Branching Out to Speciation in a Model of Fractionation: The Malvaceae

Yue Zhang, Chunfang Zheng, Sindeed Islam, Yong-Min Kim, and David Sankoff

Abstract—Fractionation is the genome-wide process of losing one gene per duplicate pair following whole genome doubling (WGD). An important type of evidence for duplicate gene loss is the frequency distribution of similarities between paralogous gene pairs in a genome or orthologous gene pairs in two species. We extend a previous branching process model for fractionation, originally accounting for paralog similarities, to encompass the distribution of ortholog similarities, after multiple rounds of whole genome doubling and fractionation, with the speciation event occurring at any point. We estimate the fractionation rates during all the inter-event periods in each lineage of the plant family Malvaceae. We suggest a major correction of the phylogenetic position of the durian sub-family, and discover a new triplication event in this lineage.

Index Terms—Whole genome duplication, fractionation, branching process, gene pair similarity distribution, Malvaceae, durian

1 Introduction

The evolutionary history of the flowering plants (angiosperms) is punctuated with numerous whole genome doubling and tripling (WGD) events, a phenomenon that has only occasionally been identified in other phylogenetic domains. In the angiosperms, every species whose genome has been sequenced has at least one such event in its history, and very often two, three, four or more, except in one early diverging lineage [1]. The doubling events in any one lineage are typically spaced tens of millions of years apart, and the tetraploid or hexaploid genome at the origin of the event re-diploidizes in relatively short order, as homeologous sequences diverge, chromosomes fuse and inversion and translocations disrupt the separate identity of the two subgenomes. However, the set of genes in the genome can remain elevated for long periods of time and if there are several WGD, the genome can accumulate 100,000 genes or more instead of the approximately 25,000 necessary for most angiosperm genomes. It is true that pairs of duplicate genes tend to lose one redundant member over time, a process called *fractionation* with some categories of genes more susceptible to loss than others [2], [3], and sometimes from one subgenome rather than the other [4], but the question remains of whether fractionation occurs at a rapid enough pace to counter the effect of recurrent WGD on gene number. This is the main focus of this paper; the rate of fractionation after a WGD or a series of WGD.

- *Y. Zhang, C. Zheng, S. Islam, and D. Sankoff are with the Department of Mathematics and Statistics, University of Ottawa, Ottawa, ON K1S 2C3, Canada. E-mail: {yzhan481, czhen033, sankoff}@uottawa.ca, sindeed.islam@gmail.com.*
- *Y.-M. Kim is with the Korean Bioinformation Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon 34141, Korea. E-mail: ymkim@kribb.re.kr.*

Manuscript received 12 July 2019; revised 25 Sept. 2019; accepted 10 Nov. 2019. Date of publication 19 Dec. 2019; date of current version 7 Oct. 2021. (Corresponding author: David Sankoff.)
Digital Object Identifier no. 10.1109/TCBB.2019.2955649

An important type of evidence in analyzing ancient polyploidization events is the distribution of coding sequence similarities between two paralogous genes in a genome. For flowering plants with one, two, or more WGD events in their history, the distribution of similarities is a mixture of distributions, each component of which is centred at a similarity value indicative of the age of one of the events. We have developed a model for predicting the shape of these distributions based on the event times, the ploidy multiplicities of the events, rates of loss of duplicate genes from the genome (fractionation), and rates of sequence divergence [5]. Underlying these predictions is a *paralog tree* generated by a discrete-time branching process with one biologically-motivated constraint, which is mathematically tractable and whose parameters are well suited to statistical inference.

The mathematics of the branching process involving recurring WGD within a *single* genome, and the associated inferential tasks associated with *paralogous* gene pairs, have been worked out, but attempts to extend this to *orthologs* in *two* genomes post-speciation [5], [6] have involved treatments not fully consistent with the spirit of the single genome model. These previous models interrupted the ongoing fractionation process in the transition from one species to two, and established a new process, with a new loss-rate parameter, to take account of the loss of genes from the two daughter species and the reduction in the number of orthologous gene pairs. In reality, however, speciation as such does not involve any sudden change in the ongoing evolution of the two diverging populations. This includes the uninterrupted continuation of the fractionation process independently in the two new species—neither of them is "aware" of what is happening in the other.

Thus a key aspect of this paper is to provide a treatment of the transition from one genome to two daughter genomes that smoothly continues the fractionation regime in place before speciation, and extends it independently in each of these species until a new WGD in that species or until the

present time of observation. With some additional details, this solution enables a complete model encompassing all the WGD in the ancestral genome, the speciation event, and all the independent WGDs in the daughter genomes. It defines all the homolog pairs at the time of observation in terms of the event at which they originated. The key period for the model, that which includes the speciation event, is often also the key period for empirical studies, since it includes the speciation plus the fractionation process before that event.

In addition our model leads to the first formal expression of the fact, often empirically observed and intuitively reasonable if not widely known, that the distribution of relative frequencies of ortholog similarities in two genomes is independent of WGD that occur after speciation. However the absolute frequencies change after speciation, they remain a constant multiple of the relative frequencies at the moment of speciation; the last component of the mixture of distributions is always due to the speciation event creating the two diverging genomes, and this is the only speciation event leaving any trace on the distribution. The pre-speciation WGD and the single speciation event are the only events that determine the relative frequencies of the ortholog similarities; further WGD change the absolute frequencies but not the relative frequencies of these similarities. In this paper, we find for the first time an expression for the *amplifying factor*, the ratio of the absolute to the relative frequencies as a function of post-speciation WGDs.

Aside from providing an insight into fractionation, the distributions of homolog similarities have phylogenomic implications. To the extent that a distribution may be decomposed into a mixture of normal components, the local modes may be found at the means of these distributions, and serve as accurate reflection of the timing of the events. We have previously [6], [7] used these modal values—or their logarithmic transforms—as a similarity or distance matrix for input into a phylogenetic algorithm, such as neighbour-joining.

We apply the model and inferential apparatus to eight genomes in the family Malvaceae. This is an ideal group of plants, since many of them, in several subfamilies and genera, have recently had their genomes sequenced, and they are very heterogeneous with respect to the number of WGD in their respective lineages since the Malvaceae emerged some 40 Mya [8]. One result is an overview of the variability in the proportional rates of fractionation in different lineages and in different periods. Another result is an unexpected phylogenetic positioning of one subfamily (Durionaceae, or Helicteroideae) within an early diverging clade of the Malvaceae containing cacao and jute rather its usual assigned position as originating close to cotton in a more recent multifurcating branching of subfamilies. Moreover, the *Durio* genome, shows a recent whole genome *triplication*, different from the WGD elsewhere in this plant family.

The next section summarizes the general model for generating the distribution of paralog similarities and its extension to the distribution of ortholog similarities. This is followed by a discussion of the Malvaceae genomes we study and a sketch of our inferential procedures. A full exposition of our results and inferences follows. The discussion of *Durio* evolution appears in the Appendix, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2019.2955649.

