Linkage analysis and haplotype phasing in experimental autopolyploid populations with high ploidy level using hidden Markov models

Marcelo Mollinari[1,*], Antonio Augusto Franco Garcia[2,*],

1 Department of Horticultural Science, Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, USA
2 Department of Genetics, University of São Paulo/ESALQ, Piracicaba, São Paulo, Brazil

* mmollin@ncsu.edu (MM), * augusto.garcia@usp.br (AAFG)

Abstract

Modern SNP genotyping technologies allow to measure the relative abundance of different alleles for a given locus, and consequently to estimate their allele dosage, opening a new road for genetic studies in autopolyploids. Despite advances in genetic linkage analysis in autotetraploids, there is a lack of statistical models to perform linkage analysis in organisms with higher ploidy levels. In this paper, we present a statistical method to estimate recombination fractions and infer linkage phases in full-sib populations of autopolyploid species with even ploidy levels in a sequence of SNP markers using hidden Markov models. Our method uses efficient two-point procedures to reduce the search space for the best linkage phase configuration and reestimates the final parameters using maximum-likelihood estimation of the Markov chain. To evaluate the method, and demonstrate its properties, we rely on simulations of autotetraploid, autohexaploid and autooctaploid populations. The results show the reliability of our approach, including situations with complex linkage phase scenarios in hexaploid and octaploid populations.

Author summary

In this paper we present a multilocus complete solution based in hidden Markov models to estimate recombination fractions and infer the linkage phase configuration in full-sib mapping populations with even ploidy levels under random chromosome segregation. We also present an efficient pairwise loci analysis to be used in cases were the multilocus analysis becomes compute-intensive.

Introduction

Polyploids are organisms with more than two sets of chromosomes. They are very important in agriculture and play a fundamental role in evolutionary processes, such as differentiation of species [1]. The number of sets of chromosomes in an organism is called *ploidy level*. These multiple sets of chromosomes in a polyploid can originate from the combination of chromosomes from different, but related species, or from the duplication of chromosomes from the same species [2, 3]. In the first scenario, they are called *allopolyploids*; in the second, *autopolyploids*. Another way to characterize polyploid organisms is according to their pattern of inheritance. In general, allopolyploids exhibits *disomic* segregation, since homologous chromosomes have more affinity than homeologous chromosomes and tend to form preferential bivalents within each sub-genome [4]. Autopolyploids, however, exhibit more than two homologous chromosomes per homology group. Thus, during the meiosis, they can form either bivalents or multivalents [4, 5]. The expected segregation ratios in autopolyploids vary depending on the type of chromosome configuration that the organism presents during meiosis. If the chromosomes pair randomly, the segregation is called *polysomic* [6–9]. In addition, the homologous chromosomes may have preferential pairing, which can vary from complete preferential (disomic segregation) to complete random (polysomic segregation). Since the molecular mechanics of polyploid organisms are quite complex, this rigid dichotomy is often broken, and organisms can exhibit intermediate modes of inheritance [4, 10]. Throughout this paper, the term autopolyploid (or autotetraploid, autohexaploid, etc.) will refer to polyploid organisms that exhibit polysomic segregation.

Despite all advances in genetic studies in autotetraploids [11–21], there is still a shortage of statistical methods to address organisms with higher ploidy levels, such as sweet potato [22–24], sugarcane [25, 26], some ornamental flowers and forage crops (reviewed in [27]). In this work, we denote as *high-level autopolyploids* those autopolyploid organisms with ploidy level greater than four. A fundamental class of statistical methods that are lagged behind in high-level autopolyploid studies is the construction of genetic maps. A reliable genetic map is a crucial step in quantitative trait loci (QTL) analysis, as well as the assembly of reference genomes and the study of evolutionary processes [28–30]. Although understanding the concept of genetic mapping is rather easy, the construction of such maps in high-level autopolyploids is challenging. Even under bivalent pairing, there is a large number of possible configurations during the meiosis, and this number gets exponentially larger as the ploidy level increases. Denoting m as the ploidy level, it is possible to find up to m different alleles for a locus in one individual. Furthermore, if some of those alleles are not distinguishable, it is necessary to consider the number of copies of each different allelic form, also known as allele *dosage*. Finally, depending on the marker system used to access the genotypic information, in the vast majority of cases, it is not possible to obtain the complete information about a particular locus.

The construction of a genetic map in a full-sib population can be summarized in five basic steps: *i*) estimation of pairwise recombination fractions and associated LOD Scores; *ii*) separation of markers into linkage groups; *iii*) order markers within each linkage group using an optimization technique; *iv*) parental phasing, recombination fraction update and likelihood computation and *v*) if the order is optimal, the map is complete, otherwise, return to step *iii*. Historically, genetic maps in high-level autopolyploids have been constructed using only alleles present in one homologous chromosome, called *single-dose markers* [31, 32]. In a full-sib population, these markers segregate in a 1:1 ratio (if they are present only in one parent), or in a 3:1 ratio (if present in both parents). Given this level of simplification, it is possible to use the five-step procedure coupled with a standard software suitable for backcross diploid populations. Nevertheless, it is well accepted that the use of single-dose markers imposes limitations on the construction of adequate genetic maps. These approaches sub-sample the genome [19, 26], which precludes further consideration of multiallelic effects in models for QTL mapping and subsequent studies. Moreover, there is low statistical power to detect linkage when markers are in repulsion phase configurations [31, 33]. Although some authors have addressed this problem by including multiple dose markers when constructing genetic maps and performing QTL mapping [33, 34], the limitations on the genotyping technologies at the time required that the allelic dosage had to be inferred based on expected segregation rates. Because of the high amount of hidden information imposed by marker systems on those studies [31, 33], the estimation of recombination fraction between multi-dose markers was highly impaired.

Quantitative genotyping technologies for single nucleotide polymorphism (SNPs) evaluation have opened the door for further genetic mapping studies in high-level autopolyploids. It is now possible to measure the abundance of specific alleles within a locus in a polyploid genome [19, 26, 36–39]. This technology, combined with the genotypic distribution in the population [37], makes it possible to infer the allelic dosage by using the ratio between the abundances of the two alternative alleles. Once the dosage of the markers is estimated, the construction of linkage maps can be significantly improved by taking this information into account. [19] and [40] presented works that take into consideration the dosage of quantitative SNP data both in linkage studies and QTL mapping for autotetraploids.

Genetic linkage maps can be constructed based on two-point or multipoint estimates of the recombination fraction. Two-point methods use information of pairs of markers, and even though they are less computationally demanding than multipoint methods, they require a higher amount of information in the markers to provide reliable results. Multipoint approaches, instead, use information of multiple makers present in a linkage group, increasing the statistical efficiency of the analysis [17, 41, 42, 53]. This feature is particularly important in polyploid linkage analysis, where markers are mostly partially informative. One widely used procedure to obtain multipoint estimates is the hidden Markov model (HMM) [41]. The construction of the genetic map using this method provides the estimates of the recombination fractions between all adjacent markers in a linkage group, as well the multipoint likelihood, which has been shown to be an excellent criterion to evaluate and compare linkage phase configurations and orders of makers [42]. [17] presented a statistical framework in which HMMs were applied to reconstruct genetic linkage maps, but it was limited to autotetraploids. Recently, [35] constructed an ultra-dense integrated linkage map for hexaploid chrysanthemum using two-point analysis. However, there is a lack of multipoint procedures that can handle cases where less marker information is available in high ploidy levels.

The main challenges we address in this paper are the inference of the haplotypes of the multiple homologous chromosomes and the multipoint estimation of recombination fractions in high-level polyploids. Although [21] proposed a probabilistic multilocus haplotype reconstruction model for autotetraploids considering double reduction, this remains as an open question for organisms with higher ploidy levels. Our method relies on an HMM and is developed for species with even ploidy levels under random chromosome segregation (complete polysomic inheritance). We also present a two-point method which is capable of dealing with hundreds of markers even in high ploidy level scenarios. Hence, we are proposing solutions for steps *i* and *iv* in high-level autopolyploids. Step *ii* is straightforward from step *i* using clusterization algorithms, as proposed by [50]. Even though step *iii* is a challenging task in genetic mapping, it can be addressed using pairwise recombination fractions or the resulting likelihood of the Markov model as it has been proposed by several studies [43–49]. To evaluate our method, and to show its properties, we rely on simulations of autotetraploid, autohexaploid, and autooctaploid data. The R computer codes to reproduce all simulations and analysis are publicly available.

Methods

In this section, we define the notation used throughout this article and present the probabilistic model for the gamete formation in autopolyploids. Then, we move to the calculation of the *transition probabilities* for adjacent marker loci (Eq 6) and follow to the *initial state* (Eq 7) and *emission probability* distributions (Eqs 8 and 9) which are fundamental in an HMM model. We conclude by explaining the complexity of estimating linkage phases between markers, presenting an efficient two-point algorithm that simplifies the problem in a way that allows the phasing to be inferred using real data.

Notation

Consider one homology linkage group in a mapping population derived from a cross between two autopolyploid individuals P and Q with the same ploidy level (full-sib family). The ploidy level is denoted by m, and can be any even number greater than zero. Let the vectors $\mathcal{P}_k^m = \{P_k^i\}$ and $\mathcal{P}_{k+1}^m = \{P_{k+1}^i\}$, and $\mathcal{Q}_k^m = \{Q_k^i\}$ and $\mathcal{Q}_{k+1}^m = \{Q_{k+1}^i\}$, $i = 1, \cdots, m$, denote the genotype of two adjacent multiallelic loci k and $k+1$ in P and Q, respectively. The superscript i indicates one of the possible alleles for the loci, and each locus has m different alleles in each parent. For example, for a cross between two autohexaploid individuals, $\mathcal{P}_k^6 = \left\{P_k^1, P_k^2, \cdots, P_k^6\right\}$; similarly, this can be done for $\mathcal{P}_{k+1}^6$, $\mathcal{Q}_k^6$ and $\mathcal{Q}_{k+1}^6$. All alleles denoted by the same superscript number are in the same homologous chromosome (e.g., P_k^1 and P_{k+1}^1 are in homologous chromosome 1, etc).

The following assumptions are made to ensure random chromosome segregation [6, 8] and no double reduction [51]: *i*) there is only formation of bivalents during the meiosis; *ii*) there is no preferential pairing during the formation of bivalents; *iii*) all bivalents have the same recombination fraction between loci k and $k+1$; *iv*) bivalents are independent and *v*) there is separation of sister chromatids during the meiosis II. Consequences of violations of these assumptions will be addressed later using simulations.

Bivalent formation

It occurs during meiosis I (more specifically, at the pachytene stage of prophase). In diploid cells, there is only one possible pairing configuration: two duplicated homologous from a homology group pair to form one bivalent. However, in autopolyploid cells, given the previous assumptions, the number of possible pairing configurations, i.e., the number of possible bivalent chromosomal pairing for a given homology group during meiosis is

$$w_m = \frac{1}{\frac{m}{2}!} \prod_{i=1}^{\frac{m}{2}} \binom{2i}{2} \qquad (1)$$

The orientation of the bivalents does not affect the expected frequencies of each gamete type, and therefore will not be considered. For example, for an autotetraploid individual, there are two bivalents and three possible bivalent configurations. Homologous chromosome pair as 1 with 2, and 3 with 4; or, 1 with 3 and 2 with 4; or 1 with 4 and 2 with 3 [52]. We denote $\Psi = \{\psi_j\}$, $j = 1, \cdots, w_m$ a set of all bivalent configurations for a given ploidy level.

