Genet. Res., Camb. (2012), **94**, *pp.* 163–177. © Cambridge University Press 2012
doi:10.1017/S0016672312000353

Estimation of recombination frequency in bi-parental genetic populations

ZIQI SUN, HUIHUI LI*, LUYAN ZHANG AND JIANKANG WANG
Institute of Crop Science, The National Key Facility for Crop Gene Resources and Genetic Improvement, Chinese Academy of Agricultural Sciences, Beijing 100081, China

(*Received 15 March 2012; revised 29 May 2012; accepted 30 May 2012*)

Summary

Linkage analysis plays an important role in genetic studies. In linkage analysis, accurate estimation of recombination frequency is essential. Many bi-parental populations have been used, and determining an appropriate population is of great importance in precise recombination frequency. In this study, we investigated the estimation efficiency of recombination frequency in 12 bi-parental populations. The criteria that we used for comparison were LOD score in testing linkage relationship, deviation between estimated and real recombination frequency, standard error (SE) of estimates and the least theoretical population size (PS) required to observe at least one recombinant and to declare the statistically significant linkage relationship. Theoretical and simulation results indicated that larger PS and smaller recombination frequency resulted in higher LOD score and smaller deviation. Lower LOD score, higher deviation and higher SE for estimating the recombination frequency in the advanced backcrossing and selfing populations are larger than those in backcross and F_2 populations, respectively. For advanced backcrossing and selfing populations, larger populations were needed in order to observe at least one recombinant and to declare significant linkage. In comparison, in F_2 and F_3 populations higher LOD score, lower deviation and SE were observed for co-dominant markers. A much larger population was needed to observe at least one recombinant and to detect loose linkage for dominant and recessive markers. Therefore, advanced backcrossing and selfing populations had lower precision in estimating the recombination frequency. F_2 and F_3 populations together with co-dominant markers represent the ideal situation for linkage analysis and linkage map construction.

1. Introduction

With the development of molecular markers, quantitative trait locus (QTL) mapping has become a routine approach for genetic studies of complex traits in plants, animals and humans, where the construction of linkage maps is a crucial step (Wu *et al.*, 2000). To construct an accurate linkage map, precisely estimating recombination frequency is the key and this has been widely studied over many years (Fisher, 1935; Haldane & Smith, 1947; Morton, 1955; Smith, 1953, 1959; Ott, 1974; Nordheim *et al.*, 1983; Ritter *et al.*, 1990; Frisch & Melchinger, 2007). In addition to linkage maps construction, recombination frequency needs to be precisely estimated for QTL fine mapping, marker-assisted backcrossing, marker-assisted selection, map-based cloning, etc.

Many factors may affect the accuracy of recombination frequency estimation. Säll & Nilsson (1994) investigated the accuracy of recombination frequency estimates with respect to (1) limited sample size, (2) heterogeneity in recombination frequency between sexes or among meioses and (3) factors that distort the segregation misclassification or differential viability. Xu & Zhou (2000) showed that linkage analysis was more reliable if the real recombination frequency between two loci was less than or equal to 0·15 when population size (PS) was 50. Hackett & Broadfoot (2003) demonstrated that missing values and/or typing errors in genotyping reduced the proportion of correctly ordered maps, and the presence of segregation distortion had little effect on marker order in the linkage maps. Frisch & Melchinger (2007) investigated the effect of mating

* Corresponding author: Institute of Crop Science, Chinese Academy of Agricultural Sciences, No. 12 Zhongguancun South Street, Beijing 100081, China. Tel: 86-10-8210 8572. E-mail: lihuihui@caas.net.cn

https://doi.org/10.1017/S0016672312000353 Published online by Cambridge University Press

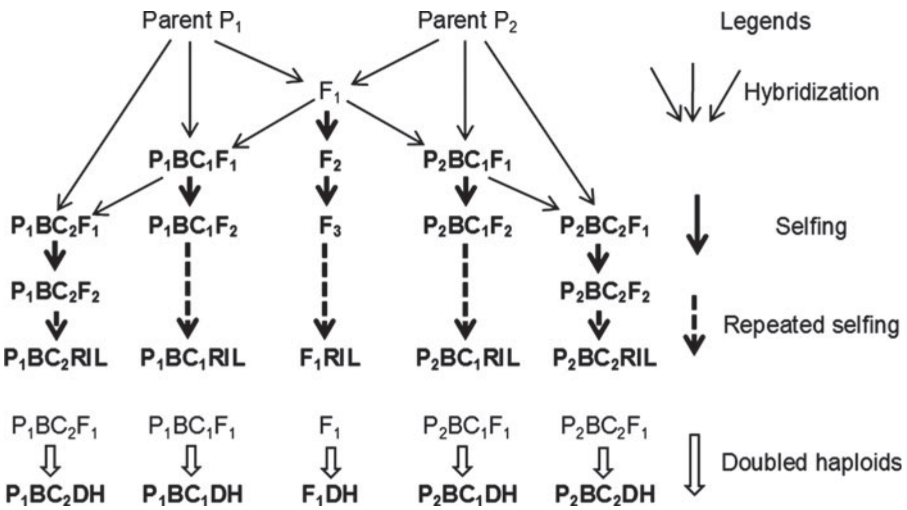

Fig. 1. Bi-parental populations commonly used in genetic studies of plants.

scheme on the precision of recombination frequency estimates.

Various types of population have been used to estimate the recombination frequency and then to construct linkage maps. For examples, types of populations in rice include F_2 (Harushima *et al.*, 1998), doubled haploids (DH; Temnykh *et al.*, 2001), recombination inbred lines (RIL; Sirithunya *et al.*, 2002), backcross (BC_1F_1; Yan *et al.*, 2003), advanced backcross (BC_3F_1; Tan *et al.*, 2008) populations, etc. In soybean, BC_1F_1 (Lu *et al.*, 2006), BC_1F_3 (Li *et al.*, 2008), RIL (Liu *et al.*, 2009), F_2 (Wang *et al.*, 2010), etc. have been used to construct linkage maps. One commonly observed problem is different populations result in inconsistent linkage maps even if the same set of markers was used in genotyping (Liang *et al.*, 2007). Antonio *et al.* (1996) compared genetic distance and the order of DNA markers in five populations of rice. The five populations consisted of three F_2, one BC_1F_1 and one RIL, which were derived from different parents. They concluded that about 170 DNA markers commonly mapped in the five populations showed the same linkage groups with conserved linkage order. However, the genetic distances between markers among the five populations were not completely consistent due to the differences in genetic backgrounds. Liang *et al.* (2007) found that the linkage maps of F_2 and F_6 populations derived from the same rice subspecies cross differed in linkage groups, linked markers, genetic orders and genetic distances.

It would be useful to know which populations were more suitable for estimating recombination frequency in order to guarantee the high efficiency of linkage analysis. Our objectives in this study were to investigate the effect of population type and size on the estimation of recombination frequency and then to determine the most suitable bi-parental populations to estimate the recombination frequency.

2. Materials and methods

(i) *Bi-parental populations in plant genetics and breeding*

Populations commonly used in plant genetics and breeding are shown in Fig. 1, which were derived from two homozygous parental lines P_1 and P_2. Backcross populations when P_1 was used as the recurrent parent had the same genetic structure as those when P_2 was used as the recurrent parent. Therefore, when P_1 was used as the recurrent parent only backcross populations were considered, i.e. 12 bi-parental populations were used to compare the precision of recombination frequency estimates in this study.

According to the frequency of P_1 alleles, these populations can be roughly classified as F_1-derived where the P_1 allele frequency was 0·5 (i.e. F_2, F_3, F_1DH and F_1RIL), BC_1F_1 and BC_1F_1-derived where the P_1 allele frequency was 0·75 (i.e. BC_1F_1, BC_1F_2, BC_1DH and BC_1RIL), and BC_2F_1 and BC_2F_1-derived where the P_1 allele frequency was 0·875 (i.e. BC_2F_1, BC_2F_2, BC_2DH and BC_2RIL). Relationship among the 12 bi-parental populations is shown in Fig. 1.

