Evolutionary Model for the Statistical Divergence of Paralogous and Orthologous Gene Pairs Generated by Whole Genome Duplication and Speciation

Yue Zhang, Chunfang Zheng, and David Sankoff

Abstract—We outline a principled approach to the analysis of duplicate gene similarity distributions, based on a model integrating sequence divergence and the process of fractionation of duplicate genes resulting from whole genome duplication (WGD). This model allows us to predict duplicate gene similarity distributions for a series of two or three WGD, for whole genome triplication followed by a WGD, and for triplication, followed by speciation, followed by WGD. We calculate the probabilities of all possible fates of a gene pair as its two members proliferate or are lost, predicting the number of surviving pairs from each event. We discuss how to calculate maximum likelihood estimators for the parameters of these models, illustrating with an analysis of the distribution of paralog similarities in the poplar genome.

Index Terms—Mixture of distributions, fractionation, probability model

1 Introduction

Speciation creates a set of orthologous gene pairs involving all or almost all genes in the two daughter genomes, and these pairs all evolve according to a dynamic of decaying similarity as gene sequence and amino acid sequences inexorably diverge through random single nucleotide mutation. Whole genome duplication (WGD) creates a set of paralogous pairs involving all genes in the affected genome, and these pairs also diverge through the same processes of random mutation. In addition, paralogous pairs may disappear through the process of fractionation, whereby one of the two genes is excised, pseudogenized or otherwise removed as a recognizable coding gene.

A widespread practice in comparative genomics is to infer the nature and timing of evolutionary events through the examination of the distribution of similarities between orthologous or paralogous gene pairs. This is done by identifying local modes or peaks in the distribution, and inferring that duplications around these points were generated by speciation or WGD events. The identification of the peaks may be accomplished by visual inspection or, if the data seem noisy, by software available for the analysis of mixture of normal distributions, such as Emmix [1].

There is, however, no rigorous methodology for interpreting the volume of the individual normal distributions inferred by such general methods. Indeed, many possible outputs may not conceivable as produced by genomic events. Moreover, there is a general tendency to overfit–to infer components of the mixture that are really just reflect statistical fluctuation in the data.

In this paper, we outline a principled approach to the analysis of duplicate gene similarity distributions. It is based on the simplest, one-parameter model of sequence divergence, as well as an equally simple, one-parameter model of the fractionation process. We extend this to build models of a series of two or three WGD, of whole genome triplication followed by a WGD, characteristic of the core eudicots, and of a triplication, followed by a speciation and then a WGD in one of the two daughter species. In all these cases, we calculate the probabilities of all possible fates of a gene pair as its two members proliferate or are lost, to predict the number of surviving pairs from each event. To our knowledge, this is the first method to account for the volume of the component normals of a distribution of similarities, preliminary to an evolutionarily meaningful inference procedure.

We also outline how to infer the parameters of a model, using maximum likelihood methods. We illustrate with an analysis of the distribution of paralog similarities in the genome of the poplar, *Populus trichocarpa*.

We will conclude with a detailed discussion of the advantages and difficulties of our approach and detailed proposals for further research.

2 The Building Blocks

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 orthologs

- *The authors are with the Department of Mathematics and Statistics, University of Ottawa, Ottawa, ON K1N 6N5, Canada. E-mail: {yzhan481, sankoff}@uottawa.ca, chunfang313@gmail.com.*

Manuscript received 3 Apr. 2017; revised 15 May 2017; accepted 22 May 2017. Date of publication 14 July 2017; date of current version 5 Oct. 2018. (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.2017.2712695

