

Distribution of gene pair similarity in syntenic regions within
and between genomes: a branching process account of the
polyploidization, speciation and fractionation cycle

Yue Zhang

Thesis submitted in partial fulfillment of the requirements
for the degree of
Doctorate in Philosophy Mathematics and Statistics¹

Department of Mathematics and Statistics
Faculty of Science
University of Ottawa

© Yue Zhang, Ottawa, Canada, 2019

¹The Ph.D. program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics

Abstract

The evolution of plant genomes is notable for manifesting a cycle of whole genome doubling, fractionation (gradual loss of redundant genes) and speciation. The thesis is based on a branching process model of the doubling and fractionation process, integrated with a standard model of sequence divergence. The immediate application of this work is to account for the distribution of sequence similarity for duplicate gene pairs, both within plant genomes and between two related plant genomes in terms of a cycle of polyploidization, fractionation and speciation.

We derive a mixture distribution for duplicate gene pair similarities generated by speciation and/or repeated episodes of polyploidization. We account not only for the timing of these events in terms of local modes or peaks of the component distributions, but also their volume, or amplitude, and variance. We outline how to infer the parameters of the model. We illustrate with analyses of the distribution of homolog similarities in a number of plant families: Brassicaceae, Solanaceae and Malvaceae.

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.

In addition, we solve the problem of identifying the ploidy level of a series of two

or three polyploidizations by invoking the observed and predicted gene triple profiles for each model, i.e., by calculating the probability of the four types of triple with origins in one or the other event, or both.

Résumé

Nous dérivons une distribution mixte pour l'identité de séquence de paires de gènes en double générées par la spéciation et / ou des épisodes de polyploïdisation. Nous prenons en compte non seulement la chronologie de ces événements en termes de modes locaux ou de pics des distributions composantes, mais également leur volume (ou amplitude) et leur variance. Nous décrivons comment estimer les paramètres du modèle. Nous illustrons par des analyses de la distribution d'identité d'homologues dans quelques familles de plantes: Brassicaceae, Solanaceae et Malvaceae.

À notre connaissance, il s'agit de la première méthode qui prend en compte le volume des composantes d'une distribution d'identités, préalable à une procédure d'inférence bien motivée du point de vue de l'évolution.

De plus, nous résolvons le problème d'identification du niveau de ploïdie d'une série de deux ou trois polyploïdisations en invoquant les profils de triplets de gènes observés et prédits pour chaque modèle, c'est-à-dire en calculant la probabilité des quatre types de triplets ayant pour origine l'un ou l'autre événement, ou tous les deux.

Acknowledgement

I would like to express my sincere gratitude to David Sankoff for the confidence in me he showed four years ago in undertaking to supervise my Ph.D. study and research, and for his patient guidance, enthusiastic encouragement and immense knowledge. This thesis would not have been possible without his inspired suggestions of research directions. My work would not have been nearly as fulfilling without the appreciation he frequently expressed for my mathematical results. I have been very fortunate to work alongside some wonderful colleagues in the Sankoff lab over the years, many of whom have contributed directly and indirectly to my thesis. Thanks to Daniella Santos Munoz, Eric Lam, Joao Meidanis, Madisyn Gabrielle Turcotte, Manuel Lafond, Mona Meghdari, Sindeed Islam, Yuji Jeong and Zhe Yu. I am particularly grateful for the assistance given by Chunfang Zheng.

Special thanks go to the Department of Mathematics and Statistics of the University of Ottawa for giving me the freedom to specialize in bioinformatics and to focus on this area, and to Aïda Ouangraoua of the University of Sherbrooke, Jason Gao of Carleton University, Kelly Burkett and Mayer Alvo of the University of Ottawa for agreeing to serve on my dissertation committee, to examine this thesis and for their advice.

Last but not least, I must express my very profound gratitude to my parents

and my younger sister, whose love and guidance are with me in whatever I pursue. I would like to express my deepest gratitude to my boyfriend, Sicong Ma, for his unstinting encouragement, unfailing support throughout my years of study and his patience with the time and effort I consecrated to my research. Thank you.

Dedication

For my mother, father and sister.

Contents

List of Figures	ix
List of Tables	xii
1 Introduction	1
2 Models for similarity distributions of syntenic homologs and applications to phylogenomics	14
3 Pinning down ploidy in paleopolyploid plants	26
4 A branching process for homology distribution based inference of polyploidy, speciation and loss	38
5 Branching out to speciation in a model of fractionation: the Malvaceae	59
6 Distinguishing successive ancient polyploidy levels based on genome-internal syntenic alignment	71
7 Conclusions	84

List of Figures

1.1	Origin of orthologs and paralogs in speciation and WGD. . . .	5
1.2	Gene loss from homologous chromosomes.	6
1.3	Similarity Distribution.	7
2.1	Components of the number of surviving pairs created by WGDs at t_1 and t_2	17
2.2	Components of the number of surviving pairs created by WGDs at t_1, t_2 and t_3	18
2.3	Components of the number of surviving pairs created by a whole genome tripling at t_1 and a WGD at t_2	18
2.4	The distribution of duplicate gene similarities in poplar	19
2.5	Ortholog similarities in the <i>B. rapa</i> – <i>B. oleracea</i> comparison .	19
2.6	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	20
2.7	Phylogenetic relationship of six species in the family Brassicaceae	22
2.8	Gene similarity distribution between 15 pairs of genomes in the Brassicaceae	22

LIST OF FIGURES

2.9	Brassicales phylogeny derived from the literature	23
2.10	Phylogeny calculated by proposed algorithm	23
3.1	Event with ploidy $r_i = 4$	29
3.2	Counting t_i -pairs	29
3.3	Evolutionary history of the Brassicales	31
3.4	Duplicate gene similarities in syntenic blocks in the <i>Brassica rapa</i> genome self-comparison.	31
3.5	All paralog trees generated by a genome duplication at time t_1 followed by a triplication at time t_2 , both events followed by fractionation	32
3.6	All paralog trees generated by a genome triplication at time t_1 followed by a duplication at time t_2 , both events followed by fractionation.	33
3.7	(u, v, w) space with MLE solutions for (x, y, z) model	34
3.8	(s, w) surface	35
3.9	Regions of the 10 <i>Brassica rapa</i> chromosomes coloured by 14 triples of recently diverged regions	36
4.1	A gene tree produced by two triplings at generations 1 and 2, followed by a speciation at generation 3	47
4.2	Phylogenetic relationships among the Solanaceae	50
4.3	Distribution of ortholog similarities	52
4.4	Distribution of paralog similarities	53
4.5	Similarity of orthologs as a function of speciation time; Relation of standard deviation to component mean	55

LIST OF FIGURES

4.6	Estimating amplification factor due to <i>Solanum</i> triplication . .	56
5.1	Phylogenetic relationships among the Malvaceae	63
5.2	<i>Aquilaria</i> and <i>Hibiscus</i> similarities	64
5.3	<i>Gossypium</i> and <i>Durio</i> similarities	65
5.4	Grewioideae and Byttnerioideae similarities	65
5.5	Trends of peak modes and standard deviation	68
5.6	<i>Durio</i> synteny blocks showing clear pattern of recent triplication	69
6.1	Distribution of gene pair similarity in the <i>Durio zibethinus</i> genom	73
6.2	Sample trajectory	74
6.3	Distribution of syntenic gene pair similarities in <i>Populus tri-</i> <i>chocarpa</i>	78
6.4	Distribution of syntenic gene pair similarities in <i>Brassica oleracea</i>	79
6.5	Distribution of syntenic gene pair similarities in <i>Brassica ol-</i> <i>eracea</i> decomposed into three events	81

List of Tables

2.1	Peak similarity level, by genome	23
4.1	Characteristics inferred for speciation event distributions. . . .	54
4.2	Estimates of survival (left) and of fractionation rates (right). .	56
5.1	Statistical parameters of the speciation peaks.	66
5.2	Ranges of fractionation rates preceding speciation	67
5.3	Ranges of fractionation rates following WGD	67
6.1	Formulae for the expected numbers of triplets	75
6.2	Durian model predictions	77
6.3	<i>Populus</i> model predictions	78
6.4	<i>Brassica oleracea</i> model predictions	79
6.5	Formulae for the expected numbers of triplets after three events.	80
6.6	<i>Brassica oleracea</i> three-event model predictions	82

Preface

This dissertation is submitted for the degree of Doctor of Philosophy at the University of Ottawa. The research described herein was conducted under the supervision of David Sankoff, professor in the Department of Mathematics, University of Ottawa, between July 2015 and September 2019.

The original motivation for this research was the rampant misuse of the distribution of duplicate gene similarities in genomics to make unjustifiable claims about evolutionary history, especially in the comparative genomics of flowering plants. At the beginning the goal was simply to work out a few specific instances of models to predict a distribution, but eventually it involved formulating a full-fledged general model, realizing its limitations and moving beyond. The order of the chapters reflects this development in my understanding of the problems. Similarly, the earlier work speaks of the models in terms of a birth-and-death process, though I call on no actual properties of such processes. The later work more appropriately characterizes the models in terms of branching processes, since it is eminently reasonable to speak about siblings, and gene pairs, in terms of a branching process, whereas this has no meaning in classical birth-and-death.

This work is to the best of my knowledge original, except where acknowledgements and references are made to previous work. Neither this, nor any substantially

similar dissertation has been or is being submitted for any other degree, diploma or other qualification at any other university.

Chapter 1

Introduction

A genome contains all the genes defining a species, and is almost identical from organism to organism, the slight differences accounting for individual genetic variation within a population as well as genetic distinctions between different populations of the same species. Within each individual, the same genome resides in the DNA of almost all its cells.

The genomes of two different species contain different sets of genes, but for related genomes, nearly every gene in one genome has a similar counterpart in the other. The degree of similarity, as measured by comparison of their DNA sequences, depends in a stochastic way on the recency of the last common ancestor of the two species. Indeed, an average, such as mean or mode, of the distribution of similarities of such pairs of “duplicate” genes is often used as an index of the time of speciation, i.e., the moment the two emergent species ceased to admit gene flow between them.

In some phylogenetic domains, such as the yeast or the fish or, especially, the plants, the genome of species can be seen to have irrevocably doubled or tripled some time ago, through a process called polyploidization. This event was usually sudden,

1. INTRODUCTION

in terms of evolutionary time, i.e., within several generations, after millions or tens of millions of years of reproducing as a single-copy genome. This “whole genome doubling”, or WGD, also results in a distribution of similarities of pairs of genes *within* the genome of each species descending from the doubling event.

The statistical distribution of “duplicate” gene similarities, both within a genome and between two related genomes, is an important construct for researchers in comparative genomics, and appears in many papers reporting the first genome sequence of a species. This distribution, in plants especially, is characterized by multiple modes, or “peaks”, each deriving from a WGD event or a speciation event, with the location of the peak on the similarity axis indicating the age of the event. It is understood that the relative volumes (or areas or amplitudes), of the regions of the distribution supporting each peak is a function of the rate of duplicate gene loss, a universal process, but no formal account of this has been given in the literature. *The main theme of this thesis is the mathematical modelling, for the first time, of the cycle of WGD (or WGT: whole genome tripling,...), gene loss (called fractionation in the case of within-genome gene pairs) and speciation, giving rise to a distribution of gene similarities and allowing data-based inference of the important dynamic factors characterizing these processes.*

The body of this thesis consists of five publications, reflecting the development of the theory as illustrated by appropriate genomic data. Before discussing each of these contributions, it is necessary to establish some biological preliminaries, to discuss data sources and to review some mathematical concepts.

1. INTRODUCTION

1.1 Biological background

The contribution of this thesis is to the field known as “Comparative genomics”, which these days provides the main methodology and scientific support to evolutionary theory, but is also important for oncology and other medical research, agriculture, forestry, fishery and many fields of biology. Comparative genomics is largely based on probability modeling, statistics, data science, algorithmics and other types of mathematics, and deals with data sets that can be very large.

Comparative genomics studies similarities and differences in proteins, DNA (including genes) and other structures or quantities that differ among species. The idea is to explore evolutionary relationships between organisms, as assessed by any number of quantitative measures. We focus here on a set of interrelated genomic processes underlying the familial relationship among species, namely speciation, WGD, WGT, etc., and gene loss, particularly “fractionation”, which we shall introduce in the following paragraphs. We study these in the context of the angiosperms, or flowering plants, whose evolution evidences a widespread historical pattern of WGD. We make use of public data sources, as described in a later section. As mentioned earlier, the main statistical data we analyze are the distributions of gene similarities, a widespread but poorly understood compilation derived from annotated genomic sequences, i.e., sequences in which all or most of the genes have been identified.

To understand these distributions, we introduce a branching process model for analyzing the cycle of polyploidization processes underlying WGD, WGT, etc., and the subsequent loss of genes known as fractionation. Some of the questions we introduce cannot be addressed using gene pairs only, so we introduce new methods for the study of gene triples, allowing us to infer some important properties of polyploidization.

1. INTRODUCTION

Our theoretical results are tested throughout using angiosperm genome data, producing results that are of considerable biological interest.

1.2 Polyploidization and Whole Genome Doubling

The evolutionary history of the flowering plants, or “angiosperms”, is characterized by a continuous diversification of newly emerged species into all terrestrial ecological niches, especially over the last 145 million years, so that they displaced other widespread seed plants, such as the conifers, as the predominant group of land plants over the last 60-100 million years. This explosion of the number of different flowering plants has been mediated by hundreds of thousand of speciation events leading to around 300,000 extant species.

Besides speciation, the “family tree” or “phylogeny” of the flowering plants is punctuated with numerous WGD and WGT events, whose subsequent evolution often includes speciation into multiple distinct lineages. The biological processes responsible for the appearance of these events is polyploidization usually as a result of an error in meiosis. This produces duplicate or triplicate (“homeologous”) copies of all the chromosomes in the genome, containing exactly the same genes in the same order.

Thus both speciation and WGD produce duplicate genes for all the genes in a genome. Two genes are “homologous” if they trace their descent from a single gene in an ancestral genome, and entities related by homology are called homologs. Orthologs are homologous genes that evolved separately in two species after a speciation event. Paralogs are homologous genes that evolved within a single species, for example after a WGD event. This is represented in Figure 1.1.

In both speciation and WGD, the two genes in each pair evolve independently

1. INTRODUCTION

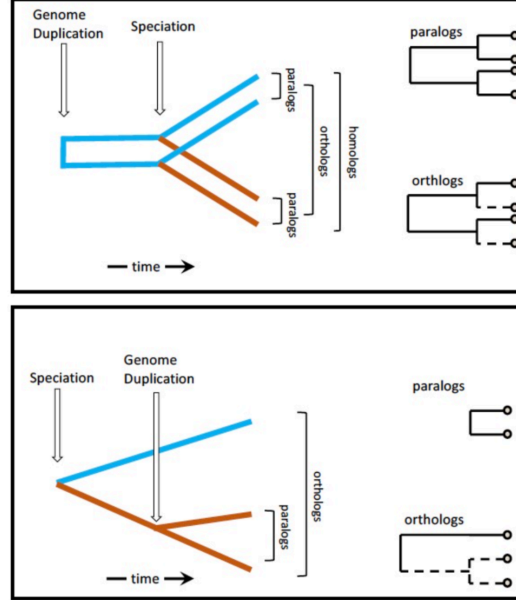


Figure 1.1: Homologs. The origin of orthologs in speciation and the origin of paralogs in WGD. Upper example represents a WGD event followed by a speciation. Red and blue lines indicate two species, also captured by solid and dotted lines in the schema on the right. The lower example represents a speciation event followed by a WGD

according to a dynamic of decaying similarity as gene sequence and amino acid sequences inexorably diverge through random single nucleotide mutation. Some genes may be lost completely. After speciation, shifting ecological conditions or reorganization of cellular networks for one or both daughter genomes, may enable the loss of genes that are no longer required. For paralogous pairs we call the loss process “fractionation”, whereby one of the two genes (or one or two of the three genes in the case of WGT) is excised, pseudogenized or otherwise removed as a recognizable coding gene, as in Figure 1.2.

The cycle of fractionation after polyploidization is the focus of all the mathematical modeling and the statistical analysis in this thesis.

1. INTRODUCTION

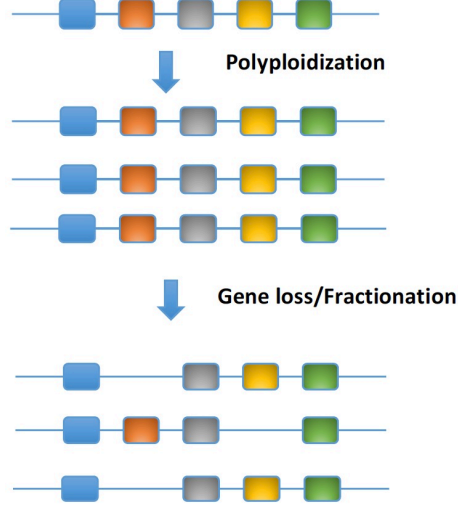


Figure 1.2: Gene loss from homologous chromosomes after speciation or from homeologous chromosomes, by fractionation, after WGD/WGT.

1.3 Mixture Distribution

A widespread practice in comparative genomics is to infer the nature and timing of evolutionary events through the examination of the distribution of similarities in homologs, as in Figure 1.3. This is done routinely in comparative genomics by finding peaks, or local modes, in order to estimate the event's age. The event time t_i can be inferred by visualizing or otherwise inferring local modes or peaks in the distribution, and identifying the doublings around these points as having been generated by speciation or WGD events. In comparing two species, there are peaks at times corresponding to all their shared polyploidization, followed by a single peak dating from their speciation event, but no further peaks. Additional polyploidizations after speciation increase the number of orthologs and paralogs in a uniform way across the board, but do not change the relative size of peaks.

The distribution of similarities is then a mixture of distributions, each of which

1. INTRODUCTION

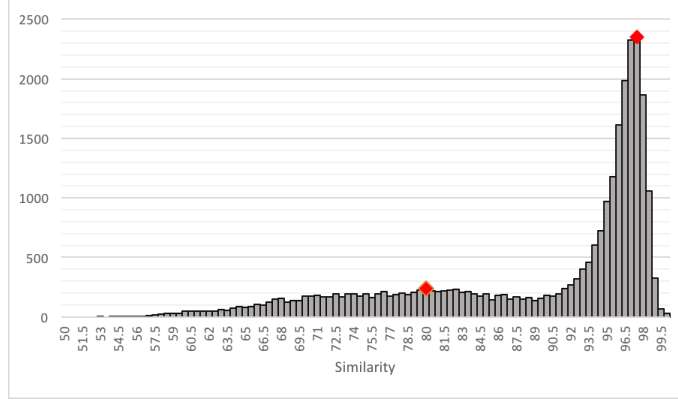


Figure 1.3: Similarity Distribution

is centered at a similarity value indicative of the age of one of the polyploidizations or speciation. We develop models for predicting the shape of these distributions and fractionation of duplicate genes based on the event times.

In this thesis, 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. The latter model produces a homolog tree in the form of a branching process, which is mathematically tractable and whose parameters are well suited to statistical inference. 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.

1. INTRODUCTION

1.4 Branching Process

One way of looking at the proliferation and loss genes in a genome is as a population of individuals evolving in a stochastic way. Of the different types of model: birth-and-death processes, branching processes, etc., the closest resemblance to the biological phenomena of speciation, WGD and gene loss, including fractionation, is a discrete time branching process. The key property of a branching process is that it allows us to keep tracks of pairs of “sibling” genes, essential for studying the distribution of gene pair similarities. We review the classical discrete-time branching process and comment on how applicable it is for generating populations of paralogs. We also discuss the possibilities and limitations of statistical inference of the parameters of the model.

1.5 Data

The CoGe (Comparative Genomics) platform is a software suite of interlinked and interconnected web-based tools, together with a very large database of genome sequences and annotations in a standardized format, for easily visualizing and comparing. For pairs of genomes for which coding sequences (including genes) are known, we mainly use the SYNMAP program in CoGe. The basic kind of data for this thesis is the similarity between the coding regions of two homologous genes - genes that originated, some time in the past from the same ancestral gene. The coding DNA sequence of these two genes can be aligned automatically according to various optimality criteria so that each aligned region in the two reflects a common ancestral region. For example, part of the two aligned genes might be

CCGACGTTACGG

1. INTRODUCTION

CGGAGGTAAAGC

The similarity between the two will be the proportion of positions where the sequences are identical, 58 % in this example. SYNMAP searches for syntenic blocks sets of five or more successive genes on two chromosomes in the two genomes that are consecutive and homologous, with allowances made for the occasional missing gene. Among the various details of each pair of genes, SYNMAP reports their similarity.

The set of all similarities over all the pairs of homologous genes in two genomes can be compiled from the output of SYNMAP. This defines the distribution of gene pair similarities; whose analysis is the main point of this thesis. We do not use any other kind of data, with the exception of the phylogenetic tree of group of organisms, drawn from the literature.

The two genomes may be distinct, producing pairs of orthologs, or we can compare a genome with itself, producing pairs of paralogs. Focusing on syntenic, or colinear, sets of duplicate genes ensures that these pairs are all produced during the same WGD or speciation event. This excludes so-called “tandem pairs” and isolated pairs out of syntenic context.

The analyses we present consist of combinatorial models, including counting procedures, equation solving and statistical inference, most of which are freely available on the R platform, although the counting sometimes requires a small amount of coding specific to individual problems. For this reason, we felt it would be redundant to assemble all of these in a unified user-friendly package; the diverse set of analytic goals would require a complex set of options, defeating the purpose of such a package.

1. INTRODUCTION

1.6 The Genomes

The genomes used in this study all belong to the core eudicots, a group of hundreds of thousands of species, with a common seven-chromosome ancestor that underwent WGT, called γ , about 120 million years ago.

1.6.1 Poplar

Populus trichocarpa has undergone WGD, following the γ whole genome tripling. This WGD, called the “salicoid”, is shared with the willow and possibly other members of the Salicaceae family. We used poplar in some of the initial stages of constructing our model, as well as to test our theory of triplets in the final chapter.

1.6.2 The Brassicaceae

This is an economically important family of plants scattered across almost the entire land surface of the planet. Many of these are analyzed in the next chapter. The genus BRASSICA has undergone not only γ , but also the so-called β and α WGD, followed by a WGT. *Brassica rapa* and *Brassica oleracea* are used to analyse the problem of determining the ploidy levels of two successive polyploidizations, using algebraic methods, and later, using statistical methods.

1.6.3 The Malvaceae

The Malvaceae are a family of plants in the rosid order Malvales. The family bifurcated early into two lineages, giving rise to the subfamily Malvoideae and to the three sister subfamilies Grewioideae, Helicteroideae and Byttnerioideae. Our Malvoideae

1. INTRODUCTION

genomes reveal repeated instances of WGD, the Grewioideae and Helicteroideae only on each, and the Byttnerioideae none.

1.6.4 The Solanaceae

The Solanaceae contains most of the economically important genera that are contained in the sub-family Solanoideae (such as: tomato, potato, tobacco, etc).

1.7 The Objectives of the chapters

This thesis is based on five papers that are published, in press or, in one case, recently submitted. An earlier publication [IEEE/ACM Transactions in Computational Biology and Bioinformatics, 1545-59 63, 2017] is not included since a corrected version forms part of the multi-author paper included here as Chapter 2. In the remaining four of these papers I am the first author. My supervisor is the senior (last) author and Dr.Chunfang Zheng, a long time associate of Dr.Sankoff's is a co-author. Dr.Zheng's role was largely in discussing research plans with Dr.Sankoff, and help with data input to CoGe, except in the first paper, where she was also responsible for the phylogenomic algorithm. Other co-authors were only involved in discussions and advice, but did not take part in the mathematical developments, nor the data analysis.

Because each paper was originally a self-contained publication, there is some overlap and redundancy among their introductory and background sections, but these are not too lengthy and each paper provides a quick review of what has been accomplished in the previous chapters. Indeed the temporal sequence of the papers is complicated by the fact that they were all conference papers before most of them were revised and

1. INTRODUCTION

accepted for journals, so that the revisions contained some improvements only developed for later papers. Fortunately, many of the papers share common terminology and notation, contributing to the coherence of the thesis as a whole.

Chapter 2 contains the first exposition of several instances of the problem, laying out a straightforward approach to their solutions. This was basically a combinatorial analysis, which has proved to be correct and opened the door to all the further work in this thesis. It did incorporate, however, an overly restrictive model of sequence divergence, especially as regards the variance estimates for cohorts of gene pairs.

Chapter 3 is the first to give a general expression for the branching process model for WGD, WGT and fractionation, although this is still referred to as a birth-and-death model, not the most precise designation. This paper tackles the question of determining the ploidy of an event, or a succession of two events, and showed by calculating the intersection of algebraic varieties defined by competing models that the pairs similarity data usually could not distinguish between the models. Additional data, such as numbers of unpaired genes, could help, but these data are generally much less reliable than pairs. The paper also suggests studying the number of similar triples originating from the two events, and carries out a rough calculation to distinguish between the two models.

Chapters 4 and 5 review the details of the branching process model for generating populations of paralogs, as summarized from Chapter 3 but derive simpler expressions (no summations, no factorials) and extend the model to introduce speciation to derive the expected number of orthologous pairs with most recent common ancestors at each ancestral time. Chapter 5 completes the development of the combined WGD and speciation model with a particular attention to the fractionation regimes immediately before and after speciation, as well as a formal demonstration of the fact that

1. INTRODUCTION

no peaks due to fractionation show up after the speciation peak in the comparison of two genomes. The methods are applied to the evolution of six genomes from the plant family Solanaceae (Chapter 4) and ten genomes from the family Malvaceae (Chapter 5). Of particular interest is the systematic inference of fractionation rates made possible for the first time by our models, and the fact that these rates in the woody, perennial Malvaceae species are approximately half of those in the annual, herbaceous Solanaceae plants. These two chapters represent the culmination of both our development of the branching process model and its application to produce original biological results.

In Chapter 6 we set out a definitive triples-based framework for establishing the ploidies of a series of polyploidization events - are they WGD, WGT or other? This kind of approach, as well as reasonable but not model-based solutions, appear in Chapter 3 as well as other papers from the Sankoff lab[[bioRxiv654285](#)], but here we develop, for the first time, a model-based (branching process) approach to predicting the number of triples according to their three pairwise similarities. Profiles of triple frequency are derived for two and for three successive events, for all combinations of WGD and WGT, and are successfully applied to three genomes.

1.8 The Objectives of the thesis

The primary goals of this thesis are to account for the distribution of gene pair similarities through a mathematical model and to apply this to derive results of biological interest, such as fractionation rates, on real plant genomes.

Chapter 2

Models for Similarity Distributions of Syntenic Homologs and Applications to Phylogenomics

This chapter contains the first exposition of several instances of the problem, laying out a straightforward approach to their solutions. This is basically a combinatorial analysis, which has proved to be correct and opened the door to all the further work in this thesis. It does incorporate, however, an overly restrictive model of sequence divergence, especially as regards the variance estimates for cohorts of gene pairs.

