Genet. Res., Camb. (2010), **92**, *pp.* 283–294. © Cambridge University Press 2010
doi:10.1017/S0016672310000273

An investigation of the power for separating closely linked QTL in experimental populations

CHEN-HUNG KAO* AND MIAO-HUI ZENG
Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan, Republic of China

(*Received 11 March 2010 and in revised form 20 May 2010*)

Summary

Hu & Xu (2008) developed a statistical method for computing the statistical power for detecting a quantitative trait locus (QTL) located in a marker interval. Their method is based on the regression interval mapping method and allows experimenters to effectively investigate the power for detecting a QTL in a population. This paper continues to work on the power analysis of separating multiple-linked QTLs. We propose simple formulae to calculate the power of separating closely linked QTLs located in marker intervals. The proposed formulae are simple functions of information numbers, variance inflation factors and genetic parameters of a statistical model in a population. Both regression and maximum likelihood interval mappings suitable for detecting QTL in the marker intervals are considered. In addition, the issue of separating linked QTLs in the progeny populations from an F_2 subject to further self and/or random mating is also touched upon. One of the primary keys to our approach is to derive the genotypic distributions of three and four loci for evaluating the correlation structures between pairwise unobservable QTLs in the model across populations. The proposed formulae allow us to predict the power of separation when several factors, such as sample sizes, sizes and directions of QTL effects, distances between QTLs, interval sizes and relative QTL positions in the intervals, are considered together at a time in different experimental populations. Numerical justifications and Monte Carlo simulations were provided for confirmation and illustration.

1. Introduction

The calculation of statistical power of quantitative trait locus (QTL) detection has been an important problem in QTL mapping. Soller *et al.* (1976) and Lander & Botstein (1989) discussed the power of QTL detection when a QTL is coincident with a genetic marker. Hu & Xu (2008) developed a simple method to calculate the statistical power of QTL in the interval flanked by its two markers in a population. On the basis of the regression (REG) interval mapping model (Haley & Knott, 1992), their method can predict the power of QTL detection given the factors, such as size and position of QTL, sample size and interval size, by evaluating a non-central *F*-distribution function. It has been noticed that closely linked QTL might be mistakenly estimated as a single (ghost) QTL with a larger effect at the wrong position if they have the same direction effects, or they might be out of detection if their effects are in the opposite direction (Lander & Botstein, 1989; Kao & Zeng, 1997; Ronin *et al.*, 1999). Therefore, the study of separating linked QTLs to improve the QTL resolution remains an important issue. Ronin *et al.* (1999) derived the asymptotic expected LOD values based on the non-central chi-squared distribution for the study of two linked QTLs coincident with markers. Mayer (2005) compared REG interval mapping and maximum likelihood (ML) interval mapping in detecting two linked QTLs using Monte Carlo simulations. So far, analytical methods for the power analysis of detecting linked QTLs situated in the marker intervals have not been fully developed. We propose statistical methods for calculating the power for separating

* Corresponding author: Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan, Republic of China. Tel: (02) 2783-5611 ext 418. Fax: (02) 2783-1523. e-mail: chkao@stat.sinica.edu.tw

closely linked QTLs located in the intervals. The proposed formulae are based on the information numbers (the inverse of the variance of the best estimated QTL effects), variance inflation factors (caused by the correlations between linked QTLs) and genetic parameters of statistical models in a population. Both REG and ML interval mapping are considered in the formulation. Further, the power analyses of the above-mentioned papers mainly focus on the genome structures of the backcross (BC) and F_2 populations (Hu & Xu also discussed the power calculation in the double haploid (DH) and recombinant inbred lines (RIL)). We also discuss the separation of closely linked QTLs in other experimental populations, such as advanced intercross (AI) and recombinant inbred (RI) populations, subject to more meiosis cycles. One important key to the proposed method is to derive the genotypic distributions of three and four loci to characterize the correlation structures between pairwise QTL variables in the model for different populations. In general, we found that, given a distance between QTLs, separation can be more powerful for QTLs of similar size, with opposite direction effects, located closer to markers and in narrow intervals, and contributing to a high proportion of trait variation. More advanced populations may facilitate the separation of linked QTLs by providing more recombinants and changing genome structures. Numerical and simulated results are presented for confirmation.

2. Methods

(i) *Test statistic for detecting a QTL*

Consider an F_2 population or its progeny populations produced by further selfing and/or intercrossing the F_2 individuals for different numbers of generations. There are three possible genotypes, P_1 homozygote, heterozygote and P_2 homozygote for any gene. Let Q_jQ_j, Q_jq_j and q_jq_j be the three possible genotypes of a QTL, say Q_j, under consideration in a population. For an individual i in a random sample with size n, let x^*_{ij} represent the coded variable of QTL genotype as

$$x^*_{ij} = \begin{cases} 1, & \text{if } Q_j \text{ is } Q_jQ_j, \\ 0, & \text{if } Q_j \text{ is } Q_jq_j, \\ -1, & \text{if } Q_j \text{ is } q_jq_j, \end{cases} \tag{1}$$

and a_j denote its additive effect. Similarly, it is straightforward to construct the coded variable x^*_{ik} for another QTL, say Q_k, with additive effect a_k for a model taking multiple QTLs into account. When Q_j, flanked by the left marker M_j and right marker N_j with alleles (M_j, m_j) and (N_j, n_j), is considered, the conditional expectation of x^*_{ij} given M_j and N_j, $w_{ij}=E(x^*_{ij}|M_j, N_j)$, is used as the predictor variable in the REG interval mapping model (Haley & Knott, 1992). For a single QTL model, Hu & Xu (2008) have shown that the test statistic

$$\lambda = \frac{\sum(w_{ij}-\overline{w}_{ij})^2}{\sigma^2} \times \hat{a}^2_j \tag{2}$$

(σ^2 is the residual error variance) follows a central F-distribution under the null hypothesis (H_0: $a_j=0$). Under the alternative hypothesis (H_1: $a_j \neq 0$), this test statistic follows a non-central F-distribution with the non-centrality parameter $\delta = n \times \text{var}(w_{ij}) \times a^2_j/\sigma^2$. The non-centrality parameter is a function of several important factors, sample size, variance of the predictor variable, QTL effect and residual error variance. By analysing these factors, the power for detecting a QTL can be predicted for different situations. For example, Hu & Xu (1998) analysed one of the key factors, the variance of the coded variable, var(w_{ij}), by deriving its different formulations for the BC, F_2, RIL and DH. When n is sufficiently large, $\sum(w_{ij}-\overline{w}_{ij})^2 = n \times \text{var}(w_{ij})$, and the variance of the estimated effect is $\text{var}(\hat{a}_j) = \sigma^2/[n \times \text{var}(w_{ij})]$. When var($w_{ij}$) is small, var($\hat{a}_j$) is large and δ is small, leading to lower power in QTL detection.

(ii) *Variances of predictor variables*

The aim of this study is to calculate the power for detecting two or more closely linked QTLs and to extend the power analysis to the populations beyond F_2 using both REG and ML interval mapping. When analysing the power for detecting one QTL, we only need to understand the asymptotic behaviour of the variances of predictor variables to construct the test statistic for power analysis as has been done by Hu & Xu (2008). For dissecting linked QTLs, we should further derive the covariances between different QTL predictor variables to obtain the asymptotic variance–covariance matrix of QTL parameters for power analysis. An important step to obtain the variances and covariances of the predictor variables is to characterize the genotypic distributions of multiple genes in the populations. For example, evaluating $E(w_{ij})$ and var(w_{ij}) in the BC between a population M_jN_j/M_jN_j on F_1 requires considering the four flanking marker genotypes of two genes, M_jN_j/M_jN_j, M_jN_j/M_jn_j, M_jN_j/m_jN_j and M_jN_j/m_jn_j with frequencies $(1-r)/2$, $r/2$, $r/2$ and $(1-r)/2$, where r is the recombination fraction between A and B (Xu, 1995). Evaluating them in the F_2 between two populations, M_jN_j/M_jN_j and m_jn_j/m_jn_j, requires taking into account ten marker genotypes of two genes, M_jN_j/M_jN_j, m_jn_j/m_jn_j, M_jn_j/M_jn_j, m_jN_j/m_jN_j, M_jN_j/M_jn_j, M_jN_j/m_jN_j, M_jn_j/m_jn_j, m_jN_j/m_jn_j, M_jN_j/m_jn_j and M_jn_j/m_jN_j with frequencies $(1-r)^2/4$, $(1-r)^2/4$, $r^2/4$, $r^2/4$, $r(1-r)/2$, $r(1-r)/2$, $r(1-r)/2$, $r(1-r)/2$, $(1-r)^2/2$ and $r^2/2$ (Hu & Xu, 2008). In the progeny populations from F_2, these ten genotypic frequencies change over populations. For AI populations subject to more cycles of random mating, the well-known formula, $P'(\underline{M_jN_j}) = (1-r) \times P(\underline{M_jN_j}) + r \times P(M_j) \times P(N_j)$, can be used to obtain the genotypic frequencies, where $P'(\underline{M_jN_j})$ is the frequency of $\underline{M_jN_j}$ in the next generation. For RI populations subject to further selfing, Haldane & Waddington's transition equations (1931) can be applied to obtain the ten frequencies. Using the same notations as in Haldane & Waddingtion's paper, we denote the frequency of M_jN_j/M_jN_j (m_jn_j/m_jn_j) genotype as C, the frequency of M_jn_j/M_jn_j (m_jN_j/m_jN_j) genotype as D, the frequency of M_jN_j/M_jn_j (M_jN_j/m_jN_j, M_jn_j/m_jn_j, or m_jN_j/m_jn_j) genotype as E, the frequency of M_jN_j/m_jn_j genotype as F and the frequency of M_jn_j/m_jN_j genotype as G, respectively, in the populations. With such settings, it is straightforward to show that $E(w_{ij}) = 0$ and to formulate the variance of w_{ij} in a population as

