Linkage Analysis and Map Construction in Genetic Populations of Clonal F₁ and Double Cross

Luyan Zhang, Huihui Li, and Jiankang Wang
The National Key Facility for Crop Gene Resources and Genetic Improvement, Institute of Crop Science, and CIMMYT China Office, Chinese Academy of Agricultural Sciences, Beijing 100081, China

ABSTRACT In this study, we considered four categories of molecular markers based on the number of distinguishable alleles at the marker locus and the number of distinguishable genotypes in clonal F₁ progenies. For two marker loci, there are nine scenarios that allow the estimation of female, male, and/or combined recombination frequencies. In a double cross population derived from four inbred lines, five categories of markers are classified and another five scenarios are present for recombination frequency estimation. Theoretical frequencies of identifiable genotypes were given for each scenario, from which the maximum likelihood estimates of one or more of the three recombination frequencies could be estimated. If there was no analytic solution, then Newton-Raphson method was used to acquire a numerical solution. We then proposed to use an algorithm in Traveling Salesman Problem to determine the marker order. Finally, we proposed a procedure to build the two haploids of the female parent and the two haploids of the male parent in clonal F₁. Once the four haploids were built, clonal F₁ hybrids could be exactly regarded as a double cross population. Extensive comparisons with software JoinMap4.1, One-Map, and R/qtl show that the methodology proposed in this article can build more accurate linkage maps in less time.

KEYWORDS clonal F₁ double cross recombination frequency linkage analysis map construction

Plant species can be divided into three groups with respect to their sexual mating and asexual reproductive systems, i.e., self-pollination, cross-pollination, and vegetative (or clonal or asexual) propagation (Allard 1999). An asexually propagated population consists of clones that are genetically identical to that of their parents. Reproduction by asexual propagation is common in higher plants, including nearly all fruit and nut trees such as strawberries, grapes, and pineapples; some field crops such as potatoes, sugarcane, yams, cassavas, and sweet potatoes; and many ornamental species (Allard 1999). Individual clonal plants usually show high heterozygosity. Once the superiority of any heterozygous clone is identified, this superiority can be protected and utilized by continued vegetative reproduction for a long period of time (Allard 1999).

Most clonal species have the problem of inbreeding depression, but hybridization between different clones, or even self-pollination of one clonal line, can produce seeds and therefore generate segregating clonal F₁ progenies. Many genetic linkage studies have been conducted in clonal species, such as potatoes (Tanksley et al. 1992; van Os et al. 2006), cassavas (Fregene et al. 1997; Kunkeaw et al. 2010), sweet potatoes (Li et al. 2010), sugarcane (Liu et al. 2010), populus (Zhang et al. 2000), pears (Yamamoto et al. 2002), apples (Hemmat et al. 1994), and pineapples (Carlier et al. 2004). Most studies focused on linkage map construction by adapting the clonal F₁ progenies into inbred line–derived populations, such as pseudo-backcrosses or pseudo-testcrosses. This is a tedious procedure, and many less informative markers may not be used. For example, Hemmat et al. (1994) only considered three groups of markers in linkage map construction: those segregating as a result of heterozygosity in the female or male parent or in both parents. Many markers were discarded in estimation of recombination frequency before linkage map construction. Some studies on clonal species used the CP model (cross pollinators) in the software JoinMap (Stam 1993; van Ooijen 2006), which translates the clonal F₁ progenies into a pseudo-backcross or pseudo-testcross...
population to estimate the recombination frequency in female and male parents.

Ritter et al. (1990) proposed a method of recombination frequency estimation between heterozygous parents based on RFLP markers, using part of the informative markers in the clonal \(F_1\) progenies. Ritter and Salamini (1996) considered more allelic configurations as an improvement of the previous work. Malepeaard et al. (1997) presented an overview of marker pair segregation configurations and then acquired the maximum likelihood estimators for the recombination frequency. Based on 18 cross types and the assumption that both parents had the same meiotic recombination, Wu et al. (2002a) proposed a methodology for linkage analysis in outcrossing species. Pairwise recombination frequency and linkage phase were estimated simultaneously by the posterior probabilities of the four different assignments conditional on the observed phenotype of the markers. Wu et al. (2002b) used the same algorithm in another study (Wu et al. 2002a), but considering the sex-specific recombination frequencies. Algorithms proposed by Wu et al. (2002a, b) were implemented in the R software (www.r-project.org) as a package called OneMap (Margarido et al. 2007). However, EM algorithm and Markov chains used in recombination frequency estimation and linkage phase determination were time-consuming. In addition, some configurations in the previous studies (Ritter and Salamini 1996; Malepeaard et al. 1997; Wu et al. 2002a, b) were identical in recombination frequency estimation. For example, Wu et al. (2002) gave 18 cross combinations based on the genotypes of the two parents. The first four each generates four genotypes, which can be properly identified in the progenies. They are identical when used in linkage analysis. Redundant configurations complicate the application of those methods in practical populations.

The R/qtl package could be used for linkage analysis in phase-known double cross (Broman et al. 2003), but it was not suitable for clonal \(F_1\) and phase-unknown double cross. It has been noted that software packages in R software were computationally slow and always failed to construct dense maps (van Ooijen 2011). Based on five segregation types of markers, van Ooijen (2011) proposed a Monte Carlo multipoint maximum likelihood algorithm to simultaneously estimate recombination frequency and determine marker order. An integrated map was generated by averaging lengths over anchored segments from two separate parental maps and by interpolating or extrapolating for markers segregating in only one parent. The methodology in van Ooijen (2011) was implemented in JoinMap4.1. The ordering algorithm used in JoinMap4.1 was called simulated annealing, which determines the best marker order by minimizing the sum of recombination frequencies in adjacent segments.

Genetic analysis methodology of clonal species is less investigated compared with self-pollinated and cross-pollinated species. In self-pollinated and cross-pollinated species, double crosses (or four-way crosses) can be made from four inbred lines to extend the genetic diversity in genetic studies and plant breeding. In clonal \(F_1\) and double cross, the number of alleles at each locus may be up to four. For each marker pair, there are four potential linkage phases in clonal \(F_1\). Once the linkage phase is determined, one clonal \(F_1\) can be viewed as a double cross population.

The unknown linkage phase and multiple alleles complicate recombination frequency estimation in clonal \(F_1\) and double cross populations. Our objectives in this study were: (1) to identify and classify informative markers based on the number of distinguishable alleles and the number of distinguishable genotypes; (2) to derive the theoretical frequencies of identifiable genotypes for each scenario of marker pairs and maximum likelihood estimates of recombination frequencies; (3) to build the female, male, and combined linkage maps; (4) to build the four haploids of the female and male parents based on the estimated recombination frequencies and the combined linkage map; and (5) to demonstrate the advantage of the proposed methods in comparison with other software.

### MATERIALS AND METHODS

#### Marker categories and coding criteria in clonal \(F_1\) progenies

Genetic studies in clonal species are normally conducted in \(F_1\) hybrids of two clonal parents, one used as female and the other as male (Figure 1). The two parents are normally heterozygous and unrelated or less related in genetics, and therefore may have up to four identifiable alleles at each polymorphism locus. In this study, \(A, B\) were used to represent the two potential alleles in the female parent; \(C\) and \(D\) represented the two potential alleles in the male parent, as indicated at two loci in Figure 1. Based on the actual number of identifiable alleles in the two parents and the actual number of identifiable genotypes in the \(F_1\) progenies, each marker locus can be classified into four categories (Figure 2).

- **Category I** (or ABCD) represents the case of fully informative markers. By fully informative, we mean the four genotypes at one locus in progenies can be clearly identified. In other words, the two alleles in any clonal progeny can be traced back to its female and male parents (Figure 2). For category I markers, two alleles can be identified in either parent. The four genotypes in progenies are coded as \(AC, AD, BC, BD\) (Figure 2). When no distortion occurs, the four genotypes will follow the Mendelian ratio of 1:1:1:1. However, it is possible that one female allele is the same as one male allele. For example, when allele \(A\) is equal to allele \(C\) at a marker locus, there is no problem assigning the two alleles in a progeny to the two parents. This marker is still classified as category I.

