A mechanistic model of linkage analysis in allohexaploids

Huan Li[*1], Xuli Zhu[*1], Qin Yan[1], Ke Mao[1], and Rongling Wu[1,2]

[1]Center for Computational Biology, College of Biological Sciences and Technology, Beijing Forestry University, Beijing 100083, China

[2]Center for Statistical Genetics, The Pennsylvania State University, Hershey, PA 17033, USA

[*]These authors contributed to this work equally.

Running Head: Linkage analysis in allohexaploids

Corresponding author:

Rongling Wu

Center for Statistical Genetics

The Pennsylvania State University

Hershey, PA 17033

USA

Tel: (717)531-2037

Email: rwu@phs.psu.edu

Abstract: Despite their pivotal role in agriculture and biological research, polyploids, a group of organisms with more than two sets of chromosomes, are very difficult to study. Increasing studies have used high-density genetic linkage maps to investigate the genome structure and function of polyploids and to identify genes underlying polyploid traits. However, although models for linkage analysis have been well established for diploids, with some essential modifications for tetraploids, no models have been available thus far for polyploids at higher ploidy levels. The linkage analysis of polyploids typically requires knowledge about their meiotic mechanisms, depending on the origin of polyplody. Here we describe a computational modeling framework for linkage analysis in allohexaploids by integrating their preferential chromosomal-pairing meiotic feature into a mixture model setting. The framework, implemented with the EM algorithm, allows the simultaneous estimates of preferential pairing factors and the recombination fraction. We investigated statistical properties of the framework through extensive computer simulation and validated its usefulness and utility by analyzing a real data from a full-sib family of allohexaploid persimmon. Our attempt in linkage analysis of allohexaploids by incorporating their meiotic mechanism lays a foundation for allohexaploid genetic mapping and also provides a new horizon to explore allohexaploid parental kinship.

Key words: allohexaploid, preferential pairing factor, recombination fraction, EM algorithm, persimmon

INTRODUCTION

Polyploidy is an important force for the evolution of plants (Otto and Whitton 2000; Soltis and Soltis 2000). It was estimated that 70 – 80% of angiosperms are polyploids or experienced phases of polyploids during their evolutionary process (Lewis 1979; Masterson 1994). Many crops, such as wheat, sugarcanes, potato, cotton and canola, are polyploids, which play a central role in agriculture (Leitch and Leitch 2008). Polyploids can be classified into two types, i.e., allopolyploids, whose chromosomes are composed of distinct genomes

through interspecific hybridization, and autopolyploids, in which the chromosome doubling of genetically similar genomes is due to the fusion of unreduced gametes (Müntzing 1936; Soltis and Soltis 2000; Soltis et al. 2004; Gaeta and Pires 2010). Of all ploypoids, more than 75% are found to be allopolyploids (Soltis and Soltis 2009). A growing body of evidence indicates that polyploids have great advantages in response to selection and adaption partly through increased rates of meiotic recombination (Soltis and Soltis 1999; Grant 2004; Comai 2005; Chen 2010; Pecinka et al. 2011).

The nature of polyploids can be depicted through how chromosomes pair at meiosis. According to this criterion, polyploids can be sorted into bivalent polyploids, multivalent polyploids and mixed polyploids (Comai 2005). In general, extreme allopolyploids present bivalent formation in which more similar chromosomes are expected to have higher pairing frequencies than less similar chromosomes, a phenomenon which can be described by the preferential pairing factor (Sybenga 1988). On the other hand, extreme autopolyploids are pervaded by multivalent formation in which more than two chromosomes pair at a time, resulting in the appearance of two sister chromatids into the same gamete, called double reduction (Hauber et al. 1999). Mixed polyploids with both bivalent and multivalent formation are confounded by both preferential pairing and double reduction (Wu et al. 2004; Burke et al. 2015).

The past two decades have witnessed increasing studies of linkage mapping in polyploids (Ripol et al. 1999; Luo et al. 2001a,b; Kriegner et al. 2003; McCord al. 2011; Hackett et al. 2013, 2014; Monden et al. 2015; Bourke et al. 2015). Linkage maps are constructed on the basis of segregation and transmission of genes into the next progeny generation. Due to some unique cytological phenomena during meiosis, e.g., double reduction and preferential pairing, statistical models for linkage analysis in polyploids should be qualitatively more complex than those in diploids. This complexity has led to tremendous development of powerful statistical models for linkage analysis and QTL mapping in tetraploid (Hackett et al. 1998, 2013; Luo et al. 2001; Rehmsmeier 2013). Sybenga (1965, 1966) recognized the event of unequal paring probability during chromosome synapsis and developed mathematical models to describe different chromatid paring probabilities in polyploids. By taking bivalent and multivalent pairing formations into account during meiosis, Wu and group developed a series of models for linkage analysis, map construction and QTL mapping (Wu et al. 2001a,b, 2002;

Ma et al. 2002; Lu et al. 2013; Yang et al. 2013; Xu et al. 2014a,b). The phenomenon of double reduction was also considered in Luo et al.'s (2001a,b) autotetraploid model and Rehmsmeier's (2013) computational model. More recently, Li et al. (2010) developed a specialized EM algorithm for QTL mapping in multivalent tetraploids, which may impact in the field of polyploid QTL mapping. Overcoming the drawback of two-point linkage analysis, Yang et al. (2013) and Lu et al. (2013) developed a three-point linkage analysis model which can not only accurately estimate the linkage between loci, but also detect genetic interference throughout the genome.

While all these polyploidy linkage models are focused on tetraploids, there is still a gap in the model development of linkage analysis in hexaploids despite their significant importance in agriculture and biology (Monden et al. 2015). In this article, we describe and assess a statistical model that embeds preferential chromosomal-pairing within the framework of linkage analysis in allohexaploids. A considerable body of evidence shows that chromosome pairing occurs between homeologues during meiosis and homeologous recombination plays an important role in chromosomal rearrangements (Nicolas et al. 2007; Lim et al. 2008; Gaeta and Pires 2010; Xiong et al. 2011). The new model allows us to simultaneously estimate the preferential pairing factor and recombination fraction between any pair of molecular markers. We outline a detailed procedure to test the significance of these two parameters, facilitating the studies of allohexaploid genome structure and organization. The model offers a useful tool for linkage mapping and population genetic studies in allohexaploids.

