DOI: 10.1534/genetics.106.064030

A New Strategy for Estimating Recombination Fractions Between Dominant Markers From an F_2 Population

Yuan-De Tan*,† and Yun-Xin Fu*,‡,[1]

*Human Genetics Center, School of Public Health, University of Texas, Houston, Texas 77030, †College of Life Science, Hunan Normal University, Changsha, 410081, Hunan, People's Republic of China and ‡Laboratory for Conservation and Utilization of Bioresources, Yunnan University, Kunming Province, 65091, Yunnan, China

Manuscript received July 27, 2006
Accepted for publication October 9, 2006

ABSTRACT

Although most high-density linkage maps have been constructed from codominant markers such as single-nucleotide polymorphisms (SNPs) and microsatellites due to their high linkage information, dominant markers can be expected to be even more significant as proteomic technique becomes widely applicable to generate protein polymorphism data from large samples. However, for dominant markers, two possible linkage phases between a pair of markers complicate the estimation of recombination fractions between markers and consequently the construction of linkage maps. The low linkage information of the repulsion phase and high linkage information of coupling phase have led geneticists to construct two separate but related linkage maps. To circumvent this problem, we proposed a new method for estimating the recombination fraction between markers, which greatly improves the accuracy of estimation through distinction between the coupling phase and the repulsion phase of the linked loci. The results obtained from both real and simulated F_2 dominant marker data indicate that the recombination fractions estimated by the new method contain a large amount of linkage information for constructing a complete linkage map. In addition, the new method is also applicable to data with mixed types of markers (dominant and codominant) with unknown linkage phase.

MOST high-density linkage maps have been constructed from codominant markers such as single-nucleotide polymorphisms (SNPs) and microsatellites because of their high linkage information, but linkage maps of dominant markers will become more and more important because such markers are often related to biological functions and are increasingly available as proteomic techniques are becoming mature. Proteomic markers include position-shift locus (PSL), presence/absence sport (PAS), and protein quantitative locus (PQL) (Thiellement *et al.* 1999; Zivy and de Vienne 2000; Consoli *et al.* 2002), of which PAS and PQL are dominant markers (Thiellement *et al.* 1999; Zivy and de Vienne 2000; Consoli *et al.* 2002). An example of a linkage map constructed from mostly dominant markers is the *Escherichia coli* bacteriophage T7 protein linkage map (Bartel *et al.* 1996). High-density linkage maps in the future will be more likely constructed from both dominant and codominant markers since such maps can provide fine genetic locations of functional markers through high-density codominant markers flanking them. Therefore, accurate estimates of recombination fractions between dominant markers and between dominant and codominant markers are important.

Due to dominance, the genotype of an individual at a dominant marker is often ambiguous, which increases the complexity of analysis. An important issue in the estimation of the recombination fraction is how to efficiently deal with different linkage phases between a pair of dominant loci (Mester *et al.* 2003a). Two different linkage phases for a double heterozygote are well recognized. One is known as the repulsion phase, which corresponds to the situation in which these two dominant alleles reside on different chromosomes; otherwise, it is known as the coupling phase. In a two-point analysis that considers two markers at a time, the repulsion phase provides much less information about linkage than the coupling phase (Allard 1956; Knapp *et al.* 1995; Liu 1998; Mester *et al.* 2003a). This is especially true for double heterozygotes from the F_2 population (Liu 1998). In reality, about half of the markers are in the coupling phase and the remaining markers are in the other coupling phase. The phase between two couplings is repulsion (Liu 1998; Mester *et al.* 2003a). This leads in practice to the construction of two separate partner linkage maps: one is called the paternal map on which markers are derived from the paternal parent and the other is called the maternal map consisting of the maternal markers (Knapp *et al.*

[1]*Corresponding author:* Human Genetics Center, School of Public Health, University of Texas, 1200 Herman Pressler, Houston TX 77030. E-mail: yunxin.fu@uth.tmc.edu

1995; Peng *et al.* 2000; Mester *et al.* 2003a). To date, there is no effective way to integrate the partner maps into a single complete map. Mester *et al.* (2003) attempted to use pairs of codominant and dominant markers to accomplish this task because such pairs of markers in the repulsion phase have higher linkage information than pairs of dominant markers in the coupling phase. However, this strategy is extremely demanding because it requires that every dominant marker be paired with a codominant marker.

The two-point analysis implemented by the expectation-maximization (EM) algorithm (Dempster *et al.* 1977; Lander and Green 1987; Ott 1991) is a powerful approach for estimating recombination fractions between codominant loci and between dominant loci in the coupling phase, but it has a poor resolution for dominant loci in the repulsion phase (see Liu 1998). This is because the two-point analysis cannot distinguish the coupling phase from the repulsion phase of dominant markers, which have rather different statistical properties. In addition to the need for treating coupling and repulsion phases separately, examining three loci at a time will lead to a better utilization of available linkage information. The problem is that not only the number of combinations of the three loci is large when the total number of loci is large, but also the complexity of the analysis increases due to the need to distinguish several types of double or triple heterozygotes. To circumvent these problems, we propose an alternative approach in this article. The new method considers three loci at a time. It first classifies phenotypes into four pairs of gamete genotypes, followed by estimating their frequencies from the sample that led to the identification of the linkage phase of the loci, then estimates recombination fractions between loci according to their linkage phase, and finally reduces the three-point estimates of the recombination fractions to two-point estimates. A key to this strategy is a fast method for estimating the frequencies of different gamete types because of the need to deal with a large number of loci combinations. We are able to develop very efficient estimators of these frequencies by taking advantage of the simplicity of their expectations. The estimates of recombination fractions obtained by this new method make it possible to integrate two separate partner linkage maps based on the EM estimates of recombination fractions into a single complete linkage map.

METHODS

Estimating the frequencies of three-locus gametes: Since the novel method to be described for estimating recombination fractions makes use of the frequencies of gametes defined by alleles from three loci, we start by presenting estimators of these frequencies. Two cases need to be considered separately. The first corresponds to the situation in which all three loci are dominant and thus is referred to as "dominant loci." The second is that only one or two loci out of three are dominant and is referred to as "mixed loci."