$$\mathrm{var}(w_{ij}) = E(w_{ij}^2) = \sum_{k=1}^{10} f_k(p_{k1} - p_{k3})^2, \tag{3}$$

where p_{k1} and p_{k3}, $k = 1, 2, \ldots, 10$, are conditional probabilities of Q_jQ_j and Q_jq_j genotypes given the ten flanking marker genotypes, and f_k are the frequencies of the ten marker genotypes for two flanking markers (C, D, E, F and G). Note that the derivation of p_{k1} and p_{k3} is not straightforward as has been done in BC and F_2 populations, and it involves using the genotypic distributions of three genes (Kao & Zeng, 2009). If the event of double recombinations is ignored within a marker interval, equation (3) can be explicitly formulated as

$$\mathrm{var}(w_{ij}) = 2(C + D + E) - 4p(1-p)(E + 2D), \tag{4}$$

where $p = r_1/r$ (r and r_1 are the recombination fractions between (M_j, N_j) and (M_j, Q_j)). It is interesting to analyse equation (4) to gain some insight into $\mathrm{var}(w_{ij})$. In equation (4), $\mathrm{var}(w_{ij})$ is bounded by $2(C + D + E)$, which is the variance of a fully observed QTL-coded variable. The term $p(1-p)$ measures the relative QTL position in a marker interval, and $E + 2D$ measures the interval size. As the marker interval becomes wider or the QTLs get closer to the centre position of the interval, $E + 2D$ or $p(1-p)$ becomes larger, and the value of $\mathrm{var}(w_{ij})$ becomes smaller. In the F_2 population, $2(C + D + E) = 1/2$, $E + 2D = r/2$ and $\mathrm{var}(w_{ij}) = 1/2 - 2rp(1-p)$, which are bounded by 1/2. In AI F_t populations, $2(C + D + E) = 1/2$ and $E + 2D = r_t/2$, where $r_t = [1-(1-r)^{t-2}(1-2r)]/2$. The variance $\mathrm{var}(w_{ij}) = 1/2 - 2r_tp(1-p)$, which is also bounded by 1/2 ($p = r_{1t}/r_t$) and decreases in the later populations. In RI populations, $2(C + D + E)$ is between 1/2 and 1, and $E + 2D$ is between $2/r$ and $2r/(1 + 2r)$. The value of $\mathrm{var}(w_{ij})$ increases as population advances. In RIL, $2(C + D + E) = 1$, $E + 2D = 2r/(1 + 2r)$ and $\mathrm{var}(w_{ij}) = 1 - (8r/(1 + 2r))p(1-p)$, which are bounded by 1. Similarly, the variance of the predictor variable for dominance effect is about $\sim 1/4 - r/2 \times \{1 - r[(1 - 2p(1-p))^2 + 2p(1-p)]\}$, which is bounded by 1/4 in the F_2 population. The variance of the predictor variable is $\mathrm{var}(w_{ij}) = 1/4 - rp(1-p)$ bounded by 1/4 in the BC population (see also Xu, 1995). Hu & Xu (2008) formulated $\mathrm{var}(w_{ij})$ in the F_2, RIL and DH populations when double recombinations in the intervals are considered. In general, the larger the variance of a predictor variable, the greater the power in QTL detection.

(iii) *Power for detecting a QTL*

When only one QTL is considered in the model, Hu & Xu (2008) showed an example that $\mathrm{var}(w_{ij})$ is 0·450 for a QTL located in the middle of a 10-cM marker interval ($r_1 = r_2 = 0{\cdot}04758$ and $r = 0{\cdot}09063$), and that 252 individuals are required to detect this QTL with 80% power under $\alpha = 0{\cdot}01$ when the QTL explains 5% of the trait variation in the F_2 population. Our formulae in equation (3) allow us to calculate the values of $\mathrm{var}(w_{ij})$ and sample sizes required in different populations under the same conditions. For the same conditions, the values of $\mathrm{var}(w_{ij})$ derived using our formulae in the different AI and RI populations are presented in Table 1. It shows that the trend in the change of variance behaves differently under selfing and random mating. When further selfing, the variance increases. When successive intercrossing, the variance tends to decrease. For example, the values are 0·651 and 0·806 in the RI F_3 and RIL (generation 10 of RI population is called RIL), respectively, and they are 0·426 and 0·271 in the AI F_3 and AI F_{10}, respectively. The different values of $\mathrm{var}(w_{ij})$ cause the non-centrality parameter to be different, thus affecting the power of detection. To guarantee an 80% power to detect this QTL under $\alpha = 0{\cdot}01$, it would require about 175, 155, 148 and 143 individuals in the RI F_3, F_4, F_5 and RIL populations, and it would require 262, 284, 302 and 426 individuals in the AI F_3, F_4, F_5 and F_{10} populations. This shows that the sample size can be saved in the more advanced RI populations and may not be saved in the later AI populations when mapping a single QTL located in the interval.

(iv) *Covariances between predictor variables*

To obtain covariances between the predictor variables, $\mathrm{cov}(w_{ij}, w_{ik})$'s, we need to understand the genotypic distributions of three and four genes in a population. For two linked QTLs, Q_j and Q_k, flanked by two marker pairs (M_j, N_j) and (M_k, N_k) they can be located in neighbouring or non-neighbouring marker intervals. For the neighbouring case, the order is

Table 1. *The values of variances, covariances and correlations of the predictor variables in the AI and RI* F_t *populations. The case considered is* $\mathrm{M_j\text{-}Q_j\text{-}N_j\text{-}Q_k\text{-}N_k}$ *with* $\mathrm{d}_{M_jQ_j}=5\,\mathrm{cM}$, $\mathrm{d}_{Q_jN_j}=5\,\mathrm{cM}$, $\mathrm{d}_{N_jQ_k}=5\,\mathrm{cM}$ *and* $\mathrm{d}_{Q_kN_k}=5\,\mathrm{cM}$

t	$V(x_{ij})$	$C(x_{ij}, x_{ik})$	$\rho(x_{ij}, \mathrm{x}_{ik})$	$V(w_{ij})$	$C(w_{ij}, w_{ik})$	$\rho(w_{ij}, w_{ik})$	$V(x_{ij}{*})$	$C(x_{ij}{*}, x_{ik}{*})$	$\rho(x_{ij}{*}, x_{ik}{*})$
AI F_t									
2	0·5	0·410	0·819	0·450	0·409	0·909	0·437	0·380	0·869
3	0·5	0·372	0·744	0·426	0·372	0·874	0·427	0·370	0·866
4	0·5	0·339	0·677	0·402	0·338	0·841	0·410	0·343	0·836
5	0·5	0·308	0·616	0·378	0·307	0·811	0·382	0·302	0·790
6	0·5	0·280	0·560	0·355	0·278	0·784	0·361	0·273	0·755
7	0·5	0·255	0·509	0·333	0·253	0·759	0·338	0·237	0·702
8	0·5	0·232	0·463	0·312	0·230	0·736	0·337	0·228	0·676
9	0·5	0·206	0·412	0·291	0·208	0·715	0·305	0·198	0·650
10	0·5	0·172	0·383	0·271	0·189	0·696	0·310	0·192	0·620
RI F_t									
2	0·5	0·410	0·819	0·450	0·409	0·909	0·437	0·380	0·869
3	0·750	0·577	0·769	0·651	0·577	0·886	0·640	0·563	0·880
4	0·875	0·658	0·741	0·739	0·644	0·872	0·734	0·637	0·868
5	0·938	0·679	0·724	0·778	0·672	0·864	0·793	0·671	0·846
6	0·969	0·692	0·714	0·794	0·682	0·859	0·807	0·685	0·849
7	0·984	0·697	0·708	0·802	0·687	0·856	0·814	0·688	0·845
8	0·992	0·699	0·705	0·804	0·687	0·855	0·814	0·690	0·848
9	0·996	0·700	0·703	0·806	0·688	0·854	0·811	0·689	0·849
10	0·998	0·901	0·702	0·806	0·688	0·854	0·815	0·691	0·848

$V(x_{ij})$: variance of $V(x_{ij})$. $C(x_{ij}, x_{ik})$ and $\rho(x_{ij}, x_{ik})$: covariance and correlation between x_{ij} and x_{ik}. x_{ij} and x_{ik} denote the predictor variables when Q_j and Q_k are fully observed, and w_{ij} and w_{ik} (x^*_{ij} and x^*_{ik} denote the predictor variables when Q_j and Q_k are not observed and constructed from their flanking markers in the REG (ML) interval mapping model.

