Zhang *et al. BMC Genomics* 2018, **19**(Suppl 5):287
https://doi.org/10.1186/s12864-018-4624-y

BMC Genomics

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 synteny blocks. Insofar as all the paralog pairs in a paralogous synteny 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

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 the 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 $t_1 < \cdots < t_{n-1}$ for some $n \geq 1$, each existing gene is replaced by $r_1, \ldots, r_{n-1}$ 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 $u_1^{(i)}, \ldots, u_{r_i}^{(i)}$.

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 $a_1^{(i)}, \ldots, a_{r_i}^{(i)}$ be the number of cases where $1, \ldots, r_i$ copies survive

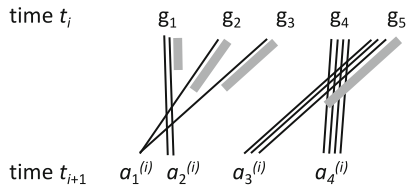

Fig. 1 Event with ploidy $r_i = 4$, showing population of $m_i = 5$ genes at time t_i, each giving rise to 4 progeny, of which $1 \le j \le 4$ survive until time t_{i+1}. $a_j^{(i)}$ is the number of times j progeny survive. In the diagram, thin solid lines represent individual progeny that survive, and thick grey lines represent the total progeny of a gene that do not survive. In this example the only gene, all of whose progeny survive, is g_4. Here $a_1^{(i)} = 2, a_2^{(i)} = a_3^{(i)} = a_4^{(i)} = 1, m_{i+1} = 2 \times 1 + 1 \times 2 + 1 \times 3 + 1 \times 4 = 11$

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 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 $j = 1, \ldots, r_i$.

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

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

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

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

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, $j \ge 2$, there are $\binom{j}{2}$ surviving pairs of genes. If $j = 1$ there are

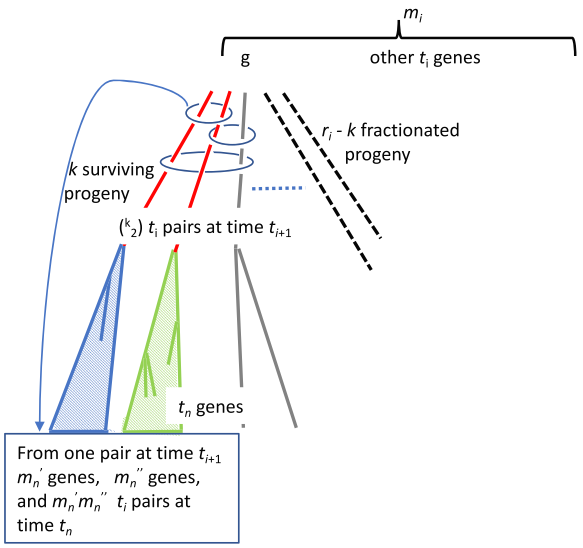

Fig. 2 Counting t_i-pairs. The three unfractionated progeny of gene g define three t_i-pairs, as indicated by three ovals. We follow the pair contained in the uppermost oval, as the two members at time t_{i+1} independently (shaded triangles) evolve into m_n' and m_n'' genes, respectively, at time t_n, defining $m_n' m_n''$ t_i-pairs at time t_n

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

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

At time t_j, for $i + 1 \le j \le n$, 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

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

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

$$m^* \in \{0, \ldots, m_1\}$$
$$\mathbf{E}(m^*) = m_1 \prod_{i=1}^{n-1} u_1^{(i)}, \tag{8}$$

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

The terms $\mathbf{E}\left(d^{(i,i+1)}\right)$ and $\mathbf{E}^{(i+1,n)}(m_n)$ in Eq. (7) both involve calculating probabilites 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 $\mathrm{B}(G, p_i)$, where

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

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

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

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

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

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

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

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

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

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

Inference

The distribution of gene pair similarities is of the form $f(k)$, where $k = k_{\min}, \ldots, k_{\max}$. 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

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

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

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

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 rosid 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