Dominant loci: Consider three dominant loci each having two alleles. Let *A* and *a* be the two alleles for the first locus, *B* and *b* be those for the second, and *C* and *c* be those for the third. Uppercase letters denote dominant alleles and lowercase letters recessive alleles. A meiosis from a triple-heterozygote individual of the F_1 population can produces eight different types of three-locus gamete: *ABC*, *ABc*, *Abc*, *AbC*, *aBC*, *abC*, *aBc*, and *abc*, where *ABC* and *abc*, *Abc* and *aBC*, *abC* and *ABc*, and *AbC* and *aBc* are, respectively, sister gametes. These sister gametes are expected to have equal frequency under the assumption of no segregation distortion during meiosis. In practice, a chi-square test can be used to remove loci that exhibit significant segregation distortion. These gametes can be grouped into four pairs of nonsister gametes. Define an F_2 population:

$$\begin{aligned} q_1 &= p(ABC) = p(abc),\\ q_2 &= p(Abc) = p(aBC),\\ q_3 &= p(ABc) = p(abC),\\ q_4 &= p(AbC) = p(aBc). \end{aligned}$$

It follows that $2q_1 + 2q_2 + 2q_3 + 2q_4 = 1$. The individuals of the F_2 population can be classified into four categories. Category i ($i = 0, \ldots, 3$) consists of individuals with exactly i loci possessing a dominant allele. To estimate gamete frequencies, it is necessary to consider the frequency of each category. Let *aabbC_* represent the phenotype in which only locus *c* exhibits a dominant phenotype. Therefore *aabbC_* represent the group of individuals from category 1 whose locus *c* has a dominant allele(s). It is obvious that there are three genotypes in category 1 and *aabbC_* can be further dissected into

$$aabbC\_ \rightarrow \begin{cases} aabbCC \rightarrow (abC)^2: & q_3^2 \\ aabbCc \rightarrow (abC)(abc): & q_3 q_1 \\ aabbcC \rightarrow (abc)(abC): & q_1 q_3. \end{cases}$$

Phenotypes *aaB_cc* and *A_bbcc* are also dissected in a similar fashion.

There are also three phenotypes in category 2, each of which can be dissected into five pairs of sister gametes. For instance, the phenotype *A_B_cc* can be dissected into

$$\begin{aligned} (ABc)(ABc):&\quad q_3^2\\ (ABc)(abc):&\ 2q_3q_1\\ (ABc)(Abc):&\ 2q_3q_2\\ (ABc)(aBc):&\ 2q_3q_4\\ (Abc)(aBc):&\ 2q_2q_4. \end{aligned}$$

Note that the phenotype for category 3 is not very informative since the single phenotype corresponds to

too many genotypes. Therefore frequencies for category 3 are not used.

Let Q_1, Q_2, Q_3, Q_4, Q_5, Q_6, and Q_7 be the expected frequencies of phenotypes *aabbcc*, *aabbC_*, *aaB_cc*, *A_bbcc*, *A_B_cc*, *A_bbC_*, and *aaB_C_* in the F_2 population, respectively. Then

$$\begin{aligned} Q_1 &= q_1^2 \\ Q_2 &= q_3^2 + 2q_1q_3 \\ Q_3 &= q_4^2 + 2q_1q_4 \\ Q_4 &= q_2^2 + 2q_1q_2 \end{aligned} \tag{1}$$

and

$$\begin{aligned} Q_5 &= q_3^2 + 2q_1q_3 + 2(q_3q_2 + q_3q_4 + q_2q_4) \\ Q_6 &= q_4^2 + 2q_1q_4 + 2(q_3q_2 + q_3q_4 + q_2q_4) \\ Q_7 &= q_2^2 + 2q_1q_2 + 2(q_3q_2 + q_3q_4 + q_2q_4). \end{aligned} \tag{2}$$

Letting $Q_0 = 2(q_2q_3 + q_2q_4 + q_3q_4)$, Equation 2 may be rewritten as

$$\begin{aligned} Q_5 &= Q_2 + Q_0 \\ Q_6 &= Q_3 + Q_0 \\ Q_7 &= Q_4 + Q_0. \end{aligned} \tag{3}$$

Moment estimates of $q_1, \ldots, q_4$ can be obtained from the above sets of equations by replacing Q_i by their moment estimates, which are simply their observed frequencies in the sample. Theoretically Equation 1 is sufficient for deriving solutions for q's. However, Equation 3 can be used to further minimize the stochastic effect in the observed frequencies. Specifically, Q_2, Q_3, and Q_4 can be estimated as

$$\begin{aligned} \hat{Q}_2^o &= \hat{Q}_5 - \hat{Q}_0 = 0.25 - (\hat{Q}_1 + \hat{Q}_6 + \hat{Q}_7) \\ \hat{Q}_3^o &= \hat{Q}_6 - \hat{Q}_0 = 0.25 - (\hat{Q}_1 + \hat{Q}_5 + \hat{Q}_7) \\ \hat{Q}_4^o &= \hat{Q}_7 - \hat{Q}_0 = 0.25 - (\hat{Q}_1 + \hat{Q}_5 + \hat{Q}_6), \end{aligned} \tag{4}$$

where $Q_0 = Q_5 + Q_6 + Q_7 + Q_1 - 0.25$ (see APPENDIX A). It follows that Q_2, Q_3, and Q_4 can alternatively be estimated from the observed frequencies of Q_1, Q_5, Q_6, and Q_7. We can combine the two sets of estimates of Q_2, Q_3, and Q_4 to obtain a more stable set of estimates as

$$\begin{aligned} \hat{Q}_2^* &= \frac{1}{a_2 + b_2}(a_2\hat{Q}_2 + b_2\hat{Q}_2^o) \\ \hat{Q}_3^* &= \frac{1}{a_3 + b_3}(a_3\hat{Q}_3 + b_3\hat{Q}_3^o) \\ \hat{Q}_4^* &= \frac{1}{a_4 + b_4}(a_4\hat{Q}_4 + b_4\hat{Q}_4^o), \end{aligned} \tag{5}$$

where a_k and b_k are weights of $\hat{Q}_k$ and $\hat{Q}_k^o$, respectively, where $k = 2, 3, 4$. $\hat{Q}_k$ is the estimate of Q_k. Our simulation study showed that $a_k = b_k$ usually gives the best result for the estimation of q_k. When the sample is small, it is possible that $\hat{Q}_k^o \leq 0$ or $\hat{Q}_k = 0$. In such a case, one can set $a_k = 1$ and $b_k = 0$ for $\hat{Q}_k^o \leq 0$, or $a_k = 0$ and $b_k = 1$ for $\hat{Q}_k^o \geq 0$ and $\hat{Q}_k = 0$.