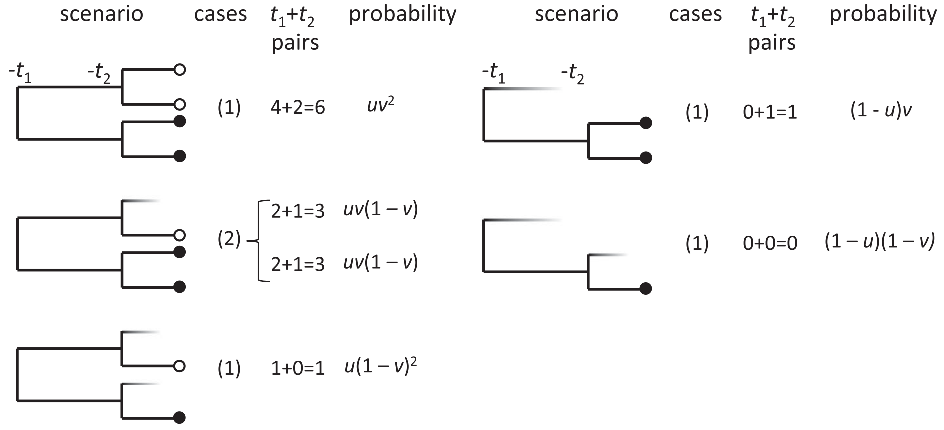

Fig. 1. Components of the number of surviving pairs created by WGDs at t_1 and t_2.

(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 [2].

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

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

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 pair, is represented by a parameter $u \in [0, 1]$, representing the probability, for a pair of 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

$$u = \mathrm{e}^{-\rho t}, \tag{2}$$

where ρ is the fractionation parameter.

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 estimate of u (and of ρ).

3 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 $\mathrm{e}^{-\rho(t_1-t_2)}$ and $\mathrm{e}^{-\rho(t_2-0)} = \mathrm{e}^{-\rho t_2}$, respectively, for a pair of genes present at the start of the time interval, that neither 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