(ii) *Theoretical frequencies of genotypes at two linked loci*

Assuming that A/a and B/b were two linked marker loci in a diploid species, and the recombination frequency per meiosis was denoted as r. In the F_1RIL population which was derived from repeated selfing since F_1, recombination frequency was denoted as r_{RIL}. The relationship between r_{RIL} and r at two linked loci was $r_{\mathrm{RIL}}=2r/(1+2r)$ (Haldane & Waddington, 1931). Assuming two markers were co-dominant, frequencies of possible genotypes in each population could be expressed by r and r_{RIL} (Table 1). Double heterozygous was distinct as AB/ab for the

https://doi.org/10.1017/S0016672312000353 Published online by Cambridge University Press

Table 1. *Genotypic frequencies for 12 bi-parental populations when alleles A and a are co-dominant at marker locus A/a and alleles B and b are co-dominant at marker locus B/b*

	AABB	AABb	AAbb	AaBB	AB/ab	Ab/aB	Aabb	aaBB	aaBb	aabb	P^a
Population	n_1	n_2	n_3	n_4	n_5		n_6	n_7	n_8	n_9	
F_2	$\frac{1}{4}(1-r)^2$	$\frac{1}{2}r(1-r)$	$\frac{1}{4}r^2$	$\frac{1}{2}r(1-r)$	$\frac{1}{2}(1-r)^2$	$\frac{1}{2}r^2$	$\frac{1}{2}r(1-r)$	$\frac{1}{4}r^2$	$\frac{1}{2}r(1-r)$	$\frac{1}{4}(1-r)^2$	$r(2-\frac{3}{2}r)$
F_3	$\frac{1}{4}(1-r)+\frac{1}{8}$ $\times(1-r)^{4+\frac{1}{8}r^4}$	$\frac{1}{2}r(1-r)$ $\times(1-r+r^2)$	$\frac{1}{4}r+\frac{1}{4}r^2(1-r)^2$	$\frac{1}{2}r(1-r)$ $\times(1-r+r^2)$	$\frac{1}{4}r^4+\frac{1}{4}(1-r)^4$	$\frac{1}{2}r^2(1-r)^2$	$\frac{1}{2}r(1-r)$ $\times(1-r+r^2)$	$\frac{1}{4}r+\frac{1}{4}r^2(1-r)^2$	$\frac{1}{2}r(1-r)$ $\times(1-r+r^2)$	$\frac{1}{4}(1-r)+\frac{1}{8}(1-r)^4$ $+\frac{1}{8}r^4$	$r\left(\frac{5}{2}-\frac{7}{2}r\right.$ $\left.+3r^2-\frac{3}{2}r^3\right)$
F_1DH	$\frac{1}{2}(1-r)$		$\frac{1}{2}r$					$\frac{1}{2}r$		$\frac{1}{2}(1-r)$	r
F_1RIL	$\frac{1}{2}(1-r_{RIL})$		$\frac{1}{2}r_{RIL}$					$\frac{1}{2}r_{RIL}$		$\frac{1}{2}(1-r_{RIL})$	$r(\frac{2}{1+2r})$
BC_1F_1	$\frac{1}{2}(1-r)$	$\frac{1}{2}r$		$\frac{1}{2}r$	$\frac{1}{2}(1-r)$						r
BC_1F_2	$\frac{1}{2}-\frac{1}{4}r+\frac{1}{8}(1-r)^3$	$\frac{1}{4}r+\frac{1}{4}r(1-r)^2$	$\frac{1}{8}r+\frac{1}{8}r^2(1-r)$	$\frac{1}{4}r+\frac{1}{4}r(1-r)^2$	$\frac{1}{4}(1-r)^3$	$\frac{1}{4}r^2(1-r)$	$\frac{1}{4}r(1-r)^2$	$\frac{1}{8}r+\frac{1}{8}r^2(1-r)$	$\frac{1}{4}r(1-r)^2$	$\frac{1}{8}(1-r)^3$	$r\left(\frac{7}{4}-\frac{7}{4}r\right.$ $\left.+\frac{3}{4}r^2\right)$
BC_1DH	$\frac{1}{2}+\frac{1}{4}(1-r)^2$		$\frac{1}{2}r-\frac{1}{4}r^2$					$\frac{1}{2}r-\frac{1}{4}r^2$		$\frac{1}{4}(1-r)^2$	$r(1-\frac{r}{2})$
BC_1RIL	$\frac{1}{2}+\frac{1}{4}(1-r)$ $\times(1-r_{RIL})$		$\frac{1}{4}r+\frac{1}{4}(1-r)r_{RIL}$					$\frac{1}{4}r+\frac{1}{4}(1-r)r_{RIL}$		$\frac{1}{4}(1-r)(1-r_{RIL})$	$r(\frac{3}{2+4r})$
BC_2F_1	$\frac{1}{2}+\frac{1}{4}(1-r)^2$	$\frac{1}{2}r-\frac{1}{4}r^2$		$\frac{1}{2}r-\frac{1}{4}r^2$	$\frac{1}{4}(1-r)^2$						$r(1-\frac{r}{2})$
BC_2F_2	$\frac{5}{8}+\frac{1}{8}(1-r)^2$ $+\frac{1}{16}(1-r)^4$	$\frac{1}{8}-\frac{1}{8}(1-r)^2$ $\times(1-r+r^2)$	$\frac{1}{16}-\frac{1}{16}(1-r)^2$ $\times(1-r^2)$	$\frac{1}{8}-\frac{1}{8}(1-r)^2$ $\times(1-r+r^2)$	$\frac{1}{8}(1-r)^4$	$\frac{1}{8}r^2(1-r)^2$	$\frac{1}{8}r(1-r)^3$	$\frac{1}{16}-\frac{1}{16}(1-r)^2$ $\times(1-r^2)$	$\frac{1}{8}r(1-r)^3$	$\frac{1}{16}(1-r)^4$	$r\left(\frac{5}{4}-\frac{7}{4}r\right.$ $\left.+\frac{5}{4}r^2-\frac{3}{8}r^3\right)$
BC_2DH	$\frac{3}{4}+\frac{1}{8}(1-r)^3$		$\frac{1}{8}-\frac{1}{8}(1-r)^3$					$\frac{1}{8}-\frac{1}{8}(1-r)^3$		$\frac{1}{8}(1-r)^3$	$r\left(\frac{3}{4}-\frac{3}{4}r\right.$ $\left.+\frac{1}{4}r^2\right)$
BC_2RIL	$\frac{3}{4}+\frac{1}{8}(1-r)^2$ $\times(1-r_{RIL})$		$\frac{1}{8}-\frac{1}{8}(1-r)^2$ $\times(1-r_{RIL})$					$\frac{1}{8}-\frac{1}{8}(1-r)^2$ $\times(1-r_{RIL})$		$\frac{1}{8}(1-r)^2(1-r_{RIL})$	$r(\frac{4-r}{4+8r})$

[a] The frequency of recombinant zygotes, blanks stand for zero, and $r_{RIL}=2r/1+2r$.

coupling linkage and Ab/aB for the repulsion linkage. In this sense, genotypes in F_1 population were AB/ab. Assuming that there is no extensive selection affecting marker-allele frequencies, the frequencies of their four gametes AB, Ab, aB and ab were $(1-r)/2$, $r/2$, $r/2$ and $(1-r)/2$, respectively. Assuming independence of meiotic behaviour from generation, genotypic frequencies of BC_1F_1 and BC_2F_1 populations can be calculated from the genotypic frequency of F_1 population through backcross transition matrix (Nelson, 2011). Through selfing transition matrix (Nelson, 2011), genotypic frequencies of F_2 and F_3 populations can be calculated from the genotypic frequency of F_1 population, and genotypic frequencies of BC_1F_2 and BC_2F_2 populations can be calculated from the genotypic frequencies of BC_1F_1 and BC_2F_1 populations, respectively. Through DH-transition matrix (Nelson, 2011), genotypic frequencies of F_1DH population can be calculated from the genotypic frequency of F_1 population, and genotypic frequencies of BC_1DH and BC_2DH populations can be calculated from the genotypic frequencies of BC_1F_1 and BC_2F_1 populations, respectively. The genotypic frequencies of F_1RIL, BC_1RIL and BC_2RIL were similar to those of F_1DH, BC_1DH and BC_2DH, except that r was replaced by r_{RIL} since the repeated selfing.

It should be noted that four genotypic frequencies of BC_1F_1 and BC_2F_1 were the same as those of F_1DH and BC_1DH, respectively (Table 1). Genotypic frequencies in F_1RIL and BC_1RIL were similar to those of the corresponding backcross (BC_1F_1 and BC_2F_1) and DH (F_1DH and BC_1DH) populations, respectively, except that r was substituted by $-r_{\mathrm{RIL}}$. Due to the backcross, frequencies of parental genotype and recombinant genotype were not balanced, especially after two rounds of backcrossing. For two co-dominant loci, there were 10 genotypes for F_2, BC_1F_2, BC_2F_2 and F_3 populations, due to the coupling and repulsive linkage for double heterozygous, and four genotypes for the other eight populations (Table 1).

When marker loci A/a and B/b were not both co-dominant, frequencies of possible genotypes in populations BC_1F_1, F_2, F_3, BC_2F_1, BC_1F_2 and BC_2F_2 (Tables S1–S5 available online at http://journals.cambridge.org/GRH) could be derived by using the frequencies in Table 1. Co-dominant marker was denoted by C, dominant marker was denoted by D and recessive marker was denoted by R in Tables S1–S5. Six cases were considered including (C, C), (C, D), (C, R), (D, D), (D, R) and (R, R). (D, C), (R, C) and (R, D) were not considered, since the same results would be retained as (C, D), (C, R) and (D, R), respectively. When allele A is dominant to allele a in marker locus A/a, no polymorphism at locus A/a for populations BC_1F_1 and BC_2F_1. Hence, BC_1F_1 and BC_2F_1 were not considered in (C, D) (Table S1), (D, D) (Table S3) and (D, R) (Table S4).

(iii) *Estimation of recombination frequency*

The maximum likelihood (ML) method was used to estimate recombination frequency (Fisher, 1935; Bailey, 1961; Wu *et al.*, 2007). For any type of population, there are K observed genotype categories ($K=10$ in Table 1; $K=6$ in Tables S1 and S2; and $K=4$ in Tables S3–S5). For each observed genotype, a probability can be derived, which is a function of r. This probability is denoted by $p_k(r)$ for the kth observed genotype. Let n_k be the observed count for the kth genotype, then the multinomial log-likelihood function is, $L(r)=\sum_{k=1}^{K} n_k \log[p_k(r)]$. To obtain the ML solution of r, we need to differentiate $L(r)$ with respect to r and set the derivative equal to zero, i.e. $L'(r)=0$. It was difficult to obtain the analytic estimate of r due to the complexity of $L'(r)=0$. However, it was feasible to have the numerical solutions by applying the Newton–Raphson algorithm.

