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

Marcelo Mollinari and Antonio Augusto Franco Garcia

File S1: Algebraic simplifications for transition probabilities

Given a vector $\mathbf{p}_k$, the number of possible vectors $\mathbf{p}_{k+1}$ which have l recombinant chromosomes between k and $k+1$ is $\binom{\frac{m}{2}}{l}$; the number of possible $\mathbf{p}_{k+1}$ which have $\frac{m}{2}-l$ parental chromosomes is $\binom{\frac{m}{2}}{\frac{m}{2}-l}$. Thus for a given $\mathbf{p}_k$ and l , the number of possibles $\mathbf{p}_{k+1}$ (i.e. all possible states in locus $k+1$ with l recombinants) is

$$\binom{\frac{m}{2}}{l}\binom{\frac{m}{2}}{\frac{m}{2}-l} = \binom{\frac{m}{2}}{l}^2$$

Thus, the probability of $\mathbf{p}_k$ is

$$\Pr(\mathbf{p}_k) = \sum_{l=0}^{\frac{m}{2}} \Pr(\mathbf{p}_k,\ \mathbf{p}_{k+1})\binom{\frac{m}{2}}{l}^2 \tag{1}$$

Using (1),

$$\Pr(\mathbf{p}_{k+1}|\mathbf{p}_k) = \frac{\Pr(\mathbf{p}_k,\ \mathbf{p}_{k+1})}{\sum_{l=0}^{\frac{m}{2}} \Pr(\mathbf{p}_k,\ \mathbf{p}_{k+1})\binom{\frac{m}{2}}{l}^2} \tag{2}$$

The numerator of (2) can be written as

$$\begin{aligned}\Pr(\mathbf{p}_k,\ \mathbf{p}_{k+1}) &= \frac{\frac{(1-r_k)^{\frac{m}{2}-l}(r_k)^l}{2^{\frac{m}{2}}} l!(\frac{m}{2}-l)!}{\frac{1}{\frac{m}{2}!}\prod_{i=2,4,\cdots,m}\binom{i}{2}} \\ &= (1-r_k)^{\frac{m}{2}-l}(r_k)^l \frac{l!(\frac{m}{2}-l)!\frac{m}{2}!}{2^{\frac{m}{2}}\prod_{i=2,4,\cdots,m}\binom{i}{2}} \end{aligned} \tag{3}$$

The denominator of (2) can be written as

$$\sum_{l=0}^{\frac{m}{2}} \Pr(\mathbf{p}_k,\ \mathbf{p}_{k+1}) \binom{\frac{m}{2}}{l}^2 = \sum_{l=0}^{\frac{m}{2}} \binom{\frac{m}{2}}{l}^2 \frac{\frac{(1-r_k)^{\frac{m}{2}-l}(r_k)^l}{2^{\frac{m}{2}}} l!(\frac{m}{2}-l)!}{\frac{1}{\frac{m}{2}!} \prod_{i=2,4,\cdots,m} \binom{i}{2}}$$

$$= \sum_{l=0}^{\frac{m}{2}} \binom{\frac{m}{2}}{l} (1-r_k)^{\frac{m}{2}-l}(r_k)^l \frac{l!(\frac{m}{2}-l)!\frac{m}{2}!}{2^{\frac{m}{2}} \prod_{i=2,4,\cdots,m} \binom{i}{2}} \frac{\frac{m}{2}!}{l!(\frac{m}{2}-l)!}$$

$$= \frac{\frac{m}{2}!^2}{2^{\frac{m}{2}} \prod_{i=2,4,\cdots,m} \binom{i}{2}} \sum_{l=0}^{\frac{m}{2}} \binom{\frac{m}{2}}{l} (1-r_k)^{\frac{m}{2}-l}(r_k)^l$$

$$= \frac{\frac{m}{2}!^2}{2^{\frac{m}{2}} \prod_{i=2,4,\cdots,m} \binom{i}{2}} \left\{(1-r_k)r_k\right\}^{\frac{m}{2}}$$

$$= \frac{\frac{m}{2}!^2}{2^{\frac{m}{2}} \prod_{i=2,4,\cdots,m} \binom{i}{2}} \tag{4}$$

Dividing (3) by (4),

$$\Pr(\mathbf{p}_{k+1}|\mathbf{p}_k) = (1-r_k)^{\frac{m}{2}-l}(r_k)^l \frac{l!(\frac{m}{2}-l)!}{\frac{m}{2}!}$$

$$= \frac{(1-r_k)^{\frac{m}{2}-l}(r_k)^l}{\binom{\frac{m}{2}}{l}} \tag{5}$$

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

Marcelo Mollinari and Antonio Augusto Franco Garcia

File S2: Algorithm for obtaining l_P and l_Q given two genotypic indices

This algorithm is used to find the number of recombinant bivalents between loci k and $k+1$ for parents P and Q (i.e. l_P and l_Q) given a ploidy level and the two genotypic indices j and j' which indicate the order of the genotypic states in the transition space in the full-sib population, defined in Eq 6. Consider a binary vector $\mathbf{a}_i$ of length m indicating the presence and absence of alleles in the gametic genotypes $\theta^m_{P,i}$ or $\theta^m_{Q,i}$. See Table S2.1 for an example of vectors in an autohexaploid with their gametic genotypes and indices. With all possible vectors listed, the number of recombinant bivalents l_P between any two gametic genotype, $\theta^m_{P,i}$ and $\theta^m_{P,i'}$ is $\frac{1}{2}\sum \mid \mathbf{a}_i - \mathbf{a}_{i'} \mid$. In real situations, for high ploidy levels the enumeration of all possible $\mathbf{a}$ vectors is a highly intensive task. Thus, Algorithm 1 returns the binary vector for any ploidy level m in any i position in the transition space. This procedure holds for both parents and can be applied to obtain l_P and l_Q given j and j'. To obtain l_P, consider $i = 1 + (j-1)\,\mathbf{div}\,\binom{m}{\frac{m}{2}}$ and $i' = 1 + (j'-1)\,\mathbf{div}\,\binom{m}{\frac{m}{2}}$, where **div** denotes the integer division operator. Then, $\mathbf{a}_i$ and $\mathbf{a}_{i'}$ can be obtained using Algorithm 1 and $l_P = \frac{1}{2}\sum \mid \mathbf{a}_i - \mathbf{a}_{i'} \mid$. To obtain l_Q, consider $h = \binom{m}{\frac{m}{2}} + j - i\binom{m}{\frac{m}{2}}$ and $h' = \binom{m}{\frac{m}{2}} + j' - i'\binom{m}{\frac{m}{2}}$. Then, $l_Q = \frac{1}{2}\sum \mid \mathbf{a}_h - \mathbf{a}_{h'} \mid$.

Table S2.1 : **Example of binary vectors in an autohexaploid with their genotypes and indices.**

i	Gametic genotype ($\theta^m_{P,i}$)	$\mathbf{a}_i$
1	$\{P^1_k, P^2_k, P^3_k\}$	$\{1,1,1,0,0,0\}$
2	$\{P^1_k, P^2_k, P^4_k\}$	$\{1,1,0,1,0,0\}$
3	$\{P^1_k, P^2_k, P^5_k\}$	$\{1,1,0,0,1,0\}$
4	$\{P^1_k, P^2_k, P^6_k\}$	$\{1,1,0,0,0,1\}$
5	$\{P^1_k, P^3_k, P^4_k\}$	$\{1,0,1,1,0,0\}$
6	$\{P^1_k, P^3_k, P^5_k\}$	$\{1,0,1,0,1,0\}$
7	$\{P^1_k, P^3_k, P^6_k\}$	$\{1,0,1,0,0,1\}$
8	$\{P^1_k, P^4_k, P^5_k\}$	$\{1,0,0,1,1,0\}$
9	$\{P^1_k, P^4_k, P^6_k\}$	$\{1,0,0,1,0,1\}$
10	$\{P^1_k, P^5_k, P^6_k\}$	$\{1,0,0,0,1,1\}$
11	$\{P^2_k, P^3_k, P^4_k\}$	$\{0,1,1,1,0,0\}$
12	$\{P^2_k, P^3_k, P^5_k\}$	$\{0,1,1,0,1,0\}$
13	$\{P^2_k, P^3_k, P^6_k\}$	$\{0,1,1,0,0,1\}$
14	$\{P^2_k, P^4_k, P^5_k\}$	$\{0,1,0,1,1,0\}$
15	$\{P^2_k, P^4_k, P^6_k\}$	$\{0,1,0,1,0,1\}$
16	$\{P^2_k, P^5_k, P^6_k\}$	$\{0,1,0,0,1,1\}$
17	$\{P^3_k, P^4_k, P^5_k\}$	$\{0,0,1,1,1,0\}$
18	$\{P^3_k, P^4_k, P^6_k\}$	$\{0,0,1,1,0,1\}$
19	$\{P^3_k, P^5_k, P^6_k\}$	$\{0,0,1,0,1,1\}$
20	$\{P^4_k, P^5_k, P^6_k\}$	$\{0,0,0,1,1,1\}$

Algorithm 1

```
1: function BolVec(m, i)
2:     s_0 ← 0
3:     s_1 ← 1
4:     increment ← 0
5:     sentinel ← 0
6:     vector a(m)                                   ▷ Vector a of size m
7:     while sentinel < m/2 do
8:         temp ← C(m − s_1, m/2 − s_0 − 1)
9:         if i > temp + increment then
10:            a(s_1 − 1) ← 0
11:            increment ← increment + temp
12:        else
13:            a(s_1 − 1) ← 1
14:            s_0 + +
15:        end if
16:        sentinel ← sentinel + a(s_1 − 1)
17:        s_1 + +
18:    end while
19:    return a
20: end function
```

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

Marcelo Mollinari and Antonio Augusto Franco Garcia

File S3: Example of usage of the two-point and multipoint procedures to infer the linkage phase configuration in both parents and estimate recombination fractions in a sequence of markers in high-level autopolyploids.

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.

Parent haplotypes and dataset simulation The mapping population comprise 30 offsprings derived from two autotetraploid parents, P and Q. We simulated one linkage group genotyped with three biallelic markers positioned at a fixed distance of 1cM between them. The linkage phase configuration of both parents is presented in Figure S3.1.

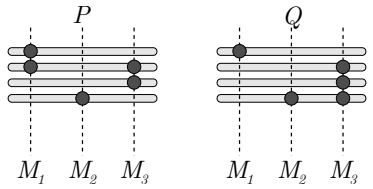

Figure S3.1 : Linkage phase configuration for three biallelic markers in two parents, P and Q.

The observed doses in parent P are $d_P^1 = 2$, $d_P^2 = 1$, $d_P^3 = 2$ and in Q are $d_Q^1 = 1$, $d_Q^2 = 1$, $d_Q^3 = 3$. The simulated dataset of 30 offsprings is shown in Table S3.1. For simplicity purposes, we simulated the dosage information with no error, thus the probability distribution of the genotypes always have the value 1 associated to the simulated dosage and 0 to the remaining ones.

Table S3.1 : Simulated data set of 30 offsprings. Since the dosage informaton has no error, $O_{k,i}$ represents the dosage of the marker at position k for individual i. We also present the vector of probabilities associated to the dosage genotypes $\boldsymbol{\pi}_{k,i}$ for marker k, individual i.