2 The General Model

In our general model for WGD and fractionation, at times $t_1 < \cdots < t_{n-1}$ each gene in the population of $M_i > 0$ genes is independently replaced by j daughter genes, where $1 \leq j \leq r_i$. The quantity $r_i \geq 2$ represents the *ploidy* of the ith event and $r_i - j \geq 1$ is considered the effect of *fractionation* on the progeny of the parent gene. The trajectory of the process is in effect a sample point from the $n-1$ probability distributions $u_1(i), \ldots, u_{r_i}(i)$ for $i = 1, \ldots, n-1$. There is no provision for $u_0(i) > 0$, a condition called "no lineage extinction". The fact that the $u_j(i)$ are defined over the progeny of each gene independently may be characterized as "sibling rivalry"; there is no constraint or relation on the survival of "cousins". Motivations for the "no lineage extinction" and "sibling rivalry" assumptions are given in [7], [9].

Let $\mathbf{a}(i) = (a_1(i), \ldots, a_{r_i}(i))$ represent the numbers of genes at time t_i, of which $1, \ldots, r_i$, respectively, survive until t_{i+1}, so that

$$M_i = \sum_{j=1}^{r_i} a_j(i), \qquad M_{i+1} = \sum_{j=1}^{r_i} j a_j(i). \tag{1}$$

Given M_i, the probability of $\mathbf{a}(i)$ is

$$P_{r_i}(\mathbf{a}(i)) = \binom{M_i}{a_1(i), \ldots, a_{r_i}(i)} (u_1(i)^{a_1(i)} \ldots u_{r_i}(i)^{a_{r_i}(i)}), \tag{2}$$

and the probability of an entire trajectory, defining a paralog gene tree is

$$P_{r_1}(\mathbf{a}(1)) \ldots P_{r_{n-1}}(\mathbf{a}(n-1)), \tag{3}$$

with $M_1 \geq 1$ given and the other M_i determined by Equation (1).

Once we know how to calculate these probabilities, it is possible to calculate the $\mathbf{E}(M_i)$. And using the independence of the trajectories starting at any two sibling genes existing at time t_i, and their independence from the trajectory between time t_1 and t_i, we can calculate $\mathbf{E}(N_i)$ the expected number of pairs of genes at time t_n originating at time t_i.

The accumulation of multinomial coefficients in Equations (2) and (3), and the potentially high degree polynomials might seem computationally formidable. In practice, however, n seldom exceeds 5 or 6, and the r_i are generally 2 or 3. Thus individual instances of the model are generally computationally tractable.

For example, suppose there is just $M_1 = 1$ gene at time t_1, and suppose all $r_i = 2$. We can write $u(i) = u_2(i), i = 1, \ldots, n-1$ for the probability that both progeny of a gene at time t_i survive until time t_{i+1}. We have previously shown [6] the expected number N_i of duplicate pairs of genes born at time t_i surviving until t_n is

$$\begin{aligned} \mathbf{E}(N_1) &= u(1) \prod_{j=2}^{n-1} (1 + u(j))^2 \\ \mathbf{E}(N_i) &= \prod_{j=1}^{i-1} (1 + u(j)) u(i) \prod_{j=i+1}^{n-1} (1 + u(j))^2 \\ \mathbf{E}(N_{n-1}) &= \prod_{j=1}^{n-2} (1 + u(j)) u(n-1). \end{aligned} \tag{4}$$

Suppose there are $n_A - 1 - s$ WGD in species A after speciation and $n_B - 1 - s$ in species B. Let

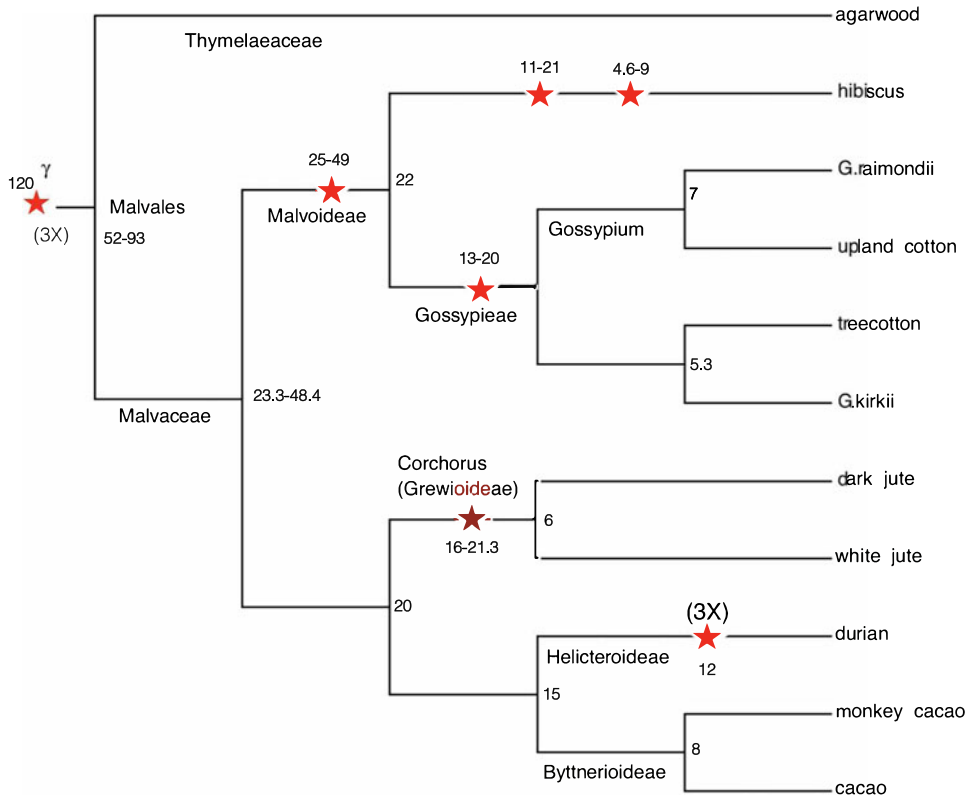


Fig. 1. Phylogenetic relationships among the Malvaceae, showing WGDs and speciation events. Numbers indicate millions of years from the event to the present, estimated from the following sources: agarwood(*Aquilaria*) divergence from Malvaceae [8], [16] and other references, and Malvoideae divergence from Grewioideae and Byttnerioideae [15], other times in the Malvoidae from [15]. Corchorus WGD is from [18], Grewioideae and Byttnerioideae divergence is from [20], and other points, like the the speciation of the two Byttnerioideae, simply interpolated. Most of these estimated times are 30–50 percent less than those given in the Time Tree of Life [13] or the Angiosperm Phylogeny Website [14], but are based on more recent technology.

$$F_A = \prod_{i=s}^{n_A-1}(1+u^A(i))$$
$$F_B = \prod_{k=s}^{n_B-1}(1+u^B(k)), \tag{5}$$

be the expectation of the "amplifying factors" affecting the distribution of orthologs due to these WGD. Then

$$\mathbf{E}(O_1) = \frac{1}{2}u(1)\prod_{j=2}^{s-1}(1+u(j))^2F_AF_B$$
$$\mathbf{E}(O_i) = \frac{1}{2}\prod_{j=1}^{i-1}(1+u(j))u(i)\prod_{j=i+1}^{s-1}(1+u(j))^2F_AF_B \tag{6}$$
$$\mathbf{E}(O_s) = \frac{1}{4}\prod_{j=1}^{s-1}(1+u(j))F_AF_B,$$

are the expected number of ortholog pairs observed after the n_A-1-s WGD in species A by which time there will have been n_B-1-s WGD in species B. (We dispense with the terminology of "outparalogs" versus "orthologs", since we are keeping track of the event at which each pair of homologous genes originates).