The Newton–Raphson algorithm for the ML solution of r is

$$r^{(t+1)}=r^{(t)}-\frac{L'(r^{(t)})}{L''(r^{(t)})},$$

where $L'(r)$ and $L''(r)$ are the first- and second-order derivatives of the log-likelihood function with respect to r, respectively. Assuming that the iteration process converges when $t=T$, the maximum likelihood estimation (MLE) of r is $\hat{r}=r^{(T)}$.

The variance of the estimated r is approximated by $V(\hat{r})=-1/E[L''(\hat{r})]$, where $L''(\hat{r})$ is the second derivative of the log-likelihood function with respect to r, evaluated at $\hat{r}$, and $E[L''(\hat{r})]$ is the expectation of $L''(\hat{r})$ with respect to n_k (see supplementary material).

(iv) *Test of the linkage relationship between two loci*

The existence of the linkage can be tested by the following hypotheses:

$$H_0\text{: } \hat{r}=0{\cdot}5 \text{ vs. } H_A\text{: } \hat{r}<0{\cdot}5,$$

where H_0 is the null hypothesis, corresponding to no linkage; H_A is the alternative hypothesis, corresponding to the genetic linkage between the two loci; and $\hat{r}$ is the estimated recombination frequency. The log-likelihood function under the null hypothesis is $L_0=\log[L(r=0{\cdot}5)]$, whereas the log-likelihood function under the alternative hypothesis is $L_A=\log[L(\hat{r})]$. The LOD score can be calculated from the log-likelihoods under the two hypotheses, i.e. L_A-L_0.

(v) *The least PS to observe one recombinant*

To facilitate our demonstration, the frequencies of each genotype in Table 1 and Tables S1–S5 were denoted as $f_1, f_2, \ldots, f_K$, and the probabilities of K genotypes to have recombinants observed were denoted as $p_1, p_2, \ldots, p_K$ ($K=10$ in Table 1; $K=6$ in Tables S1

https://doi.org/10.1017/S0016672312000353 Published online by Cambridge University Press

and S2; and $K=4$ in Tables S3–S5). Thus $(p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8, p_9, p_{10})=(0, 1, 1, 1, 0, 0, 1, 1, 1, 0)$ in Table 1, $(p_1, p_2, p_3, p_4, p_5, p_6)=(0, 1, 0, 1, 1, 0)$ in Table S1, $(p_1, p_2, p_3, p_4, p_5, p_6)=(0, 1, 1, 0, 1, 0)$ in Table S2, $(p_1, p_2, p_3, p_4)=(0, 1, 1, 0)$ in Tables S3 and S5, and $(p_1, p_2, p_3, p_4)=(0, 0, 1, 0)$ in Table S4. Therefore, the probability of a population to have at least one recombinant observed would be $p=\sum_{k=1}^{K} p_k f_k$. The recombinant probability (p) of each bi-parental population is shown in Table 1 and Tables S1–S5 as well. The theoretical PS required to observe at least one recombinant at the 95 % confidence level could be calculated from the equation $(1-p)^n=0{\cdot}05$, i.e. $n=\ln(0{\cdot}05)/\ln(1-p)$.

(vi) *The least PS to declare a significant linkage relationship*

With a given PS, say n, and given r, the expected observation of each genotype is equal to $n_k=f_k n$, where f_k ($k=1, \ldots, K$; $K=10$ in Table 1; $K=6$ in Tables S1 and S2; and $K=4$ in Tables S3–S5) is the frequency of each genotype in Table 1 and Tables S1–S5. The theoretical log-likelihood function under the alternative hypothesis is $L_A(r, n)=\log[L(r, n_k; k=1, \ldots, K)]$, and the log-likelihood function under the null hypothesis is $L_0=\log[L(r=0{\cdot}5)]$. Therefore, the theoretical LOD score for given r and n, is $L_A(r, n)-L_0$. For a given r and PS n, the expected LOD score can be calculated, where the least PS to test linkage, that is LOD$\geqslant 3$, can be determined.

(vii) *Simulation of the bi-parental genetic populations*

The simulated genome consisted of seven chromosomes, each with two linked markers. Only co-dominant markers were considered in simulation. The recombination frequencies between each marker pair were 0·01, 0·02, 0·03, 0·05, 0·10, 0·20 and 0·30, respectively, which were converted to marker distance by Haldane mapping function when conducting the simulation experiments. A thousand replications of each of the 12 bi-parental populations were simulated under five levels of size (PS = 50, 100, 200, 300 and 500). Simulated populations were generated by the integrated software for building linkage maps and mapping QTL, which is called QTL IciMapping (available from http://www.isbreeding.net). The LOD score and estimated recombination frequency were calculated by averaging the 1000 simulated runs.

3. Results

(i) *Comparison of LOD score in testing the linkage relationship*

LOD score is the test statistic for detecting the significance of recombination frequency (r) compared with independent inheritance. When LOD score is greater than a threshold value (generally 3), the two loci under consideration are declared to be linked; otherwise they are independent. The higher the LOD score, the more significant is the linkage between the two loci. Averaged LOD scores and their respective standard errors (SEs) of 12 bi-parental genetic populations under five levels of PS and seven levels of r are showed in Fig. 2.

It is clear that larger PS and smaller r resulted in higher LOD score, regardless of the population type (Fig. 2*a–g*). For each population type, LOD score reached the highest when PS = 500, but it declined sharply when $r=0{\cdot}1$. This indicated that large PS and small r were favourable to detect the linked markers. LOD scores of the 12 bi-parental populations were significantly different (Fig. 2*a–g*), but similar trend can be seen for different r and PS values. LOD scores of four F_1-derived populations (i.e. F_2, F_3, F_1DH and F_1RIL) were higher than those of BC_1F_1 and three BC_1F_1-derived populations (i.e. BC_1F_2, BC_1DH and BC_1RIL), respectively. BC_1F_1 and three BC_1F_1-derived populations have higher LOD scores than BC_2F_1 and three BC_2F_1-derived populations (i.e. BC_2F_2, BC_2DH and BC_2RIL), respectively. That is to say, LOD scores declined with the round of backcrossing. Taking $r=0{\cdot}01$ and PS = 300 for an example (Fig. 2*a*), LOD score of BC_2DH was lower than that of BC_1DH. LOD score of F_1DH was the highest among DH populations (i.e. F_1DH, BC_1DH and BC_2DH). The reason is that backcrossing makes the allele frequency in one locus be apart from 0·5, which is detrimental for recombination frequency estimation. LOD scores of BC_1F_1 and BC_2F_1 were similar to those of F_1DH and BC_1DH, respectively, due to the same genotypic frequencies (Table 1).

LOD scores of three RILs populations (i.e. F_1RIL, BC_1RIL and BC_2RIL) were lower than those of corresponding backcross (BC_1F_1 and BC_2F_1) and DH (F_1DH, BC_1DH and BC_2DH) populations, since the recombination frequency caused by meiosis in multiple generations ($-r_{\mathrm{RIL}}$) was greater than that by meiosis in one generation (r). LOD scores of F_2 and F_2-related populations (i.e. F_3, BC_1F_2 and BC_2F_2) were higher than those of the corresponding backcross (i.e. BC_1F_1 and BC_2F_1), DH (i.e. F_1DH, BC_1DH and BC_2DH) and recombinant inbred lines (i.e. F_1RIL, BC_1RIL and BC_2RIL). The possible reason is that F_2 and F_2-related populations had 10 genotypes, whereas other populations had only four genotypes (Table 1), and more genotypes might provide more recombination information. In most cases, LOD scores of F_2 population were the highest, except when $r=0{\cdot}01$, PS = 100–500 (Fig. 2*a*) and $r=0{\cdot}02$, PS = 500 (Fig. 2*b*) where LOD scores of F_3 population were the highest.

https://doi.org/10.1017/S0016672312000353 Published online by Cambridge University Press