M_j-Q_j-N_j-Q_k-N_k (N_j and M_k are the same marker). For the non-neighbouring case, the order is M_j-Q_j-N_j–M_k-Q_k-N_k order. Note that the case for QTLs located in non-neighbouring intervals may include additional markers between N_j and M_k. For the case of M_j-Q_j-N_j–M_k-Q_k-N_k order, the two predictor variables, w_{ij} and w_{ik}, are constructed using the marker pairs (M_j, N_j) and (M_k, N_k). Therefore, computing their covariance, cov(w_{ij}, w_{ik}), needs to considered all for the 136 possible genotypes of M_j, N_j, M_k and N_k markers (see the Appendix). For the case of M_j-Q_j-N_j-Q_k-N_k order, obtaining the covariance only needs to evaluate all the 36 marker genotypes of M_j, N_j and N_k markers. The latter case is more difficult to detect Q_j and Q_k simultaneously as they share the same flanking marker N_j. The covariance between w_{ij} and w_{ik} can be generally expressed as

$$\begin{aligned}\mathrm{cov}(w_{ij}, w_{ik}) &= E(w_{ij}\times w_{ik})\\ &= \sum_{k=1}^{n_g} f_k\times(p_{k1}-p_{k3})(p_{j1}-p_{j3}),\end{aligned} \qquad (5)$$

where $n_g=36$ or 136 and f_k are the genotypic frequencies of flanking markers from trigenic and tetragenic distributions. In F_2 population, the genotypic distributions of three and four markers can be obtained from the product of probability distributions of adjacent pairwise genes, i.e. $P(\underline{M_jN_jN_k})=P(\underline{M_jN_j})\times P(\underline{N_jN_k})$ and $P(\underline{M_jN_jM_kN_k})=P(\underline{M_jN_j})\times P(\underline{N_jM_k})\times P(\underline{M_kN_k})$, under the Haldane map function. For example, the gamete frequency $P(\underline{M_jN_jM_kN_k})=(1-r_1)(1-r_2)(1-r_3)/2$ in the F_2 population, where r_1, r_2 and r_3 are the recombination fractions between (M_j, N_j), (N_j, M_k) and (M_k, N_k). For the advanced populations beyond F_2, trigenic and tetragenic genotypic distributions cannot be obtained from the direct product of pairwise gene distributions. We use special devises outlined in Kao & Zeng (2009) and in the Appendix to obtain the genotypic distributions of three and four genes. Although the covariance in equation (5) does not have a simple form as in equation (4) for variance, it can be easily written into a computer programme to obtain the covariances under different situations in different populations. For example, in the case of M_j-Q_j-N_j–M_k-Q_k-N_k order with $d_{\mathrm{M}_j\mathrm{Q}_j}=5\,\mathrm{cM}$, $d_{\mathrm{Q}_j\mathrm{N}_j}=5\,\mathrm{cM}$, $d_{\mathrm{M}_k\mathrm{Q}_k}=5\,\mathrm{cM}$ and $d_{\mathrm{Q}_k\mathrm{N}_k}=5\,\mathrm{cM}$, the values of cov($w_{ij}$, w_{ik}) are 0·7445, 0·6736, 0·6095, 0·5515 and 0·4991 for $d_{\mathrm{N}_j\mathrm{M}_k}=10$, 15 20, 25 and 30 cM, respectively, in the F_2 population. In the case of M_j-Q_j-N_{jk}-Q_k-N_k order with $d_{\mathrm{M}_j\mathrm{Q}_j}=5\,\mathrm{cM}$, $d_{\mathrm{Q}_j\mathrm{N}_j}=5\,\mathrm{cM}$, $d_{\mathrm{M}_j\mathrm{Q}_k}=5\,\mathrm{cM}$ and $d_{\mathrm{Q}_k\mathrm{N}_k}=5\,\mathrm{cM}$, its covariances in different populations are presented in Table 1. Table 1 shows that the covariance increases under further selfing and decreases when subjected to more intercrossing. For example, the covariance is 0·409 in the F_2 population. The values are 0·577 and 0·688 in the RI F_3 and RIL, respectively, and they are 0·372 and 0·189 in the AI F_3 and AI F_{10}, respectively. Although the covariance can become larger or smaller, the correlations between the coded variables, $\rho(w_{ij}, w_{ik})$, all decrease in the advanced populations (Table 1). The correlation is 0·909 in the F_2 populations. It becomes 0·886 and 0·854 in the RI F_3 and RIL, and it is 0·874 and 0·696 in the AI F_3 and F_{10} populations. As will be discussed later, the detection of linked QTLs can benefit from the diminishing correlation between predictor variables in the advanced populations.

(v) *Variances of the estimated QTL effects*

For a single QTL model, we only need the variance of the coded variable, $\mathrm{var}(w_{ij})$, to construct a test statistic in power analysis (equation (2)). As the variance of the estimated effect is the inverse of the information number of QTL effect, i.e. $\mathrm{var}(\hat{a}_j) = I^{-1}(a_j)$, for n large, we have

$$\mathrm{var}(\hat{a}_j) = \frac{\sigma^2}{n \times \mathrm{var}(w_{ij})} \tag{6}$$

and $\mathrm{var}^{-1}(\hat{a}_j)/n = I(a_j)/n \sim \mathrm{var}(w_{ij})/\sigma^2$ in a single QTL model. For multiple, say p, QTLs in the model, the variance–covariance matrix of the predictor variables is required in constructing the test statistics. Similarly, for n large, we have $I(a)/n = [(W'W)/\sigma^2]/n \to V(W)/\sigma^2$, where W denotes the matrices whose i, jth entry is w_{ij} and $V(W)$ is the variance–covariance matrix with diagonal elements $\mathrm{var}(w_{ij})$'s, $j = 1, 2, \ldots, p$, and off-diagonal elements $\mathrm{cov}(w_{ij}, w_{ik})$'s. Under normal assumption, $n^{1/2}(\underline{\hat{a}} - \underline{a}) \to \mathrm{N}_p(0, V^{-1}(W) \times \sigma^2)$ (Fuller 1976). Without loss of generality, we present the case of $p = 2$ with Q_j and Q_k in the model for a better illustration. For $p = 2$, the $V^{-1}(W)$ matrix is

$$V^{-1}(W) = [1 - \rho^2(w_{ij}, w_{ik})]^{-1} \times \begin{pmatrix} \mathrm{var}^{-1}(w_{ij}) & -\dfrac{\mathrm{cov}(w_{ij}, w_{ik})}{\mathrm{var}(w_{ij}) \times \mathrm{var}(w_{ik})} \\ -\dfrac{\mathrm{cov}(w_{ij}, w_{ik})}{\mathrm{var}(w_{ij}) \times \mathrm{var}(w_{ik})} & \mathrm{var}^{-1}(w_{ik}) \end{pmatrix}, \tag{7}$$

where $\rho(w_{ij}, w_{ik}) = \mathrm{cov}(w_{ij}, w_{ik}) / \sqrt{\mathrm{var}(w_{ij}) \times \mathrm{var}(w_{ik})}$. Therefore, the variances of estimated a_j and a_k are

$$\mathrm{var}(\hat{a}_j) = \frac{1}{[1 - \rho^2(w_{ij}, w_{ik})]} \times \frac{\sigma^2}{n \times var(w_{ij})}$$

and

$$\mathrm{var}(\hat{a}_k) = \frac{1}{[1 - \rho^2(w_{ij}, w_{ik})]} \times \frac{\sigma^2}{n \times \mathrm{var}(w_{ik})}, \tag{8}$$

respectively. By comparing equations (6) with (8), it shows that the variances of the estimated QTL effects are not only affected by $\mathrm{var}(w_{ij})$ and $\mathrm{var}(w_{ik})$ but also by $\mathrm{cov}(w_{ij}, w_{ik})$ through $\rho(w_{ij}, w_{ik})$. The first term on the right-hand side of equation (8) is usually called variance inflation factor (VIF), which can be also expressed in terms of information numbers, $I(a_j)$, $I(a_k)$ and $I(a_j, a_k)$, as

$$\mathrm{VIF} = [1 - \rho^2(w_{ij}, w_{ik})]^{-1} = \left[1 - \frac{I(a_j) \times I(a_k)}{I^2(a_j, a_k)}\right]^{-1}. \tag{9}$$

