Integrated synteny- and similarity-based inference on the polyploidization–fractionation cycle

Yue Zhang, Zhe Yu, Chunfang Zheng and David Sankoff

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

DS, 0000-0001-8415-5189

Whole-genome doubling, tripling or replicating to a greater degree, due to fixation of polyploidization events, is attested in almost all lineages of the flowering plants, recurring in the ancestry of some plants two, three or more times in retracing their history to the earliest angiosperm. This major mechanism in plant genome evolution, which generally appears as instantaneous on the evolutionary time scale, sets in operation a compensatory process called fractionation, the loss of duplicate genes, initially rapid, but continuing at a diminishing rate over millions and tens of millions of years. We study this process by statistically comparing the distribution of duplicate gene pairs as a function of their time of creation through polyploidization, as measured by sequence similarity. The stochastic model that accounts for this distribution, though exceedingly simple, still has too many parameters to be estimated based only on the similarity distribution, while the computational procedures for compiling the distribution from annotated genomic data is heavily biased against earlier polyploidization events—syntenic ‘crumble’. Other parameters, such as the size of the initial gene complement and the ploidy of the various events giving rise to duplicate gene pairs, are even more inaccessible to estimation. Here, we show how the frequency of unpaired genes, identified via their embedding in stretches of duplicate pairs, together with previously established constraints among some parameters, adds enormously to the range of successive polyploidization events that can be analysed. This also allows us to estimate the initial gene complement and to correct for the bias due to crumble. We explore the applicability of our methodology to four flowering plant genomes covering a range of different polyploidization histories.

1. Introduction

Two orthogonal approaches to the study of fractionation—duplicate gene loss after polyploidization—focus on one hand on the decrease over time of the number of surviving duplicate pairs [1–7] and, on the other hand, the number of syntenically consecutive pairs lost after the event [8–12]. In this paper, we integrate the two in a single model, enabling for the first time inference of all parameters, with wide application to flowering plant genomes.

The basic model of the cycle between whole-genome replication (the result of polyploidization) and fractionation is a discrete-time branching process, reviewed in §2. Each branching event represents a polyploidization, at which time every member of the population gives rise to a variable number of offspring, interpreted as survivors of the fractionation process. Only the current (final) state of the process is observed.

The main theoretical construct is the prediction of the expected number of gene pairs (paralogs) generated at each branching event, but only observed at the current time. Grafted onto the branching process model is a way of identifying which of the events gave rise to each gene pair. This is based on a...
mutational model of gene sequence divergence, causing a
decay over time in the similarity between the genes in a pair.

The model enables us to quantitatively account for a major
type of comparative genomics data, discussed in §6, the dis-
tribution of gene pair similarities in syntenic blocks (collinear
runs of genes on two chromosomes) either within a genome
or between two genomes, as can be compiled by methods
like SyntenyMap on the CoGe platform [13,14]. For example,
based on the parameters of the branching process, we can cal-
culate rates of fractionation after each polyploidization, and
examine the extent it varies from species to species, and on
whether it is clocklike within genera, families or orders. We
have previously applied this approach to flowering plant
families that have been affected by more than one polyploid-
ization event over many tens of millions of years: the
Brassicaceae [2,3,6], Solanaceae [4], Malvaceae [5,6] and others.

A major limitation, not of the model, but of the previous
analyses based on it, is that the distribution of gene pair simi-
larities contains only enough information to estimate one
fractionation parameter per branching event, which is not
sufficient for most uses. The model, however, also predicts
the number of unpaired genes, or singletons, generated by
the process at each branching event, which can also be
observed in the very same SyntenyMap synteny blocks defined
by the gene pairs. As the first novel contribution of this paper
presented in §3, we show how these additional data
on syntenic structure greatly expand the scope of the analyses
based on the branching process model.

The parameters used in synteny block construction are set to
control the trade-off between accidental short runs of collinear
gene pairs arising through coincidental tandem duplication,
non-homologous recombination, gene movement, common
domain structure, assembly errors and other factors, on one
hand, versus runs genuinely associated with polyploidization
events, on the other hand, but shortened over time due to
chromosomal rearrangements, individual gene movements
and loss of both members of non-essential gene pairs.

These latter processes of erosion over evolutionary time of
the number of gene pairs (and, proportionately, of singletons)
belonging to blocks, summarized in §5, which may be subsumed
the number of gene pairs (and, proportionately, of singletons)
and loss of both members of non-essential gene pairs.