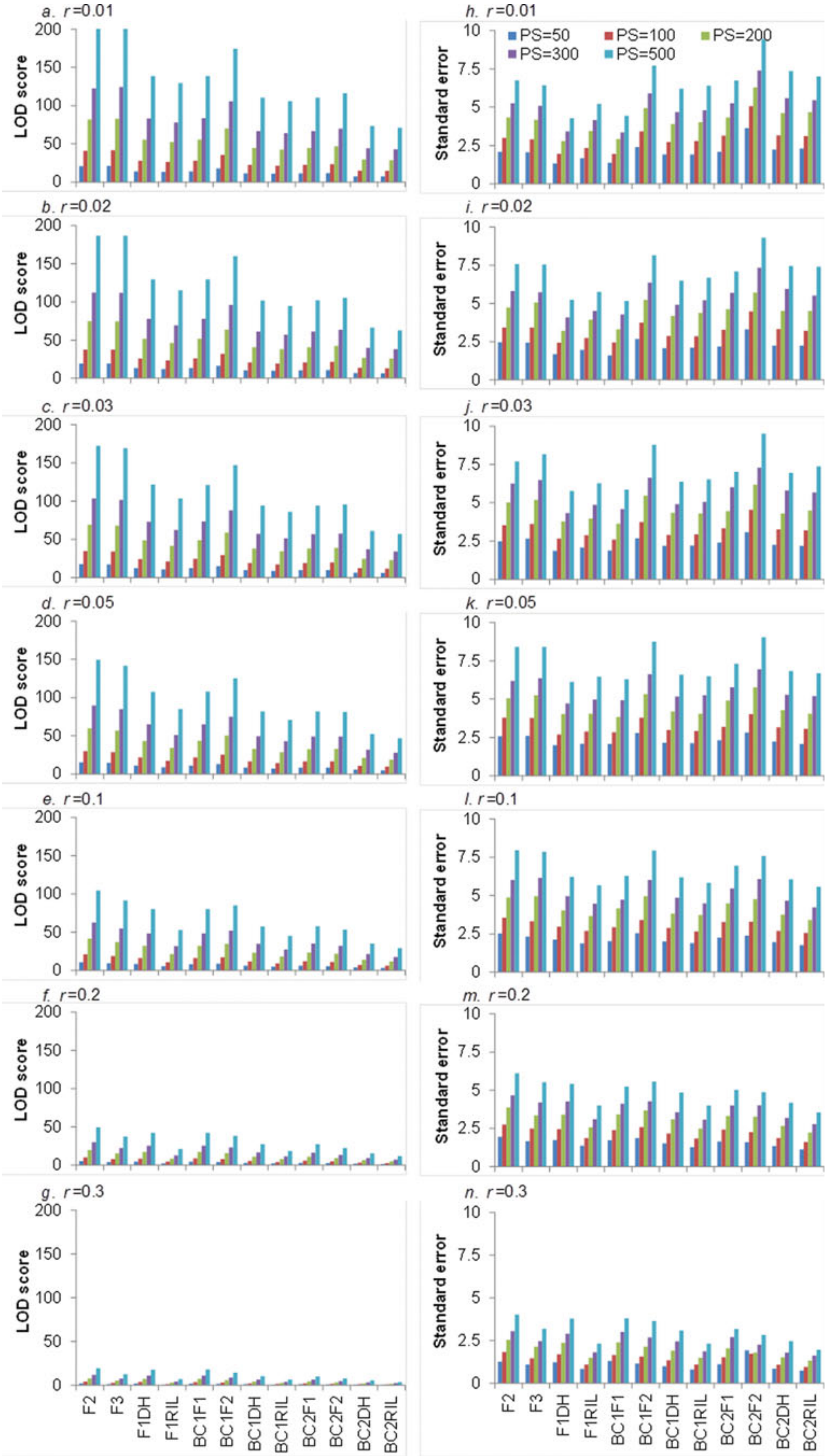

Fig. 2. Average LOD scores (*a*–*g*) and their respective SEs (*h*–*n*) from 1000 simulations in 12 bi-parental populations corresponding to seven levels of recombination frequencies ($r = 0{\cdot}01$, 0·02, 0·03, 0·05, 0·1, 0·2 and 0·3) and five PS (PS = 50, 100, 200, 300 and 500). Only co-dominant markers were considered.

https://doi.org/10.1017/S0016672312000353 Published online by Cambridge University Press

Higher PS resulted in higher SE of LOD scores, which was similar to the trend of LOD score (Fig. 2*h*–*n*). For the increased rounds of backcrossing and *r*, SE was generally decreased, due to the decreased LOD score. However, when $r \leqslant 0{\cdot}05$ the SE of LOD scores slightly increased as the round of backcrossing and *r* increased (Fig. 2*h*–*k*). The SE of F_2 population slightly increased from 6·767 ($r = 0{\cdot}01$) to 8·408 ($r = 0{\cdot}05$), and then decreased to 4·018 ($r = 0{\cdot}3$).

(ii) *Comparison of accuracy in estimating recombination frequency*

Absolute deviation between estimated and real recombination frequencies and their respective SE under seven levels of *r* and five levels of PS are shown in Fig. 3. Small deviation and SE implied accurate recombination frequency estimation. As expected, the deviation and SE of estimated recombination frequency decreased with the increase of PS. When $PS \geqslant 200$ the deviation and SE for the 12 populations were almost equal to zero. It indicated that increasing sample sizes always led to high precision of estimation. Generally, the deviations became large with the increase of *r* (Fig. 3*e*–*g*). For $r = 0{\cdot}01$, the deviations were close to zero except for those under BC_2F_2 population (Fig. 3*a*). When *r* increased, the deviations became significantly large (Fig. 3*e*–*g*). In terms of the population type, deviations for F_2 population were almost equal to zero regardless of PS and *r*. The deviations of BC_2F_1-related populations (i.e. BC_2F_1, BC_2DH, BC_2RIL and BC_2F_2) were generally higher than those of other eight populations (Fig. 3*a*–*g*), which indicated that more rounds of backcrossing were not favoured for precision estimation of recombination frequency. SE of estimated recombination frequencies increased with the increase of *r* and rounds of backcrossing, which were similar to the trends in deviations (Fig. 3*h*–*n*). Among the 12 populations, SE of F_2 population was generally the smallest.

To further investigate the effect of population type on the precision of recombination frequency estimation, we compared the theoretical SE under seven levels of *r* and $PS = 50$ in 12 bi-parental populations (Table 2). The theoretical SE were calculated from second derivative of the log-likelihood function $E[L''(\hat{r})]$ (for details, see Materials and methods section) that is $\sqrt{V_{\hat{r}} = -1/E[L''(\hat{r})]}$, where $\hat{r}$ is the estimate of *r* and *n* is the sample size. The theoretical SE when two markers were co-dominant were consistent with the averaged SE obtained from 1000 simulation runs, due to the large sample property of population mean (Table 2 and Fig. 3). The theoretical SE increased with the increase in *r* and rounds of backcrossing (Table 2), which was consistent with what we had seen in the simulated results (Fig. 3). SE of BC_1F_1 and BC_2F_1 were equal to those of F_1DH and BC_1DH (Table 2 and Fig. 3), respectively, due to the same genotypic frequencies when two markers were co-dominant (Table 1). The theoretical SE of BC_2RIL was the highest among the 12 populations regardless of the level of *r*. When $r \leqslant 0{\cdot}1$ the theoretical SE of F_3 was the lowest, whereas when $r > 0{\cdot}1$ those of F_2 was the lowest.

To evaluate the effect of dominant and recessive markers on the estimation of recombination frequency, we considered dominant and recessive markers in populations F_2, F_3, BC_1F_1, BC_1F_2, BC_2F_1 and BC_2F_2. Generally, theoretical SE increased when markers were dominant and/or recessive, since fewer genotypes were observed (Tables S1–S5). The advantage of using co-dominant markers became obvious when *r* was getting large. SEs under (D, R) were the largest (Table 2) among the six cases (C, C), (C, D), (C, R), (D, D), (D, R) and (R, R) for populations F_2, F_3, BC_2F_1 and BC_2F_2. The reason is that only one recombinant genotype can be observed under (D, R), whereas in the other three genotype groups we cannot distinguish recombinant and non-recombinant from observations (Table S4).

(iii) *PS required to observe at least one recombinant and to declare the significant linkage relationship*

To detect the linkage between two loci, we need the PS to be large enough to guarantee that (1) at least one recombinant can be observed for tight linkage (Table 3) and (2) LOD score is greater than a threshold of 3 for loose linkage (Table 4). Small *r* implied a tight linkage between two loci. From the results in previous sections, LOD score for tight linkage was always high, which indicated that there was more chance to detect linkage, while recombinant was difficult to occur. In this sense, for a specific population type with *r* increased, PS required to make at least one recombinant observed decreased sharply (Table 3), while to make the linkage statistically significant increased conspicuously (Table 4). Therefore, the maximum value of the two PS in Tables 3 and 4 should be used in the application. Taking BC_1F_1 under two co-dominant markers for example, we needed 299 individuals to be 95 % sure to observe at least one recombinant for $r = 0{\cdot}01$, while 11 individuals were enough to make it statistically significant. Therefore, when we develop a BC_1F_1 for linkage analysis, at least 299 individuals should be included.