The VIF can measure the inflation level of the variance of an estimate (Marquardt, 1970). When $\rho(w_{ij}, w_{ik}) = 0$, $\mathrm{VIF} = 1$ and there is no variance inflation. If $\rho(w_{ij}, w_{ik}) \neq 0$, $\mathrm{VIF} > 1$ indicating that the inflation of variances occurs. In general, large VIF indicates seriously inflated variances and a severe collinearity problem, and the linked QTL are not likely to be detected statistically. For the same M_j-Q_j-N_j-Q_k-N_k order considered in Table 1, the value of VIF in $\mathrm{var}(\hat{a}_j)$ or $\mathrm{var}(\hat{a}_k)$ is 5·750 ($(1 - 0{\cdot}909^2)^{-1}$), implying that its variance is inflated by 5·750 times as compared to when they are unlinked. The values of VIF are 4·651 and 3·694 in the RI F_3 and RIL, respectively, and they are 4·650 and 1·940 in the AI F_3 and AI F_{10}, respectively. The values of VIF become smaller in the more advanced populations. Therefore, advanced populations have the ability to provide smaller VIF values for more powerful QTL detection (more explanation is given below). Also, the VIF is generally larger when interval sizes become wider or the putative QTL move towards the centres of intervals (not shown). With VIF, $V^{-1}(W)$ in equation (7) can be simplified in expression as $V^{-1}(W) = \mathrm{VIF} \times A_0$, where $A_0 = [a_{ij}]_{2 \times 2}$ denotes the 2×2 matrix in the equation.

(vi) *Test statistics for detecting linked QTL*

We now derive the test statistics for analysing the separation of linked QTL and calculating the separating power. Let

$$t_j = (\hat{a}_j - a_j)/\sigma_j \quad \text{and} \quad t_k = (\hat{a}_k - a_k)/\sigma_k \tag{10}$$

be the standardized estimated QTL effects, where $\sigma_j^2 = \mathrm{VIF} \times a_{11} \times \sigma^2/n$ and $\sigma_k^2 = \mathrm{VIF} \times a_{22} \times \sigma^2/n$ are the variances of the estimated effects ($a_{11} = \mathrm{var}^{-1}(w_{ij})$ and $a_{22} = \mathrm{var}^{-1}(w_{ij})$). As $I^{-1}(a_j) = (a_{11} \times \sigma^2)/n$ and $I^{-1}(a_k) = (a_{22} \times \sigma^2)/n$, it is more convenient and succinct to express σ_j^2 and σ_k^2 as

$$\sigma_j^2 = \mathrm{VIF} \times I^{-1}(a_j) \quad \text{and} \quad \sigma_k^2 = \mathrm{VIF} \times I^{-1}(a_k) \tag{11}$$

in a population. Accordingly, the joint distribution of t_j and t_k follows a bivariate normal distribution with mean zero and covariance matrix with diagonal elements, one, and off-diagonal elements, $\rho(w_{ij}, w_{ik})$, as

$$\begin{bmatrix} t_j \\ t_k \end{bmatrix} \sim N_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -\rho(w_{ij}, w_{ik}) \\ -\rho(w_{ij}, w_{ik}) & 1 \end{bmatrix} \right). \tag{12}$$

Given a pre-specified critical value c at the significance level α, the power of separation is the sum of probabilities that t_j and t_k are simultaneously different from zeros:

$$P\left(t_j > c - \frac{a_j}{\sigma_j},\ t_k > c - \frac{a_k}{\sigma_k}\right) + P\left(t_j < -c - \frac{a_j}{\sigma_j},\ t_k < -c - \frac{a_k}{\sigma_k}\right) + P\left(t_j > c - \frac{a_j}{\sigma_j},\ t_k < -c - \frac{a_k}{\sigma_k}\right) + P\left(t_j < -c - \frac{a_j}{\sigma_j},\ t_k > c - \frac{a_k}{\sigma_k}\right) \tag{13}$$

in the bivariate normal distribution. Note that the sum of four probabilities is equivalent to Type I error α under the null hypothesis (H_0: $a_j = 0$ and $a_k = 0$). Under the alternative hypothesis (H_1: $a_j \neq 0$ and $a_k \neq 0$), equation (13) is the power to reject H_0 and allows us to evaluate the power of separation for different values of a_j and a_k in different populations (see section 4).

When an ML interval mapping is implemented in separating linked QTLs, the model is a normal mixture model under the assumption of normal errors. We use x^*_{ij}'s to denote the predictor variables in the ML interval mapping models. By treating x^*_{ij}'s as missing data and y_i as observed data, we can apply the EM algorithm to obtain the MLE and information matrix by operating on the complete-data likelihood

$$L(\mathrm{Y}_{\mathrm{com}}|\theta) = f(y_i|\theta, x^*_{i1}, \ldots, x^*_{ip})\ g(x^*_{i1}, \ldots, x^*_{ip}). \tag{14}$$

For p QTL, there are 3^p QTL genotypes, and let μ_j, $j = 1, 2, \ldots, 3^p$, denote their genotypic values. In the complete-data likelihood, the conditional distribution of the observed data given missing data, $f(y_i|\theta, x^*_{i1}, \ldots, x^*_{ip})$, follows a normal distribution $N(\mu_j, \sigma^2)$, and $g(x^*_{i1}, \ldots, x^*_{ip})$ is a 3^p-nomial distribution depending on the values of x^*_{ij}'s (QTL genotypes). Let q_{ij}'s be the 3^p-nomial probabilities derived from the conditional probabilities of QTL genotypes given the flanking marker genotypes. Both MLE and observed information matrix involve the posterior probabilities of the QTL genotypes, $\pi_{ij} = [q_{ij} \times N(\mu_j, \sigma^2)] / [\sum_{j=1}^{3^p} q_{ij} \times N(\mu_j, \sigma^2)]$ (please see Kao & Zeng (1997) for more details about the derivations). Therefore, for $p = 2$, evaluating the (expected) information numbers, $I(a_j)$, $I(a_k)$ and $I(a_j, a_k)$, needs to integrate the distribution of markers and traits, and thus is more challenging. Here, we suggest a Monte Carlo simulation approach to evaluate the expected π_{ij} by simulating, say 10 000, individuals to approximate the expected π_{ij} as $E(\pi_{ij}) = \left(\sum_{i=1}^{10\,000} \hat{\pi}_{ij}\right)/10000$, where $\hat{\pi}_{ij}$ denotes the value of π_{ij} of each individual. In turn, the information numbers can be obtained. Similarly, to those outlined in REG interval mapping, we can denote $I(a_j)/n = \mathrm{var}(x^*_{ij})/\sigma^2$ and $I(a_j, a_k)/n = \mathrm{cov}(x^*_{ij}, x^*_{ik})/\sigma^2$ for sufficiently large n in ML interval mapping. Table 1 presents the values of $I(a_j)$ and $I(a_j, a_k)$ for the same case of M_j-Q_j-N_j-Q_k-N_k order. The values are obtained by simulating trait values governed by two QTLs with equal effects, and the heritability is $h^2 = 0{\cdot}05$ with $\sigma^2 = 1$. As $\sigma^2 = 1$, $I(a_j) = \mathrm{var}(x^*_{ij})$ and $I(a_j, a_k) = \mathrm{cov}(x^*_{ij}, x^*_{ik})$. The values of $\mathrm{var}(x^*_{ij})$ are 0·437, 0·640 and 0·815 in the F_2, RI F_3 and RIL, respectively, and are 0·428 and 0·310 in the AI F_3 and F_{10}, respectively (the values of $\mathrm{var}(x_{ik})$ are of very similar size and not presented). As compared to $\mathrm{var}(w_{ij})$ in REG interval mapping, these variances are of similar sizes. The values of $\mathrm{cov}(x^*_{ij}, x^*_{ik})$ are 0·380, 0·563, 0·691, 0·370 and 0·192 in the F_2, RI F_3, RIL, AI F_3 and AI F_{10} populations. Except for the value in generation 10, these values are smaller as compared to the values of $\mathrm{cov}(w_{ij}, w_{ik})$ in REG interval mapping (the values of $\mathrm{cov}(w_{ij}, w_{ik})$ are 0·409, 0·577, 0·688, 0·392 and 0·189, respectively). Also, the values of correlation between the QTL-coded variables can be also obtained (Table 1). In general, the predictor variables in ML interval mapping have smaller covariances (correlations). Therefore, the ML method will have smaller VIF values when fitting closely linked QTL together. The values of VIF are 4·084, 4·299 and 3·232 in the F_2, RI F_3 and RIL, respectively, and are 3·999 and 1·655 in the AI F_3 and AI F_{10}, respectively. These results indicate that the ML interval mapping suffers a low collinearity problem, and it can be more efficient and powerful in detecting linked QTLs as will be further validated in sections 3 and 4. By obtaining the information numbers of the QTL effects for ML interval mapping, the components in equation (12) can be updated to construct test statistics, $t_j = (\hat{a}_j - a_j)/\sigma_j$ and $t_k = (\hat{a}_k - a_k)/\sigma_k$ for ML interval mapping. Then, using the bivariate normal distributions, the hypothesis H_0: $a_j = 0$ and $a_k = 0$ can be tested for calculating the power of ML interval mapping.