Figure 1. Event with ploidy \( t_i = 4 \), showing population of \( m_i = 5 \) genes at
time \( t_i \), each giving rise to 4 progeny, of which \( 1 \leq j \leq 4 \) survive until time
\( t_{i+1} \), \( a_j^{(i)} \) is the number of times \( j \) progeny survive. Black lines represent indi-
vidual progeny that survive, and grey lines represent the total progeny of a
gene that do not survive. Here, \( a_1^{(i)} = 2 \), \( a_2^{(i)} = a_3^{(i)} = a_4^{(i)} = 1 \). From [2].

\[ m_i = \sum_j a_j^{(i)} = 5 \]
\[ m_{i+1} = \sum_j ja_j^{(i)} = 11 \]

\[ g \]

other \( t_j \) genes

\( j \) surviving
progeny

\( (2,3) \) \( t_j \) pairs at time \( t_{i+1} \)

from one pair at time \( t_i \)
\( m_i \) \( t_j \) genes

\( m_i \) \( t_j \) pairs at time \( t_{i+1} \)

Figure 2. Counting \( t_j \)-pairs. The three unfractionated progeny of gene \( g \)
define three \( t_j \)-pairs, as indicated by three ovals. We follow the pair contained
in the uppermost oval, as the two members at time \( t_{i+1} \) independently
(shaed triangles) evolve into \( m_i' \) and \( m_i'' \) genes, respectively, defining
\( m_i' t_j n_i' \), \( t_j \)-pairs at time \( t_{i+1} \). From [2].

\[ \begin{align*}
mi' & = m_i + j \cdot a_j^{(i)} \leq 4, \\
ni' & = m_i + ja_j^{(i)} \\
mi'' & = m_i + j \cdot a_j^{(i)} \\
ni'' & = m_i + ja_j^{(i)}
\end{align*} \]

\( mi' - k \) fractionated progeny

\[ g \]

Figure 3. Synteny block on homologous fragments of two chromosomes.
Dark circles indicate retained genes, white circles deleted genes. There
are five retained gene pairs, four singletons on chromosome B and one singleton
on chromosome A.

\[ chromosome A \]
\[ chromosome B \]

in §6, we illustrate with four flowering plant genomes: poplar
(Populus trichocarpa), scarlet sage (Salvia splendens), durian
(Durio zibethinus) and black pepper (Piper nigrum). Each of
these genomes exemplifies a different history of two or
three stages of ancient tetraploidy and/or hexaploidy.

2. The branching process model

The model, expounded most completely in [4,5], consists of
successive branching events at times \( t_1 < \cdots < t_{n-1} \), and
observation time \( t_n > t_{n-1} \). The population size, \( 'gene compi-
lement' \), at \( t_1 \) is \( m_1 \) but only \( m_0 \) is observed. At each
branching time \( t_i \) every member of the population gives
rise to some number \( j \) of offspring, where \( 1 \leq j \leq t_i \), with
probability terms, every member has exactly \( r_i \) offspring and \( r_i - 1 \) of these are lost to fractionation.) The replication process corresponds to the concept of ‘2\( r_i \)-ploidization’, as in tetraploidization \( (r_i = 2) \) or hexaploidization \( (r_i = 3) \). (Note that while ancient hexaploidy can be inferred for many flowering plants, the process of engendering this state is understood to involve a succession of events, not a single ‘hexaploidization event’.)

The trajectory of the branching process is in effect a sample point from the \( n - 1 \) probability distributions \( u_i(\cdot), \ldots, u_n(\cdot) \) for \( i = 1, \ldots, n - 1 \). There is no provision for \( u_0(\cdot) \rightarrow 0 \), for reasons of inference—any model with one or more non-zero \( u_i(\cdot) \) is the same as some model with all \( u_0(\cdot) = 0 \) that has the same probability structure on the observations at \( t_0 \). (For purposes of modelling alone, forgoing empirical application, allowing non-zero \( u_0(\cdot) \) may be interesting, e.g. for studying limit behaviour. For example, the existing branching/fractionation process is supercritical, but allowing non-zero \( u_0(\cdot) \) can change this to critical or subcritical.)

Let \( a(\cdot) = (a_1(\cdot), \ldots, a_n(\cdot)) \) represent the numbers of genes at time \( t_0 \) with \( 1, \ldots, r_i \) offspring, so that

\[
m_i = \sum_{j=1}^{r_i} a_j(\cdot) \quad \text{and} \quad m_{i+1} = \sum_{j=1}^{r_i} a_j(\cdot), \quad (2.1)
\]
as in figure 1. Given \( m_i \), the probability of \( a(\cdot) \) is

\[
P_{r_i}(a(\cdot)) = (m_i a_1(\cdot), \ldots, a_{r_i}(\cdot)u_1(\cdot)^{v_1(\cdot)} \cdots u_n(\cdot)^{v_n(\cdot)}), \quad (2.2)
\]
and the probability of an entire trajectory, defining a paralog gene tree is

\[
P_{r_1}(a(1)) \cdots P_{r_n}(a(n - 1)), \quad (2.3)
\]
with \( m_1 \geq 1 \) given and the other \( m_i \) determined by equation (2.1).