For a specific level of *r*, the least PS required for observing at least one recombinant and to detect the linkage for BC_2F_1-derived populations was always large. Taking $r = 0{\cdot}03$ and case (C, C) as an example, the least PS required to observe at least one recombinant was 136 for BC_2DH (Table 3), which was the largest among the 12 populations if markers were

https://doi.org/10.1017/S0016672312000353 Published online by Cambridge University Press

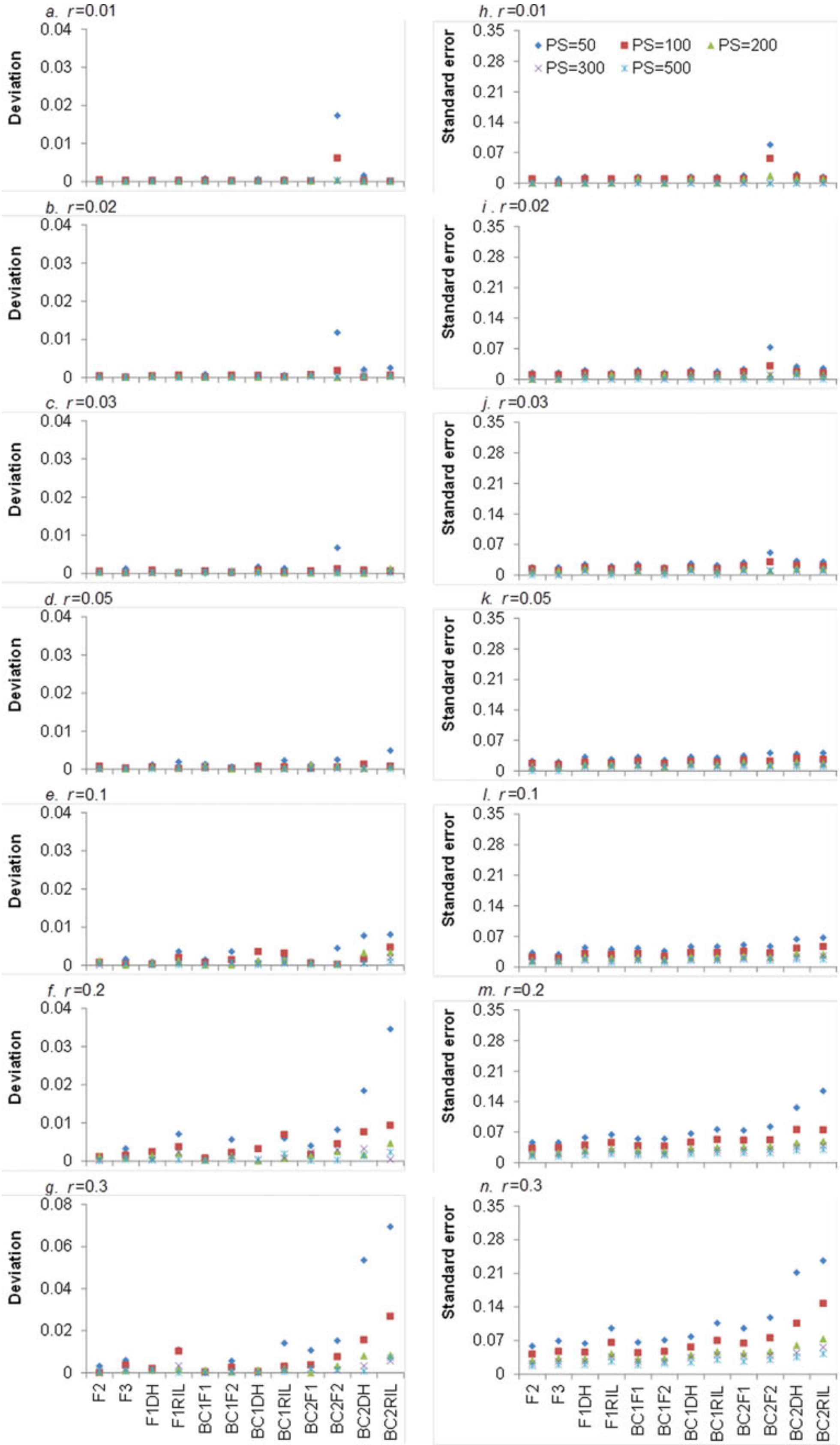

Fig. 3. Average deviations between estimated recombination frequencies and real recombination frequencies (*a*–*g*) and SEs of estimating recombination frequencies (*h*–*n*) from 1000 simulations in 12 bi-parental populations corresponding to seven levels of recombination frequencies ($r = 0{\cdot}01$, 0·02,0·03, 0·05, 0·1, 0·2 and 0·3) and five PS (PS = 50, 100, 200, 300 and 500). Only co-dominant markers were considered.

https://doi.org/10.1017/S0016672312000353 Published online by Cambridge University Press

Table 2. *Theoretical SEs under seven levels of recombination frequency in 12 bi-parental populations when PS=50*

Population[a]	$r=0{\cdot}01$	$r=0{\cdot}02$	$r=0{\cdot}03$	$r=0{\cdot}05$	$r=0{\cdot}1$	$r=0{\cdot}2$	$r=0{\cdot}3$
F_2 (C, C)	0·0100	0·0141	0·0173	0·0224	0·0318	0·0457	0·0574
F_2 (C, D)	0·0141	0·0200	0·0244	0·0315	0·0443	0·0623	0·0751
F_2 (C, R)	0·0141	0·0200	0·0244	0·0315	0·0443	0·0623	0·0751
F_2 (D, D)	0·0142	0·0200	0·0246	0·0318	0·0451	0·0646	0·0801
F_2 (D, R)	0·1414	0·1414	0·1413	0·1410	0·1397	0·1347	0·1270
F_2 (R, R)	0·0142	0·0200	0·0246	0·0318	0·0451	0·0646	0·0801
F_3 (C, C)	0·0090	0·0128	0·0158	0·0207	0·0304	0·047	0·064
F_3 (C, D)	0·0116	0·0165	0·0204	0·0266	0·039	0·059	0·0776
F_3 (C, R)	0·0116	0·0165	0·0204	0·0266	0·039	0·059	0·0776
F_3 (D, D)	0·0116	0·0166	0·0204	0·0268	0·0392	0·0599	0·0798
F_3 (D, R)	0·0276	0·0381	0·0456	0·0565	0·0736	0·0935	0·108
F_3 (R, R)	0·0116	0·0166	0·0204	0·0268	0·0392	0·0599	0·0798
F_1DH	0·0141	0·0198	0·0241	0·0308	0·0424	0·0566	0·0648
F_1RIL	0·0102	0·0147	0·0184	0·0246	0·0379	0·0626	0·0876
BC_1F_1 (C, C)	0·0141	0·0198	0·0241	0·0308	0·0424	0·0566	0·0648
BC_1F_1 (C, R)	0·0141	0·0198	0·0241	0·0308	0·0424	0·0566	0·0648
BC_1F_1 (R, R)	0·0141	0·0198	0·0241	0·0308	0·0424	0·0566	0·0648
BC_1F_2 (C, C)	0·0107	0·0152	0·0187	0·0243	0·0351	0·0519	0·0666
BC_1F_2 (C, D)	0·0164	0·0234	0·0288	0·0376	0·0547	0·0817	0·1036
BC_1F_2 (C, R)	0·0127	0·0180	0·0220	0·0286	0·0408	0·0589	0·0734
BC_1F_2 (D, D)	0·0164	0·0234	0·0288	0·0377	0·0552	0·0836	0·1100
BC_1F_2 (D, R)	0·0391	0·0540	0·0648	0·0802	0·1032	0·1248	0·1350
BC_1F_2 (R, R)	0·0127	0·0180	0·0221	0·0286	0·0409	0·0592	0·0739
BC_1DH	0·0142	0·0200	0·0246	0·0318	0·0451	0·0646	0·0801
BC_1RIL	0·0118	0·0170	0·0212	0·0284	0·0436	0·0708	0·0965
BC_2F_1 (C, C)	0·0142	0·0200	0·0246	0·0318	0·0451	0·0646	0·0801
BC_2F_1 (C, R)	0·0142	0·0200	0·0246	0·0318	0·0451	0·0646	0·0801
BC_2F_1 (R, R)	0·0142	0·0200	0·0246	0·0318	0·0451	0·0646	0·0801
BC_2F_2 (C, C)	0·0127	0·0182	0·0224	0·0294	0·0434	0·0669	0·0896
BC_2F_2 (C, D)	0·0208	0·0307	0·0391	0·0547	0·0939	0·1841	0·2635
BC_2F_2 (C, R)	0·0142	0·0203	0·0250	0·0327	0·0478	0·0725	0·0948
BC_2F_2 (D, D)	0·0202	0·0290	0·0359	0·0475	0·0713	0·1137	0·1573
BC_2F_2 (D, R)	0·0394	0·0549	0·0663	0·0833	0·1114	0·1453	0·1697
BC_2F_2 (R, R)	0·0142	0·0203	0·0250	0·0327	0·0479	0·0726	0·0951
BC_2DH	0·0164	0·0234	0·0288	0·0377	0·0552	0·0836	0·1100
BC_2RIL	0·0288	0·0414	0·0515	0·0684	0·1032	0·1597	0·2039

[a] Assuming A and a are the two marker alleles at one locus, B and b are the two marker alleles at the other locus. (C, C) denotes that A and a are co-dominant, and B and b are co-dominant; (C, D) denotes that A and a are co-dominant, and B is dominant to b; (C, R) denotes that A and a are co-dominant, and B is recessive to b; (D, D) denotes that A is dominant to a, and B is dominant to b; (D, R) denotes that A is dominant to a, and B is recessive to b; and (R, R) denotes that A is recessive to a, and B is recessive to b.

co-dominant. The least PS required to detect the linkage was 27 for BC_2RIL (Table 4), which was the largest among the 12 populations. The results indicated that more rounds of backcrossing were not favoured in linkage analysis.

In terms of marker type, similar trends can be seen in Tables 3 and 4 as those in Table 2. The least PS required to observe one recombinant and to detect linkage increased when dominant and recessive markers were considered, especially when r is large. Under (D, R), a huge population was needed to observe one recombinant and to detect linkage (Tables 3 and 4). The required PS for F_2 and F_3 populations under (C, C) (Tables 3 and 4) was always the smallest regardless of the level of r, which showed the advantages of using F_2 and F_3 to estimate the recombination frequency for co-dominant markers.