Individual	$\boldsymbol{\pi}_{1,i}$		$\boldsymbol{\pi}_{2,i}$		$\boldsymbol{\pi}_{3,i}$	
(i)	$O_{1,i}$		$O_{2,i}$		$O_{3,i}$	
1	2	{0, 0, 1, 0, 0}	0	{1, 0, 0, 0, 0}	3	{0, 0, 0, 1, 0}
2	1	{0, 1, 0, 0, 0}	1	{0, 1, 0, 0, 0}	2	{0, 0, 1, 0, 0}
3	3	{0, 0, 0, 1, 0}	0	{1, 0, 0, 0, 0}	2	{0, 0, 1, 0, 0}
4	1	{0, 1, 0, 0, 0}	0	{1, 0, 0, 0, 0}	3	{0, 0, 0, 1, 0}
5	2	{0, 0, 1, 0, 0}	1	{0, 1, 0, 0, 0}	3	{0, 0, 0, 1, 0}
6	1	{0, 1, 0, 0, 0}	2	{0, 0, 1, 0, 0}	3	{0, 0, 0, 1, 0}
7	2	{0, 0, 1, 0, 0}	1	{0, 1, 0, 0, 0}	3	{0, 0, 0, 1, 0}
8	2	{0, 0, 1, 0, 0}	1	{0, 1, 0, 0, 0}	3	{0, 0, 0, 1, 0}
9	2	{0, 0, 1, 0, 0}	0	{1, 0, 0, 0, 0}	3	{0, 0, 0, 1, 0}
10	3	{0, 0, 0, 1, 0}	1	{0, 1, 0, 0, 0}	2	{0, 0, 1, 0, 0}
11	1	{0, 1, 0, 0, 0}	1	{0, 1, 0, 0, 0}	2	{0, 0, 1, 0, 0}
12	2	{0, 0, 1, 0, 0}	0	{1, 0, 0, 0, 0}	2	{0, 0, 1, 0, 0}
13	1	{0, 0, 0, 1, 0}	1	{0, 1, 0, 0, 0}	3	{0, 0, 0, 1, 0}
14	2	{0, 0, 1, 0, 0}	0	{1, 0, 0, 0, 0}	2	{0, 0, 1, 0, 0}
15	1	{0, 1, 0, 0, 0}	2	{0, 0, 1, 0, 0}	2	{0, 0, 1, 0, 0}
16	1	{0, 1, 0, 0, 0}	2	{0, 0, 1, 0, 0}	2	{0, 0, 1, 0, 0}
17	3	{0, 0, 0, 1, 0}	1	{0, 1, 0, 0, 0}	2	{0, 0, 1, 0, 0}
18	2	{0, 0, 1, 0, 0}	0	{1, 0, 0, 0, 0}	3	{0, 0, 0, 1, 0}
19	2	{0, 0, 1, 0, 0}	1	{0, 1, 0, 0, 0}	3	{0, 0, 0, 1, 0}
20	2	{0, 0, 1, 0, 0}	1	{0, 1, 0, 0, 0}	3	{0, 0, 0, 1, 0}
21	2	{0, 0, 1, 0, 0}	0	{1, 0, 0, 0, 0}	3	{0, 0, 0, 1, 0}
22	3	{0, 0, 0, 1, 0}	0	{1, 0, 0, 0, 0}	2	{0, 0, 1, 0, 0}
23	2	{0, 0, 1, 0, 0}	0	{1, 0, 0, 0, 0}	2	{0, 0, 1, 0, 0}
24	2	{0, 0, 1, 0, 0}	1	{0, 1, 0, 0, 0}	1	{0, 1, 0, 0, 0}
25	1	{0, 1, 0, 0, 0}	1	{0, 1, 0, 0, 0}	4	{0, 0, 0, 0, 1}
26	3	{0, 0, 0, 1, 0}	0	{1, 0, 0, 0, 0}	2	{0, 0, 1, 0, 0}
27	2	{0, 0, 1, 0, 0}	1	{0, 1, 0, 0, 0}	2	{0, 0, 1, 0, 0}
28	2	{0, 0, 1, 0, 0}	1	{0, 1, 0, 0, 0}	2	{0, 0, 1, 0, 0}
29	1	{0, 1, 0, 0, 0}	2	{0, 0, 1, 0, 0}	3	{0, 0, 0, 1, 0}
30	1	{0, 1, 0, 0, 0}	2	{0, 0, 1, 0, 0}	2	{0, 0, 1, 0, 0}

Two-point recombination fraction The efficient construction of the genetic map starts with the two-point analysis which allow a fast estimation of recombination fraction and linkage phase configuration between all possible pairs of markers in the dataset. The key point in this analysis is the reduction of the full transition space given a ploidy level, the dosage and linkage phase configuration of the markers in the parents. For each possible combination of dosage and linkage phase configuration, a specific reduction of dimension must be done. After that, for a given pair of markers with their specific doses, the best linkage phase configuration is assessed based on the likelihood of the evaluated models. Moreover, genetic linkage can also be tested.

Since $d_P^1 = 2$, $d_P^2 = 1$, $d_P^3 = 2$, $d_Q^1 = 1$, $d_Q^2 = 1$ and $d_Q^3 = 3$, we can write:

$$
\begin{aligned}
\phi_P^1 &= \left\{\{P_1^1, P_1^2\}, \{P_1^1, P_1^3\}, \{P_1^1, P_1^4\}, \{P_1^2, P_1^3\}, \{P_1^2, P_1^4\}, \{P_1^3, P_1^4\}\right\} \\
\phi_P^2 &= \left\{\{P_2^1\}, \{P_2^2\}, \{P_2^3\}, \{P_2^4\}\right\} \\
\phi_P^3 &= \left\{\{P_3^1, P_3^2\}, \{P_3^1, P_3^3\}, \{P_3^1, P_3^4\}, \{P_3^2, P_3^3\}, \{P_3^2, P_3^4\}, \{P_3^3, P_3^4\}\right\} \\
\phi_Q^1 &= \left\{\{Q_1^1\}, \{Q_1^2\}, \{Q_1^3\}, \{Q_1^4\}\right\} \\
\phi_Q^2 &= \left\{\{Q_2^1\}, \{Q_2^2\}, \{Q_2^3\}, \{Q_2^4\}\right\} \\
\phi_Q^3 &= \left\{\{Q_3^1, Q_3^2, Q_3^3\}, \{Q_3^1, Q_3^2, Q_3^4\}, \{Q_3^1, Q_3^3, Q_3^4\}, \{Q_3^2, Q_3^3, Q_4^4\}\right\}
\end{aligned}
$$

Table S3.2 shows all possible linkage phase configurations for markers M_1 and M_2

Table S3.2 : $\Phi_P^{1,2} = \phi_P^1 \times \phi_P^2$ and $\Phi_Q^{1,2} = \phi_Q^1 \times \phi_Q^2$ for the simulated data. In this case $d_P^1 = 2$, $d_P^2 = 1$, $d_Q^1 = 1$ and $d_Q^2 = 1$. In both cases there are two partitions. The gray cells represent the partition where $w_P^{1,2} = 1$ and the white cells represent the partition where $w_P^{1,2} = 0$

	ϕ_P^2			
ϕ_P^1	$\{P_2^1\}$	$\{P_2^2\}$	$\{P_2^3\}$	$\{P_2^4\}$
$\{P_1^1, P_1^2\}$	$\{\{P_1^1, P_1^2\}, \{P_2^1\}\}$	$\{\{P_1^1, P_1^2\}, \{P_2^2\}\}$	$\{\{P_1^1, P_1^2\}, \{P_2^3\}\}$	$\{\{P_1^1, P_1^2\}, \{P_2^4\}\}$
$\{P_1^1, P_1^3\}$	$\{\{P_1^1, P_1^3\}, \{P_2^1\}\}$	$\{\{P_1^1, P_1^3\}, \{P_2^2\}\}$	$\{\{P_1^1, P_1^3\}, \{P_2^3\}\}$	$\{\{P_1^1, P_1^3\}, \{P_2^4\}\}$
$\{P_1^1, P_1^4\}$	$\{\{P_1^1, P_1^4\}, \{P_2^1\}\}$	$\{\{P_1^1, P_1^4\}, \{P_2^2\}\}$	$\{\{P_1^1, P_1^4\}, \{P_2^3\}\}$	$\{\{P_1^1, P_1^4\}, \{P_2^4\}\}$
$\{P_1^4, P_1^3\}$	$\{\{P_1^2, P_1^3\}, \{P_2^1\}\}$	$\{\{P_1^2, P_1^3\}, \{P_2^2\}\}$	$\{\{P_1^2, P_1^3\}, \{P_2^3\}\}$	$\{\{P_1^2, P_1^3\}, \{P_2^4\}\}$
$\{P_1^2, P_1^4\}$	$\{\{P_1^2, P_1^4\}, \{P_2^1\}\}$	$\{\{P_1^2, P_1^4\}, \{P_2^2\}\}$	$\{\{P_1^2, P_1^4\}, \{P_2^3\}\}$	$\{\{P_1^2, P_1^4\}, \{P_2^4\}\}$
$\{P_1^3, P_1^4\}$	$\{\{P_1^3, P_1^4\}, \{P_2^1\}\}$	$\{\{P_1^3, P_1^4\}, \{P_2^2\}\}$	$\{\{P_1^3, P_1^4\}, \{P_2^3\}\}$	$\{\{P_1^3, P_1^4\}, \{P_2^4\}\}$

	ϕ_Q^2			
ϕ_Q^1	$\{Q_2^1\}$	$\{Q_2^2\}$	$\{Q_2^3\}$	$\{Q_2^4\}$
$\{Q_1^1\}$	$\{\{Q_1^1\}, \{Q_2^1\}\}$	$\{\{Q_1^1\}, \{Q_2^2\}\}$	$\{\{Q_1^1\}, \{Q_2^3\}\}$	$\{\{Q_1^1\}, \{Q_2^4\}\}$
$\{Q_1^2\}$	$\{\{Q_1^2\}, \{Q_2^1\}\}$	$\{\{Q_1^2\}, \{Q_2^2\}\}$	$\{\{Q_1^2\}, \{Q_2^3\}\}$	$\{\{Q_1^2\}, \{Q_2^4\}\}$
$\{Q_1^3\}$	$\{\{Q_1^3\}, \{Q_2^1\}\}$	$\{\{Q_1^3\}, \{Q_2^2\}\}$	$\{\{Q_1^3\}, \{Q_2^3\}\}$	$\{\{Q_1^3\}, \{Q_2^4\}\}$
$\{Q_1^4\}$	$\{\{Q_1^4\}, \{Q_2^1\}\}$	$\{\{Q_1^4\}, \{Q_2^2\}\}$	$\{\{Q_1^4\}, \{Q_2^3\}\}$	$\{\{Q_1^4\}, \{Q_2^4\}\}$

In order to show the mechanics of the dimension reduction, let us consider for $w_P^{1,2} = 0$ and $w_Q^{1,2} = 0$ the configurations $\left\{\varphi_P^1 = \{P_1^1, P_1^2\}, \varphi_P^2 = \{P_2^3\}\right\}$ and $\left\{\varphi_Q^1 = \{Q_1^1\}, \varphi_Q^2 = \{Q_2^2\}\right\}$. Table S3.8 shows the full transition space for autotetraploids. Notice that for higher ploidy levels, the dimension of the transition space can be unwieldy, but in real cases we do not need to represent it in its complete form. The reduction of dimensionality of the full transition space is based on the proper combination of the rows and columns in Table S3.8. As result, instead of a space with dimensions (36×36), the dimension of the new transition space is (5×5). As pointed out in Eq 20, each elements in the reduced transition space is composed by a combination of

$$\frac{(1 - r_k)^{m - l_P - l_Q} (r_k)^{l_P + l_Q}}{\binom{\frac{m}{2}}{l_P}\binom{\frac{m}{2}}{l_Q}}$$

with $l_P, l_Q = \{0, \cdots, m\}$. The weight of each element corresponding to the ordered pair (l_P, l_Q) is given by the variable $\zeta_{D_1,D_2}(l_P, l_Q)$. Table S3.3 shows all possible $\zeta_{D_1,D_2}(l_P, l_Q)$ given $\varphi_P^1 = \{P_1^1, P_1^2\}$, $\varphi_Q^1 = \{Q_1^1\}$, $\varphi_P^2 = \{P_2^3\}$ and $\varphi_Q^2 = \{Q_2^2\}$ for all (l_P, l_Q) for each dosage pair $(\mathcal{D}_1^4, \mathcal{D}_2^4)$.