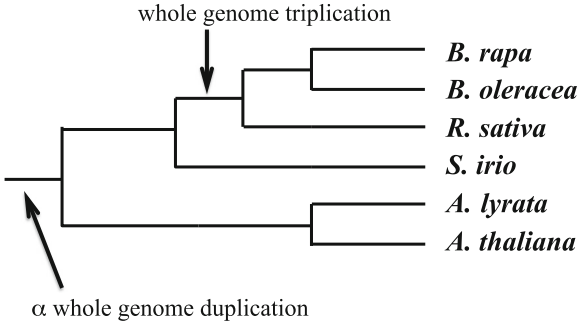

Fig. 3 Evolutionary history of the Brassicales, showing the α genome duplication in the lineage of all the species and the more recent *Brassica* triplication

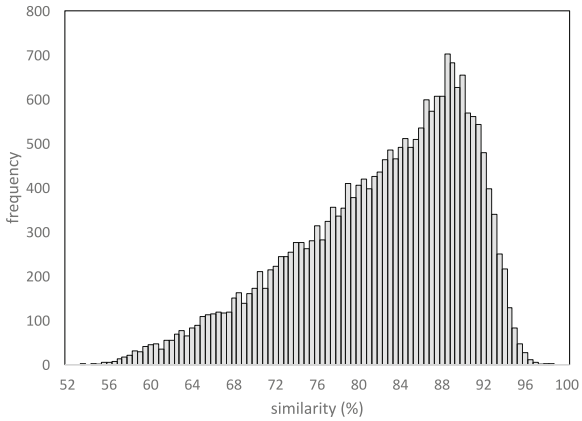

Fig. 4 Duplicate gene similarities in syntenic blocks in the *Brassica rapa* genome self-comparison

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

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:

$$
\begin{aligned}
\mathbf{E}\,(t_1 \text{ pairs}) &= \left(4u^2+4uv+4u+v^2+2v+1\right)w \qquad (17) \\
\mathbf{E}(t_2 \text{ pairs}) &= 3wu+wv+3u+v \\
\mathbf{E}(\text{unpaired}) &= (1-w)(1-u-v).
\end{aligned}
$$

