Yang, M. E. Nasrallah, J. Bergelson, J. C. Carrington, B. S. Gaut, J. Schmutz, K. F. X. Mayer, Y. Van de Peer, I. V. Grigoriev, M. Nordborg, D. Weigel, and Y.-L. Guo, "The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change," *Nature Genetics*, vol. 43, pp. 476–481, 2011.
- [15] T. A. G. Initiative, "Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*," *Nature*, vol. 408, pp. 796–815, 2000.
- [16] A. Haudry, A. E. Platts, E. Vello, D. R. Hoen, M. Leclercq, R. J. Williamson, E. Forczek, Z. Joly-Lopez, J. G. Steffen, K. M. Hazzouri, K. Dewar, J. R. Stinchcombe, D. J. Schoen, X. Wang, J. Schmutz, C. D. Town, P. P. Edger, J. C. Pires, K. S. Schumaker, D. E. Jarvis, T. Mandakova, M. A. Lysak, E. van den Bergh, M. E. Schranz, P. M. Harrison, A. M. Moses, T. E. Bureau, S. I. Wright, and M. Blanchette, "An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions," *Nature Genetics*, vol. 45, pp. 891–898, 2013.

- [17] M. Dassanayake, D.-H. Oh, J. S. Haas, A. Hernandez, H. Hong, S. Ali, D.-J. Yun, R. A. Bressan, J.-K. Zhu, H. J. Bohnert, and J. M. Cheeseman, "The genome of the extremophile crucifer *Thellungiella parvula*," *Nature Genetics*, vol. 43, pp. 913–918, 2011.
- [18] R. Yang, D. E. Jarvis, H. Chen, M. A. Beilstein, J. Grimwood, J. Jenkins, S. Shu, S. Prochnik, M. Xin, C. Ma et al., "The reference genome of the halophytic plant *Eutrema salicagineum*," *Frontiers Plant Sci.*, vol. 4, 2013, Art. no. 46.
- [19] T. Slotte, K. M. Hazzouri, J. A. Agren, D. Koenig, F. Maumus, Y.-L. Guo, K. Steige, A. E. Platts, J. S. Escobar, L. K. Newman, W. Wang et al., "The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution," *Nature Genetics*, vol. 45, pp. 831–835, 2013.
- [20] S. Cheng, E. van den Bergh, P. Zeng, X. Zhong, J. Xu, X. Liu, J. Hofberger, S. de Bruijn, A. S. Bhide, C. Kuelahoglu, C. Bian, et al., "The *Tarenaya hassleriana* genome provides insight into reproductive trait and genome evolution of crucifers," *The Plant Cell*, vol. 25, pp. 2813–2830, 2013.
- [21] S. Kagale, S. J. Robinson, J. Nixon, R. Xiao, T. Huebert, J. Condie, D. Kessler, W. E. Clarke, P. P. Edger, M. G. Links et al., "Polyploid evolution of the Brassicaceae during the Cenozoic era," *The Plant Cell*, vol. 26, pp. 2777–2791, 2014.
- [22] M. Koch and D. German, "Taxonomy and systematics are key to biological information: *Arabidopsis*, *Eutrema (Thellungiella)*, *Noccaea* and *Schrenkiella* (Brassicaceae) as examples," *Frontiers Plant Sci.*, vol. 4, 2013, Art. no. 267.



David Sankoff received the PhD degree in mathematics from McGill University, and has been a member of the Centre de Recherches Mathématiques in Montreal for many years. He currently holds the Canada research chair in Mathematical Genomics in the Mathematics and Statistics Department, University of Ottawa, and is cross appointed to the Biology and the Computer Science Departments. His research interests include comparative genomics, particularly probability models, statistics, and algorithms for genome rearrangements, with a focus on the genomes of flowering plants.



Chunfang Zheng received the master's and PhD degrees in biology from the University of Ottawa, where she has been a research associate in Dr. Sankoff's lab. She has published extensively on algorithms for genome rearrangements and participated in the evolutionary analysis for many flowering plant genome sequencing projects.



Yue Zhang received the master's degree in statistics from Carleton University in Ottawa and is currently working toward the PhD degree in Dr. Sankoff's Lab, University of Ottawa, focusing on the statistical analysis of subgenome evolution after paleopolyploidy.



João Meidanis received the PhD degree from the University of Wisconsin-Madison, with a thesis on DNA fragment assembly and algorithms for gene functional prediction. He is a faculty member of the University of Campinas where he founded the Computational Biology Group. He has coordinated the bioinformatic analysis for several genome projects and is the director of the consulting enterprise Scylla Bioinformatics.



Eric Lyons received the PhD degree in plant biology from the University of California, Berkeley. Previously, he spent five years working in industry at biotech, pharmaceutical, and software companies. He was responsible for developing the CoGe platform for comparative genomics data management and analysis. He has published extensively on these topics and has been invited to teach numerous workshops on the analysis of genomic data to plant, vertebrate, invertebrate, microbe, and health researchers. He is currently an assistant professor with the School of Plant Sciences, CyVerse.



Haibao Tang received the PhD degree in plant biology from the University of Georgia. He is currently a professor at Fujian Agricultural and Forestry University and senior scientist at the University of Arizona, working on genome assembly, annotation, and analyses of plant genomes. He has devised novel computational methods to clarify the evolution of flowering plants, which he has applied to papaya, sorghum, Medicago, Brassica, tomato, Amborella, and numerous other genomes. He has two pending patents for isolating genes responsible for important agronomic traits in the cereal crops.

▷ For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.