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

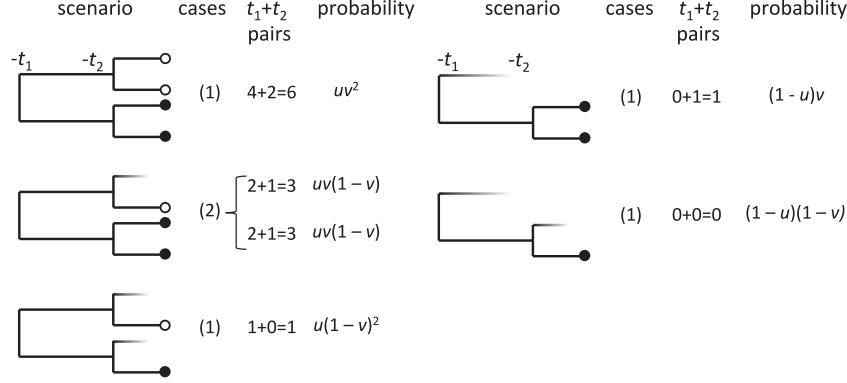


Fig. 1. Components of the number of surviving pairs created by WGDs at t_1 and t_2 . Fading lines illustrate gene loss, but play no role in the calculation of the number of cases.

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 .

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

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

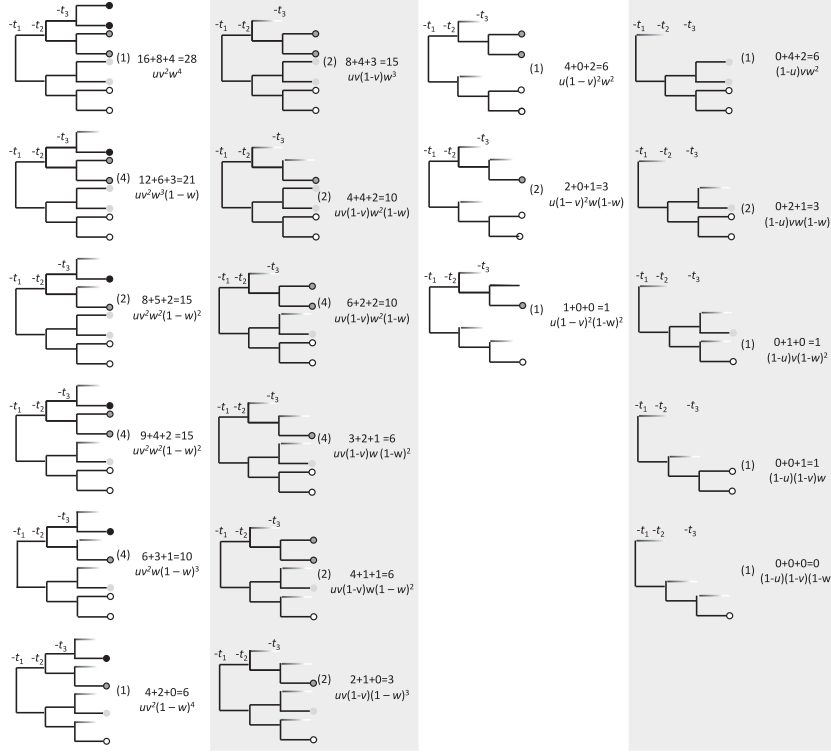


Fig. 2. Components of the number of surviving pairs created by WGDs at t_1 , t_2 , and t_3 . See Fig. 1 to interpret the content of the various columns.

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

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

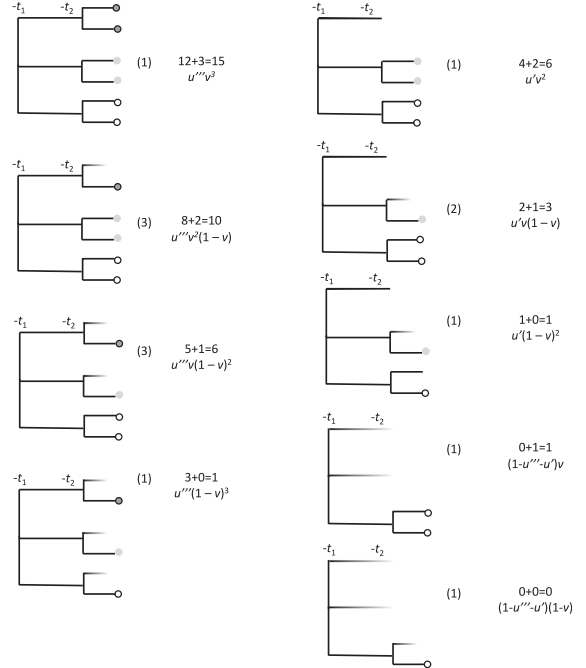


Fig. 3. Components of the number of surviving pairs created by a whole genome tripling at t_1 and a WGD at t_2 .

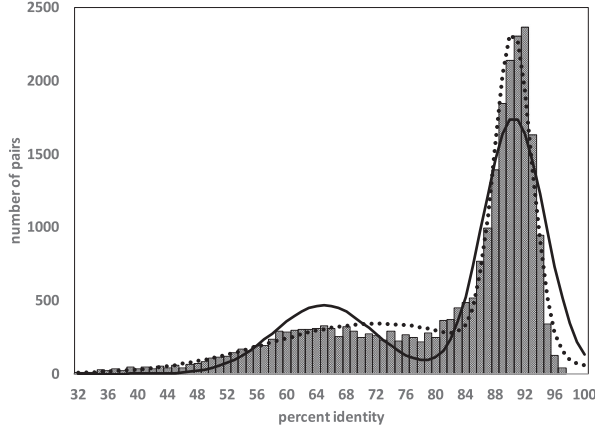


Fig. 4. The distribution of duplicate gene similarities in poplar. Some pairs with around 99 percent similarity were excluded, since they represent slightly divergent alleles of the same gene, not two distinct genes. Histogram: empirical data. Solid curve: fit of model with estimated fractionation rates.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

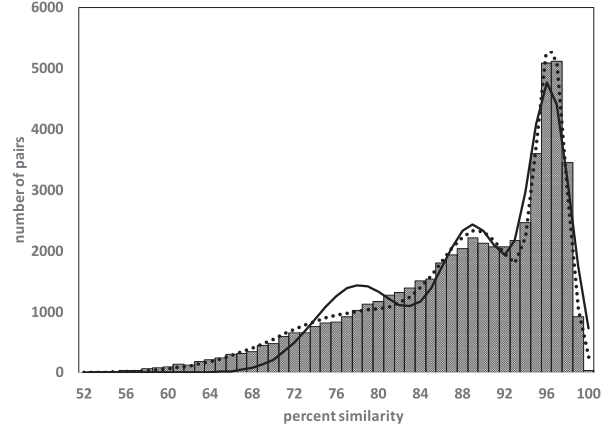


Fig. 5. Ortholog similarities in the *B. rapa* – *B. oleracea* comparison. Histogram: empirical data. Solid curve: fit of model with estimated fractionation rates. Dotted line curve: EMMIX fit.

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

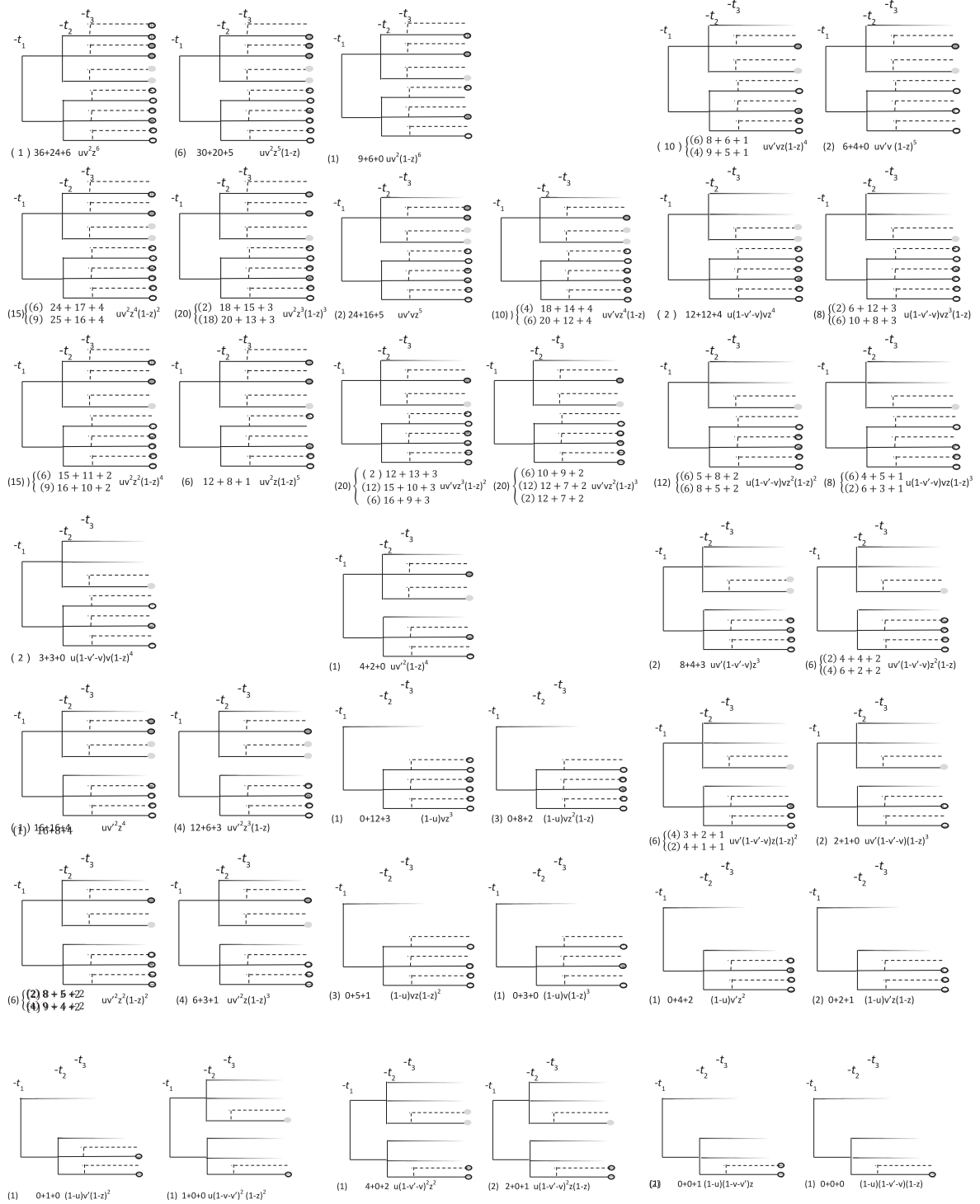


Fig. 6. Components of the set of surviving orthologous pairs in two species diverging at time t_3 after a (shared) WGD at t_1 , and a shared WGT at t_2 . Number of cases of same component with different labelling in parentheses. $^a w + x + y$ indicates w pairs dating from the shared WGD, x pairs created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

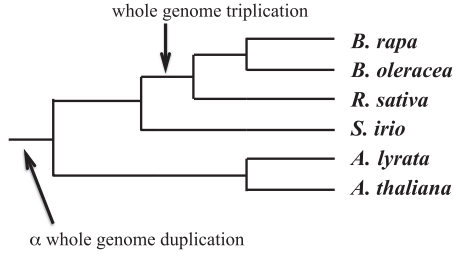


Fig. 7. Phylogenetic relationship of six species in the family Brassicaceae, showing lineages affected by WGD and WGT events.

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 spikerush) [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]).

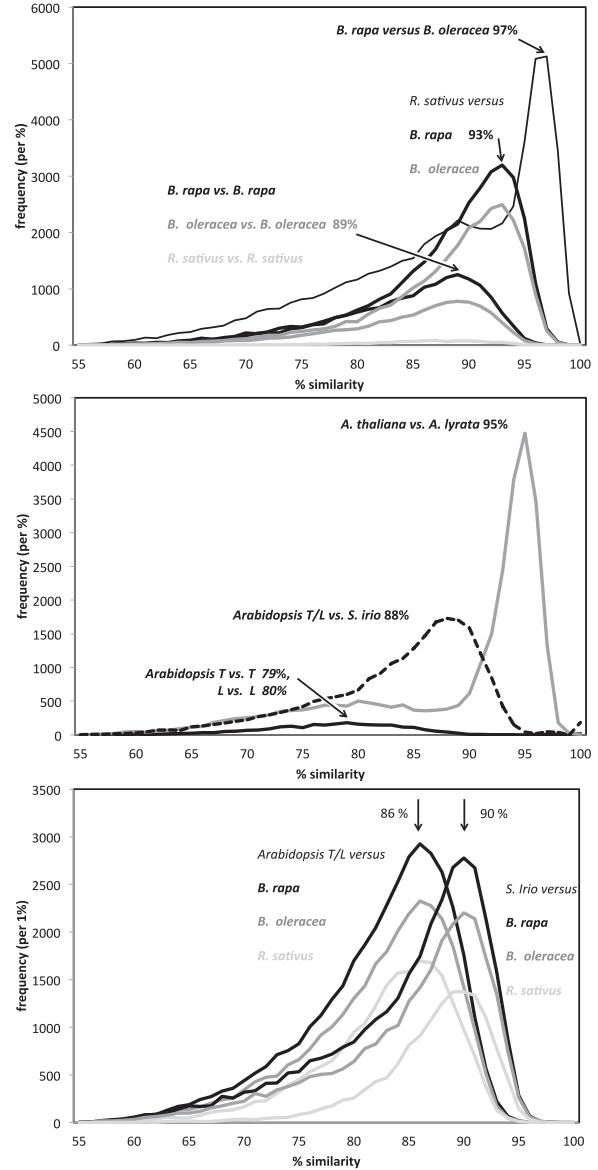


Fig. 8. Gene similarity distribution between 15 pairs of genomes in the Brassicaceae and five self comparisons. Local modes ("peaks") are indicated. Only one of each comparison is shown for *Arabidopsis*, the other is superimposed and indistinguishable.

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

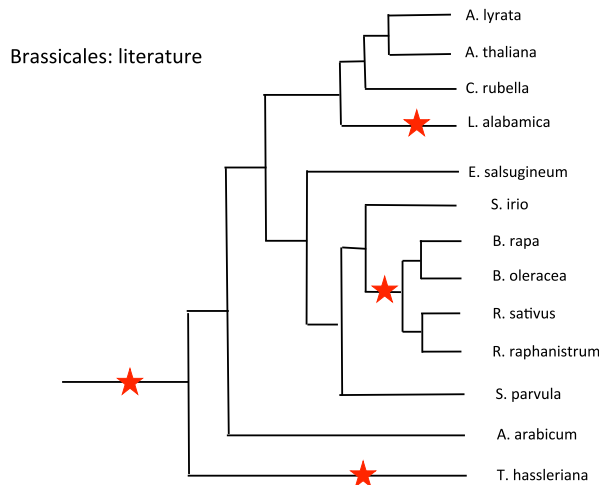


Fig. 9. Brassicales phylogeny derived from the literature.

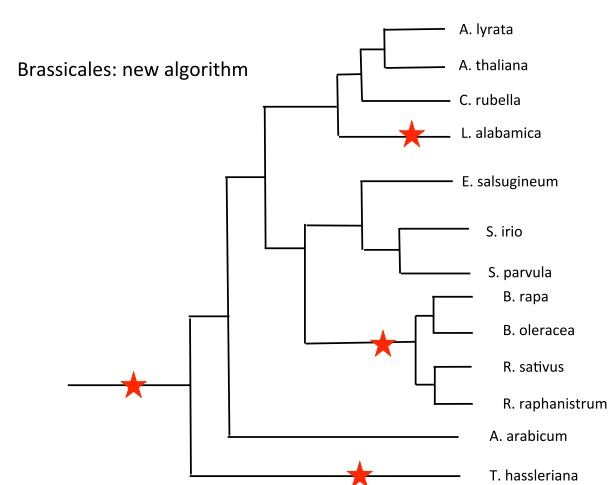


Fig. 10. Phylogeny calculated by proposed algorithm.

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 synteny 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, S. 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. Poullain, 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. Ruperio, F. Cheng, N. E. Waminal, Y. Huang, C. Dong, L. Wang, J. Li, Z. Hu, M. Zhang, 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. Shiu, "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. Ottilar, 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 salicgineum*," *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.

Chapter 3

Pinning down ploidy in paleopolyploid plants

This chapter is the first to give a general expression for the branching process model for WGD, WGT and fractionation, although this is still referred to as a birth-and-death model, not the most precise designation. This chapter tackles the question of determining the ploidy of an event, or a succession of two events, and shows by calculating the intersection of algebraic varieties defined by competing models that the pairs similarity data usually cannot distinguish between the models. Additional data, such as numbers of unpaired genes, can help, but these data are generally much less reliable than pairs. This chapter also suggests studying the number of similar triples originating from the two events, and carries out a rough calculation to distinguish between the two models. This problem is taken up again in Chapter 6 in a much more rigorous way.

RESEARCH

Open Access



Pinning down ploidy in paleopolyploid plants

Yue Zhang, Chunfang Zheng and David Sankoff*

From RECOMB-CG - 2017 : The Fifteenth RECOMB Comparative Genomics Satellite Conference
Barcelona, Spain. 04-06 October 2017

Abstract

Background: Fractionation is the genome-wide process of losing one gene per duplicate pair following whole genome multiplication (doubling, tripling, . . .). This is important in the evolution of plants over tens of millions of years, because of their repeated cycles of genome multiplication and fractionation. One type of evidence in the study of these processes is the frequency distribution of similarities between the two genes, over all the duplicate pairs in the genome.

Results: We study modeling and inference problems around the processes of fractionation and whole genome multiplication focusing first on the frequency distribution of similarities of duplicate pairs in the genome. Our birth-and-death model accounts for repeated duplication, triplication or other multiplication events, as well as fractionation rates among multiple progeny of a single gene specific to each event. It also has a biologically and combinatorially well-motivated way of handling the tendency for at least one sibling to survive fractionation. The method settles previously unexplored questions about the expected number of gene pairs tracing their ancestry back to each multiplication event. We exemplify the algebraic concepts inherent in our models and on *Brassica rapa*, whose evolutionary history is well-known. We demonstrate the quantitative analysis of high-similarity gene pairs and triples to confirm the known ploidies of events in the lineage of *B. rapa*.

Conclusions: Our birth-and-death model accounts for the similarity distribution of paralogs in terms of multiple rounds of whole genome multiplication and fractionation. An analysis of high-similarity gene triples confirms the recent *Brassica* triplication.

Keywords: Whole genome duplication, Gene loss, Birth and death process, Multinomial model, Paralog gene tree, Sequence divergence, *Brassica rapa*

Background

While the genomic multiplicity of recent polyploids is accessible through cytogenetics and other methodologies, the nature of early large-scale genome events like auto- and allopolyploidization is obscured by interchromosomal translocations, chromosome fusions and other chromosomal rearrangements, by gene family expansions and fractionated gene loss and by sequence divergence between paralogs. One valuable line of evidence about these ancestral events is the discovery of two sets of at

least four or five collinear pairs of highly related genes – paralogs – in close succession in different regions of the genome, known as paralogous syntenic blocks. Insofar as all the paralog pairs in a paralogous syntenic block resemble each other to the same extent, this indicates that there was a duplication of the chromosomal region containing them, which can then be dated approximately according to the degree of DNA sequence divergence. If there are many syntenic blocks of the same age throughout the genome, this is suggestive of a whole genome duplication at that point in time.

In 2007 Jaillon et al. noted that syntenic regions in the genome of grape (*Vitis vinifera*) were distributed as triples, not just duplicates [1]. Much of the genome could

*Correspondence: sankoff@uottawa.ca

Department of Mathematics and Statistics, University of Ottawa, 585 King Edward, Ottawa K1N 6N5, Canada



© The Author(s). 2018 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

be partitioned into seven sets of three syntenic regions, indicative of a whole genome triplication over 100 M years ago, producing a 21-chromosome grape ancestor from a 7-chromosome precursor. Of interest is that in each triplet of regions, forming three pairs of regions, there were many duplicate gene pairs – involving just two regions – but very few actual triples of three highly related genes, one in each region. In addition there were very few duplications within a single region, or between genes in two different triplets among the seven sets of grape triplets of regions. The 21-chromosome construct has since been widely recognized as the ancestor of the core eudicots. The principle of three-way similarities among syntenic regions, indicated by

- some duplicated pairs between each two of the three regions,
- with or without any triplicated genes,
- no pairs within a single region and
- no pairs between different triples of regions,

is the signature pattern for ancient whole genome triplication, or paleohexaploidy. This may be generalized in straightforward ways to octoploidy and higher multiplicities of polyploidization. For example, an ancient octoploidization would be reflected in 4-tuples of regions, where there would be some duplicated gene pairs between each of the $\binom{4}{2} = 6$ pairs of regions, but no gene pairs within regions and no gene pairs between different 4-tuples.

Another type of important evidence in analyzing ancient polyploidization events is the distribution of coding sequence similarities between two paralogous genes. All flowering plants, and indeed most land plants, have at least one, and generally two, three or more polyploidizations in their history. The distribution of similarities is then a mixture of distributions, each of which is centered at a similarity value indicative of the age of one of the polyploidizations. 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 [2]. This model produces a *paralog tree* in the form of a birth and death process with one biologically-motivated constraint, which remains mathematically tractable and whose parameters are well suited to statistical inference. Because of a trade-off between ploidy and fractionation rates, however, in many instances the multiplicity of the various ploidy events in the evolution of a genome cannot be determined uniquely, which is a severe problem for understanding its history.

One goal of this paper is to remedy this shortcoming by combining the syntenic approach pioneered in [1] with the

paralog tree model of [2] to produce a method capable of estimating the multiplicity of the polyploidization events, as well as the fractionation parameters.

The next section summarizes the general model for generating the distribution of paralog similarities. This is followed by a brief section describing the inference of the parameters. We then focus on two particular instances, one where a hexaploidization (whole genome triplication) precedes a tetraploidization (whole genome duplication), and the other where the triplication follows the duplication. The difficulty of ploidy inference is illustrated with data from the turnip, or Napa cabbage (*Brassica rapa*) genome, and investigated in algebraic detail. In a section entitled “Counting triples”, we introduce the method inspired by [1] for distinguishing whole genome triplication from whole genome duplication, given the distribution of duplicate gene similarity, and we apply this to confirm the known sequence of events in the ancestral history of this species.

Methods

The general model

We summarize and correct a new and general model [3] for the repeated cycle of polyploidization events, each followed by fractionation. This model allows an arbitrary number of events and rates of fractionation of the progeny of any gene holding across the entire genome after each event. From this we calculate expected numbers of duplicate gene pairs, at the time of observation (i.e., the present time), originating at each of the historical polyploidization events, leading to the prediction of the entire distribution of similarities, using standard models of mutational processes.

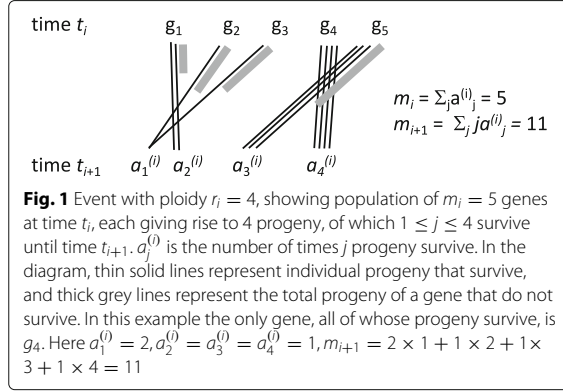
The model is a continuous-time birth-and-death process with the entire population synchronized as to birth times and number of progeny, but with the number of deaths of the siblings in each individual “litter” determined probabilistically.

The birth-and-death process

The process starts with $m_1 \geq 1$ genes at time t_1 ; at times $t_1 < \dots < t_{n-1}$ for some $n \geq 1$, each existing gene is replaced by $r_1, \dots, r_{n-1}$ progeny, respectively, where each $r_i \geq 2$. As illustrated in Fig. 1, for each gene's progeny, at least one and at most r_i genes survive until time t_{i+1} , as governed by a probability distribution $u_1^{(i)}, \dots, u_{r_i}^{(i)}$.

The results are observed at time t_n , namely a measure of similarity (e.g., coding sequence similarity) between all pairs of genes in the population, where the m_1 original genes are considered to be unrelated or too remotely related to be considered.

Let there be m_i genes at time t_i , let $a_1^{(i)}, \dots, a_{r_i}^{(i)}$ be the number of cases where $1, \dots, r_i$ copies survive



fractionation until time t_{i+1} , so that $\sum_{j=1}^{r_i} a_j^{(i)} = m_i$. Note that there is no provision for g to have zero surviving descendants (i.e., $a_0^{(i)} \equiv 0$); these genes would be considered as leaving no evidence for their existence and are not counted in m_i . Note that $m_{i+1} = \sum_{j=1}^{r_i} j a_j^{(i)}$.

We use $u_j^{(i)}$ to represent the probability that j of the r_i potential copies survive to time t_{i+1} , for $j = 1, \dots, r_i$.

Thus the probability distribution of the evolutionary histories represented by $\mathbf{r} = \{r_i\}_{i=1 \dots n-1}$ and the variable $\mathbf{a} = \{a_j^{(i)}\}_{j=1 \dots r_i}^{i=1 \dots n-1}$ is

$$P(\mathbf{r}; \mathbf{a}) = \prod_{i=1}^{n-1} \left[\binom{m_i}{a_1^{(i)}, \dots, a_{r_i}^{(i)}} \prod_{j=1}^{r_i} (u_j^{(i)})^{a_j^{(i)}} \right]. \quad (1)$$

The expected number of genes at time t_n is then

$$\mathbf{E}(m_n) = \sum_{\mathbf{a}} P(\mathbf{r}; \mathbf{a}) m_n. \quad (2)$$

Similarly, we write

$$P^{(j,k)}(\mathbf{r}; \mathbf{a}) = \prod_{i=j}^{k-1} \left[\binom{m_i}{a_1^{(i)}, \dots, a_{r_i}^{(i)}} \prod_{h=1}^{r_i} (u_h^{(i)})^{a_h^{(i)}} \right] \quad (3)$$

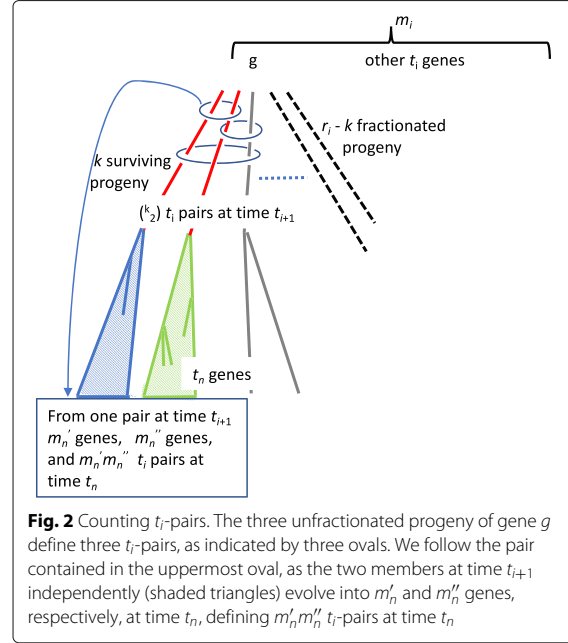
for the probability measure over all events starting at time t_j with m_j genes, and preceding time t_k . In this case the expected number of genes at time t_k is

$$\mathbf{E}^{(j,k)}(m_k) = \sum_{\mathbf{a}} P^{(j,k)}(\mathbf{r}; \mathbf{a}) m_k. \quad (4)$$

The paralog pairs

Having characterized the origin and survival of individual genes and their descendants in the environment of recurrent polyploidization and fractionation, we can now focus on the pairs of genes observed at time t_n . Our discussion is illustrated by Fig. 2.