Since $Q_2 = q_3^2 + 2q_1q_3 + q_1^2 - q_1^2 = (q_3 + q_1)^2 - q_1^2$, therefore q_3 can be expressed as

$$q_3 = \sqrt{Q_2 + Q_1} - \sqrt{Q_1}. \tag{6a}$$

Similarly we have

$$q_2 = \sqrt{Q_4 + Q_1} - \sqrt{Q_1}, \tag{6b}$$

$$q_4 = \sqrt{Q_3 + Q_1} - \sqrt{Q_1}. \tag{6c}$$

Q_2 and Q_1 are estimated by $\hat{Q}_2^*$ and $\hat{Q}_1$, so q_3 is estimated by

$$\hat{q}_3 = \sqrt{\hat{Q}_2 + \hat{Q}_1} - \sqrt{\hat{Q}_1}. \tag{7a}$$

Similarly

$$\hat{q}_2 = \sqrt{\hat{Q}_4 + \hat{Q}_1} - \sqrt{\hat{Q}_1}, \tag{7b}$$

$$\hat{q}_4 = \sqrt{\hat{Q}_3^* + \hat{Q}_1} - \sqrt{\hat{Q}_1}. \tag{7c}$$

q_1 is estimated by

$$\hat{q}_1 = \sqrt{\hat{Q}_1}. \tag{7d}$$

Mixed loci: Two configurations in the case of the mixed loci need to be considered. The first is two codominant loci and one dominant locus (2C1D), and the second is one codominant locus and two dominant loci (1C2D) (see Figure 1). For a codominant locus, "0" and "1" represent two parental types of homozygotes and "2" represent heterozygote. While for the dominant locus, "*A*" and "*a*" represent a dominant phenotype and a recessive phenotype, respectively. Without loss of generality, we assume in the following discussion the order of loci in the case of 2C1D is DCC. The 12 phenotypes are informative for linkage analysis, which are *a*00, *a*01, *a*02, *a*10, *a*11, *a*12, *a*20, *a*21, *A*00, *A*01, *A*10, and *A*11, while phenotypes *A*20, *A*21, and *A*02 and *A*12, *a*22, and *A*22 are much less informative because they are double (or potentially) and triple (or potentially) heterozygotes. In the F_2 population, similar to phenotype *aabbcc* in dominant loci, phenotypes *a*00, *a*01, *a*10, and *a*11 are homozygous and have the expected frequencies $Q_1 = q_1^2$, $Q_2 = q_2^2$, $Q_3 = q_3^2$, and $Q_4 = q_4^2$, respectively, and *A*00, *A*01, *A*10, and *A*11 are similar to *A_bbcc* in dominant loci and have the expected frequencies $Q_{21} = q_2^2 + 2q_2q_1$, $Q_{43} = q_4^2 + 2q_4q_3$, $Q_{34} = q_3^2 + 2q_3q_4$, and $Q_{12} = q_1^2 + 2q_2q_1$, respectively. The frequencies of *a*02, *a*12, *a*20, and *aa*21 are expected to have $P_{13} = 2q_1q_3$, $P_{24} = 2q_2q_4$, $P_{14} = 2q_1q_4$, and $P_{23} = 2q_2q_3$, respectively. Thus, for any nonsister gamete type,

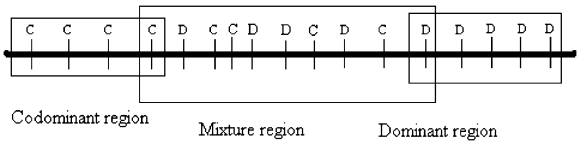

FIGURE 1.—Three marker regions on a chromosome. C, codominant marker; D, dominant marker.

there are three ways to estimate these gamete frequencies. For example, q_1 can be estimated by the following three equations:

$$q_1 = \sqrt{Q_{12} + Q_2} - \sqrt{Q_2}, \tag{8a}$$

$$q_1 = \sqrt{Q_3 + P_{13} + Q_1} - \sqrt{Q_3}, \tag{8b}$$

$$q_1 = \sqrt{Q_4 + P_{14} + Q_1} - \sqrt{Q_4}. \tag{8c}$$

A simple single estimate can be obtained by taking the average of the three. The approach is also used for other gametes, resulting in the estimates

$$\hat{q}_1 = \frac{1}{3}\left[\left(\sqrt{\hat{Q}_{12} + \hat{Q}_2} - \sqrt{\hat{Q}_2}\right) + \left(\sqrt{\hat{Q}_3 + \hat{P}_{13} + \hat{Q}_1} - \sqrt{\hat{Q}_3}\right) + \left(\sqrt{\hat{Q}_4 + \hat{P}_{14} + \hat{Q}_1} - \sqrt{\hat{Q}_4}\right)\right], \tag{9a}$$

$$\hat{q}_2 = \frac{1}{3}\left[\left(\sqrt{\hat{Q}_{21} + \hat{Q}_1} - \sqrt{\hat{Q}_1}\right) + \left(\sqrt{\hat{Q}_3 + \hat{P}_{23} + \hat{Q}_2} - \sqrt{\hat{Q}_3}\right) + \left(\sqrt{\hat{Q}_4 + \hat{P}_{24} + \hat{Q}_2} - \sqrt{\hat{Q}_4}\right)\right], \tag{9b}$$