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

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

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

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

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

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

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

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

and the probability of an unpaired gene is

$$Q^* = P(C). \tag{10}$$

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

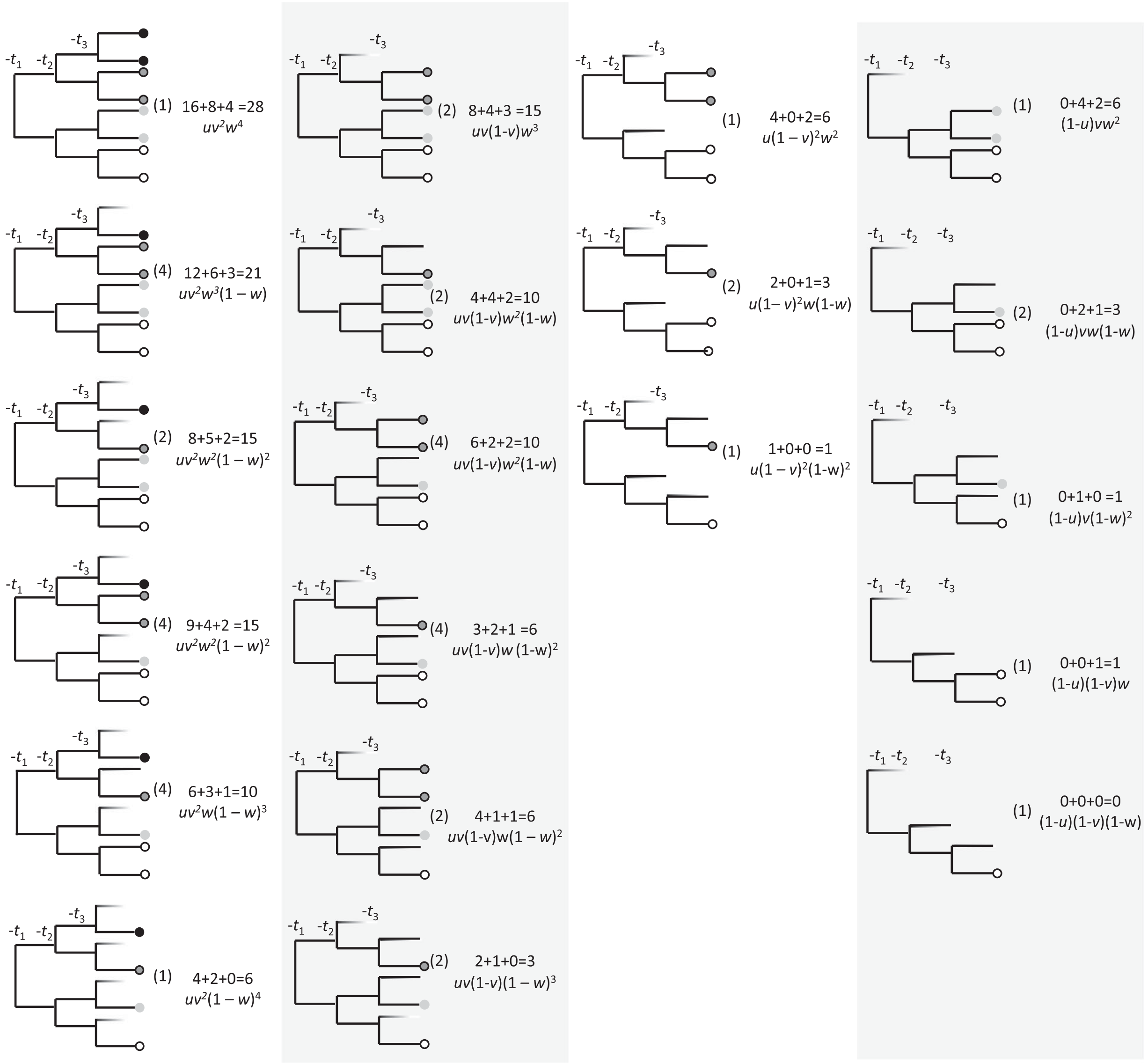

Fig. 2. Components of the number of surviving pairs created by WGDs at t_1, t_2 and t_3. See Fig. 1 to interpret the content of the various columns.

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

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

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

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.

4 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 [3]). 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.

$$\begin{aligned} E(t_1 \text{ pairs}) &= (1 - 3w^2 + 2w)uv^2 + (2 + 6w^2 + 4w)uv \\ &\quad + (1 + w^2 + 2w)u \\ E(t_2 \text{ pairs}) &= ((1 + w^2 + 2w)u + 1 + w^2 + 2w)v \\ E(t_3 \text{ pairs}) &= -2uv^2w^2 + ((2w^2 - w)u + w)v \\ &\quad + uv + w \\ E(\text{unpaired}) &= (1 - u)(1 - v)(1 - w). \end{aligned} \tag{13}$$

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.

5 Whole Genome Triplication Followed by WGD

The core eudicots contain more species than all the other groups of flowering plants combined. A whole genome triplication, called the "γ" event, occurred in the eudicot lineage just before the emergence of the core eudicots, and a large proportion of these have undergone further WGD. The model analyzed in Fig. 3 is appropriate for this case. Here

$$\begin{aligned} E(t_1 \text{ pairs}) &= (u' + 3u''')v2 + (2u' + 6u''')v + b + 3u''' \\ E(t_2 \text{ pairs}) &= -3u'''v3 + 3u'''v2 + (1 + 2u''' - u')v \\ E(\text{unpaired}) &= (1 - u''' - u')(1 - v). \end{aligned} \tag{14}$$

6 The Effect of Speciation

Up to now we have considered only WGD events, including triplications. In comparing two species, there are peaks at times corresponding to their shared WGD, followed by a single peak dating from their speciation event, but no further peaks. Fig. 4 contains the analysis of a whole genome triplication, followed by a speciation event, and a further WGD in one of the daughter genomes. Here the distribution of homologous gene pair similarities is predicted by