Expected gametic frequency for a given bivalent configuration

We will present the expected gametic frequencies considering parent P. Since parent Q undergoes a similar process, it is possible to combine the expected gametic frequencies to obtain the expected genotypic frequency in the full-sib population. Each of the bivalents obtained for a given configuration ψ_j can result in two types of chromosomes for loci k and $k+1$: *parental*, which results from bivalents with zero or any other even number of recombinations between k and $k+1$; and *recombinants*, which results from bivalents with any odd number of recombinations. As presented by [34], the probabilities of all chromosome types for any single bivalent can be represented always as

$$\mathbf{V} = \begin{bmatrix} \Pr(P_k^i, P_{k+1}^i) & \Pr(P_k^i, P_{k+1}^{i'}) \\ \Pr(P_k^{i'}, P_{k+1}^i) & \Pr(P_k^{i'}, P_{k+1}^{i'}) \end{bmatrix} = \begin{bmatrix} \frac{1-r_k}{2} & \frac{r_k}{2} \\ \frac{r_k}{2} & \frac{1-r_k}{2} \end{bmatrix}$$

where r_k is the recombination fraction between k and $k+1$, $i \neq i'$. For a given configuration ψ_j, the expected frequencies for all possible gametes derived from that configuration is

$$\mathbf{V}_1 \otimes \cdots \otimes \mathbf{V}_{\frac{m}{2}}$$

where $\otimes$ denotes the Kronecker product of matrices and subscripts in $\mathbf{V}$ indicate the corresponding bivalent. All elements of this product are of the form

$$\frac{(1-r_k)^{\frac{m}{2}-l}(r_k)^l}{2^{\frac{m}{2}}}$$

where l denotes the number of total recombinant bivalents between loci k and $k+1$, $l \in \{0, \cdots, m/2\}$. From this, we can define the probability of observing any gamete (for two loci) given a bivalent configuration ψ_j as

$$\Pr(\mathbf{p}_k,\ \mathbf{p}_{k+1}|\psi_j) = \begin{cases} \frac{(1-r_k)^{\frac{m}{2}-l}(r_k)^l}{2^{\frac{m}{2}}} & \text{if } \psi_j \text{ is } \textit{consistent} \text{ with the gamete } \{\mathbf{p}_k,\ \mathbf{p}_{k+1}\} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where vectors $\mathbf{p}_k$ and $\mathbf{p}_{k+1}$ denote a subset of $\frac{m}{2}$ alleles present in $\mathcal{P}_k^m$ and $\mathcal{P}_{k+1}^m$, respectively; $\{\mathbf{p}_k, \mathbf{p}_{k+1}\}$ indicates a gamete for loci k and $k+1$ from parental P. *Consistent* means that the gamete can be produced from bivalent configuration ψ_j. Notice that some gametes cannot be obtained from ψ_j once the bivalents are formed.

Since we assume that alleles with the same superscript are in the same homologous chromosome, l can be obtained by a simple examination of superscripts of elements contained in $\mathbf{p}_k$ and $\mathbf{p}_{k+1}$. Consider, for example, $\psi_1 = \{(1,2),(3,4),(5,6)\}$ ($m = 6$, Fig 1). If one observes $\mathbf{p}_k = \left\{P_k^1, P_k^3, P_k^5\right\}$ and $\mathbf{p}_{k+1} = \left\{P_{k+1}^1, P_{k+1}^4, P_{k+1}^6\right\}$, the number of recombinant chromosomes is $l = 2$. Therefore, $\Pr\left(\{P_k^1, P_k^3, P_k^5\}, \{P_{k+1}^1, P_{k+1}^4, P_{k+1}^6\} \mid \psi_1\right) = \frac{(1-r_k)(r_k)^2}{2^3}$. On the other hand, $\Pr\left(\{P_k^1, P_k^2, P_k^5\}, \{P_{k+1}^1, P_{k+1}^2, P_{k+1}^5\} \mid \psi_1\right) = 0$, since it is impossible to obtain this gamete from configuration ψ_1, i.e., it is not *consistent* with ψ_1.

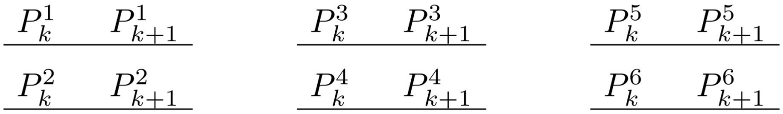

Figure 1. One possible pairing configuration in an autohexaploid, namely ψ_1. P_k^i denotes one allele present in homologous chromosome i for loci k in parent P. Notice that some allelic configurations, such as $\left(\{P_k^1, P_k^2, P_k^5\}, \{P_{k+1}^1, P_{k+1}^2, P_{k+1}^5\}\right)$, are impossible to be obtained in this bivalent pairing. In this case, the homologous chromosomes containing alleles P_k^1 and P_k^2 will migrate to opposite poles of the cell during meiosis I. Therefore, P_k^1 and P_k^2 will not be present in the same gamete.

Gametic frequency unconditional to bivalent configurations

In reality ψ_j is unknown, thus the conditional probability given by Eq (2) must be considered for all possible ψ_j. The probability of observing a gamete $\{\mathbf{p}_k,\ \mathbf{p}_{k+1}\}$, unconditional to ψ_j, can be expressed as

$$\Pr(\mathbf{p}_k,\ \mathbf{p}_{k+1}) = \sum_{j=1}^{w_m} \Pr(\mathbf{p}_k,\ \mathbf{p}_{k+1}|\psi_j)\Pr(\psi_j) \tag{3}$$

It is important to notice that only a subset of Ψ is *consistent* with the observed gamete, and consequently $\Pr(\mathbf{p}_k,\ \mathbf{p}_{k+1} \mid \psi_j) > 0$ only for some ψ_j's. Fig 2 shows a graphical representation of Eqs 2 and 3 for autohexaploid gametes.

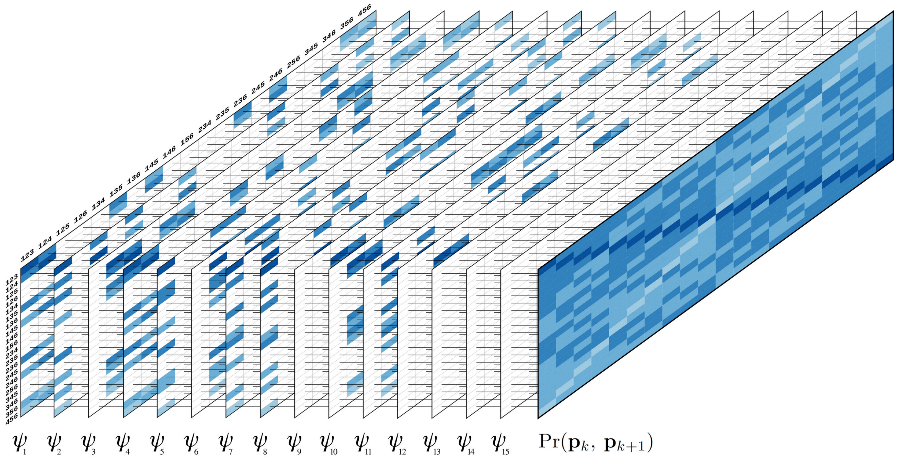

Figure 2. Graphical representation of Eqs 2 and 3 for autohexaploid gametes. The first 15 tables represent the gametic probabilities given different bivalent configurations $\psi_.$ (Eq 2). The rows and the columns indicate gametic configurations for loci k and $k+1$, respectively. For simplification, only the superscripts of the gametic configurations were presented. For example, row 123, column 123, represent the gamete $\left(\{P_k^1, P_k^2, P_k^3\}; \{P_{k+1}^1, P_{k+1}^2, P_{k+1}^3\}\right)$. Colored cells indicate the probability of gametic configurations *consistent* with the bivalent configuration $\psi_.$. The color scale indicates the number of recombinant bivalents associated to the gametic probability varying from 0 (dark blue) to 3 (light blue). Blank cells indicate *non-consistent* configurations. The far right full table represents the sum over all ψ configurations, weighted by their probability (Eq 3).

The probability of observing a specific gamete is always the same for each ψ_j in this consistent subset (Eq 2). Therefore, under random pairing (assumption *ii*), our task reduces to finding the number of elements in this subset that are consistent with the observed gamete and multiply $\Pr(\mathbf{p}_k,\ \mathbf{p}_{k+1}|\psi_j)\Pr(\psi_j)$ by this number. The result is the probability of observing a gamete unconditional to the bivalent configuration.

For every gamete, l can change from zero to $m/2$ recombinant homologous chromosomes. The observed gamete is the result of homologous chromosomes that migrate to one pole of the cell at anaphase I. Since we are assuming that there is separation of sister chromatids during anaphase II, if $l = 0$ (all chromosomes are of parental type), there is no information about the pairing configuration of the homologous chromosomes that migrate to the opposite pole of the cell. In this situation, there are $\left(\frac{m}{2}\right)!$ possible pairing configurations, and the number of possible ψ_j that can produce gametes with $l = 0$ is $\left(\frac{m}{2}\right)!$. Therefore, for $l > 0$, there are $\left(\frac{m}{2} - l\right)!$ possible pairing configurations of parental chromosomes. For the remaining l recombinant chromosomes, the number of possible pairing configurations is $l!$. Thus, the total number of possible pairing configurations that can produce a specific gamete is

$l!\left(\frac{m}{2}-l\right)!$. This is precisely the number of elements in the subset of Ψ consistent with the observed gamete. Given the assumption of no preferential pairing during the formation of bivalents, $\Pr(\psi_j) = \frac{1}{w_m}$, the probability of a gamete $\{\mathbf{p}_k,\ \mathbf{p}_{k+1}\}$, unconditional to ψ_j, can be simplified to

$$\Pr(\mathbf{p}_k,\ \mathbf{p}_{k+1}) = \frac{l!\left(\frac{m}{2}-l\right)!}{w_m}\frac{(1-r_k)^{\frac{m}{2}-l}(r_k)^l}{2^{\frac{m}{2}}} \tag{4}$$

Map reconstruction via hidden Markov model

The construction of a genetic map involves the estimation of the genetic distance and order between markers within linkage groups. If the origin of the haplotypes (i.e., linkage phase) for the parents of the mapping population is unknown, it also needs to be estimated. For several years, hidden Markov models have been proven to be an excellent avenue for obtaining these estimates [17, 41, 42, 53]. The multipoint likelihood obtained using HMMs is employable as a criterion to compare marker orders and judge which one is best, and also to provide a reliable estimation of recombination fraction and linkage phases. [54] defines an HMM as a generative process composed of three well-defined probability distributions: *transition*, *initial state* and *emission*. In genetic mapping context, the transition probability distribution is defined as the probability of having a particular genotype at position $k+1$, given the genotype at position k. Using Eq (4) the gametic transition probabilities $\Pr(\mathbf{p}_{k+1}|\mathbf{p}_k)$, or the conditional probability of a gamete genotype at loci $k+1$ given the gamete genotype at loci k, is simply

$$\Pr(\mathbf{p}_{k+1}|\mathbf{p}_k) = \frac{\Pr(\mathbf{p}_k,\ \mathbf{p}_{k+1})}{\Pr(\mathbf{p}_k)}$$

Under random chromosome segregation, both $\mathbf{p}_k$ and $\mathbf{p}_{k+1}$ can have $\binom{m}{\frac{m}{2}}$ different genotypes. Let $\boldsymbol{\Theta}_P^m = \left\{\theta_{P,i}^m\right\}$, $i = 1, \cdots, \binom{m}{\frac{m}{2}}$ denote all possible genotypes that $\mathbf{p}_k$ can assume for loci k. Also, assume that genotypes in $\boldsymbol{\Theta}_P^m$ are arranged according to the lexicographical order of their superscripts. For example, in an autotetraploid, $\boldsymbol{\Theta}_P^4 = \{(P_k^1, P_k^2), (P_k^1, P_k^3), (P_k^1, P_k^4), (P_k^2, P_k^3), (P_k^2, P_k^4), (P_k^3, P_k^4)\}$ for locus k. After some simplifications (see S1 Appendix) the transition probability, i.e., the conditional probability of a gametic genotype $\theta_{P,i}^m$ in locus $k+1$ given the gametic genotype $\theta_{P,i'}^m$ in locus k, is

$$\Pr(\mathbf{p}_{k+1} = \theta_{P,i'}^m|\mathbf{p}_k = \theta_{P,i}^m) = \frac{(1-r_k)^{\frac{m}{2}-l}(r_k)^l}{\binom{\frac{m}{2}}{l}} \tag{5}$$

where $i, i' \in \{1, \cdots, \binom{m}{\frac{m}{2}}\}$. The initial state and the emission probability distributions will be addressed in the next section (Eqs 7 to 9).