The Model

Preferential pairing factor

The probability, with which more similar chromosomes pair more frequently than less similar chromosomes, is defined as the preferential pairing factor (Sybenga 1966). For an allopolyploid system exhibiting bivalent formation, it is expected that the preferential pairing factor influences gamete formation and frequencies. Consider a heterozygous allohexaploid derived from the chromosomal combination of three distinct diploid genomes **A**, **B** and **C**. Six sets of chromosomes in this allohexaploid are labeled as **1**, **2**, **3**, **4**, **5** and **6**, respectively. Assume that chromosomes **1** and **2** are homologous, as are chromosomes **3** and **4**, as well as chromosomes **5** and **6**. Under bivalent pairing, there are a total of 15 possible pair-wise

combinations among six chromosomes, expressed as **1** and **2**, **1** and **3**, …, **5** and **6**. Denote the preferential pairing factor as θ_1 between chromosome **1** and **2**, θ_2 between chromosome **3** and **4**, and θ_3 between chromosome **5** and **6**. Thus, the frequencies of pairing of any two chromosomes are derived as

$$\begin{aligned} &p_{12} = \tfrac{1}{15} + \theta_1,\ p_{34} = \tfrac{1}{15} + \theta_2,\ p_{56} = \tfrac{1}{15} + \theta_3; \\ &p_{13} = p_{14} = p_{15} = p_{16} = p_{23} = p_{24} = p_{25} = \\ &p_{26} = p_{35} = p_{36} = p_{45} = p_{46} = \tfrac{1}{15} - \tfrac{1}{12}(\theta_1 + \theta_2 + \theta_3) \end{aligned} \quad (1)$$

where the subscripts stand for 15 possible chromosome pairs.

For an allohexaploid genotype **123456**, its bivalent pairing takes place in three different ways: fully preferential pairing, partially preferential pairing and no preferential pairing, each of which is derived from a particular chromosomal-pairing configuration that occurs with a different frequency and generates different groups of triploid gametes at meiosis (Table 1). A chromosomal-pairing configuration is defined by using ‖ to separate pairing chromosomes. For example, **12‖34‖56** is a chromosomal-pairing configuration in which chromosome 1 pairs with 2, 3 with 4 and 5 with 6. Expressions for the frequencies of chromosomal-pairing configurations are given in supplementary materail 1. Each chromosomal-pairing configuration produces eight triploid gametes, leading to 15 × 8 = 120 gametes in total. Virtually, these gametes are distinguished by 20 types, i.e., 123, 124, …, 456, whose frequencies are expressed as g_{123}, g_{124}, $\cdots$, g_{456}, respectively.

Meiotic chromosomal segregation

Our model focuses on an allohexaploid that undergoes only bivalent pairing during meiosis. Suppose that a heterozygous allohexaploid line is crossed with a homozygous line to generate a pseudo-test backcross in which the genotypes of the progeny are consistent with the genotypes of gametes produced by the heterozygous parent. Assume that there are two fully informative markers, **A** and **B**, which are both heterozygous in one parent but homozygous in another parent. We denote six different alleles as a_1, …, a_6 at marker **A** and b_1, …, b_6 at marker **B**. Nonalleles at the two markers are located in six chromosomes with $6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$ possible linkage phases, expressed as

$$\begin{matrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 \\ b_1 & b_2 & b_3 & b_4 & b_5 & b_6 \end{matrix} \quad (2\text{-}1)$$

$$\begin{matrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 \\ b_1 & b_2 & b_3 & b_4 & b_5 & b_6 \end{matrix} \tag{2-2}$$

$$\vdots \qquad\qquad \vdots$$

$$\begin{matrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 \\ b_6 & b_5 & b_4 & b_3 & b_2 & b_1 \end{matrix}. \tag{2-720}$$

A heterozygous allohexaploid with one particular linkage phase above has 15 possible chromosomal-pairing configurations. For linkage phase (2-1), these configurations can be expressed, in the order shown in Table 1, as

$$\begin{matrix} a_1 & a_2 \\ b_1 & b_2 \end{matrix} \begin{matrix} a_3 & a_4 \\ b_3 & b_4 \end{matrix} \begin{matrix} a_5 & a_6 \\ b_5 & b_6 \end{matrix} \tag{3-1}$$

$$\begin{matrix} a_1 & a_2 \\ b_1 & b_2 \end{matrix} \begin{matrix} a_3 & a_5 \\ b_3 & b_5 \end{matrix} \begin{matrix} a_4 & a_6 \\ b_4 & b_6 \end{matrix} \tag{3-2}$$

$$\vdots \qquad\qquad \vdots$$

$$\begin{matrix} a_1 & a_6 \\ b_1 & b_6 \end{matrix} \begin{matrix} a_2 & a_4 \\ b_2 & b_4 \end{matrix} \begin{matrix} a_3 & a_5 \\ b_3 & b_5 \end{matrix} \tag{3-15}$$

where we assume that chromosomes **1** and **2**, **3** and **4**, and **5** and **6** are each homologous. Under fully preferentially pairing (3-1), marker **A** produces eight triploid gametes $a_1a_3a_5$, $a_1a_3a_6$, $a_1a_4a_5$, $a_1a_4a_6$, $a_2a_3a_5$, $a_2a_3a_6$, $a_2a_4a_5$, and $a_2a_4a_6$ with the same thing as marker **B**. Let r denote the recombination fraction between markers **A** and **B**. According to the number of crossover between markers **A** and **B**, meiotic gametes fall into 4 categories: no crossover, one crossover, two crossovers and three crossovers, with frequencies denoted as p_0, p_2, p_2, p_3, and p_4, respectively, which are expressed, in terms of r, as follows:

$$p_0 = \left(\frac{1-r}{2}\right)^3$$

$$p_1 = \frac{r}{2}\left(\frac{1-r}{2}\right)^2$$

$$p_2 = \left(\frac{r}{2}\right)^2\left(\frac{1-r}{2}\right)$$

$$p_3 = \left(\frac{r}{2}\right)^3$$

By combining three haploid gametes each from a different chromosome, triploid gametes are generated. Table 1 lists all possible triploid gametes and their probabilities produced under fully preferentially pairing (3-1) of linkage phase (2-1).