		(l_P, l_Q)								
$\mathcal{D}_1^4$	$\mathcal{D}_2^4$	$(0,0)$	$(0,1)$	$(0,2)$	$(1,0)$	$(1,1)$	$(1,2)$	$(2,0)$	$(2,1)$	$(2,2)$
0	0	0	0	0	1/18	1/3	1/9	1/36	1/6	1/18
0	1	1/36	1/6	1/18	1/6	2/3	1/6	1/18	1/6	1/36
0	2	1/18	1/6	1/36	1/9	1/3	1/18	0	0	0
0	3	0	0	0	0	0	0	0	0	0
0	4	0	0	0	0	0	0	0	0	0
1	0	1/18	1/3	1/9	1/3	5/3	1/2	1/9	1/2	5/36
1	1	2/9	5/6	7/36	5/6	10/3	5/6	7/36	5/6	2/9
1	2	5/36	1/2	1/9	1/2	5/3	1/3	1/9	1/3	1/18
1	3	0	0	0	0	0	0	0	0	0
1	4	0	0	0	0	0	0	0	0	0
2	0	5/36	1/2	1/9	1/2	5/3	1/3	1/9	1/3	1/18
2	1	2/9	5/6	7/36	5/6	10/3	5/6	7/36	5/6	2/9
2	2	1/18	1/3	1/9	1/3	5/3	1/2	1/9	1/2	5/36
2	3	0	0	0	0	0	0	0	0	0
2	4	0	0	0	0	0	0	0	0	0
3	0	1/18	1/6	1/36	1/9	1/3	1/18	0	0	0
3	1	1/36	1/6	1/18	1/6	2/3	1/6	1/18	1/6	1/36
3	2	0	0	0	1/18	1/3	1/9	1/36	1/6	1/18
3	3	0	0	0	0	0	0	0	0	0
3	4	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0
4	1	0	0	0	0	0	0	0	0	0
4	2	0	0	0	0	0	0	0	0	0
4	3	0	0	0	0	0	0	0	0	0
4	4	0	0	0	0	0	0	0	0	0

Table S3.3 : $\zeta_{D_1,D_2}(l_P, l_Q)$ given $\varphi_P^1 = \{P_1^1, P_1^2\}$, $\varphi_Q^1 = \{Q_1^1\}$, $\varphi_P^2 = \{P_2^3\}$ and $\varphi_Q^2 = \{Q_2^2\}$. The results would be the same for any other combination of φ's that result in $(w_P^{1,2} = 0, w_Q^{1,2} = 0)$.

Thus, $\mathbf{A}_{\varphi_P^1=\{P_1^1,P_1^2\},\varphi_P^2=\{P_2^3\},\varphi_Q^1=\{Q_1^1\},\varphi_Q^2=\{Q_2^2\}}(r_k)$ can be easily obtained using article's Eq. 21

$$\begin{bmatrix} \frac{r(1+r)}{36} & \frac{1+2r-2r^2}{36} & \frac{(r-2)(r-1)}{36} & 0 & 0 \\ \frac{2+4r-r^2}{36} & \frac{4-r+r^2}{18} & \frac{5-2r-r^2}{36} & 0 & 0 \\ \frac{5-2r-r^2}{36} & \frac{4-r+r^2}{18} & \frac{2+4r-r^2}{36} & 0 & 0 \\ \frac{(r-2)(r-1)}{36} & \frac{1+2r-2r^2}{36} & \frac{r(1+r)}{36} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Using article's Eq 22, one can specify the likelihood function $L(r_1 \mid w_P^{1,2}, w_Q^{1,2})$ which will be maximized in order obtain $\hat{r}$. Since the simulated data set contains no error, it is possible to simply count the number of individuals in each one of the dosage category and represent them in a double entry table

Table S3.4 : Number of individuals in each one of the dosage-based genotypic classes for markers M_1 and M_2

	d^{M_2}		
d^{M_1}	0	1	2
0	0	0	0
1	1	4	5
2	7	8	0
3	3	4	0

For this data set, the likelihood function is

$$L(r_1 \mid w_P^{1,2} = 0, w_Q^{1,2} = 0) \propto \left[\frac{2 + 4r_1 - {r_1}^2}{36}\right] \left[\frac{4 - r_1 + {r_1}^2}{18}\right]^{12} \left[\frac{5 - 2r_1 - {r_1}^2}{36}\right]^{12} \left[\frac{(r_1 - 2)(r_1 - 1)}{36}\right]^3 \left[\frac{1 + 2r_1 - 2{r_1}^2}{36}\right]^2 \quad (1)$$

The maximum likelihood estimator of r_1 can be obtained using any iterative procedure. In this example, we used the golden section search combined with parabolic interpolation described in [1] and implemented in the function `optim` in R software. Given $w_P^{1,2} = 0$ and $w_Q^{1,2} = 0$, $\hat{r} = 4.58 \times 10^{-5}$ and the associated natural logarithm of the likelihood is -60.47. Traditionally, genetic mapping procedures use LOD Scores (base-10 log likelihood ratio) as a means for decision making. In our method we use two LOD Scores: (i) first takes into account the logarithm of

the ratio between the highest likelihood among all linkage phase configurations and the likelihood of all possible linkage phase configurations; (*ii*) the second uses 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. The results for the pair of markers (M_1, M_2) are presented in Table S3.5

$w_P^{1,2}$	$w_Q^{1,2}$	$\hat{r}$	log-likelihood	LOD_{ph}	$LOD_{\hat{r}}$
0	0	0.00	−60.47	0.00	2.49
0	1	0.50	−66.20	2.49	0.00
1	0	0.50	−66.20	2.49	0.00
1	1	0.50	−66.20	2.49	0.00

Table S3.5 : Recombination fraction estimates and associated LOD Scores between markers (M_1, M_2). LOD_{ph} indicates the logarithm to base 10 of the ratio between the highest likelihood among all linkage phase configurations and the likelihood of all possible linkage phase configurations. $LOD_{\hat{r}}$ uses the ratio between the model under $H_a : r = \hat{r}$ and under the null hypothesis of no linkage $H_o : r = 0.5$.

The same reasoning can be applied for pairs (M_1, M_3) and (M_2, M_3):

$w_P^{1,3}$	$w_Q^{1,3}$	$\hat{r}$	log-likelihood	LOD_{ph}	$LOD_{\hat{r}}$
1	0	0.00	−58.61	0.00	2.21
0	1	0.00	−59.06	0.20	2.18
1	1	0.50	−63.68	2.21	0.00
0	0	0.42	−63.77	2.24	0.13
2	0	0.50	−64.08	2.38	0.00
2	1	0.50	−64.08	2.38	0.00

$w_P^{2,3}$	$w_Q^{2,3}$	$\hat{r}$	log-likelihood	LOD_{ph}	$LOD_{\hat{r}}$
0	1	0.00	−60.59	0.00	0.34
1	0	0.00	−60.68	0.04	0.30
1	1	0.50	−61.37	0.34	0.00
0	0	0.50	−61.37	0.34	0.00

Table S3.6 : Recombination fraction estimates and associated LOD Scores between markers (M_1, M_3) and (M_2, M_3). LOD_{ph} indicates the logarithm to base 10 of the ratio between the highest likelihood among all linkage phase configurations and the likelihood of all possible linkage phase configurations. $LOD_{\hat{r}}$ uses the ratio between the model under $H_a : r = \hat{r}$ and under the null hypothesis of no linkage $H_o : r = 0.5$.

Analyzing Tables S3.5 and S3.6, it is possible to narrow down the number of linkage phase configurations that need to be evaluated using the multipoint procedure. Let us assume $LOD_{ph} > 2.00$ the threshold for the linkage phase elimination criteria. For the first two markers M_1 and M_2, $\Phi_P^{1,2}(\eta = 2)$ contains only situations where $w_P^{1,2} = 0$ and $w_Q^{1,2} = 0$ (white cells in Table S3.2). In this case, the size of $\Phi_P^{1,2}(\eta = 2)$ is $U = 12$ and the $\mathbf{H}_2^u$ matrices are

$$\mathbf{H}_2^1 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \mathbf{H}_2^2 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{H}_2^3 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \mathbf{H}_2^4 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{H}_2^5 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad \mathbf{H}_2^6 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}$$

$$\mathbf{H}_2^7 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \mathbf{H}_2^8 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{H}_2^9 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad \mathbf{H}_2^{10} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \mathbf{H}_2^{11} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix}, \quad \mathbf{H}_2^{12} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix}$$

Notice that any of the matrices can be obtained from any other just by permuting the rows. Thus, we chose the configuration related to the first matrix ($\mathbf{H}_2^1 \rightarrow \big\{\{P_1^1, P_1^2\}, \{P_2^3\}\big\}$) to proceed with the analysis. The same reasoning applies to parent Q. In that case, we choose the configuration $\big\{\{Q_1^1\}, \{Q_2^2\}\big\}$. Notice that, at threshold level $\eta = 2$, the two-point analysis did not provide any information between markers M_2 and M_3. In any case, it is possible to remove redundant configurations when evaluating $\mathbf{H}_3$. For markers M_1 and M_3 the procedure is similar to markers M_1 and M_2. Figure shows possible configurations for parents P and Q using two-point information for pair of markers M_1-M_2 and M_2-M_3. Using informations of M_1 and M_3, it is possible to narrow down the number of configurations to be analyzed by multipoint procedures.

In linkage groups with more markers, this procedure can be applied until there is no more information to be extracted from the two-point analysis at the assumed threshold level. In this simulation, only four linkage phase configurations needed to be evaluated using multipoint procedure.