Including information of both parents

Any given individual in a full-sib population is formed by the union of gametes from both parents, P and Q. Each parent can form $\binom{m}{\frac{m}{2}}$ different gametes for locus k. Since the formation of gametes in both parents is independent, the genotypic transition probability distribution can be written as

$$\begin{aligned}\Pr(\mathcal{G}_{k+1,j'}^m|\mathcal{G}_{k,j}^m) &= \Pr(\mathbf{p}_{k+1} = \theta_{P,i'}^m|\mathbf{p}_k = \theta_{P,i}^m)\Pr(\mathbf{q}_{k+1} = \theta_{Q,h'}^m|\mathbf{q}_k = \theta_{Q,h}^m) \\ &= \frac{(1-r_k)^{m-l_P-l_Q}(r_k)^{l_P+l_Q}}{\binom{\frac{m}{2}}{l_P}\binom{\frac{m}{2}}{l_Q}}\end{aligned} \tag{6}$$

where $\mathcal{G}^m_{k,j}$ denotes the genotype of an individual derived from the union of gametes $\theta^m_{P,i}$ and $\theta^m_{Q,h}$ at locus k. The same reasoning applies to $\mathcal{G}^m_{k+1,j'}$; $i, i', h, h' \in \{1, \cdots, \binom{m}{\frac{m}{2}}\}$, $j = (i-1)\binom{m}{\frac{m}{2}} + h$ and $j' = (i'-1)\binom{m}{\frac{m}{2}} + h'$. l_P and l_Q denote the number of recombinant bivalents between loci k and $k+1$ in parents P and Q, respectively. Let $g_m = \binom{m}{\frac{m}{2}}^2$ denote the number of possible genotypes derived from the cross between individuals P and Q. For simplification and without loss of generality, let $t_k(j, j') = \Pr(\mathcal{G}^m_{k+1,j'} | \mathcal{G}^m_{k,j})$. For a comprehensive example of the transition probabilities and the indexation used in Eq. 6, see Table 8 in S3 Appendix.

Given a ploidy level m and a recombination fraction r_k, the only information required to obtain $t_k(j, j')$ in Eq (6) is l_P and l_Q. Since the genotypes in $\mathbf{\Theta}^m_P$ and $\mathbf{\Theta}^m_Q$ are arranged according to the lexicographical order of their superscripts, it is possible to obtain (l_P, l_Q) for any given pair (j, j') using the algorithm presented in S2 Appendix. Although the number of possible transitions between positions k and $k+1$ is $(g_m)^2$, which can be a very large number even for modest ploidy levels, it is possible to obtain the transition between any specific genotypes in j and j' without computing the entirety of the transition space.

The initial state distribution is the probability of observing a specific genotype. Given the assumption that there is no preferential pairing during the formation of bivalents, a uniform probability density function can be employed as the initial state probability function

$$\gamma_j = \Pr(\mathcal{G}^m_{1,j}) = \frac{1}{g_m}, \; j \in \{1, \cdots, g_m\} \tag{7}$$

To this point, both transition and initial state distributions consider different allelic variants for all m homologous chromosomes in both parents. This scenario can only be achieved when using fully informative markers. In reality, autopolyploid species may have the same allelic variant in some homologous chromosomes. Besides, even if all homologous have different allelic forms, modern genotyping platforms are usually capable of detecting polymorphisms at the nucleotide level (SNPs), which are essentially biallelic. Due to this lack of identity between the observed data and the full transition space, we make use of the emission function, which is defined as the probability of observing a molecular phenotype given a genotype $\mathcal{G}^m_{k,j}$.

The detection of the allelic variants in modern genotyping platforms is based on the abundance of different alternative nucleotides. In the autopolyploid setting, this can be translated as the *dosage* of a SNP at a specific locus. The dosage of a SNP can be estimated using the ratio between the abundance of its two allelic forms. Several methods were proposed to perform this task including [36], [37] and [38]. Here we introduce a biallelic derivation of the emission probability distribution. Although the function presented here use biallelic information, other distributions can be derived for partial informative multiallelic marker systems following the same reasoning.

Let $d^k_P, d^k_Q \in \{0, \cdots, m\}$ denote the *observed dosage* of one allelic form in locus k for parents P and Q, respectively. The choice of the allelic form denoted by d^k_P is arbitrary, as long as the same allelic form is used in d^k_Q. The dosage observed in parent P can be originated from alleles present in d^k_P of the m homologous chromosomes. Let $\phi^k_P = \{\varphi^k_P : \varphi^k_P \subseteq \mathcal{P}^m_k, \#\{\varphi^k_P\} = d^k_P\}$ denote a set of size $\binom{m}{d^k_P}$ containing all possible subsets in $\mathcal{P}^m_k$ that originate the observed dosage d^k_P. The operator $\#\{.\}$ is the cardinality of a set. The same reasoning applies for ϕ^k_Q. For instance, in an autotetraploid, if $d^k_P = 3$, the three doses present in locus k can be derived from four distinct subsets $\phi^k_P = \big\{(P^1_k, P^2_k, P^3_k), (P^1_k, P^2_k, P^4_k), (P^1_k, P^3_k, P^4_k), (P^2_k, P^3_k, P^4_k)\big\}$. Given two particular subsets φ^k_P and φ^k_Q in ϕ^k_P and ϕ^k_Q, each one of the g_m genotypic states in the full transition space can be associated to a dosage. The dosage associated to the j-th state is obtained by counting the number of alleles present in the intersection between the parental allelic set $(\varphi_P^k \cup \varphi_Q^k)$ and $\mathcal{G}_{k,j}^m$. Thus, the emission function can be defined as

$$b_j(O) = \Pr(O|\mathcal{G}_{k,j}^m, \varphi_P^k, \varphi_Q^k) = \begin{cases} 1-\epsilon & \text{if} \quad O = \delta(k,j) \\ \frac{\epsilon}{m} & \text{otherwise} \end{cases} \tag{8}$$

where $\delta(k,j) = |(\varphi_P^k \cup \varphi_Q^k) \cap \mathcal{G}_{k,j}^m|$ and ϵ denotes the global genotype error rate. In addition to the punctual estimate of the dosage, the genotyping calling methods cited above also provide the probability distribution of the dosages for a particular marker for all individuals of the biparental population. If this information is available, a more general emission function can be derived. Instead of modeling a global error rate ϵ, we use the prior information provided by the genotyping calling procedure. Let $\boldsymbol{\pi}_k = \{\pi_i^k\}_{(1\times m+1)}$ denote the probability distribution vector associated to the dosages $0, \cdots, m$ at position k for a particular individual in the biparental population. For example, $\boldsymbol{\pi}_k = \left\{0, \frac{1}{6}, \frac{2}{3}, \frac{1}{6}, 0\right\}$ denotes a tetraploid individual with probabilities $\frac{1}{6}$, $\frac{2}{3}$ and $\frac{1}{6}$ of having one, two and three doses, respectively, and zero for the remaining ones. Then, the emission probability function can be written as

$$b_j(O) = \Pr(O|\mathcal{G}_{k,j}^m, \varphi_P^k, \varphi_Q^k, \boldsymbol{\pi}_k) = \pi_{\delta(k,j)+1}^k \tag{9}$$

In this case, the observation O can be any dosage from 0 to m and the information about the genotypes will be contained in the probability distribution of the dosages $\boldsymbol{\pi}_k$. Thus, the probability of observing any dosage given a genotype $\mathcal{G}_{k,j}^m$ associated to a particular dosage $\delta(k,j)$ can be obtained by simply assessing the corresponding value in the probability distribution provided by the genotype calling procedure. Notice that Eq 8 can be reduced to Eq 9 using the appropriate $\boldsymbol{\pi}_k$. For example, in autotetraploids, when the observed dosage for locus k is one, $O = 1$, $\boldsymbol{\pi}_k = \left\{\frac{\epsilon}{m}, 1-\epsilon, \frac{\epsilon}{m}, \frac{\epsilon}{m}, \frac{\epsilon}{m}\right\}$. Moreover, for missing values, it is possible to use the probability distribution of the genotypic classes under polysomic segregation, as presented by [37].

Multipoint likelihood and the estimation of recombination fraction

Suppose there are z markers in a homology group in a known order represented by $M_1, \cdots, M_k, \cdots, M_z$. Let $\mathbf{r} = (r_1, \cdots, r_k, \cdots, r_{z-1})$ denote the recombination fraction vector between all marker intervals in this sequence. Also, assume linkage phase configurations in parents P and Q denoted respectively by $\Phi_P = (\varphi_P^1, \cdots, \varphi_P^k, \cdots, \varphi_P^z)$ and $\Phi_Q = (\varphi_Q^1, \cdots, \varphi_Q^k, \cdots, \varphi_Q^z)$. The sequence of observations for the z markers is denoted by $(O_1, \cdots, O_k, \cdots, O_z)$ and its underlying probability distributions is denoted by $\Pi = (\boldsymbol{\pi}_1, \cdots, \boldsymbol{\pi}_k, \cdots, \boldsymbol{\pi}_z)$. The likelihood of $M_1, \cdots, M_k, \cdots, M_z$ can be obtained using Eqs (6), (7) and (9) following the classical *forward procedure* [54]. Let $\alpha_k(j) = \Pr(O_1, \cdots, O_k; \mathcal{G}_{k,j}^m \mid \mathbf{r}, \Phi_P, \Phi_Q, \Pi)$ denote the probability of the partial observation sequence $(O_1, \cdots, O_k)$ and genotype $\mathcal{G}_{k,j}^m$, $j \in \{1, \cdots, g_m\}$ given the sequence of recombination fractions $\mathbf{r}$, the linkage phase configurations Φ_P and Φ_Q and the probability distributions for the sequence of observations Π. The forward procedure follows the steps below:

1. Initialization:

$$\alpha_1(j) = \gamma_j b_j(O_1),\; j = 1, \cdots, g_m \tag{10}$$

2. Induction:

$$\alpha_{k+1}(j') = \left[\sum_{j}^{g_m} \alpha_k(j)\, t_k(j,j')\right] b_{j'}(O_{k+1}) \tag{11}$$

where $\quad k = 1, \cdots, z-1$ and $j' = 1, \cdots, g_m$

3. Termination:

$$\Pr(O_1, \cdots O_z | \mathbf{r}, \Phi_P, \Phi_Q, \Pi) = \sum_{j=1}^{g_m} \alpha_z(j) \tag{12}$$

Then, the likelihood of the model is defined as

$$\prod_{i=1}^{n} \Pr(O_{1,i}, \cdots, O_{z,i} | \mathbf{r}, \Phi_P, \Phi_Q, \Pi_i) \tag{13}$$

where n is the number of individuals in the full-sib population, $O_{1,i}, \cdots, O_{z,i}$ is the sequence of marker observations for individual i and Π_i is a $(m+1) \times z$ matrix where the k-th column denotes the probability distributions associated to the marker M_k, individual i. The multipoint maximum likelihood estimate of $\mathbf{r}$ can be obtained using the *forward-backward* procedure coupled with the EM algorithm [54]. For the backward procedure, consider the variable $\beta_k(j) = \Pr(O_{k+1}, \cdots, O_z \mid \mathcal{G}^m_{k,j}, \mathbf{r}, \Phi_P, \Phi_Q, \Pi)$ as the probability of the partial observation sequence from $k+1$ to z, given the genotype $\mathcal{G}^m_{k,j}$, the recombination fraction vector $\mathbf{r}$, the linkage phase configurations Φ_P and Φ_Q and the probability distributions for the sequence of observations Π. The solution to $\beta_k(j)$ was also described by [54] as follows:

1. Initialization:

$$\beta_z(j) = 1, \; j = 1, \cdots, g_m \tag{14}$$

2. Induction:

$$\beta_k\,(j) = \sum_{j'}^{g_m} t_k(j, j') b_{j'}(O_{k+1}) \beta_{k+1}\,(j') \tag{15}$$

where $\quad k = z-1, z-2, \cdots, 1$ and $j = 1, \cdots, g_m$

To estimate the recombination fraction for all intervals in the marker sequence we need to define $\xi_k(j, j')$ as the probability of state $\mathcal{G}^m_{k,j}$ at position k and state $\mathcal{G}^m_{k+1,j'}$ at position $k+1$ given the sequence of observations $O_1, \cdots O_z$ and their underlying probability distributions Π, the recombination fraction vector $\mathbf{r}$ and the linkage phase configurations Φ_P and Φ_Q

$$\begin{aligned} \xi_k(j, j' \mid \mathbf{r}) &= \Pr(\mathcal{G}^m_{k,j}, \mathcal{G}^m_{k+1,j'} \mid O_1, \cdots O_z, \Pi, \mathbf{r}, \Phi_P, \Phi_Q) \\ &= \frac{\alpha_k(j) t_k(j, j') b_{j'}(O_{k+1}) \beta_{k+1}(j')}{\sum_{j=1}^{g_m} \sum_{j'=1}^{g_m} \alpha_k(j) t_k(j, j') b_{j'}(O_{k+1}) \beta_{k+1}(j')} \end{aligned} \tag{16}$$

The recombination frequency r_k can be estimated through an iterative process using

$$r_k^{s+1} = \sum_{i=1}^{n} \sum_{j=1}^{g_m} \sum_{j'=1}^{g_m} \frac{\xi_k(j, j' \mid \mathbf{r}^s) \phi(j, j')}{n} \tag{17}$$

where $\xi_k(j, j' \mid \mathbf{r}^s)$ is calculated for individual i, $\phi(j, j') = \frac{(l_P + l_Q)}{m}$ is the proportion of recombinations between markers k and $k+1$ for individuals with genotypes $\mathcal{G}^m_{k,j}$ and $\mathcal{G}^m_{k+1,j'}$ and $\mathbf{r}^s$ is the vector of recombination fractions in the iteration (s) and $\mathbf{r}^{s+1}$ is the updated recombination fraction vector [55].