Likelihood, estimation and tests

Consider a full-sib family derived from two allohexaploid parents in which markers may be segregating in two manners. One is the intercross segregation at which both parents are heterozygous. The second is the testcross segregation at which one parent is heterozygous whereas the second is homozygous. As the demonstration of model derivation, we consider two testcross markers by crossing the parents

$$\begin{vmatrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 \\ b_1 & b_2 & b_3 & b_4 & b_5 & b_6 \end{vmatrix} \times \begin{vmatrix} a_1 & a_1 & a_1 & a_1 & a_1 & a_1 \\ b_1 & b_1 & b_1 & b_1 & b_1 & b_1 \end{vmatrix}. \quad (4)$$

Let $n_{k_1k_2k_3/l_1l_2l_3}$ denote the observation of progeny with triploid gamete genotype $k_1k_2k_3$ ($k_1 < k_2 < k_3 = 1, \ldots, 6$) at marker **A** and triploid gamete genotype $l_1l_2l_3$ ($l_1 < l_2 < l_3 = 1, \ldots, 6$) at marker **B** derived from the heterozygous parent. Correspondingly, the probability of a two-marker triploid gamete genotype is denoted as $p_{k_1k_2k_3/l_1l_2l_3}$. Then, we formulate a likelihood for observed genotype data, expressed as

$$\log L = \sum_{k_1<k_2<k_3} \sum_{l_1<l_2<l_3} n_{k_1k_2k_3/l_1l_2l_3} \log p_{k_1k_2k_3/l_1l_2l_3}. \quad (5)$$

If the heterozygous parent has a certain chromosomal-pairing configuration (Table 1), $p_{k_1k_2k_3/l_1l_2l_3}$ only contains the unknown recombination fraction r. By maximizing the likelihood (5), the maximum likelihood estimate (MLE) of r can be obtained by an explicit expression.

In practice, the chromosomal-pairing configuration of an allohexaploid is unknown. However, given that it presents a mix of 15 possible configurations (Table 1), for the same triploid gamete genotype, we can derive $p_{k_1k_2k_3/l_1l_2l_3}$ as a sum of its frequencies (determined by p_0, p_1, p_2, or p_3) weighted by chromosomal-pairing configuration frequencies f_1, …, f_{15}. With the derived $p_{k_1k_2k_3/l_1l_2l_3}$, the likelihood (5) was reformulated as a mixture model. The EM

algorithm (given in supplementary material 2) can be implemented to obtain the estimates of preferential pairing factors θ_1, θ_2 and θ_3 and recombination fraction r. It is shown that the estimation of r can be obtained from a closed form.

For a practical marker data, we do not know the linkage phase and chromosomal homology of an allohexaploid a priori. We can infer these two uncertainties from the marker genotype data to obtain the correct estimate of recombination fraction r. First, an allohexaploid may have 180 types of linkage phases over two markers each with six different alleles. Second, there are 15 possible homologuous relationships among six chromosomes, i.e.,

$$\begin{cases} 1) & \textbf{1} \text{ to } \textbf{2}, \textbf{3} \text{ to } \textbf{4}, \text{ and } \textbf{5} \text{ to } \textbf{6} \\ 2) & \textbf{1} \text{ to } \textbf{2}, \textbf{3} \text{ to } \textbf{5}, \text{ and } \textbf{4} \text{ to } \textbf{6} \\ 3) & \textbf{1} \text{ to } \textbf{2}, \textbf{3} \text{ to } \textbf{6}, \text{ and } \textbf{4} \text{ to } \textbf{5} \\ 4) & \textbf{1} \text{ to } \textbf{5}, \textbf{3} \text{ to } \textbf{4}, \text{ and } \textbf{2} \text{ to } \textbf{6} \\ 5) & \textbf{1} \text{ to } \textbf{6}, \textbf{3} \text{ to } \textbf{4}, \text{ and } \textbf{2} \text{ to } \textbf{5} \\ 6) & \textbf{1} \text{ to } \textbf{3}, \textbf{2} \text{ to } \textbf{4}, \text{ and } \textbf{5} \text{ to } \textbf{6} \\ 7) & \textbf{1} \text{ to } \textbf{4}, \textbf{2} \text{ to } \textbf{3}, \text{ and } \textbf{5} \text{ to } \textbf{6} \\ 8) & \textbf{1} \text{ to } \textbf{3}, \textbf{2} \text{ to } \textbf{5}, \text{ and } \textbf{4} \text{ to } \textbf{6} \\ 9) & \textbf{1} \text{ to } \textbf{3}, \textbf{2} \text{ to } \textbf{6}, \text{ and } \textbf{4} \text{ to } \textbf{5} \\ 10) & \textbf{1} \text{ to } \textbf{4}, \textbf{2} \text{ to } \textbf{5}, \text{ and } \textbf{3} \text{ to } \textbf{6} \\ 11) & \textbf{1} \text{ to } \textbf{4}, \textbf{2} \text{ to } \textbf{6}, \text{ and } \textbf{3} \text{ to } \textbf{5} \\ 12) & \textbf{1} \text{ to } \textbf{5}, \textbf{2} \text{ to } \textbf{3}, \text{ and } \textbf{4} \text{ to } \textbf{6} \\ 13) & \textbf{1} \text{ to } \textbf{5}, \textbf{2} \text{ to } \textbf{4}, \text{ and } \textbf{3} \text{ to } \textbf{6} \\ 14) & \textbf{1} \text{ to } \textbf{6}, \textbf{2} \text{ to } \textbf{3}, \text{ and } \textbf{4} \text{ to } \textbf{5} \\ 15) & \textbf{1} \text{ to } \textbf{6}, \textbf{2} \text{ to } \textbf{4}, \text{ and } \textbf{3} \text{ to } \textbf{5} \end{cases} \tag{6}$$

Thus, we need to choose an optimal combination of linkage phase and chromosomal homology from a total of $180 \times 15 = 2700$ possibilities. By obtaining the corresponding 2700 plug-in likelihood values, we select the largest one that corresponds to the optimal combination fromn which the MLEs of preferential pairing factors and recombination fraction can be obtained.