For each of the $a_j^{(i)}$ genes with j surviving copies, $j \geq 2$, there are $\binom{j}{2}$ surviving pairs of genes. If $j = 1$ there are



no pairs. The total number of pairs created at time t_i and surviving to time t_{i+1} is thus

$$d^{(i,i+1)} = \sum_{j=2}^{r_i} \binom{j}{2} a_j^{(i)}. \quad (5)$$

These are called the t_i -pairs at time t_{i+1} . The expected number of such pairs is

$$\mathbf{E}(d^{(i,i+1)}) = \sum_{\mathbf{a}} P^{(1,i+1)}(\mathbf{r}; \mathbf{a}) \sum_{j=2}^{r_i} \binom{j}{2} a_j^{(i)}. \quad (6)$$

At time t_j , for $i+1 \leq j \leq n$, any two descendants of the two genes making up a t_i -pair with no more recent common ancestor is also called a t_i -pair (at time t_j). In other words, for any two genes at time t_j , they form a t_i -pair if their most recent common ancestor underwent polyploidization at time t_i .

For a given t_i -pair g' and g'' at time t_{i+1} , where $i < n-1$, the expected number of pairs of descendants $d^{(i,n)}$ having no more recent common ancestor than g' and g'' , will be

$$\mathbf{E}(d^{(i,n)}) = \mathbf{E}(d^{(i,i+1)}) (\mathbf{E}^{(i+1,n)}(m_n))^2 \quad (7)$$

where $m_{i+1} = 1$. This follows from the independence of the fractionation process between time t_i and time t_{i+1} and both parts of the process starting with g' and g'' .

Not all the m_n genes in Eq. (2) are in pairs. Because of fractionation after every polyploidization event, some genes will remain unpaired. We have

$$m^* \in \{0, \dots, m_1\}$$

$$\mathbf{E}(m^*) = m_1 \prod_{i=1}^{n-1} u_1^{(i)}, \quad (8)$$

where m^* is the current number of unpaired genes.

The terms $\mathbf{E}(d^{(i,i+1)})$ and $\mathbf{E}^{(i+1,n)}(m_n)$ in Eq. (7) both involve calculating probabilities with Eq. (3). As n and the r_i increase, so do the m_i , and this becomes computationally very expensive, due to the product of multinomial coefficients in Eq. (3) and the sum of many such probabilities in Eq. (4). Nevertheless, making use of the recursive nature of these calculations allows for more efficiency than the explicit generation of evolutionary histories and the counting of pairs within each one.

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 F of similarities. The second step is to derive the actual distribution of gene pair similarities, or an appropriate approximation to it, for t_i -pairs.

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).

Besides p , the other important parameter is G , reflecting average gene length in terms of the number of nucleotides in the genes' coding region. Because this length varies greatly, in practice G needs to be estimated.

In the simplest case, the distribution of similarities is the binomial $B(G, p_i)$, where

$$p_i = \frac{1}{4} + \frac{3}{4}e^{-\lambda t_i} \in [0, 1], \quad (9)$$

and is related to the time $t_i \in [0, \infty)$ elapsed since the event that gave rise to the pair. This distribution has

$$\begin{aligned} \text{mean : } \mu_i &= Gp_i \\ &= G \left(\frac{1}{4} + \frac{3}{4}e^{-\lambda t_i} \right) \in [0, 1] \\ \text{variance : } \sigma_i^2 &= Gp_i(1 - p_i) \\ &= \frac{3}{16}G(1 + 3e^{-\lambda t_i})(1 - e^{-\lambda t_i}), \end{aligned} \quad (10)$$

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

The densities of similarities of t_i -pairs can be approximated by a normal distribution $\mathbf{N}(\mu_i, \sigma_i^2)$ (as long as p_i is not too close to 1.0), and the expected frequency by

$$F_i = \mathbf{E}(d^{(i,n)}) \mathbf{N}(\mu_i, \sigma_i^2). \quad (11)$$

We can predict the entire frequency distribution over all t_i as:

$$F(t) = \sum_{i=1}^{n-1} F_i(t), \quad (12)$$

keeping in mind that the model also predicts unpaired genes according to Eq. (8). Then the predicted relative frequencies become

$$\begin{aligned} q(i) &= \frac{F(i)}{\mathbf{E}(m^*) + \sum_j F(j)} \\ q^* &= \frac{\mathbf{E}(m^*)}{\mathbf{E}(m^*) + \sum_j F(j)}. \end{aligned} \quad (13)$$

Inference

The distribution of gene pair similarities is of the form $f(k)$, where $k = k_{\min}, \dots, k_{\max}$. The data may also include f^* , the frequency of unpaired genes. The value of $k_{\min}$ is set to eliminate pairs due to noise or to polyploidization events earlier than those of immediate interest. At the other extreme, $k_{\max}$ is set somewhat lower than 100%, in order to remove any effects of heterozygosity, whereby an apparent duplicate gene pair actually consists of two alleles of a single gene, rather than two genes at different positions in the genome.

The likelihood of a model, given some data set is

$$L = C \prod_{i=k_{\min}}^{k_{\max}} q(i)^{f(i)} (q^*)^{f^*}, \quad (14)$$

where q depends only on the parameters of the model, and the maximum likelihood estimators of the parameters of a model can be found by maximizing

$$f^* \log q^* + \sum_{i=k_{\min}}^{k_{\max}} f(i) \log q(i) \quad (15)$$

with respect to these parameters.

Results and discussion

Two models for one dataset

For a given instance of the above model, if we know some of the parameters, we can infer the others. This includes

- the t_i : the times of each event,
- the fractionation rates,
- λ : the divergence rate, and
- G : the gene length parameter.

However, we cannot easily estimate the r_i , the event ploidies, from the distribution of paralog pair similarities. To understand why, we consider an important example, namely the outcome of two events, a tetraploidy leading to a whole genome duplication and hexaploidy, leading to

a whole genome triplication. We will illustrate with data on the *Brassica rapa* genome [4], member of the *Brassica* genus, known to have undergone a triplication after an earlier duplication shared with other Brassicales genera, such as *Arabidopsis*, as shown in Fig. 3.

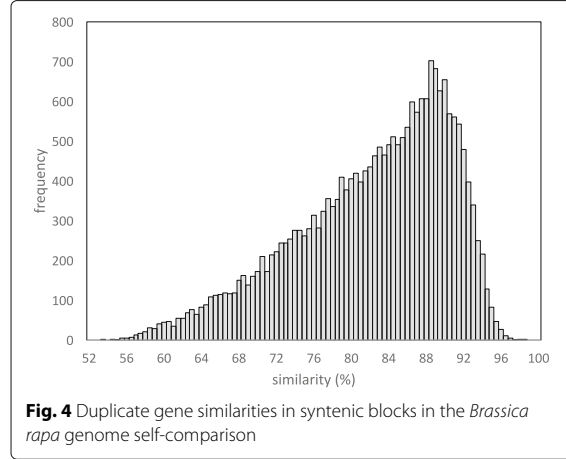
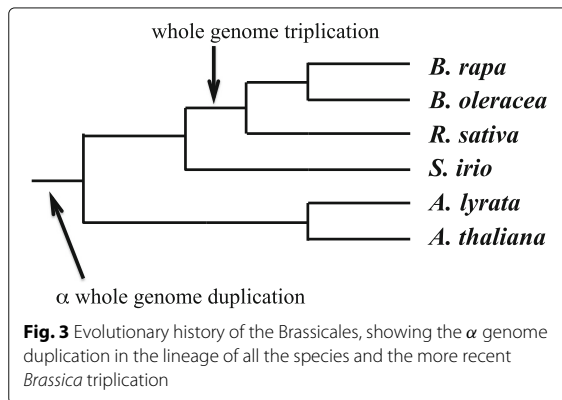
The distribution of gene pair similarities derived from SYNMAP (on the COGE platform [5, 6]) is shown in Fig. 4. Only the recent event, the *Brassica* triplication, is clearly visible as a distinct peak in the histogram, but the voluminous tail at early similarities attests to the effect of the earlier Brassicales duplication. Brassicales is a rosoid order and as such also descends from the γ core eudicot triplication, which would have produced pairs with around 70% similarity, but very few remained in synteny blocks, so for the purposes of our subsequent analysis, we ignore this event. Indeed, we imposed no bounds $k_{\min}$ or $k_{\max}$ on the data produced by SYNMAP.

We can explore the discriminatory power of our method by fitting two models to these data, one where a whole genome duplication precedes a triplication, known to be true, and an incorrect one where the duplication follows the triplication.

The calculations leading to Eqs. (7) and (8) are not lengthy in the case of these two models, and are portrayed schematically in Figs. 5 and 6, where u, v, w, x, y and z are probabilities that can be fixed independently of each other, as long as $u + v < 1$ and $x + y < 1$. (To avoid trivial models in either case, u, w, x and z must be greater than zero and less than 1, while v and y must be greater or equal to zero and less than 1. We will term these *valid models*.)

In Fig. 5, $w = u_2^{(1)}$, the probability that both offspring survive until time t_2 after the first duplication at time t_1 , so that $1 - w$ is the probability that only one survives. After the triplication at time t_2 , the probabilities are $u = u_3^{(2)}$ and $v = u_2^{(2)}$ that three offspring or two offspring, respectively, survive until time t_3 .

Looking at the second paralog tree in the left-hand column of Fig. 5, for example, Eq. (1) becomes



$$\begin{aligned}
 P(\mathbf{r}; \mathbf{a}) &= \prod_{i=1}^{n-1} \left[\binom{m_i}{a_1^{(i)}, \dots, a_{r_i}^{(i)}} \prod_{j=1}^{r_i} \left(u_j^{(i)} \right)^{a_j^{(i)}} \right] \\
 &= \binom{1}{a_1^{(1)}, a_2^{(1)}} w^{a_2^{(1)}} (1-w)^{a_1^{(1)}} \left(\frac{2a_2^{(1)} + a_1^{(1)}}{a_1^{(2)}, a_2^{(2)}, a_3^{(2)}} \right) \\
 &\quad \times u^{a_3^{(2)}} v^{a_2^{(2)}} (1-u-v)^{a_1^{(2)}} \\
 &= \binom{1}{0, 1} w \binom{2}{0, 1, 1} uv \\
 &= 2wvu.
 \end{aligned} \tag{16}$$

The coefficient 2 in this expression corresponds to the two “versions” of the diagram with the colours of the gene pair and gene triple permuted. Note that the branches without coloured dots in most of the paralog trees are simply meant to be suggestive of the fractionation process, do not reflect anything in Eq. (1), and are not involved in the colour permutations in counting the number of versions. The number of t_1 and t_2 pairs at time t_3 , as calculated in Eqs. (5-8), can be counted directly for each tree in the figure.

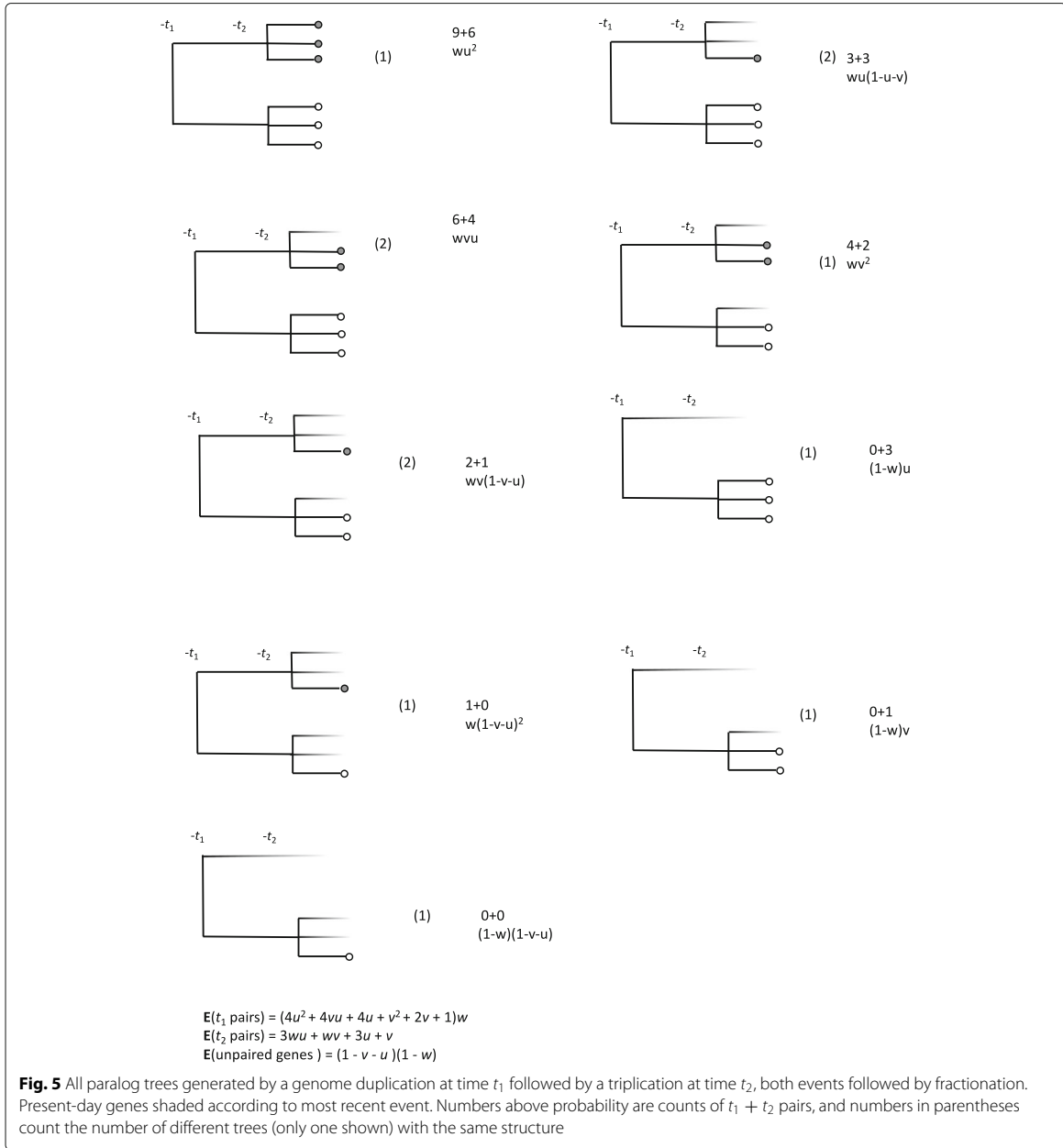
Turning to Fig. 6, $z = u_2^{(2)}$, the probability that both offspring survive until time t_3 after the second event (duplication) at time t_2 , so that $1 - z$ is the probability that only one survives. After the triplication at time t_1 , the probabilities are $x = u_3^{(1)}$ and $y = u_2^{(1)}$ that three offspring or two offspring, respectively, survive until time t_2 .

The expected number of pairs in the duplication precedes triplication model (Fig. 5) is given by:

$$E(t_1 \text{ pairs}) = (4u^2 + 4uv + 4u + v^2 + 2v + 1)w \tag{17}$$

$$E(t_2 \text{ pairs}) = 3wu + wv + 3u + v$$

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



The same quantities in the triplication-first model (Fig. 6) are:

$$E(t_1 \text{ pairs}) = (x + 3y)(1 + z)^2 \quad (18)$$

$$E(t_2 \text{ pairs}) = 2yz + xz + z$$

$$E(\text{unpaired}) = (1 - z)(1 - x - y)$$

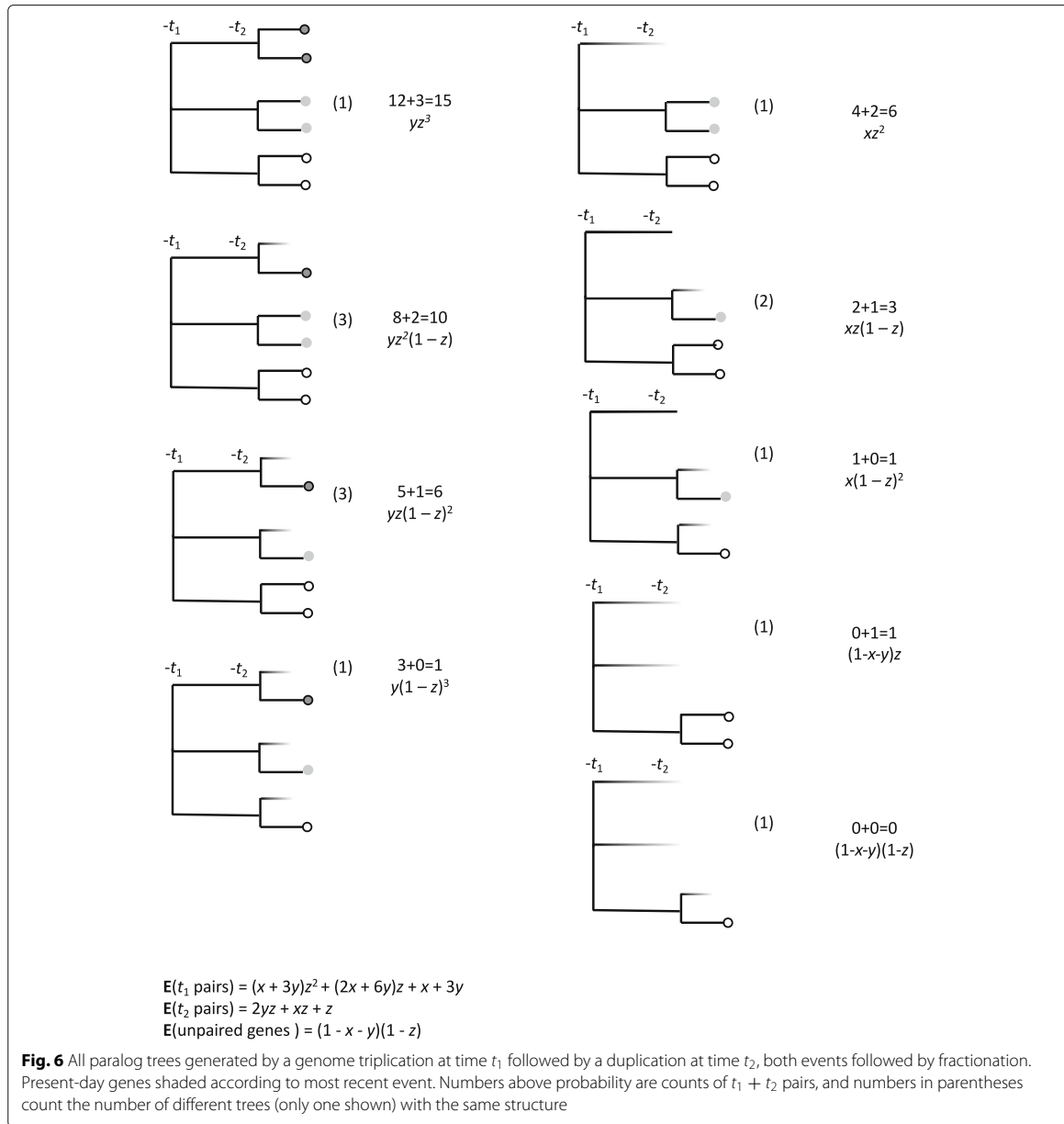
Can the principle of maximum likelihood discriminate between the two models? The likelihood of either model depends only on the $q(i)$ and q^* in Eq. (14). Then the

parameters of one model can be related to a set of parameter values in the complex field (i.e., not necessarily probabilities) with the same likelihood in the other model through the equations:

$$(4u^2 + 4uv + 4u + v^2 + 2v + 1)w = (x + 3y)(1 + z)^2$$

$$3wu + wv + 3u + v = 2yz + xz + z$$

$$(1 - w)(1 - u - v) = (1 - z)(1 - x - y), \quad (19)$$



This implies that all maximum likelihood solutions in both models have the same likelihood. Furthermore, for large enough samples we can expect at least one maximum likelihood solution in each valid model. Because the parameters are underdetermined, the likelihood depending only on the $q(i)$ and q^* and not the absolute frequencies $f(i)$ and f^* , there may be several solutions. In addition, if in one model the maximum likelihood solution involves parameters which satisfy the conditions of a valid

model, this is not necessarily true of the corresponding parameters in the other model.

We may then ask, do Eqs. (17) and (18) determine a bijection between some valid model in (u, v, w) space and some valid model in (x, y, z) space? The answer is determined by the intersection of $(u, v, w) \in [0, 1]^3 \cap \{u + v < 1\}$ and $(x, y, z) \in [0, 1]^3 \cap \{x + y < 1\}$ and the algebraic variety determined by the system in Eq. (19).

By systematically exploring a three-dimensional grid in each cube, we located all points in the valid region of the (u, v, w) cube for which Eq. (19) produced points in the valid region of the (x, y, z) cube. This produced two surfaces as depicted in Fig. 7 between which the two models have a correspondence. Outside of this volume, Eq. (19) have only solutions that are complex or outside one or both valid regions.

In Fig. 7, we see that the multiple maximum likelihood solutions for the *B. rapa* data in (u, v, w) space form a linear subspace, only part of which also contains solutions for the (x, y, z) model.

To restrict the set of solutions, we may make use of f^* , the observed number of unpaired genes, which is not directly involved in the likelihood maximization - only q^* is. This number is 17,751. Unfortunately, we have no access to the number of genes in the ancestral genome preceding the two polyploidization events, but we can guess, based on core eudicots that have escaped polyploidization after γ , such as grape [1] and Robusta coffee [7], that 25,000 is a reasonable value. Of genomes that have polyploidized and fractionated, many, such as *Utricularia*, papaya, *Mimulus* or most pertinent, *Arabidopsis*, have returned currently to gene numbers less than 29,000 [8]. This means that of 25,000-29,000 ancestral genes, the average number of currently unpaired genes per original gene is 0.61-0.71. In Fig. 7, the grey dots at the extreme left represent valid solutions in the (u, w, v) space but not the (x, y, z) space, predicting 0.65-0.68 unpaired genes while for the blue dots corresponding to valid solutions in both spaces the predicted number is only 0.53-0.55, suggesting an ancestral gene complement of

32,000-34,000, which seems unlikely. These calculations thus lend more credence to the model where duplication precedes triplication.

In this model, we have used u and v as two independent parameters controlling fractionation for the triplication event, and similarly x and y in the triplication-first model. This may represent excessive parametrization, however, since there are very likely biological constraints on such pairs of parameters, though this has not yet been modeled or studied empirically. A reasonable way of modelling this is to postulate a constrained binomial process for the fractionation loss of one or two genes of each triple generated by the triplication event. Thus we may replace u and v by using a single parameter s and replace x and y by a single parameter h as follows:

$$\begin{aligned} u &= \frac{s^2}{3(1-s) + s^2}, & v &= \frac{3s(1-s)}{3(1-s) + s^2} \\ x &= \frac{3h(1-h)}{3(1-h) + h^2}, & y &= \frac{h^2}{3(1-h) + h^2}. \end{aligned} \quad (20)$$

The investigation into the connection between the two models starts with

$$\begin{aligned} \frac{9w}{(3-3s+s^2)^2} &= \frac{(3h+6hz+3hz^2)}{(3-3h+h^2)} \\ \frac{(3ws+3s)}{(3-3s+s^2)} &= \frac{3z}{(3-3h+h^2)} \\ \frac{(1-w)(3(1-s)^2)}{(3-3s+s^2)} &= \frac{(1-z)(3(1-h)^2)}{(3-3h+h^2)} \end{aligned} \quad (21)$$

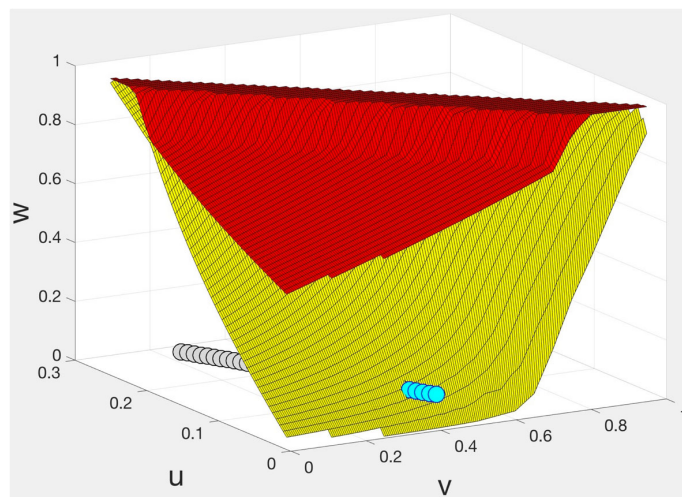


Fig. 7 (u, v, w) space with MLE solutions for (x, y, z) model (blue dots) and (u, v, w) model (blue and grey dots). Corresponding (x, y, z) models exist only for the (u, v, w) values in the volume bounded by the red and yellow sheets

There being only two parameters in each model, we can use only two of the Eqs. (21) to understand the correspondences between these models. In this case there is a bijection between $(s, w) \in [0, 1]^2$ and $(h, z) \in [0, 1]^2$. However, corresponding points in the two spaces reflect different numbers of unpaired genes (and of t_1 pairs and t_2 pairs).

The vertical axis in Fig. 8 represents the difference between the predictions of unpaired genes in the two models, with the red line tracing the values where the difference is zero. The blue line represents the maximum likelihood solutions for the *B. Rapa* data. The yellow dots identify (s, w) solutions that predict 0.6-0.68 currently unpaired genes per ancestral gene. Far from the red line, we conclude that the triplication-first model does not represent reality.

Counting triples

We have seen that in contrast to the divergence and fractionation parameters, the ploidy of whole genome multiplication events is not easy to infer from the distribution of gene pair similarities. There are more direct ways, however, to establish the ploidy of these events. Most obvious in our case is the detection of highly similar, i.e., recently diverged, triples of genes or triples of chromosomal regions, as evidence of the late triplication model.

The total number of genes in the CoGe *B. rapa* data set is 41,020. The application of SYNMAP to compare *B. rapa* with itself (cf. Fig. 4) shows there are 14 triples of paralogous regions made of long synteny blocks with relatively little interruption. These regions cover 80% of the genome, as can be seen in Fig. 9. For some of these triples, one or

more of the three regions are divided among two or three chromosomes, due to genome rearrangement processes. But they all display the signature pattern of recent triplication enunciated by Jaillon et al. [1], namely large numbers of highly similar gene pairs *among* the three regions and relatively few highly similar gene pairs *within* each region, or between the regions and other parts of the genome.

The total number of gene pairs detected by SYNMAP is 22,406, including 13,716 (61%) where the members are in two paralogous regions and have high similarity, defined as 81% or higher. Of these, for the overwhelming majority, both members are in the 14 triples of regions.

Looking at all the high-similarity pairs, a large number of these form gene triples, 2392 of them, that are not part of a 4-tuple or higher. The great majority of these gene triples, 2118, are located in the 14 triples of high-similarity regions.

We can contrast this situation with the predictions of the triplication-first model. If the time of the duplication corresponded to $p = 0.87$, then there would be many pairs with similarity > 0.81 , but few triples where all the pairs satisfied > 0.81 . There would, however, be many triples where all three pairs had similarity < 0.81 . This is clearly not the case.