$$\hat{q}_3 = \frac{1}{3}\left[\left(\sqrt{\hat{Q}_{34} + \hat{Q}_4} - \sqrt{\hat{Q}_4}\right) + \left(\sqrt{\hat{Q}_3 + \hat{P}_{13} + \hat{Q}_1} - \sqrt{\hat{Q}_1}\right) + \left(\sqrt{\hat{Q}_3 + \hat{P}_{23} + \hat{Q}_2} - \sqrt{\hat{Q}_2}\right)\right], \tag{9c}$$

$$\hat{q}_4 = \frac{1}{3}\left[\left(\sqrt{\hat{Q}_{43} + \hat{Q}_3} - \sqrt{\hat{Q}_3}\right) + \left(\sqrt{\hat{Q}_2 + \hat{P}_{24} + \hat{Q}_4} - \sqrt{\hat{Q}_2}\right) + \left(\sqrt{\hat{Q}_4 + \hat{P}_{14} + \hat{Q}_1} - \sqrt{\hat{Q}_1}\right)\right], \tag{9d}$$

where $\hat{Q}_1$, $\hat{Q}_3$, $\hat{Q}_4$, $\hat{Q}_2$, $\hat{Q}_{21}$, $\hat{Q}_{43}$, $\hat{Q}_{34}$, $\hat{Q}_{12}$, $\hat{P}_{13}$, $\hat{P}_{24}$, $\hat{P}_{14}$, and $\hat{P}_{23}$ are estimates of Q_1, Q_3, Q_4, Q_2, Q_{21}, Q_{43}, Q_{34}, Q_{12}, P_{13}, P_{24}, P_{14}, and P_{23}, respectively.

Similarly, we can obtain estimates of the frequencies of these four types of nonsister gametes in 1C2D from

$$\hat{q}_2 = \frac{1}{2}\left[\left(\sqrt{\hat{Q}_{21} + \hat{Q}_1} - \sqrt{\hat{Q}_1}\right) + \left(\sqrt{\hat{Q}_{24} + \hat{Q}_4} - \sqrt{\hat{Q}_4}\right)\right], \tag{10a}$$

$$\hat{q}_3 = \frac{1}{2}\left[\left(\sqrt{\hat{Q}_{31} + \hat{Q}_1} - \sqrt{\hat{Q}_1}\right) + \left(\sqrt{\hat{Q}_{34} + \hat{Q}_4} - \sqrt{\hat{Q}_4}\right)\right], \tag{10b}$$

$$\hat{q}_1 = \frac{1}{2}\left[\left(\sqrt{\hat{Q}_1 + \hat{P}_{14} + \hat{Q}_4} - \sqrt{\hat{Q}_4}\right) + \sqrt{\hat{Q}_1}\right], \tag{10c}$$

$$\hat{q}_4 = \frac{1}{2}\left[\left(\sqrt{\hat{Q}_1 + \hat{P}_{14} + \hat{Q}_4} - \sqrt{\hat{Q}_1}\right) + \sqrt{\hat{Q}_4}\right], \tag{10d}$$

where $\hat{Q}_1$, $\hat{Q}_4$ $\hat{Q}_{21}$, $\hat{Q}_{24}$, $\hat{Q}_{31}$, $\hat{Q}_{34}$, and $\hat{P}_{14}$ are the estimated frequencies of phenotypes $a0c$, $a1c$, $A0c$, $a1C$, $a0C$, $A1c$, and $a2c$, respectively.

Three-point estimates of recombination fractions between loci: Recombination fractions between loci can be estimated from q's. Since q's are estimated separately, their sum does not always satisfy the equation $q = q_1 + q_2 + q_3 + q_4 = 0.5$. Therefore, before estimating the recombination fraction, we obtain normalized estimates of q's as

$$p_1 = \frac{\hat{q}_1}{2\hat{q}}, \quad p_3 = \frac{\hat{q}_3}{2\hat{q}}$$

$$p_2 = \frac{\hat{q}_2}{2\hat{q}}, \quad p_4 = \frac{\hat{q}_4}{2\hat{q}}.$$

It is obvious that three loci are viewed to be independent if the null hypothesis $p_1 = p_2 = p_3 = p_4$ holds at the significance level of 0.05, two loci are believed to be linked with each other, and the rest is independent if two of four types of nonsister gametes have equal estimated frequencies at the 0.05 significance level.

For linked loci, the frequencies of the four pairs of nonsister gametes can be used to distinguish the coupling phase from the repulsion phase between loci and consequently lead to proper estimates of the recombination fraction between loci according to whether they are in the coupling phase or in the repulsion phase. For example, suppose the order of the three loci is *a*–*b*–*c*. Then if p_4 is the smallest and p_1 is the largest, each pair of the three loci is in the coupling phase, and if p_4 is the largest and p_1 is the smallest, then loci *a* and *c* are in the coupling phase but loci *a* and *b* and loci *b* and *c* are in the repulsion phase. On the other hand, if p_2 is the largest and p_3 is the smallest, then loci *a* and *b* are in coupling phase but loci *a* and *c* and loci *b* and *c* are in repulsion phase. Similarly if p_2 is the smallest and p_3 is the largest, then loci *b* and *c* are in coupling phase but loci *a* and *b* and loci *a* and *c* are in repulsion.

In the coupling phase p_4 is the frequency of double crossover in the F_2 progeny. Thus, the recombination fractions between *a* and *b*, between *b* and *c*, and between *a* and *c* can be estimated by

$$\begin{aligned} r_{ab} &= 2(p_2 + p_4) \\ r_{bc} &= 2(p_3 + p_4) \\ r_{ac} &= 2(p_2 + p_3). \end{aligned} \tag{11}$$

Estimates of the recombination fractions between loci in the other orders in the coupling phase are also obtained in a similar manner.

