Zhang *et al. BMC Bioinformatics* 2019, **20**(Suppl 20):635
https://doi.org/10.1186/s12859-019-3202-x

BMC Bioinformatics

RESEARCH Open Access

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

Yue Zhang, Chunfang Zheng and David Sankoff*

From 17th RECOMB Satellite Conference on Comparative Genomics
Montpellier, France. 1-4 October 2019

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.

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

Conclusions: We have developed a way of inferring the ploidy of up to three successive WGD and/or WGT events by estimating the time of origin of each of the similarities in triples of genes. This may be generalized to a larger number of events and to higher ploidies.

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,...). This is illustrated in Fig. 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

*Correspondence: sankoff@uottawa.ca
Department of Mathematics and Statistics, University of Ottawa, 150 Louis Pasteur pvt, Ottawa K1N 6N5, Canada

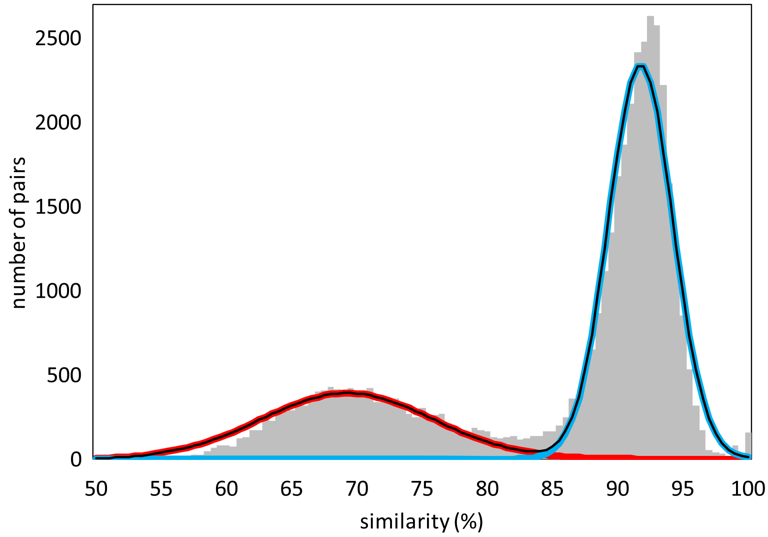

Fig. 1 Distribution of gene pair similarities. Pairs in the *Durio zibethinus* (CoGe ID 51764) genome after two rounds of whole genome tripling. Discrimination point $H = 85.2\%$. Cut-off for pairs not originating in polyploidization > 98%

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 Fig. 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* [14], 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 $t_i, i = 1, \ldots, n$. Set $m_1 = 1$. At time t_i, $i = 1, \ldots n-1$, 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

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

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

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

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 Fig. 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 Fig. 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.

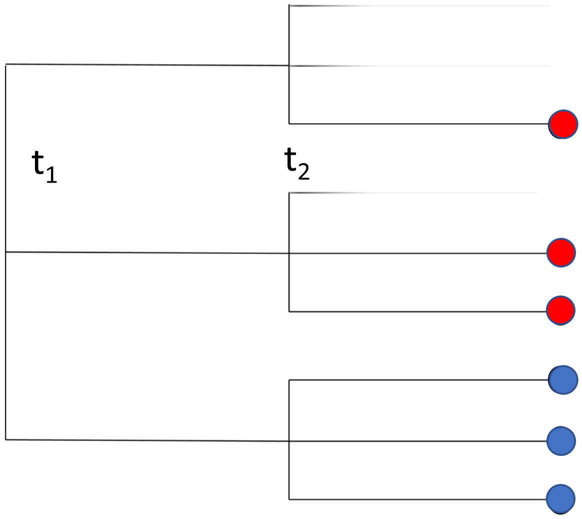

Fig. 2 Sample trajectory, starting from a single gene, of branching process based on two whole genome triplings

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 4 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

Table 1 Formulae for the expected numbers of triplets $W_M(\Delta)$ of each type Δ, by branching model M

Model $M \in \mathcal{M}$	Tripling- tripling	Tripling- doubling	Doubling- tripling	Doubling- doubling
Triplet $\Delta \in \mathcal{T}$	(3,3)	(3,2)	(2,3)	(2,2)
$\{t_1, t_1, t_1\}$	$u'(1 + 2v' + v)^3$	$u'(1 + v)^3$	-	-
$\{t_1, t_1, t_2\}$	$2(3u' + u)(3v' + v)$ $\times(1 + 2v' + v)$	$2(3u' + u)$ $\times v(1 + v)$	$2u(3v' + v)$ $\times(1 + 2v' + v)$	$(2u)v(1 + v)$
$\{t_1, t_2, t_2\}$	-	-	-	-
$\{t_2, t_2, t_2\}$	$(1 + 2u' + u)v'$	-	$(1 + u)v'$	-