Conclusion

The decomposition of gene pair similarity distributions into a number of normal distributions has been staple of comparative genomics. Statistical mixture of distributions methods [9] have been used extensively, to detect the distributions, to find their means and to test their significance. Because these are general methods, they do not take into account the biological processes that gave

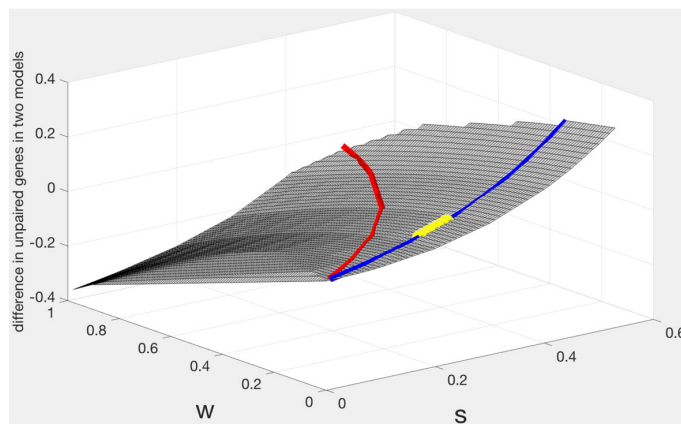
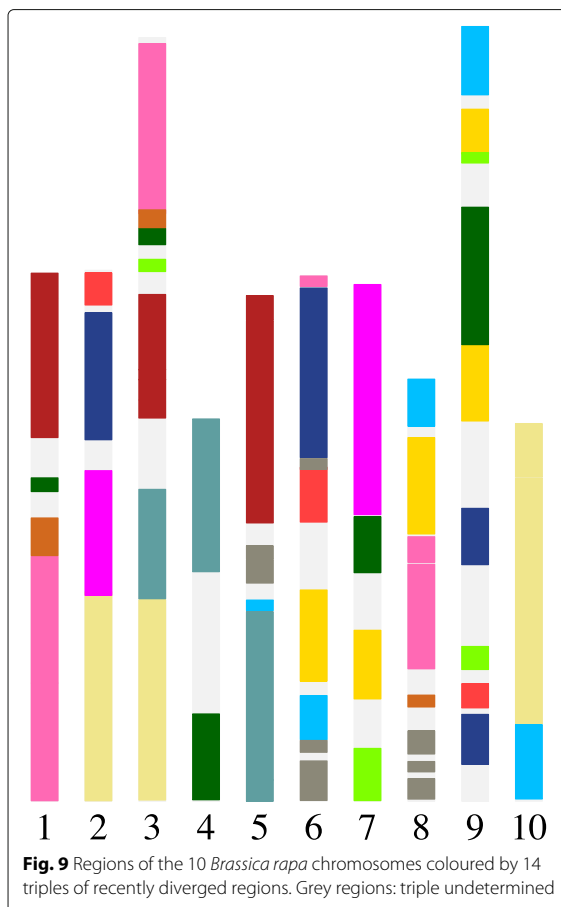


Fig. 8 (s, w) surface showing the difference in the number of unpaired genes between corresponding points in the two models. The red line indicates points where the two models agree on the number of genes, but this does not intersect with the set of MLE solutions on the blue line). In particular, the yellow dots indicating realistic numbers of unpaired genes are far from the red line



rise to the distribution and thus may lead to meaningless results. For example, they can find significance in a two-distribution decomposition, in which the one reflecting an earlier event has smaller variance than the most recent one, a biological impossibility. They can produce a decomposition where a peak with a large amplitude is succeeded by one with very small amplitude, again biologically implausible.

For distributions reflecting genuine events, mixture methods may provide accurate measures of the timing of these events, but offer little else of biological interest. Our models go further, allowing, for the first time, the estimation of fractionation rates from pair similarity distributions. We have proposed algebraic machinery for comparing competing models, and as an illustrative test, used it to confirm what was already well-known, that the *B. rapa* genome triplication is more recent than its duplication event.

At the end we must conclude, despite the unexpected insights provided by mathematically modeling the

genome multiplication-fractionation cycle, that to decide on the ploidy of the multiplication events, the strongest evidence, at least for the most recent events, is found in the tabulation of high-similarity pairs, triples, or other multiples. If few of the high-similarity pairs are in triples or other tuples, then the most recent event is likely to have been a tetraploidization. If a large proportion of the pairs are in triples but not in higher tuples, the event must have been a hexaploidization.

By judiciously parsing the similarity axis using cut-off values between peaks of the distribution, or between the mean values of inferred normal components of the overall distribution, we might hope in some cases to extend this simple approach to find the multiplicity of the earlier events,

Acknowledgements

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

Funding

The publication cost of this article was partly funded by Natural Sciences and Engineering Research Council of Canada Discovery Grant number 8867-2008.

Availability of data and materials

Brassica rapa genome available on Coge platform, as cited.

About this supplement

This article has been published as part of *BMC Genomics* Volume 19 Supplement 5, 2018: Proceedings of the 15th Annual Research in Computational Molecular Biology (RECOMB) Comparative Genomics Satellite Workshop: genomics. The full contents of the supplement are available online at <https://bmcgenomics.biomedcentral.com/articles/supplements/volume-19-supplement-5>.

Authors' contributions

The study was planned by all three authors, who also wrote the paper. YZ and DS developed the probabilistic models, did the algebraic analysis of the polyploidization-fractionation models and the statistical fit to the *B. rapa* data. CZ carried out the syntenic analysis and identification of the current chromosomal distribution of the *B. rapa* subgenomes. All authors read and approved the paper.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Published: 8 May 2018

References

1. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Huguency P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyère C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gasparo G, Dumas V, Felice N, Paillard S, Juman I, Moroldo M,

- Scalabrín S, Canaguier A, Le Clairche I, Malacrida G, Durand E, Pesole G, Laucou V, Chatelet P, Merdinoglu D, Delledonne M, Pezzotti M, Lecharny A, Scarpelli C, Artiguenave F, Pè ME, Valle G, Morgante M, Caboche M, Adam-Blondon AF, Weissenbach J, Quétier F, Wincker P, French-Italian Public Consortium for Grapevine Genome Characterization. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*. 2007;449:463–7.
2. Zhang Y, Zheng C, Sankoff D. Evolutionary model for the statistical divergence of paralogous and orthologous gene pairs generated by whole genome duplication and speciation. *IEEE/ACM Trans Comput Biol Bioinformatics*. 2017. <https://doi.org/10.1109/TCBB.2017.2712695>.
3. Zhang Y, Sankoff D. The similarity distribution of paralogous gene pairs created by recurrent alternation of polyploidization and fractionation. In: Meidanis J, Nakhleh L, editors. *Comparative Genomics. RECOMB-CG*. vol. 10562. Lecture Notes in Computer Science. Heidelberg: Springer; 2017. p. 1–13.
4. Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun JH, Bancroft I, Cheng F, Huang S, Li X, Hua W, Wang J, Wang X, Freeling M, Pires JC, Paterson AH, Chalhou B, Wang B, Hayward A, Sharpe AG, Park BS, Weissshaar B, Liu B, Li B, Liu B, Tong C, Song C, Duran C, Peng C, Geng C, Koh C, Lin C, Edwards D, Mu D, Shen D, Soumpourou E, Li F, Fraser F, Conant G, Lassalle G, King GJ, Bonnema G, Tang H, Wang H, Belcram H, Zhou H, Hirakawa H, Abe H, Guo H, Wang H, Jin H, Parkin IA, Batley J, Kim JS, Just J, Li J, Xu J, Deng J, Kim JA, Li J, Yu J, Meng J, Wang J, Min J, Poulain J, Wang J, Hatakeyama K, Wu K, Wang L, Fang L, Trick M, Links MG, Zhao M, Jin M, Ramchiary N, Drou N, Berkman PJ, Cai Q, Huang Q, Li R, Tabata S, Cheng S, Zhang S, Zhang S, Huang S, Sato S, Sun S, Kwon SJ, Choi SR, Lee TH, Fan W, Zhao X, Tan X, Xu X, Wang Y, Qiu Y, Yin Y, Li Y, Du Y, Liao Y, Lim Y, Narusaka Y, Wang Y, Wang Z, Li Z, Wang Z, Xiong Z, Zhang Z, Brassica rapa Genome Sequencing Project Consortium. The genome of the mesopolyploid crop species *Brassica rapa*. *Nat Genet*. 2011;43:1035–9.
5. Lyons E, Freeling M. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J*. 2008;53:661–73.
6. Lyons E, Pedersen B, Kane J, Alam M, Ming R, Tang H, Wang X, Bowers J, Paterson A, Lisch D, Freeling M. Finding and comparing syntenic regions among *Arabidopsis*, and the outgroups papaya, poplar and grape: CoGe with rosids. *Plant Physiol*. 2008;148:1772–81.
7. Denoeuf F, Carretero-Paulet L, Dereeper A, Droc G, Guyot R, Pietrella M, Zheng C, Alberti A, Anthony F, Aprea G, Aury JM, Bento P, Bernard M, Bocs S, Campa C, Cenci A, Combes MC, Crouzillat D, Da Silva C, Daddiego L, De Bellis F, Dussert S, Garsmeur O, Gayraud T, Guignon V, Jahn K, Jamilloux V, Joët T, Labadie K, Lan T, Leclercq J, Lepelley M, Leroy T, Li LT, Librado P, Lopez L, Muñoz A, Noel B, Pallavicini A, Perrotta G, Poncet V, Pot D, Priyono, Rigoreau M, Rouard M, Rozas J, Tranchant-Dubreuil C, VanBuren R, Zhang Q, Andrade AC, Argout X, Bertrand B, de Kochko A, Graziosi G, Henry RJ, Jayarama, Ming R, Nagai C, Rounsley S, Sankoff D, Giuliano G, Albert VA, Wincker P, Lashermes P. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science*. 2014;345:1181–4.
8. Ibarra-Laclette E, Lyons E, Hernández-Guzmán G, Anahí Pérez-Torres C, Carretero-Paulet L, Chang T-H, Lan T, Welch AJ, Abraham Juárez MJ, Simpson J, Fernández-Cortés A, Arteaga-Vázquez M, Góngora-Castillo E, Acevedo-Hernández G, Schuster SC, Himmelbauer H, Minoche AE, Xu S, Lynch M, Oropeza-Aburto A, Cervantes-Pérez SA, Ortega-Estrada MJ, Cervantes-Luevano JJ, Michael TP, Mockler T, Bryant D, Herrera-Estrella A, Albert VA, Herrera-Estrella L. Architecture and evolution of a minute plant genome. *Nature*. 2013;498:94–8.
9. McLachlan GJ, Peel D, Basford KE, Adams P. The **EMMIX**, software for the fitting of mixtures of normal and t-components. *J Stat Softw*. 1999;4:1–14.

Submit your next manuscript to BioMed Central
and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit



Chapter 4

A branching process for homology distributionbased inference of polyploidy, speciation and loss

Chapters 4 and 5 review the details of the branching process model for generating populations of paralogs , as summarized from Chapter 3 but derive simpler expressions(no summations, no factorials) and extend the model to introduce speciation to derive the expected number of orthologous pairs with most recent common ancestors at each ancestral time. Chapter 5 completes the development of the combined WGD and speciation model with a particular attention to the fractionation regimes immediately before and after speciation, as well as a formal demonstration of the fact that no peaks due to fractionation show up after the speciation peak in the comparison of two genomes. The methods are applied to the evolution of six genomes from the plant family Solanaceae (Chapter 4) and ten genomes from the family Malvaceae (Chapter 5). Of particular interest is the systematic inference of fractionation rates made

4. A BRANCHING PROCESS FOR HOMOLOGY DISTRIBUTIONBASED INFERENCE OF POLYPLOIDY, SPECIATION AND LOSS

possible for the first time by our models, and the fact that these rates in the woody, perennial Malvaceae species are approximately half of those in the annual, herbaceous Solanaceae plants. These two chapters represent the culmination of both our development of the branching process model and its application to produce original biological results.

RESEARCH

A branching process for homology distribution-based inference of polyploidy, speciation and loss

Yue Zhang, Chunfang Zheng and David Sankoff*

*Correspondence:
sankoff@uottawa.ca
Department of Mathematics and
Statistics, University of Ottawa,
150 Louis Pasteur pvt, Ottawa,
Canada K1N 6N5
Full list of author information is
available at the end of the article
†Equal contributor

Abstract

Background: The statistical distribution of the similarity or difference between pairs of paralogous genes, created by whole genome doubling, or between pairs of orthologous genes in two related species is an important source of information about genomic evolution, especially in plants.

Methods: We derive the mixture of distributions of sequence similarity for duplicate gene pairs generated by repeated episodes of whole gene doubling. This involves integrating sequence divergence and gene pair loss through fractionation, using a branching process and a mutational model. We account not only for the timing of these events in terms of local modes, but also the amplitude and variance of the component distributions. This model is then extended to orthologous gene pairs.

Results: We apply the model and inference procedures to the evolution of the Solanaceae, focusing on the genomes of economically important crops. We assess how consistent or variable fractionation rates are from species to species and over time.

Keywords: Branching process; Whole genome doubling; Fractionation; Solanaceae

Background

An important source of information in the study of genomic evolution is the statistical distribution of the similarity or difference between pairs of *paralogous* genes, created by one or more rounds of polyploidization and resulting in whole genome doubling (WGD), tripling (WGT),..., or between pairs of *orthologous* genes, a consequence of speciation. In comparative genomics we try to identify peaks or local modes of these distributions, in order to assign chronological dates to each of the WGD or speciation events. Our approach has been to account for these data through processes of paralogous gene pair divergence by point mutation, and by gene pair loss through duplicate gene deletion – *fractionation*, in terms of a succession of multinomial samples integrated with a standard model of sequence divergence. This tries to account not only for the timing of peaks, but also their amplitude and how compact or diffuse they are [1, 2, 3]. In the present paper, we reformulate this model in terms of branching processes and extend it to the study of orthologous gene pairs, so that we can apply it to the evolution of the Solanaceae, focusing on the genomic comparisons among tomato, potato, eggplant, pepper, tobacco and petunia genomes. Our main goal is to systematically and quantitatively analyze the process of gene loss, using this family as an example, to assess how consistent or variable fractionation rates are from species to species and how they change over time.

We first review the classical discrete-time branching process and comment on how applicable it is for generating populations of paralogs. We also discuss the possibilities and limitations of statistical inference of the parameters of the model.

We then derive the expected counts of present-day paralogous pairs created at each ancestral time. These results are then reduced to simpler expressions (no summations, no factorials) for several important cases. We extend our model to introduce speciation, which allows us to derive the expected number of orthologous pairs with most recent common ancestors at each ancestral time.

In order to account for genomic data, we can observe all the paralogous pairs, as well as the orthologous pairs if two species are involved, but we cannot directly observe at which WGD or speciation time each pair originated. Here is where the mutational model plays a role. A paralog or ortholog pair does not consist of two identical genes, in terms of identical DNA sequence, but they are considerably more similar than two random sequences. However, the similarity decreases as the time from pair origination increases; nucleotide changes affect the DNA of both genes independently according to a relatively constant rate parameter. The set of pairs generated by a single WGD or speciation event displays a distribution of similarities, whose mean is inversely related to the time from that event to the present and whose variance reflects the degree of randomness of the process of similarity decay. The similarities of all the pairs originating from all the events thus constitute a mixture of distributions. The means of the component distributions cannot usually be estimated by averaging, because of extensive overlap, but can be identified as local modes in the distribution of gene pair similarities. Maximum likelihood methods can then fill out the remaining information about the variances of each component distribution and their proportions in the mixture.

We apply our model and methodology to six genomes from the Solanaceae (“nightshade”) family of flowering plants using the grapevine genome as an outgroup. We compare all the genomes to each other (21 comparisons) and five of the six to themselves, using the SYNMAP tool on the COGE platform [4, 5] to obtain the distribution of paralogous and orthologous gene pair similarities, resulting from WGD and speciation event. The goal is to estimate rates of fractionation, based on the information previously derived about the component distributions. We then compare the results from the 26 distributions for consistency and for variation between genomes.

Methods

The Classical Branching Process in WGD Context

In our process the discrete time parameter $i = 1, 2, \dots, n$ is interpreted as the generation number and X_i denotes the number of genes present in the i^{th} generation. Generation $i + 1$, for $1 < i < n$ consists of the copies of genes in the i^{th} generation as follows:

Each gene j in the i^{th} generation produces a random number ξ_j with distribution

$$u_k^{(i)} = P[\xi_j = k], \text{ for } k = 0, \dots, r_i, \quad (1)$$

where $r_i \in \{2, 3, \dots\}$ is the *ploidy* of the i^{th} whole genome event. The distribution $u_k^{(i)}$ depends on i and so may differ from generation to generation.

Let $X_1 = 1$, then for $i \geq 1$

$$X_{i+1} = \sum_{k=1}^{X_i} \xi_k. \quad (2)$$

The mean and variance of the number of copies in the $i + 1^{st}$ generation per gene in the i^{th} generations are then

$$\mu_i = \sum_{k=0}^{r_i} k u_k^{(i)}, \quad \sigma_i^2 = \sum_{k=0}^{r_i} (k - \mu_i)^2 u_k^{(i)}. \quad (3)$$

The generating function for this event, defined for $s \in [0, 1]$, is

$$f(s) = E[s^\xi] = \sum_{k=0}^{r_i} u_k^{(i)} s^k, \quad (4)$$

and so

$$f'(1) = \mu_i, \quad f''(1) = \sigma_i^2 + \mu_i^2. \quad (5)$$

Suppose r_i and the u_k are the same for every generation. The basic result on branching processes, dating from the 19th century work of Galton and Watson, is that the probability of eventual extinction is the smallest positive root of the fixed point equation

$$f(s) = s, \quad (6)$$

which, in the biologically most relevant case $r_i \equiv 2$, becomes,

$$u_2 s^2 - (1 - u_1)s + 1 - u_1 - u_2 = 0, \quad (7)$$

whose roots are

$$s = \frac{1 - u_1 \pm \sqrt{(1 - u_1)^2 - 4u_2(1 - u_1 - u_2)}}{2u_2} \quad (8)$$

$$= \left\{ 1, \frac{1 - u_1}{u_2} - 1 \right\}. \quad (9)$$

This implies that the probability of extinction is less than 1 if and only if $u_0 < u_2$.

In the other important case, $r_i \equiv 3$, we have

$$u_3 s^3 + u_2 s^2 - (1 - u_1)s + 1 - u_1 - u_2 = 0, \quad (10)$$

where the pertinent root is

$$\frac{-(u_3 + u_2) + \sqrt{((u_3 + u_2)^2 + 4u_3u_0)}}{2u_3}. \quad (11)$$

Applicability of a Branching Process Model

A clear difference between classical branching processes and the WGD-fractionation process lies in the role of the time scale. Branching processes have a time scale made up of the positive integers, and all individuals in the population “reproduce” at the same time. WGD also affects all the genes in a genome synchronously, but it is a critical aspect for the analysis of fractionation that the $n - 1$ WGD times are not limited to integers but may take on any real values between the starting time and the time of observation (or current time). We will circumvent this problem by considering the integer time scale of the branching process to represent the succession of generations in the population of genes, and by introducing a vector of event times, independent of the branching process. These event times will interact with the other model parameters during the inference procedures, but are not properly part of the model itself.

Another contrast between classical branching processes and the WGD-fractionation process, is that whereas the focus of branching process theory is the prediction of extinction over the long term, with our genome level studies we are generally interested in as few as one, but generally two, three or four events. And we are interested in fractionation in each generation and not the cumulative probability of eventual extinction.

Furthermore, our motivation is essentially an inference problem based on present-day genomes, but we have no access to gene families that have gone extinct; we cannot observe them in current genomes in order to analyze their genes.

Yet another difference is in the interpretation of the probabilities u_k . In the branching process model, these are the probabilities that any particular individual has k offspring. In WGD, on the other hand, all genes simultaneously give rise to exactly r copies, but the number that survive until the next event is governed by u_k . This reflects the fact that branching processes do not refer to anything between one branching event and the next, whereas after a WGD, fractionation takes place in the interval between that event and the next one. Despite this difference, on the formal level, there is no mathematical difference between the abstract model and the biological description.

Thus, though there are differences between branching processes and the biological phenomena of WGD and fractionation, the model fits the basic biology very well. A WGD occurs within an infinitesimal time period, a few generations, on the scale of evolutionary history spanning millions, tens of millions and hundreds of millions years, so the branching process with all individuals reproducing at the same time, is realistic. And the fractionation process, where many or most of the r duplicate genes are lost before the next WGD, can be nicely modelled by considering the u_k to be survival probabilities rather than probabilities of offspring numbers.

Branching processes provide a realistic conceptual framework for the biological phenomena, but the biology in turn offers a novel kind of mathematical problem, namely to account for the ages of all the *pairs* of genes, i.e., the time they diverged from their latest common ancestor.

The Inference Problem

This work is motivated by an interest in extracting information about evolutionary history from frequency distributions of homolog similarity scores. These distribu-

tions, depending on the particular type of score (similarity, K_s , 4DTv, ...), are in fact mixtures of normals, or mixtures of some other kind of distribution, with non-negligible levels of noise, and subject to distortions and false signals of various types.

Dissecting mixtures of normals is a statistical problem that crops up in a number of fields, and there are standard techniques for carrying this out [6]. These methods, however, cannot necessarily adapt to field-specific constraints, not to mention noise inhomogeneous with respect to the similarity score, and other problems with the data. In comparative genomics, the distributions in the mixture tend to overlap to a large extent, the quality of the data diminishes and noise level increases with lower similarity score. The variance of the component distributions increases with lower similarity score, but not in an entirely predictable manner.

Nevertheless, as we shall try to demonstrate, it is feasible to pin down the dates of at least four WGD and speciation events in the history of a genome. For each of the mixture components originating with these events we can estimate a mean, a variance, and a proportion, the latter referring to the area under the component distribution as a proportion of the total area. Only the mean and the proportion turn out to be useful statistics in our eventual analysis of fractionation, meaning that we can only estimate two parameters in the model for each component in the mixture of distributions. In addition, another data item is sometimes available, the current number of unpaired genes, which should allow the estimation of an additional parameter affecting the most recent WGD or speciation event [2]. The biologically motivated constraint $u_0 = 0$ can be invoked to allow estimation of u_2 , even though this “no lineage extinction” condition is an overstatement, given that not all genes are essential, and that occasionally both copies of a gene are lost.

Details of the Branching Process - The Evolution of Population Size

Denote by $M_1, \dots, M_n$ the total number of individuals (genes) existing in the population at generation $1, 2, \dots, n$. To get from generation i to generation $i + 1$, for $i = 1, \dots, n - 1$, each of the population of M_i genes is first replaced by $r_i \geq 2$ progeny. We call r_i the *ploidy* of the event. We denote by $u_j^{(i)}$ the probability that j of these r_i progeny survive until generation $i + 1$, for $j = 0, \dots, r_i$. (In applying this model we often assume $u_0^{(i)} = 0$ – “no lineage extinction” – so that we gain a degree of freedom for estimating other parameters. But this constraint is not really required in the model.) There is no replacement event at the n^{th} and final generation; this is simply the point at which the population is observed.

Let $a_0^{(i)}, \dots, a_{r_i}^{(i)}$ be the number of genes at generation i , of which $0, \dots, r_i$, respectively, survive until generation $i + 1$, so that

$$M_i = \sum_{j=0}^{r_i} a_j^{(i)}, \quad M_{i+1} = \sum_{j=0}^{r_i} j a_j^{(i)}. \quad (12)$$

The probability distribution of the evolutionary histories represented by $\mathbf{r} = \{r_i\}_{i=1 \dots n-1}$ and the variable $\mathbf{a} = \{a_j^{(i)}\}_{j=0 \dots r_i}^{i=1 \dots n-1}$ is

$$P(\mathbf{r}; \mathbf{a}) = \prod_{i=1}^{n-1} \left[\binom{M_i}{a_0^{(i)}, \dots, a_{r_i}^{(i)}} \prod_{j=0}^{r_i} (u_j^{(i)})^{a_j^{(i)}} \right], \quad (13)$$

as can be proved by induction on i . The expected number of genes at generation n is

$$\mathbf{E}(M_n) = \sum_{\mathbf{a}} P(\mathbf{r}; \mathbf{a}) M_n. \quad (14)$$

Similarly, for the events starting at generation j with M_j genes, up to generation k , we write

$$\begin{aligned} P^{(j,k)}(\mathbf{r}; \mathbf{a}) &= \prod_{i=j}^{k-1} \left[\binom{M_i}{a_0^{(i)}, \dots, a_{r_i}^{(i)}} \prod_{h=1}^{r_i} (u_h^{(i)})^{a_h^{(i)}} \right] \\ \mathbf{E}^{(j,k)}(M_k) &= \sum_{\mathbf{a}} P^{(j,k)}(\mathbf{r}; \mathbf{a}) m_k. \end{aligned} \quad (15)$$

Paralogous Gene Pairs

Having described the origin and survival of individual genes, we now summarize the analysis in [2] of the *pairs* of genes observed at generation n whose most recent common ancestor was replaced by r_i progeny at some generation i . For each of the $a_j^{(i)}$ genes with $j \geq 2$ surviving copies, there are $\binom{j}{2}$ surviving pairs of genes at generation $i+1$. The total number of pairs created at generation i and surviving to generation $i+1$ is thus

$$d^{(i,i+1)} = \sum_{j=2}^{r_i} \binom{j}{2} a_j^{(i)}. \quad (16)$$

These are called the i -pairs at generation $i+1$. The expected number of such pairs is

$$\mathbf{E}(d^{(i,i+1)}) = \sum_{\mathbf{a}} P^{(1,i+1)}(\mathbf{r}; \mathbf{a}) \sum_{j=2}^{r_i} \binom{j}{2} a_j^{(i)}. \quad (17)$$

At generation j , for $i+1 \leq j \leq n$, any two descendants of the two genes making up a i -pair *with no more recent common ancestor* is also called a i -pair (at generation j). In other words, for any two genes at generation j , they form an i -pair if their most recent common ancestor underwent replacement at generation i .

For a given i -pair g' and g'' at generation $i+1$, where $i < n-1$, the expected number of pairs of descendants $d^{(i,n)}$ having no more recent common ancestor is

$$\mathbf{E}(d^{(i,n)}) = \mathbf{E}(d^{(i,i+1)}) (\mathbf{E}^{(i+1,n)}(M_n))^2 \quad (18)$$

where $M_{i+1} = 1$ in both factors representing the descendants of an i -pair. This follows from the independence among the fractionation process between generation i and $i+1$ and both processes starting with g' and g'' .

Of the M_n genes in Equation (14), the expected number of unpaired genes is

$$\mathbf{E}(M^*) = M_1 \prod_{i=1}^{n-1} u_1^{(i)}. \quad (19)$$

Reductions to Simple Form

The accumulation of multinomial coefficients in Equation (13), and the potentially high degree polynomials might seem computationally formidable. In practice, however, n seldom attains 5 or 6, and the r_i are generally 2 or 3. Thus individual instances of the model are generally computationally tractable. In addition, though Equation (18) would seem to entail an increasing complexity of formulae as n increases, in many important cases this reduces to simple expressions.

Successive Doublings (Tetraploidizations) For example if all $r_i = 2$ for $1 \leq i \leq n-1$, we have by induction that Equation (18) reduces to

Theorem 1

$$\begin{aligned} E(N_1) &= u_2^{(1)} \Pi_{j=2}^{n-1} (1 + u_2^{(j)})^2 \\ E(N_i) &= \Pi_{j=1}^{i-1} (1 + u_2^{(j)}) u_2^{(i)} \Pi_{j=i+1}^{n-1} (1 + u_2^{(j)})^2 \\ E(N_{n-1}) &= u_2^{(n-1)} \Pi_{j=1}^{n-2} (1 + u_2^{(j)}), \end{aligned} \quad (20)$$

where N_i is the expected number of duplicate pairs of genes produced at generation i surviving until generation n .