The three key factors in our improved model, terms in Equations (5) and (6), are $(1+u^A(s)), (1+u^B(s))$ and $(1+u(s-1))$. Between the two successive WGD, at time t_{s-1} in the pre-speciation genome, and t^A_{s+1} in genome A and t^B_{s+1} in genome B, the same fractionation regime should hold, despite the speciation at t_s. Writing

$$-\log u(s-1) = \rho(t_s - t_{s-1})),$$
$$-\log u^A(s) = \rho_A(t^A_{s+1} - t_s)), \tag{7}$$
$$-\log u^B(s) = \rho_B(t^B_{s+1} - t_s)),$$

our model presumes $\rho = \rho_A = \rho_B$. The same proportional rate should hold before and after speciation, since speciation is a population-level event in the first instance, not involving any genome-level changes, in contrast with WGD.

3 The Malvaceae

The Malvaceae are family of plants in the rosid order Malvales. Considering only the genomes studied here, this family bifurcated early into two lineages, as indicated in Fig. 1, giving rise the subfamily Malvoideae and to the three sister subfamilies Grewioideae, Helicteroideae and Byttnerioideae. Our Malvoideae genomes reveal repeated instances of WGD, the Grewioideae and Helicteroideae only one each, and the Byttnerioideae none. Of note is that the genera that boast sequenced genomes are almost all economically important ones (cf [10]).

We use the SynMap software on CoGe [11], [12], and have direct access to some of the data, in an appropriate format, among those available on the CoGe platform. Those genome data gathered elsewhere (cited below) were uploaded to a temporary private account on CoGe for purposes of the present research.

The hibiscus, or "Rose of Sharon", (*Hibiscus syriacus*) genome [15] has undergone at least three WGD events. Two

cotton genomes (*Gossypium raimondii* and *Gossypium hirsutum*) [16] are in a genus closely related to *Hibiscus*, i.e., in the same subfamily, the Malvoideae, and share the first of the ancestral hibiscus WGD events. They also have experienced a more recent WGD proper to the *Gossypium* genus. *G. hirsutum* is a recent tetraploid (4MYa) of *G. raimondii* and *G. arboreum*, so that the comparison of *G. raimondii* with *G. hirsutum* can be viewed as a proxy for the comparisons between *G. raimondii* and the *G. arboreum* subgenome of *G. hirsutum*. We also made some use of two other genomes [17] from the Gossypieae tribe, *Kokia drynarioides* and *Gossypioides kirkii*, although we did not include them in the full comparative protocol we applied to the other genomes mentioned here.

The Malvaceae species in our collection include two jute genomes (*Corchorus olitorius* and *Corchorus capsularis* [18], [19], which share a WGD event, as well as cacao (*Theobroma cacao)* [21], and the closely related "monkey cacao" (*Herrania umbratica*) [22] which have been WGD-free over the last 120 My.

Of particular interest is the recently sequenced *Durio zibethinus* genome [23], which has been considered relatively close phylogenetically to the cotton genomes, but which is actually far from cotton and close to cacao. (See the Appendix, available in the online supplemental material.)

As an outgroup, we use the agarwood (*Aquilaria agallocha*) genome [24], not in the Malvaceae family, but in the same order, Malvales.

The ancestor of the core eudicots contained 21 chromosomes, resulting from a tripling of a seven-chromosome precursor [25], [26]. This is known as the "γ" tripling. Over half of the known flowering plants, including the Malvaceae, belong to this group. It is important to note that evidence of this event shows up in all our analyses.

4 Inference on the Distribution of Similarities

Knowing the expected number of pairs of genes originating at each WGD in the past is the first step in predicting the full distribution of similarities. The second step is to derive the actual distribution of gene pair similarities, or an appropriate approximation to it, for each of the $n-1$ WGD events.

One way gene pair divergence may be measured is in terms of a probability p reflecting *similarity* – the proportion of nucleotide positions that are occupied by the same base in the two orthologs (or paralogs). In analogy to radioactive decay, we relate p to time t as a negative exponential: $p = e^{-\lambda t}$, for some constant λ.

The densities of similarities of pairs generated by the ith WGD can be approximated by a normal distribution $\mathbf{N}(p_i, \sigma_i^2)$, and the expected frequency by

$$F_i = \mathbf{E}(O_i)\mathbf{N}(p_i, \sigma_i^2). \tag{8}$$

We can predict the entire frequency distribution over all events as

$$F(p) = \sum_{i=1}^{n-1} F_i(p_i), \tag{9}$$

Were we to try to decompose a mixture of distributions arising from a single comparison of two genomes, we would likely use standard software such as Emmix [27] or the R package *mixtools*. Such approaches, however, suffer from a number of problems, stemming from deviations from normality, very large differences among the O_i, disproportionate noise at lower values of p, and reliance on biologically unaware significance testing.

We nevertheless use mixture-of-distribution estimation, but our approach tries to attenuate some of their problems through pairwise comparison of all the genomes in a group of plants, a group phylogenetically diverse enough to include lineages with different WGD histories, but not so scattered that fine-grained relationships within the group will be missed. In our experience, the appropriate phylogenetic level is the plant family. The present study of the family Malvaceae builds on our previous work on the Brassicaceae [7] and on the Solanaceae [6].

The advantage in a family-based approach is that many pairs of genomes will share the same events, WGD or speciation, and hence component distributions showing the same average p. Indeed, knowing the phylogeny of the family helps us to discard spurious component distributions caused by small sample size fluctuations and to detect other component distributions that may be below the level of significance.

We first locate candidate p_i in each comparison by picking out local modes in the distribution of similarities. We are guided by the phylogenetically-based knowledge that some of these p_i are shared among several genome comparisons, since they reflect the same events. (We have developed a phylogenetic method based on modes and WGD [7], but the point of the present paper being the estimation of fractionation rates, we simply incorporated a phylogeny most in keeping with the biological literature, though this is confirmed by neighbour-joining in the Appendix, available in the online supplemental material. In each comparison, we fix these p_i in a maximum likelihood mixture of distributions analysis to produce the amplitude and variance of each component. We then iterate, allowing the p_i to shift a slight amount to improve the visual fit of the whole distribution. The amplitude and variance inferred for each component are estimates of the number of O_1 pairs, O_2 pairs, etc. These numbers, together with the p_i, can then be used to produce estimates of the $u(i)$.