- **Category II** (or \(A = B\)) represents the case of male polymorphism markers. By male polymorphism markers, we mean they show no polymorphism in the female parent, but they show polymorphism in the male parent. For category II markers, only two genotypes can be observed in the clonal \(F_1\) progenies (Figure 2). Genotypes \(AC\) and \(BC\) cannot be separated; neither can genotypes \(AD\) and \(BD\). In this category, \(X\) is used to code genotypes \(AC\) and \(BC\); \(XD\) is used to code genotypes \(AD\) and \(BD\), where \(X\) stands for either allele \(A\) or allele \(B\) (Figure 2). When no distortion occurs, the two genotypes will follow the Mendelian ratio of 1:1.

- **Category III** (or \(C = D\)) represents the case of female polymorphism markers. By female polymorphism markers, we mean they
show polymorphism in the female parent, but they show no polymorphism in the male parent. For category III markers, only two genotypes can be observed in the clonal F1 progeny (Figure 2). When no distortion occurs, the two genotypes will follow the Mendelian ratio of 1:1.

Category IV (or AB = CD) represents the case of co-dominant markers. By co-dominant markers, we mean they show the same two heterozygous genotype. The two alleles in parents are represented by A and B, and three genotypes in their progenies are represented by AA, AB, and BB.

Nine scenarios between two loci in recombination frequency estimation in clonal F1 progenies

Assuming that locus 1 and locus 2 are two linked polymorphism markers, falling into one of the four categories in Figure 2, let A1, B1, C1, and D1 be the four alleles at locus 1 and let A2, B2, C2, and D2 be the four alleles at locus 2. Recombination frequencies in the female and male parents were denoted as rF and rM, which can be used to construct the female and male linkage maps, respectively. The combined recombination frequency is denoted as r, which can be used to construct the combined map. Due to the symmetry of marker pairs, we consider nine scenarios between two loci in clonal F1 populations where at least one of the above three recombination frequencies can be estimated (Table 1). Scenario 1 represents the most ideal situation where all recombination frequencies can be properly estimated. If one locus is category II and the other one is category III (not included in Table 1), then the four genotypes at the two loci have an equal theoretical frequency of 0.25. In this scenario, none of rF, rM, and r can be estimated.

When one locus is category II, there is no polymorphism in the female parent; therefore, rF cannot be estimated (Table 1). Similarly, when one locus is category III, there is no polymorphism in the male parent; therefore, rM cannot be estimated (Table 1). In scenario 4, only half of samples can be used to estimate rF and rM (Table 1). In scenario 9, the linkage information in the two parents is confounded. It is impossible to estimate rF and rM. However, the combined r can still be estimated (Table 1).

### Table 1: The nine scenarios between two linked loci in the clonal F1 population for estimating the recombination frequency

| Scenario | Marker Category | Locus 1 | Locus 2 | rf | rm | r |
|----------|----------------|--------|--------|----|----|----|
| 1        | I (ABCD)       | I (ABCD) | ✓      | ✓  | ✓  |
| 2        | I (ABCD)      | II (A = B) | ✓     |     |     |
| 3        | I (ABCD)      | III (C = D) | ✓     |     |     |
| 4        | I (ABCD)      | IV (AB = CD) | 1/2✓  | 1/2✓ | ✓  |
| 5        | II (A = B)   | II (A = B) | ✓      | ✓  | ✓  |
| 6        | II (A = B)   | IV (AB = CD) | ✓     |     |     |
| 7        | III (C = D)  | III (C = D) | ✓      | ✓  | ✓  |
| 8        | III (C = D)  | IV (AB = CD) | ✓      |     |     |
| 9        | IV (AB = CD) | IV (AB = CD) | ✓     |     |     |

The symbol ✓ is used to indicate that recombination frequency rf, rm, or r could be estimated, and 1/2 is used to indicate that only half of the observed samples are used in estimating recombination frequency. When one marker is category II and the other one marker is category III, recombination frequency between them cannot be estimated and therefore it is not included.
lower than 0.5 if the two loci are linked. In the case of genotype $A_1B_2/B_1A_2$, gametes $A_1A_2$ and $B_1B_2$ are the two crossover types with a frequency of $(1-r_F^2)$, and $A_1B_2$ and $B_1A_2$ are the two non-crossover types with a frequency of $r_F$. The estimated $r_F$ will be more than 0.5 when the two loci are linked. Obviously, whether the estimated $r_F$ is less or more than 0.5 could help to determine the linkage phase and genotype of the female parent. Similarly, whether the estimated $r_M$ is less or more than 0.5 could help to determine the linkage phase and genotype of the male parent.

Therefore, linkage phases and genotypes of both parents can be determined by their estimated recombination frequencies, respectively. If estimated $r_F$ is less than 0.5, then the female parent will be in linkage phase $A_1A_2/B_1B_2$; otherwise, it will be in linkage phase $A_1B_2/B_2A_2$. If estimated $r_M$ is less than 0.5, then the male parent will be in linkage phase $C_1C_2/D_1D_2$; otherwise, it will be in linkage phase $C_1D_2/D_1C_2$.

Considering the two phases to be determined in both parents together, four potential linkage phases of the two parents can be defined. In phase I, the female parent has genotype $A_1A_2/B_1B_2$ and the male parent has genotype $C_1C_2/D_1D_2$. In phase II, the female parent has genotype $A_1A_2/B_1B_2$ and the male parent has genotype $C_1D_2/D_2C_2$. In phase III, the female parent has genotype $A_1B_2/B_2A_2$ and the male parent has genotype $C_1C_2/D_1D_2$. In phase IV, the female parent has genotype $A_1B_2/B_2A_2$ and the male parent has genotype $C_1D_2/D_2C_2$. The four phases will be used later for some scenarios in estimating the combined recombination frequency $r$, to make sure the estimated $r$ is less than 0.5, and the estimation will not be affected by the linkage information confounding in one or both parents.

Recombination frequency estimation in scenario 1 in clonal F₁ progenies

We begin with the most ideal situation where locus 1 has four identifiable genotypes $A_1C_1$, $A_1D_1$, $B_1C_1$, and $B_1D_1$, and locus 2 has four identifiable genotypes $A_2C_2$, $A_2D_2$, $B_2C_2$, and $B_2D_2$. The first row and first column of Table S1 show the four female and male gametes and their frequencies, from which we can easily derive theoretical frequencies of the 16 identifiable genotypes at the two linked loci.

| Genotype | Locus 1 | Locus 2 | Frequency | Sample Size |
|----------|---------|---------|-----------|-------------|
| $A_1C_1$ | $A_2C_2$ | $\frac{1}{4}(1-r_F)(1-r_M)$ | $n_1$ |
| $A_1C_1$ | $A_2D_2$ | $\frac{1}{4}(1-r_F)r_M$ | $n_2$ |
| $A_1D_1$ | $A_2C_2$ | $\frac{1}{4}(1-r_F)(1-r_M)$ | $n_3$ |
| $A_1D_1$ | $A_2D_2$ | $\frac{1}{4}(1-r_F)r_M$ | $n_4$ |
| $A_1C_1$ | $B_2C_2$ | $\frac{1}{4}r_F(1-r_M)$ | $n_5$ |
| $A_1C_1$ | $B_2D_2$ | $\frac{1}{4}r_Fr_M$ | $n_6$ |
| $A_1D_1$ | $B_2C_2$ | $\frac{1}{4}r_F(1-r_M)$ | $n_7$ |
| $A_1D_1$ | $B_2D_2$ | $\frac{1}{4}r_Fr_M$ | $n_8$ |
| $B_1C_1$ | $A_2C_2$ | $\frac{1}{4}(1-r_F)(1-r_M)$ | $n_9$ |
| $B_1D_1$ | $A_2C_2$ | $\frac{1}{4}(1-r_F)r_M$ | $n_{10}$ |
| $B_1C_1$ | $A_2D_2$ | $\frac{1}{4}(1-r_F)(1-r_M)$ | $n_{11}$ |
| $B_1D_1$ | $A_2D_2$ | $\frac{1}{4}(1-r_F)r_M$ | $n_{12}$ |
| $B_1C_1$ | $B_2C_2$ | $\frac{1}{4}(1-r_F)(1-r_M)$ | $n_{13}$ |
| $B_1D_1$ | $B_2C_2$ | $\frac{1}{4}(1-r_F)r_M$ | $n_{14}$ |
| $B_1D_1$ | $B_2C_2$ | $\frac{1}{4}(1-r_F)(1-r_M)$ | $n_{15}$ |
| $B_1D_1$ | $B_2D_2$ | $\frac{1}{4}(1-r_F)r_M$ | $n_{16}$ |

Four alleles can be clearly identified at each of the two linked loci (scenario 1 in Table 1). $A_i$, $B_i$, $C_i$, and $D_i$ are the four alleles at locus 1. $A_j$, $B_j$, $C_j$, and $D_j$ are the four alleles at locus 2. Recombination frequencies in the female and male parents are denoted as $r_F$ and $r_M$, respectively. The last column gives the symbol of observed sample size of each genotype.