When more, say p, QTLs are considered in the REG interval mapping model, the information matrix of parameters is $I(\underline{a}) = (W'W)/\sigma^2$. It can shown that $I(\underline{a})/n \sim V(W)/\sigma^2$. As $V(W)$ is invertible, we can express $V^{-1}(W) = \mathrm{VIF} \times A_0$, where $A_0 = [a_{ij}]_{p \times p}$. For ML interval mapping, the information matrix can be obtained by using the general formulae of Kao & Zeng (1997). Similarly, when sample size grows large, $I(\underline{a})/n$ can be expressed as $V(X^*)/\sigma^2$ (X^* denotes the matrix whose i, jth entry is x^*_{ij}), whose diagonal elements are the expected $I(a_j)$'s, $j = 1, 2, \ldots, p$ and off-diagonal elements are expected $I(a_j, a_k)$'s. The $V(X^*)$

matrix is also invertible and can be formulated as $V^{-1}(X^*) = \mathrm{VIF} \times A_0$. For both REG and ML interval mapping, we can define $\sigma_j^2 = \mathrm{VIF} \times a_{jj} \times \sigma^2/n$, where a_{jj}, $j = 1, 2, \ldots, p$ denote the diagonal elements in A_0. Then, we can construct the standardized estimated effects as $t_j = (\hat{a}_j - a_j)/\sigma_j$, $j = 1, 2, \ldots, p$, and $(t_1, t_2, \ldots, t_p)'$ follows a p-variate normal distribution. Given specified critical values, the probability of significance can be calculated (Genz & Bretz, 2009) to evaluate the power of separating more linked QTLs.

(vii) *Genetic parameters and residual variances*

Further, we know that the relationship between environmental variance, σ^2, and genetic variance, V_{G}, can be formulated as $\sigma^2 = \frac{1-h^2}{h^2} \times V_{\mathrm{G}}$, where h^2 is the heritability of quantitative trait variation. The genetic variance can be decomposed into components of genotypic frequencies and QTL effects. For two QTLs with additive effects only, $V_{\mathrm{G}} = \mathrm{var}(x_{ij}) \times a_j^2 + \mathrm{var}(x_{ik}) \times a_k^2 + 2 \times \mathrm{cov}(x_{ij}, x_{ik}) \times a_j \times a_k$, where x_{ij} and x_{ik} denote the coded variables of the two fully observed QTLs (see Kao & Zeng, 2009 for the components of V_{G} with complete effects and contributed by more QTLs). As $\mathrm{var}(x_{ij}) = \mathrm{var}(x_{ik}) = 2(C+D+E)$ and $\mathrm{cov}(x_{ij}, x_{ik}) = 2(C-D)$ depend on the genotypic distribution of experimental populations, given specific QTL effects, V_{G} is population dependent. For example, $V(x_{ij}) = 1/2$ and $\mathrm{cov}(x_{ij}, x_{ik}) = (1-2r_t)/2$ in AI F_t populations, and $1/2 < V(x_{ij}) < 1$ and $(1-2r)/2 < \mathrm{cov}(x_{ij}, x_{ik}) < (1-2r)/(1+2r)$ in RI F_t populations. Therefore, a more detailed formulation of V_{G} can be also expressed as

$$V_{\mathrm{G}} = \frac{1-h^2}{h^2} \times 2[(C+D+E) \times (a_j^2 + a_k^2) + (C-D) \times a_j \times a_k]. \qquad (15)$$

Xu (1995) pointed out that the residual variance in REG interval mapping inflates, due to the uncertainty of the QTL genotype, and that the amount of inflation parameter is about $[\mathrm{var}(x_{ij}) - \mathrm{var}(w_{ij})] \times a_j^2$ in a single QTL model. For a multiple QTL model, the amount is about $\sum_{j=1}^{p} [\mathrm{var}(w_{ij}) \times \mathrm{var}(w_{ij})] \times a_j^2$ ignoring covariance parts. If the event of double recombinations in the interval is negligible, this amount can be expressed as $4p(1-p)(E+2D) \times a_j^2$ in a single QTL model. For p QTL, Q_j, in p distinct intervals, (M_j, N_j), $j = 1, 2, \ldots, p$, the amount of inflation is about $\sum_{j=1}^{p} 4p_j(1-p_j)(E_j + 2D_j) \times a_j^2$, where $p_j = r_{1j}/r_j$ (r_j and r_{1j} are the recombination fractions between (M_j, N_j) and between (M_j, Q_j), and E_j (D_j) is the frequency of M_jN_j/M_jn_j (M_jn_j/M_jn_j) in the population. There is no inflation if QTLs are completely observed (coincident with markers). Therefore, when QTLs are located at intervals and inferred from flanking markers, the inflation of residual variance reduces the QTL detection power as compared to the power of detecting completely observed QTL (see also section 5).

The above analyses decompose equation (12) into components related to sample size, QTL effects, distance between genes, interval size and genotypic distribution of a population. They pave the way to predict and analyse the power of separation under these factors, across populations and using different methods, and to conduct the QTL analysis when QTLs are completely observed (coincident with markers) or not observable (located in the markers intervals). The validity of proposed formulae in predicting the power of separating linked QTLs is first checked by Monte Carlo simulations, and then the formulae are applied to the power analysis under several mapping factors in different populations.

3. Simulation

We consider the case of M_j-Q_j-N_j-M_k-Q_k-N_j order in the F_2 population. We assume that all markers are 10 cM apart, and the two QTLs are located in the middle of their intervals. We set $h^2 = 0{\cdot}2$, $a_j = 1$ and $a_k = 1$. With such a setting, $\mathrm{var}(w_{ij}) = \mathrm{var}(w_{ik}) = 0{\cdot}4522$ and $\rho(w_{ij}, w_{ik}) = 0{\cdot}7445$ in REG interval mapping. The predicted powers by REG interval mapping are 2·34, 8·70, 20·09, 34·27, 48·35, 60·63 and 70·61% for $n = 200$, 250, 300, 350, 400, 450 and 500 at $\alpha = 0{\cdot}005$. For ML interval mapping, $\mathrm{var}(x_{ij}^*) = \mathrm{var}(x_{ik}^*) = 0{\cdot}4515$ and $\rho(x_{ij}^*, x_{ik}^*) = 0{\cdot}7272$. The predicted powers by ML interval mapping are 3·90, 12·59, 26·18, 41·45, 55·54, 67·19 and 76·27% for the six different sample sizes at the same α level. Under each case, 200 simulated replicates were generated to obtain the observed powers. The observed power is the proportion of replicates with both test statistics larger than the critical value. For both methods, their observed powers are compared with the predicted powers for each case and plotted in Fig. 1. It indicates that the observed and predicted powers by REG and ML interval mapping are reasonably close to each other under the given sample sizes. Thus, simulation results validate our proposed formulae.