$$\begin{aligned} \mathbf{E}[t_1 \text{ pairs}] = {} & (6u'''z^2(1-z) - 30u'''z(1-z)^2 \\ & + 12u'''z(1-z) + 30u'z(1-z)^2)v3 \\ & + (30u'''z(1-z)^2 - 60u'z(1-z)^2)v^2 \\ & + (2u'z(1-z) + 2u'z^2 + 6u'''z^2(1-z) \\ & + 6u'''z^3 + 30u'z(1-z)^2)v + 2u'z^2 \\ & + 12u'''z^2(1-z) + 6u'''z^3 + 2u'z(1-z) \\ \mathbf{E}[t_2 \text{ pairs}] = {} & (3u'''z^3 + 12u'''z^2(1-z) - 9u'''z(1-z)^2 \\ & + 6u'z(1-z)^2)v^3 + (-9u'''z^3 \\ & + 15u'''z(1-z)^2 - 18u'''z^2(1-z) \\ & - 24u'z(1-z)^2)v^2 + (1 - u''' - u' \\ & + 2u'z(1-z) + 12u'z(1-z)^2 + 2u'z^2 \\ & + 12u'''z^3 + 24u'''z^2(1-z))v \\ & + 1 - u''' - u' + 2u'z^2 + 2u'z(1-z). \end{aligned} \tag{15}$$

Note that all of the t_2 pairs in equation (6) are orthologs, but the t_1 pairs contain a mixture of orthologs and paralogs.

7 The Case of *Populus Trichocarpa*

Though the work we have presented consists of combinatorial models, and the inference procedures are not implemented in a user-friendly package, we did analyze one data set using functions on the R platform according to the model in the previous section. We extracted data on the poplar genome [4] from the CoGe platform [5], [6], and calculated gene pairs, producing the distribution of similarities in Fig. 5. In estimating the parameters using our preliminary code, we were confined to sampling 500 out of the 13,000 pairs. Running the program repeatedly, the results were quite reproducible.

In Fig. 5, if the early event is identified with the γ triplication 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 [4]. It is also consistent with a constant fractionation rate ρ over the whole time period covered in the analysis.

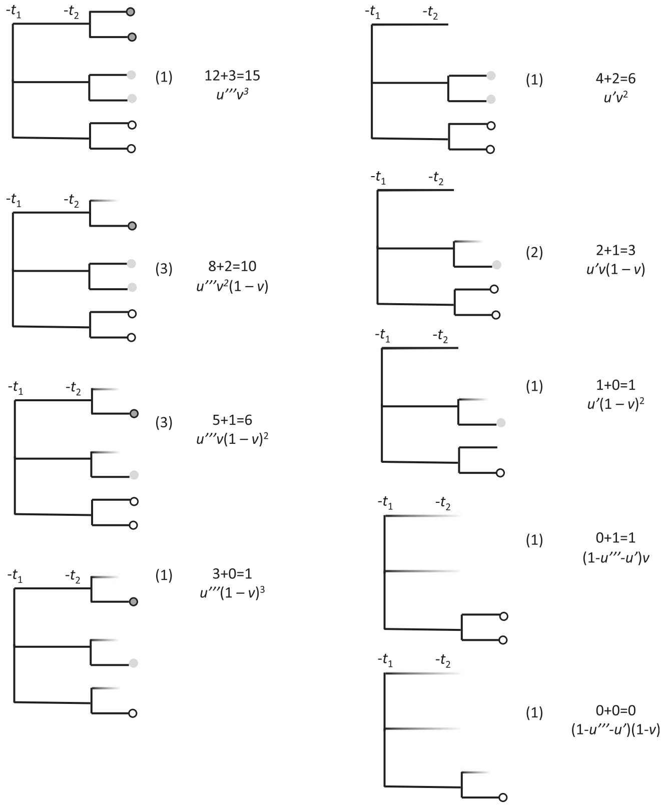

Fig. 3. Components of the number of surviving pairs created by a whole genome triplication at t_1 and a WGD at t_2.

8 Conclusions and Directions for Further Work

We have presented the first model of the simultaneous processes of duplicate gene divergence and fractionation of in the evolution of one or more species affected by WGD. This allows the prediction of both the location, shape and amplitude of the evolutionary signals, both speciation and WGD, contained in pairwise genome comparisons.

The parameter G affects the spread of the normally distributed contribution by an individual event to the overall distribution of gene pair similarities. It reflects the length of a gene, in terms of the number of nucleotides in the coding sequence, or of the number of synonymous sites, or of the number of four-fold degenerate sites. Length, however, is variable, from gene to gene, and from genome to genome, degrading the signal in the empirical distribution of similarities. One important direction for further work would be to incorporate the known properties of gene length distribution into the theory. Conceptually, this should pose little problem since since G is approximately log-normally distributed [7], so that σ^2 in equation (1) can be adjusted directly.