The remaining linkage phase configurations were analyzed using the full HMM procedure. The log-likelihood and LOD Scores of the models are presented in Table S3.7

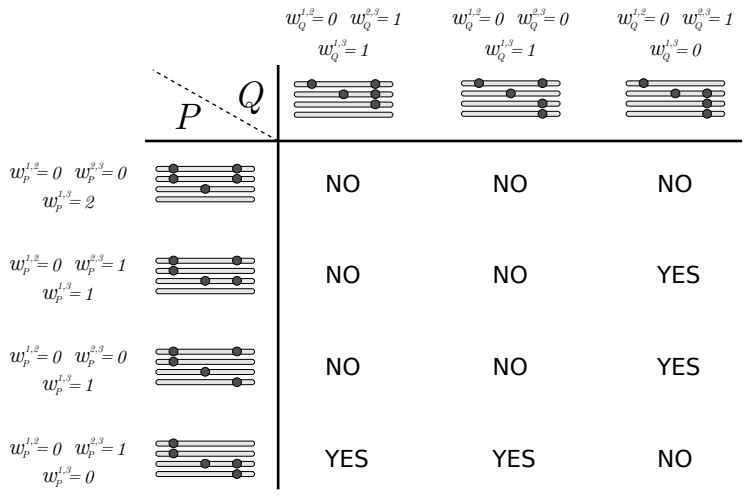

Figure S3.2 : Possible linkage phase configurations in parents P and Q eliminating the ones with $LOD_{ph} > 2.00$ for pairs of markers (M_1, M_2) and (M_2, M_3). $w_P^{k,k'}$ and $w_Q^{k,k'}$ indicate the number of homologous chromosomes that share allelic variants for loci k and k' in parents P and Q, respectively. Some of the combinations can be eliminated using information from the pair (M_1, M_3). For instance, with a threshold of $LOD_{ph} > 2.00$, any of the combination in the first line of the table is not possible since $w_P^{1,3} = 2$ and the associated LOD_{ph} regardless the configuration in Q is 2.38. Using this reasoning it is possible to eliminate several configurations (marked with a NO) based on the information from the pair (M_1, M_3). Thus, there are only four configurations left to be tested using the multipoint procedure.

Table S3.7

Linkage phase			*LOD*	log-likelihood
M_1	M_2	M_3		
$\{P_1^1 P_1^2\}$ $\{Q_1^1\}$	$\{P_2^3\}$ $\{Q_2^2\}$	$\{P_3^1 P_3^4\}$ $\{Q_3^2 Q_3^3 Q_3^4\}$	0.00	−81.90
$\{P_1^1 P_1^2\}$ $\{Q_1^1\}$	$\{P_2^3\}$ $\{Q_2^2\}$	$\{P_3^3 P_3^4\}$ $\{Q_3^1 Q_3^3 Q_3^4\}$	1.75	−85.93
$\{P_1^1 P_1^2\}$ $\{Q_1^1\}$	$\{P_2^3\}$ $\{Q_2^2\}$	$\{P_3^1 P_3^3\}$ $\{Q_3^2 Q_3^3 Q_3^4\}$	3.07	−88.96
$\{P_1^1 P_1^2\}$ $\{Q_1^1\}$	$\{P_2^3\}$ $\{Q_2^2\}$	$\{P_3^3 P_3^4\}$ $\{Q_3^2 Q_3^3 Q_3^4\}$	3.45	−89.86

Finally, the best configuration obtained using both two-point and multipoint procedure is

Figure S3.3

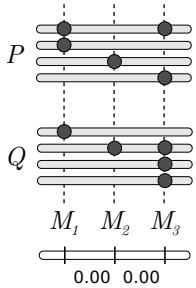

Figure S3.3

This configuration is a permutation of the homologous chromosomes of the simulation. In practical terms it consists in the same linkage phase configuration. The codes to perform the analysis step by step, as presented in this supplement, are available at `https://github.com/mmollina/Autopolyploid_Linkage/blob/master/src/SI3_example.R`

References

[1] Brent, R. P., 1973 *Algorithms for Minimization without Derivatives*. Prentice-Hall, Englewood Cliffs, New Jersey.

				j'																																			
j	i	h	$\mathcal{G}^4_{k,j} = \{\theta^4_{P,i}, \theta^4_{Q,h}\}$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
1	1	1	$\{(P^1_k, P^2_k), (Q^1_k, Q^2_k)\}$	a^2	ab	ab	ab	ab	ac	ab	b^2	b^2	b^2	b^2	bc	ab	b^2	b^2	b^2	b^2	bc	ab	b^2	b^2	b^2	b^2	bc	ab	b^2	b^2	b^2	b^2	bc	ac	bc	bc	bc	bc	c^2
2	1	2	$\{(P^1_k, P^2_k), (Q^1_k, Q^3_k)\}$	ab	a^2	ab	ab	ac	ab	b^2	ab	b^2	b^2	bc	b^2	b^2	ab	b^2	b^2	bc	b^2	b^2	ab	b^2	b^2	bc	b^2	b^2	ab	b^2	b^2	bc	b^2	bc	ac	bc	bc	c^2	bc
3	1	3	$\{(P^1_k, P^2_k), (Q^1_k, Q^4_k)\}$	ab	ab	a^2	ac	ab	ab	b^2	b^2	ab	bc	b^2	b^2	b^2	b^2	ab	bc	b^2	b^2	b^2	b^2	ab	bc	b^2	b^2	b^2	b^2	ab	bc	b^2	b^2	bc	bc	ac	c^2	bc	bc
4	1	4	$\{(P^1_k, P^2_k), (Q^2_k, Q^3_k)\}$	ab	ab	ac	a^2	ab	ab	b^2	b^2	bc	ab	b^2	b^2	b^2	b^2	bc	ab	b^2	b^2	b^2	b^2	bc	ab	b^2	b^2	b^2	b^2	bc	ab	b^2	b^2	bc	bc	c^2	ac	bc	bc
5	1	5	$\{(P^1_k, P^2_k), (Q^2_k, Q^4_k)\}$	ab	ac	ab	ab	a^2	ab	b^2	bc	b^2	b^2	ab	b^2	b^2	bc	b^2	b^2	ab	b^2	b^2	bc	b^2	b^2	ab	b^2	b^2	bc	b^2	b^2	ab	b^2	bc	c^2	bc	bc	ac	bc
6	1	6	$\{(P^1_k, P^2_k), (Q^3_k, Q^4_k)\}$	ac	ab	ab	ab	ab	a^2	bc	b^2	b^2	b^2	b^2	ab	bc	b^2	b^2	b^2	b^2	ab	bc	b^2	b^2	b^2	b^2	ab	bc	b^2	b^2	b^2	b^2	ab	c^2	bc	bc	bc	bc	ac
7	2	1	$\{(P^1_k, P^3_k), (Q^1_k, Q^2_k)\}$	ab	b^2	b^2	b^2	b^2	bc	a^2	ab	ab	ab	ab	ac	ab	b^2	b^2	b^2	b^2	bc	ab	b^2	b^2	b^2	b^2	bc	ac	bc	bc	bc	bc	c^2	ab	b^2	b^2	b^2	b^2	bc
8	2	2	$\{(P^1_k, P^3_k), (Q^1_k, Q^3_k)\}$	b^2	ab	b^2	b^2	bc	b^2	ab	a^2	ab	ab	ac	ab	b^2	ab	b^2	b^2	bc	b^2	b^2	ab	b^2	b^2	bc	b^2	b^2	bc	ac	bc	bc	c^2	bc	b^2	ab	b^2	b^2	bc	b^2
9	2	3	$\{(P^1_k, P^3_k), (Q^1_k, Q^4_k)\}$	b^2	b^2	ab	bc	b^2	b^2	ab	ab	a^2	ac	ab	ab	b^2	b^2	ab	bc	b^2	b^2	b^2	ab	bc	bc	b^2	b^2	b^2	bc	bc	ac	c^2	bc	bc	b^2	b^2	ab	bc	b^2	b^2
10	2	4	$\{(P^1_k, P^3_k), (Q^2_k, Q^3_k)\}$	b^2	b^2	bc	ab	b^2	b^2	ab	ab	ac	a^2	ab	ab	b^2	b^2	bc	ab	b^2	b^2	b^2	b^2	bc	ab	b^2	b^2	bc	bc	c^2	ac	bc	bc	b^2	b^2	bc	ab	b^2	b^2
11	2	5	$\{(P^1_k, P^3_k), (Q^2_k, Q^4_k)\}$	b^2	bc	b^2	b^2	ab	b^2	ab	ac	ab	ab	a^2	ab	b^2	bc	b^2	b^2	ab	b^2	b^2	bc	b^2	b^2	ab	b^2	bc	c^2	bc	bc	ac	bc	b^2	bc	b^2	b^2	ab	b^2