Once we know how to calculate these probabilities, it is possible to calculate the \( E(m_i \) ). And using the independence of the trajectories starting at any two sibling genes existing at time \( t_0 \) and their independence from the trajectory between time \( t_1 \) and \( t_2 \), we can calculate \( E(N) \) the expected number of pairs of genes at time \( t_0 \) originating at time \( t_1 \), as summarized in figure 2.

The accumulation of multinomial coefficients in equations (2.2) and (2.3), and the potentially high degree polynomials might seem computationally formidable. In practice, however, 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 \( u(\cdot) = u_2(\cdot), i = 1, \ldots, n - 1 \) for the probability that both progeny of a gene at time \( t_1 \) survive until time \( t_{i+1} \). We have previously shown [4] the expected number \( N_i \) of duplicate pairs of genes born at time \( t_1 \) and observed at time \( t_k \) is

\[
E(N_1) = m_1 u(\cdot) \sum_{j=1}^{r_1} u(\cdot)^{v_1(\cdot)} \quad \text{and} \quad E(N_{n-1}) = \sum_{j=1}^{r_1} u(\cdot)^{v_1(\cdot)} m_1 u(\cdot)(n - 1).
\]

There are \( n - 1 \) parameters in the vector \( u(\cdot) \), and \( n - 1 \) equations in equation (2.4). The presence of an \( m \)th

\[
E(S_i) = m_1 (1 - u(\cdot)) \quad \text{and} \quad E(S_i) = m_1 \sum_{j=1}^{r_1} u(\cdot)^{v_1(\cdot)} (1 - u(\cdot)). \quad (2.5)
\]
3. Singletons in synteny blocks

The estimation of the fractionation rates, total gene complement sizes and crumble coefficients associated with the \( t_i \) depends on accurate values for the means of the \( N_i \) and \( S_i \) to substitute in equations such as (2.4) and (2.5). For the \( N_i \), this is ensured by the analysis of counting the gene pairs in synteny blocks (cf. §6), and calculating the sequence similarity of each pair to determine the appropriate \( t_i \). Singletons, on the other hand, by their nature are not comparable to any other gene, and thus would not seem to be directly associated with any \( t_i \).

One way to approach the number of singleton genes might be to subtract the number of genes in all \( t_i \) pairs from the total number of genes in the genome. Since a gene may be in several pairs, in synteny blocks corresponding to different \( t_i \), however, this calculation requires a more detailed data analysis than is possible from the distribution of gene pair similarities alone. More important, relying on the total number of genes in the genome is very misleading, since many or most of these will have been generated in the time elapsed between \( t_{n-1} \) and \( t_n \) by gene family expansion, tandem duplications and other processes.

It is the singletons in the synteny blocks, not the genome total minus the paired genes, that we will use here in the inference of retention rates. Because of their association with the pairs in the blocks, we can pinpoint when a singleton was created, from a pair arising at a specific \( t_i \). This results in additional independent observations to help in parameter estimation.

In the simplest model of fractionation [9], at each step, a random gene pair is selected to lose one member. In a competing class of models [8], gene loss is effected by excision of a variable length fragment of a chromosome, often formulated in terms of a gamma distribution. The study of the internal structure of syntenic blocks, illustrated in figure 3, arose as an indirect way of determining whether fractionation is basically ‘functional’ or ‘structural’. The former posits that fractionation targets specific gene pairs, inactivating or deleting one member of one pair, to redress dosage imbalances or other problems with synthetic or metabolic processes created by whole-genome doubling. The latter, structural explanation represents fractionation as a process of random excision of excess DNA with, say, geometrically distributed length, and which may involve one or more genes, as long as this is not lethal.

Empirically, both types of process play a substantive role [12]. Whatever their relative importance, the expected number of singletons in a synteny block is the sum of the expectations of number of singletons caused by either or both processes.

The number of singletons in a synteny block produced at \( t_i \) constitutes the appropriate comparison for the number of pairs in that block, because the singletons were produced by the same branching process as the pairs (or, in the alternative interpretation, during the period between \( t_i \) and \( t_{i+1} \)).