Furthermore, we calculated the ratios (p/r) of the frequency of recombinant zygotes (p; Table 1 and Tables S1–S5) and recombination frequency (r) for two linked loci. $p/r>1$ implies that r was magnified by the estimated proportion of recombinant zygotes in the corresponding population; $p/r=1$ implies that r was unbiased by the estimated proportion of recombinant zygotes in the corresponding population; and $p/r<1$ implies that r was reduced by the estimated proportion of recombinant zygotes in the corresponding population. When r was magnified, it will be easy to observe the recombinant, and detect the linkage. p/r under two co-dominant markers are shown in

https://doi.org/10.1017/S0016672312000353 Published online by Cambridge University Press

Table 3. *Theoretical PS required to have at least one recombinant observed under 95% confidence level for seven levels of recombination frequencies in 12 bi-parental populations*

Population[a]	$r = 0{\cdot}01$	$r = 0{\cdot}02$	$r = 0{\cdot}03$	$r = 0{\cdot}05$	$r = 0{\cdot}1$	$r = 0{\cdot}2$	$r = 0{\cdot}3$
F_2 (C, C)	150	75	50	30	15	8	5
F_2 (C, D)	299	149	99	60	31	16	11
F_2 (C, R)	299	149	99	60	31	16	11
F_2 (D, D)	299	149	99	61	31	16	11
F_2 (D, R)	149 786	29 956	13 616	4754	1197	299	132
F_2 (R, R)	299	149	99	61	31	16	11
F_3 (C, C)	121	61	41	25	13	7	5
F_3 (C, D)	199	99	67	41	21	11	8
F_3 (C, R)	199	99	67	41	21	11	8
F_3 (D, D)	213	99	67	41	21	11	8
F_3 (D, R)	998	598	373	229	110	52	34
F_3 (R, R)	213	99	67	41	21	11	8
F_1DH	299	149	99	59	29	14	9
F_1RIL	152	77	52	32	17	9	7
BC_1F_1 (C, C)	299	149	99	59	29	14	9
BC_1F_1 (C, R)	299	149	99	59	29	14	9
BC_1F_1 (R, R)	299	149	99	59	29	14	9
BC_1F_2 (C, C)	172	86	58	35	18	9	7
BC_1F_2 (C, D)	427	199	135	82	43	24	17
BC_1F_2 (C, R)	249	119	80	48	24	12	8
BC_1F_2 (D, D)	373	213	135	82	43	24	17
BC_1F_2 (D, R)	2995	998	748	427	213	102	66
BC_1F_2 (R, R)	249	124	82	49	24	12	8
BC_1DH	300	150	100	60	31	16	11
BC_1RIL	203	103	70	43	23	13	10
BC_2F_1 (C, C)	300	150	100	60	31	16	11
BC_2F_1 (C, R)	300	150	100	60	31	16	11
BC_2F_1 (R, R)	300	150	100	60	31	16	11
BC_2F_2 (C, C)	242	122	82	50	27	15	11
BC_2F_2 (C, D)	748	332	213	129	70	39	31
BC_2F_2 (C, R)	299	157	99	61	32	17	12
BC_2F_2 (D, D)	748	299	213	124	70	39	31
BC_2F_2 (D, R)	2995	1497	748	498	249	124	85
BC_2F_2 (R, R)	299	149	99	61	32	17	12
BC_2DH	403	203	136	83	43	24	17
BC_2RIL	305	156	106	66	36	21	16

[a] Assuming A and a are the two marker alleles at one locus, B and b are the two marker alleles at the other locus. (C, C) denotes that A and a are co-dominant, and B and b are co-dominant; (C, D) denotes that A and a are co-dominant, and B is dominant to b; (C, R) denotes that A and a are co-dominant, and B is recessive to b; (D, D) denotes that A is dominant to a, and B is dominant to b; (D, R) denotes that A is dominant to a, and B is recessive to b; and (R, R) denotes that A is recessive to a, and B is recessive to b.

Fig. 4. p/r decreased, while all were greater than 1 under populations F_3, F_2, F_1RIL, BC_1F_2, BC_1RIL and BC_2F_2. Thus, the power to estimate r was high, and small samples were needed to observe recombinant and detect the linkage in those populations, especially in F_3 population, which were consistent with the results we have seen in Fig. 2, and Tables 3 and 4. $p/r < 1$ under populations BC_1DH, BC_2F_1, BC_2RIL and BC_2DH, which indicated that these four populations were inappropriate to estimate r. p/r under BC_2F_1-derived populations were smaller than those under F_1-derived populations and BC_1F_1-derived populations, which indicated the inferior r estimation from two rounds of backcrossing populations once again. On the other hand, p/r was high when r was small, say $r < 0{\cdot}1$, which was consistent with the results in Fig. 2, and Tables 3 and 4. The advantage of using F_3 population was weakened to that of F_2 population when $r > 0{\cdot}25$.

It should be noted that p/r under F_2 populations across different r was higher than those under RIL population, but lower than those under F_3 population (Fig. 4). In addition to the 12 bi-parental populations described in Fig. 1, one more selfing pollination after F_3 (i.e. F_4), BC_1F_2 (i.e. BC_1F_3) and BC_2F_2 (i.e. BC_2F_3) were included to evaluate the trend of p/r for repeated selfing. It turns out that from F_4, BC_1F_3 and BC_2F_3 populations, p/r became lower and lower, and approached RIL, BC_1RIL and BC_2RIL populations, respectively, which indicated the less efficiency of

https://doi.org/10.1017/S0016672312000353 Published online by Cambridge University Press

Table 4. *Theoretical PS required to detect linkage between two loci (LOD ⩾ 3) for seven levels of recombination frequencies in 12 bi-parental populations*

Population[a]	$r=0{\cdot}01$	$r=0{\cdot}02$	$r=0{\cdot}03$	$r=0{\cdot}05$	$r=0{\cdot}1$	$r=0{\cdot}2$	$r=0{\cdot}3$
F_2 (C, C)	8	9	9	11	15	31	78
F_2 (C, D)	14	15	16	19	26	51	123
F_2 (C, R)	14	15	16	19	26	51	123
F_2 (D, D)	14	15	16	19	27	56	147
F_2 (D, R)	82	83	83	86	96	138	262
F_2 (R, R)	14	15	16	19	27	56	147
F_3 (C, C)	8	9	9	11	17	41	121
F_3 (C, D)	12	13	15	17	26	58	162
F_3 (C, R)	12	13	15	17	26	58	162
F_3 (D, D)	12	14	15	17	26	62	179
F_3 (D, R)	31	33	35	39	52	100	246
F_3 (R, R)	12	14	15	17	26	62	179
F_1DH	11	12	13	14	19	36	84
F_1RIL	12	14	15	18	29	73	219
BC_1F_1 (C, C)	11	12	13	14	19	36	84
BC_1F_1 (C, R)	11	12	13	14	19	36	84
BC_1F_1 (R, R)	11	12	13	14	19	36	84
BC_1F_2 (C, C)	9	10	11	12	18	40	107
BC_1F_2 (C, D)	21	23	25	29	42	90	236
BC_1F_2 (C, R)	12	13	14	16	23	49	125
BC_1F_2 (D, D)	21	23	25	29	44	101	289
BC_1F_2 (D, R)	54	57	59	65	84	150	343
BC_1F_2 (R, R)	12	13	14	16	23	49	128
BC_1DH	14	15	16	19	27	56	147
BC_1RIL	15	16	18	22	34	83	238
BC_2F_1 (C, C)	14	15	16	19	27	56	147
BC_2F_1 (C, R)	14	15	16	19	27	56	147
BC_2F_1 (R, R)	14	15	16	19	27	56	147
BC_2F_2 (C, C)	13	15	16	19	29	68	199
BC_2F_2 (C, D)	34	37	41	48	72	166	469
BC_2F_2 (C, R)	17	18	20	23	34	78	218
BC_2F_2 (D, D)	34	38	41	49	76	193	606
BC_2F_2 (D, R)	66	70	75	84	114	229	585
BC_2F_2 (R, R)	17	18	20	23	34	79	220
BC_2DH	21	23	25	29	44	101	289
BC_2RIL	22	24	27	33	52	133	406

[a] Assuming A and a are the two marker alleles at one locus, B and b are the two marker alleles at the other locus. (C, C) denotes that A and a are co-dominant, and B and b are co-dominant; (C, D) denotes that A and a are co-dominant, and B is dominant to b; (C, R) denotes that A and a are co-dominant, and B is recessive to b; (D, D) denotes that A is dominant to a, and B is dominant to b; (D, R) denotes that A is dominant to a, and B is recessive to b; and (R, R) denotes that A is recessive to a, and B is recessive to b.

advanced selfing populations in estimating recombination frequency.

4. Discussion and conclusion

Linkage analysis is fundamental in genetic studies. An accurate estimation of recombination frequency is essential for gene mapping, marker-assisted selection, map-based cloning, etc. Bi-parental populations have been widely used to estimate the recombination frequency in the application, but few studies can be identified in which populations were more suitable for estimating recombination frequency. In this study, 12 bi-parental populations were considered to investigate the efficiency in estimating the recombination frequency in theory and by extensive simulations. Actually, those 12 populations included most of the commonly used mating schemes of bi-parental populations in genetics and breeding.