12	2	6	$\{(P^1_k, P^3_k), (Q^3_k, Q^4_k)\}$	bc	b^2	b^2	b^2	b^2	ab	ac	ab	ab	ab	ab	a^2	bc	b^2	b^2	b^2	b^2	ab	bc	b^2	b^2	b^2	b^2	ab	c^2	bc	bc	bc	bc	ac	bc	b^2	b^2	b^2	b^2	ab
13	3	1	$\{(P^1_k, P^4_k), (Q^1_k, Q^2_k)\}$	ab	b^2	b^2	b^2	b^2	bc	ab	b^2	b^2	b^2	b^2	bc	a^2	ab	ab	ab	ac	ac	bc	bc	bc	bc	c^2	ab	b^2	b^2	b^2	b^2	b^2	bc	ab	b^2	b^2	b^2	b^2	bc
14	3	2	$\{(P^1_k, P^4_k), (Q^1_k, Q^3_k)\}$	b^2	ab	b^2	b^2	bc	b^2	b^2	ab	b^2	b^2	bc	b^2	ab	a^2	ab	ab	ac	ab	bc	ac	bc	bc	c^2	bc	b^2	ab	b^2	b^2	bc	b^2	b^2	ab	b^2	b^2	bc	b^2
15	3	3	$\{(P^1_k, P^4_k), (Q^1_k, Q^4_k)\}$	b^2	b^2	ab	bc	b^2	b^2	b^2	b^2	ab	bc	b^2	b^2	ab	ab	a^2	ac	ab	ab	bc	bc	ac	c^2	bc	bc	b^2	b^2	ab	bc	b^2	b^2	b^2	b^2	ab	bc	b^2	b^2
16	3	4	$\{(P^1_k, P^4_k), (Q^2_k, Q^3_k)\}$	b^2	b^2	bc	ab	b^2	b^2	b^2	b^2	bc	ab	b^2	b^2	ab	ab	ac	a^2	ab	ab	bc	bc	c^2	ac	bc	bc	b^2	b^2	ab	bc	b^2	b^2	b^2	b^2	bc	ab	b^2	b^2
17	3	5	$\{(P^1_k, P^4_k), (Q^2_k, Q^4_k)\}$	b^2	bc	b^2	b^2	ab	b^2	b^2	bc	b^2	b^2	ab	b^2	ab	ac	ab	ab	a^2	ab	bc	c^2	bc	bc	ac	bc	b^2	bc	b^2	b^2	bc	b^2	b^2	bc	b^2	b^2	ab	b^2
18	3	6	$\{(P^1_k, P^4_k), (Q^3_k, Q^4_k)\}$	bc	b^2	b^2	b^2	b^2	ab	bc	b^2	b^2	b^2	b^2	ab	ac	ab	ab	ab	a^2	c^2	bc	bc	bc	bc	ac	bc	b^2	b^2	b^2	b^2	ab	bc	b^2	b^2	b^2	b^2	b^2	ab
19	4	1	$\{(P^2_k, P^3_k), (Q^1_k, Q^2_k)\}$	ab	b^2	b^2	b^2	b^2	bc	ab	b^2	b^2	b^2	b^2	bc	ac	bc	bc	bc	bc	c^2	a^2	ab	ab	ab	ab	ac	ab	b^2	b^2	b^2	b^2	bc	ab	b^2	b^2	b^2	b^2	bc
20	4	2	$\{(P^2_k, P^3_k), (Q^1_k, Q^3_k)\}$	b^2	ab	b^2	b^2	bc	b^2	b^2	ab	b^2	b^2	bc	b^2	bc	ac	bc	bc	c^2	bc	ab	a^2	ab	ab	ac	ab	b^2	ab	b^2	b^2	bc	b^2	b^2	ab	b^2	b^2	bc	b^2
21	4	3	$\{(P^2_k, P^3_k), (Q^1_k, Q^4_k)\}$	b^2	b^2	ab	bc	b^2	b^2	b^2	b^2	ab	bc	b^2	b^2	bc	bc	ac	c^2	bc	bc	ab	ab	a^2	ac	ab	ab	b^2	b^2	ab	bc	b^2	b^2	b^2	b^2	ab	bc	b^2	b^2
22	4	4	$\{(P^2_k, P^3_k), (Q^2_k, Q^3_k)\}$	b^2	b^2	bc	ab	b^2	b^2	b^2	b^2	bc	ab	b^2	b^2	bc	bc	c^2	ac	bc	bc	ab	ab	ac	a^2	ab	ab	b^2	b^2	bc	ab	b^2	b^2	b^2	b^2	bc	ab	b^2	b^2
23	4	5	$\{(P^2_k, P^3_k), (Q^2_k, Q^4_k)\}$	b^2	bc	b^2	b^2	ab	b^2	b^2	bc	b^2	b^2	ab	b^2	bc	c^2	bc	bc	ac	bc	ab	ac	ab	ab	a^2	ab	b^2	bc	b^2	b^2	ab	b^2	b^2	bc	b^2	b^2	ab	b^2
24	4	6	$\{(P^2_k, P^3_k), (Q^3_k, Q^4_k)\}$	bc	b^2	b^2	b^2	b^2	ab	bc	b^2	b^2	b^2	b^2	ab	c^2	bc	bc	bc	bc	ac	ac	ab	ab	ab	ab	a^2	bc	b^2	b^2	b^2	b^2	ab	bc	b^2	b^2	b^2	b^2	ab
25	5	1	$\{(P^2_k, P^4_k), (Q^1_k, Q^2_k)\}$	ab	b^2	b^2	b^2	b^2	bc	ac	bc	bc	bc	bc	c^2	ab	b^2	b^2	b^2	b^2	bc	ab	b^2	b^2	b^2	b^2	bc	a^2	ab	ab	ab	ab	ac	ab	b^2	b^2	b^2	b^2	bc
26	5	2	$\{(P^2_k, P^4_k), (Q^1_k, Q^3_k)\}$	b^2	ab	b^2	b^2	bc	b^2	bc	ac	bc	bc	c^2	bc	b^2	ab	b^2	b^2	bc	b^2	b^2	ab	b^2	b^2	bc	b^2	ab	a^2	ab	ab	ac	ab	b^2	ab	b^2	b^2	bc	b^2
27	5	3	$\{(P^2_k, P^4_k), (Q^1_k, Q^4_k)\}$	b^2	b^2	ab	bc	b^2	b^2	bc	bc	ac	c^2	bc	bc	b^2	b^2	ab	bc	b^2	b^2	b^2	ab	bc	b^2	b^2	ab	ab	a^2	ac	ab	ab	b^2	b^2	ab	bc	b^2	b^2
28	5	4	$\{(P^2_k, P^4_k), (Q^2_k, Q^3_k)\}$	b^2	b^2	bc	ab	b^2	b^2	bc	bc	c^2	ac	bc	bc	b^2	b^2	bc	ab	b^2	b^2	b^2	b^2	bc	ab	b^2	b^2	ab	ab	ac	a^2	ab	ab	b^2	b^2	bc	ab	b^2	b^2
29	5	5	$\{(P^2_k, P^4_k), (Q^2_k, Q^4_k)\}$	b^2	bc	b^2	b^2	ab	b^2	bc	c^2	bc	bc	ac	bc	b^2	bc	b^2	b^2	ab	b^2	b^2	bc	b^2	b^2	ab	b^2	ab	ac	ab	ab	a^2	ab	b^2	bc	b^2	b^2	ab	b^2
30	5	6	$\{(P^2_k, P^4_k), (Q^3_k, Q^4_k)\}$	bc	b^2	b^2	b^2	b^2	ab	c^2	bc	bc	bc	bc	ac	bc	b^2	b^2	b^2	b^2	ab	bc	b^2	b^2	b^2	b^2	ab	ac	ab	ab	ab	ab	a^2	bc	b^2	b^2	b^2	b^2	ab
31	6	1	$\{(P^3_k, P^4_k), (Q^1_k, Q^2_k)\}$	ac	bc	bc	bc	bc	c^2	ab	b^2	b^2	b^2	b^2	bc	ab	b^2	b^2	b^2	b^2	bc	ab	b^2	b^2	b^2	b^2	bc	ab	b^2	b^2	b^2	b^2	bc	a^2	ab	ab	ab	ab	ac
32	6	2	$\{(P^3_k, P^4_k), (Q^1_k, Q^3_k)\}$	bc	ac	bc	bc	c^2	bc	b^2	ab	b^2	b^2	bc	b^2	b^2	ab	b^2	b^2	bc	b^2	b^2	ab	b^2	b^2	bc	b^2	b^2	ab	b^2	b^2	bc	b^2	ab	a^2	ab	ab	ac	ab
33	6	3	$\{(P^3_k, P^4_k), (Q^1_k, Q^4_k)\}$	bc	bc	ac	c^2	bc	bc	b^2	b^2	ab	bc	b^2	b^2	b^2	b^2	ab	bc	b^2	b^2	b^2	b^2	ab	bc	b^2	b^2	b^2	b^2	ab	bc	b^2	b^2	ab	ab	a^2	ac	ab	ab
34	6	4	$\{(P^3_k, P^4_k), (Q^2_k, Q^3_k)\}$	bc	bc	c^2	ac	bc	bc	b^2	b^2	bc	ab	b^2	b^2	b^2	b^2	bc	ab	b^2	b^2	b^2	b^2	bc	ab	b^2	b^2	b^2	b^2	bc	ab	b^2	b^2	ab	ab	ac	a^2	ab	ab
35	6	5	$\{(P^3_k, P^4_k), (Q^2_k, Q^4_k)\}$	bc	c^2	bc	bc	ac	bc	b^2	bc	b^2	b^2	ab	b^2	b^2	bc	b^2	b^2	ab	b^2	b^2	bc	b^2	b^2	ab	b^2	b^2	bc	b^2	b^2	ab	b^2	ab	ac	ab	ab	a^2	ab
36	6	6	$\{(P^3_k, P^4_k), (Q^3_k, Q^4_k)\}$	c^2	bc	bc	bc	bc	ac	bc	b^2	b^2	b^2	b^2	ab	bc	b^2	b^2	b^2	b^2	ab	bc	b^2	b^2	b^2	b^2	ab	bc	b^2	b^2	b^2	b^2	ab	ac	ab	ab	ab	ab	a^2