### Table 2. Statistics and parameter estimates for the black pepper genome.

| block length | cutoff | \( r_1 \) pairs | \( r_2 \) pairs | \( t_1 \) singles | \( t_2 \) singles | \( c \) | \( u \) | \( v \) | \( m_1 \) |
|--------------|--------|----------------|----------------|-----------------|----------------|-----|-----|-----|-----|
| \( \geq 3 \)  | 89.4%  | 18 898         | 15 646         | 23 637          | 23 206         | 1.09| 0.29| 0.40| 30 446 |
| \( \geq 4 \)  | 89.3%  | 13 593         | 14 244         | 19 875          | 22 773         | 0.92| 0.26| 0.38| 29 311 |
| \( \geq 5 \)  | 89.1%  | 11 067         | 13 711         | 19 995          | 23 657         | 0.85| 0.23| 0.37| 30 417 |

### Table 3. Equations for rates, initial population and crumble for a tripling followed by a doubling.

| event | observed | expected number |
|-------|----------|-----------------|
| \( t_1 \) pairs | \( cm_1(u + 3u)(1 + v)^2 \) |
| \( t_2 \) pairs | \( m_1(1 + 2u + o)(1 + v) \) |
| \( t_1 \) singletons | \( cm_1(1 - u - u') \) |
| \( t_2 \) singletons | \( m_1(1 + 2u + o)(1 - v) \) |

![Figure 6. Distribution of sequence similarity of duplicate gene pairs in the poplar genome.](image)

![Figure 7. Distribution of sequence similarity of duplicate gene pairs in the durian genome.](image)
run) of length expected number of runs of 1 chosen at random among the remaining pairs at each step. It is important to understand the relationship between neighbouring expected values, such non-independence does not matter to the gene pair to a neighbouring pair. Since our model only calculates the retention of both is not statistically independent from one block at the same time. If this is the case, the loss of one copy or fractionation may affect several duplicate gene pairs in a synteny block.

3.1. Synteny and fractionation

Equations for rates, initial population and crumble for a tripling process, where the probability of success at each

event observed expected number

| event | observed | expected number |
|-------|----------|----------------|
| $t_1$ | pairs    | $cm_1(u + 3u'/1 + 2u' + v)^2$ |
| $t_2$ | pairs    | $m_1(1 + 2u' + u)(v + 3u')$ |
| $t_3$ | singletons | $cm_1(1 - u - u')$ |
| $t_4$ | singletons | $m_1(1 + 2u' + u)(1 - v - v')$ |

The number of successes is $x$ trials with a probability of success $q = 1$. However, the variance of the number of successes is non-negligible, whereas it is zero for our process, and the variance of the number of runs of a given length is also greater than our process. Thus our interest in the fractionation process, where the probability of success at each position depends on the total number of successes already achieved.

In [17], we showed how this model was deficient in predicting longer run and gap lengths in the Coffea arabica tetraploid genome. We estimated this one gene pair at a time model accounted for about 70% of fractionation events, while a geometric distribution of deletion lengths with mean 3.5 accounted for the remaining 30%.

4. Constraints on rates

Under the assumption that the event that each offspring gene is deleted, or survives, is an independent binomial trial, conditioned on at least one such gene surviving, we avoid having to estimate more than one parameter in $u(\cdot)$ for each replication event. The $\Sigma r_i - 1$ ploidy parameters tend to be too numerous when the $r_i$ are larger than 2. As first suggested in [15] and verified in [16], we can circumvent this by assuming gene loss is independent among all the copies, conditional on at least one surviving. For $r_i = 3$, if $p$ is the probability one gene is lost, the probability that

— all three genes survive is $(1 - p)^3/(1 - p^3) = u'$
— two of the three survive is $3p(1 - p)^2/(1 - p^3) = u$
— only one survives is $3p^2(1 - p)/(1 - p^3) = 1 - u' - u'$

Let

$$E = \sqrt{3(3 - 6u - u^2)}.$$ (4.1)

Then

$$u_1 = \frac{u^2 - (u + 1)E + 3}{12 + 2E} \quad \text{or} \quad u_1' = \frac{u^2 + (u + 1)E + 3}{12 - 2E}. \quad (4.2)$$

As can be seen in figure 4, this relationship—the left-hand formula in (4.2)—is indistinguishable for practical purposes from $u' = u^{2.5}$ as long as $u < 0.37$. While we will not incorporate this constraint into our estimation procedures directly, we will use it to choose among alternative analyses when there are too many parameters compared to equations in the branching process.

5. A model for the erosion of synteny blocks over time

The fractionation process has the effect of eroding and completely losing synteny groups over long periods of time, partly because of biological processes like chromosomal rearrangement and gene pair divergence, and partly because of necessary technical limitations on the software detecting
Table 6. Statistics and parameter estimates for the durian genome.

| length | cut-off | $t_1$ | $t_2$ | $t_1$ | $t_2$ | $c$ | $u$ | $u'$ | $v$ | $v'$ | $m_1$ |
|--------|---------|-------|-------|-------|-------|-----|-----|-----|-----|-----|-------|
| ≥3     | 85.8%   | 11 472| 14 854| 7876  | 10 109| 0.75| 0.27| 0.04| 0.40| 0.11| 15 200|
| ≥4     | 85.5%   | 8081  | 14 538| 6965  | 10 602| 0.60| 0.25| 0.03| 0.39| 0.10| 16 000|
| ≥5     | 85.5%   | 6704  | 14 242| 6419  | 10 691| 0.53| 0.23| 0.03| 0.39| 0.10| 16 300|