4. Numerical analysis

On the basis of our proposed formulae, numerical analyses of the power of dissecting closely linked QTLs under various mapping factors and in different experimental populations are shown in Figs 2(*a–d*). The factors considered are sample size, QTL effect, interval size and distance between QTLs, and the populations considered include the F_2, AI and RI. Also, both REG and ML interval mapping are applied to the power analysis. In all the cases, we assume $h^2 = 0{\cdot}2$. Figure 2(a) shows the power curves of

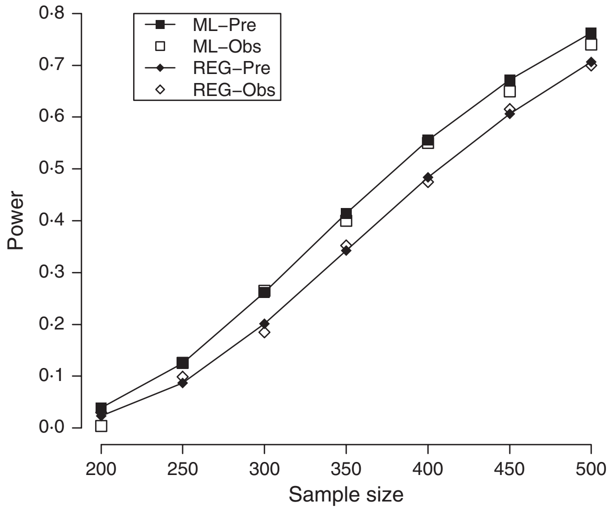

Fig. 1. The predicted and observed powers obtained by ML and REG interval mapping under different sample sizes in the F_2 population. The order considered is M_j-Q_j-N_j–M_k-Q_k-N_k. The two QTLs have equal effects and are located in the middle of the 10 cM spaced intervals. The distance between QTLs is 20 cM and $h^2=0{\cdot}2$.

separating two QTLs located in 10 or 20 cM spaced marker intervals under different distances. The order considered is M_j-Q_j-N_j–M_k-Q_k-N_k, and both QTL are located right in the middle of their intervals ($d_{\mathrm{M_jQ_j}}=5$ cM, $d_{\mathrm{Q_jN_j}}=5$ cM, $d_{\mathrm{M_kQ_k}}=5$ cM and $d_{\mathrm{Q_kN_k}}=$ 5 cM in the case of the 10 cM intervals, and $d_{\mathrm{M_jQ_j}}=10$ cM, $d_{\mathrm{Q_jN_j}}=10$ cM, $d_{\mathrm{M_kQ_k}}=10$ cM and $d_{\mathrm{Q_kN_k}}=10$ cM in the case of the 20 cM intervals). The distances between QTLs are 20, 25, 30, 35, 40, 45 and 50 cM, respectively ($d_{\mathrm{N_jM_k}}=10$, 15, 20, 25 and 30 cM in the 10 cM intervals, and $d_{\mathrm{N_jM_k}}=0$, 5, 10, 15 and 20 cM in the 20 cM intervals). The two QTLs have equal effects and the sample size is 200. It shows that, given a distance between QTLs, the powers of separation are larger when they are in the narrow intervals. Also, the powers by ML interval mapping is higher than those by REG interval mapping. As mentioned earlier, separating linked QTLs is the most difficult for the case of M_j-Q_j-N_j-Q_k-N_k (M_j-Q_j-N_k-Q_k-N_k) order, because they share a common flanking marker. Figures 1 *b–d* present the powers of separating two 10-cM-apart QTLs in the F_2, AI and RI populations for this order. Assume that $d_{\mathrm{M}_j}\mathrm{Q}_j=$ 5 cM, $d_{\mathrm{Q}_j}\mathrm{N}_j=5$ cM, $d_{\mathrm{N}_j}\mathrm{Q}_k=5$ cM and $d_{\mathrm{Q}k}\mathrm{N}_k=5$ cM, and that the QTLs have equal effects. In Fig. 2 *b*, with 500 sample size, the powers of REG and ML interval mapping are very low (close to zeros) in the F_2 and F_3 populations. But, the powers increase in the more advanced populations. The powers increase to 0·238 and 0·670 using REG interval mapping in AI F_6 and RI F_6 populations, and they increase to 0·367 and 0·741, respectively, using the ML method. Figure 2 *b* also presents the powers of separation when Q_j and Q_k are completely observed (and fitted into the model). As expected, the powers are greater when they are completely observed (the curves with solid and empty triangles). For example, the power is 0·427 in F_2, and it becomes 0·732 and 0·925 in AI F_3 and RI F_3 populations, respectively. The powers gradually attain more than 0·99 for more advanced populations. Figure 2 *c* shows the powers of separating two fully observed linked QTLs under different sample sizes. The QTLs have equal effects. The powers are about 0, 0·001, 0·037, 0·198 and 0·427 for $n=100$, 300 and 500, respectively, in the F_2 population, and are 0·059 (0·032), 0·194 (0·518), 0·565 (0·862), 0·815 (0·968) and 0·930 (0·994), respectively, in the AI F_5 (RI F_5) populations. This shows that advanced populations can be much more efficient, and that the RI populations can be more powerful than the AI populations in separation. Figure 2 *d* illustrates the relations between power and sample size when separating 10-cM-apart QTL with different sizes in the F_2 population. The QTLs are assumed to be completely observed. The powers of separating QTLs with similar size (e.g. $a_j : a_k = 1 : 1$) are higher than those of separating QTLs with different size (e.g. $a_j : a_k = 2 : 1$), and that the powers for separating QTLs with different direction of effects (e.g. $a_j : a_k = 1 : -1$) is much higher than those with the same direction of effects (e.g. $a_j : a_k =$ $1 : 1$). For example, the powers are 0·236, 0·344, 0·427, 0·981 and 1·000 (0·298) for the effect ratio 1 : 2, 1 : 1·5, 1 : 1, 1 : −1·5 and 1 : −1 with $n=500$, respectively. In general, an effective separation of closely linked QTLs requires large n, high h^2, and small ρ and more QTL information in a population.

5. Discussion

QTL mapping is a key approach to the understanding and estimation of the genetic architectures of quantitative traits in quantitative genetics (Zeng *et al.*, 1999). In QTL mapping, when QTLs are tightly linked, the estimation of QTL parameters could be easily biased, and the power of detection could be reduced. Therefore, the study of detecting and separating the linked QTLs correctly and efficiently remains an important issue in QTL mapping (Lander & Botstein, 1989, Ronin *et al.*, 1999; Hu & Xu, 2008). We tackle this issue by developing test statistics to test the effects of QTLs located at the markers or in the intervals. Both the REG and ML interval mapping models are considered. By well characterizing the genotypic distributions of three and four genes, we are able to evaluate the variances and covariances of the predictor variables of QTL in the models, and then to construct test statistics for detecting linked QTLs under more wide-ranging situations. Our proposed test statistics are simple functions of information numbers, VIF and genetic parameters in the models in the populations. They allow us to predict the power