We use modes to estimate the p_i because the overlapping tails of the component distributions preclude their estimation by averaging. Furthermore, the fact that hundreds or thousands of syntenically validated orthologs support the position of the mode, even if they are not exactly the same set of orthologs in each comparison, means we are not limited to the small set of genes that are single copy in all the genomes, and which in any case are susceptible to non-parallel fractionation.

Once we have decomposed the mixed distribution of similarities into its component distributions, the area under the distribution due to each component is an estimate of $O_i, i = 1, \ldots, s$, the number of ortholog pairs due to the WGD at time t_i, whose expectation we derived in Equation (6). Then

$$\frac{O_1}{\frac{1}{2}u(1)\Pi_{j=2}^{s-1}(1+u(j))^2}, \quad \frac{O_i}{\frac{1}{2}\Pi_{j=1}^{i-1}(1+u(j))u(i)\Pi_{j=i+1}^{s-1}(1+u(j))^2}, \quad \frac{O_s}{\frac{1}{4}\Pi_{j=1}^{s-1}(1+u(j))}, \tag{10}$$

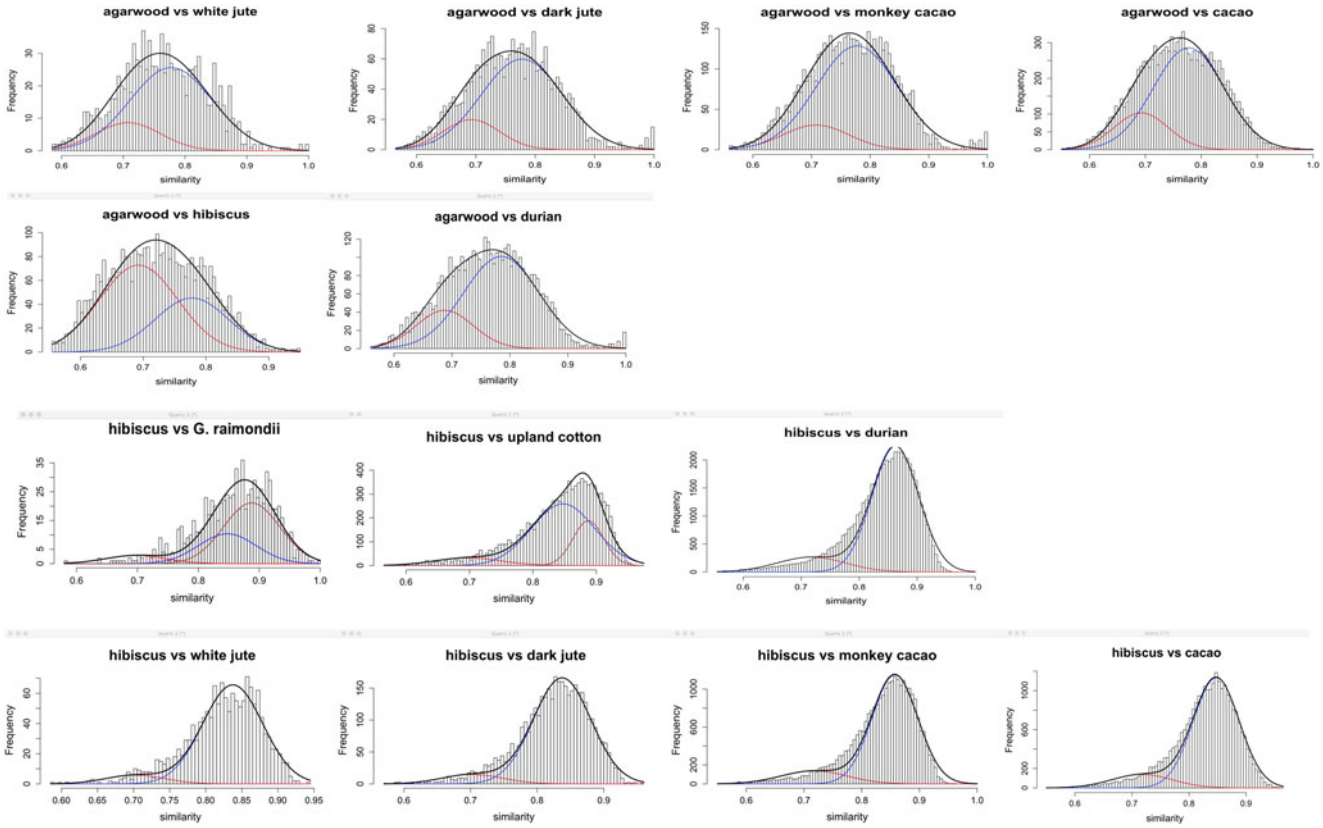


Fig. 2. Aquilaria and Hibiscus similarities.

are estimates of the proportion of the area under the original distribution due to only the WGD at time t_i or the speciation at time t_s. There are the same number of independent proportions $(s-1)$ as parameters $u(i), i = 1\ldots, s-1$, so that we can estimate all the parameters.

Although the roles of F_A and F_B are an important analytical result, the great disparity in the quantity of data produced by the various comparisons, due entirely to methodological disparities, preclude any systematic attempt to analyze them. We can nevertheless suggest a procedure for further work. First, for any two genomes A and B, an estimate of the product F_AF_B can be obtained by averaging all the quantities in (10) above. Suppose the lineage of genome A consists of a series of WGD interspersed with speciations giving rise to genomes $B_1, B_2, \ldots$, themselves with no post-speciation WGD, then the successive values of F_A at these speciation points could be compared in order to obtain some of the factors $(1 + u(i))$ in Equations (5) at the intervening WGD.

If, on the other hand, the genomes $B_1, B_2, \ldots$ themselves underwent WGD after divergence from the A lineage, there is no direct way of factoring the product F_AF_B.

5 Results

Figs. 2, 3 and 4 depict the distributions and their decompositions for the 38 comparisons, including seven self-comparisons, we succeeded in producing by CoGe. Technical difficulties with CoGe, beyond our control, such as the unavailability of coding sequence for *Aquilaria* and the Gossypieae genomes, or the small numbers of gene pairs detected, prevented us from obtaining five pairwise genomic comparisons and two self-comparisons. Table 1 presents details about the components distribution associated with the ancient or recent speciation in the lineages of every pair of genomes.

The individual graphs all share the same x-axis, ranging from 0.5 to 1.0. However, the y-axis scale differs greatly among the comparisons. Part of the reason for this is no doubt the different lengths of the time period since speciation, allowing fractionation to reduce the number of gene pairs. More important in many cases are difficulties with the genome assembly or, more crucially, its annotation. Among the genomes we analyzed, those with the most results included the hibiscus, upland cotton, cacao, monkey cacao, and durian genomes. The most problematic were the *G. raimondii* genome, the two jute genomes and the two Gossypieae. This does not necessarily speak to the quality of these genome sequences nor their annotation; there are a number of possible vulnerabilities beyond our control in the pipeline between the original database, through the uploading to CoGe, to the SynMap computations and display. These problems do not extend to the other genomes and though they make affect the precision of some of the results, should not result in any biases.