Table 7. Equations for rates, initial population and crumble for two successive triplings followed by a doubling.

| event | observed | expected number |
|-------|----------|-----------------|
| $t_1$ | pairs    | $c_1 m_1 (u + 3u')(1 + 2u'(1 + w)^2$ |
| $t_2$ | pairs    | $c_2 m_1 (1 + 2u' + u)(1 + w)^2$ |
| $t_3$ | pairs    | $m_1 (1 + 2u' + u)(1 + 2u' + 2v')$ |
| $t_1$ | singletons| $c_1 m_1 (1 - u - u')$ |
| $t_2$ | singletons| $c_2 m_1 (1 - u' - u)(1 + v')$ |
| $t_3$ | singletons| $m_1 (1 + 2u' + u)(1 + 2v' + v)(1 - w)$ |

Table 8. Equations for rates, initial population and crumble for a tripling followed by two doublings.

| event | observed | expected number |
|-------|----------|-----------------|
| $t_1$ | pairs    | $c_1 m_1 (u + 3u')(1 + v)^2(1 + w)^2$ |
| $t_2$ | pairs    | $c_2 m_1 (1 + 2u' + u)(1 + v + w)^2$ |
| $t_3$ | pairs    | $m_1 (1 + 2u' + u)(1 + v + w)$ |
| $t_1$ | singletons| $c_1 m_1 (1 - u - u')$ |
| $t_2$ | singletons| $c_2 m_1 (1 - u' - u)(1 + v')$ |
| $t_3$ | singletons| $m_1 (1 + 2u' + u)(1 + v')(1 - w)$ |

6. Four plant genomes

We explore four genomes with various histories of genome replication. We assume that the historical polyploidy events were correctly established for each species, although we could also find them using the method in [6]. The history determines a number of equations similar to (2.4) linking the fractionation rates to the expected values of singletons and pairs observed from each event.

The construction of datasets for our analysis, embodied in software such as SynMap applied to genomes available on the CoGe platform [13,14], involves scanning a genome for pairs of similar genes, then searching for runs of collinear such pairs in two different genome locations. Each run, or ‘synten block’, must contain a preset minimum number of pairs and have no more than a certain number of consecutive unpaired genes. That the level of similarity of the pairs is relatively uniform in a block, together with the collinearity, lends credence to the conclusion that the pairs were all created simultaneously at one of the replication (branching) times $t_i$, both locations inheriting the pre-replication gene order, and that the interspersed singletons are the remnants of fractionated contemporaneous pairs.

In each case, we

- compare the genome to itself, using SynMap with default parameters,
- construct the distribution of similarities of gene pairs in the synteny blocks,
- find the singleton genes embedded in each synteny block,
basically as a label for the authors, or simply local assembly issues. duplications or high heterozygosity mentioned by the pairs closer to 100% similarity may reflect the segmental with mean values around 78% and 94%. Additional duplicate cate genes in synteny blocks in figure 5 is indicative of two, suggested one doubling event, but the distribution of dupli-

In our analyses, we use \( u, u', v, v', c \) and \( m_1 \) to refer to the survival of two or three (if pertinent) copies instead of one after the first polyploidization event, the survival of two or three (if pertinent) copies instead of one after the second polyploidization event, the crumble constant and the initial gene complement size, respectively. Note that although our theoretical discussions in §§2, 3 and 5 were phrased in terms of the branching times \( t_i \), the equations describing the individual models involved only the \( u(\cdot) \), which are really retention probabilities, not fractionation rates. In the following examples, the term \( t_i \) serves basically as a label for the \( i \)th branching event.

### 6.1. Black pepper (Piper nigrum)

We choose to analyse the black pepper genome (CoGe ID 56158) since it has undergone the simplest series of whole-genome replications, namely two successive doublings. As a magnoliid, it diverged from the eudicots before the ‘gamma’ whole-genome tripling common to our four other examples in this section. The original report [20] only suggested one doubling event, but the distribution of duplicate genes in synteny blocks in figure 5 is indicative of two, with mean values around 78% and 94%. Additional duplicate pairs closer to 100% similarity may reflect the segmental duplications or high heterozygosity mentioned by the authors, or simply local assembly issues.