Regarding the detection power, we compared the LOD scores under five levels of PS and seven levels of recombination frequency across 12 bi-parental populations. In terms of estimation precision, we compared deviations between estimated and real recombination frequencies under five PS, and the theoretical SEs under seven levels of recombination frequency across 12 bi-parental populations. Theoretically, we evaluated the least PS needed to observe at least one recombinant and to make the linkage detected (i.e. LOD score was not less than 3) for

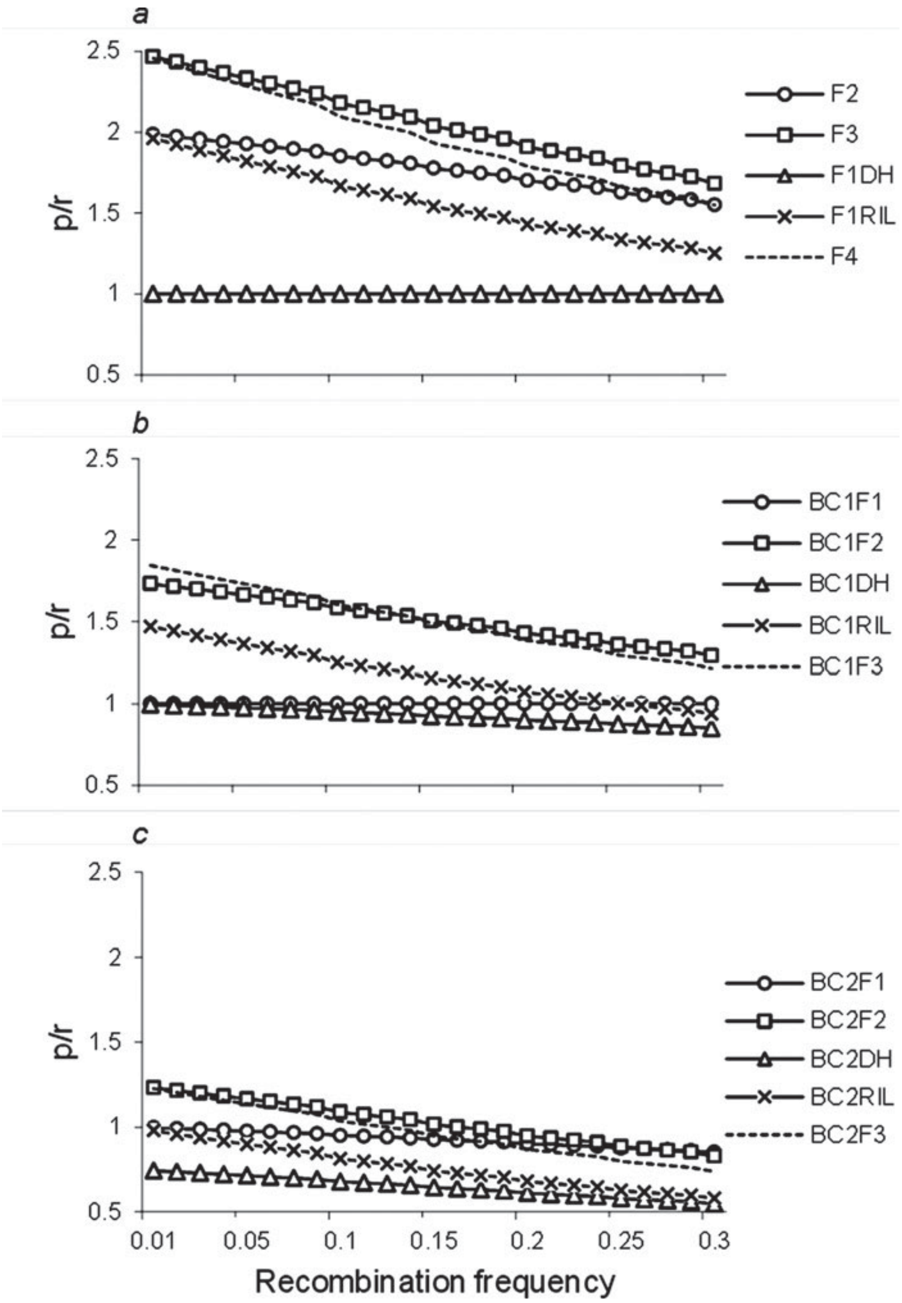

Fig. 4. Ratio (p/r) of the frequency of recombinant zygotes (p) and the recombination frequency (r) of two linked co-dominant marker loci. In addition to the 12 bi-parental populations described in Fig. 1, F_4, BC_1F_3 and BC_2F_3 were included to evaluate the trend of p/r for repeated selfing.

co-dominant, dominant and recessive markers. In summary, detection power and estimation precision of two tightly linked loci (i.e. small recombination frequency) were high. Large size of population is critical for recombination frequency estimation. Advanced backcrossing and selfing populations reduce the precision in estimating the recombination frequency. For dominant and recessive markers, large-sized populations were needed to observe at least one recombinant and detect the genetic linkage, indicating that dominant and recessive markers should be avoided as much as possible in recombination frequency estimation. For the 12 populations considered in this study, F_2 and F_3 populations with co-dominant markers had the highest power and precision to estimate the recombination frequency.

For co-dominant markers, F_2 and F_3 populations showed advantages over backcross (i.e. BC_1F_1 and BC_2F_1), DH (i.e. F_1DH, BC_1DH and BC_2DH) and recombinant inbred lines (i.e. F_1RIL, BC_1RIL and BC_2RIL) (Figs. 2–4) on recombination frequency estimation. F_2 and F_3 populations had 10 genotypes, whereas backcross, DH and recombinant inbred lines had only four genotypes and no heterozygous

https://doi.org/10.1017/S0016672312000353 Published online by Cambridge University Press

(Table 1). For F_2 and F_3 populations, the theoretical standard deviations of estimated recombination frequency (Table 2), and the least PS needed to observe at least one recombinant (Table 3) and to make the linkage detected (Table 4) for co-dominant markers were smaller than those for dominant and recessive markers. Compared with co-dominant marker, using dominant and recessive markers merged some genotypes together, and cannot distinguish recombinant and non-recombinant from observations in some cases, which decreased the efficiency in recombination frequency estimation (Tables S1–S5). For example, for F_2 and F_3 populations under (C, C) (Table 1), there were 10 genotypes where AABb, AAbb, AaBB, Aabb, aaBB and aaBb were observed recombinant, and under (D, R) (Table S4) there were four genotypes where only aaBB can be observed as recombinant. Therefore, F_2 and F_3 populations with co-dominant markers have more genotypes than those with dominant and recessive markers, and those under backcross, DH and recombinant inbred lines.

For BC_1F_1 and BC_2F_1 the number of genotype using co-dominant markers (i.e. (C, C); Table 1) was the same as that using dominant and recessive markers (Tables S2 and S5). The theoretical least PS needed to observe at least one recombinant (Table 3) and to make the linkage detected were the same under (C, C), (C, R) and (R, R). It implied that marker type did not play an effect on recombination frequency estimation (Tables 3 and 4) for backcross populations. However, backcrossing made the allele frequency in one locus apart from 0·5, thus genotypic frequency among each genotype was greatly different, especially for multiple rounds of backcrossing. This difference was detrimental for recombination frequency estimation as well (Figs. 2–4 and Tables 2–4). Taking $r = 0{\cdot}01$ for example, the theoretical SE for the estimates of recombination frequency for F_2, BC_1F_2 and BC_2F_2 populations using co-dominant markers were 0·0100, 0·0107 and 0·0127, respectively (Table 2). The theoretical PS required to have at least one recombinant observed for F_2, BC_1F_2 and BC_2F_2 populations using co-dominant markers were 150, 172 and 242, respectively (Table 3). The theoretical PS required for detecting linkage for F_2, BC_1F_2 and BC_2F_2 populations using co-dominant markers were 8, 9, and 13, respectively (Table 4).

Few advantages for advanced selfing populations to estimate recombination frequency were observed (Figs 2–4 and Tables 2–4). Taking $r = 0{\cdot}01$ for example, the theoretical SE for the estimates of recombination frequency for F_2 and F_1RIL, and BC_1F_2 and BC_1RIL populations using co-dominant markers were 0·0100 and 0·0102, and 0·0107 and 0·0118, respectively (Table 2). The theoretical PS is required to have at least one recombinant observed for F_2 and F_1RIL, and BC_1F_2 and BC_1RIL populations using co-dominant markers were 150 and 152, and 172 and 203, respectively (Table 3). The theoretical PS required to detect linkage for F_2 and F_1RIL, and BC_1F_2, and BC_1RIL populations using co-dominant markers were 8 and 12, and 9 and 15, respectively (Table 4). Therefore, the advanced backcrossing and selfing populations were less efficient to estimate recombination frequency than backcross and F_2 populations, respectively. Consequently, for a given PS F_2 and F_3 populations with co-dominant markers would have more information on recombination, and represent the ideal situation for recombination frequency estimation.