In general, we note some moderate problems of fit between the model and the data. Many of these have to do with the evident non-normality of the component subgenomes. This is clearest in the speciation components, where there is skewness involving a steeper slope on the side closest to 100 percent similarity. Thus the data are fewer than expected from a normal approximation on this side, and contribute to a "shoulder" that exceeds the normal curve on the side closer to 50 percent similarity. In some cases, this partially obscures a

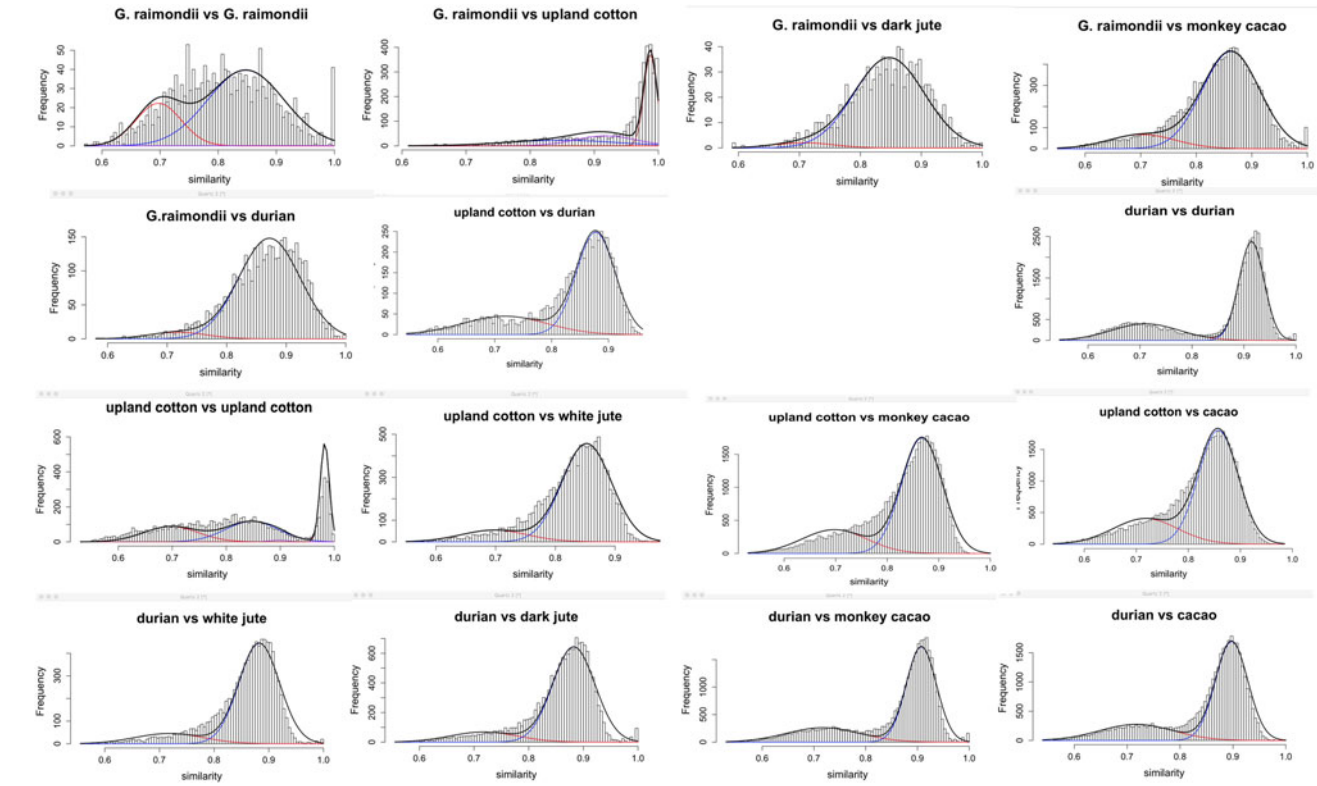


Fig. 3. Gossypium and Durio similarities.

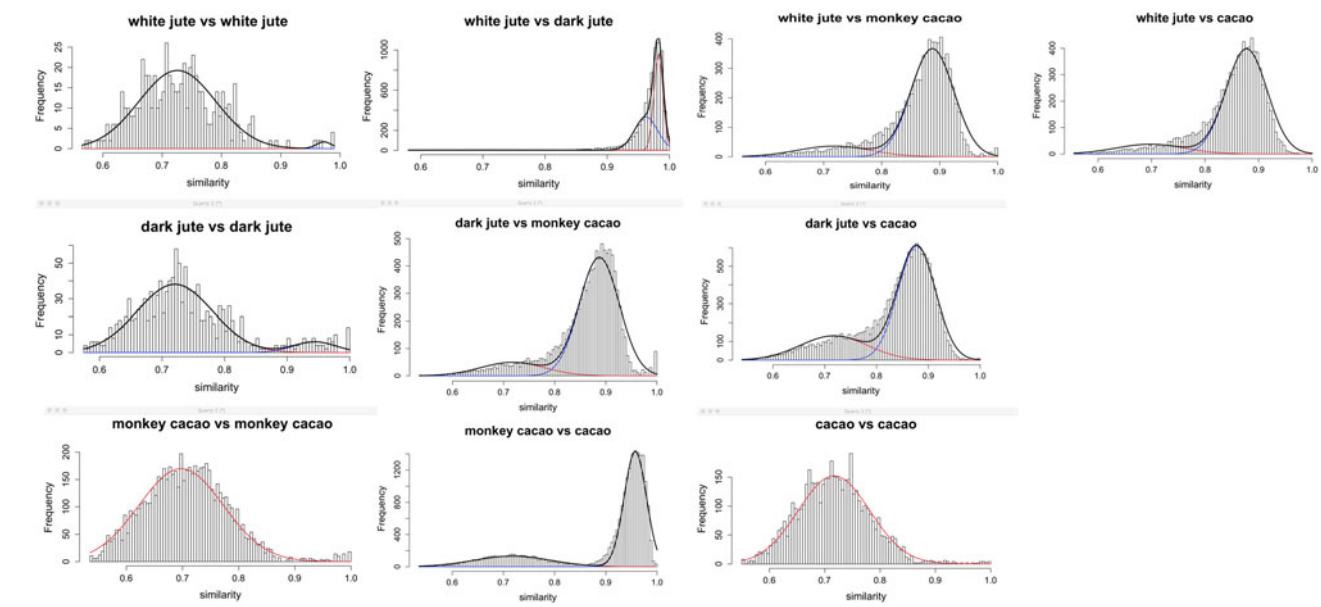


Fig. 4. Grewioideae and Byttnerioideae similarities.

slightly earlier peak, but careful inspection usually reveals a local mode where the peak is expected.