The equations where we substitute observed values for expected ones in expressions deriving from the branching process model include those for pairs (cf. equation (2.4)) plus those for singletons (cf. equation (2.5)), as in table 1. To take into account the syntenic crumble process, we repeated the SYNMAP search for synteny blocks with three different values of the minimum block size parameter: 5 (the default), 4 and 3. The results in table 2 confirm this effect, with over 70% more \( t_1 \) pairs and 18% more singletons when the block size criterion is relaxed from 5 to 3. This is substantial, even allowing for some noise with the less stringent criterion. The crumble constant, which estimates the loss of synteny due solely to the block size criterion, is moderate for size 5 and 4, and undetectable for size 3 (\( c \approx 1 \)).

Of note is the stability of the estimates of \( m_1 \), the number of genes in the genome before \( t_1 \). Also, the cutoff between the two components of the distribution does not vary, suggesting that the additional gene pairs generated by the less stringent criterion come from the same two events as with the default configuration.

### 6.2. Poplar (Populus trichocarpa)

Poplar (CoGe ID 25127) descends from the important whole-genome tripling (known as ‘gamma’) at the origin of the core eudicots. As a member of the Salicaceae family, it has undergone a further whole-genome doubling [21] (the ‘Salicoid’ doubling). The equations for a tripling followed by a doubling are given in table 3.

Figure 6 shows a clear separation between the gene pairs created by the two events.

In contrast to the black pepper analysis, we now have more parameters (five) to determine, with only four equations. Here, we make use of the constraint derived from the conditioned binomial analysis developed in §4. Rather than enter the constraint as an additional equation, which would lend it too much weight in simultaneously solving for the other parameters, we simply solved the four equations for a range of values of \( c \), namely each value

| block length | \( c_1 \) | \( c_2 \) | \( u \) | \( u' \) | \( v \) | \( v' \) | \( w \) | \( m_1 \) |
|--------------|--------|--------|--------|--------|--------|--------|--------|--------|
| ≥3           | 1.1    | 0.7    | 0.31   | 0.31   | 0.24   | 0.07   | 0.69   | 13 333 |
| ≥4           | 0.9    | 0.8    | 0.19   | 0.19   | 0.29   | 0.04   | 0.65   | 13 760 |
| ≥5           | 0.8    | 0.8    | 0.15   | 0.15   | 0.27   | 0.05   | 0.64   | 13 821 |

---

Table 9. Statistics for the scarlet sage genome.

| block length | \( t_1 \) | \( t_2 \) | \( t_3 \) | \( t_1 \) | \( t_2 \) | \( t_3 \) |
|--------------|--------|--------|--------|--------|--------|--------|
| ≥3           | 27 837 | 15 640 | 16 801 | 9941   | 8632   | 7726   |
| ≥4           | 18 826 | 15 628 | 15 515 | 9576   | 9303   | 8288   |
| ≥5           | 15 265 | 14 864 | 14 786 | 8942   | 9342   | 8441   |

---

Table 10. Parameter estimates for the scarlet sage genome according to the tripling–tripling–doubling model.
between 0 and 1, in steps of 0.01. Then we picked out the value of \( c \) that resulted in the closest match to equation (4.2).

Table 4 again shows the stability of \( m_1 \) and the cutoff between the two components, despite the 60% increase in the number of \( t_1 \) pairs and 32% increase in the singletons when the block stringency is reduced, due to the use of the crumble constant.

6.3. Durian (Durio zibethinus)

When first sequenced the durian genome (CoGe ID 51764) was thought to have undergone a further doubling after the gamma tripling [22]. Subsequent work by ourselves [5] and others [23] showed that the second event was clearly also a tripling (figure 7).

In the case of two triplings, there are still only four equations based on the similarity distribution, two for the pairs, and two for the singletons. But now there are six parameters to find: \( u, u', v, v', c \) and \( m_1 \). Again, we relied on the conditioned binomial model for the relationship between the two-copy and three-copy survival parameters. We defined a two-dimensional grid for \( m_1 \) from 10 000 to 30 000 in steps of 100, and \( c \) from 0 to 1 in steps of 0.01, and solved the equations for each point on the grid. We then retained all the combinations that closely approximated the constraint in equation (4.2) between \( u \) and \( u' \). Among these solutions, we then chose the one where \( v \) and \( v' \) most closely satisfied the same constraint.

For the second version, the search array involved only \( c_2 \), there being enough equations to directly solve the six equations for \( u, u', v, w, m_1 \) and \( c_1 \).

In the first history, two triplings and a doubling, the solutions were assessed to find the combinations of \( c_1 \) and \( c_2 \) where \( u \) and \( u' \) were close to the predictions of equation (4.2). Among these solutions, we then chose the one where \( v \) and \( v' \) most closely satisfied the same constraint.

For the second history, a tripling and two doublings, the search array involved only \( c_2 \) there being enough equations to directly solve the six equations for \( u, v, w, m \) and \( c_1 \). In tables 10 and 11, we note consistency throughout in the survival rates in the two histories, and the crumble constants \( c_1 \), which are similar.