For convenience, the 16 genotypes were rearranged in Table 2, and sample sizes of the 16 genotypes were represented by $n_1$, $n_2$, . . . , and $n_{16}$. Based on theoretical frequencies and sample sizes in Table 2, the likelihood function ($L$) and logarithm likelihood ($\log L$) can be constructed in Equation (1).

$$L = \frac{n!}{n_1! \cdots n_{16}!} \left[ \frac{1}{4}(1-r_F)(1-r_M) \right]^{n_1+n_2+n_3+n_4} \times \left[ \frac{1}{4}(1-r_F)r_M \right]^{n_5+n_6+n_7+n_8} \times \left[ \frac{1}{4}r_Fr_M \right]^{n_9+n_{10}+n_{11}+n_{12}} \times \left[ \frac{1}{4}r_F(1-r_M) \right]^{n_13+n_{14}+n_{15}+n_{16}}$$

$$\log L = C + (n_{14} + n_{13:16}) \log (1 - r_F) + n_{5:12} \log r_F$$

$$+ (n_1 + n_{4:5} + n_{8:9} + n_{12:13} + n_{16})$$

$$\times \log (1 - r_M) + (n_{2:3} + n_{6:7} + n_{10:11} + n_{14:15}) \log r_M,$$

Equation (1), MLE of $r_F$ and $r_M$ can be directly calculated from Equation (2).

$$\hat{r}_F = \frac{n_{5:12}}{n} \quad \hat{r}_M = \frac{n_{3:13} + n_{6:7} + n_{10:11} + n_{14:15}}{n}$$

where $C$ is a constant independent of the unknown recombination frequencies. The maximum likelihood estimates (MLE) of recombination frequencies can be calculated either by solving the likelihood equation (i.e., $\frac{\partial \log L}{\partial \theta} = 0$) or by some approximate algorithms where there is no analytic solution to the likelihood equation. From Equation (1), MLE of $r_F$ and $r_M$ can be directly calculated from Equation (2).

$$\begin{align*}
\hat{r}_F &= \frac{1}{2}(\hat{r}_F + \hat{r}_M) & \text{if } \hat{r}_F \leq 0.5, \quad \hat{r}_M \leq 0.5 \quad \text{(i.e. linkage phase I)} \\
\hat{r}_F &= \frac{1}{2} \hat{r}_F + \frac{1}{2} (1 - \hat{r}_M) & \text{if } \hat{r}_F \leq 0.5, \quad \hat{r}_M > 0.5 \quad \text{(i.e. linkage phase II)} \\
\hat{r}_F &= \frac{1}{2} (1 - \hat{r}_F) + \frac{1}{2} \hat{r}_M & \text{if } \hat{r}_F > 0.5, \quad \hat{r}_M \leq 0.5 \quad \text{(i.e. linkage phase III)} \\
\hat{r}_F &= 1 - \frac{1}{2} (\hat{r}_F + \hat{r}_M) & \text{if } \hat{r}_F > 0.5, \quad \hat{r}_M > 0.5 \quad \text{(i.e. linkage phase IV)}
\end{align*}$$

It can be easily seen that the estimate thus defined in Equation (3) is always less than 0.5. In addition, it can be proved that the estimate in Equation (3) is also MLE of $r$, when directly calculated from its likelihood function.

Recombination frequency estimation in scenarios 2 and 3 in clonal F₁ progenies

In scenario 2, locus 1 has four genotypes $A_1C_1$, $A_1D_1$, $B_1C_1$, and $B_1D_1$, and locus 2 has two genotypes $X_1C_2$ and $X_2D_2$. In scenario 3, locus 1 has four genotypes $A_1C_1$, $A_1D_1$, $B_1C_1$, and $B_1D_1$, and locus 2 has two genotypes $A_2X_2$ and $B_2X_2$. Table 3 shows theoretical frequencies of the eight identifiable genotypes at the two loci. The theoretical frequencies do not contain the female recombination frequency in scenario 2, and they do not contain the male recombination frequency in scenario 3. Therefore, $r_F$ cannot be estimated in scenario 2; $r_M$ cannot be estimated in scenario 3. MLE of $r_M$ in
Table 3 Theoretical frequencies of the eight identifiable genotypes in the clonal F1 population

| Genotype | Locus 1 | Locus 2 (X₂ = A₂ or B₂) | Frequency |
|----------|---------|--------------------------|-----------|
| 1        | A₁C₁    | X₀C₂                     | ½(1 - rₐ) |
| 2        | A₁C₁    | X₀D₂                     | ½(1 - rₐ) |
| 3        | A₁D₁    | X₀C₂                     | ½rₐ       |
| 4        | A₁D₁    | X₀D₂                     | ½rₐ       |
| 5        | B₁C₁    | X₀C₂                     | ½(1 - rₐ) |
| 6        | B₁C₁    | X₀D₂                     | ½(1 - rₐ) |
| 7        | B₁D₁    | X₀C₂                     | ½rₐ       |
| 8        | B₁D₁    | X₀D₂                     | ½rₐ       |

For scenarios 2 and 3 (Table 1), A₁, B₁, C₁, and D₁ are the four alleles at locus 1. For scenario 2, X₂ = A₂ or B₂, C₂, and D₂ are the three alleles at locus 2. For scenario 3, A₂, B₂, and X₂ = C₂ or D₂ are the three alleles at locus 2. Recombination frequencies in the female and male parents are denoted as rₐ and rₘ, respectively. The last column gives the symbol of observed sample size of each genotype.

For scenarios 2 and 3 (Table 1), A₁, B₁, C₁, and D₁ are the four alleles at locus 1. For scenario 2, X₂ = A₂ or B₂, C₂, and D₂ are the three alleles at locus 2. For scenario 3, A₂, B₂, and X₂ = C₂ or D₂ are the three alleles at locus 2. Recombination frequencies in the female and male parents are denoted as rₐ and rₘ, respectively. The last column gives the symbol of observed sample size of each genotype.

3.2 Combined Recombination Frequency

The combined recombination frequency, \( \hat{r} \), can be calculated from its likelihood functions, given in Equation (4).

\[
\hat{r}_M = \frac{\hat{n}_{23} + \hat{n}_{47}}{\hat{n}}. \tag{4}
\]

where \( \hat{n}_i \) is the observed sample size for the \( i \)th genotype (Table 3), \( \hat{n}_{ij} \) is the sum of \( \hat{n}_i \) to \( \hat{n}_j \), and \( \hat{n} \) is the total sample size (i.e., \( \hat{n} = \sum \hat{n}_i \)). Define the estimate of \( \hat{r} \) in Equation (5).

\[
\hat{r} = \begin{cases} 
\hat{r}_M & \text{if } \hat{r}_M \leq 0.5 \\
1 - \hat{r}_M & \text{otherwise} 
\end{cases}. \tag{5}
\]

It can be easily seen that the estimate thus defined is less than 0.5. In addition, the estimate in Equation (5) is MLE of \( \hat{r} \), when directly calculated from its likelihood function.