The two Gossypieae were not included in most of our analyses, largely because no annotations were available. Without CDS information, these genomes could only be compared to other genomes by using the capacity of SynMap to locate likely homology regions in sequence data based on the CDS in another genome. This problem was also encountered with the agarwood genome, but comparisons were much more productive in this case.

In the cases of *G. raimondii* genome, the two jute genomes and the two Gossypieae genomes, we were obliged to lower the minimum syntenic block size for the SynMap run from 5 to 3, in order to get even a minimal number of orthologous or paralogous pairs for further analysis. Even so, there were not enough pairs in some of the comparisons to detect some known events. For example, the white jute self-comparison does not detect the *WGD*, though it shows up both in the self-comparison of dark jute and in the comparison of the two jute genomes.

TABLE 1
Statistical Parameters of the Speciation Peaks

mean	agar-wood	hibiscus	G.rai-mondii	upland cotton	white jute	dark jute	monkey cacao	cacao	durian
agarwood		0.780	–	–	0.780	0.780	0.780	0.780	0.788
hibiscus	0.780		0.890	0.850	0.840	0.840	0.860	0.850	0.865
G.raimondii	–	0.890		0.990	–	0.850	0.865	–	0.874
upland cotton	–	0.850	0.990		0.855	–	0.870	0.860	0.880
white jute	0.780	0.840	–	0.855		0.985	0.890	0.895	0.885
dark jute	0.780	0.840	0.850	–	0.985		0.890	0.880	0.885
monkey cacao	0.780	0.860	0.865	0.870	0.890	0.890	–	0.960	0.910
cacao	0.780	0.850	–	0.860	0.895	0.880	0.960		0.900
durian	0.788	0.865	0.874	0.880	0.885	0.885	0.910	0.900	
standard deviation	**agar-wood**	**hibiscus**	**G.rai-mondii**	**upland cotton**	**white jute**	**dark jute**	**monkey cacao**	**cacao**	**durian**
agarwood		0.053	–	–	0.067	0.071	0.069	0.066	0.065
hibiscus	0.053		0.046	0.024	0.042	0.043	0.040	0.040	0.040
G.raimondii	–	0.046		0.010	–	0.057	0.044	–	0.051
upland cotton	–	0.024	0.010		0.042	–	0.039	0.039	0.034
white jute	0.067	0.042	–	0.042		0.008	0.037	0.040	0.037
dark jute	0.071	0.043	0.057	–	0.008		0.039	0.036	0.039
monkey cacao	0.069	0.040	0.044	0.039	0.037	0.039		0.022	0.029
cacao	0.066	0.040	–	0.039	0.040	0.036	0.022		0.030
durian	0.065	0.040	0.051	0.034	0.037	0.039	0.029	0.030	
component proportion	**agar-wood**	**hibiscus**	**G.rai-mondii**	**upland cotton**	**white jute**	**dark jute**	**monkey cacao**	**cacao**	**durian**
agarwood		0.514	–	–	0.789	0.731	0.839	0.703	0.761
hibiscus	0.514		0.606	0.238	0.920	0.914	0.853	0.862	0.852
G.raimondii	–	0.606		0.491	–	0.954	0.850	–	0.942
upland cotton	–	0.238	0.491		0.863	–	0.753	0.754	0.719
white jute	0.789	0.920	–	0.863		0.566	0.856	0.813	0.853
dark jute	0.731	0.914	0.954	–	0.566		0.845	0.723	0.856
monkey cacao	0.839	0.853	0.850	0.753	0.856	0.845		0.759	0.730
cacao	0.703	0.862	–	0.754	0.813	0.723	0.759		0.729
durian	0.761	0.852	0.942	0.719	0.853	0.856	0.730	0.729	

Modes identified visually, assuring rough concordance with other comparisons sharing the same event history. Standard deviations and component proportions determined by maximum likelihood, assuming means coincide with modes. Incremental adjustments to the means applied to correct regions that fit less well due to component overlap and non-normality followed by further rounds of likelihood maximization.

Despite these problems, some clear patterns emerge. The figures demonstrate a strong resemblance among the six comparisons with agarwood. Hibiscus shows a disproportionate number of pairs surviving from the γ event, but lacking comparisons with cotton, we can only attribute this to statistical fluctuations in the relatively low number of pairs in the agarwood comparisons.

A key trend in these figures is the compact recent speciation peaks in the comparisons of the non-Malvinon-Malvoideaeoideae genomes: the jutes, durian, cacao and monkey cacao. This contrasts consistently with the more dispersed distributions comparing the non-Malvoideae genomes to the Malvoideae. The grouping of the three non-Malvoideae subfamilies as sister groups seems confirmed by the mean similarities in Table 1 (top section). This contrasts with suggestions in [19] that *Corchorus* is the earliest branch in the Malvaceae, and in [23] that *Durio* is a sister group to *Gossypium*. Our suggestion is confirmed by a neighbor-joining analysis (not shown), which exactly reproduces the tree in Fig. 1, with the exception that *Durio* branches off slightly before *Corchorus* on the lineage leading to *Theobroma*.

As was recently shown [6] the mean similarity scores at speciation in Table 1 drop exponentially with the assigned times in Fig. 1, except for a lower similarity than expected for the divergence time of the Malvaceae species from the Malvales outgroup – see Fig. 5. Also in this figure the standard deviation of the component distribution is linearly related to its mean, again with the exception of the original Malvaceae emergence.

We used the data in the three sections of Table 1 to estimate the $u(i)$ according to Equation (6), and transformed these according to the relationship $\rho = -\frac{\log u}{t}$, to estimate proportional fractionation rates, shown in Table 2. We performed a similar analysis on the self comparisons, shown in Table 3. In both cases a range of values is shown, corresponding to the range of time estimates shown in Fig. 1. Of note is the relative constancy of rates except after recent WGD events. The larger recent rates may be methodological artifacts or may indicate a higher rate of fractionation, immediately after the WGD.