where $a = (1-r)^2$, $b = \frac{(1-r)r}{2}$ and $c = r^2$

Table S3.8 : Full transition space in autotetraploids. The table contain all possible transitions $\Pr(\mathcal{G}^4_{k',j'}|\mathcal{G}^4_{k,j})$ between the 36 possible genotypic states for markers k and k'. For simplicity, apart from j', the name of the columns i', k' and $\mathcal{G}^4_{k',j'} = \{\theta^4_{P,i'}, \theta^4_{Q,h'}\}$ were omitted but they are analogous and in the same order as the name of the rows.

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

Marcelo Mollinari and Antonio Augusto Franco Garcia

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

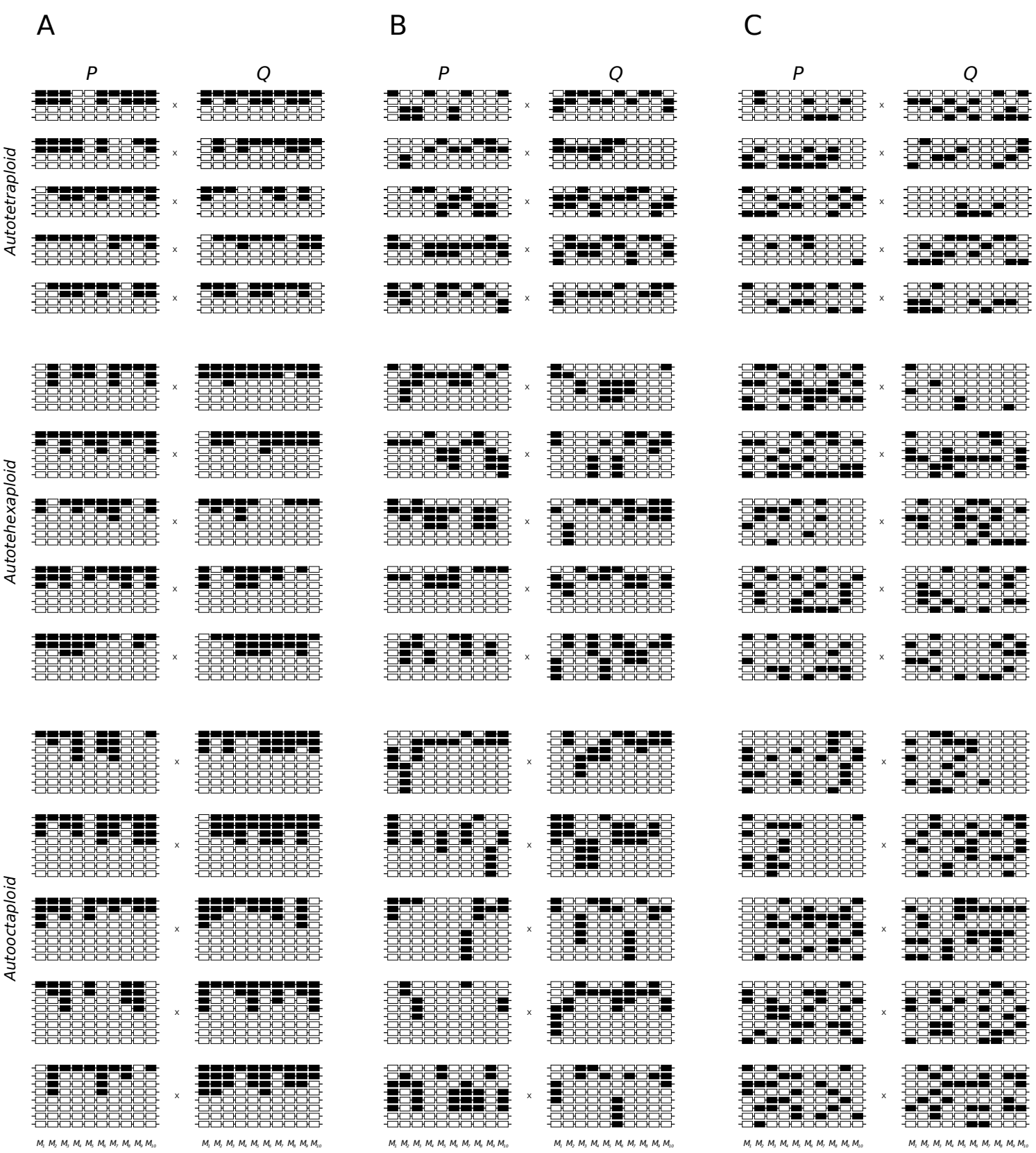

Figure S4 : Simulated haplotypes with 10 markers and three ploidy levels, namely autotetraploid ($m = 4$), autohexaploid ($m = 6$) and autooctaploid ($m = 8$). Black and white rectangles indicate two allelic variants in each marker. In all scenarios, the marker doses varied from zero to $\frac{m}{2}$. Although the method can cope with dosages greater than $\frac{m}{2}$, they are equivalent to lower doses in different linkage phases, therefore they were not considered. Each horizontal line indicates homologous chromosomes which are grouped in homology groups. Three linkage phase scenarios were simulated: In scenario A, allelic variants represented by black rectangle were assigned to the first homologous chromosome in the homology group. The remaining variants of the same type where assigned to the subsequent homologous chromosomes. This yields patterns where allelic variants of the same type were mostly concentrated in the same homologous chromosomes. In B, one type of allelic variant was randomly assigned to one of the first $\frac{m}{2}$ homologous chromosome and the remaining allelic variants of the same type were assigned to the subsequent homologous chromosomes; in C, the allelic variants were randomly assigned to the m homologous chromosomes. Five different parental haplotypes were considered for each linkage phase scenario. In total, 45 parental linkage phase configurations were considered ($3 \times 3 \times 5 = 45$).

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

Marcelo Mollinari and Antonio Augusto Franco Garcia

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

Autotetraploid

Figure S5 : Simulated haplotypes with 200 markers and two ploidy levels, namely autotetraploid ($m = 4$) and autohexaploid ($m = 6$). Black and white rectangles indicate two allelic variants in each marker. In all configurations, marker doses varied from zero to $\frac{m}{2}$. Each horizontal line indicates homologous chromosomes which are grouped in homology groups. We used scenario C, from simulation study 1, as the template for this second simulation study, i.e., the allelic variants were randomly assigned to the m homologous chromosomes. Five different parental haplotypes were considered for each ploidy level.

Autohexaploid

S5 Figure (cont.) : Simulated haplotypes with 200 markers and two ploidy levels, namely autotetraploid ($m = 4$) and autohexaploid ($m = 6$). Black and white rectangles indicate two allelic variants in each marker. In all configurations, marker doses varied from zero to $\frac{m}{2}$. Each horizontal line indicates homologous chromosomes which are grouped in homology groups. We used scenario C, from simulation study 1, as the template for this second simulation study, i.e., the allelic variants were randomly assigned to the m homologous chromosomes. Five different parental haplotypes were considered for each ploidy level.

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

Marcelo Mollinari and Antonio Augusto Franco Garcia

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

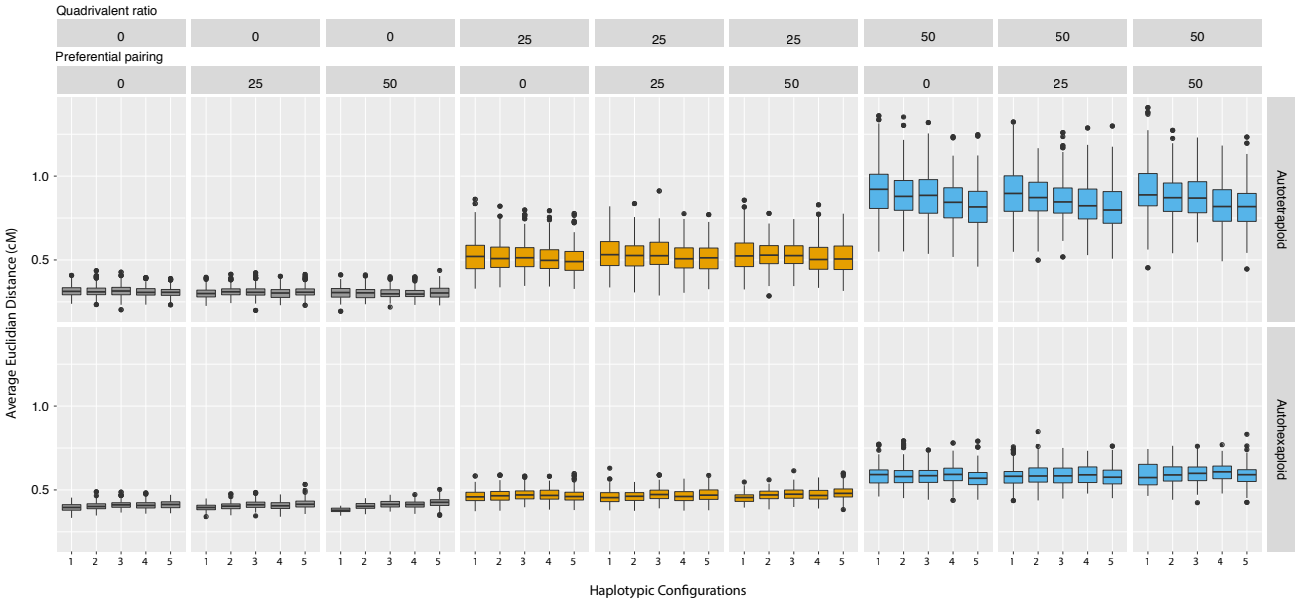

Figure S6 : Boxplots of the average Euclidean distances between the estimated and simulated distance vectors for simulation study 2. Only recombination fraction verctors from correctly estimated linkage phase configurations are considered. Each column contains a set of 5 haplotypic configurations and indicates a combination of quadrivalent formation rate and preferential pairing for autotetraploid configurations (top row) and for autohexaploid configurations (bottom row)

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

Marcelo Mollinari and Antonio Augusto Franco Garcia

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

References

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

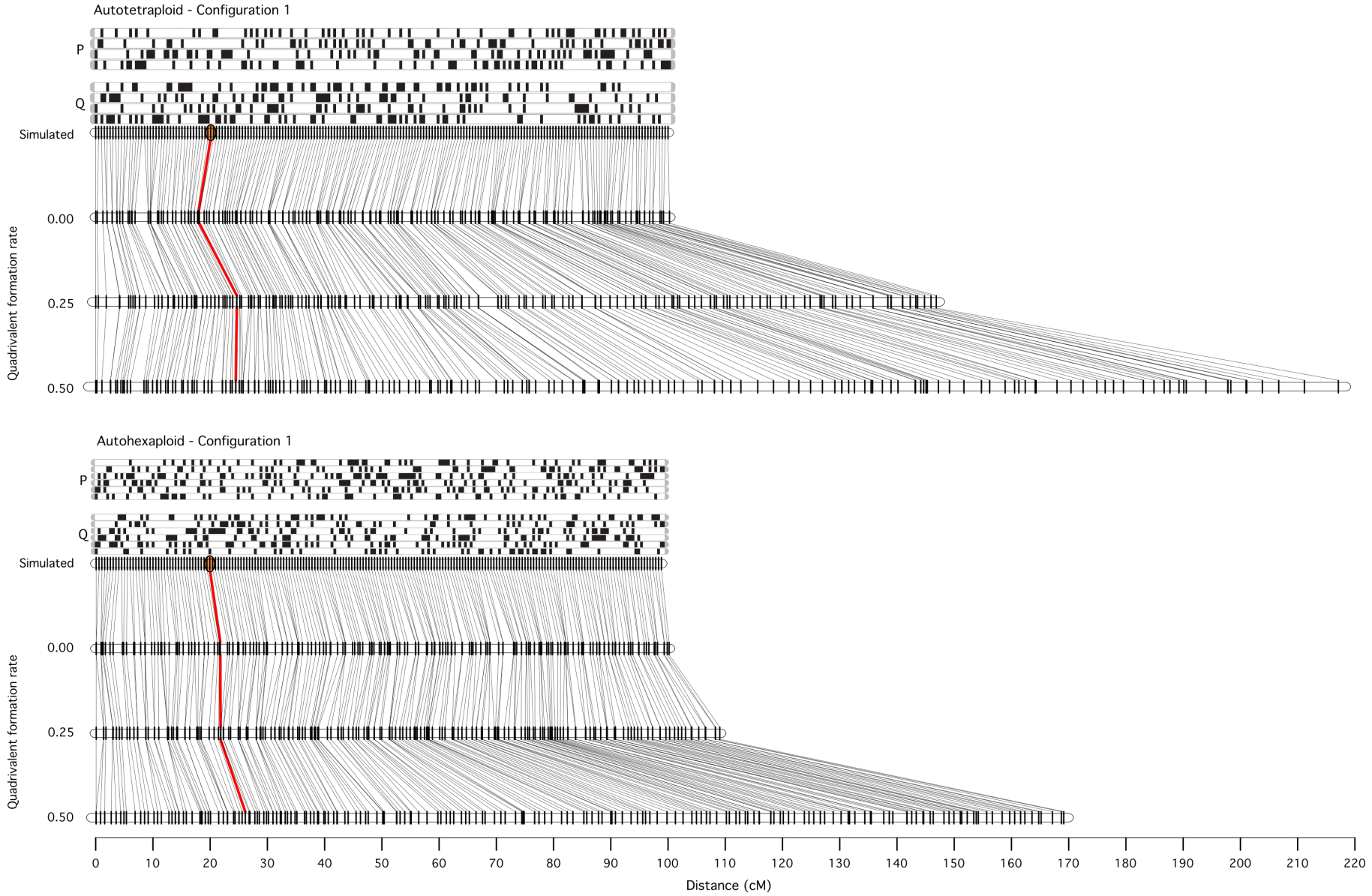

Figure S7 : Two examples of the effect of increasing quadrivalent formation rate in genetic maps selected from simulation 2. Black and white rectangles indicate two allelic variants in each marker. Each horizontal line indicates homologous chromosomes which are grouped in homology groups in parents P and Q. Autotetraploid and autohexaploid examples are presented. Both examples show the haplotypic configuration used to simulate 200 equally spaced markers with a final length of 100 cM (configuration 1, simulation 2). The centromere was positioned at 20.0 cM from the beginning of the chromosome (orange dot). Datasets were simulated with three levels of quadrivalent formation rate (0.00, 0.25 and 0.50) and random chromosome segregation (non preferential pairing) using the software PedigreeSim [1]. Maps were estimated for each dataset and are presented here. The red line connects the marker simulated at the centromere with the position of the same marker at the estimated maps.