In the repulsion phase, the order (*a–b–c*) leads to p_1 due to double crossover, and thus the recombination fractions between *a* and *b*, between *b* and *c*, and between *a* and *c* are estimated by

$$\begin{aligned} r_{ab} &= 2(p_3 + p_1) \\ r_{bc} &= 2(p_2 + p_1) \\ r_{ac} &= 2(p_2 + p_3). \end{aligned} \tag{12}$$

The recombination fractions between three loci in the other orders in the repulsion phase can be estimated in a similar fashion.

Reduction of the three-point estimates of recombination fractions to the two-point estimates: If n loci on a chromosome are genotyped in the mapping study, there are $\binom{n}{3} = n(n-1)(n-2)/6$ combinations of three loci, each of which results in three estimates of the recombination fraction. Therefore a total of $\frac{1}{2}n(n-1)(n-2)$ recombination fractions are being estimated. When n is large, it will be difficult to compare all these combinations for building a linkage map of n loci even on a modern computer. Moreover, the $\frac{1}{2}n(n-1)(n-2)$ recombination fractions contain coupling and repulsion linkage information. To avoid these complex comparisons, it is necessary to reduce the three-point estimates to two-point estimates. Although loci *i* and *j* would be configured with $n-2$ other loci to form $n-2$ three-point combinations, the linkage phase between loci *i* and *j* has already been fixed regardless of the other locus. Estimates of the recombination fraction between loci *i* and *j* may vary slightly with the other loci due to their respective different double-exchange frequencies and sampling error; hence, it needs to be adjusted with $n-2$ other loci. For convenience, let the estimate of recombination fraction between loci *i* and *j* in a three-point combination (*i, j, k*) be referred to as a three-point estimate and denoted by r_{ijk}, where *k* is called a reference locus and $k \neq i \neq j$. Thus, for n loci on a chromosome or a fragment, recombination fractions between loci *i* and *j* have $n-2$ three-point estimates. The order of loci *i, j,* and *k* in r_{ijk} has been determined previously; that is, r_{ijk} contains the order information of these three loci according to Equations 11 and 12. On the other hand, there are $n-2$ estimates of the recombination fraction between loci *i* and *j*. These $n-2$ estimates fluctuate with sampling errors and different double-exchange values, which depends upon the distances of locus *i* or/and locus *j* from locus *k*. Three cases for the variation of double-exchange values with respect to the estimate of the recombination fraction between loci *i* and *j* are considered: (1) loci *i* and *j* are adjacent loci, and all reference loci are out of interval *i–j*; (2) loci *i* and *j* are two terminal loci on a chromosome or a fragment, and all reference loci are within interval *i–j*; and (3) loci *i* and *j* are nonadjacent loci and the reference loci are either within or out of interval *i–j*. In the first case, the double exchanges dealing with all reference loci are detected and measured but different from one reference locus to another reference locus. For the second case, the double exchanges dealing with reference loci do not contribute to the recombination fraction between loci *i* and *j*. There is only one type in this case: loci *i* and *j* are two terminal loci but the $n-2$ estimates are also different with different reference loci because the double-exchange frequency is different with the reference locus; for example, a reference locus near locus *i* or *j* has less double-exchange frequency than a reference locus a distance from loci *i* and *j*. In other words, the former loses smaller double exchanges than the latter. Therefore, the former has a larger estimate value than the latter. The third case is in between the first and second cases, which is seen in the next section. Thus, the recombination fraction between loci *i* and *j* is estimated by an average estimate over $n-2$ reference loci:

$$\theta_{ij} = \frac{1}{n-2}\sum_{k=1}^{n-2} r_{ijk}. \tag{13}$$

It is obvious that θ_{ij} contains not only information of the linkage phase but also the average double-exchange frequency over all reference loci and, in addition, balances sampling errors. Therefore, θ_{ij} *is* closer to its true value than that obtained by using an EM algorithm.

AN EXAMPLE

As an example to illustrate the construction of linkage maps by MAPMAKER/EXP (version 3.0b), Lander *et al.* (1987) provided a RFLP data set of 333 F_2 mice. Since RFLP markers are codominant, A, H, and B are used in the data set for each locus to denote homozygotes of type A, heterozygotes (type H), and homozygotes of type B, respectively. To evaluate our new method, we converted these codominant marker data into dominant marker data by changing A to H and applied our new method to the dominant marker data set of the first six markers in the unknown linkage phase. Table 1 provides the estimates of the four pairs of nonsister gametes in the three-point combinations in the sample of 333 F_2 individuals. It is clear that the frequencies of the four pairs of nonsister gametes containing both loci 4 and 6 all fit the ratios of 1:1:1:1 very well, which indicates that loci 4 and 6 are independent of each other and unlinked to the other four loci. Thus, these two loci are excluded. By using Equations 11 and 12, we obtained estimates of the recombination fractions in three-point combinations (123), (125), (135), and (235). The procedure is as follows: the first step is to determine the linkage order of three loci in a combination; for example, for combination (123), $p_1 = p(R_1R_2R_3) = 0.418598 > p_2 = p(R_1D_2R_3) = 0.064757 > p_3 = p(D_1R_2R_3) = 0.011861 > p_4 = p(R_1R_2D_3) = 0.004784$ indicates that $(R_1R_2R_3)$is the parental type and $(R_1R_2D_3)$is the type