6 Discussion and Conclusion

A major methodological problem in estimating proportional fractionation rates is deciding on the time estimates for speciation and WGD events. The times given by the Timetree of Life

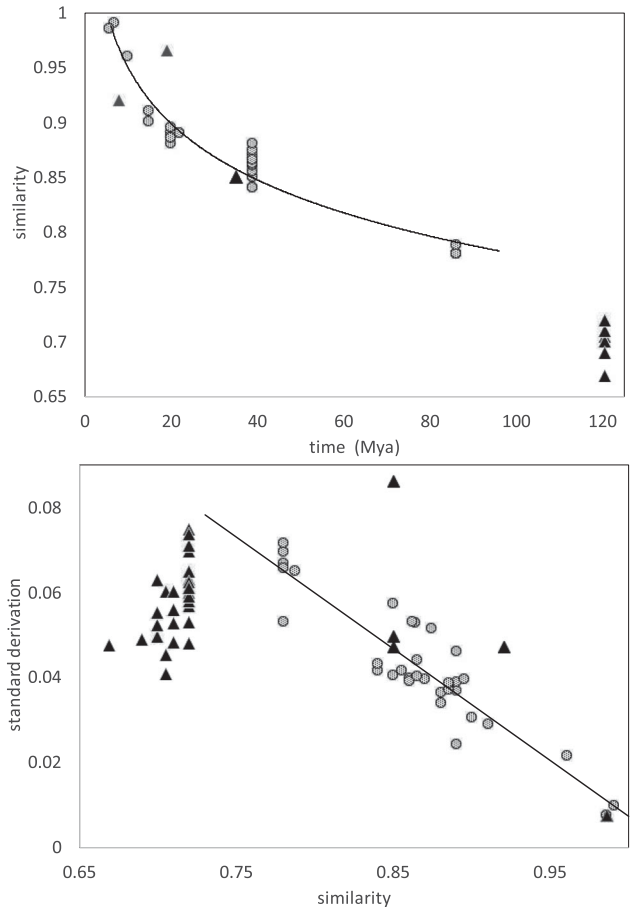


Fig. 5. Trends of peak modes and standard deviation. Trend lines fitted to Malvaceae speciation times and similarities only (dots). WGD times and similarities represented by triangles, with the large cluster pertaining to the γ triplication.

[13] or the Angiosperm Phylogeny Website [14] are based on a wide range of data and are systematically much older than those provided by more recent molecular phylogenies, often by a factor of two or more. And the more recent results usually involve a wide range of values. The accuracy of the rate estimations suffers accordingly, and can only be reported as a similarly wide range of values. This difficulty is partially due to the very different quality of the genome sequences and annotations. Over less than ten years, technology has improved by orders of magnitude, but there is still great variability among laboratories, reporting styles and journals. The studies we consulted performed "phylogenomic" estimations based on very few genes, chosen according to a variety of criteria, such as wide presence in many genomes or single-copy in all genomes. How these criteria affect the time estimates is unknown, but certainly create inconsistencies between different studies.

The principled extension of our model from WGD to speciation highlights inference difficulties due to gene loss other than fractionation. Our model focuses on gene loss within syntenic blocks and, up to the moment of speciation, disregards genes that are not in such blocks. But there are at least four reasons why genes do not show up in syntenic blocks aside from fractionation. The very criteria for defining a block, including a minimum number of gene pairs in a block with a maximum number of unpaired genes between successive pairs, necessarily excludes many pairs. Second, local rearrangements may disrupt blocks so that the resulting fragments are too small. Third, related to the previous, one member of a gene pair may be displaced, so that the pair no longer remains in syntenic context. Finally, and most important, both members of a pair may be lost, by silencing, pseudogenization, excision or other mechanism, transcending the "no lineage extinction" constraint on our model. A good proportion of genes in most organisms can be considered non-essential. Even those genes deemed essential under usual conditions, may be lost, and this has become of increasing interest as a mechanism of evolution [30].

How pervasive this non-fractionation loss may be is not known, but it will be necessary at some point to incorporate

TABLE 2
Ranges of Fractionation Rates Preceding Speciation, Starting After the Most Recent Preceding WGD, be it γ, the Malvoideae Event, the More Recent *Gossypium* Event or the *Corchorus* Duplication

ρ	agar-wood	hibiscus	G.rai-mondii	upland cotton	white jute	dark jute	monkey cacao	cacao	durian
agarwood		0.002 −0.004			0.027 −0.069	0.022 −0.055	0.033 −0.083	0.019 −0.049	0.025 −0.062
hibiscus	0.002 −0.004		0.040 −0.364		0.032 −0.043	0.031 −0.042	0.024 −0.033	0.025 −0.034	0.024 −0.033
G.raimondii		0.040 −0.364		0.082 −0.177		0.038 −0.052	0.024 −0.033		0.036 −0.048
upland cotton			0.082 −0.177		0.025 −0.034		0.017 −0.023	0.017 −0.023	0.015 −0.020
white jute	0.027 −0.069	0.032 −0.043		0.025 −0.034		0.042 −0.064	0.024	0.020	0.024
dark jute	0.022 −0.055	0.031 −0.042	0.038 −0.052		0.042 −0.064		0.023	0.014	0.024
monkey cacao	0.033 −0.083	0.024 −0.033	0.024 −0.033	0.017 −0.023	0.024	0.023		0.015	0.014
cacao	0.019 −0.049	0.025 −0.034		0.017 −0.023	0.020	0.014	0.015		0.014
durian	0.025 −0.062	0.024 −0.033	0.036 −0.048	0.015 −0.020	0.024	0.024	0.014	0.014	

TABLE 3
Ranges of Fractionation Rates Following WGD Until the Next WGD in a Lineage, from Self-Comparisons

WGD	G.raimondii	upland cotton	white jute	dark jute	durian
gamma	0.047-0.063	0.033-0.044	0.028-0.030	0.024-0.025	0.02
Malvoideae	0.093-0.670	0.094-0.677			
Gossypium	0.405-0.622				
jute			0.332-0.442	0.234-0.312	
durian					0.063

it into our models. From a mathematical point of view, our current branching process is super-critical, which is not realistic, even if some organisms, like hibiscus, have accumulated very large numbers of genes through WGD.

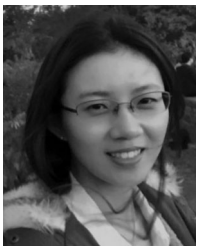
The fractionation rates found in the present study are of the order of a half those calculated for the Solanaceae [6]. This correlates with the somewhat faster evolutionary rate of the latter [8] and may represent the longer generation time of the woody tree mode prevalent in the Malvaceae compared to the annual herbaceous growth mode of the Solanaceae. If higher fractionation rates for recent WGD prove not to be methodological artifacts, this may be the result of long-lasting pairs settling into a stable subfunctionalization, particularly pertinent in the case of the γ event.

Acknowledgments

Research and publication costs were supported in part by Discovery Grant No. 8867-200 from the Natural Sciences and Engineering Research Council of Canada. DS holds the Canada Research Chair in Mathematical Genomics.