Estimation of linkage phase

Let the Cartesian product $\phi_P^1 \times \cdots \times \phi_P^k \times \cdots \times \phi_P^z = \{(\varphi_P^1, \cdots, \varphi_P^k, \cdots, \varphi_P^z) \mid \varphi_P^i \in \phi_P^i, i = 1, \cdots, z\}$ denotes a set containing all possible linkage phase configurations in parent P. Also, let $\Phi = \{\Phi^u\} = (\phi_P^1 \times \cdots \times \phi_P^k \times \cdots \times \phi_P^z) \times (\phi_Q^1 \times \cdots \times \phi_Q^k \times \cdots \times \phi_Q^z)$, $u = 1, \cdots, \prod_{k=2}^{z} \binom{m}{d_P^k}\binom{m}{d_Q^k}$, denote a set containing all possible linkage phase configurations in both parents. The probability of the linkage phase configurations can be obtained using Bayes' rule

$$\Pr(\Phi^u \mid \mathbf{O}, \mathbf{\Pi}, \mathbf{r}) = \frac{\prod_{i=1}^{n} \Pr(O_{1,i}, \cdots, O_{z,i} | \mathbf{r}, \mathbf{\Pi}_i, \Phi^u) \Pr(\Phi^u)}{\sum_{\Phi^u \in \Phi} \prod_{i=1}^{n} \Pr(O_{1,i}, \cdots, O_{z,i} | \mathbf{r}, \mathbf{\Pi}_i, \Phi^u) \Pr(\Phi^u)} \tag{18}$$

where $\mathbf{O}$ is an array containing the observation for z markers in n individuals, and $\mathbf{\Pi}$ is the underlying probability distribution for all marker observations. Since the prior probability $\Pr(\Phi^u)$ can be assumed to be uniform, the posterior probability is proportional to the likelihood of the model, which can be used to select the best linkage phase configuration. Depending on the dosage and number of markers, some of these configurations are equivalent and will result in the same likelihood. The search space for the best linkage phase configuration can be unwieldy depending on the ploidy level, dosage and number of markers. Also, the transition space on the HMM gets larger as the ploidy level increases. To circumvent these problems, we propose a very efficient two-point procedure to reduce the search space for linkage phases.

Two-point algorithm for high-level autopolyploids

When the linkage analysis is conducted only in two markers (two-point analysis), the information contained in these markers does not propagate into the rest of the chain. Thus, based on the dosage and linkage phase configuration of the markers involved in the analysis, the g_m genotypic states present in the full transition space can be collapsed into a small number of states, and a straightforward likelihood function can be derived. It is worthwhile to mention that the estimates obtained using the two-point procedure are the same as those obtained using the multipoint algorithm for two markers. However, the computation is extremely faster.

Consider a biallelic marker in an autopolyploid biparental cross with ploidy m. The number of possible genotypic states in the progeny for a given locus at position k is $u(d_P^k) + u(d_Q^k) + 1$, where the operator $u(x) = \left|\left|x - \frac{m}{2}\right| - \frac{m}{2}\right|$ and $|.|$ denotes module. For example, in an autohexaploid biparental cross, if the dosage of the marker at position k in parent P is two ($d_P^k = 2$) and in parent Q is three ($d_Q^k = 3$), the number of possible genotypic classes expected in the progeny is six. Depending on the linkage phase configuration, each of the g_m genotypic states in the full transition space corresponds to one of these expected genotypic classes, as presented in the emission function (Eqs 8 and 9). Thus, in the previous example, all the g_m states could be collapsed into six different classes. To perform this reduction of dimensionality, let $\mathcal{D}_k^m \in \{0, \cdots, m\}$ denote one of the possible genotypes based on the dosage of one individual in the progeny of an autopolyploid biparental cross for position k with ploidy m. The joint probability of $\mathcal{D}_k^m$ and $\mathcal{D}_{k'}^m$, for a given genotypic configuration at positions k and k' can be written as

$$\Pr(\mathcal{D}_k^m, \mathcal{D}_{k'}^m \mid \varphi_P^k, \varphi_Q^k, \varphi_P^{k'}, \varphi_Q^{k'}) = \sum_{j \in \mathrm{T}_k} \sum_{j' \in \mathrm{T}_{k'}} \Pr(\mathcal{G}_{k',j'}^m | \mathcal{G}_{k,j}^m) \Pr(\mathcal{G}_{k,j}^m) \tag{19}$$

where $\mathrm{T}_k = \{j \mid \delta(k,j) = \mathcal{D}^m_k \ , j = 1, \cdots, g_m\}$ and $\delta(k,j)$ was defined in Eq 8; the same applies to $\mathrm{T}_{k'}$. Since in a two-point analysis the probability distribution of the genotypic states in locus k can be assumed to be uniform, i.e., $\Pr(\mathcal{G}^m_{k,j}) = \frac{1}{g_m}$, Eq (19) can be rewritten as a sum of weighted terms from Eq (6)

$$\Pr(\mathcal{D}^m_k, \mathcal{D}^m_{k'} \mid r_k, \varphi^k_P, \varphi^{k'}_P, \varphi^k_Q, \varphi^{k'}_Q) = \sum_{l_P=0}^{\frac{m}{2}} \sum_{l_Q=0}^{\frac{m}{2}} \zeta_{\mathrm{T}_k,\mathrm{T}_{k'}}(l_P, l_Q) \frac{(1-r_k)^{m-l_P-l_Q}(r_k)^{l_P+l_Q}}{\binom{\frac{m}{2}}{l_P}\binom{\frac{m}{2}}{l_Q}} \tag{20}$$

where

$$\zeta_{\mathrm{T}_k,\mathrm{T}_{k'}}(l_P, l_Q) = \frac{1}{g_m} \sum_{j \in \mathrm{T}_k} \sum_{j' \in \mathrm{T}_{k'}} h(j, j'; l_P, l_Q)$$

$h(j,j';l_P,l_Q)$ is 1 if (j,j') corresponds to (l_P,l_Q) according to the procedure described in S2 Appendix and zero otherwise. Eq 20 can be expressed in matrix form as

$$\mathbf{A}_{\varphi^k_P,\varphi^{k'}_P,\varphi^k_Q,\varphi^{k'}_Q}(r_k) = \left\{\Pr(\mathcal{D}^m_k = i-1, \mathcal{D}^m_{k'} = j-1 \mid r_k, \varphi^k_P, \varphi^{k'}_P, \varphi^k_Q, \varphi^{k'}_Q)_{i,j}\right\} \tag{21}$$

where $\mathbf{A}_{\varphi^k_P,\varphi^{k'}_P,\varphi^k_Q,\varphi^{k'}_Q}(r_k)$ is a $(m+1)\times(m+1)$ matrix. Yet, in a two-point analysis with biallelic markers, the linkage phase configuration can be summarized in an ordered pair $(w^{k,k'}_P, w^{k,k'}_Q)$ indicating the number of homologous chromosomes that share allelic variants for loci k and k' in parents P and Q, respectively. For a given pair $(\varphi^k_P, \varphi^{k'}_P)$, $w^{k,k'}_P = \#\{x^k_P \cap x^{k'}_P\}$, where x^k_P and $x^{k'}_P$ denote the set of homologous chromosomes inherited by parent P in positions k and k', which can be assessed using the superscripts in φ^k_P and $\varphi^{k'}_P$. $\#\{.\}$ indicates the cardinality of the set. Notice that φ^k_P and $\varphi^{k'}_P$ can assume several linkage phase configurations resulting in the same $w^{k,k'}_P$. Let $\Phi^{k,k'}_P = \phi^k_P \times \phi^{k'}_P$ denote a set containing all possible pairs $(\varphi^k_P, \varphi^{k'}_P)$ for a given pair $(d^k_P, d^{k'}_P)$. In this set, there are $\min\{u(d^k_p), u(d^{k'}_p)\} + 1$ partitions, each one corresponding to a different $w^{k,k'}_P$. Fig 3 shows an example of $\Phi^{k,k'}_P$ for $(d^k_P = 2, d^{k'}_P = 2)$ in an autotetraploid homology group. The size of the set is 36, and it can be subdivided into three partitions where $w^{k,k'}_P = 2$, $w^{k,k'}_P = 1$ and $w^{k,k'}_P = 0$.

In a two-point context, the likelihood function derived from any of the configurations belonging to the same partition (same $w^{k,k'}_P$) will be the same. Thus, any of them can be used to obtain the likelihood function for a given $w^{k,k'}_P$. Let $(\varphi^k_P, \varphi^{k'}_P)^*$ denote one of the possible pairs $(\varphi^k_P, \varphi^{k'}_P)$ that correspond to $w^{k,k'}_P$. The same reasoning applies to parent Q. Without loss of generality, the two-point likelihood function of biallelic observed molecular phenotypes for markers k and k' given $w^{k,k'}_P$ and $w^{k,k'}_Q$ is

$$L(r_k \mid w^{k,k'}_P, w^{k,k'}_Q) = \prod_{i=1}^{n} \boldsymbol{\pi}^k \mathbf{A}_{(\varphi^k_P,\varphi^{k'}_P)^*,(\varphi^k_Q,\varphi^{k'}_Q)^*}(r_k)(\boldsymbol{\pi}^{k'})^T \tag{22}$$

where n is the number of individuals and T denotes transposition of a vector. In Eq (22), r_k can be estimated using iterative procedures such as EM or Newton-Raphson. As in Eq (18), it is possible to list all linkage phase configurations and evaluate them based on their likelihood. Here we use the LOD Score (base-10 logarithm of likelihood ratios) in relation to the highest likelihood. Thus, models with high likelihoods will yield LOD Scores close to zero. We also use the LOD Score to asses the evidence for linkage between the two markers using the ratio between the model under $H_a : r = \hat{r}$ and under the null hypothesis of no linkage $H_o : r = 0.5$, given a linkage phase configuration.

As previously shown, it is possible to enumerate all linkage phase configurations for parent P using the Cartesian product $\phi^1_P \times \phi^2_P \times \cdots \times \phi^z_P$. To reduce this Cartesian

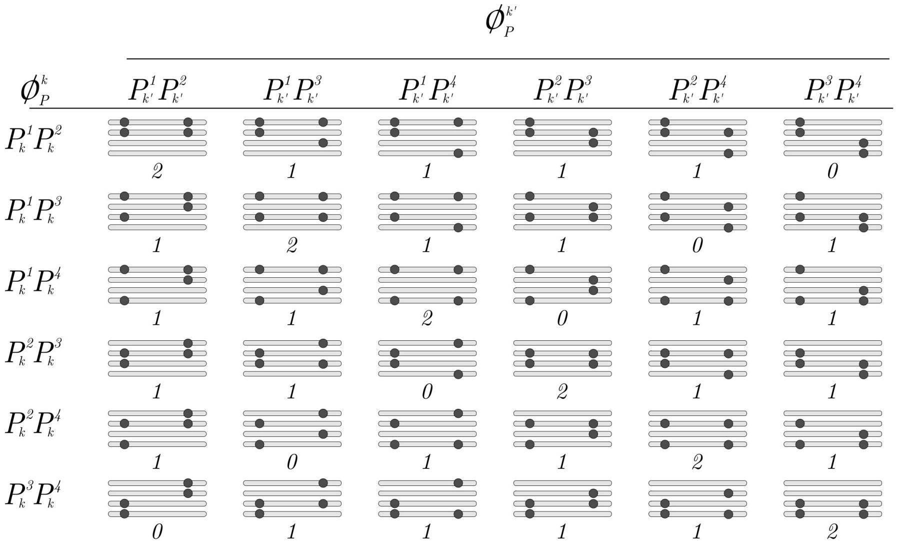

Figure 3. Example of $\Phi_P^{k,k'} = \phi_P^k \times \phi_P^{k'}$ for an autotetraploid homology group with observed dosages $d_P^k = 2$ and $d_P^{k'} = 2$ homologous chromosomes sharing alleles. In this case, ϕ_P^k denotes a set of size six, containing all possible subsets of size two in $\mathcal{P}_k^4 = \{P_k^1, P_k^2, P_k^3, P_k^4\}$. The same reasoning applies to $\phi_P^{k'}$. The horizontal bars represent homologous chromosomes forming a homology group and the dots represent allelic variations of a biallelic marker. The number below each homology group represents the number of homologous chromosomes that share allelic variants ($w_P^{k,k'}$). This defines three partitions: $w_P^{k,k'} = 2$, $w_P^{k,k'} = 1$ and $w_P^{k,k'} = 0$. Notice that, from a homology group within a specific partition, it is possible to obtain the same linkage phase configuration observed in another homology group within that partition by permuting the its homologous chromosomes

space based on two-point analysis, we add a restriction where all pairs $(\varphi_P^k, \varphi_P^{k'})$ in a sequence of configurations $\left(\varphi_P^1, \cdots, \varphi_P^z\right)$ must be contained in $\Phi_p^{k,k'}(\eta)$, where $\Phi_p^{k,k'}(\eta)$ is a subset of all partitions in $\Phi_p^{k,k'}$ in which the associated LOD Sore is smaller than η. Thus, a reduced subset of linkage phases in parent P based on two-point analysis can be obtained using

$$\Phi_P(\eta) = \left\{(\varphi_P^1, \cdots, \varphi_P^z) \mid \varphi_P^i \in \phi_P^i \wedge (\varphi_P^k, \varphi_P^{k'}) \in \Phi_P^{k,k'}(\eta), \forall\, k, k' \in (1, \cdots, z), k > k'\right\} \tag{23}$$