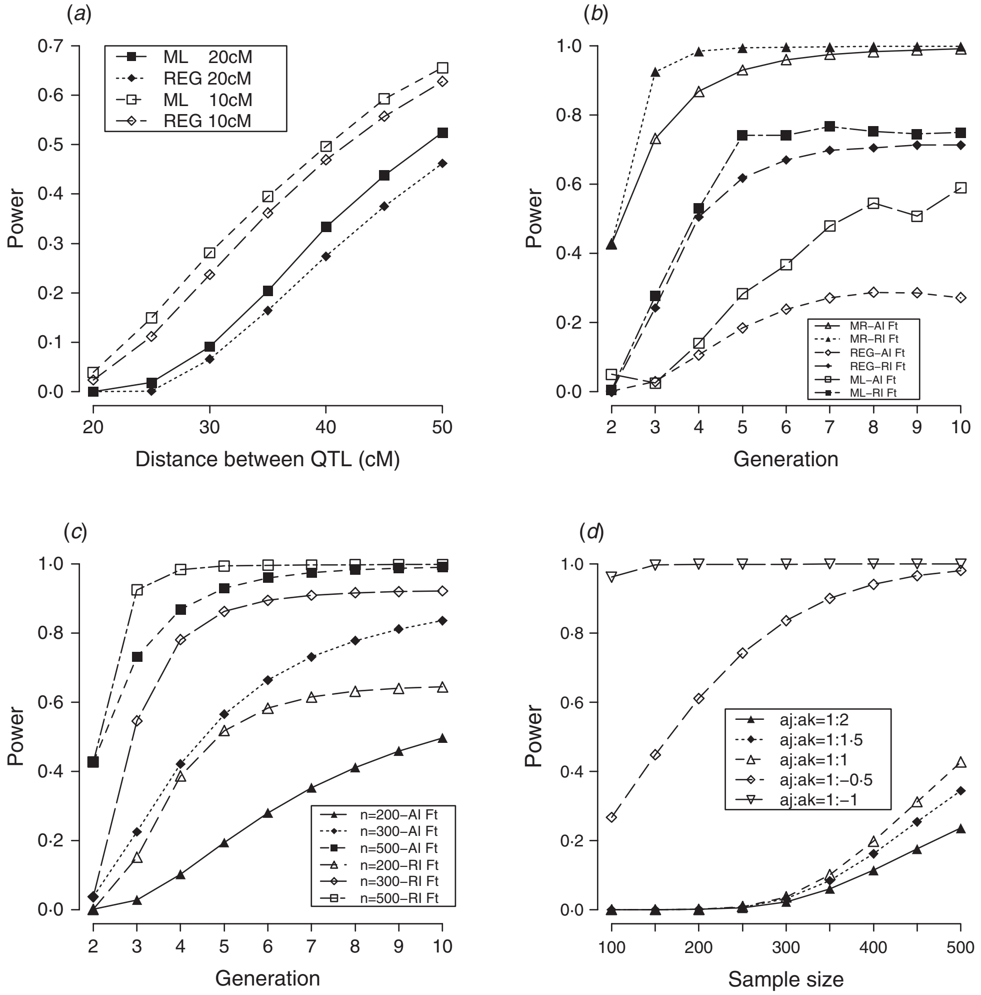

Fig. 2. (a) Power curves of separating two linked QTLs located in the middle of the 10- or 20-cM-spaced marker intervals under various distances in the F_2 population. The order considered is M_j-Q_j-N_j–M_k-Q_k-N_k. The distances between QTLs are 20, 25, 30, 35, 40, 45 and 50 cM, respectively. The two QTLs have equal effects, and $n = 200$. (b) Power curves of separating two 10-cM-apart QTLs when QTLs are coincident with markers (MR) or located in the intervals (REG and ML) in the AI and RI populations. QTLs have equal effects and $n = 500$. The order considered is M_j-Q_j-N_j-Q_k-N_k. (c) Power curves of separating two 10-cM-apart QTLs under different sample sizes in the AI and RI populations. QTLs have equal effects and are located at markers. (d) Power curves of separating two 10-cM-apart QTLs with different sizes of effects under different sample sizes in the F_2 population. QTLs are assumed to be located at markers. In all cases, $h^2 = 0{\cdot}2$. $\alpha = 0{\cdot}005$ is chosen as the significant level.

of separating linked QTLs under different mapping factors and across different populations. The direct application of our approach to QTL mapping requires the intervals potentially localizing QTL are known for testing. However, those potential intervals are not known before implementing the preliminary analysis. To identify the potential intervals, the use of multi-dimensional search, such as screening all pairs of close intervals, along the whole genomes may not be appropriate, as it can be subjected to a substantial computational burden. In practice, one suggestion is to first use one-QTL model analysis (one-dimensional search) to identify the regions containing potential intervals. In the likelihood profiles of the one-dimensional search, the regions showing significant sign changes in the estimated QTL effects or showing wide and significant peaks (ghost QTL) may indicate containing potential intervals (Haley & Knott, 1992; Kao *et al.*, 1999; Zeng *et al.*, 1999). Then our approach can be applied to these potential intervals for further analysis of closely linked QTLs.

The different advanced populations have different population structures, such as homozygosities, linkage disequilibria (correlations between genes) and genotypic frequencies (Weir, 1996). Therefore, they will show different properties in the resolution of closely linked QTLs. When QTLs are linked, their correlation can be generally formulated as $1-2R$, where R is the proportion of recombinants in a population. In a population, the closer they are linked, the less recombinants are produced and the stronger the correlation is. Fitting linked QTLs is equivalent to fitting correlated variables into the model, which cause the problems of collinearity in statistical estimation. Consequently, the separation becomes more difficult for closer QTLs as the collinearity problem becomes more severe. The obvious way to relieve the collinearity problem is to increase the proportion of recombinant in a population. In the BC or F_2 populations, the proportion of recombinants is equivalent to the recombination fraction between QTLs ($R=r$). In the AIL and RIL populations, more recombinants can be produced and accumulated, so that $R>r$ as generation proceeds. Then, these advanced populations would provide smaller VIF and reduce correlations for the QTL parameters to facilitate QTL detection. Nevertheless, we should know that the sizes of marker intervals localizing QTLs may expand (relative to that in the backcross or F_2 population) in the more advanced AI populations (Lynch & Walsh, 1998; Kao & Zeng, 2009) so that the benefit may be offset. Greatly improving separation in the AI populations requires denser markers around the detected QTL (QTL located in narrow intervals), and the improvement would be limited if QTLs are in the sparse marker region (wide intervals). The more powerful separation in the later RI population is also due to the increase of additive variances (accumulation of homozygotes). For example, the additive variance of a QTL in RIL can be twice of that in the F_2 population, and the power of separation can be much higher in the RIL (see Fig. 2 *b*, *c*). By well utilizing the properties of genome structures in the later advanced populations, it is possible to improve the resolution of closely linked QTLs in QTL detection.

Given a distance between QTLs, the powers of separating QTL at the markers are greater than those in the intervals (Fig. 2 *b*). To detect QTLs located in the intervals, the REG or ML interval mapping models have been very popular and used in the separation. In either one of the two statistical models, when the flanking marker intervals become wider or the locations of QTLs are closer to the middle of the intervals, the variances of predictor variables become smaller and their correlations become larger (not shown). Consequently, their detection would be more difficult (Fig. 2 *a*). Our proposed formulase can take the parameters of QTLs positions and effects and the population structures together into account to predict the power of separation. In general, given a distance between QTL, separation can be more effective for QTLs of similar size, located closer to markers and in narrow intervals, with opposite direction effects, and contributing to a high proportion of trait variation. Also, it is possible to gain more power in QTL detection by utilizing more advanced populations. The results may facilitate the analysis of QTL resolution in the genetic study of quantitative traits.

The authors are grateful to the editor and one anonymous reviewer for helpful comments. We are also grateful to Hsiang-An Ho for writing the computer programme. This work was supported by a grant NSC98-2118-M-001-018 from the National Science Council, Taiwan, Republic of China.

References

Fuller, W. A. (1976). *Introduction to Statistical Time Series*. New York: Wiley.

Genz, A. & Bretz, F. (2009). *Computation of Multivariate Normal and t probabilities*. New York: Springer.

Geiringer, H. (1944). On the probability theory of linkage in Mendelian heredity. *The Annals of Mathematical Statistics* **15**, 25–57.

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

Haley, C. S. & Knott, S. A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**, 315–324.

Hu, Z. & Xu, S. (2008). A simple method for calculating the statistical power for detecting a QTL located in a marker interval. *Heredity* **101**, 48–52.

Kao, C. H. & Zeng, Z. B. (1997). General formulas for obtaining the MLE and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm. *Biometrics* **53**, 653–655.

Kao, C.-H., Zeng, Z.-B. & Teasdale, R. (1999). Multiple interval mapping for quantitative trait loci. *Genetics* **152**, 1203–1216.

Kao, C. H. & Zeng, M. H. (2009). A study on mapping quantitative trait loci in the advanced populations derived from two inbred lines. *Genetics Research* **91**, 85–99.

Lander, E. S. & Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185–199.

Lynch, M. & Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sunderland, MA: Sinauer Associates.

Marquardt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and Nonlinear Estimation. *Technometrics* **12**, 591–612.

Mayer, M. (2005). A comparison of regression interval mapping and maximum likelihood interval mapping for linked QTL. *Heredity* **94**, 599–905.

Ronin, Y. I., Korol, A. B. & Nevo, E. (1999). Single- and multiple-trait mapping analysis of linked quantitative trait loci: Some asymptotic analytical approximation. *Genetics* **151**, 387–396.

Soller, M., Brody, T. & Genizi, A. (1976). On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. *Theoretical and Applied Genetics* **47**, 35–39.

Weir, B. S. (1996). *Genetic Data Analysis II*. Sunderland, MA: Sinauer Associates.

Xu, S. (1995). A comment on the simple regression method for interval mapping. *Genetics* **141**, 1657–1659.

Zeng, Z. B., Kao, C. H. & Basten, C. (1999). Estimating the genetic architecture of quantitative traits. *Genetics Research* **74**, 279–289.