Bi-parental populations derived by random intermating from F_2 generation, such as advanced intercross line (AIL; Darvasi & Soller, 1995) and Intermated B73 × Mo17 (IBM; Lee *et al.*, 2002) population, were not considered in this study. However, the basic principles and conclusions shown in this paper should apply to these populations as well. Continued intercrossing of a population would reduce linkage disequilibrium and cause the proportion of recombinants between any linked loci to asymptotically approach 0·5. Beavis *et al.* (1992) showed that the recombination on chromosome 1 in maize, a 2·7-fold increase in the recombination frequency was observed in the IBM population after five generations of intermating. The expected recombination frequencies after *t* generations of random mating were calculated as $r_{(t)} = r_{(t-1)} + [r_{(t-1)} - 2r^2_{(t-1)}]/2$, where $r_{(t)}$ is the frequency of recombinants after *t* generations of random mating and $r_{(t-1)}$ is the frequency of recombinants in the prior generation (Beavis *et al.*, 1992). Genotypic frequencies in AIL or IBM population can be determined by random mating transition matrix (Falconer & Mackay, 1996). Therefore, similar strategies as those used in this study can be utilized to evaluate the efficiency of AIL and IBM populations on recombination frequency estimation.

In this study, we evaluated the efficiency of 12 bi-parental populations to estimate the recombination frequency, and concluded that F_2 and F_3 populations with co-dominant markers would be superior to the other 10 bi-parental populations in estimating the recombination frequency. It should be noted that advanced selfing, intercross and intermated populations have ideal properties for genetic study as well. For example, each line in these populations is homozygous so there is no within-line genetic variance; and they can be easily bulked and assessed in multiple sites and seasons in replicated trials, etc. So phenotype can be measured precisely; and genotype by environment interactions can be studied. Therefore, these populations have advantages in identifying genotype to phenotype relationship.

We have implemented the recombination frequency estimation of the bi-populations as a tool in the

https://doi.org/10.1017/S0016672312000353 Published online by Cambridge University Press

integrated software QTL IciMapping (available from http://www.isbreeding.net), which is named 2pointREC. There are three parts on the 2pointREC interface (Fig. S1): (1) to specify population and marker character; (2) to specify the sample size of each marker class; and (3) to view the estimation of recombination frequency. One out of the 20 bi-parental populations (12 bi-parental populations in this study and 8 additional backcrossing populations when P_2 was used as recurrent parent; Fig. 1) can be specified (Fig. S2). Considering co-dominance, dominance and recessive, there are six scenarios for a pair of markers (Fig. S3). After population and marker characters have been specified, all potential marker classes that occurred in the specified population will be shown and the users can specify the observed sample size for each marker class. In the 'Results' window, e.g. Fig. S4, the first item is the estimated recombination frequency, followed by variance and standard deviation of the estimate. LOD score and the significance probability are shown for testing genetic linkage. Finally, the estimated recombination frequency was converted to map distance in cM by Haldane and Kosambi mapping functions.

This work was supported by the Natural Science Foundation of China (No. 31000540).

5. Supplementary material

The online data are available at http://journals.cambridge.org/GRH.

References

Antonio, B. A., Inoue, T., Kajiya, H., Nagamura, Y., Kurata, N., Minobe, Y., Yano, M., Nakagahra, M. & Sasaki, T. (1996). Comparison of genetic distance and order of DNA markers in five populations of rice. *Genome* **39**, 346–956.

Bailey, N. T. J. (1961). *Introduction to the Mathematical Theory of Genetic Linkage*. Oxford: Oxford University Press.

Beavis, W. D., Lee, M., Grant, D., Hallauer, A. R., Owens, T., Katt, M. & Blair, D. (1992). The influence of random mating on recombination among RFLP loci. *Maize Genetics Cooperation Newsletter* **66**, 52–53.

Darvasi, A. & Soller, M. (1995). Advanced intercross lines, an experimental population for fine genetic mapping. *Genetics* **141**, 1199–1207.

Falconer, D. S. & Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*. 4th edition. Harlow: Longman.

Fisher, R. A. (1935). The detection of linkage with 'dominant' abnormalities. *Annals of Eugenics* **6**, 187–201.

Frisch, M. & Melchinger, A. E. (2007). Precision of recombination frequency estimates after random intermating with finite population sizes. *Genetics* **178**, 597–600.

Hackett, C. A. & Broadfoot, L. B. (2003). Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. *Heredity* **90**, 33–38.

Haldane, J. B. S. & Smith, C. A. B. (1947). A new estimate of the linkage between the genes for haemophilia and colour-blindness in man. *Annals of Eugenics* **14**, 10–31.

Haldane, J. B. S. & Waddington, C. H. (1931). Inbreeding and linkage. *Genetics* **16**, 357–374.

Harushima, Y., Yano, M., Shomura, A., Sato, M., Shimano, T., Kuboki, Y., Yamamoto, T., Lin, S. Y., Antonio, B. A. & Parco, A. (1998). A high-density rice genetic linkage map with 2275 markers using a single F_2 population. *Genetics* **148**, 479–494.

Lee, M., Sharopova, N., Beavis, W. D., Grant, D., Katt, M., Blair, D. & Halauer, A. (2002). Expanding the genetic map of maize with the intermated B73 × Mo17 (IBM) population. *Plant Molecular Biology* **48**, 453–461.

Li, C., Jiang, H., Zhang, W., Qiu, P., Liu, C., Li, W., Gao, Y., Chen, Q. & Hu, G. (2008). QTL analysis of seed and pod traits in soybean. *Molecular Plant Breeding* **6**, 1091–1100.

Liang, Y., Peng, Y., Ye, S., Li, P., Sun, L., Ma, Z. & Li, Y. (2007). Comparison of genetic linkage maps based on F_2, F_6 populations derived from rice subspecies cross. *Hereditas* (*Beijing*) **29**, 1110–1120.

Liu, S., Zhou, R., Yu, D., Chen, S. & Gai, J. (2009). QTL mapping of protein related traits in soybean [*Glycine max* (L.) Merr.]. *Acta Agronomica Sinica* **35**, 2139–2149.

Lu, W., Gai, J., Zheng, Y. & Li, W. (2006). Construction of a soybean genetic linkage map and mapping QTLs resistant to soybean cyst nematode (*Heterodera glycines* Ichinohe). *Acta Agronomica Sinica* **32**, 1272–1279.

Morton, N. E. (1955). Sequential tests for the detection of linkage. *American Journal of Human Genetics* **7**, 277–318.

Nelson, J. C. (2011). Linkage analysis in unconventional mating designs in line crosses. *Thoeretical and Applied Genetics* **123**, 897–906.

Nordheim, E. V., O'Malley, D. M. & Guries, R. P. (1983). Estimation of recombination frequency in genetic linkage studies. *Theoretical and Applied Genetics* **66**, 313–321.

Ott, J. (1974). Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. *American Journal of Human Genetics* **26**, 588–597.

Ritter, E., Gebhardt, C. & Salamini, F. (1990). Estimation of recombination frequencies and construction of RFLP linkage maps in plants from crosses between heterozygous parents. *Genetics* **125**, 645–654.

Säll, T. & Nilsson, N. O. (1994). The robustness of recombination frequency estimates in intercrosses with dominant markers. *Genetics* **137**, 589–596.

Sirithunya, P., Tragoonrung, S., Vanavichit, A., Pa-In, N., Vongsaprom, C. & Toojinda, T. (2002). Quantitative trait loci associated with leaf and neck blast resistance in recombination inbred line population in rice (*Oryza sativa*). *DNA Research* **9**, 79–88.

Smith, C. A. B. (1953). The detection of linkage in human genetics. *Journal of the Royal Statistical Society Series B* **15**, 153–184.

Smith, C. A. B. (1959). Some comments on the statistical methods used in linkage investigations. *American Journal of Human Genetics* **11**, 289–304.

Tan, L., Zhang, P., Liu, F., Wang, G., Ye, S., Zhu, Z., Fu, Y., Cai, H. & Sun, C. (2008). Quantitative trait loci underlying domestication and yield-related traits in an *Oryza sativa* × *Oryza rufipogon* advanced backcross population. *Genome* **51**, 692–704.

Temnykh, S., DeClerck, G., Lukashova, A., Lipovich, L., Cartinhour, S. & McCouch, S. (2001). Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Research* **11**, 1441–1452.

Wang, X., Xu, Y., Li, G., Li, H., Gen, W. & Zhang, Y. (2010). Mapping quantitative trait loci for 100-seed weight in soybean (*Glycine max* L. Merr.). *Acta Agronomica Sinica* **36**, 1674–1682.

Wu, R., Han, Y., Hu, J., Fang, J., Li, L., Li, M. & Zeng, Z. (2000). An integrated genetic map of *Populus deltoids* based on amplified fragment length polymorphism. *Theoretical and Applied Genetics* **100**, 1249–1256.

Wu, R., Ma, C. & Casella, G. (2007). *Statistical Genetics of Quantitative Traits*. New York: Springer, pp. 43–56.

Xu, N. & Zhou, Z. (2000). The change of ELOD score and power under the real recombination rates. *Hereditas* (*Beijing*) **22**, 233–235.

Yan, C., Liang, G., Chen, F., Li, X., Tang, S., Yi, C., Tian, S., Lu, J. & Gu, M. (2003). Mapping quantitative trait loci associated with rice grain shape based on an *indica*/*japonica* backcross population. *Acta Genetica Sinica* **30**, 711–716.

https://doi.org/10.1017/S0016672312000353 Published online by Cambridge University Press