MLE of \( \hat{r}_F \) in scenario 3 can be calculated from its likelihood function, given in Equation (6).

\[
\hat{r}_F = \frac{\hat{n}_2 + \hat{n}_{45} + \hat{n}_7}{\hat{n}}. \tag{6}
\]

where \( \hat{n}_i \) is the observed sample size of the \( i \)th genotype (Table 3), \( \hat{n}_{ij} \) is the sum of \( \hat{n}_i \) to \( \hat{n}_j \), and \( \hat{n} \) is the total sample size (i.e., \( \hat{n} = \sum \hat{n}_i \)). Define the estimate of \( \hat{r} \) in Equation (7).

\[
\hat{r} = \begin{cases} 
\hat{r}_F & \text{if } \hat{r}_F \leq 0.5 \\
1 - \hat{r}_F & \text{otherwise}
\end{cases}. \tag{7}
\]

Similar to Equation (5), the estimate thus defined is less than 0.5, and is MLE of \( r \).

Recombination frequency estimation in scenario 4 in clonal F₁ progenies

In this scenario, locus 1 has four genotypes A₁C₁, A₁D₁, B₁C₁, and B₁D₁, and locus 2 has three genotypes A₂A₂, A₂B₂, and B₂B₂. Table 4 shows theoretical frequencies of the 12 identifiable genotypes at the two loci. Information on \( r_F \) and \( r_M \) is confounded in half of the genotypes. MLE of \( r_F \) and \( r_M \) using the other half of the genotypes are given in Equation (8).

\[
\hat{r}_F = \frac{\hat{n}_3 + \hat{n}_{47} + \hat{n}_{10}}{\hat{n}_1 + \hat{n}_3 + \hat{n}_{47} + \hat{n}_{9} + \hat{n}_{10} + \hat{n}_{12}}, \tag{8}
\]

\[
\hat{r}_M = \frac{\hat{n}_{34} + \hat{n}_{69} + \hat{n}_{10}}{\hat{n}_1 + \hat{n}_3 + \hat{n}_{47} + \hat{n}_{9} + \hat{n}_{10} + \hat{n}_{12}},
\]

where \( \hat{n}_i \) is the observed sample size of the \( i \)th genotype and \( \hat{n}_{ij} \) is the sum of \( \hat{n}_i \) to \( \hat{n}_j \).

As stated, estimated \( r_F \) and \( r_M \) in Equation (8) can be used in determining the linkage phases in both parents. Then, the theoretical
Recombination frequency estimation in scenarios 5 and 7 in clonal F₁ progenies

In scenario 5, locus 1 has two genotypes X₁C₁ and X₁D₁, and locus 2 has two genotypes X₂C₂ and X₂D₂. In scenario 6, locus 1 has two genotypes A₁X₁ and B₁X₁, and locus 2 has two genotypes A₂X₂ and B₂X₂. Table 5 shows theoretical frequencies of the four identifiable genotypes at the two loci. Obviously, theoretical frequencies do not contain the female recombination frequency in scenario 5 and do not contain the male recombination frequency in scenario 7. Thus, r_F cannot be estimated in scenario 5; r_M cannot be estimated in scenario 7. MLE of r_M in scenario 5 can be calculated from its likelihood functions, given in Equation (9).

\[ r_M = \frac{n_{j3}}{n} \]  \hspace{1cm} (9)

where nᵢ is the observed sample size of the ith genotype (Table 5), nᵢj is the sum of nᵢ to nᵢj, and n is the total sample size \( i.e., \ n=n_{t,x} \). Define the estimate of r in Equation (10).

\[ r = \begin{cases} r_M & \text{if } r_M \leq 0.5 \\ 1 - r_M & \text{otherwise} \end{cases} \]  \hspace{1cm} (10)

MLE of r_F in scenario 7 can be calculated from its likelihood functions, given in Equation (11). Define the estimate of r in Equation (12).

\[ r_F = \frac{n_{j3}}{n} \]  \hspace{1cm} (11)

\[ r = \begin{cases} r_F & \text{if } r_F \leq 0.5 \\ 1 - r_F & \text{otherwise} \end{cases} \]  \hspace{1cm} (12)

Similar to Equation (5), Equation (7), Equation (10), and Equation (12), the estimates defined in Equation (14), Equation (15), and Equation (16) are less than 0.5 and are MLE of r for scenarios 5 and 7, respectively.

Recombination frequency estimation in scenarios 6 and 8 in clonal F₁ progenies

In scenario 6, locus 1 has two genotypes X₁C₁ and X₁D₁, and locus 2 has three genotypes A₂A₂, A₂B₂, and B₂B₂. In scenario 8, locus 1 has two genotypes A₁X₁ and B₁X₁, and locus 2 has three genotypes A₃A₃, A₃B₃, and B₃B₃. Table 6 shows theoretical frequencies of the six identifiable genotypes at the two linked loci. The theoretical frequencies do not contain the female recombination frequency in scenario 6 and do not contain the male recombination frequency in scenario 8. Thus, r_F cannot be estimated in scenario 6, and r_M cannot be estimated in scenario 8. MLE of r_M in scenario 6 can be calculated from its likelihood function, given in Equation (13).

\[ r_M = \frac{n_{j3}}{n_1 + n_{j3} + n_6} \]  \hspace{1cm} (13)

where nᵢ is the observed sample size of the ith genotype (Table 6) and nᵢj is the sum of nᵢ to nᵢj. Define the estimate of r in Equation (14).

\[ r = \begin{cases} r_M & \text{if } r_M \leq 0.5 \\ 1 - r_M & \text{otherwise} \end{cases} \]  \hspace{1cm} (14)

Maximum likelihood estimates of r_F in scenario 8 can be calculated from its likelihood function, given in Equation (15). Define the estimate of r in Equation (16).

\[ r_F = \frac{n_{j3}}{n_1 + n_{j3} + n_6} \]  \hspace{1cm} (15)

\[ r = \begin{cases} r_F & \text{if } r_F \leq 0.5 \\ 1 - r_F & \text{otherwise} \end{cases} \]  \hspace{1cm} (16)

Similar to Equation (5), Equation (7), Equation (9), and Equation (10), the estimates defined in Equation (14), Equation (15), and Equation (16) are less than 0.5 and are MLE of r for scenarios 6 and 8, respectively.

Table 5 Theoretical frequencies of the four identifiable genotypes in the clonal F₁ population

| Genotype | Scenario 5 (Table 1) | Scenario 7 (Table 1) | Frequency | Scenario 7 (Table 1) | Frequency |
|----------|----------------------|----------------------|-----------|----------------------|-----------|
| Locus 1  (X₁ = A₁ or B₁) | Locus 2  (X₂ = A₂ or B₂) | Frequency | Locus 1  (X₁ = C₁ or D₁) | Locus 2  (X₂ = C₂ or D₂) | Frequency |
|         |                      |                      |          |                      |          |
| 1       | X₁C₁                | X₂C₂                | \( \frac{1}{2}(1 - r_M) \) | A₁X₁    | A₂X₂                | \( \frac{1}{2}(1 - r_F) \) | n₁        |
| 2       | X₁C₁                | X₂D₂                | \( \frac{1}{2}r_M \)         | A₁X₁    | B₂X₂                | \( \frac{1}{2}r_F \)         | n₂        |
| 3       | X₁D₁                | X₂C₂                | \( \frac{1}{2}r_M \)         | B₁X₁    | A₂X₂                | \( \frac{1}{2}r_F \)         | n₃        |
| 4       | X₁D₁                | X₂D₂                | \( \frac{1}{2}(1 - r_M) \) | B₁X₁    | B₂X₂                | \( \frac{1}{2}(1 - r_F) \) | n₄        |

For scenarios 5 and 7 (Table 1). For scenario 5, \( X₁ = \{A₁ or B₁ \} \), C₁ and D₁ are the three alleles at locus 1; \( X₂ = \{A₂ or B₂ \} \), C₂, and D₂ are the three alleles at locus 2. For scenario 7, A₁, B₁, and \( X₁ = \{C₁ or D₁ \} \) are the three alleles at locus 1; \( A₂, B₂, \) and \( X₂ = \{C₂ or D₂ \} \) are the three alleles at locus 2. Recombination frequencies in the female and male parents are denoted as \( r_f \) and \( r_m \), respectively. The last column gives the symbol of observed sample size of each genotype.