After the unknown parameters (θ_1, θ_2 and θ_3 and r) are estimated under an optimal linkage phase and homology, we formulate a series of hypotheses to test if each of these parameters is significant. These tests include those of whether the recombination fraction is different from

0.5, whether there is no preferential pairing between more similar chromosomes during meiosis, i.e. $\theta_1 = \theta_2 = \theta_3 = 0$, and whether one of the preferential pairing does not exist between a particular pair of more similar chromosomes during meiosis, i.e. $\theta_1 = 0$ or $\theta_2 = 0$ or $\theta_3 = 0$. To test these hypotheses, we need to calculate the log-likelihood ratios from the likelihoods under the null and alternative hypotheses and compare it against a critical value obtained from a chi-square distribution with three or one degree of freedom.

Linkage model for partially informative markers

In the preceding section, a procedure was described for linkage analysis of fully informative markers (with six distinct alleles at each marker) in allohexaploids, but a consideration should be taken for those partially informative markers which have multiple copies of the same alleles at one or two markers. For partially informative markers, a mixture likelihood constructed under a particular allelic configuration can be similarly constructed, but with a more complex structure due to their inconsistency between observed genotypes and real configurations. For example, a five-allele genotype observed as $a_1a_2a_3a_4a_5$, may have five possible configurations; i.e., $a_1|a_1|a_2|a_3|a_4|a_5|$, $a_1|a_2|a_2|a_3|a_4|a_5|$, $a_1|a_2|a_3|a_3|a_4|a_5|$, $a_1|a_2|a_3|a_4|a_4|a_5|$, and $a_1|a_2|a_3|a_4|a_5|a_5|$, but for a four-allele genotype observed as $b_1b_2b_3b_4$, it has as many as 10 possible configurations, such as $b_1|b_1|b_1|b_2|b_3|b_4|$, $b_1|b_1|b_2|b_2|b_3|b_4|$, $b_1|b_1|b_2|b_3|b_3|b_4|$, $b_1|b_1|b_2|b_3|b_4|b_4|$, $b_1|b_2|b_2|b_2|b_3|b_4|$, $b_1|b_2|b_2|b_3|b_3|b_4|$, $b_1|b_2|b_2|b_3|b_4|b_4|$, $b_1|b_2|b_3|b_3|b_3|b_4|$, $b_1|b_2|b_3|b_3|b_4|b_4|$, and $b_1|b_2|b_3|b_4|b_4|b_4|$. A three- or two-allele genotype has 10 and five different configurations, respectively. An extra difficulty for linkage analysis of partially informative markers lies in the estimation of the probability at which each allelic configuration occurs and then the determination of the most likely configuration.

For two fully informative markers, there are 15 distinguishable chromosomal homologies (6). But some of these homologies are collapsed into the same group for partially informative markers. Also, in such a case, the triplotypes of gametes are collapsed because of indistinguishable types of recombinants and non-recombinants. These two types of collapses together make it more difficult to estimate the preferential pairing factors and recombination fraction the EM algorithm. We have derived a general procedure for estimating these two parameters and testing their significance when two markers are partially informative. It should be noted that, for fully informative markers and five- and four-allele partially informative markers, all three preferential pairing factors (θ_1, θ_2 and θ_3) that determine chromosomal pairing types can be estimated, but because of reduced degrees of freedom,

only one and two preferential pairing factors can be estimated for two- and three-allele partially informative markers, respectively.

Results

Computer simulation

We performed Monte Carlo simulation to investigate the statistical properties of the allohexaploid linkage analysis model. The simulation experiments were designed to reflect ranges of the preferential pairing factor θ_1 = 0.15, 0.10 and 0.05 and recombination fraction r = 0.05, 0.15 and 0.30. The data of marker genotypes were simulated for two testcross markers in a full-family of size N = 100, 200, or 400 by assuming a particular linkage phase for the two markers and chromosomal homology.

Table 3 shows the results about parameter estimation from 1000 simulation replicates by the new model. It appears that a small sample size 100 can provide reasonably good estimates of the preferential pairing factor and recombination fraction, but the accuracy and precision of parameter estimates increase dramatically with increasing sample size. The linkage of two highly linked markers (r = 0.05) can be better estimated than that of two loosely linked markers (r = 0.30). The preferential pairing factors can be well estimated, not depending on the degree of linkage between two markers. The power to detect the correct linkage phase and homology is quite high. This is not surprising because the segregation of marker genotypes is very sensitive to the pattern of linkage phase and homology.

We performed an additional simulation study to examine the power of detecting the linkage and preferential pairing factors. The empirical power of the detection of these parameters was calculated by considering different sample sizes and different degrees of linkage (Table 2). In general, the power to detect the linkage is very high, which is not surprised because the MLE of the recombination fraction was based on an explicit expression. The power to jointly detect all possible preferential pairing is also very high, but reduced for the detection of individual preferential pairing. When the preferential pairing factor is low (say 0.05), the power to detect it becomes very low especially when sample size is modest (100). As shown in Table 2, if the preferential pairing factor is about 0.10, sample size of 200 – 400 is required to detect

its occurrence. When the preferential pairing factor is low, e.g., 0.05, sample size of over 400 should be used.

To investigate how the model performs for the linakge analysis of partially informative markers, a simulation study was carried out by hypothosizing two markers **A**: $a_1a_2a_3a_3a_4a_5$ and **B**: $b_1b_2b_2b_3b_4b_4$ of different strengths of linkage (r = 0.05, 0.15 and 0.30). The preferential pairing factors were assumed as θ_1=0.15 , θ_2=0.10 and θ_3=0.05. We assume a backcross design of different sample sizes n = 100, 200 and 400. Table 5 gives the estimates of the preferential pairing factors and recombination fraction under these scenarios. The model can provide reasonably accurate estimates of these parameters. As expected, the precision of parameter estimation increases with increasing sample size. The estimation of the preferential pairing factors is generally independent of the degree of linkage. For partially informative markers, the power to correctly detect both allelic configuration and chromosomal homology is about 0.5 with a modest sample size. However, we noted that power would increase to > 0.95 if the selected allelic configuration differs by one chromosome from the true configuration. Table 6 lists the power of linkage detection and preferential pairing detection under different simulation scenarios. In general, all the power is quite high even for a modest sample size (100), but to detect preferential pairing, a large sample size (300) is needed if the markers are loosely linked.