It is important to note that it is not necessary to represent the whole Cartesian space $\{\Phi_P\}$ to restrict the linkage phase configurations to the condition $(\varphi_P^k, \varphi_P^{k'}) \in \Phi_P^{k,k'}(\eta)$. This procedure can be done through the sequential addition of markers from M_1 to M_z. For each marker $M_{k'}$ added to the end of the chain, the ordered pair (k, k'), $k' = 2, \cdots, z$ and $k = k' - 1, \cdots, 1$, is evaluated and only linkage phase configurations that meet the condition $(\varphi_P^k, \varphi_P^{k'}) \in \Phi_P^{k,k'}(\eta)$ $\forall k \in \{k' - 1, \cdots, 1\}$ are considered.

Some of the configurations selected using the previous procedure can be equivalent once they are products of a permutation of the same set of homologous chromosomes. In order to remove this redundancy, let each one of the selected configurations be represented as a binary matrix of dimensions $(m \times k')$ such as

$$\mathbf{H}_{k'}^u = \{h_{i,j}\}_{(m \times k')} = \begin{cases} 1 & \text{if } P_j^i \in \varphi_P^j \\ 0 & \text{otherwise} \end{cases} \tag{24}$$

where $u \in \{1, \cdots, U\}$, U is the number of selected linkage phase configurations, and k' indicates that $M_{k'}$ was the last marker inserted in the chain. The rows of matrix $\mathbf{H}_{k'}^{u}$ represent the homologous chromosomes for the u-th linkage phase configuration with the insertion of the k'-th marker at the end of chain; 1 denotes the presence of an allelic variation, and 0 denotes its absence. If a matrix $\mathbf{H}_{k'}$ could be obtained from a matrix $\mathbf{H}_{k'}^{u'}$ just by permuting the rows (permuting the order of the homologous chromosomes), these two linkage configurations yield the same likelihood. Thus, one of the configurations should be excluded from consideration. The same reasoning applies to parent Q. This procedure can be done recursively until all redundancy is eliminated. The reduced linkage phase configurations search space considering both parents is obtained using $\Phi(\eta) = \Phi_P(\eta) \times \Phi_Q(\eta)$, such as $\#\{\Phi(\eta)\} \ll \#\{\Phi\}$, combined with the redundancy elimination for homology groups. This sequential procedure results in a set of linkage phase configurations containing markers up to $M_{k'}$, which are evaluated using the HMM likelihood. A LOD Score threshold in relation to the most likely configuration is assumed to determine which configurations should be taken into consideration in the next round of marker inclusion (Fig. 4).

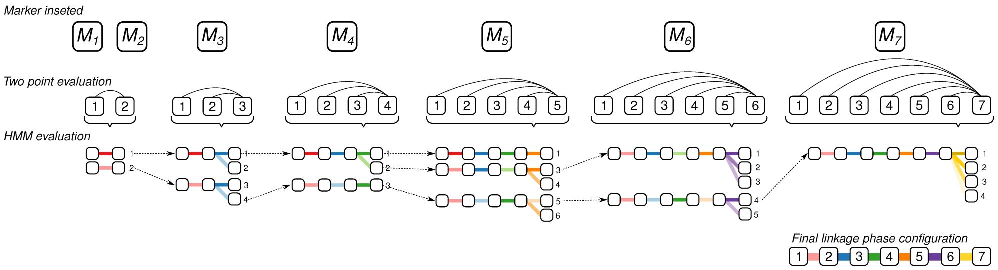

Figure 4. Example of linkage phase configuration estimation using two-point based sequential space reduction and HMM evaluation. Only one parent is presented. The two-point search reduction is composed of two parts: the first one evaluates the LOD Scores obtained through pairwise recombination fraction likelihoods. The second detects equivalent configurations by performing all possible permutations of the homologous chromosomes. The remaining configurations are evaluated using the HMM-based likelihood. In the first step, linkage phase configurations of M_1 and M_2 are evaluated using the two-point analysis. Color shades indicate different linkage phase configurations provided by the two-point analysis. In this example, there are two possible linkage phases represented by two shades of red. These configurations are not evaluated using the HMM, once the outcome would be the same obtained using two-point analysis. In the second step, we evaluate the linkage phases between markers M_3 and M_2, and M_3 and M_1. Configurations with LOD scores smaller than η are maintained to be evaluated by HMM. There are two possible linkage phases given a certain η, represented by two shades of blue. These two configurations are combined with the configurations from the previous step, resulting in four configurations evaluated using HMM likelihood. Given a likelihood threshold, only configurations 1 and 4 are eligible for the next step. The same reasoning applies for the remaining markers. A final linkage phase configuration is obtained after inserting the last marker and choosing the one that yields the highest HMM-based likelihood.

Finally, with all markers inserted, the multipoint likelihood of the whole map is used to find the best configuration among the remaining ones, and the recombination fractions are reestimated. To demonstrate the mechanics of the two-point analysis coupled with the multipoint procedure, a simple example is presented in S3 Appendix. All the methods and procedures described here are available in a software called

MAPPoly, which can be accessed at `https://github.com/mmollina/mappoly`.

Simulations

Simulation 1 - local performance under random bivalent pairing: the aim of this simulation study was to evaluate the local performance of the algorithm considering three ploidy levels ($m = 4$, $m = 6$ and $m = 8$) under the mapping model assumptions (i.e., random pairing and bivalent formation). To be in accordance with molecular data that have been made available through sequence technologies, we simulated bi-allelic markers that can be observed in terms of dosage in parents and progeny. Three different linkage phase scenarios were simulated: In scenario A, for each marker, if the dosage was greater than zero, one of the allelic variants was assigned to the first homologous chromosome in the homology group and the remaining variants of the same type were assigned to the subsequent homologous chromosomes. In B, the allelic variant was randomly assigned to one of the first $\frac{m}{2}$ homologous chromosome and the remaining were assigned to the subsequent homologous chromosomes; in scenario C the allelic variants were randomly assigned to the m homologous chromosomes. Thus, it is expected an increasing difficulty to detect recombination events from scenario A, where the allelic variants were concentrated in the same homologous chromosomes, to scenario C, where they are randomly distributed. Consequently, the phasing and recombination fraction estimation become more challenging from scenario A to scenario C. In real situations, scenarios A and B could occur locally due to lack of recombination between homologous chromosomes since their polyploid formation, whereas scenario C represents regions with higher recombination rates.

For each combination of ploidy level and linkage phase scenario, we simulated five different parental haplotypes. In total, 45 parental configurations were considered ($3 \times 3 \times 5$, S4 Figure). For autotetraploid and autohexaploid configurations, we simulated 1000 full-sib populations. For autooctaploids, this number was reduced to 200 due to the high demand of computer processing required to reconstruct such maps. Each population was comprised of 200 individuals with one linkage group containing 10 markers positioned at a fixed distance of 1 cM between them. For each combination, the percentage of correctly estimated linkage phase configuration in each parent was recorded. Also, for the cases where the linkage phases were correctly estimated, we calculated the average Euclidean distance between the distances of the estimated and simulated maps using $\left\{\frac{(\hat{\mathbf{d}}-\mathbf{d})^T(\hat{\mathbf{d}}-\mathbf{d})}{z-1}\right\}^{-\frac{1}{2}}$ where $\hat{\mathbf{d}}$ is the vector of distances for a estimated map, $\mathbf{d}$ is the vector of distances for the simulated map, z is the number of markers and T indicates vector transposition. For example, a value of 1 cM indicates that the maps differ 1 cM in average from each other [42]. We used the sequential two-point procedure to reduce the search space assuming that linkage phase configurations with associated $LOD < 3.0$ should be investigated using HMM multipoint strategies ($\eta = 3$). For the remaining configurations evaluated using HMM, we kept those with $LOD < 10.0$ to be evaluated in the next round of marker insertion. Notice that, although the likelihood obtained for each map could be used as a criterion to evaluate the order of the markers, this was not considered in this simulation due to the computational demanding nature of the multiple simulations added to high ploidy levels, specially $m = 8$.

Simulation 2 - chromosome-wise performance under preferential pairing and multivalent formation: In this simulation study, we evaluated the performance of the algorithm in dense maps, allowing for multivalent formation and preferential pairing. We used Scenario C from the previous study as a template to simulate five tetraploid and five hexaploid parental haplotypic configurations, each one comprising 200 equally spaced markers with a final length of 100.0 cM (S5 Figure). For each parental configuration, we simulated 200 full-sib populations of 200 offspring considering a combination of three levels of preferential pairing (0.00, 0.25 and 0.50) and three levels of cross-like quadrivalent formation proportion (0.00, 0.25 and 0.50). No hexavalents were simulated in this study. For autohexaploids, the multivalent configurations were always composed by a cross-like quadrivalent plus a bivalent. The centromere was positioned at 20.0 cM from the beginning of the chromosome (subtelocentric centromere with arms ratio 1:4) to study the effect of the double reduction at the distal end of both chromosome arms. All simulations were conducted using the software PedigreeSim [56]. In addition to the statistics recorded in Simulation 1, we computed the rate of double reduction observed in each marker for all constructed maps using the "founderalleles" file provided by PedigreeSim. We also evaluate two values for the LOD Score threshold associated to the two-point analysis ($\eta = 3$ and $\eta = 5$). We used a multipoint LOD Score threshold of 10.0. The R scripts to perform the simulations presented here can be accessed at `https://go.ncsu.edu/mappoly-support-info`.

Simulation results

Simulation 1: Table 1 shows the percentage of data sets where the linkage phase configuration was correctly estimated in both parents P and Q. In scenario (A) the method was capable of recovering the correct linkage phase configuration in all situations for all ploidy levels. In scenarios (B) and (C) there was a slight decrease on the ability to correctly estimate the linkage phase configuration, especially for $m = 6$ and $m = 8$. Although in these cases the percentage of correctly estimated linkage phases was lower, the numbers are considerably high, varying from 100% to 88.8%. This indicates a very good performance to estimate the linkage phase configurations, even using the two-point procedure to narrow the search space.

Table 1. Percentage of data sets where linkage phase configuration was correctly estimated for parents P and Q in simulation 1.

Ploidy level	A		B		C	
	P	Q	P	Q	P	Q
Autotetraploid ($m = 4$)	100	100	99.7	99.8	100	100
	100	100	99.7	99.7	99.9	99.7
	100	100	100	100	99.7	99.8
	100	100	99.9	99.7	99.9	99.9
	100	100	99.9	99.8	100	100
Autohexaploid ($m = 6$)	100	100	96.6	97.2	96.1	94.6
	100	100	97.3	97.5	95.8	95.6
	100	100	96.5	96.7	94.7	94.6
	100	100	97.3	97.4	96.1	94.7
	100	100	97.2	97.4	95.2	94.5
Autooctaploid ($m = 8$)	100	100	93.6	94.4	93.2	95.7
	100	100	97.6	96.8	92.1	93.9
	100	100	96.8	97.6	90.4	89.2
	100	100	97.7	98.4	90.6	90.0
	100	100	96.9	94.6	88.8	90.6

Fig 5 shows the distributions of the average Euclidean distances between the estimated and simulated distance vectors for the correctly estimated linkage phase configuration. In all cases, the majority of the recombination fractions were consistently estimated once the medians of all distributions are very close 0.5, with no practical problems in terms of mapping construction. These results show that, apart from a relatively small percentage of entangled linkage phase configurations, the method successfully performed the phasing and managed to estimate the recombination fraction of 10 markers in all situations evaluated.