432 | L. Zhang, H. Li, and J. Wang
III. If the true linkage phase was IV (Figure 3C), then the estimated recombination frequency is the observed sample size of the segregating progenies from the estimated recombination frequencies and constructed linkage phase. Obviously, if the experimental phase coincides at 0.8 in linkage phase I, at 0.5 in linkage phases II and III, and at 0.2 when both marker loci are assigned to one of the four phases. When true recombination frequency is 0.5, then recombination frequency is estimated at approximately 0.5 in all linkage phases (Figure 3D). In this case, linkage phase does not make any sense and is randomly assigned to one of the four phases.

Consistent with previous scenarios, \( r_F \) and \( r_M \) need to be defined to reflect the identified linkage phase after \( r \) and linkage phase are determined. For this purpose, \( r_F \) and \( r_M \) are both assigned to \( r \) in linkage phase I, assigned to \( 1 - r \) in linkage phase II, assigned to \( 1 - r \) and \( r \) in linkage phase III, and assigned to \( 1 - r \) in linkage phase IV. For convenience, estimates of \( r_F \) and \( r_M \) are given in Equation 18.

\[
\tilde{r}_F = \begin{cases} 
\hat{r} & \text{for phase I or II} \\
1 - \hat{r} & \text{for phase III or IV}
\end{cases}
\]

\[
\tilde{r}_M = \begin{cases} 
\hat{r} & \text{for phase I or III} \\
1 - \hat{r} & \text{for phase II or IV}
\end{cases}
\]  

### Table 6: Theoretical frequencies of the six identifiable genotypes in the clonal F₁ population

| Genotype | Locus 1 | Frequency | Sample Size |
|----------|---------|-----------|-------------|
|          | Scenario 6 | Scenario 8 | Locus 2 | Scenario 6 | Scenario 8 | Sample Size |
| 1        | X₁C₁     | A₁X₁      | A₂A₂      | \( \frac{1}{2}(1 - r) \) | \( \frac{1}{2}(1 - r) \) | \( n₁ \) |
| 2        | X₁C₁     | A₁X₁      | A₂B₂      | \( \frac{1}{2}(1 - r) \) | \( \frac{1}{2}(1 - r) \) | \( n₂ \) |
| 3        | X₁C₁     | A₁X₁      | B₂B₂      | \( \frac{1}{2}r \) | \( \frac{1}{2}r \) | \( n₃ \) |
| 4        | X₁D₁     | B₁X₁      | A₂A₂      | \( \frac{1}{2}r \) | \( \frac{1}{2}r \) | \( n₄ \) |
| 5        | X₁D₁     | B₁X₁      | A₂B₂      | \( \frac{1}{2}r \) | \( \frac{1}{2}r \) | \( n₅ \) |
| 6        | X₁D₁     | B₁X₁      | B₂B₂      | \( \frac{1}{2}(1 - r) \) | \( \frac{1}{2}(1 - r) \) | \( n₆ \) |

For scenarios 6 and 8 (Table 1). \( X₁, X₂, C₁, C₂, D₁, \) and \( D₂ \) are the three alleles at locus 1; \( A_1 \) and \( B_2 \) are the two alleles at locus 2. For scenario 6, \( A₁ \) and \( B₂ \) are the three alleles at locus 1; \( A₂ \) and \( B₁ \) are the two alleles at locus 2.

### Table 7: Theoretical frequencies of the nine identifiable genotypes in the clonal F₁ population

| Genotype | Locus 1 | Locus 2 | Expected Frequency | Sample Size |
|----------|---------|---------|--------------------|-------------|
|          | Scenario 6 | Scenario 8 | Phase I | Phase II and III | Phase IV | Sample Size |
| 1        | A₁A₁     | A₂A₂    | \( \frac{1}{2}(1 - r) \)^2 | \( \frac{1}{2}r(1 - r) \) | \( \frac{1}{2}r^2 \) | \( n₁ \) |
| 2        | A₁A₁     | A₂B₂    | \( \frac{1}{2}r(1 - r) \) | \( \frac{1}{2}(1 - 2r + 2r^2) \) | \( \frac{1}{2}r(1 - r) \) | \( n₂ \) |
| 3        | A₁A₁     | B₂B₂    | \( \frac{1}{2}r^2 \) | \( \frac{1}{2}r(1 - r) \) | \( \frac{1}{2}(1 - r) \) | \( n₃ \) |
| 4        | A₁B₁     | A₂A₂    | \( \frac{1}{2}(1 - r) \) | \( \frac{1}{2}(1 - 2r + 2r^2) \) | \( \frac{1}{2}r(1 - r) \) | \( n₄ \) |
| 5        | A₁B₁     | A₂B₂    | \( \frac{1}{2}(1 - r) \) | \( \frac{1}{2}(1 - 2r + 2r^2) \) | \( \frac{1}{2}r(1 - r) \) | \( n₅ \) |
| 6        | A₁B₁     | B₂A₂    | \( \frac{1}{2}r(1 - r) \) | \( \frac{1}{2}(1 - 2r + 2r^2) \) | \( \frac{1}{2}(1 - 2r + 2r^2) \) | \( n₆ \) |
| 7        | B₁B₁     | A₂A₂    | \( \frac{1}{2}r^2 \) | \( \frac{1}{2}r(1 - r) \) | \( \frac{1}{2}(1 - r) \) | \( n₇ \) |
| 8        | B₁B₁     | A₂B₂    | \( \frac{1}{2}r(1 - r) \) | \( \frac{1}{2}(1 - 2r + 2r^2) \) | \( \frac{1}{2}r(1 - r) \) | \( n₈ \) |
| 9        | B₁B₁     | B₂B₂    | \( \frac{1}{2}(1 - r) \) | \( \frac{1}{2}r(1 - r) \) | \( \frac{1}{2}(1 - r) \) | \( n₉ \) |

For scenario 9 (Table 1). \( A₁ \) and \( B₂ \) are the two alleles at locus 1; \( A₂ \) and \( B₁ \) are the two alleles at locus 2. The combined recombination frequency is denoted as \( r \). The last column gives the symbol of observed sample size of each genotype. For linkage phase I, female and male parents have the same genotype \( A₁A₂/B₁B₂ \). For linkage phase II, female and male parents have genotypes \( A₁A₂/B₁B₂ \) and \( A₂B₁/B₂A₂ \), respectively. For linkage phase III, female and male parents have genotypes \( A₁B₁/B₂A₂ \) and \( A₂A₂/B₁B₂ \), respectively. For linkage phase IV, female and male parents have the same genotype \( A₁B₁/B₂A₂ \).
haploid building tries to assign the two alleles A and B at each locus on the female map to haploids HapA and HapB. Male haploid building tries to assign the two alleles C and D at each locus on the male map to haploids HapC and HapD. Haploid building of ordered markers on one chromosome is similar for both parents. We use the female parent as an example to explain the building procedure.

Step 1: At the first ordered locus, allele A is assigned to HapA; allele B is assigned to HapB.

Step 2: For the second ordered locus, if estimated rf with the first locus is lower than 0.5, then allele A is assigned to HapA; allele B is assigned to HapB. Otherwise, allele B is assigned to HapA and allele A is assigned to HapB.

Step 3: For the next ordered locus, if estimated rf with its previous locus is lower than 0.5, then allele A is assigned to HapA, and allele B is assigned to HapB if allele A at the previous locus is on HapA; allele B is assigned to HapA and allele A is assigned to HapB if allele B at the previous locus is on HapA. If estimated rf with its previous locus is more than 0.5, then allele B is assigned to HapA, and allele A is assigned to HapB if allele A at the previous locus is on HapA; allele A is assigned to HapA and allele B is assigned to HapB if allele B at the previous locus is on HapA.

Step 4: Repeat the process from step 3 until the last ordered locus on the chromosome.