7. Conclusion

We have described a comprehensive account of the similarity distribution of duplicate gene pairs as a function of the time since their creation by whole-genome doubling, as measured by sequence similarity. A branching process model for generating this distribution has too many rate parameters to be estimated based only on the distribution itself. We mitigate this problem by using the frequency of unpaired genes, distinguished from other single-copy genes by their embedding in paralogous synteny blocks, stretches made up largely of duplicate parts. However, the computational procedures for constructing synteny blocks from annotated genomic data are heavily biased against earlier polyploidization events. We have shown here how to quantify this syntenic ‘crumble’, and how to correct the bias caused by it. Other parameters, such as the size of the initial gene complement, are less accessible. We showed how previously established constraints among some parameters add substantially to the range of successive polyploidization events that can be analysed. In particular, this also allows us to estimate the initial gene complement and helps correct for the bias due to crumble. Finally, we demonstrated the applicability of our methodology to four flowering plant genomes with various doubling and tripling histories.

Table 11. Parameter estimates for the scarlet sage genome according to the tripling–doubling–doubling model.

| block length | \( c_1 \) | \( c_2 \) | \( u \) | \( u' \) | \( v \) | \( w \) | \( m_1 \) |
|--------------|---------|---------|------|------|-----|-----|-------|
| \( \geq 3 \)  | 1.06    | 0.80    | 0.26 | 0.03 | 0.39| 0.69| 13 297|
| \( \geq 4 \)  | 0.93    | 0.87    | 0.22 | 0.02 | 0.38| 0.65| 13 626|
| \( \geq 5 \)  | 0.85    | 0.88    | 0.21 | 0.02 | 0.37| 0.64| 13 601|
The importance of singletons in our analysis prompts concerns of whether they may originate, not from fractionation of their paralogs, but from their insertion into one of the homologous chromosomes, such as through the transposon activity rife in plant genomes [25]. However, the major plant transposon families are all well characterized, and transposons are routinely not annotated as genes, and would not show up in the synteny blocks detected by SynMap. Even if the annotation were faulty, masking routines would eliminate transposons, but the genomes we have verified, such as the Populus we studied in §6.2, as well as linen (Linum usitatissimum) that have unmasked and masked versions of the same assembly in CoGe, show no fewer genes after masking than before. Thus we can be confident in the origin of our singletons in the fractionation process.

Even if we can estimate the retention rates and the gene complement at each event, one critical model parameter cannot be derived from the frequency distribution of gene pair similarities and the number of singletons, namely the ploidy level r. Though we may sometimes be able to guess r by visual inspection of the output of SynMap, this is not usually the case for earlier events. We have previously shown how to derive additional information from the raw gene pair data in order to construct informative gene triples [6]. Statistics on the configurations of similarities within these triples can then be used to deduce r.

Data accessibility. The annotated genome data used in this paper is freely available on the CoGe website. The SynMap program is also available online on that site. The equation solving and other calculations were carried out using the mpl and maxLik packages in R.

**Authors’ contributions.** Y.Z., Z.Y. and D.S. conceived of and designed the study and wrote the manuscript. C.Z. participated in data collection and analysis, including writing scripts for data conversion. All authors gave final approval for publication and agree to be held accountable for the work performed therein.

Competing interests. We declare we have no competing interests.

Funding. This work has been supported by the Natural Science and Engineering Research Council of Canada and the Canada Research Chair programme.

References

1. Zhang Y, Sankoff D. 2017 The similarity distribution of paralogous gene pairs created by recurrent alternation of polyploidization and fractionation. In Comparative genomics. RECOMB-CG 2017 (eds J Meidanis, L Nakhleh). Lecture Notes in Computer Science, vol. 10562, pp. 1–13. (doi:10.1007/978-3-319-67979-2_1)

2. Zhang Y, Zheng C, Sankoff D. 2018 Pinning down ploidy in paleopolyploid plants. BMC Genomics 19, 28. (doi:10.1186/s12864-018-4624-y)

3. Sankoff D, Zheng C, Zhang Y, Meidanis J, Lyons E, Tang H. 2019 Models for similarity distributions of syntenic homology and applications to phylogenetics. IEEE/ACM Trans. Comput. Biol. Bioinform. 16, 727–737. (doi:10.1109/TCBB.2018.2849373)

4. Zhang Y, Zheng C, Sankoff D. 2019 A branching process for homology distribution-based inference of polyploidy, speciation and loss. Algorithms Mol. Biol. 14, 16. (doi:10.1186/s13015-019-0153-8)

5. Zhang Y, Zheng C, Islam C, Kim Y-M, Sankoff D. In press. Branching out to speciation in a model of fractionation: the Malvaceae. IEEE/ACM Trans. Comput. Biol. Bioinform. (doi:10.1109/TCBB.2019.2955649)