Worked example

Currently, we have a small real dataset to test the usefulness of our model. A full-sib family of persimmon was derived the cross between an allohexaploid tree (male) and a diploid tree (female) at Shandong Research Institute of Pomology, Taishang, China. The family contains 106 progeny, genotyped for several dozens of SSR markers screened from published EST-SSR primers. We analyzed four randomly chosen markers, whose mating types are detected, on the basis of Mendelian segregation law, as

$$a_1a_1a_1a_1a_1a_1 \times a_1a_2a_0a_0a_0a_0 \quad \text{for marker DKYQ200}$$
$$a_1a_2a_0a_0a_0a_0 \times a_1a_1a_1a_1a_1a_1 \quad \text{for marker DKYQ248}$$
$$a_1a_1a_1a_1a_1a_1 \times a_1a_2a_0a_0a_0a_0 \quad \text{for marker DKYQ252}$$
$$a_1a_2a_3a_0a_0a_0 \times a_3a_3a_3a_3a_3a_3 \quad \text{for marker DKYQ257}$$

where some alleles cannot be determined precisely as a_1 or a_2, which are denoted by a_0. Our model was equipped with a function to discern these alleles for the accurate estimation of the linkage and preferential pairing factor.

We performed pair-wise linkage analysis by first determining the most likely linkage phase and homology under which the recombination fraction and preferential pairing factors were estimated, with results given in Table 7. By a re-sampling approach, the standard errors of each estimate were obtained. The estimated recombination fractions between these two markers range from 0.056 to 0.134, showing mutual highly linked relationships. Because a few number of alleles at each marker, we can only estimate a couple of preferential pairing factors. It is interesting to see that preferential pairing does occur among chromosomes in the allohexaploid persimmon, which suggests that this hexaploid woody plant probably has experienced the combination of distinct genomes through interspecific hybridization. Different values of the preferential pairing factors estimated from different marker pairs may indicate varying degrees of relatedness among different regions of chromosomes.

Discussion

Current statistical models are mainly focused on linkage analysis for diploids (Lander and Green 1987; Stam 1993; Maliepaard et al. 1997; Wu et al. 2002). Many models for linkage analysis of polyploids are generally borrowed from diploids, which may produce misleading results because polyploids undergo qualitatively different meiotic mechanisms from diploids. For example, the frequencies of gamete formation are not only influenced by the recombination fraction, but also influenced by the relative frequencies of different chromosome pairing mechanisms, such as preferential pairing that has a widespread occurrence in allopolyploids (Sybenga 1965, 1966, 1988) and double reduction in autopolyploids (Luo et al. 2004; Wu and Ma 2005). The preferential pairing factor is an important parameter that describes the cytological characteristic of allopolyploids thought to play a key role in plant evolution. Sybenga (1992, 1998) used the preferential pairing factor to describe the homology in allopolyploids. Wu et al. (2001a) proposed that the preferential pairing factor could explain the difference between pairing formation derived from bivalent and multivalent pairings. Here, we describe and assess a model for allohexaploid linkage analysis incorporating the preferential pairing factor. Our model built upon preferential

pairing and chromosomal homology shows good power to obtain more realistic results than existing linkage analysis models.

The statistical model proposed can simultaneously estimate the recombination fraction and preferential pairing factor in allohexaploids. Our model can handle the linkage of any types of markers, such as testcross markers (at which only one of the two parents is heterozygous) and intercross markers (at which both parents are heterozygous), segregating in a full-sib family of two heterozygous parents. Simulation studies were performed to investigate the statistical behavior of our model. It was found that the model displays high precision for estimating the recombination fraction and preferential pairing factor over a range of sample sizes and parameter values. By estimating the preferential pairing factor, the model helps geneticist to determine whether a particular allohexaploid undergoes preferential pairing or random pairing during meiosis and relate this information to understand the evolutionary diversity of polyploids (Nicolas et al. 2007; Lim et al. 2008; Gaeta and Pires 2010; Xiong et al. 2011).

Existing linkage analysis models for polyploids are mainly focused on triploids and tetraploids, with hexaploids being never touched before. Our model derived here fills a gap in this area. We provided a general framework for allohexaploid linkage analysis and its principle can be extended to octoploid and dexaploid species, but their increasing complexity of model derivation deserves an independent study. Meanwhile, the model focuses on the marker segregation and recombination in a full-sib family, but its principle can also be extended to consider an open-pollinated natural population used to study the genetic structure and evolution of natural populations (Sun et al. 2015).

Our model is based on two-point linkage analysis of fully informative markers. There is still a plenty of room to modify and comprehend the model. First, we assume that an allohexaploid only undergoes a bivalent pairing, but this assumption may be too strong in some situations in which both bivalent and multivalent pairing may occur at the same time. Wu et al. (2004) proposed a mixture model that allows these two types of pairing to be separated in tetraploids from the EM algorithm. More complex EM algorithms should be derived to accommodate to this information in hexaploids. Second, it deserves being extended into three-point linkage analysis because this can not only estimate the combination fraction between two loci, but also examine the influence of genetic interference on the linkage estimation (Wu et al., 2002;

Lu et al., 2004; Hou et al., 2009; Liu et al., 2012). Third, subsequent work is needed for QTL mapping to understand the genetic architecture of quantitatively inherited traits in allohexaploids. Despite these extensions being made, our current model provides a general platform to study the linkage and homology of allohexaploids. We have packed the model into computer software at http://ccb.bjfu.edu.cn/program.html (available upon the acceptance of this manuscript) which can be freely used by other researchers.

Acknowledgements

We thank Xiaoming Pang for his contribution to this work and Shandong Research Institute of Pomology for supplying their parsimony data to validate our model. This work is supported by Special Fund for Forest Scientific Research in the Public Welfare (201404102), Changjiang Scholars Award and “Thousand-person Plan” Award.

Authors’ contributions: H.L. wrote the manuscript. X.Z. and Q.Y. derived the model and performed data analysis. Q.Y. and K.M. conducted marker experiment. R.W. conceived of the idea and wrote the manuscript. All authors read and approved the final manuscript.