TABLE 1

Estimation of frequencies of four types of nonsister gametes

Combination: Position in combination			Frequencies of four gametes				
1	2	3	$p(R_1R_2R_3) = p_1$	$p(D_1R_2R_3) = p_2$	$p(R_1D_2R_3) = p_4$	$p(R_1R_2D_3) = p_3$	Expected ratio
1	2	3	0.418598	0.011861	0.064757	0.004784	
1	2	4	0.196761	0.016656	0.051207	0.235376	
1	2	5	0.3761	0.057600	0.006113	0.060262	
1	2	6	0.191609	0.013230	0.031609	0.263553	
1	3	4	0.233690	0.000000	0.012201	0.254109	
1	3	5	0.370070	0.007329	0.007329	0.115272	
1	3	6	0.193669	0.000000	0.010476	0.295854	
1	4	5	0.191721	0.000000	0.228865	0.079413	
1	4	6	0.147948	0.108306	0.095798	0.147948	1:1:1:1 ($p = 0.3817$)
1	5	6	0.175535	0.007168	0.051080	0.266218	
2	3	4	0.194303	0.041324	0.021653	0.242720	
2	3	5	0.416944	0.002734	0.021395	0.058926	
2	3	6	0.202325	0.017127	0.022547	0.258002	
2	4	5	0.191569	0.000000	0.262388	0.046043	
2	4	6	0.130069	0.136493	0.116719	0.116719	1:1:1:1 ($p = 0.3820$)
2	5	6	0.220542	0.000000	0.024577	0.254881	
3	4	5	0.188681	0.006188	0.236025	0.069106	
3	4	6	0.135443	0.117948	0.105580	0.141029	1:1:1:1 ($p = 0.3819$)
3	5	6	0.191077	0.018237	0.035008	0.255678	
4	5	6	0.140382	0.102766	0.154085	0.102766	1:1:1:1 ($p = 0.3817$)

R, recessive; *D*, dominant.

due to double exchange. Those remaining are recombinants where R_i and D_i, respectively, represent recessive and dominant alleles in locus i ($i = 1, 2, 3$) in a combination. These three loci have the linkage order of 1–3–2. The second step is to determine the linkage phase: since gamete $(R_1R_2R_3)$ is recessive at all three loci and has the largest frequency among these four types of nonsister gametes, we can determine that loci 1, 2, and 3 are in the coupling phase. The third step is to estimate recombination fractions in combination (123) by applying Equation 11 for the case of the coupling phase to the data in Table 1; that is,

$$r_{132} = 2 \times (0.011861 + 0.004784) = 0.0333,$$
$$r_{231} = 2 \times (0.064757 + 0.004784) = 0.1391,$$
$$r_{123} = 2 \times (0.011861 + 0.064757) = 0.1532.$$

TABLE 2

Estimation of recombination fractions by using the new method

Three-point combinations			Recombination fractions between loci		
a	*b*	*c*	*a–b*	*b–c*	*a–c*
1	2	3	0.1532	0.1391	0.0333
1	2	5	0.1274	0.1328	0.2357
1	3	5	0.0293	0.2452	0.2452
2	3	5	0.1254	0.1606	0.2005

Similarly, we also obtained estimates of the recombination fractions in combinations (125), (135), and (235) (see Table 2).

Finally, the three-point estimates of the recombination fractions were incorporated into two-point estimates by applying Equation 13 to the data in Table 2:

$$\theta_{12} = (r_{123} + r_{125})/2 = (0.1532 + 0.1274)/2 = 0.1403,$$
$$\theta_{13} = (r_{132} + r_{135})/2 = (0.0333 + 0.0293)/2 = 0.0313,$$
$$\theta_{15} = (r_{152} + r_{153})/2 = (0.2357 + 0.2452)/2 = 0.2405,$$
$$\theta_{23} = (r_{231} + r_{235})/2 = (0.1391 + 0.1254)/2 = 0.1323,$$
$$\theta_{25} = (r_{251} + r_{253})/2 = (0.1328 + 0.2005)/2 = 0.1667,$$
$$\theta_{35} = (r_{351} + r_{352})/2 = (0.2452 + 0.1606)/2 = 0.2029.$$

On the basis of the two-point estimates of recombination fractions, the best linkage map for these four loci under study was found to be 1–3–2–5, using a novel approach called the unidirectional growth method (Tan and Fu 2006), where loci 1, 2, 3, and 5 correspond to markers T175, T93, C35, and C66, respectively, in the original data set. The same linkage map (see Figure 2A) was obtained when only some of the markers were converted to dominant markers and is also the same linkage map that was obtained by MAPMAKER (at LOD = 3.0) in the original data. However, when all markers are converted to the dominant type, MAPMAKER yielded a linkage map 1–3–2–5–6–4 (at LOD = 3.0) where locus 6 corresponding to marker T209 was linked to locus 5

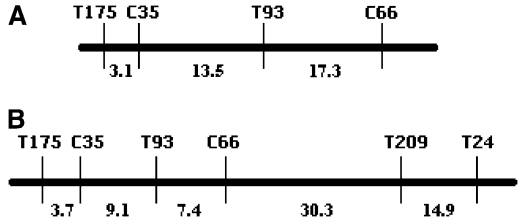

FIGURE 2.—Two linkage maps of loci built by the unidirectional growth method (TAN and FU 2006) on the basis of the new estimates of recombination fractions (A) and by MAPMARKER on the basis of the EM estimates (B), where the data of the RFLP markers provided in MAPMAKER/EXP (version 3.0, LANDER *et al.* 1987) were converted into dominant markers by replacing B with H.

(C66) at map distance 30.3 cM and locus 4 corresponding to T24 was linked to locus T209 at map distance 14.9 cM (see Figure 2B). These observations indicate that the new method leads to a better estimate of recombination than the maximum-likelihood method between dominant markers in the case of unknown phase in F_2 progeny.

SIMULATION STUDY

Since real data are not the best for fully evaluating a method because of unknown recombination fractions between loci, we used a computer simulation to generate data so that estimates of the recombination fraction can be compared to their true values. In addition to the new method, we also implemented the EM algorithm (see LIU 1998 for a detailed description of the process). To avoid potential unknown bias of a map-making method, we implemented the exhaustive search method to make maps (LIU 1998). Since the exhaustive search is extremely time consuming (MESTER *et al.* 2003b), we examined only two short linkage maps, composed of 6 and 11 dominant loci, respectively. Five map distances 10, 15, 20, 25, and 30 cM (1 cM = 1%) were randomly assigned to each adjacent interval. This setting makes it more difficult to estimate recombination fractions than in the case of a single fixed distance for all adjacent loci.