Corollary 1 If all the $u_2^{(j)} = u$, then for $1 \leq i \leq n-1$,

$$E(N_i) = u(1 + u)^{2n-i-1}. \quad (21)$$

Successive Triplings (Hexaploidizations) In the case all $r_i = 3$ for $1 \leq i \leq n-1$,

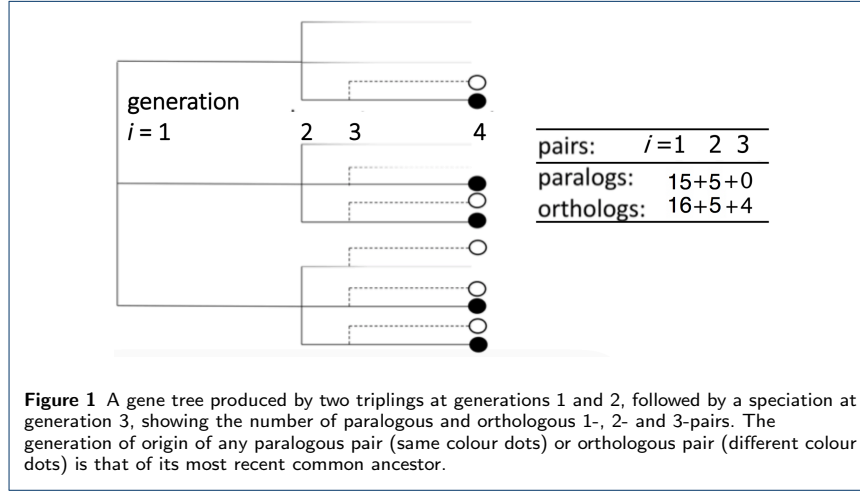
$$\begin{aligned} E(N_1) &= (3u_3^{(1)} + u_2^{(1)}) \Pi_{j=2}^{n-1} (1 + 2u_3^{(j)} + u_2^{(j)})^2 \\ E(N_i) &= \Pi_{j=1}^{i-1} (1 + 2u_3^{(j)} + u_2^{(j)}) (3u_3^{(i)} + u_2^{(i)}) \Pi_{j=i+1}^{n-1} (1 + 2u_3^{(j)} + u_2^{(j)})^2 \\ E(N_{n-1}) &= (3u_3^{(n-1)} + u_2^{(n-1)}) \Pi_{j=1}^{n-2} (1 + 2u_3^{(j)} + u_2^{(j)}). \end{aligned} \quad (22)$$

General r . For $r \geq 2$ the same for all generations, and $u_j^{(i)} = u_j$ for $j = 1, \dots, r$ and $i = 1, \dots, n-1$, there will be $K \geq 0$ and $K' \geq 0$, depending on the distribution of u_j , such that

$$E(N_i) = K' K^{2n-i-1}. \quad (23)$$

Introducing Speciation into the Model

When two populations of a species evolve into two daughter species, we may assume that they initially have the same gene complement, and share identical paralog trees. We can no longer observe the state of the paralog tree at generation n – that event is in the past – instead we observe the current set of orthologous gene pairs at generation $n+1$. Obviously, if such a tree has M_n genes at generation n , this will create at most M_n different orthologous n -pairs at generation $n+1$, the time of observation, taking into account the possibility of fractionation between the n -th and $n+1$ -st generations. Thus in Figure 1, though there are six genes in generation



3, we only observe four pairs of orthologs surviving fractionation after the speciation event.

One way to allow fractionation to continue beyond the speciation event is to extend the branching process, treating speciation as another WGD event, though the counting of orthologs is necessarily different than the counting of i -paralogs, as illustrated in Figure 1. For this sequence of events, the same logic behind Equations (20-23) allows us to write

$$\begin{aligned}
 E(O_1) &= 0.5(3u_3^{(1)} + u_2^{(1)})(1 + 2u_3^{(2)} + u_2^{(2)})^2(1 + u_2^{(3)})^2 \\
 E(O_2) &= 0.5(1 + 2u_3^{(1)} + u_2^{(1)})(3u_3^{(2)} + u_2^{(2)})(1 + u_2^{(3)})^2 \\
 E(O_3) &= u_2^{(3)}(1 + 2u_3^{(1)} + u_2^{(1)})(1 + 2u_3^{(2)} + u_2^{(2)}),
 \end{aligned} \tag{24}$$

where O stands for the number of ortholog pairs.

This approach is more general than simply counting two pairs of orthologs for every pair of paralogs required by the no fractionation assumption, since $u_2^{(3)}$ can be less than 1. However, even this is not really satisfactory, since it treats gene loss in one of the two genomes created at generation n as if it were the product of fractionation within a single genome, when in fact the two genomes are entirely independent of each other. The “correct” way of proceeding would be to allow the fractionation regime operative between the $n-1$ -st and n -th generations to continue independently between the n -th and $n+1$ -st generations in each of the two genomes until the observation step $n+1$, or until this is interrupted by new WGDs in the two species. This is done as follows

For example, suppose there is just $M_1 = 1$ gene at generation 1, and suppose all $r_i = 2$. We can write $u(i) = u_2(i)$, $i = 1, \dots, n-1$ for the probability that both progeny of a gene at generation i survive until generation $i+1$. We rewrite

Equations (20) as

$$\begin{aligned}
\mathbf{E}(N_1) &= u(1)\Pi_{j=2}^{n-1}(1+u(j))^2 \\
\mathbf{E}(N_i) &= \Pi_{j=1}^{i-1}(1+u(j))u(i)\Pi_{j=i+1}^{n-1}(1+u(j))^2 \\
\mathbf{E}(N_{n-1}) &= \Pi_{j=1}^{n-2}(1+u(j))u(n-1)
\end{aligned} \tag{25}$$

Set $s = n$ to emphasize that this is a speciation event, and not a WGD or observation event. Suppose there are $n_A - 1 - s$ WGD in species A after speciation and $n_B - 1 - s$ in species B . Let

$$\begin{aligned}
F_A &= \Pi_{i=s}^{n_A-1}(1+u^A(i)) \\
F_B &= \Pi_{k=s}^{n_B-1}(1+u^B(k))
\end{aligned} \tag{26}$$

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

$$\begin{aligned}
\mathbf{E}(O_1) &= \frac{1}{2}u(1)\Pi_{j=2}^{s-1}(1+u(j))^2F_AF_B \\
\mathbf{E}(O_i) &= \frac{1}{2}\Pi_{j=1}^{i-1}(1+u(j))u(i)\Pi_{j=i+1}^{s-1}(1+u(j))^2F_AF_B \\
\mathbf{E}(O_s) &= \frac{1}{4}\Pi_{j=1}^{s-1}(1+u(j))F_AF_B
\end{aligned} \tag{27}$$

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 .

The three key factors in our improved model, terms in Equation (26) and (27), are $(1+u^A(s))$, $(1+u^B(s))$ and $(1+u(s-1))$. Between the two successive WGD, at generation $s-1$ in the pre-speciation genome, and $s+1$ in genome A and also $s+1$ in genome B – though the two events are generally not synchronous, the same fractionation regime, in terms of rates, should hold, despite the speciation at generation s . Calculation of rates requires not only the u , but also a time t_i associated with each event i . Writing

$$\begin{aligned}
-\log u(s-1) &= \rho(t_s - t_{s-1}), \\
-\log u^A(s) &= \rho_A(t_{s+1}^A - t_s), \\
-\log u^B(s) &= \rho_B(t_{s+1}^B - t_s),
\end{aligned} \tag{28}$$

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.

The Distribution of Similarities

The goal of this work is to understand fractionation, so that if at the time of observation we could count the i -pairs for $i \geq 1$, we could use Equations like (20-24) as a basis for making inferences about the $u_j^{(i)}$. But although we can observe all the paralogous pairs, as well as the orthologous pairs if two species are involved, we cannot *directly* observe at which WGD or speciation event each pair originated.

Instead, what we observe at generation n (or $n + 1$ in the case of orthology, or later if there have been WGD in the daughter species) is a measure p of similarity (e.g., the proportion of identical nucleotides in the aligned coding sequences) between each pair of genes in the population. Because of how sequence similarity decays by random substitutions of nucleotides, we can expect an approximately exponential decline in p with time.

Thus if the distribution of gene pair similarities clusters around values $p_1 < p_2 < \dots < p_{n-1}$, we can infer that these correspond to WGD events at some time $t_1 < t_2 < \dots < t_{n-1}$ at generations $1, \dots, n - 1$ of the branching process. And assuming a large sample of gene pairs, each of these clusters can be modelled by a normal distribution. The distribution of gene pairs is thus a mixture of $n - 1$ normals.

Previous work assumed that the variance of the similarity of a gene pair was proportional to $p(1 - p)$, but this did not provide a very good fit in practice. In the present paper, we do not assume any such relationship. Indeed, our strategy will be to identify the t_i by a combination of techniques described in the next paragraph, and fix these in a standard maximum likelihood estimate of the variance and amplitude of each component of the mixture. This enables us to calculate the proportion of all the gene pairs in each component. We use these proportions, or frequencies derived by multiplying by the total number of pairs, as the numbers of i pairs, from which we can estimate the survival proportions using Equations (20-24).

The Mode as an Estimator of t_i

The $t_i, i = 1, \dots, n - 1$ are not inherent parts of the branching process model. But they are of course very important for the study of evolution and the estimation of rates.

There are well-established methods for decomposing a mixture of normals (or other predetermined distributions) into their component distributions [6]. Experience shows, however, that these methods, despite their built-in validation criteria, are not robust against non-normality, especially with genomic data, and tend to deliver spurious extra components, and components located in unlikely places. We will nevertheless make use of these methods, but in a way constrained to give appropriate results.

We will compare several genomes to each other. Our strategy is first to locate the t_i in each comparison by picking out local modes in the distribution of similarities, guided by the knowledge that some of these t_i are shared among several genome comparisons, since they reflect the same events. Then for each comparison, some of these estimates are refined by maximum likelihood methods, which also produce the amplitude and variance of the component. From these we can directly estimate how many gene pairs are 1-pairs, 2-pairs, etc. These numbers can then be used to produce estimates of the $u_j^{(i)}$.

Why use the mode? Because of overlapping tails, reminiscent of the mixing of generations, i.e., the decay of synchrony, in initially synchronized population, studied in the antediluvian literature [7], the means of the component distributions cannot be estimated by averaging, but can be identified as local modes in the overall distribution of gene pair similarities.

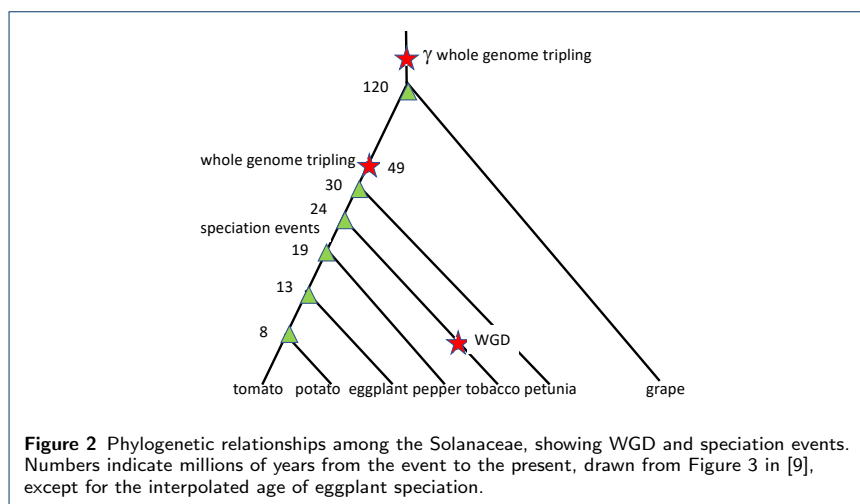
Estimating the local modes of an underlying distribution by using the modes of the sample involves a trade-off between precision and a proliferation of misleading

modes. With gene pair similarities grouped into large bins, or averaged among moving windows of large size, the empirical distribution will be relatively smooth, and bonafide modes will be easily noticed. But a large bin size only indicates that the mode is somewhere in a large interval. With small bin sizes, or sliding window sizes, the position of the nodes are more precisely determined, but more subject to a proliferation of spurious nodes due to statistical fluctuation. Again, we control this problem by considering several related comparisons at a time.

Results

The Evolution of the Family Solanaceae

The Solanaceae is a family of plants in the asterid order Solanales. This family is distinguished biologically by its early whole genome tripling, as indicated in Figure 2, and scientifically by the fact that many of its species boast sequenced genomes, namely all the economically most important ones (cf [8]).



The Genomes

We use the SYNMAP software on CoGE, and thus have direct access to most 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 tomato (*Solanum lycopersicum*) genome sequence and annotation [10] are considered the gold standard among the asterid genome projects. Although there is a recent update to version 3, we used the more familiar (from previous work) version 2.40.

The potato (*Solanum tuberosum*) genome [11] is also a high quality sequence has now been fully assembled into pseudomolecules (version 4.03).

The tobacco (*Nicotiana benthamiana*) genome was sequenced some years ago [12], but its sequence and annotation have been updated and made available for comparative purposes, together with the petunia (*Petunia hybrida*) genome [9], both via SGN—the Sol Genomics Network <https://solgenomics.net>. Among the

Solanaceae genomes studied here, only tobacco has undergone a WGD since the original Solanaceae tripling.

The pepper genome (*Capiscum annuum* version 1.55) [13] is drawn from a genus closely related to *Solanum*. We had no access to any updated version of this, and the quality of assembly and annotation is not as complete as those listed above.

A draft version of the eggplant genome (*Solanum melongena*) has also been available for some time [14], and this is what we use here despite its quality not measuring up to more recent standards, although a new version is available for browsing via SGN, with restrictions against comparative use awaiting the writing up and publication of the project.

As an outgroup, we use the grapevine (*Vitis vinifera*) genome [15], one of the first flowering plant genomes to be sequenced (in 2007), and one that has proven to be extraordinarily conservative, both with respect to mutational rate and to rearrangement of chromosomal structure. Indeed, the structure of the 19 grape chromosomes resembles in large measure that of the 21 chromosomes of the ancestor of the core eudicots, resulting from a tripling of a seven-chromosome precursor [16]. This is known as the “ γ ” tripling. Over half of the known flowering plants, including the Solanaceae, belong to this group.

The Comparisons

We applied SYNMAP to all pairs of the seven genomes and also compared each genome with itself (with the exception of eggplant, because of technical difficulties). We used the default parameters, which are fairly strict in ensuring that all pairs were part of a syntenic block, and thus created at the same time. This excluded duplicate gene pairs that may have been created individually, at some time other than during a WGD event.

The results are shown in Figures 3 and 4. In Figure 3, we note the relative stability of the γ and Solanaceae tripling-based distributions, but the narrowing of the speciation-based distributions as speciation time approaches the present.

In Figure 4, we note the conservatism of grape, which retains higher similarities for γ paralogs than the Solanaceae. That the γ -based orthologs in the Solanaceae comparisons with grape all suggest equally remote speciation times, rather than manifesting a compromise with the more recent grape-versus-grape values indicates that the Solanaceae ancestor underwent a period of relatively rapid evolution.

We compiled the characteristics - p, σ , number (and overall proportion) of pairs - for each component in each of the analyses in Figures 3 and 4. Of those in Figure 3, only the results for the speciation (most recent) event are displayed in Table 1. Figure 5 shows the relation between p and divergence time for the speciation event pertinent to each pair of genomes, and their common earlier WGD.

On the left of Figure 5, the cluster of points around 120 My represents the gene pairs generated by the γ tripling event pre-dating all core eudicots, too remote in time to be distinguished from the speciation of the ancestor of grape and the ancestor of the Solanaceae. Points near the centre represent the Solanaceae tripling. Scattered points at more recent times indicate the speciation events among the six Solanaceae species.

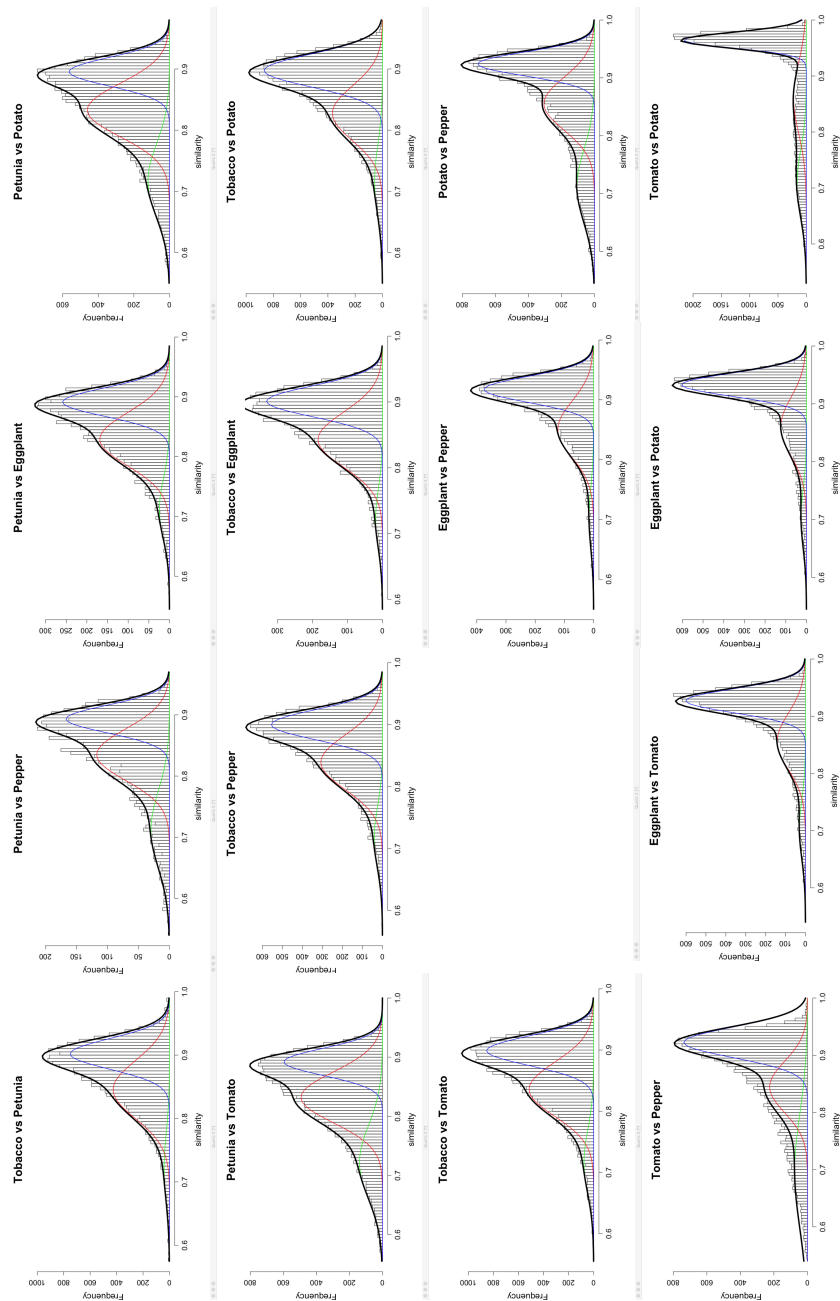


Figure 3 Distribution of ortholog similarities in comparisons among six Solanaceae genomes, with normal distributions fitted to similarities generated by each WGD and speciation event.

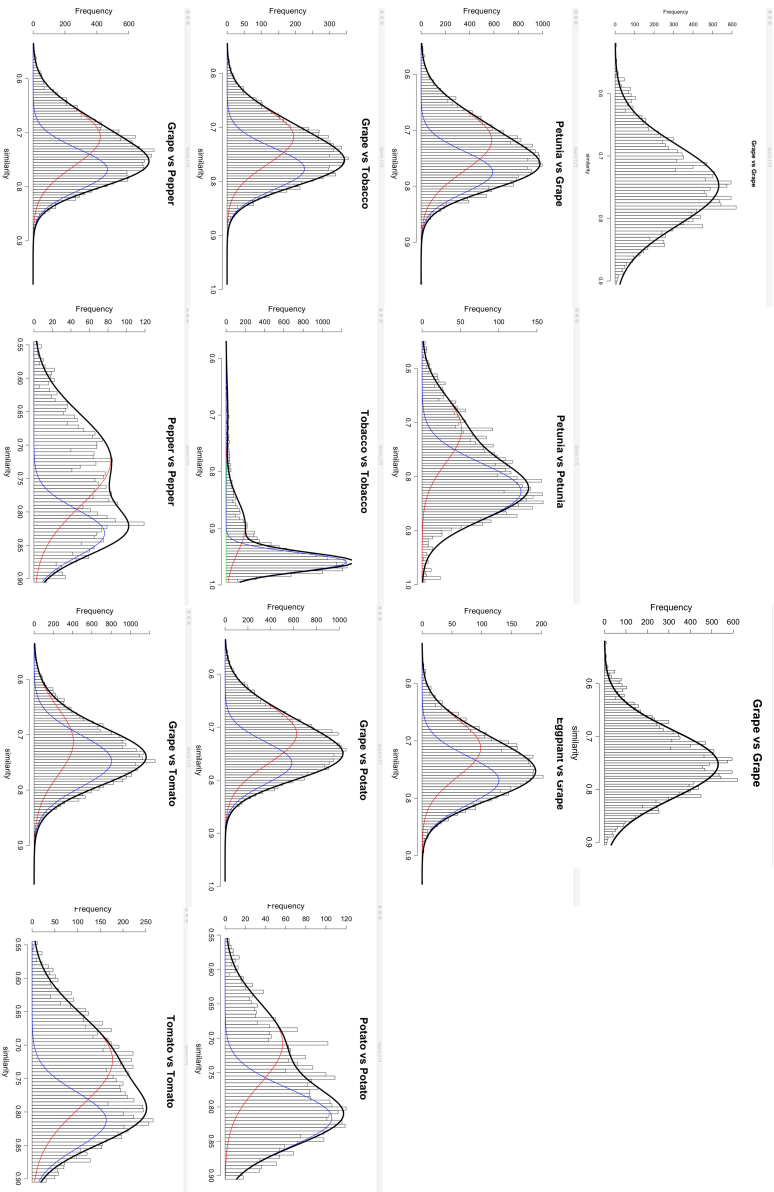


Figure 4 Distribution of paralog similarities in five* Solanaceae genomes and in grape, with normal distributions fitted to similarities generated by each WGD. This is compared to ortholog similarities in each Solanaceae genome versus grape. Two grape panels represent two slightly different fits to the data. Note the Y-axis in the tobacco self-comparison is out of proportion with the rest, because of its recent WGD.

*We were unable to run SYNMAP for eggplant self-comparison.

modal similarity p	Grape	Tomato	Potato	Eggplant	Pepper	Tobacco	Petunia
Grape		0.769	0.770	0.772	0.771	0.779	0.777
Tomato	0.769		0.966	0.930	0.925	0.901	0.895
Potato	0.770	0.966		0.936	0.927	0.900	0.898
Eggplant	0.772	0.930	0.936		0.920	0.904	0.894
Pepper	0.771	0.925	0.927	0.920		0.902	0.896
Tobacco	0.779	0.901	0.900	0.904	0.902		0.904
Petunia	0.777	0.895	0.898	0.894	0.896	0.904	

component σ	Grape	Tomato	Potato	Eggplant	Pepper	Tobacco	Petunia
Grape		0.039	0.040	0.041	0.038	0.039	0.037
Tomato	0.039		0.014	0.021	0.027	0.024	0.023
Potato	0.040	0.014		0.020	0.020	0.027	0.023
Eggplant	0.041	0.021	0.020		0.021	0.023	0.024
Pepper	0.038	0.027	0.020	0.021		0.024	0.022
Tobacco	0.039	0.024	0.027	0.023	0.024		0.024
Petunia	0.037	0.023	0.023	0.024	0.022	0.024	

component proportion	Grape	Tomato	Potato	Eggplant	Pepper	Tobacco	Petunia
Grape		0.45	0.40	0.51	0.44	0.47	0.40
Tomato	0.45		0.57	0.59	0.53	0.44	0.31
Potato	0.40	0.57		0.61	0.41	0.54	0.33
Eggplant	0.51	0.59	0.61		0.55	0.48	0.42
Pepper	0.44	0.53	0.41	0.55		0.46	0.36
Tobacco	0.47	0.44	0.54	0.48	0.46		0.45
Petunia	0.40	0.45	0.33	0.42	0.36	0.45	

Table 1 Characteristics inferred for speciation event distributions.

The trend line in the figure is $p = 1.2e^{-0.09t}$, which fits well, although the coefficient of the exponential is greater than expected (i.e., 1.0). The right of Figure 5 suggests that the standard deviation of the component normals are linearly related to their modes (and hence their means). The speciation data for modal values unequivocally support the phylogeny in Figure 2, e.g., as calculated by neighbour joining (not shown).

Fractionation rates

We calculated maximum likelihood estimates for $u_2^{(1)}$, $u_2^{(2)}$ and $u_2^{(3)}$, based on component proportions like those in the bottom section of Table 1. Because there are only two independent proportions per comparison, pertaining to t_1, t_2 and t_3 , and an estimate of the number of unpaired genes (predicted by the model in Equation (19)), we could not also infer the $u_3^{(i)}$, and simply assumed $u_3^{(1)} = (u_2^{(1)})^2$ and $u_3^{(2)} = (u_2^{(2)})^2$, on the premise that the small probability of two additional progeny surviving (beyond the one essential to avoid extinction) would be approximately the product of their individual probabilities. These event-specific and species-specific survival parameters $u_j^{(i)}$ on the left of Table 2 are directly estimable from the distribution statistics, and reveal much about the difference between the event and the species pairs, but our ultimate interest is in fractionation rates, which we denote ρ , and their consistency or variability. In general,

$$\begin{aligned} u(t) &= e^{-\rho t} \\ \rho &= \frac{-\ln u(t)}{t}. \end{aligned} \quad (29)$$

When we apply this rule to the survival rates in the table, using the time intervals derived from [9], we derive the fractionation rates on the right of the table. From the sections of Table 2 on survival we observe:

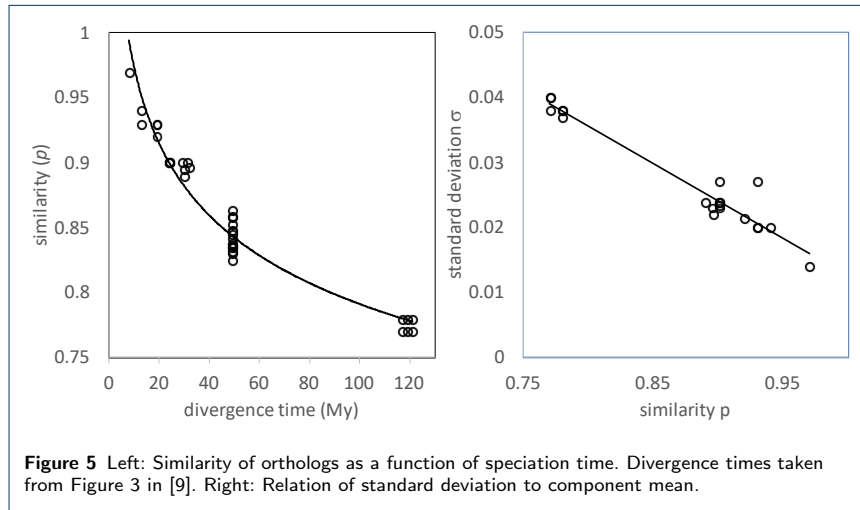


Figure 5 Left: Similarity of orthologs as a function of speciation time. Divergence times taken from Figure 3 in [9]. Right: Relation of standard deviation to component mean.