Duplicate genes are also produced by mechanisms other than WGD. These can be largely avoided by requiring pairs to be corresponding syntenic contexts when extracting them from the data. This eliminates most of the problems due to tandem gene duplicates.

The assumption of a constant rates of gene divergence is another first order simplification made for analytical tractability, reducing as far as possible the number of parameters to be estimated. Rates in fact are variable among genes,

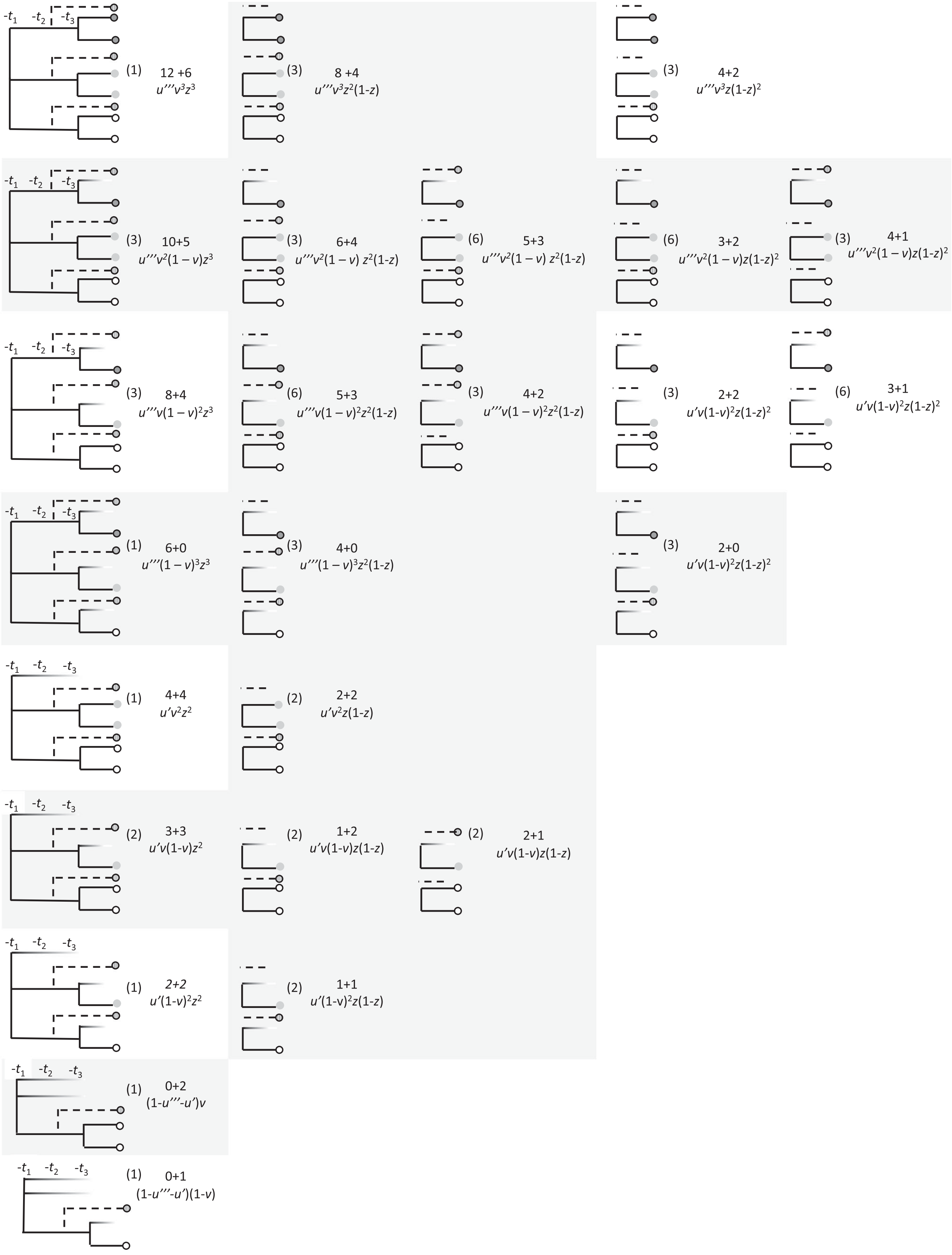
$-t_1$ $-t_2$ $-t_3$
(1) 12 +6 $u'''v^3z^3$
(3) 8 +4 $u'''v^3z^2(1\text{-}z)$
(3) 4+2 $u'''v^3z(1\text{-}z)^2$
(3) 10+5 $u'''v^2(1-v)z^3$
(3) 6+4 $u'''v^2(1-v)\, z^2(1\text{-}z)$
(6) 5+3 $u'''v^2(1-v)\, z^2(1\text{-}z)$
(6) 3+2 $u'''v^2(1-v)z(1\text{-}z)^2$
(3) 4+1 $u'''v^2(1-v)z(1\text{-}z)^2$
(3) 8+4 $u'''v(1-v)^2z^3$
(6) 5+3 $u'''v(1-v)^2z^2(1\text{-}z)$
(3) 4+2 $u'''v(1-v)^2z^2(1\text{-}z)$
(3) 2+2 $u'v(1\text{-}v)^2z(1\text{-}z)^2$
(6) 3+1 $u'v(1\text{-}v)^2z(1\text{-}z)^2$
(1) 6+0 $u'''(1-v)^3z^3$
(3) 4+0 $u'''(1-v)^3z^2(1\text{-}z)$
(3) 2+0 $u'v(1\text{-}v)^2z(1\text{-}z)^2$
(1) 4+4 $u'v^2z^2$
(2) 2+2 $u'v^2z(1\text{-}z)$
(2) 3+3 $u'v(1\text{-}v)z^2$
(2) 1+2 $u'v(1\text{-}v)z(1\text{-}z)$
(2) 2+1 $u'v(1\text{-}v)z(1\text{-}z)$
(1) 2+2 $u'(1\text{-}v)^2z^2$
(2) 1+1 $u'(1\text{-}v)^2z(1\text{-}z)$
(1) 0+2 $(1\text{-}u'''\text{-}u')v$
(1) 0+1 $(1\text{-}u'''\text{-}u')(1\text{-}v)$