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

Marcelo Mollinari and Antonio Augusto Franco Garcia

File S8: Summary of results from B2721 population map construction

Table S8.1 : Summary of the B2721 population genetic map comprising 12 linkage groups

Chromosome number	map length (cM)		LOD[1]	map length (cM)		LOD[1]	map length (cM)		LOD[1]	Number of markers
	de novo	genomic		*de novo* + fitTetra proportions	genomic + fitTetra proportions		*de novo* + global error	genomic + global error		
1	332.5	264.2	549.5	294.3	230.3	576.6	159.2	151.4	307.3	368
2	255.3	173.7	701.8	231.9	147.9	728.2	127.3	105.4	491.0	225
3	238.7	165.2	649.4	211.3	140.0	692.1	112.8	96.0	474.5	274
4	290.6	185.7	945.1	257.8	152.2	1016.9	119.6	95.5	683.6	373
5	183.1	164.1	169.0	151.8	127.9	207.2	91.7	84.5	154.9	248
6	251.4	156.8	939.8	230.8	136.1	1019.4	114.1	88.5	808.2	371
7	234.7	160.3	673.2	201.8	122.0	753.3	111.3	91.3	509.9	365
8	165.2	124.5	388.5	144.1	103.7	424.1	89.7	80.3	270.3	242
9	219.5	175.2	384.9	184.9	136.3	418.7	117.6	104.2	295.5	261
10	186.1	154.9	172.0	168.9	135.1	206.3	104.7	95.3	191.5	200
11	185.8	157.3	269.6	158.5	126.0	292.8	105.7	96.0	210.3	234
12	187.6	157.4	237.9	151.3	119.1	273.9	88.0	88.4	134.5	187

1: LOD Scrore was computed using the base-10 logarithm of ratio between the multilocus likelihood of the *de novo* ordered map and the genomic ordered map (i.e., the likelihood of the genomic ordered map was higher in all cases).

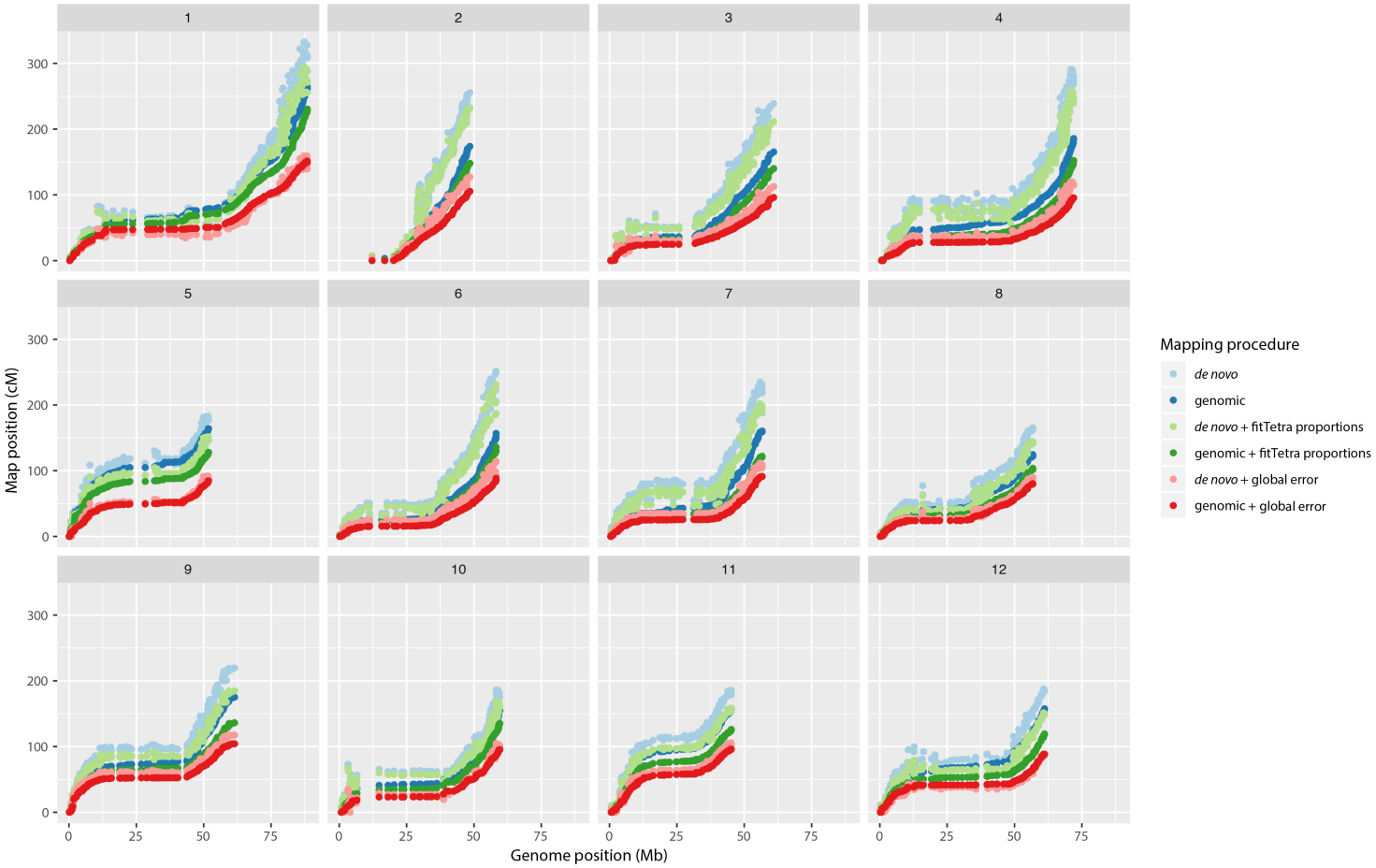

Figure S8.1 : Scatter plot of map distances versus genomic positions for 12 linkage groups in B2721 population. We used two different SNP orders: : a *de novo* order provided by MDS algorithm [1] and the order obtained from the *Solanum tuberosum* genome version 4.03 [2]. For each order, we applied three recombination fraction multilocus estimation procedures: (*i*) using marker dosage, (*ii*) using the proportions provided by the R package fitTetra [3] in the HMM model and (*iii*) using a global error in the HMM model.

References

1. K. F. Preedy and C. A. Hackett. A rapid marker ordering approach for high-density genetic linkage maps in experimental autotetraploid populations using multidimensional scaling. *Theor. Appl. Genet.*, 129(11):2117–2132, 2016.

2. Sanjeev Kumar Sharma, Daniel Bolser, Jan de Boer, Mads Sønderkær, Walter Amoros, Martin Federico Carboni, Juan Martín D'Ambrosio, German de la Cruz, Alex Di Genova, David S. Douches, Maria Eguiluz, Xiao Guo, Frank Guzman, Christine A. Hackett, John P. Hamilton, Guangcun Li, Ying Li, Roberto Lozano, Alejandro Maass, David Marshall, Diana Martinez, Karen McLean, Nilo Mejía, Linda Milne, Susan Munive, Istvan Nagy, Olga Ponce, Manuel Ramirez, Reinhard Simon, Susan J. Thomson, Yerisf Torres, Robbie Waugh, Zhonghua Zhang, Sanwen Huang, Richard G. F. Visser, Christian W. B. Bachem, Boris Sagredo, Sergio E. Feingold, Gisella Orjeda, Richard E. Veilleux, Merideth Bonierbale, Jeanne M. E. Jacobs, Dan Milbourne, David Michael Alan Martin, and Glenn J. Bryan. Construction of Reference Chromosome-Scale Pseudomolecules for Potato: Integrating the Potato Genome with Genetic and Physical Maps. *G3*, 3(11):2031–2047, 2013.

3. Roeland E Voorrips, Gerrit Gort, and Ben Vosman. Genotype calling in tetraploid species from bi-allelic marker data using mixture models. *BMC Bioinformatics*, 12(1):172, 2011.

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

Marcelo Mollinari and Antonio Augusto Franco Garcia

Figure S9: Simulated haplotypes for comparison between polymapR and HMM-based method

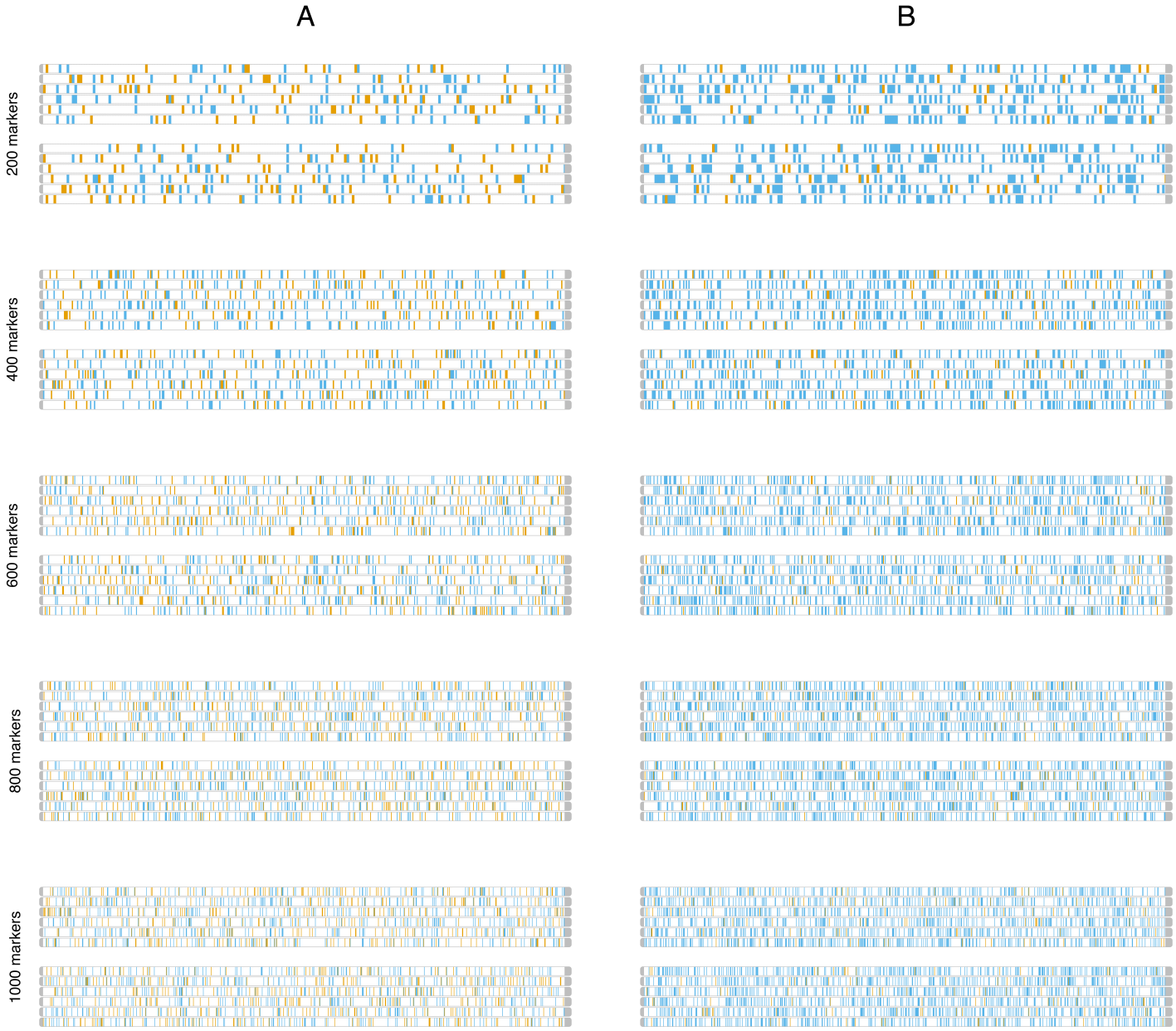

Figure S9 : Simulated haplotypes with 200, 400, 600, 800 and 1000 markers in autohexaploid ($m = 6$) parents. Orange rectangles indicate alternative allelic variants in simplex and double simplex markers; blue rectangles indicate alternative allelic variants in the remaining dosage configurations. (A) proportions of markers with 0 and 1 dose: 40%; markers with 2 and 3 doses: 10%; (B) proportions for all dosage types from 0 to 3: 25%. Proportions simulated for both parents. Each horizontal line indicates homologous chromosomes which are grouped in homology groups. Similarly to Simulation 2, the allelic variants were randomly assigned to the m homologous chromosomes.