- the 15 estimates of survival between γ and the Solanaceae tripling are systematically much lower than the survival between the latter tripling and speciation, and after speciation.
- The early survival figures are quite variable; a major cause of this is the quality of the genome sequencing, assembly and annotation, so that comparisons of the draft genome sequence of eggplant, for example, apparently miss many of the gene pairs generated by γ .
- The high rates of survival in the comparisons involving petunia or tobacco over the time interval between the Solanaceae tripling and speciation clearly reflect the shorter time interval before their respective speciation events.
- The speciation survival results reflect, as expected, phylogenetic relationships, though imperfectly, due in part to sequence and annotation quality, and in part due to the amplification of the number of pairs in the recent tobacco WGD.

From the sections of Table 2 on fractionation rates we observe:

- A large reduction of variability (compared to survival) in the results for the inter-tripling interval, due only to the logarithmic transform.
- A large, but not complete, reduction in the difference between the two periods of fractionation, due to the normalization by the time span. This is compatible with the idea that fractionation rates may be universally constrained to a relatively narrow range of values.
- The high rates of post-speciation ortholog loss within *Solanum*, and the relatively low rates for the comparisons involving petunia or tobacco, suggest that the process initially proceeds more quickly than fractionation, or levels off after a certain point, or both.

The modeling leading to Equation (27) suggests that if only one of species A or B, undergoes another post-speciation WGD, we should be able to estimate the amplifying factor. Figure 6 suggests that tomato, which has undergone a WGT since its ancestral speciation from grape, has $F_A = 1.75$. This results is confirmed if we substitute potato instead of tomato, but the great variability in genome quality

γ survival x 100						
	Tomato	Potato	Eggplant	Pepper	Tobacco	Petunia
Tomato		5.12	1.64	8.34	4.05	7.89
Potato	5.12		1.42	7.13	2.68	6.58
Eggplant	1.64	1.42		0.98	1.05	1.93
Pepper	8.34	7.13	0.98		1.95	3.23
Tobacco	4.05	2.68	1.05	1.95		1.87
Petunia	7.89	6.58	1.93	3.23	1.87	

γ fractionation rate						
	Tomato	Potato	Eggplant	Pepper	Tobacco	Petunia
Tomato		0.042	0.058	0.035	0.045	0.036
Potato	0.042		0.060	0.037	0.051	0.038
Eggplant	0.058	0.060		0.065	0.064	0.056
Pepper	0.035	0.037	0.065		0.055	0.048
Tobacco	0.045	0.051	0.064	0.055		0.056
Petunia	0.036	0.038	0.056	0.048	0.056	

Solanaceae survival x100						
	Tomato	Potato	Eggplant	Pepper	Tobacco	Petunia
Tomato		15.6	7.22	12.32	26.79	22.92
Potato	15.6		6.43	16.6	13.59	20.97
Eggplant	7.22	6.43		6.79	8.92	11.78
Pepper	12.32	16.6	6.79		13.15	8.65
Tobacco	26.79	13.59	8.92	13.15		16.83
Petunia	22.92	20.97	11.78	8.65	16.83	

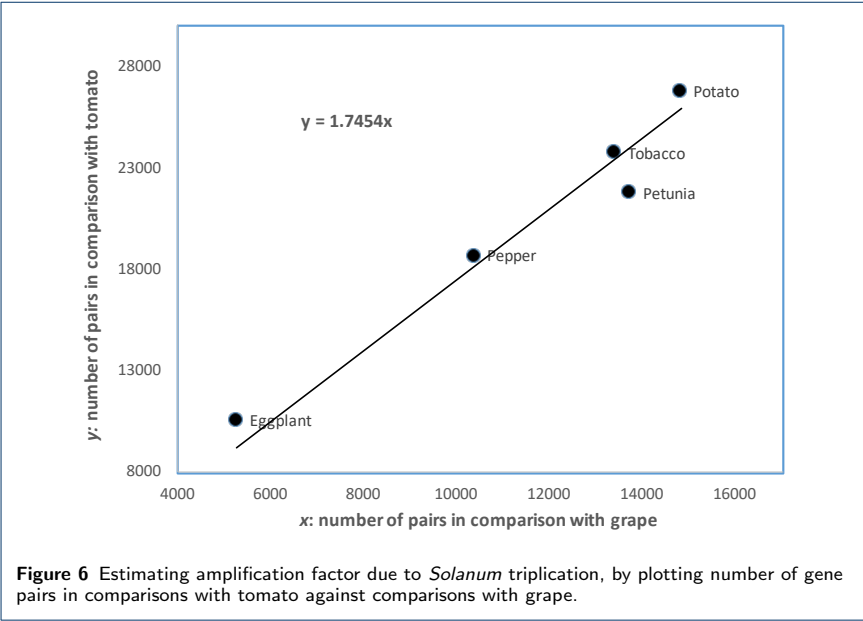
Solanaceae rate						
	Tomato	Potato	Eggplant	Pepper	Tobacco	Petunia
Tomato		0.045	0.073	0.070	0.053	0.078
Potato	0.045		0.076	0.060	0.080	0.082
Eggplant	0.073	0.076		0.090	0.097	0.113
Pepper	0.070	0.060	0.090		0.081	0.129
Tobacco	0.053	0.080	0.097	0.081		0.094
Petunia	0.078	0.082	0.113	0.129	0.094	

Speciation survival x100						
	Tomato	Potato	Eggplant	Pepper	Tobacco	Petunia
Tomato		27.52	8.53	20.66	23.84	11.68
Potato	27.52		7.95	13.56	15.55	10.61
Eggplant	8.53	7.95		5.97	5.80	6.53
Pepper	20.66	13.56	5.97		9.31	4.09
Tobacco	23.84	15.55	5.80	9.31		12.02
Petunia	11.68	10.61	6.53	4.09	12.02	

Speciation rate						
	Tomato	Potato	Eggplant	Pepper	Tobacco	Petunia
Tomato		0.161	0.189	0.083	0.060	0.072
Potato	0.161		0.195	0.105	0.078	0.075
Eggplant	0.189	0.195		0.148	0.119	0.091
Pepper	0.083	0.105	0.148		0.099	0.107
Tobacco	0.060	0.078	0.119	0.099		0.071
Petunia	0.072	0.075	0.091	0.107	0.071	

Table 2 Estimates of survival (left) and of fractionation rates (right).

precludes any meaningful results in other comparisons. In particular, we could not detect an effect of the recent tobacco WGD. Thus this kind of analysis must await the availability of a collection of related genomes with comparably high quality genome sequence.



Conclusions

We modelled the process of fractionation to account for the distribution of similarities between paralog or ortholog gene pairs after a number of whole genome doublings, triplings, etc., each followed by a period of duplicate gene loss. The model is

a discrete-time branching process, with its synchronous reproduction events across the population. Fractionation over the inter-generation interval is accounted for by the probability distribution on the number of offspring, interpreted instead as survival probabilities applied to a fixed number of offspring.

The observations of gene pair similarities consist of a mixture of normals, each component generated by one event, with the event time estimated by the sequence divergence from the event to the present. Despite the overlapping distributions, we can estimate the mean (*via* a local mode), standard deviation and proportion of the sample.

We then use these parameters to estimate survival probabilities for gene pairs from one event to the next. From the survival data we can then estimate fractionation rates, the number of gene pairs lost per unit time.

We apply our ideas to six genomes from the family Solanaceae and outlier grape. The SYNMAP program on the COGE platform produces the distribution of similarities of syntenically validated paralogs and orthologs to feed into our analysis. The 21 pairwise genome comparisons produce a highly consistent picture of the creation and loss of duplicate gene pairs. The survival probabilities and fractionation rates are eminently interpretable in terms of phylogenetic considerations. This work has now been replicated for the family Malvaceae [17].

Based on our methods and results, we can accurately characterize fractionation rates, something first attempted some years ago [18]. Indeed, we are now in a position to question to what extent fractionation embodies clocklike behaviour.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

The datasets generated during and/or analysed during the current study are available in the COGE repository, <https://genomeevolution.org/coge/>

Competing interests

The authors declare that they have no competing interests.

Funding

The research described in this article as well as publication costs were partly funded by Natural Sciences and Engineering Research Council of Canada Discovery Grant number 8867-2008.

Authors' contributions

The study was planned by all three authors, who also wrote the paper. YZ and DS developed the probabilistic models, did the statistical analysis. CZ participated in all aspects of the data handling and statistics. All authors read and approved the paper.

Acknowledgements

Not applicable.

Authors' information

DS holds the Canada Research Chair in Mathematical Genomics.

References

1. Zhang Y, Zheng C, Sankoff D. Evolutionary model for the statistical divergence of paralogous and orthologous gene pairs generated by whole genome duplication and speciation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2018;15(05):1579–1584.
2. Zhang Y, Zheng C, Sankoff D. Pinning down ploidy in paleopolyploid plants. *BMC Genomics*. 2018;19(5):287.
3. Sankoff D, Zheng C, Zhang Y, Meidanis J, Lyons E, Tang H. Models for similarity distributions of syntenic homologs and applications to phylogenomics. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2018;.
4. Lyons E, Freeling M. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *The Plant Journal*. 2008;53:661–673.
5. Lyons E, Pedersen B, Kane J, Freeling M. The value of nonmodel genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates rosids. *Tropical Plant Biology*. 2008;1:181–190.
6. McLachlan GJ, Peel D, Basford KE, Adams P. The EMMIX software for the fitting of mixtures of normal and t-components. *Journal of Statistical Software*. 1999;4(2):1–14.
7. Sankoff D. Duration of detectible synchrony in a binary branching process. *Biometrika*. 1971;58(1):77–81.
8. Vallée GC, Muñoz DS, Sankoff D. Economic importance, taxonomic representation and scientific priority as drivers of genome sequencing projects. *BMC Genomics*. 2016;17(10):782.
9. Bombarely A, Moser M, Amrad A, Bao M, Bapaume L, Barry CS, et al. Insight into the evolution of the Solanaceae from the parental genomes of *Petunia hybrida*. *Nature Plants*. 2016;2(6):16074.
10. Tomato Genome Consortium, et al. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*. 2012;485(7400):635.
11. Potato Genome Sequencing Consortium, et al. Genome sequence and analysis of the tuber crop potato. *Nature*. 2011;475(7355):189.
12. Bombarely A, Rosli HG, Vrebalov J, Moffett P, Mueller LA, Martin GB. A draft genome sequence of *Nicotiana benthamiana* to enhance molecular plant-microbe biology research. *Molecular Plant-Microbe Interactions*. 2012;25(12):1523–1530.
13. Kim S, Park M, Yeom SI, Kim YM, Lee JM, Lee HA, et al. Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. *Nature Genetics*. 2014;46(3):270.
14. Hirakawa H, Shirasawa K, Miyatake K, Nunome T, Negoro S, Ohyama A, et al. Draft genome sequence of eggplant (*Solanum melongena* L.): the representative *Solanum* species indigenous to the old world. *DNA Research*. 2014;21(6):649–660.
15. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*. 2007;449(7161):463.
16. Zheng C, Chen E, Albert VA, Lyons E, Sankoff D. Ancient eudicot hexaploidy meets ancestral eusoid gene order. *BMC Genomics*. 2013;14(7):S3.
17. Zhang Y, Zheng C, Islam S, Kim YM, Sankoff D. Branching out to speciation with a model of fractionation: the Malvaceae. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2019;.
18. Sankoff D, Zheng C, Zhu Q. The collapse of gene complement following whole genome duplication. *BMC Genomics*. 2010;11(1):313.

Chapter 5

Branching out to Speciation in a Model of Fractionation: the Malvaceae

Branching out to Speciation in a Model of Fractionation: the Malvaceae

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

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.

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.

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.

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

Manuscript received May 10, 2019; revised May 10, 2019.

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 than 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.

2 THE GENERAL MODEL

In our general model for WGD and fractionation, at times $t_1 < \dots < t_{n-1}$ 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 i -th 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 $u_1(i), \dots, u_{r_i}(i)$ for $i = 1, \dots, n - 1$. 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 $\mathbf{a}(i) = (a_1(i), \dots, a_{r_i}(i))$ represent the numbers of genes at time t_i , of which $1, \dots, r_i$, respectively, survive until t_{i+1} , so that

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

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

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

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

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

with $M_1 \geq 1$ given and the other M_i determined by Equations (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 $u(i) = u_2(i)$, $i = 1, \dots, n - 1$ 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:

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

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

$$\begin{aligned} F_A &= \Pi_{i=s}^{n_A-1} (1 + u^A(i)) \\ F_B &= \Pi_{k=s}^{n_B-1} (1 + u^B(k)) \end{aligned} \quad (5)$$

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

$$\begin{aligned} \mathbf{E}(O_1) &= \frac{1}{2}u(1)\prod_{j=2}^{s-1}(1+u(j))^2F_AF_B \\ \mathbf{E}(O_i) &= \frac{1}{2}\prod_{j=1}^{i-1}(1+u(j))u(i)\prod_{j=i+1}^{s-1}(1+u(j))^2F_AF_B \\ \mathbf{E}(O_s) &= \frac{1}{4}\prod_{j=1}^{s-1}(1+u(j))F_AF_B \end{aligned} \quad (6)$$

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 Equation (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_{s+1}^A in genome A and t_{s+1}^B in genome B , the same fractionation regime should hold, despite the speciation at t_s . Writing

$$\begin{aligned} -\log u(s - 1) &= \rho(t_s - t_{s-1}), \\ -\log u^A(s) &= \rho_A(t_{s+1}^A - t_s), \\ -\log u^B(s) &= \rho_B(t_{s+1}^B - t_s), \end{aligned} \quad (7)$$

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 rosoid order Malvales. Considering only the genomes studied here, this family bifurcated early into two lineages, as indicated in Figure 3, 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 *Gossypoides 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.)

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 i -th WGD can be approximated by a normal distribution $\mathbf{N}(p_i, \sigma_i^2)$, and the expected frequency by

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

We can predict the entire frequency distribution over all events as:

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

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

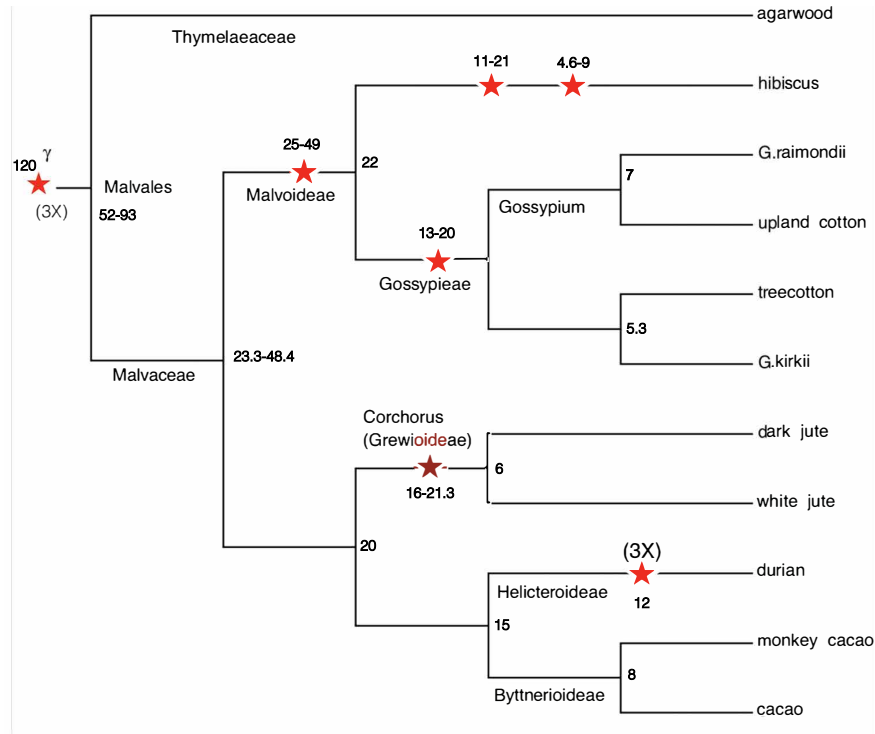


Fig. 1. Phylogenetic relationships among the Malvaceae, showing WGDs and speciation events. Numbers indicate millions of years from the event to the present, estimated from the following sources: agarwood (*Aquilaria*) divergence from Malvaceae [8], [16] and other references, and Malvoideae divergence from Grewioideae and Byttnerioideae [15], other times in the Malvoideae 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 % less than those given in the Time Tree of Life [13] or the Angiosperm Phylogeny Website [14], but are based on more recent technology.

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. In each comparison, we fix these p_i in a maximum likelihood mix-

ture 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 , $i = 1, \dots, s$, the number of ortholog pairs due to the WGD at time t_i , whose expectation we derived in Equations

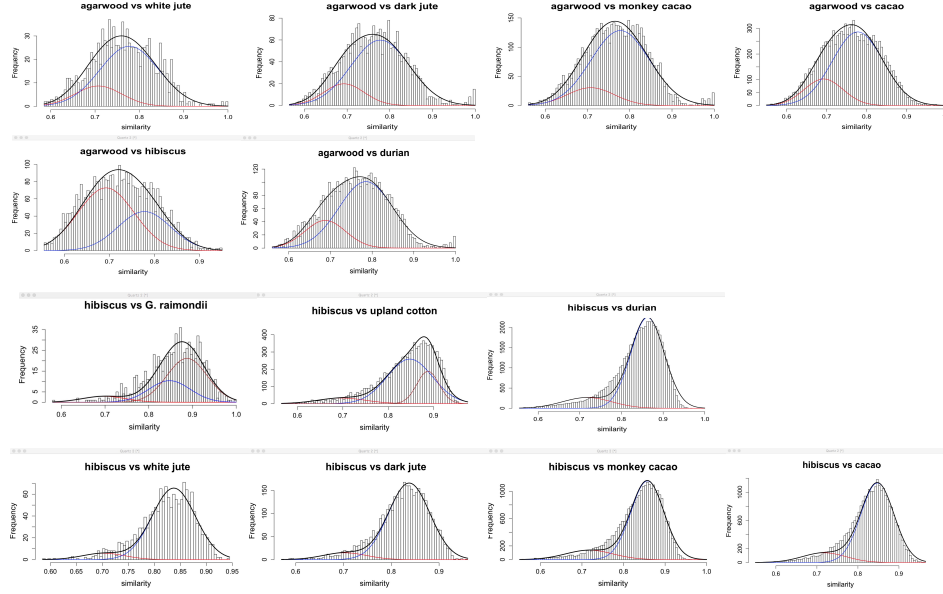


Fig. 2. Aquilaria and Hibiscus similarities

(6). Then

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

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 $u(i)$, $i = 1 \dots, s-1$, 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 $F_A F_B$ 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 $B_1, B_2, \dots$, 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 $B_1, B_2, \dots$ themselves underwent WGD after divergence from the A lineage, there is no direct way of factoring the product $F_A F_B$.

5 RESULTS

Figures 4, 4 and 4 depict the distributions and their decompositions for the 38 comparisons, including seven self-

comparisons, we succeeded in producing by COGE. Technical difficulties prevented us from obtaining five pairwise genomic comparisons and two self-comparisons. Table 5 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 in our pipeline between the original database, through the uploading to COGE, to the SYNMAP computations and display.

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% 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% similarity. In some cases, this partially obscures a 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

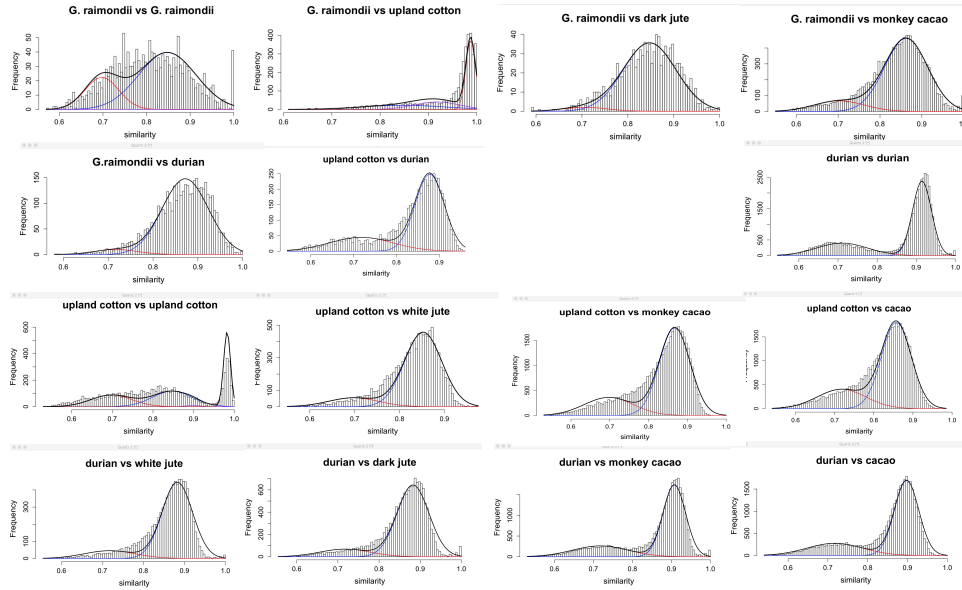


Fig. 3. Gossypium and Durio similarities

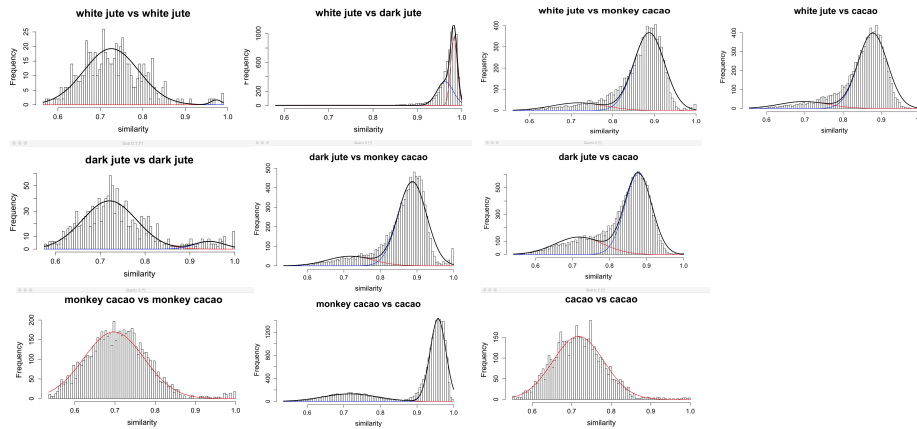


Fig. 4. Grewioideae and Byttnerioideae similarities

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.

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-Malvoideae 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 5 (top section). This contrasts

	agar-wood	hibiscus	G.raimondii	upland cotton	white jute	dark jute	monkey cacao	cacao	durian
mean									
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									
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									
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	

TABLE 1
Statistical parameters of the speciation peaks.

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 Figure 3, 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 5 drop exponentially with the assigned times in Figure 3, except for a lower similarity than expected for the divergence time of the Malvaceae species from the Malvales outgroup – see Figure 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 5 to estimate the $u(i)$ according to Equations (6), and transformed these according to the relationship $\rho = -\frac{\log u}{t}$, to estimate proportional fractionation rates, shown in Table 5. We performed a similar analysis on the self comparisons, shown in Table 5. In both cases a range of values is shown, corresponding to the range of time estimates shown in Figure 3. 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 Time-tree of Life [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

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 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.

WGD	G.rai-mondii	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

TABLE 3

Ranges of fractionation rates following WGD until the next WGD in a lineage, from self-comparisons.

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 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.

APPENDIX THE DURIAN GENOME

The original publication of the *Durio zibethinus* draft genome sequence [23] noted evidence of a paleoploidy event and suggested that this was identical with the event that affected the genus *Gossypium* and that this confirmed its phylogenetic grouping with *Gossypium* as distinct from *Theobroma*

cacao. In our SYNMAP analysis however, it became clear that the *Durio* event was much more recent than the *Gossypium* WGD (which itself is shared with its Malvoideae sister *Hibiscus*), and that the *Durio* event was in fact a triplication, and not a duplication.

In Figure 6, output by SYNMAP, we see that the ancient durian polyploidy was in fact a hexaploidy. This is unambiguous because of the remarkably low degree of rearrangement the chromosomes have undergone since the polyploidization event.

In Figure 4, in the distributions arising from the self-comparisons of uplands cotton, *G. raimondii* and durian, and the pairwise comparisons between durian and the other two, we see that the phylogenetic divergence of cotton and durian preceded the two cotton paleoploidy events, for which there is no trace in durian. Table 5 (top section, rightmost column) shows that durian is in fact more closely related to all four cacao and jute genomes than to the any of the three Malvoideae genomes (or to the Gossypieae genomes, not shown).

These results do not substantially contradict the phylogenetic literature, e.g., [28], [29]. Although the Grewioideae and Byttnerioideae are generally recognized as a distinct grouping among the nine subfamilies of the Malvaceae [14], their early branching is not strongly supported by bootstrap scores in any of the single-gene or small gene sample studies that have been published. Nor are the various groupings of *Durio* or the Helicteroideae with other subfamilies consistent

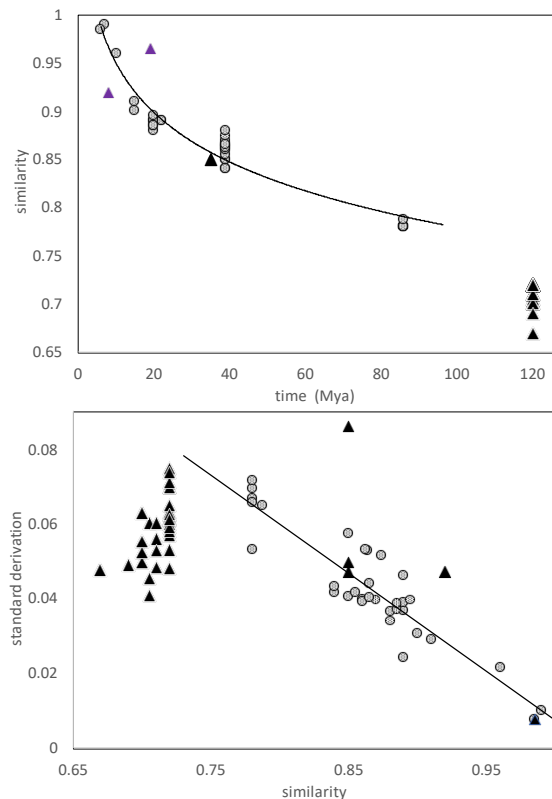


Fig. 5. Trends of peak modes and standard deviation. Trend lines fitted to Malvaceae speciation times and similarities only (dots). WGD times and similarities represented by triangles, with the large cluster pertaining to the γ triplication.

or strongly supported by molecular evidence.

Note: After the present article was accepted and in the publication pipeline, we noted that the ancient hexaploidization of durian was independently discovered and analyzed in much greater detail by Wang *et al.* [31].