Fig. 4. Components of the set of surviving orthologous pairs in two species diverging at time t_2 after a (shared) WGD at t_1, where one species undergoes a second WGD at t_3. Components ordered vertically according to increased fractionation in the genome with the additional WGD, and ordered horizontally according to increased fractionation in the genome with no additional WGD (dashed lines). Number of cases of same component with different labelling in parentheses. "$x + y$" indicates x pairs dating from the common WGD and y pairs created by the speciation event.

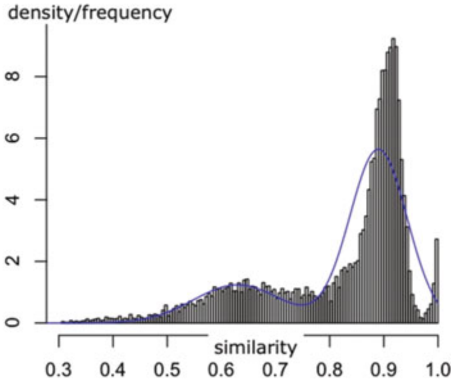

Fig. 5. The distribution of duplicate gene similarities in poplar. The estimation giving rise to the density depicted (in blue) is based on a sample of 500 pairs out of the 13,000 in the histogram. The broadening of the peak representing a recent WGD (at $p = 0.89$) is partially due to the inclusion of highly similar pairs, possibly alleles of the same gene. The small volume of the earlier peak is reflected in very small estimates of u''' and u', less than 0.1, compared to about 0.5 for v.

between lineages, and over time [8]. Though small differences in rates may not be a overriding concern in the study of a sequence of events in a single genome, neglect of these differences may lead to serious errors in speciation-WGD-based phylogenetics [9]. Our approach allows for the introduction of rate variation, while controlling the number of parameters to be estimated.