References

Bourke, P.M., Voorrips, R.E., Visser, R.G.F. and Maliepaard, C. (2015) The double-reduction lanscape in tetraploid potatos as revealed by a high-density linkage map. Genetics 210: 853-863.

Chen, Z.J. (2010) Molecular mechanisms of polyploidy and hybrid vigor. Trends Plant Sci. 15: 57-71.

Comai, L. (2005) The advantages and disadvantages of being polyploid. Nat. Rev. Genet. 6: 836-846.

Gaeta, R.T. and Chris-Pires J. (2010) Homoeologous recombination in allopolyploids: the polyploid ratchet. New Phytol. 186: 18-28.

Grant, V. (2004) Plant speciation, the book: perspectives and paradigms. New Phytol. 161: 8-11.

Hauber, D.P., Reeves, A., and Stack, S.M. (1999) Synapsis in a natural autotetraploid. Genome 42: 936-949.

Hackett, C.A, Bradshaw, J.E., Meyer, R.C., McNicol, J.W. and Waugh, R. (1998) Linkage analysis in tetraploid species: a simulation study. Genet. Res. 71: 143-153.

Hackett, C.A., McLean, K., and Bryan, G.J. (2013) Linkage analysis and QTL mapping Using SNP dosage data in a tetraploid potato mapping population. PLoS ONE 8(5): e63939.

Hackett, C.A., Bradshaw, J.E. and Bryan, G.J. (2014) QTL mapping in autotetraploids using SNP dosage information. Theor. Appl. Genet. 127: 1885-1904.

Hou, W., Liu, T., Li, Y., Li, Q., Li, J.H., Das, K., Berg, A. and Wu, R.L. (2009) Multilocus genomics of outcrossing plant populations. Theor. Popul. Biol. 76: 68-76.

Kriegner, A., Cervantes, J.C., Burg, K., Mwanga, R.O.M. and Zhang, D. (2003) A genetic linkage map of sweet potato [*Ipomoea batatas* (L.) Lam.] based on AFLP markers. Mol. Breed. 11: 169–185.

Lander, E.S. and Geen, P. (1987) Construction of multilocus genetic linkage maps in humans. Proc. Natl. Acad. Sci. U S A 84: 2363-2367.

Leitch, A.R. and Leitch, I.J. (2008) Perspective - Genomic plasticity and the diversity of polyploid plants, Science 320: 481-483.

Lewis, W.H. (1979) Polyploidy in angiosperms: dicotyledons. Basic Life Sci. 13: 241-268.

Li, J., Das, K., Fu, G.F, Tong, C.F., Li, Y., Tobias, C. and Wu R. (2010) EM algorithm for mapping quantitative trait Loci in multivalent tetraploids. Intl. J. Plant Genom. 2010, 216547.

Lim, K.Y., Soltis, D.E., Soltis, P.S., Tate, J., Matyasek, R., Srubarova, H., Kovarik, A., Pires, J,C., Xiong, Z. and Leitch, A.R (2008) Rapid chromosome evolution in recently formed polyploids in Tragopogon (Asteraceae). PLoS ONE 3(10): e3353.

Liu, J., Wang, Z., Wang, Y.Q., Li, R. and Wu, R. (2012) Model and algorithm for linkage disequilibrium analysis in a non-equilibrium population. Front. Genet. 3, 78.

Lu, Q., Cui, Y.H. and Wu, R. (2004) A multilocus likelihood approach to joint modeling of linkage, parental diplotype and gene order in a full-sib family. BMC Genet 5.

Lu, Y.F., Yang, X.X., Tong, C.F., Li, X., Feng, S.S., Wang, Z., Pang, X.M,, Wang, Y.Q., Wang. N.T., Tobias, C.M. and Wu, R.L. **(**2013) A multivalent three-point linkage analysis model of autotetraploids. Brief Bioinform 14: 460-468.

Luo, Z.W., Hackett, C.A., Bradshaw, J.E., McNicol J.W. and Mibourne, D. (2001a) Construction of a genetic linkage map in tetraploid species using molecular markers. Genetics 157: 1369-1385.

Luo, Z.W., Zhang, R.M. and Kearsey, M.J. (2001b) Theoretical basis for genetic linkage analysis in autotetraploid species. Proc. Natl. Acad. Sci. U S A 101:7040-7045.

Maliepaard, C., Jansen, J. and Van Ooijen, J.W. (1997) Linkage analysis in a full-sib family of an outbreeding plant species: overview and consequences for applications. Genet. Res. 70: 237-250.

Masterson, J. (1994) Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. Science 264: 421-424.

McCord, P.H., Sosinski, B.R., Haynes, K.G., Clough, M.E. and Yencho GC (2011) QTL mapping of internal heat necrosis in tetraploid potato. Theor. Appl. Genet. 122, 129–142.

Monden, Y., Hara, T., Okada, Y., Jahana, O., Kobayashi, A., Tabuchi, H., Onaga, S., Tahara M. (2015) Construction of a linkage map based on retrotransposon insertion polymorphisms in sweetpotato via high-throughput sequencing. Breed. Sci. 65: 145-153.

Müntzing, A. (1936) The evolutionary significance of autopolyploidy. Hereditas 21: 363-378.

Nicolas, S.D., Le Mignon, G., Eber, F., Coriton, O., Monod, H., Clouet, V., Huteau, V., Lostanlen, A., Delourme, R., Chalhoub, B., Ryder, C.D., Chèvre, A.M. and Jenczewski, E. (2007) Homeologous recombination plays a major role in chromosome rearrangements that occur during meiosis of Brassica napus haploids. Genetics 175: 487-503.

Otto, S.P. and Whitton, J. (2000) Polyploid incidence and evolution, Annu. Rev. Genet. 34: 401-437.

Pecinka, A., Fang, W., Rehmsmeier, M., Levy, A.A. and Mittelsten Scheid, O. (2011) Polyploidization increases meiotic recombination frequency in Arabidopsis. BMC Biol. 9: 24.

Ripol, M.I., Churchill, G.A., da Silva, J.A. and Sorrells, M. (1999) Statistical aspects of genetic mapping in autopolyploids, Gene 235: 31-41.