ACKNOWLEDGMENTS

Research and publication costs were supported in part by Discovery Grant number 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 (2013) The Amborella genome and the evolution of flowering plants. *Science* 342:1241089
- [2] Chen ECH, Najjar CBA, Zheng C, Brandts A, Lyons E, Tang H, Carretero- Paulet L, Albert VA, Sankoff D (2013) The dynamics of functional classes of plant genes in rediploidized ancient polyploids. *BMC Bioinformatics* 14(S-15):19.
- [3] Gout JF, Kahn D, Duret L, Paramecium Post-Genomics Consortium (2010) The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. *PLoS Genetics* 6:1000944.
- [4] Thomas BC, Pedersen B, Freeling M (2006) Following tetraploidy in an *Arabidopsis* ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Research* 16(7):934-946.
- [5] Zhang Y, Sankoff D (2017) The similarity distribution of paralogous gene pairs created by recurrent alternation of polyploidization and fractionation. In: Meidanis J, Nakhleh L (eds) *Comparative Genomics. RECOMB-CG 2017. Lecture Notes in Computer Science* 10562: 1–13.
- [6] Zhang Y, Zheng C, Sankoff D (2018) Speciation and rate variation in a birth-and-death account of WGD and fractionation; the case of Solanaceae. to appear in the proceedings of the 16th RECOMB Comparative Genomics Satellite Workshop (Blanchette M, Ouangraoua A, eds) *Lecture Notes in Computer Science*.
- [7] Sankoff D, Zheng C, Zhang Y, Meidanis J, Lyons E, Tang H (2018) Models for similarity distributions of syntenic homologs and applications to phylogenomics. *IEEE/ACM Transactions in Computational Biology and Bioinformatics* doi: 10.1109/TCBB.2018.2849377
- [8] De La Torre AR, Li Z, Van de Peer Y, Ingvarsson PK (2017) Contrasting rates of molecular evolution and patterns of selection among Gymnosperms and flowering Plants. *Molecular Biology and Evolution* 34: 1363–1377, <https://doi.org/10.1093/molbev/msx069>
- [9] Zhang Y, Zheng C, Sankoff D (2018). Pinning down ploidy in paleopolyploid plants. *BMC genomics* 19: 287.
- [10] Vallée GC, Santos Muñoz D, Sankoff D (2016) Economic importance, taxonomic representation and scientific priority as drivers of genome sequencing projects, *BMC Genomics* 17, (Suppl 10):782.
- [11] Lyons E and Freeling M (2008) How to usefully compare homologous plant genes and chromosomes as DNA sequences. *The Plant Journal* 53: 661–673.
- [12] Lyons E, Pedersen B, Kane J, Alam M, Ming R, Tang H, Wang X, Bowers J, Paterson A, Lisch D and Freeling M (2008) Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar and grape: CoGe with rosids. *Plant Physiology* 148: 1772–1781.
- [13] <http://timetree.org>
- [14] Stevens PF (2017) Angiosperm Phylogeny Website. Version 14. <http://www.mobot.org/MOBOT/research/APweb/>
- [15] Kim YM, Kim S, Koo N, Shin AY, Yeom SI, Seo E, Park SJ, Kang WH, Kim MS, Park J, Jang I, Kim PG, Byeon I, Kim MS, Choi J, Ko G, Hwang J, Yang TJ, Choi SB, Lee JM, Lim KB, Lee J, Choi IY, Park BS, Kwon SY, Choi D, Kim RW (2017) Genome analysis of *Hibiscus syriacus* provides insights of polyploidization and indeterminate flowering in woody plants. *DNA Research* 24:71–80.
- [16] Li F, Fan G, Lu C, Xiao G, Zou C, Kohel RJ, Ma Z, Shang H, Ma X, Wu J, Liang X, Huang G, Percy RG, Liu K, Yang W, Chen W, Du X, Shi C, Yuan Y, Ye W, Liu X, Zhang X, Liu W, Wei H, Wei S, Huang G, Zhang X, Zhu S, Zhang H, Sun F, Wang X, Liang J, Wang J, He Q, Huang L, Wang J, Cui J, Song G, Wang K, Xu X, Yu JZ, Zhu Y, Yu S (2015) Genome sequence of cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. *Nature Biotechnology* 33:524–30.
- [17] Grover CE, Arick MA, Conover JL, et al. (2017) Comparative genomics of an unusual biogeographic disjunction in the cotton tribe (Gossypieae) yields insights into genome downsizing. *Genome Biology and Evolution* 9:3328–3344. doi:10.1093/gbe/evx248.
- [18] Sarkar D, Mahato AK, Satya P, Kundu A, Singh S, Jayaswal PK, Singh A, Bahadur K, Pattnaik S, Singh N, Chakraborty A, Mandal NA, Das D, Basu T, Sevanti AM, Saha D, Datta S, Kar CS, Mitra J, Datta K, Karmakar PG, Sharma TR, Mohapatra T, Singh NK (2017) The draft genome of *Corchorus olitorius* cv. JRO-524 (Navin). *Genome Data* 12:151–154.
- [19] Islam MS, Saito JA, Emdad EM, Ahmed B, Islam MM, Halim A, Hossen QMM, Hossain MZ, Ahmed R, Hossain MS, Kabir SMT, Khan MSA, Khan MM, Hasan R, Aktar N, Honi U, Islam R, Rashid MM, Wan X, Hou S, Haque T, Azam MS, Moosa MM, Elias SM, Hasan AMM, Mahmood N, Shafiuddin M, Shahid S, Shommu NS, Jahan S, Roy S, Chowdhury A, Akhand AI, Nisho GM, Uddin KS, Rabeya T, Hoque SME, Snigdha AR, Mortoza S, Matin SA, Islam MK, Lashkar MZH, Zaman M, Yuryev A, Uddin MK, Rahman MS, Haque MS, Alam MM, Khan H, Alam M (2017) Comparative genomics of two jute species and insight into fibre biogenesis. *Nature Plants* 3: 16223.
- [20] Baraloto C, Hardy OJ, Paine CET, Dexter KG, Craud C, Dunning LT, Gonzalez MA, Molino JF, Sabatier D, Savolainen V a Chave J (2012) Using functional traits and phylogenetic trees to examine the assembly of tropical tree communities. *Journal of Ecology* 100: 690–701.
- [21] Argout X, Salse J, Aury JM, Guiltinan MJ, Droc G, Gouzy J, Allegre M, Chaparro C, Legavre T, Maximova SN, Abrouk M, Murat F, Fouet O, Poulain J, Ruiz M, Roguet Y, Rodier-Goud M, Barbosa-

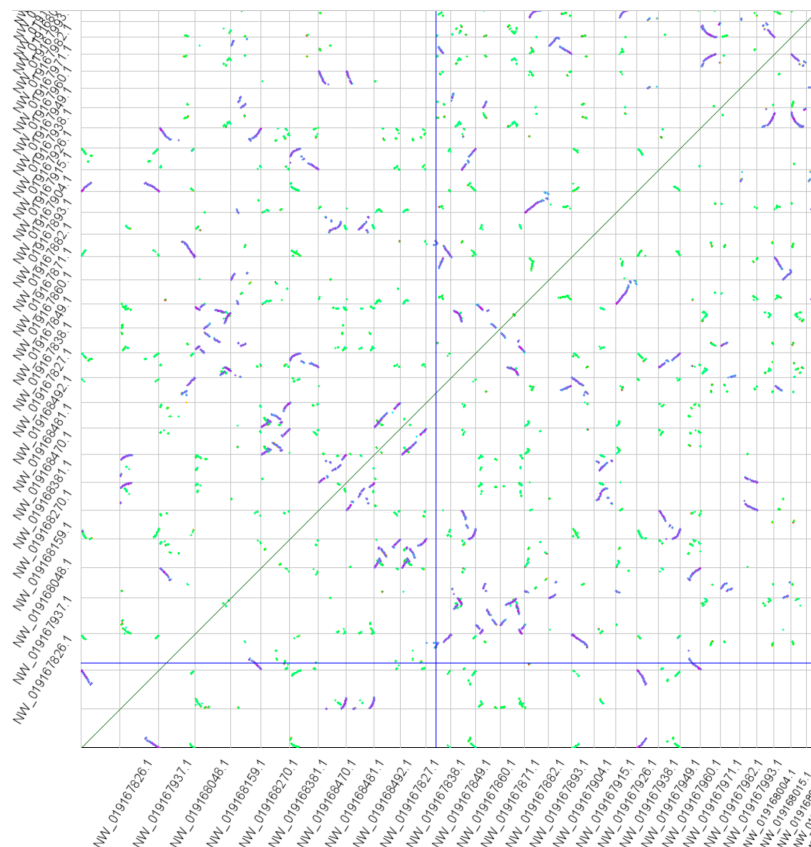
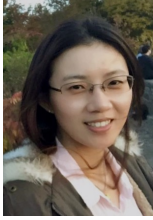


Fig. 6. Durio synteny blocks showing clear pattern of recent triplication (blue dots) covering most of the genome, as well as the γ triplication (green dots).

- Neto JF, Sabot F, Kudrna D, Ammiraju JS, Schuster SC, Carlson JE, Sallet E, Schiex T, Dievart A, Kramer M, Gelley L, Shi Z, Bérard A, Viot C, Boccara M, Risterucci AM, Guignon V, Sabau X, Axtell MJ, Ma Z, Zhang Y, Brown S, Bourge M, Golser W, Song X, Clement D, Rivallan R, Tahi M, Akaza JM, Pitollat B, Gramacho K, D'Hont A, Brunel D, Infante D, Kebe I, Costet P, Wing R, McCombie WR, Guiderdoni E, Quétier F, Panaud O, Wincker P, Bocs S, Lanaud C (2011) The genome of *Theobroma cacao* Nature Genetics 43:101–8.
- [22] NCBI (2017) *Herrania umbratica* Annotation Release 100. https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Herrania_umbratica/100/
- [23] Teh BT, Lim K, Yong CH, Ng CCY, Rao SR, Rajasegaran V, Lim WK, Ong CK, Chan K, Cheng VKY, Soh PS, Swarup S, Rozen SG, Nagarajan N, Tan P (2017) The draft genome of tropical fruit durian (*Durio zibethinus*) Nature Genetics 49, 1633–164.
- [24] Chen CH, Kuo TC, Yang MH, Chien TY, Chu MJ, Huang LC, Chen CY, Lo HF, Jeng ST, Chen LF (2014) Identification of cucurbitacins and assembly of a draft genome for *Aquilaria agallocha* BMC Genomics 15:578.
- [25] Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyère C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V, Felice N, Paillard S, Juman I, Moroldo M, Scalabrin S, Canaguier A, Le Clainche I, Malacrida G, Durand E, Pesole G, Laucou V, Chatelet P, Merdinoglu D, Delledonne M, Pezzotti M, Lecharny A, Scarpelli C, Artiguenave F, Pè ME, Valle G, Morgante M, Caboche M, Adam-Blondon AF, Weissenbach J, Quétier F, Wincker P and French-Italian Public Consortium for Grapevine Genome Characterization (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449: 463–467.
- [26] Zheng C, Chen E, Albert VA, Lyons E, Sankoff D (2013) Ancient eudicot hexaploidy meets ancestral eucosid gene order. BMC Genomics 14, S7:S3.
- [27] McLachlan GJ, Peel D, Basford KE, Adams P (1999) The EMMIX software for the fitting of mixtures of normal and t-components. Journal of Statistical Software 4, No. 2.
- [28] Sun M, Naeem R, Su JX, Cao ZY, Burleigh JG, Soltis PS, Soltis DE, Chen ZD (2016) Phylogeny of the Rosidae: A dense taxon sampling analysis. J. Syst. Evol. 54: 363–391.
- [29] Richardson JE, Whitlock BA, Meerow, AW, Madriñán S. (2015) The age of chocolate: A diversification history of *Theobroma* and Malvaceae. Front. Ecol. Evol. 3: 120. doi: 10.3389/fevo.2015.00120
- [30] Albalat R, Cañestro C (2016) Evolution by gene loss. Nature Reviews Genetics 17: 379–391.
- [31] Wang J, Yuan J, Yu J, Meng F, Sun P, Li Y, Yang N, Wang Z, Pan Y, Ge W, Wang L, Li J, Liu C, Xi Z, Zhao Y, Luo S, Ge D, Cui X, Feng G, Wang Z, Ji L, Qin J, Li XQ, Wang X (2019) Recursive paleohexaploidization shapes the durian genome. Plant Physiology. doi: 10.1104/pp.18.00921.



Yue Zhang 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 at the University of Ottawa, focusing on the statistical analysis of subgenome evolution after paleopolyploidy.



Yong-Min Kim Yong-Min Kim studied at Seoul National University and received the PhD degree from Chonnam National University. 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



Chunfang Zheng Chunfang Zheng received the master's and PhD degree 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 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 at 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.



Sindeed Islam Sindeed Islam is an undergraduate student at the University of Ottawa, and has been a volunteer student and now a research assistant in the Sankoff lab.

Chapter 6

Distinguishing successive ancient polyploidy levels based on genome-internal syntenic alignment

In Chapter 6 we set out a definitive triples-based framework for establishing the ploidies of a series of polyploidization events - are they WGD, WGT or other? This kind of approach, as well as reasonable but not model-based solutions, appear in Chapter 3 as well as other papers from the Sankoff lab[[bioRxiv654285](#)], but here we develop, for the first time, a model-based (branching process) approach to predicting the number of triples according to their three pairwise similarities. Profiles of triple frequency are derived for two and for three successive events, for all combinations of WGD and WGT, and are successfully applied to three genomes.

RESEARCH

Distinguishing successive ancient polyploidy levels based on genome-internal syntenic alignment

Yue Zhang, Chunfang Zheng and David Sankoff*

* Correspondence:
sankoff@uottawa.ca
Department of Mathematics and
Statistics, University of Ottawa,
150 Louis Pasteur pvt, Ottawa,
Canada K1N 6N5
Full list of author information is
available at the end of the article

Abstract

Background: A basic tool for studying the polyploidization history of a genome, especially in plants, is the distribution of duplicate gene similarities in syntenically aligned regions of a genome. This distribution can usually be decomposed into two or more components identifiable by peaks, or local maxima, each representing a different polyploidization event. The distributions may be generated by means of a discrete time branching process, followed by a sequence divergence model. The branching process, as well as the inference of fractionation rates based on it, requires knowledge of the ploidy level of each event, which cannot be directly inferred from the pair similarity distribution.

Methods: For a sequence of two events of unknown ploidy, either tetraploid, giving rise to whole genome doubling (WGD), or hexaploid, giving rise to whole genome tripling (WGT), we base our analysis on triples of similar genes. We calculate the probability of the four triplet types with origins in one or the other event, or both, and impose a mutational model so that the distribution resembles the original data. Using a ML transition point in the similarities between the two events as an discriminator for the hypothesized origin of each similarity, we calculate the predicted number of triplets of each type for each model combining WGT and/or WGD. This yields a predicted profile of triplet types for each model.

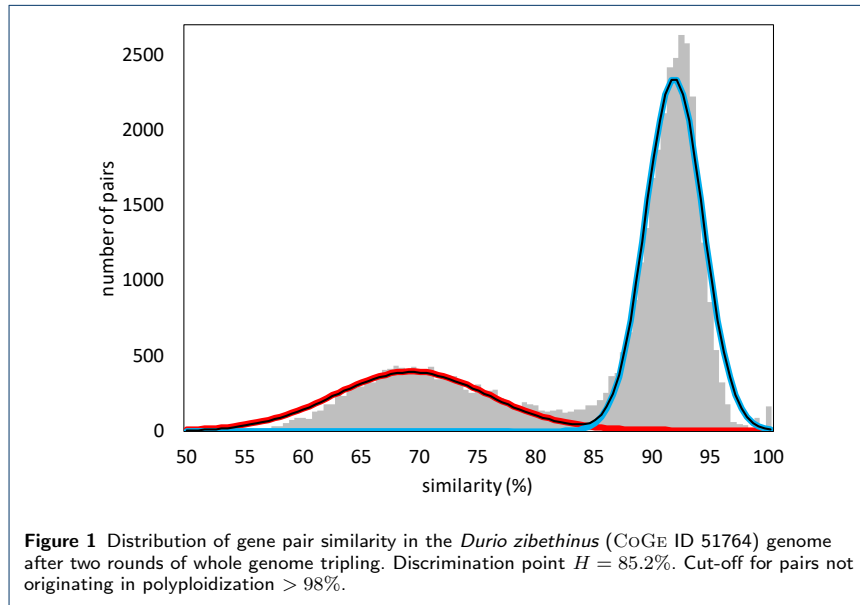
Results: We compare the observed and predicted triplet profiles for each model to confirm the polyploidization history of durian, poplar and cabbage.

Keywords: polyploidy; whole genome doubling; whole genome tripling; gene triples; plant genomes; branching process

Background

Given the pervasiveness of whole genome doubling (WGD) and tripling (WGT) in the ancestral lineages of plant species, a widespread feature of plant genome publications is the display of the distribution of duplicate gene identities (or similarities, distances, $K_s, \dots$). This is illustrated in Figure 1, which represents the distribution of similarities between syntenically aligned duplicate genes [1, 2] in the durian (*Durio zibethinus*) genome [3]. The two visually distinguishable but overlapping parts of the distribution are the legacy of two successive polyploidization events. The means (t_1 and t_2), variances and proportion of the total sample of each component of the distribution can be estimated by mixtures of models techniques such as EMMIX [4].

These distributions can be explained and generated by a discrete-time branching process model of polyploidization and fractionation (not time-homogeneous), mathematically represented by the product of successive r -nomial distributions,

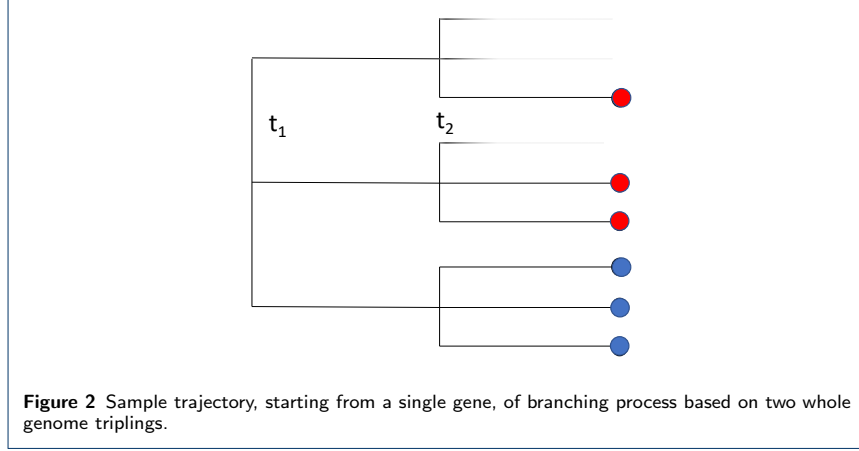


the output of one being the input of the next, where the parameters, namely the r -nomial probabilities, express resistance to gene loss through fractionation [5–9].

These parameters, which are key to understanding the cycle of polyploidization and fractionation, can be estimated using the information inferred from the distribution of gene pair similarities.

One aspect of the branching process that cannot be inferred from the study of *pairs* only is the ploidy r of the various events. Thus, durian is known to have undergone two whole genome triplings [9, 10], the γ tripling almost 120 million years ago [11], ancestral to most flowering plants, and a more recent tripling (10-20 Mya) not shared by even closely related species, like cacao. By just looking at the distribution of similarities in Figure 1 engendered by these events, however, there is no direct way of knowing whether one or both of the two component distributions represent WGD, WGT, or other polyploid events. To resolve this problem, the main point of this presentation, we propose to add *triples* of similar ($> 50\%$) genes to the study of gene pairs. Our technique, based on the branching process, responds to our concern in previous *ad hoc* treatments [7, 12] of how to use triples rigorously from a statistical point of view.

In the next section, we summarize the general branching process approach to analyzing the distribution of gene pair similarities. We then focus on four competing two-event models involving WGD and/or WGT. We define four types of gene triplet according to whether the gene pairs within them were created by the first event, the second event, or both. Within each model, we calculate the expected number of triplets of each type. Thus creates an “underlying” profile of triplet distribution to compare to the “observed” profile of triplets in the data. Because of the way the two components of the pair similarity distribution overlap, however, the origin of each triplet in the data is not always obvious. Thus we create a “predicted” profile of triplet distribution by grafting a paralog divergence model onto the branching



process, making use of a maximum likelihood dividing point between the two components. We apply this analysis to the genomes of durian, poplar *Populus trichocarpa* [13] and cabbage *Brassica oleracea* [15], each of which has a different sequence of polyploidization events. These histories are captured correctly for the first two, but the results for *B. oleracea* prompt an extension to three-event models, which we carry out, and suggest further work to higher numbers of events.

The branching process and two-event models

Denote by m_i the total number of individuals (genes) at time $t_i, i = 1, \dots, n$. Set $m_1 = 1$. At time $t_i, i = 1, \dots, n-1$, each of the m_i genes is replaced by $r_i \geq 2$ progeny, but only $j \geq 1$ of them survive until time t_{i+1} , with probability $u_j^{(i)}$.

Of the total of m_i genes at time t_i , let $a_j^{(i)}$ be the number for which j progeny survive until time t_{i+1} , so that

$$m_i = \sum_{j=1}^{r_i} a_j^{(i)}, \quad m_{i+1} = \sum_{j=1}^{r_i} j a_j^{(i)}. \quad (1)$$

The probability distribution of the evolutionary histories represented by the given $\mathbf{r} = \{r_i\}_{i=1}^{n-1}$ and the variable $\mathbf{a} = \{a_j^{(i)}\}_{j=1 \dots r_i}^{i=1 \dots n-1}$ is

$$P(\mathbf{r}; \mathbf{a}) = \prod_{i=1}^{n-1} \left[\binom{m_i}{a_1^{(i)}, \dots, a_{r_i}^{(i)}} \prod_{j=1}^{r_i} (u_j^{(i)})^{a_j^{(i)}} \right]. \quad (2)$$

The expected number of genes at time t_n is

$$\mathbf{E}(m_n) = \sum_{\mathbf{a}} P(\mathbf{r}; \mathbf{a}) m_n. \quad (3)$$

This is illustrated by the sample trajectory in Figure 2, in which a WGT at time t_1 , with all three progeny surviving – the 3-nomial sample has value 3 – is followed by another independent WGT at time t_2 where the three lineages show one, two or all three offspring surviving, i.e., the independent 3-nomials samples have values

1,2 and 3, respectively. We will study the case of two successive polyploidy events, with r_1 and r_2 taking on values 2 or 3, i.e., WGD or WGT, in all four combinations, i.e., in the set $\mathcal{M}$ of “models”, denoted $(3, 3)$, $(3, 2)$, $(2, 3)$ and $(2, 2)$. Because of the limited information that can be inferred about each component of the distribution of similarities, we can only infer the probabilities of samples of value 1, 2, or 3, so that we are limited to 2-nomials and, with some assumptions, 3-nomials, by far the biologically most important cases.

To infer parameters like fractionation rate in the polyploidization history of a genome, based on the distribution of gene pair similarities, we need to know the ploidies r_i of the various events. This motivates us to extend our study from gene pairs only to also include gene triplets.

Methods

Triplet probabilities in four models

With no loss of generality, we study triplets of similarities among three genes, rather than the triplets of genes themselves. A triplet is a (multi-)set $\{t_i, t_j, t_k\}$, where each of i, j and k may be 1 or 2. Let $\mathcal{T} = \{\{t_1, t_1, t_1\}, \{t_1, t_1, t_2\}, \{t_1, t_2, t_2\}, \{t_2, t_2, t_2\}\}$. We classify each kind of triplet we according to whether each of the three paralogies among the three pairs of gene originates in the first event or the second event. Thus in the branching process illustration in Figure 2 the blue dots represent genes that form a $\{t_2, t_2, t_2\}$ triplet and the red dots form a $\{t_1, t_1, t_2\}$ triplet. The single red dot combines with the three pairs of blue genes to form three additional $\{t_1, t_1, t_2\}$ triplets. And there are a further nine $\{t_1, t_1, t_2\}$ triplets in the sample. We can calculate the expected number of triplets of each type by enumerating the triplets of each type in each possible trajectory of the process, and multiplying by the probability of this trajectory from Expression (2). The enumeration within a trajectory is easily done by considering every triple of genes at time t_n and identifying the last common ancestor of each pair. The probabilities of the trajectories have previously been calculated [5–7]. We then sum these results over all trajectories. These are listed in Table 1.

model $M \in \mathcal{M}$	tripling- tripling (3,3)	tripling- doubling (3,2)	doubling- tripling (2,3)	doubling- doubling (2,2)
$\{t_1, t_1, t_1\}$	$u'(1 + 2v' + v)^3$	$u'(1 + v)^3$	-	-
$\{t_1, t_1, t_2\}$	$2(3u' + u)(3v' + v) \times (1 + 2v' + v)$	$2(3u' + u) \times v(1 + v)$	$2u(3v' + v) \times (1 + 2v' + v)$	$(2u)v(1 + v)$
$\{t_1, t_2, t_2\}$	-	-	-	-
$\{t_2, t_2, t_2\}$	$(1 + 2u' + u)v'$	-	$(1 + u)v'$	-

Table 1 Formulae for the expected numbers of triplets $W_M(\Delta)$ of each type Δ , by branching model M . u = probability that two progeny survive after the first polyploidization event. u' = probability that three survive. Similarly v and v' are the probabilities that two or three progeny survive, respectively, after the second event.

This table provides the expected number of triplets $W_M(\Delta)$ of each type $\Delta \in \mathcal{T}$ produced by the branching process for each model $M \in \mathcal{M}$. For a given genome, the four numerical values of $W_M(\cdot)$ constitute the *underlying profile* of the model M . The underlying profiles for each model based on maximum likelihood values of u and v are given in the top half of Tables 2, 3, and 6 below. Because of the limited

number of parameters that can be inferred from the distribution of similarities, we assume $u' = u^2$ and $v' = v^2$.

It can be seen in Table 1 that the profiles of triplet types produced by the different models are very different. If we could observe the triplet profile produced by the branching process underlying a given set of data, we could easily identify which model was responsible. However, we only see the data after mutational processes have applied. When a mutational divergence model is applied to the similarities, a single trajectory of the branching process, producing one ideal type of triplet, can produce many very different data triplets.

We can try to categorize the set of triplets in a set of data by how closely they resemble one of the four basic types. If the two components of the similarity distribution were completely separate, this would also be an easy matter. But the usual large overlap between the components means that we cannot automatically ascribe any data triplet to any particular underlying triplet.

A statistical approach

As a solution to this problem, we first try to find the best transition, or cutoff point H somewhere between the peaks of the two components. For this we compute the product of the probability density at each similarity value less than H , according to the component with mean at t_1 , and the density at each similarity value greater than H , according to the component with mean at t_2 , and maximize with respect to H . I.e.,

$$H = \max_{h \in (0,1)} \prod_{x \leq h} \frac{1}{\sqrt{\sigma_1^2}} \exp\left[-\frac{(x - \mu_1)^2}{2\sigma_1^2}\right] \prod_{x > h} \frac{1}{\sqrt{\sigma_2^2}} \exp\left[-\frac{(x - \mu_2)^2}{2\sigma_2^2}\right] \quad (4)$$

We then categorize the triplets in the data according to the transition value H . If a similarity x is less than H we classify it as being produced at time t_1 , and we write $x \in I$. If it is greater than H , we classify it as being created at time t_2 and we write $x \in J$. This creates eight “octants”, defined by the $8 = 2^3$ combination of the three triplet similarities, which in turn are collapsed into the four types of triplet in $\mathcal{T}$ tabulated in Table 1, $\{t_1, t_1, t_1\}$, $\{t_1, t_1, t_2\}$ (representing three octants), $\{t_1, t_2, t_2\}$ (representing three octants) and $\{t_2, t_2, t_2\}$. The number of triplets in of the four types we call the *observed profile*.