Marker categories and linkage analysis in double cross populations

Double cross populations in plants have four inbred lines, A, B, C, and D, as parents that are homozygous at most chromosomal locations (Figure S1). First, one F1 hybrid is made between inbred lines A and B; the other F1 hybrid is made between inbred lines C and D. Then, a double cross is made between the two F1 hybrids; one is used as female and the other one is used as male. When polymorphism markers are screened in the four inbred lines, the four alleles in double cross populations can be clearly assigned. In this case, five marker categories can be differentiated on the number of identifiable alleles in the four original lines and the number of identifiable genotypes in their double cross progenies (Figure S2). Categories I to III are similar to those in clonal F1. Category IV in clonal F1 can be further divided into two categories in double cross, which are denoted as categories IV and V. For category IV and V, theoretical genotypic frequencies and formulas in estimating recombination frequency (r) when category V is absent, scenarios 1 to 9 in clonal F1 are still applicable in double cross populations. For these scenarios, theoretical genotypic frequencies and formulas in estimating rf, rM, and r are the same as those for clonal F1 in the case of linkage phase I, i.e., rf, rM, and r are both smaller than 0.5 if they can be estimated.

There are five new scenarios for recombination frequency estimation when category V is present. In scenario 10, locus 1 is category I and locus 2 is category V. In scenario 11, locus 1 is category II and locus 2 is category V. In scenario 12, locus 1 is category III and locus 2 is category V. In scenario 13, locus 1 is category IV and locus 2 is category V. In scenario 14, the two loci are category V.

In scenario 10, the 12 identifiable genotypes are the same as scenario 4 in Table 4. Theoretical frequency of each genotype is equal to the corresponding value in Table 4 by substituting rM with 1−rM (see Table S2). In scenario 11, the six identifiable genotypes are the same as scenario 6 in Table 6. Theoretical frequency of each genotype is equal to the corresponding value of scenario 6 in Table 6 by substituting rM with 1−rM (Table S3). In scenario 12, the six identifiable genotypes and their theoretical frequencies are the same as scenario 8 in Table 6 (Table S3). In scenario 13, genotypes and their theoretical frequencies are the same as linkage phase II of scenario 9 in Table 7 (Table S4). In scenario 14, genotypes and their theoretical frequencies are the same as linkage phase I of scenario 9 in Table 7 (Table S4). Methods for estimating r are similar to the corresponding scenarios in clonal F1. For convenience, theoretical genotypic frequencies at two loci for scenarios 10 to 14 are given in Table S2, Table S3, and Table S4.
**LOD score in testing the linkage relationship between two loci**

The existence of the linkage can be tested by the following two hypotheses.

\[ H_0 : r = 0.5 \text{ vs. } H_A : r < 0.5, \]

where \( H_0 \) is the null hypothesis corresponding to no genetic linkage, \( H_A \) is the alternative hypothesis corresponding to the linkage relationship between two loci, and \( r \) is the combined recombination frequency. The log-likelihood function under the null hypothesis is \( \log L_0 = \log L(r = 0.5) \), whereas the log-likelihood function under the alternative hypothesis is \( \log L_A = \log L(r = r) \). The LOD score can be calculated from the log-likelihoods under the two hypotheses, i.e.,

\[ LOD = \log L_A - \log L_0, \]

where \( \log \) is the logarithm function of base 10.

**One simulated population and one actual population**

We considered one chromosome with 20 evenly distributed markers in simulation. Recombination frequencies between any two neighboring markers were set at 0.05, equivalent to a genetic distance of 5.27 cM using Haldane mapping function (Haldane 1919).

One population with 200 clonal \( F_1 \) progenies was simulated by the genetics and breeding simulation tool of QuLine (Wang et al. 2003, 2004). Five markers were randomly chosen and assigned to each of the four categories (Figure 2). Markers 8, 11, 14, 17, and 19 were assigned to category III. Alleles I; markers 1, 2, 13, 15, and 20 were assigned to category II; markers 4, 6, 7, 9, 18 were assigned to category IV. The theoretical recombination frequencies between markers were shown in the upper triangular matrix (Table S5). The closer to the diagonal, the lower the recombination frequencies would be. For example, recombination frequencies between marker pair 1 and markers 2, 8, and 19 were 0.05, 0.26, and 0.42, respectively. Recombination frequencies of marker pairs 8 and 9, 8 and 15, and 8 and 20 were 0.05, 0.26, and 0.36 (Table S5), respectively.

The lower triangular matrix of Table S5 showed the estimated recombination frequencies between the 20 markers. Combined recombination frequencies cannot be estimated if one marker is category II and the other one is category III. For example, recombination frequencies between marker pair 1 and 4 and marker pair 5 and 13 cannot be estimated, which were left as blank in Table S5. When the combined recombination frequencies could be estimated, the estimates were close to their true values. For example, marker 1 was category II, its true recombination frequencies with markers 2, 8, and 19 were 0.05, 0.26, and 0.42, and the estimates were 0.05, 0.22, and 0.48, respectively. Marker 8 was category I, its true recombination frequencies with markers 9, 15, and 20 were 0.05, 0.26, and 0.36, and the estimates were 0.03, 0.27, and 0.42, respectively.

If combined recombination frequency cannot be estimated, then the corresponding marker distance and LOD score cannot be calculated either. The upper triangular matrix showed the estimated map distance between the 20 markers (Table S6). The closer between two markers, the smaller the estimated distance is. For example, the true recombination frequencies of marker pairs 1 and 2, 1 and 8, and 1 and 19 were 0.05, 0.26, and 0.42 (Table S5). Their estimated distances were 5.3 cM, 29.0 cM, and 160.9 cM (Table S6), respectively. The true recombination frequencies of marker pairs 8 and 9, 8 and 15, and 8 and 20 were 0.05, 0.26, and 0.36 (Table S5). Their estimated distances were 3.1 cM, 37.8 cM, and 88.6 cM (Table S6), respectively. It should be noted that the map length of a chromosome is calculated from lengths of individual ordered intervals, rather than the recombination frequency between the first and the last markers.

The lower triangular matrix of Table S6 showed LOD score between the 20 markers. The closer between two markers, the greater the LOD score is. For example, the true recombination frequencies between marker pairs 1 and 2, 1 and 8, and 1 and 19 were 0.05, 0.26, and 0.42 (Table S5). Their LOD scores were 43.0, 14.4, and 0.1 (Table S6), respectively. The true recombination frequencies between marker pairs 8 and 9, 8 and 15, and 8 and 20 were 0.05, 0.26, and 0.36 (Table S5). Their LOD scores were 48.5, 10.0, and 1.3 (Table S6), respectively.

**Marker ordering in simulated population**

Estimates of the combined recombination frequencies were used to order the 20 markers, and the best order with the shortest map length...
was shown in Figure 4A, which was the same as the predefined order. The estimated length of the chromosome was 101.79 cM, close to the true length 100.13 cM. Average marker distance was 5.36 cM, close to the true value 5.27 cM.

The female map does not contain markers of category II, and the male map does not contain markers of category III. The order of markers in the female and male maps were the same as that in the combined map, but map distances between markers were estimated by the female and male recombination frequencies, respectively. In the simulated population, lengths of the female and male maps were 81.90 cM and 103.02 cM, respectively (Figure 4, B and C). For the 20 markers, 1, 2, 13, 15, and 20 are category II (Table S4 and Table S5) and therefore do not appear on the female map. Marker 3 was located at the beginning and marker 19 located at the end on the female map, which explained the reduced female map length. Markers, 4, 5, 7, 9, and 17 are category III (Table S4 and Table S5); therefore, they do not appear on the male map. However, marker 1 was still located at the beginning and 20 was still located at the end on the male map, which explained the map length similar to the combined one.

Four haploids of two parents in the simulated population

Using estimated $r_f$ and $r_M$ between neighboring markers, four haploids of parents at 20 marker loci were determined (Table 8). The first marker is category II, which had no polymorphism in the female parent. It was not included on the female map, but it was included on the male map (Figure 4, B and C). Alleles on HapA and HapB were represented by X, which could be either allele A or allele B. Alleles on HapC and HapD were C and D, respectively. The second marker is category II as well. The estimated $r_M$ with previous marker was 0.05 (less than 0.5). Alleles on HapA and HapB were represented by X, which could be either allele A or allele B. Alleles on HapC and HapD were C and D, respectively, which were the same haploids as those of the previous locus. Marker 3 was the first on the female map (Figure 4B). Alleles A and B were on HapA and HapB (Table 8). It was the third marker on the male map (Figure 4C). Estimated $r_M$ with previous marker was 0.975, which was more than 0.5. Alleles D and C were assigned to HapC and HapD, respectively, which were opposite to the previous locus. The four haploids in Table 8 were consistent with the predefined haploid types.