Soltis, D.E. and Soltis, P.S. (1999) Polyploidy: recurrent formation and genome evolution. Trends Ecol. Evol. 14: 348-352.

Soltis, D.E. and Soltis, P.S. (2000) The role of genetic and genomic attributes in the success of polyploids, Proc. Natl. Acad. Sci. U S A 97: 7051-7057.

Soltis, D.E. and Soltis, P.S. (2009) The role of hybridization in plant speciation. Annu. Rev. Plant. Biol. 60: 561-588.

Soltis, D.E., Soltis, P.S. and Tate, J.A. (2004) Advances in the study of polyploidy since plant speciation. New Phytol. 161: 173-191.

Stam, P. (1993) Construction of integrated genetic linkage maps by means of a new computer package: Join Map. Plant J. 3: 739-744.

Sun, L.D., Zhu, X.L., Zhang, Q.X. and Wu, R.L. (2015) A unifying experimental design for dissecting tree genomes. Trends Plant Sci 20: 473–476.

Sybenga, J. (1965) The quantitative analysis of chromosome pairing and chiasma formation based on the relative frequencies of M I configurations. II. Primary trisomics. Genetica 36: 339-350.

Sybenga, J. (1966) The quantitative analysis of chromosome pairing and chiasma formation based on the relative frequencies of MI configurations. V. Interchange trisomics. Genetica 37: 481-510.

Sybenga, J. (1988) Mathematical models for estimating preferential pairing and recombination in triploid hybrids. Genome 30: 745-757.

Sybenga, J. (1992) Cytogenetics in plant breeding. Springer-Verlag, New York.

Wu, R., Gallo-Meagher, M., Littell, R.C. and Zeng, Z.-B. (2001a) A general polyploid model for analyzing gene segregation in outcrossing tetraploid species. Genetics 159: 869-882.

Wu, R. and Ma, C.X. (2005) A general framework for statistical linkage analysis in multivalent tetraploids. Genetics 170: 899-907.

Wu, R., Ma, C.X. and Casella, G. (2002) A bivalent polyploid model for linkage analysis in outcrossing tetraploids. Theor. Popul. Biol. 62: 129-151.

Wu, R., Ma, C.X. and Casella, G. (2004) A mixed polyploid model for linkage analysis in tetraploids. J. Comp. Biol. 11: 562–580.

Wu, R., Ma, C.X., Painter, I. and Zeng, Z.B. (2002) Simultaneous maximum likelihood estimation of linkage and linkage phases in outcrossing species. Theor. Popul. Biol. 61: 349-363.

Wu, S.S., Wu, R., Ma, C.X. Zeng, Z.-B., Yang, M. and Casella, G. (2001b) A multivalent pairing model of linkage analysis in autotetraploids. Genetics 159: 1339-1350.

Xiong, Z., Gaeta, R.T. and Pires, J.C. (2011) Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid *Brassica napus*. Proc. Natl. Acad. Sci. U S A 108: 7908-7913.

Xu, F., Lyu, Y.F., Tong, C.F., Wang, W.M., Zhu, X.L., Yin, D.L., Zhang, J., Pang, X.M., Tobias, C.M. and Wu, R.L. **(**2014a) A statistical model for QTL mapping in polysomic autotetraploids underlying double reduction. Brief Bioinform 15: 1044-1056.

Xu, F., Tong, C.F., Lyu, Y.F., Bo, W.H., Pang, X.M. and Wu, R.L. (2014b) Allotetraploid and autotetraploid models of linkage analysis. Brief Bioinform 16: 32-38.

Yang, X., Lv, Y., Pang, X., Tong, C.F., Wang, Z., Li, X., Feng, S., Tobias C.M. and Wu, R.L. (2013) A unifying framework for bivalent multilocus linkage analysis of allotetraploids. Brief. Bioinform. 14: 96-108.

Table 1 Types and frequencies of chromosomal-pairing configurations, as well as triploid gametes each configuration produces during meiosis, for an allohexaploid with six chromosomes labeled as **1**, **2**, **3**, **4**, **5**, and **6**.

No.	Pairing	Configuration	Freq.	Gamete
1	Fully preferential	**12** || **34** || **56**	f_1	**135 136 145 146 235 236 245 246**
2	Partially preferential	**12** || **35** || **46**	f_2	**134 136 145 156 234 236 245 256**
3	Partially preferential	**12** || **36** || **45**	f_3	**134 135 146 156 234 235 246 256**
4	Partially preferential	**15** || **34** || **26**	f_4	**123 124 136 146 235 245 356 456**
5	Partially preferential	**16** || **34** || **25**	f_5	**123 124 135 145 236 246 356 456**
6	Partially preferential	**13** || **24** || **56**	f_6	**125 126 145 146 235 236 345 346**
7	Partially preferential	**14** || **23** || **56**	f_7	**125 126 135 136 245 246 345 346**
8	No preferential	**13** || **25** || **46**	f_8	**124 126 145 156 234 236 345 356**
9	No preferential	**13** || **26** || **45**	f_9	**124 125 146 156 234 235 346 356**
10	No preferential	**14** || **25** || **36**	f_{10}	**123 126 135 156 234 245 345 456**
11	No preferential	**14** || **26** || **35**	f_{11}	**123 125 136 156 234 245 346 456**
12	No preferential	**15** || **23** || **46**	f_{12}	**124 126 134 136 245 256 345 356**
13	No preferential	**15** || **24** || **36**	f_{13}	**124 126 134 136 245 256 345 356**
14	No preferential	**16** || **23** || **45**	f_{14}	**124 125 134 135 246 256 346 356**
15	No preferential	**16** || **24** || **35**	f_{15}	**123 125 134 145 236 256 346 456**

Table 2 The probabilities of all possible triploid gametes for markers **A** and **B** produced by an phase-known allohexaploid, diagrammed as $\begin{matrix} a_1 & a_2 \\ b_1 & b_2 \end{matrix} \| \begin{matrix} a_3 & a_4 \\ b_3 & b_4 \end{matrix} \| \begin{matrix} a_5 & a_6 \\ b_5 & b_6 \end{matrix}$, under fully preferentially pairing.