Although we can compare the observed profile with the underlying profile, this comparison is not too meaningful since it neglects the fact that many of the data triplets classified as one type may be generated by a different underlying type, not as an error, but simply as a result of the normal process of duplicate gene sequence divergence clearly operative in the more or less dispersed and overlapping components of the distribution of gene pair similarities.

We can, however, take this process into account in producing a *predicted profile* for each model. We first calculate the variance-covariance matrix Σ of the t_1 similarities in triplets containing at least of them and t_2 similarities in triplets containing at least two of these. We fixed $\text{covar}(t_1, t_2) = 0$, in accordance with the Markov nature of the branching process.

For each model $M \in \mathcal{M}$, we construct the *predicted profile* of triplet types by integrating over the trivariate normal with means drawn from the EMMIX analysis

or identified by eye with the distribution component peaks, and covariance estimated as above, restricted to the domains defined by the transition point. Thus our prediction of $\{t_1, t_1, t_1\}$ triplets would involve a restriction to the domain (I, I, I) where all three coordinates are less than or equal to H . Our prediction of apparently $\{t_1, t_1, t_2\}$ triplets would be confined to the three octants where two coordinates are less than or equal to H and one is greater. The integrals are weighted by $W(\Delta)$, the expected number of triplets. For example, the predicted number of $\{t_1, t_1, t_2\}$ triplets would be;

$$P_M = \sum_{\Delta \in \mathcal{T}} W_M(\Delta) \int_{(I,I,J) \cup (I,J,I) \cup (J,I,I)} N(\mathbf{X}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{X}, \quad (5)$$

where $\boldsymbol{\mu}$ is the vector of component means. To summarize, we have defined three types of triplet profile:

- the observed profile, based on triples of genes all having high similarity scores with each other, which can be compiled from the list of gene pairs produced by the SYNMAP function of CoGe [1, 2],
- the underlying profile for each model $M \in \mathcal{M}$, produced by the combinatorial probabilities of the branching process for each model $M \in \mathcal{M}$,
- the predicted profile for each model $M \in \mathcal{M}$, produced by grafting a gene pair divergence model on the underlying profile.

For comparative purposes we normalize the underlying and predicted profiles so that the total number of triples is the same as the observed profiles.

Results

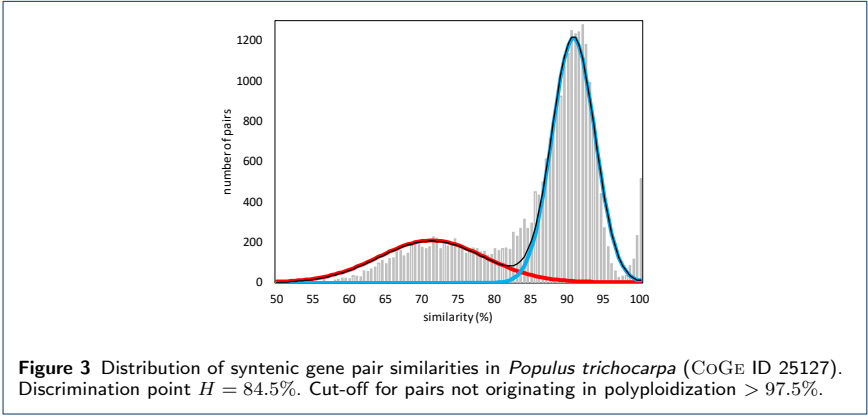
We compare the three profiles for three well-studied flowering plant genomes that are known to have undergone multiple polyploidizations in the last 120 million years, to see if our method predicts the right combination of WGT and WGD.

model $M \in \mathcal{M}$	tripling- tripling (3,3)	tripling- doubling (3,2)	doubling- tripling (2,3)	doubling- doubling (2,2)	ob- served
triplet $\Delta \in \mathcal{T}$					
AFTER FRACTIONATION ONLY					
$\{t_1, t_1, t_1\}$	1536	2027	0	0	1124
$\{t_1, t_1, t_2\}$	10734	13466	11942	15493	9368
$\{t_1, t_2, t_2\}$	0	0	0	0	212
$\{t_2, t_2, t_2\}$	3224	0	3551	0	4789
AFTER FRACTIONATION + MUTATION					
$\{t_1, t_1, t_1\}$	1534	2022	33	42	1124
$\{t_1, t_1, t_2\}$	10572	13262	11734	15219	9368
$\{t_1, t_2, t_2\}$	167	184	180	204	212
$\{t_2, t_2, t_2\}$	3220	25	3545	27	4789

Table 2 Durian model predictions before (underlying) and after imposition of mutational divergence. Shaded column indicates the model predicted by the literature and the closest fit to the observed profile.

Durian

Starting with the durian genome, the (3,3) model, known to represent true evolutionary history, is the only one with a credible prediction profile in Table 2, the only one that has reasonable values for all four triplet types. The three others all



fail to predict one or both of the $\{t_1, t_1, t_1\}$ and $\{t_2, t_2, t_2\}$ triplets. This indicates of the potential of our statistical method, since the original durian sequence article [3] did not recognize the second event as a tripling.

Poplar

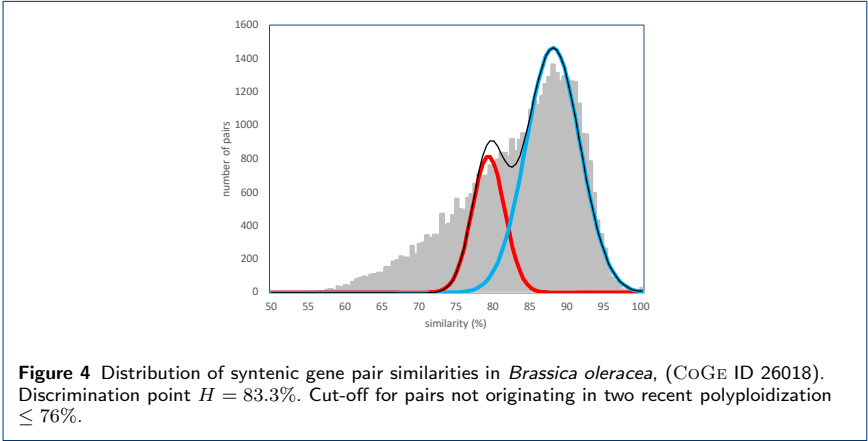
The predicted profile of the (3,2) model in Table 3 summarizes the true history of the *Populus trichocarpa* genome whose gene pair similarity distribution is displayed in Figure 3. Along with γ , this shares the ancient “salicoid” WGD with other members of the Salicaceae family [14]. (3,2) is the only model that correctly identifies both the γ event as a WGT, and the more recent event as a WGD.

model $M \in \mathcal{M}$	tripling- tripling (3,3)	tripling- doubling (3,2)	doubling- tripling (2,3)	doubling- doubling (2,2)	ob- served
AFTER FRACTIONATION ONLY					
$\{t_1, t_1, t_1\}$	300	529	0	0	390
$\{t_1, t_1, t_2\}$	1725	2817	1991	3346	2877
$\{t_1, t_2, t_2\}$	0	0	0	0	34
$\{t_2, t_2, t_2\}$	1321	0	1355	0	45
AFTER FRACTIONATION + MUTATION					
$\{t_1, t_1, t_1\}$	293	514	25	41	390
$\{t_1, t_1, t_2\}$	1611	2634	1832	3079	2877
$\{t_1, t_2, t_2\}$	172	179	186	204	34
$\{t_2, t_2, t_2\}$	1270	19	1303	21	45

Table 3 *Populus* model predictions before (underlying) and after imposition of mutational divergence. Shaded column indicates the model predicted by the literature and the closest fit to the observed profile.

Cabbage

The recent ancestor of *Brassica oleracea* genome underwent a WGT that gave rise not only to the crucifers and mustard genera, but also radishes and other related genera. Early than that a WGD called the α doubling is apparent in the whole range of family Brassicaceae genera, including *Arabidopsis*. A still earlier WGD, the β doubling, occurred in the order Brassicales lineage that includes the Brassicaceae. Thus the cabbage genome counts γ, β, α and a Brassica WGT in its evolutionary history [16, 17]. In Figure 4, we see that at least the two recent components are



clearly distinguishable, so we first carried out an analysis excluding gene pairs of similarity less than 76%.

model $M \in \mathcal{M}$	triplication-triplication (3,3)	triplication-duplication (3,2)	duplication-triplication (2,3)	duplication-duplication (2,2)	observed
triplet $\Delta \in \mathcal{T}$					
AFTER FRACTIONATION ONLY					
$\{t_1, t_1, t_1\}$	586	1337	0	0	508
$\{t_1, t_1, t_2\}$	3509	7999	3744	9336	3277
$\{t_1, t_2, t_2\}$	0	0	0	0	1379
$\{t_2, t_2, t_2\}$	5241	0	5592	0	4172
AFTER FRACTIONATION + MUTATION					
$\{t_1, t_1, t_1\}$	654	1477	137	325	508
$\{t_1, t_1, t_2\}$	3249	7271	3411	8347	3277
$\{t_1, t_2, t_2\}$	690	542	728	613	1379
$\{t_2, t_2, t_2\}$	4743	46	5060	51	4172

Table 4 *Brassica oleracea* model predictions before (underlying) and after imposition of mutational divergence.

This analysis in Table does not give satisfactory results. Indeed, the (3,3) profile matches the observed profile much better than the (2,3) profile does. This can be partially attributed to substantial number of similarities generated by the β and even γ doublings greater than 76%.

We can partially correct this by adding a third event to our branching process. This leads to eight models instead of four, and ten kinds of triplet, summarized in Table 5.

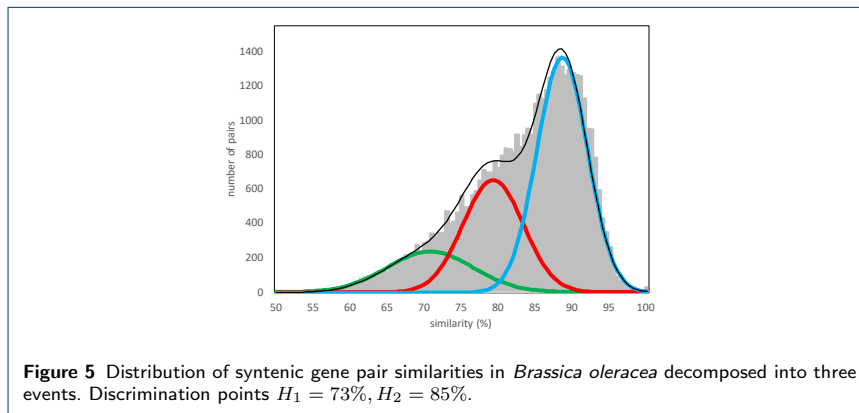
model $M \in \mathcal{M}$ triplet $\Delta \in \mathcal{T}$	(2,2,2)	(3,2,2)	(2,3,2)
$\{t1, t1, t1\}$	-	$u'(1+v)^3(1+w)^3$	-
$\{t1, t1, t2\}$	$2uv(1+v)(1+w)^3$	$2(3u' + u)v(1+v)(1+w)^3$	$2u(3v' + v) \times (1 + 2v' + v)(1 + w)^3$
$\{t1, t2, t2\}$	-	-	-
$\{t2, t2, t2\}$	-	-	$(1 + u)v'(1 + w)^3$
$\{t2, t2, t3\}$	$2(1 + u)vw(1 + w)$	$2(1 + 2u' + u)vw(1 + w)$	$2(1 + u)(3v' + v)w(1 + w)$
$\{t2, t3, t3\}$	-	-	-
$\{t3, t3, t3\}$	-	-	-
$\{t1, t1, t3\}$	$2u(1 + v)^2w(1 + w)$	$2(3u' + u)(1 + v)^2w(1 + w)$	$2u(1 + 2v' + v)^2w(1 + w)$
$\{t1, t3, t3\}$	-	-	-
$\{t1, t2, t3\}$	-	-	-
model $M \in \mathcal{M}$ triplet $\Delta \in \mathcal{T}$	(2,2,3)	(2,3,3)	(3,2,3)
$\{t1, t1, t1\}$	-	-	$u'(1+v)^3$
$\{t1, t1, t2\}$	$2uv(1+v) \times (1 + 2w' + w)^3$	$2u(3v' + v)(1 + 2v' + v) \times (1 + 2w' + w)^3$	$2(3u' + u)v(1 + v) \times (1 + 2w' + w)^3$
$\{t1, t2, t2\}$	-	-	-
$\{t2, t2, t2\}$	-	$(1 + u)v' \times (1 + 2w' + w)^3$	-
$\{t2, t2, t3\}$	$2(1 + u)v \times (3w' + w)(1 + 2w' + w)$	$2(1 + u)(3v' + v) \times (3w' + w)(1 + 2w' + w)$	$2(1 + 2u' + u)v \times (3w' + w)(1 + 2w' + w)$
$\{t2, t3, t3\}$	-	-	-
$\{t3, t3, t3\}$	$(1 + u)(1 + v)w'$	$(1 + u)(1 + 2v' + v)w'$	$(1 + 2u' + u)(1 + v)w'$
$\{t1, t1, t3\}$	$2u(1 + v)^2 \times (3w' + w)(1 + 2w' + w)$	$2u(1 + 2v' + v)^2 \times (3w' + w)(1 + 2w' + w)$	$2(3u' + u)(1 + v)^2 \times (3w' + w)(1 + 2w' + w)$
$\{t1, t3, t3\}$	-	-	-
$\{t1, t2, t3\}$	-	-	-
model $M \in \mathcal{M}$ triplet $\Delta \in \mathcal{T}$	(3,3,2)	(3,3,3)	
$\{t1, t1, t1\}$	$u'(1 + 2v' + v)^3(1 + w)^3$	$u'(1 + 2v' + v)^3(1 + 2w' + w)^3$	
$\{t1, t1, t2\}$	$2(3u' + u)(3v' + v)(1 + 2v' + v)(1 + w)^3$	$2(3u' + u)(3v' + v)(1 + 2v' + v)(1 + 2w' + w)^3$	
$\{t1, t2, t2\}$	-	-	
$\{t2, t2, t2\}$	$(1 + 2u' + u)v'(1 + w)^3$	$(1 + 2u' + u)v'(1 + 2w' + w)^3$	
$\{t2, t2, t3\}$	$2(1 + 2u' + u)(3v' + v)w(1 + w)$	$2(1 + 2u' + u)(3v' + v)(3w' + w)(1 + 2w' + w)$	
$\{t2, t3, t3\}$	-	-	
$\{t3, t3, t3\}$	-	$(1 + 2u' + u)(1 + 2v' + v)w'$	
$\{t1, t1, t3\}$	$2(3u' + u)(1 + 2v' + v)^2w(1 + w)$	$2(3u' + u)(1 + 2v' + v)^2(3w' + w)(1 + 2w' + w)$	
$\{t1, t3, t3\}$	-	-	
$\{t1, t2, t3\}$	-	-	

Table 5 Formulae for the expected numbers of triplets after three events.

We fix the mean of the first component at 71% to account for the γ event, that for the second component, representing the α event, at 79.5 % and we find two ML discrimination points, as in Figure 5.

The results of this are shown in Table 6. Here the (3,2,3) model is just as close to the observed profile than the competing (3,3,3) model, with the notable exception of $\{t2, t2, t2\}$ triplets. The absence of a distinction between the α and β events means that the similarities they generate are all conflated to yield an excess of $t2$, and consequently an excess of $t2$ triples, so that a WGT is inferred rather than a WGD.

The obvious remedy for this would be to construct four-event models (sixteen of them), with profiles consisting of 20 different triplets. We leave this for further work. In general the number of models is exponential: 2^m for m events, while the



number of triples follows the polynomial (i.e., cubic) tetrahedral sequence (A00292 in [18]) $\frac{1}{6}m(m+1)(m+2)$, so that eventually there would not be enough data to discriminate among the models.

Conclusions

We model the process of fractionation to account for the distribution of gene pair similarities after a number of whole genome doublings, triplings, etc., each followed by a period of duplicate gene loss. The model is a discrete-time branching process, with synchronous birth number $r_i \geq 2$ across the i -th generation population and deaths determined by a r_i -nomial law conditioned on at least one survivor.

The observations of gene pair similarities consist of a mixture of normals, each component generated by one event, with the event time estimated by the sequence divergence from the event to the present. Despite the overlapping distributions, we can estimate the mean (*via* a local mode), standard deviation and proportion of the sample.

Statistics on gene pairs alone do not allow us to infer r_i , so we introduce the study of gene triplets. We find formulae for the expected number of each kind of triplet, categorized as to which events produced the similarities among the three pairs of genes.

We develop a way of grafting a gene divergence model on this underlying profile of triplets to produce a predicted profile of the number of triplets of each kind. This can then be observed with the observed number of triplets.

Further work

Distinguishing among the four models combining tetraploidization or hexaploidization in two successive events is the simplest example of a more general problem. The theoretical way is clear to extending these ideas to include, for example, octoploidization through the extraction of quadruplets instead of triplets. In addition, there is a straight forward extension to the case of three or more successive polyploidization events, which we have undertaken in the study of *Brassica oleracea*. Here, three events, three normal components and two transition points are estimated from the distribution of similarities. The combinatorial probabilities have

model $M \in \mathcal{M}$								obs- erved
triplet $\Delta \in \mathcal{T}$	(2,2,2)	(3,2,2)	(2,3,2)	(2,2,3)	(2,3,3)	(3,2,3)	(3,3,2)	(3,3,3)
AFTER FRACTIONATION ONLY								
$\{t_1, t_1, t_1\}$	0	613	0	0	0	423	537	386
$\{t_1, t_1, t_2\}$	2433	2333	2119	1667	5456	1620	2043	1471
$\{t_1, t_2, t_2\}$	0	0	0	0	0	0	0	0
$\{t_2, t_2, t_2\}$	0	0	1936	0	1606	0	1866	1335
$\{t_2, t_2, t_3\}$	8210	7875	7152	5627	4953	5468	6896	4966
$\{t_2, t_3, t_3\}$	0	0	0	0	0	0	0	0
$\{t_3, t_3, t_3\}$	0	0	0	4726	364	4593	0	4215
$\{t_1, t_1, t_3\}$	4379	4200	3814	3001	2642	2916	3678	2648
$\{t_1, t_3, t_3\}$	0	0	0	0	0	0	0	0
$\{t_1, t_2, t_3\}$	0	0	0	0	0	0	0	0
AFTER FRACTIONATION + MUTATION								
$\{t_1, t_1, t_1\}$	58	248	52	40	131	172	219	158
$\{t_1, t_1, t_2\}$	1221	1390	1091	837	2335	964	1243	895
$\{t_1, t_2, t_2\}$	1172	1264	1253	803	2279	877	1331	957
$\{t_2, t_2, t_2\}$	1224	1222	2303	882	2302	889	2262	1661
$\{t_2, t_2, t_3\}$	6070	5827	5638	4452	4060	4330	5441	4177
$\{t_2, t_3, t_3\}$	1080	1036	999	1913	795	1860	964	1740
$\{t_3, t_3, t_3\}$	93	89	86	3282	309	3189	83	2930
$\{t_1, t_1, t_3\}$	1787	1717	1557	1225	1232	1192	1504	1083
$\{t_1, t_3, t_3\}$	56	54	50	38	37	37	48	35
$\{t_1, t_2, t_3\}$	2261	2174	1992	1549	1542	1510	1926	1386

Table 6 *Brassica oleracea* three-event model predictions before (underlying) and after imposition of mutational divergence. The expected (3,2,3) profile in the red-shaded column fits the observed pattern as well as the (3,3,3) profile in the blue-shaded column.

been worked out for this case and many others, and the methodology is available to complete this.

It is true that in some cases, such as that we presented in [9], concerning *Durio zibethinus*, the ploidy is evident from the clear presence in the SYNMAP self-comparison dotplots of sets of r regions covering a large proportion of the genome, each set represented by exactly $\binom{r}{2}$ synteny blocks showing synteny among all r regions. ($r = 3$ in the case of *Durio*.) Clear cases like this are rare, however, especially for genomes where the last polyploidization is more remote in time.

In previous work [7], we used additional information, beyond that contained in the similarity distribution, to confirm the recent hexaploidization of *Brassica rapa* against the alternative of tetraploidization. This kind of data, however, namely speculation about the number of single-copy genes in the current genome, was extremely subjective in that report, and is unreliable even when assessed by the best available methods on well-assembled and annotated genomes.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

The study was planned by all three authors, who also carried out the research and wrote the paper. All authors read and approved the paper.

Acknowledgements

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

References

1. How to usefully compare homologous plant genes and chromosomes as DNA sequences (E Lyons, M Freeling), The Plant Journal 53, 661–673 (2008)
2. Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar and grape: CoGE with rosids (E Lyons, B Pedersen, J Kane, M Alam, R Ming, H Tang et al.), Plant Physiology 148, 1772–1781(2008)

3. The draft genome of tropical fruit durian (*Durio zibethinus*) (BT Teh, K Lim, CH Yong, CCY Ng, SR Rao *et al.*), *Nature Genetics* 49, 1633–164 (2017)
4. The EMMIX software for the fitting of mixtures of normal and t-components (GJ McLachlan, D Peel, KE Basford, P Adams), *Journal of Statistical Software* 4, 1–14 (1999)
5. Evolutionary model for the statistical divergence of paralogous and orthologous gene pairs generated by whole genome duplication and speciation (Y Zhang, C Zheng, D Sankoff), *IEEE/ACM Transactions in Computational Biology and Bioinformatics* 1545–5963 (2017)
6. Models for similarity distributions of syntenic homologs and applications to phylogenomics (D Sankoff, C Zheng, Y Zhang, J Meidanis, E Lyons, H Tang), *IEEE/ACM Transactions in Computational Biology and Bioinformatics* (2018), doi:10.1109/TCBB.2018.2849377.
7. Pinning down ploidy in paleopolyploid plants (Y Zhang, C Zheng, D Sankoff), *BMC Genomics* 19(Suppl 5), 287 (2018)
8. A branching process for homology distribution-based inference of polyploidy, speciation and loss (Y Zhang, C Zheng, D Sankoff), *Algorithms in Molecular Biology* (2019, in press)
9. Branching out to speciation in a model of fractionation: the Malvaceae (Y Zhang, C Zheng, S Islam, Y-M Kim, D Sankoff), *IEEE/ACM Transactions in Computational Biology and Bioinformatics* (2019, in press)
10. Recursive paleohexaploidization shapes the durian genome (J Wang, J Yuan, J Yu, F Meng, P Sun, Y Li *et al.*), *Plant Physiology*. doi: 10.1104/pp.18.00921 (2019)
11. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla (O Jaillon, JM Aury, B Noel, A Policriti, C Clepet, *et al.*) *Nature* 449, 463–467 (2007)
12. The avocado genome informs deep angiosperm phylogeny, highlights introgressive hybridization, and reveals pathogen-influenced gene space adaptation (M Rendón-Anaya, E Ibarra-Laclette, A Méndez Bravo, T Lan, C Zheng, L Carretero Paulet *et al.*), *PNAS* (2019, in press)
13. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray) GA Tuskan, S DiFazio, S Jansson, J Bohlmann, I Grigoriev *et al.*, *Science* 313, 1596–1604 (2006)
14. The willow genome and divergent evolution from poplar after the common genome duplication (X Dai, Q Hu, Q Cai, K Feng, N Ye, GA Tuskan *et al.*) *Cell Research* 24, 1274–1277 (2014)
15. The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes (S Liu, Y Liu, X Yang, C Tong, D Edwards, IAP Parkin *et al.*), *Nature Communications* 5, 3930 (2014)
16. Paleopolyploidy in the Brassicales: analyses of the Cleome transcriptome elucidate the history of genome duplications in *Arabidopsis* and other Brassicales (MS Barke, H Vogel, ME Schranz), *Genome Biology and Evolution* 1, 391–9 (2009) doi: 10.1093/gbe/evp040.
17. Brassicales phylogeny inferred from 72 plastid genes: A reanalysis of the phylogenetic localization of two paleopolyploid events and origin of novel chemical defenses (PP Edger, JC Hall, A Harkess, M Tang, J Coombs, S Mohammadin *et al.*), *American Journal of Botany* 105, 463–469 (2018)
18. The Online Encyclopedia of Integer Sequences, published electronically at <https://oeis.org> (NJA Sloane) accessed June 18 (2019)

Chapter 7

Conclusions

The work in this thesis fulfills the goal of understanding the mixture of gene pair similarity distributions in terms of a model of WGD or higher copy event, fractionation and speciation. The adoption of our methods, using either available combinatorial, algebraic and statistical resources or the eventual professional design and implementation of a dedicated pipeline, should preclude the type of unwarranted interpretation of these distributions frequently encountered in the genomics literature.

Equally important, this research opens the door to a number of new modeling problems as well as new approaches to comparative genomics data.

One fact that has become increasingly clear over the course of this research is that less information can be extracted from the distribution of similarities than was first thought. The location of the peaks with relation to each other and to the origin, the standard deviation of the distributions they subtend, and the volume or amplitude of these components, seemed to provide an ample number of statistics from which to infer parameters of the models. It was only by Chapters 4 and 5 that it was clear that our attempts to model the variance was counterproductive, and de-stabilizing

7. CONCLUSIONS

for inference procedures, since it is too closely related to the mean (the location) to provide any independent information. Aside from the location of the peaks the only statistics of any use are the relative volumes of the component distributions. Moreover, since these volumes must sum to one, the number of model parameters that can be estimated is one less than the number of peaks. Nevertheless, a number of techniques can compensate for this; in Chapter 3 we explored the possibility of incorporating the number of unpaired genes, though taking into account that this is an unreliable statistic, into the estimation. In several of the chapters, we assumed a fixed (well-motivated) relationship among parameters for single gene loss and double gene loss after WGT. In Chapter 5 we assumed fractionation rates were equal over after two events. And other techniques would be available, such as constraining certain parameters to be equal across several species having undergone the same WGD event.

Our models do not allow for the loss of all the offspring of a gene. We offered the biological argument that this reflected a biological assumption that all genes are essential, thus at least one copy must remain after fractionation. This, however, is too stringent an assumption, and a more cogent reason is simply that nothing is observable about genes lineages that peter out. Indeed our model could include the option that some or many genes have zero offspring surviving until the next event. However we could not estimate the relevant parameters since there is no data observable to estimate the pertinent probability. For any such model, by taking the probabilities conditioned on there being at least one survivor, we simply return to our current models. Further research could examine the case where several genomes descend from the same ancestor, and genes with no descendants in one genome could be “observed” via their presence in other genomes.

7. CONCLUSIONS

In the course of applying our models to real genomes, it has become clear that the components of similarity distributions often deviate from normality. This has consequences for model fitting. Further work on mutational models would be required to study this.

Our work on gene triples in Chapter 6 described a statistical analysis that enables us to identify the ploidy level of a series of two or three polyploidization events, not possible with gene pairs alone. This difficulty was illustrated by our earlier paper in Chapter 3 on the intersection of algebraic varieties representing different sequences of events. In the later paper, we also suggested generalizations and extensions in several directions. Simply increasing the number of events to four, and even five, would be useful as numerous genomes with this kind of history exist. Investigating quadruples as well as triples would be another interesting extension for these genomes.

This thesis being devoted to building and testing models, in a mathematical and statistical context, we did not attempt to detour into software engineering to produce a high-quality pipeline for the study of fractionation. A future project, however, could involve the recruitment of a computer science collaborator to develop a package implementing our methodology.