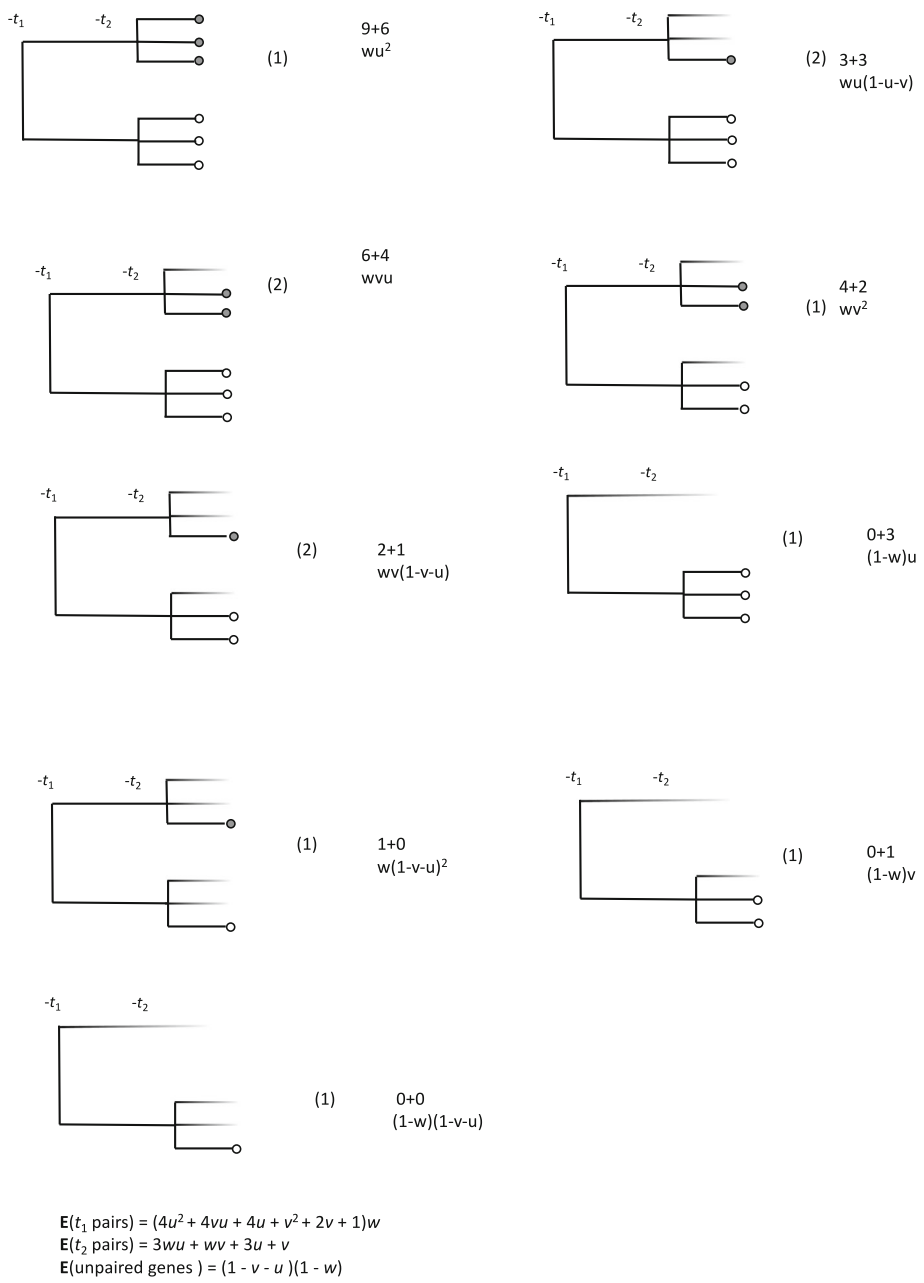

Fig. 5 All paralog trees generated by a genome duplication at time t_1 followed by a triplication at time t_2, both events followed by fractionation. Present-day genes shaded according to most recent event. Numbers above probability are counts of $t_1 + t_2$ pairs, and numbers in parentheses count the number of different trees (only one shown) with the same structure

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

$$\begin{aligned}\mathbf{E}(t_1 \text{ pairs}) &= (x + 3y)(1 + z)^2 \\ \mathbf{E}(t_2 \text{ pairs}) &= 2yz + xz + z \\ \mathbf{E}(\text{unpaired}) &= (1 - z)(1 - x - y)\end{aligned} \tag{18}$$

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:

$$\begin{aligned}\left(4u^2 + 4uv + 4u + v^2 + 2v + 1\right) w &= (x + 3y)(1 + z)^2 \\ 3wu + wv + 3u + v &= 2yz + xz + z \\ (1 - w)(1 - u - v) &= (1 - z)(1 - x - y),\end{aligned} \tag{19}$$