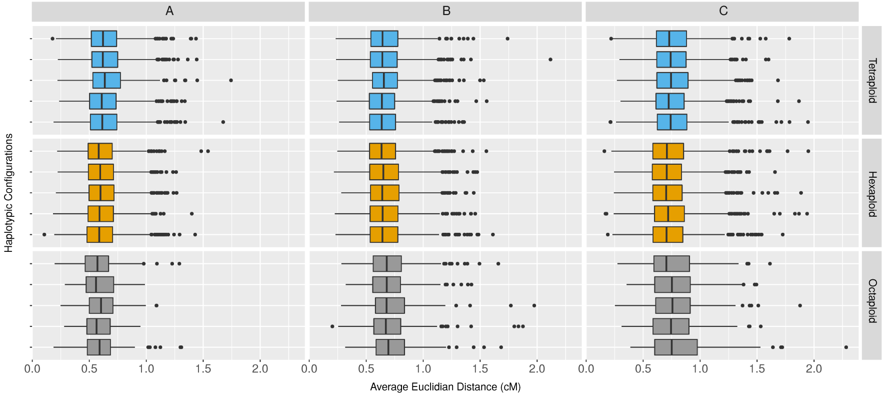

Figure 5. Distributions of the average Euclidean distances between the estimated and simulated distance vectors considering correctly estimated linkage phase configurations. The order of boxplots is the same as the order of haplotypes in S4 Figure. Each column indicates the results for different linkage phase configuration scenarios, namely, A, B and C, and each row indicates a different haplotypic configuration within three ploidy levels.

Simulation 2 The proportion of correctly estimated linkage phase configurations for the dense chromosome-wise map is shown in Table 2. In general, results for tetraploid maps were superior when compared to results for hexaploid maps. It is also possible to observe a better performance for the threshold level $\eta = 5$ in comparison to $\eta = 3$. Similarly to Simulation 1, maps resulting from configurations with no preferential pairing or quadrivalent formation showed a high proportion of correctly estimated linkage phase configurations. Results ranged from 100% to 99% for tetraploid maps and from 100% to 84% for hexaploid maps. Different levels of quadrivalent formation rate had no substantial influence in estimating the correct linkage phase configurations in tetraploids. Within the preferential pairing level 0.0, the percentage of maps with correct linkage phases varied from 100% to 90%. For hexaploids, there was a decrease in this percentage as the quadrivalent formation increases from 0.0 to 0.50, with proportions varying from 100% to 70.5%. Especially for autohexaploids, there was a considerable variation between the five simulated configurations. This occurred, because the effect of the quadrivalent formation can be more pronounced depending on the level of information contained in a particular configuration. Also, the use of a more stringent two-point threshold $\eta = 5$, improved the performance of the phasing algorithm.

Within the preferential pairing level 0.25, results showed decay of correctly estimated linkage phases, which was more pronounced for hexaploid cases with threshold level $\eta = 3$, reaching a minimum value of 52.5% for parent Q in configuration 1. Again, the use of a higher two-point threshold level, $\eta = 5$, helped to improve this number to 68.5%. For preferential pairing level 0.50, there was a clear distinction between the results in tetraploid and hexaploid cases. In the former, the effect was not as pronounced as it was in the latter, where in several cases, the proportion of correctly estimated linkage phases was close to zero. As expected, the usage of a higher threshold level of $\eta = 5$ helped to improve the number of corrected estimated linkage phase configurations. Interestingly, for both cases with preferential pairing (0.25 and 0.50), the formation of quadrivalents had an overall tendency to improve the algorithm's performance. This improvement was expected because when a quadrivalent is formed, each chromosome involved can exchange segments with two others, providing more information regarding their phase configuration.

Given a correctly estimated linkage phase, the recombination fractions were consistently estimated for all levels of preferential pairing with no quadrivalent

Table 2. Percentage of data sets where linkage phase configuration was correctly estimated for parents P and Q in simulation 2.

Preferential pairing		0.00			0.25			0.50		
Quadrivalent formation		0.00	0.25	0.50	0.00	0.25	0.50	0.00	0.25	0.50
Autotetraploid										
$\eta = 3$	P	100.0	99.0	91.5	98.5	98.5	90.0	80.5	93.0	87.5
		100.0	99.5	99.5	98.5	99.5	97.5	57.5	88.5	97.0
		99.5	97.5	98.5	100.0	98.5	94.0	55.0	85.5	94.5
		100.0	100.0	99.5	99.0	98.0	98.0	60.5	86.5	93.0
		99.5	99.5	97.0	98.5	97.0	95.5	67.5	84.5	97.5
	Q	100.0	98.5	90.0	100.0	97.0	90.0	60.0	91.5	86.0
		100.0	100.0	98.0	99.5	100.0	99.0	65.0	89.0	93.5
		100.0	98.5	98.0	97.0	98.5	94.5	41.0	82.0	93.5
		100.0	100.0	99.0	99.5	98.0	98.0	56.5	84.5	90.0
		99.5	99.5	98.0	99.0	98.5	94.5	58.0	82.0	94.0
$\eta = 5$	P	100.0	99.5	93.0	100.0	99.5	95.0	98.0	99.0	95.0
		100.0	100.0	100.0	100.0	100.0	100.0	90.0	99.5	99.0
		100.0	99.5	100.0	100.0	100.0	99.5	86.0	98.5	100.0
		100.0	100.0	100.0	99.5	100.0	99.5	86.5	98.5	100.0
		100.0	100.0	100.0	100.0	100.0	100.0	90.5	96.0	100.0
	Q	100.0	99.5	93.0	100.0	99.0	94.0	88.0	98.5	95.5
		100.0	100.0	100.0	100.0	100.0	100.0	91.5	99.5	99.5
		100.0	99.5	100.0	99.5	100.0	99.0	85.0	98.0	100.0
		100.0	100.0	100.0	99.5	100.0	99.5	86.0	97.5	98.5
		100.0	100.0	99.5	100.0	100.0	100.0	92.0	96.0	99.0
Autohexaploid										
$\eta = 3$	P	84.0	78.5	70.5	69.0	63.5	61.0	2.5	10.5	19.0
		99.0	94.0	91.0	93.0	84.5	80.0	6.5	16.0	22.0
		89.0	94.0	88.0	80.0	84.0	80.5	10.5	16.0	32.5
		93.0	90.5	86.0	88.5	84.0	80.0	9.0	16.5	28.5
		96.0	92.5	91.5	89.5	94.0	87.5	19.0	30.5	44.5
	Q	85.0	81.0	71.0	68.0	52.5	57.5	1.5	3.5	8.5
		99.0	95.0	91.0	86.5	90.0	88.5	9.0	28.0	37.5
		90.0	90.0	86.0	79.0	82.0	77.0	9.5	18.0	28.0
		96.5	92.5	89.5	90.0	89.0	89.0	25.5	35.5	41.0
		95.0	92.0	92.5	89.5	91.0	88.0	16.0	23.0	39.0
$\eta = 5$	P	86.0	84.5	75.5	77.5	69.5	72.5	27.0	36.5	52.5
		100.0	97.5	96.5	98.5	98.0	91.0	55.5	70.5	74.5
		91.5	95.5	93.0	90.5	94.5	89.5	68.0	68.5	77.5
		96.5	94.0	91.0	99.5	99.0	96.5	65.0	78.5	85.0
		98.0	98.5	100.0	97.5	99.0	99.0	73.0	87.5	91.0
	Q	86.5	83.5	75.0	69.5	68.5	72.0	17.5	20.0	39.5
		100.0	99.5	99.0	100.0	99.5	100.0	74.0	81.0	92.5
		91.5	95.5	93.0	91.0	95.0	89.5	67.5	71.5	77.0
		99.0	97.5	93.5	100.0	100.0	99.5	80.0	89.0	92.0
		98.0	98.5	100.0	97.5	99.0	99.0	83.0	83.0	90.5

formation. However, they were overestimated in the presence of quadrivalent formation. This effect was mainly observed at the terminal regions of the chromosome, especially in the long arm, where double reduction is more pronounced (Fig. 6). In this case, tetraploid maps were the most affected. This is in agreement with our expectations since in autohexaploid simulations, there was always the formation of a bivalent which was not involved in the double reduction process (although the rates of double reduction were very similar in both ploidy levels, Fig. 6). In addition to the quadrivalent, the bivalent serves as an extra source of information to access the recombination events.

The average Euclidean distances reflect the overestimation of recombination fractions in cases with quadrivalent formation, showing distributions with higher medians and interquartile ranges in tetraploid cases when compared to hexaploids (S6 Figure). Nevertheless, all the Euclidean distances distributions were located relatively close to zero, with a maximum value of 1.41 cM, indicating that although we observed overestimated recombination fractions towards the terminal ends of the chromosome, they were equally distributed, causing no severe disturbances in the final map. S7 Figure shows an example of the effect of increasing quadrivalent formation rate in autotetraploid and autohexaploid maps. As the markers get further away from the centromere, the recombination fractions become overestimated.

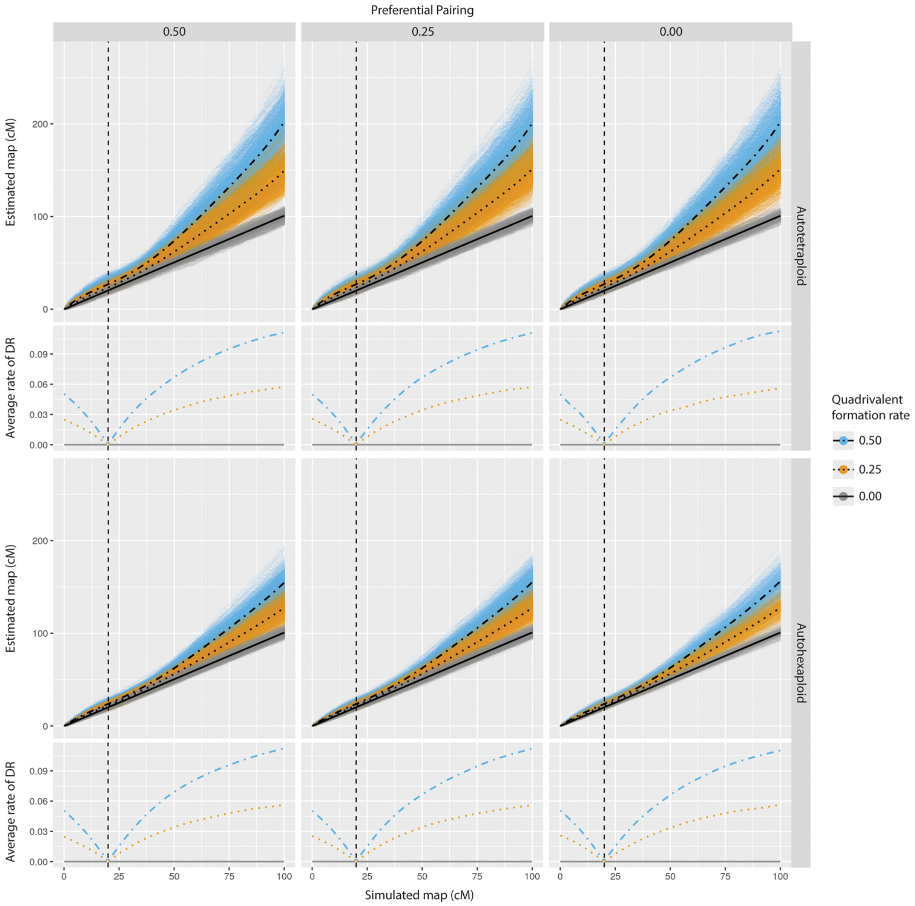

Figure 6. Comparison of estimated versus simulated maps given a correct estimation of linkage phases in simulation 2. Smoothed conditional means of the observed average rate of double reduction is presented along with the simulated chromosome. The centromere was positioned at 20 cM from its beginning (vertical dashed line). Upper panels show the results for tetraploid simulations while lower panels show the results for hexaploid simulations. Three levels of preferential pairing (0.00, 0.25, 0.50) and three levels of quadrivalent formation rate (0.00, 0.25, 0.50) were simulated. The lines superimposed to the scatter plots are smoothed conditional means of the distances using a generalized additive model. Both two-point thresholds were considered since they only affect the phasing procedure.