Marker category IV in clonal F1 can be further divided into two categories, i.e., categories IV and V in double cross (Figure S2). In a simulated population, markers 3, 6, 10, 12, and 16 were category IV. Taking marker 3 as an example, alleles on HapA, HapB, HapC, and HapD were A, B, D, and C, respectively. Its category was redefined as category V of double cross (Table 8).

For HapA and HapB of the female parents (Table 8), if we exchange alleles A and B at loci 5, 8, 12, 16, and 18, then HapA will have A alleles at all loci and HapB will have B alleles at all loci. For HapC and HapD of the male parents (Table 8), if we exchange alleles C and D at loci 3, 6, 8, 12, 14, 15, and 20, then HapC will have C alleles at all loci and HapD will have D alleles at all loci. If the four haploids built earlier could be viewed as haploids of the four inbred lines in a double cross, then clonal F1 is equivalent to double cross!

Comparison with JoinMap, OneMap, and R/qtl for linkage map construction

General information of combined linkage maps of the simulated population built by GACD, JoinMap4.1, OneMap, and R/qtl were shown (Table S7). R/qtl can only conduct linkage mapping in phase-known double cross, so marker categories and genotypes after haploid building were imported into R/qtl. Marker orders given by GACD, OneMap, and R/qtl were the same as the predefined order in the simulated model. However, marker order given by JoinMap4.1 was far from the predefined (Table S7). The first and last markers were Marker 12 and Marker 18, respectively. The true map length was 100.13 cM. Length was estimated at 101.79 cM from GACD, 15211.04 cM from JoinMap, 103.83 cM from OneMap, or 104.22 cM from R/qtl. The reason for the extremely large map length from JoinMap was the estimated value of 0.5 of recombination frequency between some neighboring markers in the female or male maps, which was converted to a distance of 10,000.0 cM in JoinMap. For example, recombination frequency between markers 3 and 5 belonging to category V and III was estimated at 0.5 on the female map, corresponding to a distance of 10,000.0 cM on the female map and 5007.99 cM on the combined map. Time spent for building the maps was 8 sec by GACD, 30 sec by JoinMap, 455 sec by OneMap, and 63 sec by R/qtl on a computer with 1.60 GHz CPU and 3.00 GB RAM.

Comparison of different software packages was also conducted in a simulated clonal F1 population with distorted markers (Supplementary Materials, see File S4) and a simulated clonal F1 population with 200 individuals and 200 markers belonging to category IV (Supplementary Materials, see File S5). A greater advantage was observed for the marker number 200 in one single chromosome (Table S8). GACD took 0.5 min, JoinMAP took 5 min, OneMAP took 537 min, and R did not output any results. GACD results in the shortest linkage map
closest to the true length in the shortest time (Table S8). The reason may be as follows. Previous studies tried to estimate recombination frequency, determine linkage phase, and build linkage map simultaneously. In our study, we first estimate all pair-wise recombination frequencies (i.e., step 1). Linkage phases were determined from the estimated recombination frequencies (i.e., step 2). Linkage map was built based on the matrix of all pair-wise recombination frequencies (i.e., step 3). Finally, the four haploids were built from the completed linkage maps (i.e., step 4). Separating a complicated genetic question into four clearly defined steps results in more accurate genetic linkage maps in shorter time. In addition, we believe the adoption of the optimization algorithm in solving the Traveling Salesman Problems also contributes to the ordering efficiency.

**Linkage maps in actual double cross population**

In the actual population, the missing marker rate was at 6.49%. Among the 220 markers, 60 markers showed segregation distortion under significance level 0.05. Recombination frequencies of all marker pairs were estimated and then used for linkage map construction. The combined genetic linkage map was constructed by 219 SSR molecular markers using the software GACD. One marker cannot be linked with any other markers and was deleted. The 10 chromosomes had 25, 28, 25, 24, 21, 19, 18, 16, 25, and 18 relatively evenly distributed markers, respectively (Figure S3). The whole genome was 1778.09 cM in length, and the average marker distance was 8.51 cM.

The 10 female chromosomes (Figure S3) had 19, 19, 20, 13, 16, 13, 12, 14, 17, and 15 markers, respectively, with a total of 158 markers. The total female map length was 1796.92 cM. The 10 male chromosomes (Figure S3) had 18, 19, 22, 21, 17, 14, 14, 9, 19, and 15 markers, respectively, with a total of 168 markers. The total male map length was 1599.13 cM.

Li *et al.* (2013) used JoinMap4.0 to build the linkage maps for this actual population. Kosambi mapping function was used to convert recombination frequency to genetic distance. As indicated in their study, 213 makers were included in the 11 linkage groups of the combined map. The other seven markers were not linked. The whole genome was 1626.3 cM, and the average marker distance was 1626.3/ (213–1) = 8.05 cM. Compared with the map by JoinMap, our method provided a methodology that has the following advantages. First, the number of linkage groups from GACD was the same as the number of chromosomes in maize genome. Second, GACD links more markers than JoinMap. One marker was identified by GACD to be unlinked, but seven markers were unlinked by JoinMap. The length of genome from GACD was slightly longer than that from JoinMap. This may be caused by two possible reasons: more markers were included on the linkage maps by GACD and chromosome 2 was split into two by JoinMap.

**DISCUSSION**

**Linkage analysis in clonal F1 progenies using all informative markers**

Linkage analysis and map construction are crucial steps in genetic studies of quantitative traits and provide the basis for map-based gene cloning and marker-assisted breeding. A key to linkage map construction is the accurate estimation of recombination frequency, which has been widely studied for various populations in plants over a long period of time (Fisher 1935; Haldane and Smith 1947; Morton 1955; Smith 1959; Bailey 1961; Ott 1974; Nordheim *et al.* 1983; Ritter *et al.* 1990, 1996; Wu *et al.* 2002a, b; van Ooijen 2011). Sill and Nilsson (1994) showed that the accuracy of recombination frequency estimation was affected by limited sample size, heterogeneity in recombination frequency between sexes or among meiosis, and factors that distort the segregation classification or differential viability. Hackett and Broadfoot (2003) investigated that accuracy of linkage maps was reduced by missing values and/or typing errors in genotyping, but segregation distortion had little effect on marker order. Sun *et al.* (2012) investigated the estimation efficiency of recombination frequency in 12 bi-parental populations. They concluded that larger population size and smaller recombination frequency resulted in higher LOD score and smaller deviation. Advanced backcrossing and selfing populations had lower precision in estimating the recombination frequency due to the enlarged recombination frequency.

The four marker categories (Figure 2) considered in this study represented all polymorphism markers that could provide the required information for genetic studies. Linkage analysis was conducted for markers not only in the same category but also in different categories. Three sets of recombination frequencies could be estimated accordingly to build the female, male, and combined linkage maps simultaneously. Results from simulated populations and one actual maize population demonstrated the accuracy of the proposed method and its advantages over other software packages. Methodology developed in this study, together with the freely available GACD software, provides an integrated and convenient approach that will greatly facilitate the genetic research of clonal species and double crosses.