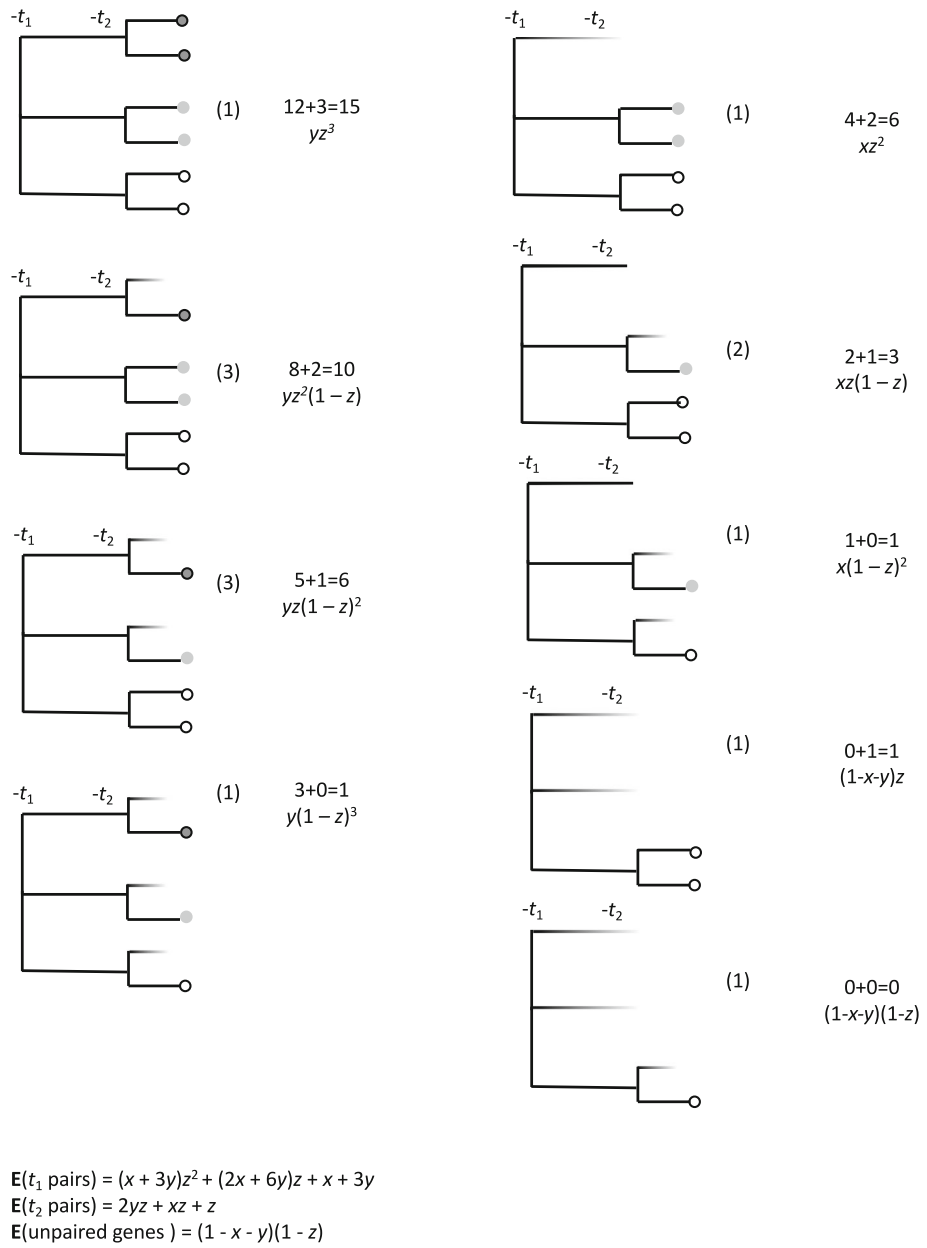

Fig. 6 All paralog trees generated by a genome triplication at time t_1 followed by a duplication at time t_2, both events followed by fractionation. Present-day genes shaded according to most recent event. Numbers above probability are counts of $t_1 + t_2$ pairs, and numbers in parentheses count the number of different trees (only one shown) with the same structure

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:

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

The investigation into the connection between the two models starts with

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

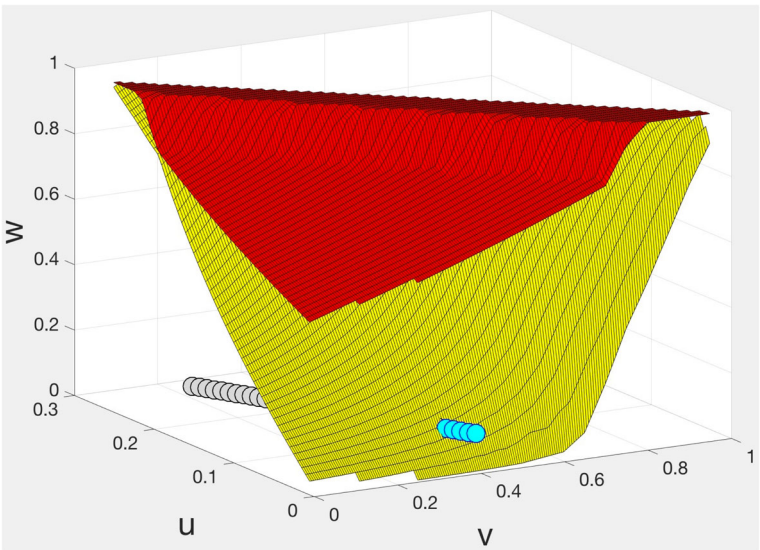

Fig. 7 (u, v, w) space with MLE solutions for (x, y, z) model (blue dots) and (u, v, w) model (blue and grey dots). Corresponding (x, y, z) models exist only for the (u, v, w) values in the volume bounded by the red and yellow sheets