Discussion

Although the concept of linkage mapping is relatively simple, the combinatorial properties and increasingly missing information that arise from the multiple sets of chromosomes make the construction of genetic maps in high-level autopolyploids extremely challenging. In this work, we frame and solve two fundamental steps towards the construction of such maps, namely multipoint recombination fraction estimations and linkage phase estimation. Our method can be applied to biallelic codominant markers and, due to the flexibility of the HMM framework upon which it was derived, it can be extended to any type of molecular marker. The HMM used in this work takes into account the linkage phase configuration of the whole linkage group to estimate the recombination fractions between adjacent markers. An efficient two-point approach was also presented to reduce the search space of linkage phase configurations. As result, our method provides the likelihood of the model, which can be used as an objective function to compare different map configurations, including linkage phases and marker order. When considering experimental populations, our method is a generalization, for any even ploidy level, of well established genetic linkage mapping methods. For diploid $(m = 2)$ populations derived from biparental crosses, our method is equivalent to the influential Lander and Green algorithm [41]; considering full-sib phase-unknown crosses, it is equivalent to [57]. For tetraploids $(m = 4)$ the method is equivalent to [17], disregarding double reduction. Thus, it encapsulates the essence of the HMM-based genetic mapping methods in a single one.

To assess the statistical power of our method, we conducted two simulation studies. Simulation 1 comprised three ploidy levels and three linkage phase configuration scenarios with ten markers. We demonstrated that our model was capable of correctly estimating the majority of parental linkage phase configurations and recombination fractions, even for complex linkage phase configurations and high ploidy levels. These well-assembled regions could function as multiallelic codominant markers which propagate their information through the HMM to the rest of the chain, improving the quality of the final map. In simulation 2, we analyzed a sequence of 200 markers in combinations of different levels of preferential pairing and rates of quadrivalent formation. In this situation, quadrivalent formation rate had a marginal effect on the phasing procedure, whereas preferential pairing reduced its performance, especially for autohexaploids. The usage of a higher two-point threshold (η) improved the linkage phase estimation in all cases. This fact indicates that the haplotype phasing is more accurate when HMM-based likelihood is used as objective function to evaluate linkage phases. We also observed that quadrivalent formation yield overestimated recombination fractions between adjacent markers located further away from the centromere. This behavior was expected since our model disregards double reduction and, consequently, was not able to correctly estimate the number of crossing over events when this phenomenon was present. Although our model is robust enough to cope with low levels of preferential pairing and tetravalent rate formation, it is possible to include both phenomena in specific points of its derivation. Preferential paring can be included in Eq 4 by not considering $\Pr(\psi_j)$ as uniformly distributed. Double reduction can be included in the definition of the genotypic states in the full transition space (Eq 5). These two phenomena add extra layers of complexity to the genetic mapping of polyploid organisms with high ploidy levels and should be addressed in future studies.

The difficulty in correctly estimating entangled linkage phase configurations lies in two major aspects of the experiments studied here: (i) the outbred nature of the experimental crosses and (ii) the incomplete information of the markers based on dosage (i.e., by not being multiallelic). In experimental population derived from inbred lines, the origin of the haplotypes can be easily inferred from the genetic design. However, obtaining pure inbred lines in high-level autopolyploids has been proven to be impractical due to the high number of crosses and generations necessary to achieve homozygous genotypes and to the inbred depression which some species undergo [61]. In our method, the linkage phase configuration is obtained by comparing the likelihood of a set of models with different linkage phase configurations (Eq 18). The capability of estimating the correct configuration is directly related to the information contained in the marker data. Some of these limitations can be overcome through the use of HMMs which take into account the information of a whole linkage group.

HMMs provide an excellent avenue to assemble genetic maps in complex scenarios, but they are remarkably computational demanding and, in some cases, unfeasible to use. Apart from parallel computing, which can greatly speed up the estimation process and is ubiquitous nowadays, the usage of two-point approaches is a viable option to reduce the dimension of the original problem efficiently. The dimension reduction is achieved by collapsing genotypic states in the full transition space according to the marker information. However, in several cases, the two-point based method can result in low statistical power which is related to the amount of information contained in markers in certain combinations of allelic dosage and linkage phase configurations. This lack of information is exacerbated as markers get distant from each other. Fig 7 shows eight possible configurations of pairs of markers in one autohexaploid parent. Considering the other parent non-informative, we computed the Fisher's information equations based on the likelihood Eq (22) [15, 33, 62]. The equations were plotted as a function of the recombination fraction. The information profiles are related to the number of different haplotypes present on the parental configuration for a given marker dosage. For instance, for two single-dose markers (Fig 7, panel I), when the alleles share the same homologous chromosome ($w_k = 1$), it is always possible to detect if the gamete contains at least one recombinant chromosome. However, when the alleles are in different homologous chromosomes ($w_k = 0$), the detection of recombination events is limited to meiotic configurations containing a bivalent where these chromosomes paired to each other. Additionally, the model proposed here contemplates both parents on the analyses, leading to more complicated linkage phase configurations and information equations.

The multipoint procedure improves the power to detect genetic linkage since the information on the markers depends not only on the observed molecular phenotype for the locus in question but also on the accumulated information along the Markov chain. Fig 7(I) shows that maps using only single-dose markers are limited to the detection of markers whose allelic variants are the same homologous chromosome ($w_k = 1$). Thus, the homologous chromosomes are treated as separate entities, instead of belonging to a homology group, and it is not possible to assemble haplotypes on the parents considering all homologous chromosomes (i.e., linkage phase estimation). Due to the lack of appropriate statistical methods, the use of diploid approximations considering single-dose markers has been the method of choice to build genetic maps in high-level autopolyploids. In our experience with construction of genetic maps in sugarcane [63–66], it is possible to anticipate a great gain of quality in those maps when using the new method proposed in this work. We also expect the same improvement for other high-level autopolyploid species.

The intrinsic lack of information in biallelic markers can be circumvented using multiple markers clustered in linkage disequilibrium (LD) blocks to assemble multiallelic marker data. Two different approaches can be used: the first one relies on the usage of high throughput molecular data and subsequent estimation of pairwise recombination fraction between the markers. In this case, due to the density of the data, closely linked markers are expected, and the Fisher's information for the two-point maximum likelihood estimator is high (Fig 7). Thus, the determination of linkage phase configurations between markers in small blocks can be successfully achieved by using two-point methods (for a detailed example, see S3 Appendix). Once these LD blocks are

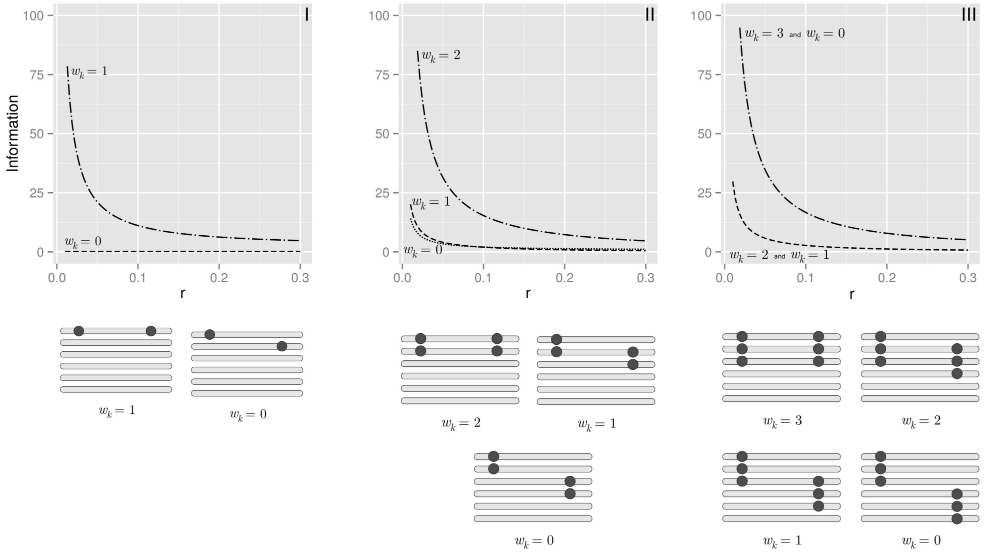

Figure 7. Fisher's information for the two-point maximum likelihood estimators in different combinations of dosages and linkage phases configurations considering one informative hexaploid parent. (I) single-dose markers; alleles share 1 and 0 homologous. (II) double-dose markers; alleles share 2, 1 and 0 homologous. (III) triple dose markers; alleles share 3, 2, 1 and 0 homologous.

well assembled, including the correct linkage phase configuration of both parents, they can be regarded as multiallelic markers. Simulation 1 showed that using two-point procedures coupled with the multipoint analysis is a trustworthy way to assemble haplotypes with closely linked markers. Another approach relies on *a priori* information about markers belonging to the same genomic region where recombination events can be neglected. This information can be obtained using any reference such as genomic or transcriptomic information. In this case, the recombination fraction can be assumed to be $r = 0$ for any pair of markers belonging to the LD block and the linkage phase configuration can be obtained using a trivial Markovian process, with transition probabilities $t_k(j, j') = 1, \forall\, j = j'$ and $t_k(j, j') = 0$ otherwise. Therefore, the biallelic information contained in SNP markers can be combined to assemble haplotypes which will represent alleles allocated in different homologous chromosomes.

The multipoint method proposed herein rely on biallelic marker information. However, the emission function (Eq 9) can be modified to incorporate multiallelic observations. When using multiallelic markers, the number of states that should be visited in the Markov model can be significantly reduced, making the HMM procedure much more efficient. Ideally, in a full-sib population, the number of different alleles should be as high as two times the ploidy level (fully informative). In this case, the Markov model would be fully observed and, the task of estimating recombination fraction reduces to count the number of recombinant events given a linkage phase configuration. Since our algorithm does not need the entire transition space to work, only a subset of states should be visited, making the calculation much faster when compared to the biallelic case.

It is worthwhile to mention that, in this paper we do not address the step *iii* mentioned in the Introduction section, namely, ordering of genetic markers. The genetic mapping literature has an extensive body of methods to address the problem of ordering markers. Several works evaluated some of these methods [42, 67, 68] and others were proposed since then [47–49]. A fundamental lesson learned from these works is that, in complex linkage phase configurations with partially informative markers, methods based on multipoint likelihood provide better results when compared with two-point based methods. However, the multipoint procedures are highly compute-intensive. In the case of high-level autopolyploids, while it is important to rely on the multipoint estimates to recover the lack of information in the biallelic markers, it is also fundamental that the method is fast enough to cope with hundreds of markers per linkage group. One possible solution to these problems is to use two-point information to build marker blocks with a small number of SNPs in high linkage disequilibrium using some clusterization process. The linkage phase within these blocks can be estimated using a combination of two-point and HMM procedures. Then, these marker blocks can be used as multiallelic markers to reduce the number of states that need to be visited in the HMM. The more informative the assembled marker blocks are, the faster is the reconstruction of the mapping using the HMM. Moreover, in several situations, genomic and transcriptomic references are available and often provide, at least, the local physical order of SNPs. Thus, instead of using two-point information to cluster the SNPs into marker blocks, they can be assembled using genomic or transcriptomic references. While this paper provides fundamental steps towards the construction of complete genetic maps in high-level autopolyploids using both multipoint and two-point procedures, the practical aspects and implications will be addressed in future studies.

Once the map is assembled, it is a trivial exercise to obtain the probability of a specific genotype at any map position, conditioned on the whole linkage group. Using this information, it is possible to compute the probability of any unobserved genotype given the genetic map. These conditional probabilities are the basis for answering a series of fundamental questions about quantitative trait loci analysis in high-level autopolyploids, such as the effect of the dosage level on the variation of quantitative traits, the interaction of the alleles within (dominance effects) and between loci (epistatic effects). Therefore, the present study will provide a sound basis for the next step of genetic studies in high-level autopolyploids, trying to unveil the complex structure of autopolyploid genomes through genetic mapping and genome assembling, and even for studying the genetic architecture of quantitative traits based on QTL mapping.

Supporting information

S1 Appendix. Algebraic simplifications for transition probabilities.

S2 Appendix. Algorithm for obtaining l_P and l_Q given two genotypic indices.

S3 Appendix. Example of usage of the two-point and multipoint procedures. In order to show the mechanics of the mapping reconstruction using the combination of two-point and multipoint strategies, we present a simple full-bib autotetraploid mapping population example. This example is easily extendable to higher ploidy levels, since it does not involve matrix forms whose high dimensions would preclude the operations.

S4 Figure. Haplotypes for simulation study 1 Simulated haplotypes with 10 markers and three ploidy levels, namely autotetraploid ($m = 4$), autohexaploid ($m = 6$) and autooctaploid ($m = 8$).

S5 Figure. Haplotypes for simulation study 2 Simulated haplotypes with 200 markers and two ploidy levels, namely autotetraploid ($m = 4$) and autohexaploid ($m = 6$).