We took two cases of linkage phases into account in the simulation: (1) coupling phase (CP), 1 allelic statuses at all loci are assigned to a parental (P_1) chromosome and all 0 allelic statuses to the other parental (P_2) chromosome; and (2) unknown phase (UP), 1 or 0 allelic status at each locus is at random allocated to each of two parental chromosomes with equal probability. We used the point process crossover model (FOSS *et al.* 1993; MCPEEK and SPEED 1995) to generate recombinants. In each of F_1 meioses, recombination events occur at random between two adjacent loci. We considered both crossover-independent and complete crossover interference (but in separate simulations). For the complete crossover interference, we assumed that crossover cannot occur within an interval and between two nonsister chromatids when there is already a crossover within its adjacent interval and between the same two nonsister chromatids in the case of which the sum of distances over two adjacent intervals is ≤40 cM.

TABLE 3

Variances of estimates of recombination fractions between adjacent dominant loci in the unknown phase (UP) deviated from their respective true values in 500 simulated samples

		Sample sizes		
Methods	Adjacent loci	100	200	300
EM algorithm	1–2	0.015	0.011	0.010
	2–3	0.016	0.013	0.012
	3–4	0.019	0.015	0.013
	4–5	0.020	0.016	0.014
	5–6	0.021	0.015	0.014
New method	1–2	0.009	0.009	0.008
	2–3	0.009	0.007	0.007
	3–4	0.008	0.009	0.008
	4–5	0.010	0.009	0.009
	5–6	0.010	0.010	0.009

The expected ratio of alleles 1 and 0 for each locus is 3:1 among F_2 individuals. The simulations were carried out with sample sizes $N = 100$, 200, and 300 F_2 individuals, and loci that exhibited significant segregation distortion as revealed by chi-square test were removed. For each parameter set, 500 replicates were generated. Two criteria were used to evaluate these methods. One is the bias of the estimates of recombination fractions between two adjacent loci, which is defined as the average squared distance of the estimate to its true value, and the other is the accuracy of a method in recovering the true linkage map of given loci.

Table 3 shows the biases of estimates in the case of UP obtained by the two methods. In all the cases, the new method has a much smaller bias than the EM algorithm, which is a good indication that the new method is a better approach. However, the ultimate measure of usefulness of a method for estimating recombination fractions is to see if it leads to more accurate linkage map estimation. Table 4 summarizes the results of linkage map estimation by applying the exhaustive search method to the estimated recombination fraction data obtained by using both the EM algorithm and the new methods. It can be seen from Table 4 that both the EM and the new estimators have a very high accuracy in the case of CP even in a relatively small sample of 100 F_2 individuals. However, the new estimator has a much higher accuracy than the EM estimator in the case of UP, as expected. Furthermore, the new method improves its accuracy rapidly with sample size. It has an accuracy of 50.5% with a sample size of 100 F_2 individuals and 85.1% with a sample size of 300 F_2 individuals. The accuracy of both estimators decreases as the

TABLE 4

Efficiencies of two recombination fraction estimators in recovering the true linkage orders of 6 and 11 linked dominant loci in 500 samples generated by simulations on the basis of crossover independence

		Linkage map of 6 loci: Sample sizes			Linkage map of 11 loci: Sample sizes		
Estimators	Linkage Phases	100	200	300	100	200	300
EM algorithm	CP	92.3	97.8	100.0	86.1	98.4	100.0
	UP	15.7	22.9	23.4	5.7	5.7	6.3
New method	CP	91.4	98.1	100.0	85.9	96.9	100.0
	UP	45.6	61.4	75.7	19.86	34.1	47.6

CP, coupling phase; UP, unknown phase.

number of dominant loci increases. Table 5 shows the results of accuracy under the assumption of crossover interference. As expected, both methods have poorer performance than under the assumption of crossover interference. Although complete crossover interference in general likely occurs only between two very small adjacent intervals. The results in Table 5 suggest that crossover interference has in general a negative impact on the estimate of the recombination fraction.

DISCUSSION

We showed in this article, using both real and simulated data, that the widely used EM algorithm for estimating the recombination fraction between a pair of loci performs poorly for dominant markers because it fails to distinguish the coupling phase from the repulsion phase. We also found (results not shown) that similar to those shown in Tables 4 and 5 MAPMAKER/EXP performed poorly (<10% accuracy) for dominant markers in the unknown linkage phase, regardless whether a two-point or a three-point approach was used to estimate recombination fractions. The excellent performance of our new method may be due to several factors: (a) improved accuracy of the estimates of the gamete frequencies, (b) three-point analysis in which coupling and repulsion phases of loci are effectively distinguished, and (c) reduction of three-point estimates to two-point estimates resulting in more stable estimates of the recombination fractions.

Although the new method appears to have a shortcoming in that good accuracy of recovering true linkage maps using its estimates requires a reasonably large sample size, it does provide a promising approach that can lead to a better estimation of linkage maps from either dominant loci or mixed loci when the sample size is ~300 F_2 individuals. One likely application of the new method is to supplement the EM method. More specifically, one can apply both methods to the same data set and obtain two sets of estimates of recombination fractions. The EM estimates are used to build two partner linkage maps in which all linked loci are in the coupling phase. The new method's estimates can be used to integrate these two partner linkage maps into a single linkage map.

This study also indicates that examination of three loci at a time does provide additional information for estimating both recombination fractions and linkage maps. Since there are on the order of n^3 combinations of three loci, any approach that analyzes three loci at a time will be demanding computationally, particularly when the number of loci is large. It will be practical only when the speed of analyzing each combination of the three loci is sufficiently fast. The new method is practical even for a large number of loci since the amount of computation for each triplet of loci is minimal.