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

Marcelo Mollinari and Antonio Augusto Franco Garcia

File S10: Results of comparison between polymapR and HMM-based method

Table S10.1 : Percentage of correctly phased markers in parents P and Q across 50 simulations. Numbers in parentheses indicate the percentage of markers included in the resulting phased map.

Dosage proportions	Number of markers		LOD = 3^i				LOD = 5^i			
			polymapR		HMM-based		polymapR		HMM-based	
40%, 40%, 10%, 10%^{ii}	200	P	100.0	(90.5)	100.0	(100.0)	100.0	(82.3)	100.0	(99.5)
		Q	99.7		100.0		99.7		100.0	
	400	P	100.0	(94.0)	99.8	(99.8)	100.0	(81.5)	100.0	(99.6)
		Q	100.0		100.0		100.0		100.0	
	600	P	100.0	(95.5)	100.0	(100.0)	100.0	(86.5)	100.0	(100.0)
		Q	100.0		100.0		100.0		100.0	
	800	P	100.0	(95.2)	100.0	(100.0)	100.0	(86.1)	100.0	(99.9)
		Q	100.0		99.8		100.0		100.0	
	1000	P	100.0	(96.1)	99.7	(100.0)	100.0	(88.5)	100.0	(99.9)
		Q	100.0		99.8		100.0		100.0	
25%, 25%, 25%, 25%^{ii}	200	P	100.0	(16.9)	99.7	(99.3)	100.0	(14.4)	100.0	(97.8)
		Q	99.9		100.0		99.9		100.0	
	400	P	98.5	(32.4)	98.9	(99.5)	98.6	(26.1)	100.0	(99.0)
		Q	96.2		99.1		96.2		100.0	
	600	P	97.7	(63.6)	100.0	(99.9)	97.5	(47.3)	100.0	(99.7)
		Q	100.0		99.6		100.0		100.0	
	800	P	100.0	(68.6)	100.0	(99.9)	100.0	(51.3)	100.0	(99.8)
		Q	100.0		99.6		100.0		100.0	
	1000	P	100.0	(77.5)	99.2	(99.9)	100.0	(57.0)	100.0	(99.9)
		Q	100.0		98.8		100.0		100.0	

i: For the HMM-based method, the LOD threshold indicates the value from which the multipoint likelihood was used to chose the best phase configuration, polymapR uses LOD thresholds to make decisions whether a marker should be used in a certain mapping context, such as, clustering homologous chromosomes or to assign it into assembled linkage groups. Thus, they are not directly comparable.
ii: (40%, 40%, 10%, 10%) represents higher proportion of simplex and double simplex markers, with 40% of the simulated markers being nulliplex, 40% simplex, 10% duplex and 10% triplex in both parents; (25%, 25%, 25%, 25%) represents equal proportions for all doses, with 25% for all dosage types, from nulliplex to triplex (see Figure S9)

Table S10.2 : Average length and standard deviation (in parentheses) of the maps across the correctly phased simulations.

Dosage proportion	Number of markers	LOD = 3				LOD = 5			
		polymapR[i]	HMM-based[ii]			polymapR[i]	HMM-based[ii]		
		MDS + PC	simulated	MDS	MDS + global error	MDS + PC	simulated	MDS	MDS + global error
40%, 40%, 10%, 10%[iii]	200	87.8(4.5)	96.7(3.0)	143.8(7.1)	99.2(3.9)	87.5(4.3)	96.6(3.0)	143.0(7.3)	99.0(4.0)
	400	88.8(3.5)	99.0(2.6)	212.4(12.3)	103.3(3.5)	89.9(10.3)	98.9(2.7)	211.8(12.3)	103.3(3.5)
	600	87.2(2.9)	98.9(2.6)	276.9(13.8)	106.9(3.9)	87.3(3.0)	98.9(2.6)	277.0(13.8)	107.0(3.9)
	800	89.5(14.0)	99.7(2.5)	350.7(16.3)	109.0(4.0)	87.7(3.2)	99.6(2.4)	350.1(16.4)	109.0(4.0)
	1000	87.0(3.3)	100.0(2.9)	413.7(20.0)	114.2(4.4)	87.0(3.2)	100.1(3.0)	413.1(19.4)	114.1(4.3)
25%, 25%, 25%, 25%[iii]	200	79.6(4.4)	97.7(3.1)	172.1(9.4)	101.2(3.9)	78.7(5.4)	97.7(3.1)	169.3(9.0)	100.9(4.0)
	400	76.9(1.9)	99.4(2.8)	254.8(12.7)	104.3(3.3)	76.7(2.7)	99.5(2.7)	254.2(13.4)	104.5(3.3)
	600	78.0(2.9)	99.2(2.6)	334.5(13.9)	106.4(2.9)	78.5(2.9)	99.2(2.6)	334.4(14.8)	106.2(3.3)
	800	79.1(2.7)	99.7(2.2)	423.7(19.9)	110.8(3.5)	78.9(2.6)	99.7(2.3)	423.4(19.8)	110.9(3.4)
	1000	80.1(8.8)	100.1(3.1)	513.1(23.1)	115.9(4.5)	78.5(3.2)	100.2(3.0)	511.2(20.8)	115.5(4.3)

i: For polymapR, the marker positions were estimated using the projection of the MDS result onto a single dimension principal curve (PC).
ii: For HMM-based method, we present the map length given the simulated and the MDS order. For the MDS order, we present the results of the map re-estimation using a 5% global error.
iii: (40%, 40%, 10%, 10%) represents higher proportion of simplex and double simplex markers, with 40% of the simulated markers being nulliplex, 40% simplex, 10% duplex and 10% triplex in both parents; (25%, 25%, 25%, 25%) represents equal proportions for all doses, with 25% for all dosage types, from nulliplex to triplex (see Figure S9)

Table S10.3 : Average of the elapsed time (in minutes) across 50 simulations.

Dosage proportions	Number of markers	LOD = 3		LOD = 5	
		polymapR	HMM-based	polymapR	HMM-based
40%, 40%, 10%, 10%	200	1.8	4.9	1.2	8.9
	400	2.0	11.6	1.8	14.2
	600	3.9	22.1	3.7	26.2
	800	7.0	37.5	6.5	43.8
	1000	10.0	59.1	9.5	65.1
25%, 25%, 25%, 25%	200	2.3	6.6	0.5	14.4
	400	4.2	17.0	2.4	22.1
	600	11.2	25.6	7.9	29.6
	800	21.4	44.0	14.9	48.8
	1000	34.5	69.3	25.4	73.3
Each map was constructed in a single Intel® Xeon® CPU E5-2670 @ 2.60GHz, 128GB RAM.					

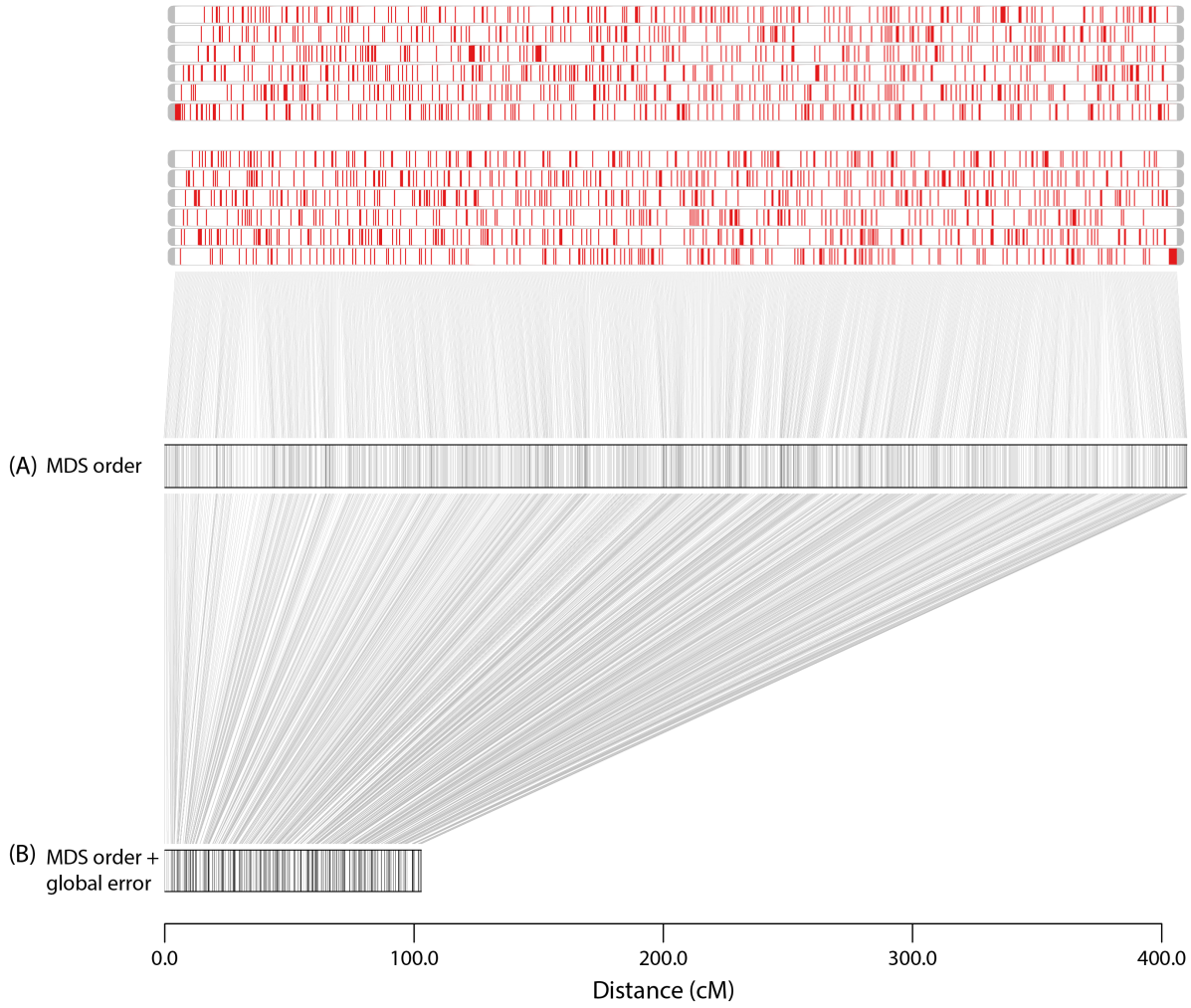

Figure S10.1 : Example of MDS ordered map estimated using marker dosages with no error modeling (A) and modeling a global error of 5% in the HMM emission function (B).