$u =$ 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

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.,

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

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 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 $covar(t_1, t_2) = 0$, 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

Table 2 Durian model predictions before (underlying) and after imposition of mutational divergence

Model $M \in \mathcal{M}$	Tripling- tripling	Tripling- doubling	Doubling- tripling	Doubling- doubling	
Triplet $\Delta \in \mathcal{T}$	(3,3)	(3,2)	(2,3)	(2,2)	Observed
	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

Shaded column indicates the model predicted by the literature and the closest fit to the observed profile

Table 3 *Populus* model predictions before (underlying) and after imposition of mutational divergence

Model $M \in \mathcal{M}$	Tripling-	Tripling-	Doubling-	Doubling-	
	tripling	doubling	tripling	doubling	
Triplet $\Delta \in \mathcal{T}$	(3,3)	(3,2)	(2,3)	(2,2)	Observed
	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

Shaded column indicates the model predicted by the literature and the closest fit to the observed profile

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;

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

where μ 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.

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

Model $M \in \mathcal{M}$	Triplication-	Triplication-	Duplication-	Duplication-	
	triplication	duplication	triplication	duplication	
Triplet $\Delta \in \mathcal{T}$	(3,3)	(3,2)	(2,3)	(2,2)	Observed
	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

In the two shaded columns, the expected (2,3) profile does not fit the observed pattern as well as the (3,3) 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 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 (CoGe ID 25127), whose gene pair similarity distribution is displayed in Fig. 3. Along with γ, this shares the ancient "salicoid" WGD with other members of the Salicaceae family [15]. (3,2) is the only model that correctly identifies both the γ event as a WGT, and the more recent event as a WGD.

Cabbage

The recent ancestor of *Brassica oleracea* genome (CoGe ID 26018), 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 Brassicacea 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 Fig. 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%.

This analysis in Table 4 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.

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 Fig. 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 $\{t_2, t_2, t_2\}$ 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 t_2, and consequently an excess of t_2 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. Choosing among models with different numbers of events would

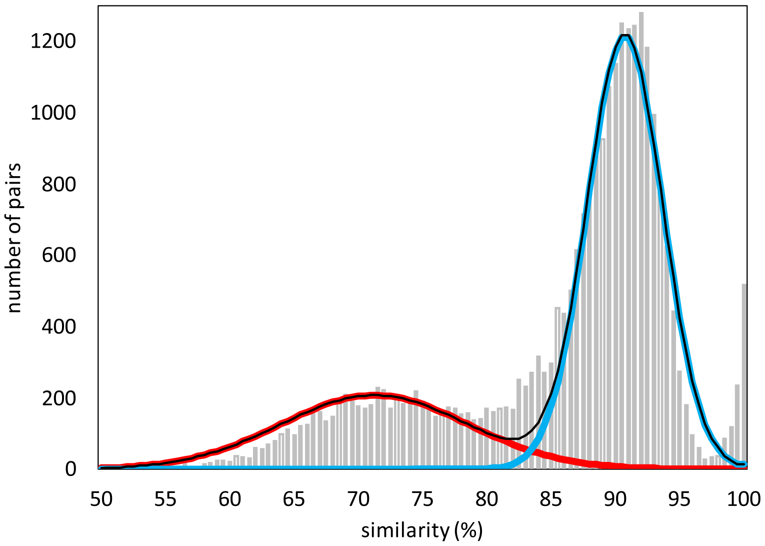

Fig. 3 Distribution of syntenic gene pair similarities in *Populus trichocarpa*. Discrimination point $H = 84.5\%$. Cut-off for pairs not originating in polyploidization $> 97.5\%$

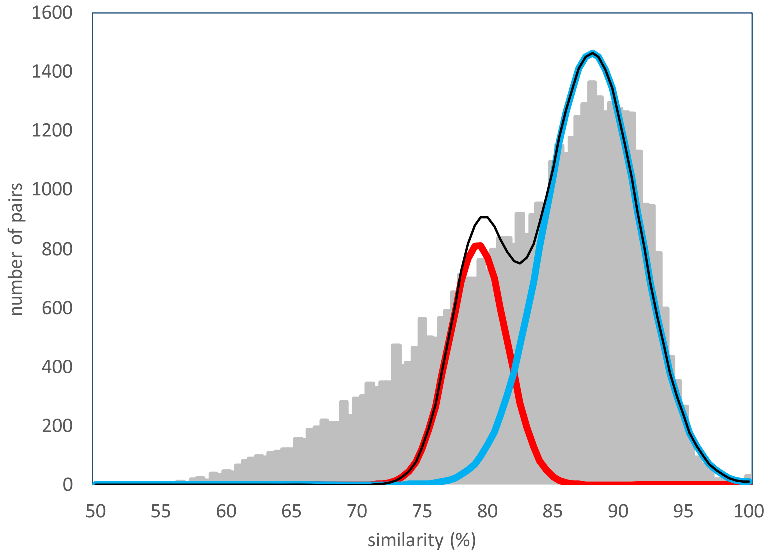

Fig. 4 Distribution of syntenic gene pair similarities in *Brassica oleracea*. Discrimination point $H = 83.3\%$. Cut-off for pairs not originating in two recent polyploidization $\leq 76\%$

require some standard for model selection such as the Akaike or Bayesian information criteria.