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

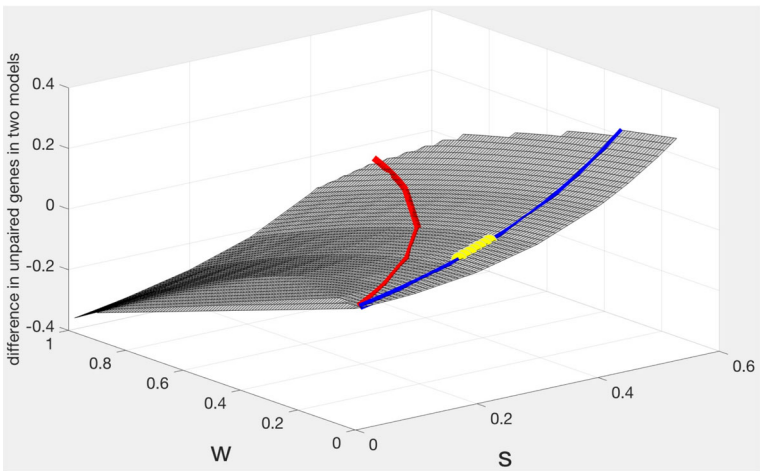

Fig. 8 (s, w) surface showing the difference in the number of unpaired genes between corresponding points in the two models. The red line indicates points where the two models agree on the number of genes, but this does not intersect with the set of MLE solutions on the blue line). In particular, the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

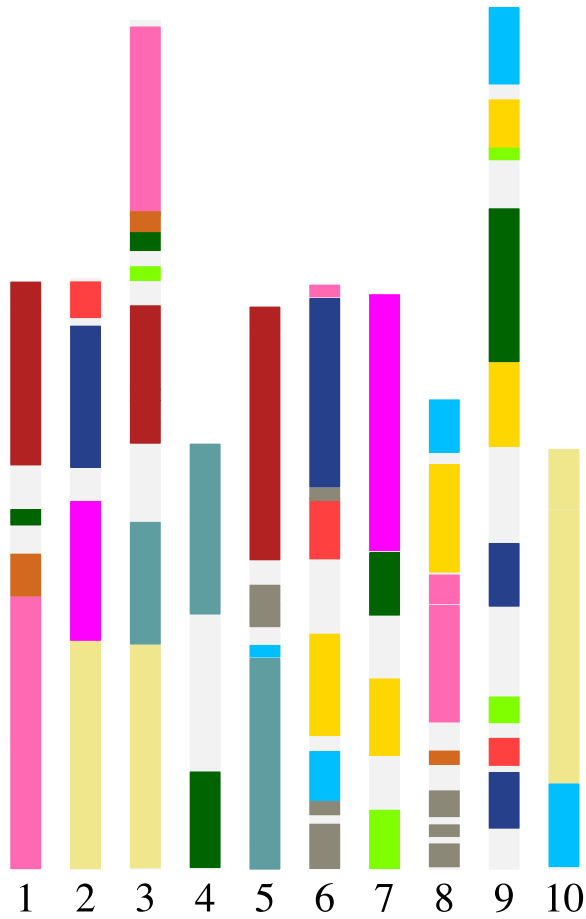

Fig. 9 Regions of the 10 *Brassica rapa* chromosomes coloured by 14 triples of recently diverged regions. Grey regions: triple undetermined

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

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, 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, 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, Chalhoub B, Wang B, Hayward A, Sharpe AG, Park BS, Weisshaar 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. Denoeud 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 JI, 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.