References

[1] Amborella genome project, "The Amborella genome and the evolution of flowering plants," *Science*, vol. 342, 2013, Art. no. 1241089.
[2] E. C. H. Chen *et al.*, "The dynamics of functional classes of plant genes in rediploidized ancient polyploids," *BMC Bioinf.*, vol. 14, no. S-15, 2013, Art. no. 19.
[3] J. F. Gout, D. Kahn, L. Duret, and Paramecium Post-Genomics Consortium, "The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution," *PLoS Genet.*, vol. 6, 2010, Art. no. 1000944.
[4] B. C. Thomas, B. Pedersen, and M. Freeling, "Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes," *Genome Res.*, vol. 16, no. 7, pp. 934–946, 2006.
[5] Y. Zhang and D. Sankoff, "The similarity distribution of paralogous gene pairs created by recurrent alternation of polyploidization and fractionation," in *Proc. RECOMB Int. Workshop Comparative Genomics*, 2017, vol. 10562, pp. 1–13.
[6] Y. Zhang, C. Zheng, and D. Sankoff, "Speciation and rate variation in a birth-and-death account of WGD and fractionation; the case of Solanaceae," *Proc. 16th RECOMB Comparative Genomics Satellite Workshop*, 2018, pp. 146–160.
[7] D. Sankoff, C. Zheng, Y. Zhang, J. Meidanis, E. Lyons, and H. Tang, "Models for similarity distributions of syntenic homologs and applications to phylogenomics," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 3, pp. 727–737, May/Jun. 2019. doi: 10.1109/TCBB.2018.2849377
[8] A. R. De La Torre, Z. Li, T. Van de Peer, and P. K. Ingvarsson, "Contrasting rates of molecular evolution and patterns of selection among Gymnosperms and flowering Plants," *Mol. Biol. Evol.*, vol. 34, pp. 1363–1377, 2017. [Online]. Available: https://doi.org/10.1093/molbev/msx069
[9] Y. Zhang, C. Zheng, and D. Sankoff, "Pinning down ploidy in paleopolyploid plants," *BMC Genomics*, vol. 19, 2018, Art. no. 287.
[10] G. C. Vallée, D. Santos Muñoz, and D. Sankoff, "Economic importance, taxonomic representation and scientific priority as drivers of genome sequencing projects," *BMC Genomics*, vol. 17, no. (Suppl 10), 2016, Art. no. 782.
[11] E. Lyons and M. Freeling, "How to usefully compare homologous plant genes and chromosomes as DNA sequences," *The Plant J.*, vol. 53, pp. 661–673, 2008.
[12] E. Lyons *et al.*, "Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar and grape: CoGe with rosids," *Plant Physiol.*, vol. 148, pp. 1772–1781, 2008.
[13] S. Kumar, G. Stecher, M. Suleski, and S. B. Hedges, "TimeTree: a resource for timelines, timetrees, and divergence times," *Mol. Biol. Evolution*, vol. 34, pp. 1812–1819. 2017. [Online]. Available: http://timetree.org
[14] P. F. Stevens, "Angiosperm phylogeny website. Version 14," 2017. [Online]. Available: http://www.mobot.org/MOBOT/research/APweb/
[15] Y. M. Kim *et al.*, "Genome analysis of *Hibiscus syriacus* provides insights of polyploidization and indeterminate flowering in woody plants," *DNA Res.*, vol. 24, pp. 71–80, 2017.
[16] F. Li *et al.*, "Genome sequence of cultivated Upland cotton (Gossypium hirsutum TM-1) provides insights into genome evolution," *Nat. Biotechnol.*, vol. 33, pp. 524–30, 2015.
[17] C. E. Grover *et al.*, "Comparative genomics of an unusual biogeographic disjunction in the cotton tribe (Gossypieae) yields insights into genome downsizing," *Genome Biol. Evol.*, vol. 9, pp. 3328–3344, 2017. doi:10.1093/gbe/evx248.
[18] D. Sarkar *et al.*, "The draft genome of *Corchorus olitorius* cv. JRO-524 (Navin)," *Genome Data*, vol. 12, pp. 151–154, 2017.
[19] M. S. Islam *et al.*, "Comparative genomics of two jute species and insight into fibre biogenesis," *Nat. Plants*, vol. 3, 2017, Art. no. 16223.
[20] C. Baraloto *et al.*, "Using functional traits and phylogenetic trees to examine the assembly of tropical tree communities," *J. Ecol.*, vol. 100, pp. 690–701, 2012.
[21] X. Argout *et al.*, "The genome of *Theobroma cacao*," *Nat. Genetics*, vol. 43, pp. 101–108, 2011.
[22] NCBI, "*Herrania umbratica* Annotation Release 100," 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Herrania_umbratica/100/
[23] B. T. Teh *et al.*, "The draft genome of tropical fruit durian (*Durio zibethinus*)," *Nat. Genetics*, vol. 49, pp. 1633–164, 2017.
[24] C. H. Chen *et al.*, "Identification of cucurbitacins and assembly of a draft genome for *Aquilaria agallocha*," *BMC Genomics*, vol. 15, 2014, Art. no. 578.
[25] O. Jaillon *et al.*, "The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla," *Nature*, vol. 449, pp. 463–467, 2007.
[26] C. Zheng, E. Chen, V. A. Albert, E. Lyons, and D. Sankoff, "Ancient eudicot hexaploidy meets ancestral eurosid gene order," *BMC Genomics*, vol. 14, p. S7:S3, 2013.
[27] 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. Statist. Softw.*, vol. 4, no. 2, pp. 1–14, 1999.
[28] M. Sun *et al.*, "Phylogeny of the Rosidae: A dense taxon sampling analysis," *J. Syst. Evol.*, vol. 54, pp. 363–391, 2016.
[29] J. E. Richardson, B. A. Whitlock, A. W. Meerow, and S. Madriñàn, "The age of chocolate: A diversification history of Theobroma and Malvaceae," *Front. Ecol. Evol.*, vol. 3, 2015, Art. no. 120. doi: 10.3389/fevo.2015.00120
[30] R. Albalat and C. Cañestro, "Evolution by gene loss," *Nat. Rev, Genet.*, vol. 17, pp. 379–391, 2016.
[31] J. Wang *et al.*, "Recursive paleohexaploidization shapes the durian genome," *Plant Physiol.*, vol. 179, pp. 209–219, 2019, doi: 10.1104/pp.18.00921.



Yue Zhang received the master's degree in statistics from Carleton University in Ottawa, Canada and the PhD degree while a member of Dr. Sankoff's Lab at the University of Ottawa, Ottawa, Canada, focusing on the statistical analysis of subgenome evolution after paleopolyploidy.

Chunfang Zheng received the master's and PhD degrees in biology from the University of Ottawa, Ottawa, Canada, 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.

Sindeed Islam is currently working toward the undergraduate degree at the University of Ottawa, Ottawa, Canada, and has been a volunteer student and now a research assistant in the Sankoff Lab.

Yong-Min Kim received the PhD degree from Chonnam National University, Gwangju, South Korea. He is currently working toward the degree at Seoul National University, Seoul, South Korea. He is currently at KOBIC/Korea Research Institute of Bioscience and Biotechnology in Daejeon, South Korea, where he is a senior researcher and team leader of the Bioinformatics Team

David Sankoff received the PhD degree in mathematics from McGill University, Quebec, Canada, and has been a member of the Centre de recherches mathematiques in Montreal for many years. He currently holds the Canada Research chair in mathematical genomics in the Mathematics and Statistics Department, University of Ottawa, Ottawa, Canada, 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. He is a member of the IEEE.