6. Zhang Y, Zheng C, Sankoff D. 2019 Distinguishing successive ancient polyploidy levels based on genome-internal syntenic alignment. BMC Bioinf. 20, 635. (doi:10.1186/s12859-019-3202-x)

7. Yu Z, Zheng C, Albert VA, Sankoff D. 2020 Excision dominates pseudogenization during fractionation after whole genome duplication and in gene loss after speciation in plants. Front. Genet. 11, 1654. (doi:10.3389/fgene.2020.00305)

8. van Hoek MJ, Hogeweg P. 2007 The role of transposable elements on polyploid plant genomes. Trends Biol. Evol. 22, 494–499. (doi:10.1093/molbev/msm058)

9. Byrnes JK, Morris GP, Li WH. 2006 Reorganization of adjacent gene relationships in yeast genomes by whole-genome (duplication) and gene deletion. Mol. Biol. Evol. 23, 1136–1143. (doi:10.1093/molbev/msj121)

10. Zheng C, Wall PK, Leebens-Mack J, Albert VA, Sankoff D. 2009 Gene loss under neighbourhood selection following whole genome duplication and the reconstruction of the ancestral Populus diploid. J. Bioinform. Comput. Biol. 7, 499–520. (doi:10.1142/S0219720009004199)

11. Sankoff D, Zheng C, Wang B, Buen Abad Najar CF. 2015 Structural vs. functional mechanisms of duplicate gene loss following whole genome doubling. BMC Genomics 16(Suppl. 17), S9. (doi:10.1186/1471-2105-16-S17-S9)

12. Yu Z, Sankoff D. 2016 A continuous analog of run length distributions reflecting accumulated fractionation events. BMC Bioinf. 17(Suppl. 14), 412. (doi:10.1186/s12859-016-1265-5)

13. Lyons E, Freeleing M. 2008 How to usefully compare homologous plant genes and chromosomes as DNA sequences. Plant J. 53, 661–673. (doi:10.1111/j.1365-313X.2007.03326.x)

14. Lyons E, Pedersen B, Kane J, Alam M, Ming R, Tang H, Wang X, Bowers J, Paterson A, Lisch D, Freeleing M. 2008 Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar and grape: CoG with rosids. Plant Physiol. 148, 1772–1781. (doi:10.1104/pp.108.124867)

15. Nadeau JH, Sankoff D. 1997 Comparable rates of gene loss and functional divergence after gene duplications early in vertebrate evolution. Genetics 147, 1259–1266. (doi:10.1093/genetics/147.3.1259)

16. Zhang C, Chen E, Albert VA, Lyons E, Sankoff D. 2013 Ancient euchromatic hexaploidy meets ancestral euroid gene order. BMC Genomics 14(Suppl. 7), 53. (doi:10.1186/1471-2164-14-S7-53)

17. Yu Z, Zheng C, Sankoff D. 2020 Gaps and runs in syntenic alignments. In Algorithms for computational biology (eds C Martin-Vide, M Vega-Rodriguez, T Wheeler). Lecture Notes in Computer Science, vol. 12099, pp. 49–60. Cham, Switzerland: Springer. (doi:10.1007/978-3-030-42266-0_5)

18. Weisstein E. 2020 Run. MathWorld—A Wolfram Web Resource. See http://mathworld.wolfram.com/ (accessed 27 August 2020).

19. McLachlan GJ, Peel D, Basford KE, Adams P. 1999 The EM package software for the fitting of mixtures of normal and t-components. J. Stat. Softw. 4, 1–14. (doi:10.18637/jss.v004.i02)

20. Hu L et al. 2019 The chromosome-scale reference genome of black pepper provides insight into piperine biosynthesis. Nat. Commun. 10, 4702. (doi:10.1038/s41467-019-12607-6)

21. Turkam GA et al. 2006 The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science 313, 596–604. (doi:10.1126/science.1128691)

22. Teh BT et al. 2017 The draft genome of tropical fruit durian (Durio zibethinus). Nat. Genet. 49, 1633–1641. (doi:10.1038/ng.3972)

23. Wang J et al. 2019 Recursive paleohexaploidization shapes the durian genome. Plant Physiol. 179, 209–219. (doi:10.1104/pp.18.00921)

24. Dong AX et al. 2018 High-quality assembly of the reference genome for scarlet sage, Salvia splendens, an economically important ornamental plant. GigaScience 7, giy068. (doi:10.1093/gigascience/giy068)

25. Vicent CM, Casacuberta JM. 2017 Impact of transposable elements on polyploid plant genomes. Ann. Bot. 120, 195–207. (doi:10.1093/ aob/mcx078)