TABLE 5

Efficiencies of two recombination fraction estimators in recovering the true linkage orders of 6 and 11 linked dominant loci in 500 samples generated by simulations on the basis of crossover interference

		Linkage map of 6 loci: Sample sizes			Linkage map of 11 loci: Sample sizes		
Estimators	Linkage phases	100	200	300	100	200	300
EM algorithm	CP	86.3	95.9	97.3	75.4	91.3	97.2
	UP	17.4	23.5	27.5	4.5	5.8	5.7
New method	CP	93.2	96.9	98.4	82.1	95.7	97.8
	UP	39.2	50.3	58.2	11.3	22.8	28.6

CP, coupling phase; UP, unknown phase.

We thank the High Performance Computer Center of Yunnan University for computational support and Sara Barton for editorial assistance. This research was supported by National Institutes of Health grant R01 GM50428 (to Y.-X. F.) and by funds from Yunnan University and a 973 project (2003CB415102).

LITERATURE CITED

Allard, R. W., 1956 Formulas and tables to facilitate the calculation of recombination values in heredity. Hilgardia **24:** 235–278.

Bartel, P. L, J. A. Roecklein, D. Sengupta and S. Fields, 1996 A protein linkage map of Escherichia coli bacteriophage T7. Nat. Genet. **12:** 72–77.

Consoli, L., A. Lefevre, M. Zivy, D. de Vienne and C. Damerval, 2002 QTL analysis of proteome and transcriptome variations for dissecting the genetic architecture of complex traits in maize. Plant Mol. Biol. **48:** 575–581.

Dempster, A. P., N. M. Laird and D. B. Rubin, 1977 Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. **39B:** 1–38.

Foss, E., R. Lander, F. W. Stahl and C. M. Steinberg, 1993 Chiasma interference as a function of genetic distance. Genetics **133:** 681–691.

Knapp, S. J, J. L. Holloway, W. C. Bridges and B. H. Liu, 1995 Mapping dominant markers using F_2 mating. Theor. Appl. Genet. **91:** 74–81.

Lander, E. S., and P. Green, 1987 Construction of multilocus linkage maps in human. Proc. Natl. Acad. Sci. USA **84:** 2363–2367.

Lander, E. S., P. Green, J. Abrahamsion, A. Barlow, M. J. Daly *et al.*, 1987 MapMaker: an interactive computer package for constructing genetic linkage maps of experimental and natural populations. Genomics **1:** 174–181.

Liu, B. H., 1998 S*tatistical Genomics. Linkage, Mapping, and QTL Analysis*, pp. 163–214. CRC Press, Cleveland/Boca Raton, FL.

Mester, D. I., Y. I. Romin, Y. Hu, E. Nevo and A. B. Korol, 2003a Efficient multipoint mapping: making use of dominant repulsion-phase markers. Theor. Appl. Genet. **107:** 1102–1112.

Mester, D. I., Y. I. Romin, Y. Hu, E. Nevo and A. B. Korol, 2003b Constructing large-scale genetic maps using an evolutionary strategy algorithm. Genetics **165:** 2269–2282.

Mcpeek, M. S., and T. P. Speed, 1995 Modeling interference in genetic recombination. Genetics **139:** 1031–1044.

Ott, G., 1991 *Analysis of Human Genetic Linkage.* Johns Hopkins University Press, Baltimore/London.

Peng, J., A. Korol, T. Fahima, M. Roder, Y. Ronin *et al.*, 2000 Molecular genetic maps in wild emmer wheat, *Triticum dicoccoides*: genome-wide coverage, massive negative interference, and putative quasi-linkage. Genome Res. **10:** 1509–1531.

Tan Y.-D., and Y.-X. Fu, 2006 A novel method for estimating linkage maps. Genetics **173:** 2383–2390.

Thiellement, H., N. Bahrman, C. Damerval, C. Plomion, M. Rossignol *et al.*, 1999 Proteomics for genetic and physiological studies in plants. Electrophoresis **20:** 2013–2026.

Zivy, M, and D. de Vienne, 2000 Proteomics: a link between genomics, genetics and physiology. Plant Mol. Biol. **44:** 575–580.

Communicating editor: N. Takahata

APPENDIX A

Since $q_1 + q_2 + q_3 + q_4 = 0.5$, an alternative expression of Q_5 is

$$\begin{aligned} Q_5 &= q_3^2 + 2q_1q_3 + 2(q_3q_2 + q_3q_4 + q_2q_4) \\ &= q_3^2 + 2q_3(q_1 + q_2 + q_4) + 2q_2q_4 \\ &= q_3^2 + 2q_3(0.5 - q_3) + 2q_2q_4 \\ &= q_3 + 2q_2q_4 - q_3^2. \end{aligned} \tag{A1}$$

Similarly, we have

$$Q_6 = q_4 + 2q_2q_3 - q_4^2, \tag{A2}$$

$$Q_7 = q_2 + 2q_3q_4 - q_2^2. \tag{A3}$$

It follows that

$$\begin{aligned} Q_5 + Q_6 + Q_7 &= q_2 + q_3 + q_4 + 2(q_2q_3 + q_2q_4 + q_3q_4) - q_2^2 - q_3^2 - q_4^2 \\ &= (0.5 - q_1) + 2(q_2q_3 + q_2q_4 + q_3q_4) - q_2^2 - q_3^2 - q_4^2 \\ &= (0.5 - q_1) + Q_0 - q_2^2 - q_3^2 - q_4^2 \end{aligned} \tag{A4}$$

and

$$\begin{aligned} (0.5 - q_1)^2 &= (q_2 + q_3 + q_4)^2 \\ &= q_2^2 + q_3^2 + q_4^2 + 2(q_2q_3 + q_2q_4 + q_3q_4) \\ &= q_2^2 + q_3^2 + q_4^2 + Q_0. \end{aligned} \tag{A5}$$

Equations A4 and A5 lead to the solution for Q_0 as

$$\begin{aligned} Q_0 &= [Q_5 + Q_6 + Q_7 + (0.5 - q_1)^2 - (0.5 - q_1)] \\ &= [Q_5 + Q_6 + Q_7 + 0.25 - q_1 + q_1^2 - (0.5 - q_1)] \\ &= Q_5 + Q_6 + Q_7 + Q_1 - 0.25. \end{aligned} \tag{A6}$$