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 compared 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 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.

A distribution of gene pair similarities is generated in the comparison of two related genomes as well as in the self-comparison of a single genome. The number of

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

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

$u =$ 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. w and w' are the probabilities that two or three progeny survive after the third event

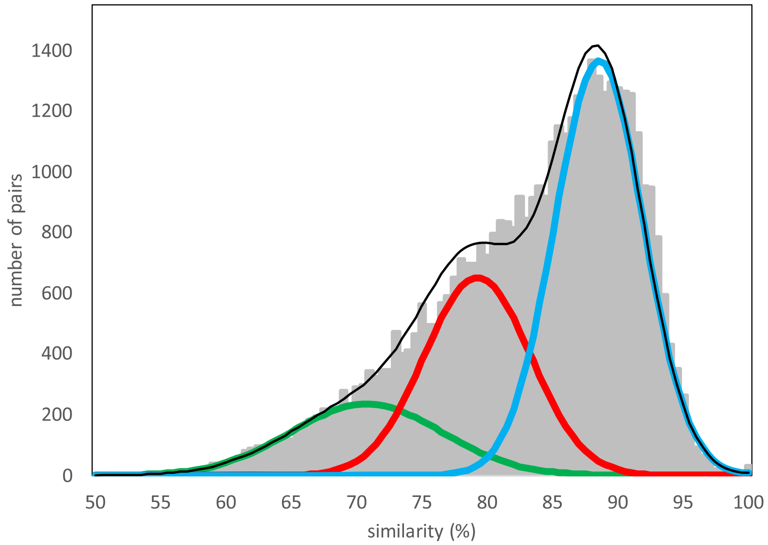

Fig. 5 Distribution decomposed into three events. Discrimination points $H_1 = 73\%, H_2 = 85\%$

Table 6 *Brassica oleracea* three-event model predictions before (underlying) and after imposition of mutational divergence

Model $M \in \mathcal{M}$									
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)	Observed
	AFTER FRACTIONATION ONLY								
$\{t_1, t_1, t_1\}$	0	613	0	0	0	423	537	386	178
$\{t_1, t_1, t_2\}$	2433	2333	2119	1667	5456	1620	2043	1471	1028
$\{t_1, t_2, t_2\}$	0	0	0	0	0	0	0	0	799
$\{t_2, t_2, t_2\}$	0	0	1936	0	1606	0	1866	1335	1712
$\{t_2, t_2, t_3\}$	8210	7875	7152	5627	4953	5468	6896	4966	5181
$\{t_2, t_3, t_3\}$	0	0	0	0	0	0	0	0	1243
$\{t_3, t_3, t_3\}$	0	0	0	4726	364	4593	0	4215	3226
$\{t_1, t_1, t_3\}$	4379	4200	3814	3001	2642	2916	3678	2648	966
$\{t_1, t_3, t_3\}$	0	0	0	0	0	0	0	0	30
$\{t_1, t_2, t_3\}$	0	0	0	0	0	0	0	0	658
	AFTER FRACTIONATION + MUTATION								
$\{t_1, t_1, t_1\}$	58	248	52	40	131	172	219	158	178
$\{t_1, t_1, t_2\}$	1221	1390	1091	837	2335	964	1243	895	1028
$\{t_1, t_2, t_2\}$	1172	1264	1253	803	2279	877	1331	957	799
$\{t_2, t_2, t_2\}$	1224	1222	2303	882	2302	889	2262	1661	1712
$\{t_2, t_2, t_3\}$	6070	5827	5638	4452	4060	4330	5441	4177	5181
$\{t_2, t_3, t_3\}$	1080	1036	999	1913	795	1860	964	1740	1243
$\{t_3, t_3, t_3\}$	93	89	86	3282	309	3189	83	2930	3226
$\{t_1, t_1, t_3\}$	1787	1717	1557	1225	1232	1192	1504	1083	966
$\{t_1, t_3, t_3\}$	56	54	50	38	37	37	48	35	30
$\{t_1, t_2, t_3\}$	2261	2174	1992	1549	1542	1510	1926	1386	658

In the two shaded columns, the expected (3,2,3) profile fits the observed pattern as well as the (3,3,3) profile orthologous gene pairs available when comparing two related genomes is generally much greater than the number of paralogous pairs identified in the self-comparison of two genomes, simply because the loss by fractionation of one copy of a duplicated gene does not eliminate all related orthology pairs: the other remaining copy and its orthology pairs remain intact. This suggests an avenue to improved accuracy of polyploidy levels inference. The larger number of data, however, may not always compensate for the fact that the speciation component of the similarity distribution is always the most recent one [9], so that the more remote (earlier) components associated with polyploidy become statistically less clear and informative.