While differences in divergence rates of gene pairs within and among genomes is relatively well understood, the same is not true of fractionation rates. There are a few quantitative studies of fractionation in the short term [10] and long term [11], but little coherent comparative literature at the whole genome level. There may be great variability of rates consequent to different events, due to the presence or absence of subgenome dominance in allopolyploids versus autopolyploids [12], the combination and timing of composite events giving rise to hexaploidy, e.g., in the Solanaceae ancestor [13], and other factors, making comparisons difficult. Our approach offers a new way of estimating fractionation rates, allowing different rates after different events or in different lineages.

The preceding considerations confirm that the most important direction for further work on this topic will be first, the elaboration of a more parametrized general model enabling the testing of questions about divergence times and rates, fractionation rates, subgenome dominance, and other evolutionary matters. Second will be the construction of a software package capable of more than the *ad hoc* analysis of individually configured data sets. Only then will we be able to do systematic simulation studies as well as comparative studies with more than a few genomes. And only then will we able to confirm the advantage of an evolutionarily principled approach over available general mixture of distributions software.

ACKNOWLEDGMENTS

Research supported in part by grants to DS from the Natural Sciences and Engineering Research Council of Canada. DS holds the Canada Research Chair in Mathematical Genomics.

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 Soft.*, vol. 4, pp. 1–14. 1999.

[2] S. Kumar and S. Subramanian, "Mutation rates in mammalian genomes," in *Proc. Nat. Academy Sci. United States America*, 2002, vol. 99, pp. 803–808.

[3] M. R. McKain, et al., "A phylogenomic assessment of ancient polyploidy and genome evolution across the Poales," *Genome Biol. Evolution*, vol. 8, pp. 1150–1164, 2016.

[4] G. A. Tuskan, et al., "The genome of black cottonwood, *Populus Trichocarpa Sci.*, vol. 313, pp. 1596–604, 2006.

[5] 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.

[6] E. Lyons, et al., "Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar and grape: CoGe with rosids," *Plant Physiology*, vol. 148, pp. 1772–1781, 2008.

[7] D. J. Lipman, A. Souvorov, E. V. Koonin, A. R. Panchenko, and T. A. Tatusova, "The relationship of protein conservation and sequence length," *BMC Evolutionary Biol.*, vol. 2, no. 20, 2002, Art. no. 20.

[8] Y. I. Wolf, P. S. Novichkov, G. P. Karev, E. V. Koonin, and D. J. Lipman, "The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages," in *Proc. Nat. Academy Sci. United States America*, 2009, vol. 106, pp. 7273–7280.

[9] D. Sankoff, C. Zheng, E. Lyons, and H. Tang, "The trees in the peaks," in *Proc. 3rd Int. Conf. Algorithms Comput. Biol.*, 2016, pp. 3–16.

[10] R. J. A. Buggs, et al., "Rapid, repeated, and clustered loss of duplicate genes in allopolyploid plant populations of independent origin," *Current Biol.*, vol. 22, pp. 248–52, 2012.

[11] D. Sankoff, C. Zheng, and Q. Zhu, "The collapse of gene complement following whole genome duplication," *BMC Genomics* vol. 11, 2010, Art. no. 1.

[12] O. Garsmeur, J. C. Schnable, A. Almeida, C. Jourda, A. D'Hont, and M. Freeling, "Two evolutionarily distinct classes of paleopolyploidy," *Molecular Biol. Evolution*, vol. 31, pp. 448–454, 2014.

[13] F. Denoeud, et al., "The coffee genome provides insight into the convergent evolution of caffeine biosynthesis," *Sci.* vol. 345, pp. 1181–1184, 2014.

Yue Zhang received the master's degree in statistics from Carleton University, Ottawa, and is currently working toward the PhD degree in Dr. Sankoff's lab with the University of Ottawa, 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, 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.

David Sankoff received the PhD degree in mathematics from McGill University, 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 with the 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.