S6 Figure. Boxplots of the average Euclidean distances between the estimated and simulated distance vectors for simulation study 2

S7 Figure. Examples of autotetraploid and autohexapoloid maps estimated from datasets with three quadrivalent formation rates: 0.00**,** 0.25 **and** 0.50

Acknowledgments

The authors wish to thank Dr. Guilherme da Silva Pereira and Dr. Zhao-Bang Zeng for their invaluable suggestions for elaboration of the manuscript.

References

1. Soltis DE, Segovia-Salcedo MC, Jordon-Thaden I, Majure L, Miles NM, Mavrodiev EV, et al. Are polyploids really evolutionary dead-ends (again)? A critical reappraisal of Mayrose et al. (2011). New Phytologist. 2014;202(4):1105–1117.

2. Birchler JA. Genetic Consequences of Polyploidy in Plants. In: Soltis PS, Soltis DE, editors. Polyploidy and Genome Evolution. Berlin: Springer-Verlag; 2012. p. 21–32.

3. Comai L. The advantages and disadvantages of being polyploid. Nature Rev Genet. 2005;6(11):836–846.

4. Osborn TC, Pires JC, Birchler JA, Auger DL, Chen ZJ, Lee HS, et al. Understanding mechanisms of novel gene expression in polyploids. Trends in Genetics. 2003;19(3):141 – 147.

5. Sybenga J. Meiotic configurations. Berlin: Springer; 1975.

6. Muller HJ. A New Mode of Segregation in Gregory's Tetraploid Primulas. Am Nat. 1914;48(572):508–512.

7. Soltis DE, Soltis PS. Molecular Data and the Dynamic Nature of Polyploidy Molecular Data and the Dynamic Nature of Polyploidy. Crit Rev Plant Sci. 1993;12(3):243–273.

8. Haldane J. Theoretical Genetics of Autopolyploids. J Genet. 1930;22(3):359–372.

9. Parisod C, Holderegger R, Brochmann C. Evolutionary consequences of autopolyploidy. New Phytol. 2010;186(1):5–17.

10. Otto SP, Whitton J. Polyploid incidence and evolution. Annu Rev Genet. 2000;34:401–437.

11. Mather K. Segregation and Linkage in Autotetraploids. J Genet. 1936;32(2):287–314.

12. Fisher RA. The Theory of Linkage in Polysomic Inheritance. Philos Trans R Soc Lond B Biol Sci. 1947;233(594):55–87.

13. Fisher RA. Allowance for double reduction in the calculation of genotype frequencies with polysomic inheritance. Ann of Eugen. 1954;12:169–171.

14. Hackett Ca, Bradshaw JE, McNicol JW. Interval mapping of quantitative trait loci in autotetraploid species. Genetics. 2001;159(4):1819–1832.

15. Luo ZW, Zhang RM, Kearsey MJ. Theoretical basis for genetic linkage analysis in autotetraploid species. Proc Natl Acad Sci USA. 2004;101:7040–7045.

16. Wu R, Ma CX, Casella G. A Bivalent Polyploid Model for Mapping Quantitative Trait Loci in Outcrossing Tetraploids. Genetics. 2004;166(1):581–595.

17. Leach LJ, Wang L, Kearsey MJ, Luo Z. Multilocus tetrasomic linkage analysis using hidden Markov chain model. Proc Natl Acad Sci USA. 2010;107:4270–4274.

18. Li J, Das K, Fu G, Tong C, Li Y, Tobias C, et al. EM Algorithm for Mapping Quantitative Trait Loci in Multivalent Tetraploids. Int J Plant Genomics. 2010;2010:216547.

19. Hackett CA, McLean K, Bryan GJ. Linkage analysis and QTL mapping using SNP dosage data in a tetraploid potato mapping population. Plos One. 2013;8(5):e63939.

20. Xu F, Lyu Y, Tong C, Wu W, Zhu X, Yin D, et al. A statistical model for QTL mapping in polysomic autotetraploids underlying double reduction. Brief Bioinform. 2013;15(6):1044–1056.

21. Zheng C, Voorrips RE, Jansen J, Hackett CA, Ho J, Bink MCAM. Probabilistic Multilocus Haplotype Reconstruction in Outcrossing Tetraploids. Genetics. 2016;203:119–131.

22. Kriegner A, Cervantes JC, Burg K, Mwanga ROM, Zhang D. A genetic linkage map of sweetpotato [*Ipomoea batatas*(L.) Lam.] based on AFLP markers. Mol Breed. 2003;11(3):169–185.

23. Arizio CM, Costa Tártara SM, Manifesto MM. Carotenoids gene markers for sweetpotato (*Ipomoea batatas* L. Lam): applications in genetic mapping, diversity evaluation and cross-species transference. Mol Genet Genomics. 2014;289(2):237–251.

24. Shirasawa K, Tanaka M, Takahata Y, Ma D, Cao Q, Liu Q, et al. A high-density SNP genetic map consisting of a complete set of homologous groups in autohexaploid sweetpotato (*Ipomoea batatas*). Sci Rep. 2017;7(February):44207.

25. Wang J, Roe B, Macmil S, Yu Q, Murray JE, Tang H, et al. Microcollinearity between autopolyploid sugarcane and diploid sorghum genomes. BMC Genomics. 2010;11(1):261.

26. Garcia AAF, Mollinari M, Marconi TG, Serang OR, Silva RR, Vieira MLC, et al. SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. Sci Rep. 2013;3(1).

27. Soltis DE, Visger CJ, Soltis PS. The polyploidy revolution then. . . and now: Stebbins revisited. Am J Bot. 2014;101(7):1057–1078.

28. Lewin HA, Larkin DM, Pontius J, O'Brien SJ. Every genome sequence needs a good map. Genome Res. 2009;19(11):1925–1928.

29. Luo MC, Gu YQ, You FM, Deal KR, Ma Y, Hu Y, et al. A 4-gigabase physical map unlocks the structure and evolution of the complex genome of Aegilops tauschii, the wheat D-genome progenitor. Proc Natl Acad Sci USA. 2013;110(19):7940–7945.

30. Lemmon ZH, Doebley JF. Genetic Dissection of a Genomic Region with Pleiotropic Effects on Domestication Traits in Maize Reveals Multiple Linked QTL. Genetics. 2014;198:345–353.

31. Wu KK, Burnquist W, Sorrells ME, Tew TL, Moore PH, Tanksley SD. The detection and estimation of linkage in polyploids using single-dose restriction fragments. Theor Appl Genet. 1992;83(3):294–300.

32. Sorrells ME. Development and Application of RFLPs in Polyploids. Crop Sci. 1992;32(5):1086.

33. Ripol MI, Churchill GA, Silva JAGD, Sorrells M. Statistical aspects of genetic mapping in autopolyploids. Gene. 1999;235:31–41.

34. Doerge RW, Craig BA. Model selection for quantitative trait locus analysis in polyploids. Proc Natl Acad Sci USA. 2000;97(14):7951–7956.

35. van Geest G, Bourke PM, Voorrips RE, et al. An ultra-dense integrated linkage map for hexaploid chrysanthemum enables multi-allelic QTL analysisTheor Appl Genet. 2017;130:2527–2541.

36. Voorrips RE, Gort G, Vosman B. Genotype calling in tetraploid species from bi-allelic marker data using mixture models. BMC Bioinformatics. 2011;12(1):172.

37. Serang O, Mollinari M, Garcia AA. Efficient Exact Maximum a Posteriori Computation for Bayesian SNP Genotyping in Polyploids. Plos One. 2012;7(2):e30906.

38. Bargary N, Hinde J, Garcia AAF. Finite Mixture Model Clustering of SNP Data. In: MacKenzie G, Peng D, editors. Statistical Modeling in Biostatistics and Bioinformatics. Switzerland: Springer; 2014. p. 139–157.

39. Mollinari M, Serang O. Quantitative SNP Genotyping of Polyploids with MassARRAY and Other Platforms. In: Batley J, editor. Plant genotyping: methods and protocols. New York: Springer; 2015. p. 215–241.

40. Hackett Ca, Bradshaw JE, Bryan GJ. QTL mapping in autotetraploids using SNP dosage information. Theor Appl Genet. 2014;127:1885–1904.

41. Lander ES, Green P. Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci USA. 1987;84:2363–2367.

42. Mollinari M, Margarido GRA, Vencovsky R, Garcia AAF. Evaluation of algorithms used to order markers on genetic maps. Heredity. 2009;103:494–502.

43. Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, et al. MAPMAKER: An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics. 1987;1(2):174–181.

44. Buetow KH, Chakravarti A. Multipoint Gene Mapping Using Seriation. I. General Methods. Am J Hum Genet. 1987;41:180–188.

45. Doerge RW. Constructing Genetic Maps By Rapid Chain Delineation. J Quant Trait Loci. 1996;2:1–14.

46. Van Os H, Stam P, Visser RG, Van Eck HJ. RECORD: a novel method for ordering loci on a genetic linkage map. Theor Appl Genet. 2005;112(1):30–40.

47. Wu Y, Bhat PR, Close TJ, Lonardi S. Efficient and Accurate Construction of Genetic Linkage Maps from the Minimum Spanning Tree of a Graph. PLoS Genet. 2008;4(10):1–11.

48. Preedy KF, Hackett CA. A rapid marker ordering approach for high-density genetic linkage maps in experimental autotetraploid populations using multidimensional scaling. Theor Appl Genet. 2016;129(11):2117–2132.

49. Wang H, van Eeuwijk FA, Jansen J. The potential of probabilistic graphical models in linkage map construction. Theor Appl Genet. 2016;130:1–12.

50. Van Ooijen JW, Jansen J. Genetic Mapping in Experimental Populations. Cambridge University Press; 2013.

51. Burnham CR. Discussions in cytogenetics. Mineapolis: Burgess Publishing; 1962.

52. Hackett CA. A comment on Xie and Xu: 'Mapping quantitative trait loci in tetraploid species'. Genet Res. 2001;78(02):187–189.

53. Jiang C, Zeng ZB. Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. Genetica. 1997;101(1997):47–58.

54. Rabiner L. A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE. 1989;77(2):257–286.

55. Broman K, Sen S. A Guide to QTL Mapping with R/qtl. New York: Springer; 2009.

56. Voorrips RE, Maliepaard CA. The simulation of meiosis in diploid and tetraploid organisms using various genetic models. BMC Bioinformatics. 2012;13(1):248.

57. Wu R, Ma CX, Painter I, Zeng ZB. Simultaneous Maximum Likelihood Estimation of Linkage and Linkage Phases in Outcrossing Species. Theor Popul Biol. 2002;61(3):349–363.

58. Cao D, Craig BA, Doerge R. A model selection-based interval-mapping method for autopolyploids. Genetics. 2005;169(4):2371–2382.

59. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. Plos One. 2011;6(5):e19379.

60. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, et al. Rapid SNP discovery and genetic mapping using sequenced RAD markers. Plos One. 2008;3(10):e3376.

61. Gallais A. Quantitative genetics and breeding methods in autopolyploids plants. Paris: INRA; 2003.

62. Mather K. The mesurement of linkage in heredity. London: Methuen & Co; 1957.

63. Garcia AAF, Kido EA, Meza AN, Souza HMB, Pinto LR, Pastina MM, et al. Development of an integrated genetic map of a sugarcane (*Saccharum* spp.) commercial cross, based on a maximum-likelihood approach for estimation of linkage and linkage phases. Theor Appl Genet. 2006;112(2):298–314.

64. Oliveira KM, Pinto LR, Marconi TG, Margarido GRA, Pastina MM, Teixeira LHM, et al. Functional integrated genetic linkage map based on EST-markers for a sugarcane (*Saccharum* spp.) commercial cross. Mol Breed. 2007;20(3):189–208.

65. Pastina MM, Malosetti M, Gazaffi R, Mollinari M, Margarido GRA, Oliveira KM, et al. A mixed model QTL analysis for sugarcane multiple-harvest-location trial data. Theor Appl Genet. 2012;124(5):835–849.

66. Palhares AC, Rodrigues-Morais TB, Van Sluys MA, Domingues DS, Maccheroni W, Jordão H, et al. A novel linkage map of sugarcane with evidence for clustering of retrotransposon-based markers. BMC Genet. 2012;13(1):51.

67. Hackett CA, Broadfoot LB. Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. Heredity. 2003;90(1):33–38.

68. Wu J, Jenkins J, Zhu J, McCarty J, Watson C. Monte Carlo simulations on marker grouping and ordering. Theor Appl Genet. 2003;107:568–573.