Single-nucleotide polymorphism (SNP) markers are more and more often being used in genetic analysis. Liu *et al.* (2014) presented a HighMap method for constructing high-density linkage maps from next-generation sequencing (NGS). HighMap used an iterative ordering and error correction strategy based on a k-nearest neighbor algorithm and a Monte Carlo multipoint maximum likelihood algorithm, which also provided an idea for dealing with NGS data. Due to the bi-allelic characteristic, individual SNP markers cannot be in category I. However, any SNP marker can be category II, III, or IV in clonal F1, or category II, III, IV, or V in double crosses. In addition, by using the

---

| Marker | Category | Female Parent | Male Parent | Updated Category |
|-------|----------|---------------|-------------|-----------------|
| 1     | II       | X             | X           | C               | D               | II              |
| 2     | II       | X             | X           | C               | D               | II              |
| 3     | IV       | A             | B           | D               | C               | V               |
| 4     | III      | A             | B           | X               | X               | III             |
| 5     | III      | A             | B           | X               | X               | III             |
| 6     | IV       | A             | B           | D               | C               | V               |
| 7     | III      | A             | B           | X               | X               | III             |
| 8     | I        | B             | A           | D               | C               | I               |
| 9     | III      | A             | B           | X               | X               | III             |
| 10    | IV       | A             | B           | C               | D               | IV              |
| 11    | I        | A             | B           | C               | D               | I               |
| 12    | IV       | B             | A           | D               | C               | IV              |
| 13    | I        | X             | X           | C               | D               | II              |
| 14    | I        | A             | B           | D               | C               | I               |
| 15    | II       | X             | X           | C               | D               | II              |
| 16    | IV       | B             | A           | C               | D               | V               |
| 17    | I        | A             | B           | C               | D               | I               |
| 18    | III      | B             | A           | X               | X               | III             |
| 19    | I        | A             | B           | C               | D               | I               |
| 20    | II       | X             | X           | X               | D               | C               |

HapA and HapB are the two haploids of the female parent. HapC and HapD are the two haploids of the male parent.
concept of haplotypes, it is possible to covert SNP markers to fully informative category I markers. For example, one haplotype is consistent of two closely linked SNP loci. Four genotypes can be identified by considering the two loci together, i.e., 11, 10, 01, and 00. Then, the haplotype can be treated as category I marker in genetic analysis.

**Difference and similarity between clonal F1 and double cross**

In clonal F1, genotype of the female parent can be either $A_1B_1/A_1B_2$ or $A_2B_1/B_2A_2$; genotype of the male parent can be either $C_1D_1/C_2D_2$ or $C_2D_1/C_2D_2$. In double cross, there are four homoyzgous inbred lines whose genotypes may be known. Alleles $A, B, C,$ and $D$ at each polymorphism locus can be traced back to the four inbred lines, when the four lines have been genotyped. In this case, genotype of the single cross between lines A and B is $A_1B_1/A_2B_2$ genotype of the single cross between lines C and D is $C_1D_1/C_2D_2$. Therefore, double cross is actually a special case of clonal F1 in which only linkage phase I is applicable (Figure S4).

In a double cross where polymorphism loci are only screened in the two single crosses, linkage phases become unknown before estimating recombination frequencies. Genotype of one single cross can be either $A_1B_1/A_2B_2$ or $A_1B_2/B_2A_2$ genotype of the other single cross can be either $C_1D_1/C_2D_2$ or $C_2D_1/C_2D_2$. In this case, the double cross must be treated as one clonal F1 population for genetic analysis (Figure S4), as is the case for the actual maize population used in this study.

Linkage phases in both parents of the clonal F1 can be determined by linkage analysis, from which four haplairs can be built. If the four haplairs could be viewed as haplairs of the four inbred lines in a double cross, then clonal F1 is equivalent to double cross. In short, there are many similarities between clonal F1 and double cross, although difference does occur (Figure S4). It is important in genetics to know when clonal F1 and double cross are equivalent and when they are not. Previous genetic studies focused on only one of clonal F1 or double cross population. To our understanding, this study is the first that tried to combine the two kinds of populations. Based on the linkage analysis, two haplairs of the female parent and two haplairs of the male parent can be built, and then the clonal F1 progenies can be viewed as a double cross population derived from four inbred lines. The unified QTL mapping method for the two kinds of populations will be fully investigated in another article (Zhang et al. 2015).

**Classification of marker categories in clonal F1 and double cross**

In clonal F1 and double crosses, both the number of identifiable alleles in parents and the number of identifiable genotypes in F1 progenies need to be considered in the classification of each marker locus. Wu et al. (2002a) only considered parents in marker classification, resulting in 18 possible cross types. However, many of them are identical in linkage analysis, and most cross types can be classified into the four marker categories in this study. For example, types $A_1$ to $A_4$ in Wu et al. (2002a) are identical to category I as defined in this study, because they all generate four genotypes that can be identified in the progenies.

Null alleles were also considered in Maliepaard et al. (1997) and Wu et al. (2002a, b). To our understanding, it is difficult to determine whether one parent carries two identical alleles or carries one allele and one null allele in practice. In the case of no missing data and no segregation distortion, type $D_1$ in Wu et al. (2002a) can be decided by the 1:1 ratio test of two marker types in the progenies, and type $A_5$ can be decided by the 1:1:1:1 ratio test of four marker types in the progenies. Unfortunately, missing data and segregation distortion are common in practical populations. In the case of type $D_1$ and a large amount of missing marker points, we may wrongly say there are three or four marker classes. Even though we do know the number of marker type classes, the segregation ratio could be seriously affected by distortion. Therefore, we do not make the difference between cross types $D_1$ and $A_5$. Instead, both types were treated as nonpolymorphism in the male parent, i.e., category III in this study.

**Wider applications of the clonal genetic analysis methods**

In practice, clonal F1 progenies may come from the selfing pollination of one clonal parent, i.e., female and male parents are from one clone population (Figure S4). In this case there are two alleles at each locus, and only marker category IV and linkage phases I and IV are applicable. Methods proposed in this study can be readily used to estimate recombination frequency, identify linkage phase, and build the two haplairs of the clonal parent. In self-pollinated and cross-pollinated species, an F2 population is the selfing generation of one F1 hybrid between two inbred parents. Linkage phases are known when both inbred parents are genotyped. In this case, the clonal F1 derived from the selfing of one clonal parent can be viewed as an F2 population, after the two parental haplairs are built.

If selfing can be viewed as a cross between the F1 hybrid and itself, the F2 population becomes a special case of clonal F1, when linkage phases are unknown, or a special case of double cross when linkage phases are known (Figure S4). In the F2 population, there are two alleles at each locus; therefore, only marker category IV is applicable. Haplairs built in clonal F1 and double cross may help to identify and correct markers that are misclassified for the two inbred parents. Moreover, genetic analysis in an F2 population can still be performed by the clonal genetic analysis methods, even when there is no genotypic data on its two parental lines or on its F1 ancestry.

More broadly, methodology proposed in this study can be applied in genetic populations derived from any two heterozygotes in animals and plants. For example, in animals, linkage analysis is normally conducted in progenies between one female parent and one male parent, both are highly heterozygous, and they are drawn from a large random-mating population. By using the methodology of clonal F1, it is possible to build the female and male linkage maps to reflect the sex-specific recombination frequencies.

**ACKNOWLEDGMENTS**

This work was supported by the National Science Foundation of China (project no. 31200917), the National 973 Research Program of China (project no. 2014CB138105), and the Generation Challenge Program (GCP) and HarvestPlus Challenge Program of CGIAR.

**LITERATURE CITED**

Allard, R. W., 1999 Principles of Plant Breeding, Ed. 2nd. John Wiley & Sons, Inc., New York, NY.

Bailey, N. T. J., 1961 Introduction to the Mathematical Theory of Genetic Linkage, Oxford University Press, Oxford, UK.

Broman, K. W., H. Wu, S. Sen, and G. A. Churchill, 2003 R/qtl: QTL mapping in experimental crosses. Bioinformatics 19: 889–890.

Carlberg, J. D., A. Reis, M. F. Duval, G. C. D’Eekenbrugge, and J. M. Leitao, 2004 Genetic maps of RAPD, AFLP and ISSR markers in Ananas bracteatus and A. comosus using the pseudo-testcross strategy. Plant Breed. 123: 186–192.

Fisher, R. A., 1935 The detection of linkage with “dominant” abnormalities. Ann. Eugen. 6: 187–201.

Fregene, M., F. Angel, R. Gomez, F. Rodriguez, P. Chavarriaga et al., 1997 A molecular genetic map of cassava (Manihot esculenta Crantz). Theor. Appl. Genet. 95: 431–441.