Abbreviations

ML: Maximum likelihood; WGD: Whole genome doubling; WGT: Whole genome tripling

Acknowledgments

The authors thank Victor A. Albert for encouraging us to explore this research topic.

About this supplement

This article has been published as part of *BMC Bioinformatics Volume 20 Supplement 20, 2019: Proceedings of the 17th Annual Research in Computational Molecular Biology (RECOMB) Comparative Genomics Satellite Workshop: Bioinformatics*. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-20-supplement-20.

Authors' contributions

DS and YZ. designed the study. YZ carried out most of the research. CZ did part of the analysis. DS and YZ wrote the paper. All authors read and approved the paper.

Funding

Publication costs are funded by the Natural Sciences and Engineering Research Council of Canada Discovery Grant RGPIN-2016-05585. The research described in this article as well as publication costs were also partly funded by this grant. DS holds the Canada Research Chair in Mathematical Genomics.

Availability of data and materials

The datasets analysed during the current study are available in the CoGe repository [1, 2], https://genomevolution.org/coge/

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Published: 17 December 2019

References

1. Lyons E, Freeling M. How to usefully compare homologous plant genes and chromosomes as DNA sequences. Plant J. 2008;53:661–73. https://doi.org/10.1111/j.1365-313X.2007.03326.x.
2. Lyons E, Pedersen B, Kane J, Alam M, Ming R, Tang H, et al. Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar and grape: CoGe with rosids. Plant Physiol. 2008;148: 1772–81.
3. Teh BT, Lim K, Yong CH, Ng CCY, Rao SR, et al. The draft genome of tropical fruit durian *Durio zibethinus*. Nat Genet. 2017;49:1633–41.
4. 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.
5. 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 Bioinforma. 2017;15(5):1545–5963.
6. 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 Trans Comput Biol Bioinforma. 2018. https://doi.org/10.1109/TCBB.2018.2849377.
7. Zhang Y, Zheng C, Sankoff D. Pinning down ploidy in paleopolyploid plants. BMC Genomics. 2018;19(Suppl 5):287.
8. Zhang Y, Zheng C, Sankoff D. A branching process for homology distribution-based inference of polyploidy, speciation and loss. Algoritm Mol Biol. 2019;14:18.
9. Zhang Y, Zheng C, Islam S, Kim Y-M, Sankoff D. Branching out to speciation in a model of fractionation: the Malvaceae. IEEE/ACM Trans Comput Biol Bioinforma. 2019. in press.
10. Wang J, Yuan J, Yu J, Meng F, Sun P, Li Y, et al. Recursive paleohexaploidization shapes the durian genome. Plant Physiol. 2019. https://doi.org/10.1104/pp.18.00921.
11. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature. 2007;449:463–7.
12. Rendón-Anaya M, Ibarra-Laclette E, Méndez Bravo A, Lan T, Zheng C, Carretero Paulet L, et al. The avocado genome informs deep angiosperm phylogeny, highlights introgressive hybridization, and reveals pathogen-influenced gene space adaptation. PNAS. 2019. in press.
13. Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, et al. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science. 2006;313:1596–604.
14. Liu S, Liu Y, Yang X, Tong C, Edwards D, Parkin IAP, et al. The *Brassica oleracea*, genome reveals the asymmetrical evolution of polyploid genomes. Nat Commun. 2014;5:3930.
15. Dai X, Hu Q, Cai Q, Feng K, Ye N, Tuskan GA, et al. The willow genome and divergent evolution from poplar after the common genome duplication. Cell Res. 2014;24:1274–7.
16. Barke MS, Vogel H, Schranz ME. Paleopolyploidy in the Brassicales: analyses of the Cleome transcriptome elucidate the history of genome duplications in *Arabidopsis* and other Brassicales. Genome Biol Evol. 2009;1:391–9. https://doi.org/10.1093/gbe/evp040.
17. Edger PP, Hall JC, Harkess A, Tang M, Coombs J, Mohammadin S, et al. Brassicales phylogeny inferred from 72 plastid genes: A reanalysis of the phylogenetic localization of two paleopolyploid events and origin of novel chemical defenses. Am J Bot. 2018;105:463–9.
18. Sloane NJA. The Online Encyclopedia of Integer Sequences. Published electronically at https//oeis.org. Accessed 18 June 2019.

Publisher's Note