	Marker A							
Marker B	$a_1\|a_3\|a_5\|$	$a_1\|a_3\|a_6\|$	$a_1\|a_4\|a_5\|$	$a_1\|a_4\|a_6\|$	$a_2\|a_3\|a_5\|$	$a_2\|a_3\|a_6\|$	$a_2\|a_4\|a_5\|$	$a_2\|a_4\|a_6\|$
$b_1\|b_3\|b_5\|$	p_0	p_1	p_1	p_2	p_1	p_2	p_2	p_3
$b_1\|b_3\|b_6\|$	p_1	p_0	p_2	p_1	p_2	p_1	p_3	p_2
$b_1\|b_4\|b_5\|$	p_1	p_2	p_0	p_1	p_2	p_3	p_1	p_2
$b_1\|b_4\|b_6\|$	p_2	p_1	p_1	p_0	p_3	p_2	p_2	p_1
$b_2\|b_3\|b_5\|$	p_1	p_2	p_2	p_3	p_0	p_1	p_1	p_2
$b_2\|b_3\|b_6\|$	p_2	p_1	p_3	p_2	p_1	p_0	p_2	p_1
$b_2\|b_4\|b_5\|$	p_2	p_3	p_1	p_2	p_1	p_2	p_0	p_1
$b_2\|b_4\|b_6\|$	p_3	p_2	p_2	p_1	p_2	p_1	p_1	p_0

The definition of p_0, p_1, p_2 and p_3 are given in the text.

Table 3 MLEs of preferential pairing factors and recombination fraction and their standard deviations over 1000 simulation replicates estimated from a simulated marker data generated by a phase- and homology-known allohexaploid under different simulation scenarios by changing the recombination fraction and sample size. The power of detecting correct linkage phase and correct homology at the same time is also given.

Simulation Scenario	θ_1 (0.15)	θ_2 (0.10)	θ_3 (0.05)	r	Power
r = 0.05					
100	0.146±0.093	0.103±0.093	0.061±0.080	0.050±0.012	1
200	0.148±0.069	0.100±0.065	0.058±0.061	0.051±0.009	1
400	0.150±0.051	0.103±0.047	0.052±0.044	0.050±0.006	1
r = 0.15					
100	0.146±0.093	0.103±0.093	0.061±0.080	0.151±0.020	1
200	0.148±0.069	0.100±0.065	0.058±0.061	0.150±0.014	1
400	0.150±0.051	0.103±0.047	0.052±0.044	0.150±0.010	1
r = 0.30					
100	0.146±0.093	0.103±0.093	0.061±0.080	0.301±0.029	0.997
200	0.148±0.069	0.100±0.065	0.058±0.061	0.300±0.018	1
400	0.150±0.051	0.103±0.047	0.052±0.044	0.301±0.014	1

Table 4 Empirical power to detect the linakge and three preferential pairing factors under different simulation scenarios by changing the recombination fraction and sample size.

Simulation Scenario	r	Power			
		$(\theta_1, \theta_2, \theta_3)$	θ_1	θ_2	θ_3
r = 0.05					
100	1	0.978	0.800	0.628	0.379
200	1	1	0.955	0.824	0.494
400	1	1	1	0.977	0.694
r = 0.15					
100	1	0.988	0.765	0.552	0.320
200	1	1	0.965	0.834	0.462
400	1	1	0.999	0.985	0.699
r = 0.30					
100	1	0.996	0.772	0.572	0.273
200	1	1	0.964	0.834	0.426
400	1	1	1	0.990	0.705

Table 5 MLE of preferential pairing factors and recombination fraction and their standard deviation over 1000 simulation replicates from a simulated marker data. The power of detecting correct linkage phase and correct homology at the same time is also given.

Number	θ_1 (0.15)	θ_2 (0.10)	θ_3 (0.05)	r	Power
r = 0.05					
100	0.150±0.098	0.098±0.086	0.070±0.096	0.066±0.052	0.447
200	0.149±0.073	0.103±0.067	0.058±0.065	0.057±0.033	0.536
400	0.146±0.052	0.100±0.049	0.057±0.045	0.056±0.020	0.429
r = 0.15					
100	0.153±0.102	0.091±0.094	0.066±0.082	0.171±0.060	0.513
200	0.150±0.073	0.103±0.067	0.058±0.065	0.156±0.044	0.442
400	0.149±0.054	0.095±0.047	0.059±0.044	0.158±0.028	0.446
r = 0.30					
100	0.166±0.098	0.088±0.096	0.064±0.085	0.301±0.065	0.395
200	0.150±0.078	0.100±0.067	0.056±0.060	0.303±0.048	0.523
400	0.144±0.050	0.105±0.051	0.054±0.045	0.296±0.031	0.476

Table 6 Power to detect whether the estimated parameters are significant under different hypotheses.

Number	Power				
	r=0.5	$\theta_1=\theta_2=\theta_3=0$	$\theta_1=0$	$\theta_2=0$	$\theta_3=0$
r = 0.05					
100	1	0.992	0.957	0.886	0.742
200	1	1	0.989	0.972	0.745
400	1	1	1	0.997	0.926
r = 0.15					
100	1	0.981	0.902	0.876	0.723
200	1	1	0.993	0.986	0.842
400	1	1	1	0.992	0.913
r = 0.30					
100	0.925	0.981	0.952	0.883	0.779
200	1	1	0.983	0.975	0.732
400	1	1	1	0.983	0.862

Table 7 Estimates of the recombination fraction and preferential pairing factors among four SSR markers genotyped for a full-sib family of persimmon.

Marker Pair	r	θ_1	θ_2
DKYQ200×DKYQ248	0.134 (±0.096)		
DKYQ200×DKYQ252	0.093 (±0.061)		
DKYQ200×DKYQ257	0.081 (±0.067)		
DKYQ248×DKYQ252	0.112 (±0.079)		
DKYQ248×DKYQ257	0.073 (±0.054)		
DKYQ252×DKYQ257	0.056 (±0.032)		
DKYQ200		0.046 (±0.031)	-
DKYQ248		0.139 (±0.072)	-
DKYQ252		-0.069 (±0.054)	-
DKYQ257		0.078 (±0.057)	0.062 (±0.045)

Note: only one or two preferential pairing factors can be estimated for two- or three-allele partially informative markers