Appendix. Genotypic distribution in advanced populations

Consider an F_2 or advanced population derived from two inbred lines P_1 and P_2. For m genes, there are 2^m different gametic genotypes and $2^{2m-1}+2^m/2$ zygotic genotypes. For example, there are 4, 8 and 16 gametic genotypes and 10, 36 and 136 zygotic genotypes for $m=2$, 3 and 4. As different populations undergo various number of meiosis cycle, the distributions of gametic and zygotic genotypes vary. For selfing, Haldane & Waddington (1931) formulated the transition equations of the ten genotypic frequencies for $m=2$. Kao & Zeng (2009) obtained the transition equations of the 36 genotypic frequencies for $m=3$. The procedures of obtaining transition equations for $m=4$ are given below. Let 1 and 0 represent the capital and small-letter alleles, respectively, from P_1 and P_2, so that the configurations of the 16 gametes can be represented as $\underline{1111}$, $\underline{0000}$, $\underline{1110}$, $\underline{0001}$, $\underline{1101}$, $\underline{0010}$, $\underline{1011}$, $\underline{0100}$, $\underline{0111}$, $\underline{1000}$, $\underline{1100}$, $\underline{0011}$, $\underline{1010}$, $\underline{0101}$, $\underline{1001}$ and $\underline{0110}$. In the F_2 population, these 16 gamete frequencies can be obtained the under Haldane map function (using the Markov property), and they are $P(\underline{1111})=P(\underline{0000})=(1-r_1)(1-r_2)(1-r_3)/2$, where r_1, r_2 and r_3 are the recombination rates between the first and second genes, between the second and third genes and the third and four genes, respectively. The other frequencies are $P(\underline{1110})=P(\underline{0001})=(1-r_1)(1-r_2)r_3/2$, $P(\underline{1101})=P(\underline{0010})=(1-r_1)r_2r_3/2$, $P(\underline{1011})=P(\underline{0100})=r_1r_2(1-r_3)/2$, $P(\underline{0111})=P(\underline{1000})=r_1(1-r_2)(1-r_3)/2$, $P(\underline{1100})=P(\underline{0011})=(1-r_1)r_2(1-r_3)/2$, $P(\underline{1010})=P(\underline{0101})=r_1r_2r_3/2$ and $P(\underline{1001})=P(\underline{0110})=r_1(1-r_2)r_3/2$, respectively. The random unification of these 16 gametes frequencies will produce the 136 different zygotes in a population. If selfing persists after F_2 to produce RI populations, the transition equation for the frequency of $\frac{1111}{1111}$ genotype is

$$
\begin{aligned}
P_{t+1}\left(\frac{1111}{1111}\right) &= P_t\left(\frac{1111}{1111}\right)+\frac{1}{4}P_t\left(\frac{1111}{1110}\right)+\frac{1}{4}P_t\left(\frac{1111}{1011}\right)+\frac{1}{4}P_t\left(\frac{1111}{1101}\right)+\frac{[(1-r_2)(1-r_3)+r_2r_3)]^2}{4}P_t\left(\frac{1111}{1010}\right)\\
&+\frac{[r_2(1-r_3)+r_3(1-r_2)]^2}{4}P_t\left(\frac{1110}{1011}\right)+\frac{(1-r_3)^2}{4}P_t\left(\frac{1111}{1100}\right)+\frac{r_3^2}{4}P_t\left(\frac{1110}{1101}\right)\\
&+\frac{(1-r_2)^2}{4}P_t\left(\frac{1111}{1001}\right)+\frac{r_2^2}{4}P_t\left(\frac{1101}{1011}\right)+\frac{(1-r_2)^2(1-r_3)^2}{4}P_t\left(\frac{1111}{1000}\right)+\frac{r_2^2(1-r_3)^2}{4}P_t\left(\frac{1100}{1011}\right)\\
&+\frac{r_2^2r_3^2}{4}P_t\left(\frac{1010}{1101}\right)+\frac{(1-r_2)^2r_3^2}{4}P_t\left(\frac{1001}{1110}\right)+\frac{1}{4}P_t\left(\frac{1111}{0111}\right)\\
&+\frac{[(1-r_1)[(1-r_2)(1-r_3)+r_2r_3]+r_1[r_2(1-r_3)+r_3(1-r_2)]]^2}{4}P_t\left(\frac{1111}{0110}\right)\\
&+\frac{[(1-r_1)[r_2(1-r_3)+r_3(1-r_2)]+r_1[(1-r_2)(1-r_3)+r_2r_3]]^2}{4}P_t\left(\frac{1110}{0111}\right)\\
&+\frac{(1-r_1)^2[(1-r_1)(1-r_2)+r_1r_2]^2}{4}P_t\left(\frac{1111}{0010}\right)+\frac{r_1^2[r_1(1-r_2)+r_2(1-r_1)]^2}{4}P_t\left(\frac{1011}{0110}\right)\\
&+\frac{r_1^2[(1-r_1)(1-r_2)+r_1r_2]^2}{4}P_t\left(\frac{1010}{0111}\right)+\frac{(1-r_1)^2[r_1(1-r_2)+r_2(1-r_1)]^2}{4}P_t\left(\frac{1110}{0011}\right)\\
&+\frac{[(1-r_1)(1-r_2)+r_1r_2]^2}{4}P_t\left(\frac{1111}{0101}\right)+\frac{[r_1(1-r_2)+r_2(1-r_1)]^2}{4}P_t\left(\frac{1101}{0111}\right)\\
&+\frac{[(1-r_1)(1-r_2)+r_1r_2]^2(1-r_3)^2}{4}P_t\left(\frac{1111}{0100}\right)+\frac{(1-r_3)^2[r_1(1-r_2)+r_2(1-r_1)]^2}{4}P_t\left(\frac{1100}{0111}\right)\\
&+\frac{r_3^2[(1-r_1)(1-r_2)+r_1r_2]^2}{4}P_t\left(\frac{1110}{0101}\right)+\frac{r_3^2[r_1(1-r_2)+r_2(1-r_1)]^2}{4}P_t\left(\frac{1101}{0110}\right)\\
&+\frac{[(1-r_1)(1-r_2)]^2}{4}P_t\left(\frac{1111}{0001}\right)+\frac{r_1^2r_2^2}{4}P_t\left(\frac{1011}{0101}\right)+\frac{(1-r_1)^2r_2^2}{4}P_t\left(\frac{1101}{0011}\right)+\frac{r_1^2(1-r_2)^2}{4}P_t\left(\frac{1001}{0111}\right)\\
&+\frac{[(1-r_1)(1-r_2)(1-r_3)]^2}{2}P_t\left(\frac{1111}{0000}\right)+\frac{[r_1(1-r_2)(1-r_3)]^2}{2}P_t\left(\frac{1000}{0111}\right)
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{[(1-r_1)r_2(1-r_3)]^2}{2}P_t\left(\frac{1100}{0011}\right)+\frac{[r_1r_2r_3]^2}{2}P_t\left(\frac{1010}{0101}\right)+\frac{[(1-r_1)(1-r_2)r_3]^2}{2}P_t\left(\frac{1110}{0001}\right)\\
&+\frac{[r_1r_2(1-r_3)]^2}{2}P_t\left(\frac{1011}{0100}\right)+\frac{[r_1(1-r_2)r_3]^2}{2}P_t\left(\frac{1001}{0110}\right)+\frac{[(1-r_1)r_2r_3]^2}{2}P_t\left(\frac{1101}{0010}\right)\\
&+\frac{(1-r_1)^2}{4}P_t\left(\frac{1111}{0011}\right)+\frac{r_1^2}{4}P_t\left(\frac{1011}{0111}\right).
\end{aligned}
$$

The above equation contains 41 terms and is derived below. Among all 136 zygotes, 41 of them are capable of producing $\underline{1111}$ gamete with different proportions. Therefore, the frequency of $\frac{1111}{1111}$ zygote in $t+1$ generation is equivalent to the sum of frequencies of $\frac{1111}{1111}$ progeny of these 41 parental zygotes in t generation. The proportion of $\frac{1111}{1111}$ progeny from $\frac{1111}{1111}$ parents is 100%. The proportion of $\frac{1111}{1111}$ progeny from $\frac{1111}{1110}$, $\frac{1111}{1011}$ and $\frac{1111}{1101}$ parents are 1/4. The proportions from $\frac{1111}{1010}$, $\frac{1110}{1011}$, $\frac{1001}{0110}$ and $\frac{1101}{0010}$ parents are $[(1-r_2)(1-r_3)+r_2r_3)]^2/4$, $[r_2(1-r_3)+r_3(1-r_2)]^2/4$, $[r_1(1-r_2)r_3]^2/4$ and $[(1-r_1)r_2r_3]^2/4$, respectively. Likewise, The proportion of $\frac{1111}{1111}$ progeny from the other genotypes can be also obtained. Because of symmetry, there are 72 transition equations in total, and the remaining 71 equations can be formulated in a similar way. The complete 72 equations and the computer programme (written in *R* language) are provided in Supplementary materials. The computer programme is also placed at http://www.stat.sinica.edu.tw/~chkao for download. If mating is random, the transition equations for obtaining the frequencies of gametic genotypes for any *m* can be derived by using Geiringer's approach (1944), and then the frequencies of zygotic genotypes can be obtained.