[Frontiers in Bioscience E4, 2670-2682, June 1, 2012]

A score-statistic approach for determining threshold values in QTL mapping

Chen-Hung Kao[1], Hsiang-An Ho[1]

[1]*Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan, R.O.C.*

TABLE OF CONTENTS

1. ABSTRACT

Issues in determining the threshold values of QTL mapping are often investigated for the backcross and F2 populations with relatively simple genome structures so far. The investigations of these issues in the progeny populations after F2 (advanced populations) with relatively more complicated genomes are generally inadequate. As these advanced populations have been well implemented in QTL mapping, it is important to address these issues for them in more details. Due to an increasing number of meiosis cycle, the genomes of the advanced populations can be very different from the backcross and F2 genomes. Therefore, special devices that consider the specific genome structures present in the advanced populations are required to resolve these issues. By considering the differences in genome structure between populations, we formulate more general score test statistics and Gaussian processes to evaluate their threshold values. In general, we found that, given a significance level and a genome size, threshold values for QTL detection are higher in the denser marker maps and in the more advanced populations. Simulations were performed to validate our approach.

2. INTRODUCTION

The statistical model of interval mapping (IM) proposed by Lander and Botstein (1) is generally a normal mixture model, as the genotypes of the quantitative trait locus (QTL) are not observable and needed to be inferred from its flanking markers. In the parameter estimation of the normal mixture model, the maximum likelihood estimation is commonly implemented to obtain the maximum likelihood estimates (MLE) through the EM algorithm (2) by treating the model as an incomplete-data problem. Typically, the presence of a QTL, i.e. the null hypothesis of no QTL, is tested over the all possible positions in the whole genome by using likelihood ratio test (LRT) statistics, and the position with the significantly maximum LRT statistic is regarded as the estimated QTL location. Under this framework, it has been recognized that the determination of the threshold values for claiming significant QTL detection (rejecting the null hypothesis) along the genomes is one of the complicated and important issues in QTL mapping (3-4) for the following reasons. One is that the QTL position is unidentified under the null hypothesis, and the maximum LRT statistic does not follow

the standard x^2 distribution asymptotically (4). Further, various factors, such as the number and size of intervals, population genome structures, and informativeness of markers, will involve in and should be considered in determining the threshold value for claiming QTL detection (5-8). Besides, because multiple correlated and uncorrelated tests are performed in searching for QTLs on the whole genome, the common pointwise significance level is not appropriate and genomewise significance level should be considered in QTL mapping (1,4,6,8).

Several theoretical and simulation approximations have been proposed to determine the threshold values of QTL detection. Lander and Botstein (1) suggested using Bonferroni argument for sparse-map case and ORENSTEIN-UHLENBECK diffusion for dense-map case to determine the threshold value. For intermediate situations, extensive numerical simulations have been used to determine the thresholds (1,9). Churchill and Doerge (10) suggested using a permutation test for determining an appropriate threshold values for specific data sets. Rebai, Goffinet and Mangin (11) used Davies's bound (12) to derive a conservative formulas for calculating the approximate thresholds for backcross and F_2 populations. They demonstrated good performance of their formulas using simulation. Dupuis and Siegmund (13) provided approximate formulas to calculate the threshold for the case of very dense markers in the population, but they did not take interval mapping into account in their approximation. Piepho (14) also used Davies's bound to propose a quick method for computing the approximate thresholds for general designs. The quick method is computationally inexpensive and claimed to be an alternative to permutation procedure. Zou *et al.* (8) and Chang *et al.* (4) proposed a score-statistic framework to assess the threshold values. As the maximum of the square of score test statistics is asymptotically equivalent to the maximum LRT statistics, the threshold values derived from the score test statistics can be used as those for the LRT statistics in the population (4, 8, 15-16).

On the basis of score test statistics, Zou *et al.* (8) proposed a resampling approach to obtain the threshold values mainly for the F_2 population by simulating the F_2 genome structure. Chang *et al.* (4) also devised a score-statistic method for computing the threshold values in a backcross population by analytically analyzing the backcross genome structure. Chang *et al.* (4) showed that score test statistics along the genomes is a Gaussian stochastic process with mean zero and well-structured covariance, and they used them to compute the threshold values of QTL detection in the backcross population. The score-statistic method not only can be less computationally demanding than the permutation test and numerical simulation, but also can be more accurate than previous approximate formulas in the computation of the threshold values (4,8). So far, most of the studies of assessing the threshold values for QTL detection are investigated in the backcross and F_2 populations (8, 11 and 14 discussed the threshold values for F_3 populations), still they are generally lacking or inadequate for the progeny populations after F_2 (advanced populations). These advanced populations, such as recombinant inbred (RI) and advanced intercrossed (AI) populations, have been well devised and implemented in genetic studies. For example, Bai *et al.* (17) used RI populations in rice and Kelly *et al.* (18) implemented AI populations in mice for investigating the genetic architectures of quantitative traits in their studies. For specific populations where time is not an issue, the advanced populations can have some very useful properties in that their genomic structures allow researchers to yield better results in their investigations. Therefore, it is of importance to address the issues of assessing the threshold values for these populations in more details in QTL mapping study. Due to the fact that these advanced populations undergo multiple meiosis cycles, their genomic structures, such as homozygosity, genotypic frequency and variance components, are differing and can be very different from the backcross and F_2 genomes. In this article, by distinguishing between different population genome structures, we formulate more general score test statistics and Gaussian processes under the framework of interval mapping to compute the threshold values and to study their behaviors for various populations. One of the keys to our approach is to devise the genotypic distributions of two, three and four genes of the populations into the formulations, so that their specific genome structures can be well described to address these issues across various populations. Simulation studies are performed to evaluate our approach to assessing the threshold values. The R program of our approach is available on http://www.stat.sinica.edu.tw/~chkao/.

3. EXPERIMENAL POPULATIONS

3.1. Advanced populations

Various experimental populations have been designed for the study of QTL mapping. Among these populations, the backcross and F_2 populations have been the most widely used designs in the studies. Besides, the progeny populations from the F_2, which are called advanced populations, are also very common. These experimental populations are produced as follows. A cross between two parental inbred lines, P_1 and P_2, is performed to produce an F_1 population. If the F_1 individuals are backcrossed to P_1 or P_2, it produces the backcross population. If the F_1 individuals are selfed or randomly mated, it produces an F_2 population. If the F_2 population is further selfed and/or randomly mated for generations, the produced progeny populations will be called advanced populations. These advanced populations include recombinant inbred (RI) and advanced intercrossed (AI) populations. As the RI populations or AI populations is obtained by repeatedly selfing (inbreeding) or randomly intermating the F2 individuals for several generations, their genomes will be subject to more meiosis cycles as compared to the backcross and F2 populations. Therefore, the genomic constitutions, such as homozygosity, genotypic distribution and linkage disequilibrium, of the advanced populations will be different from each other (19) and no longer similar to those of the backcross and F2 populations. In general, selfing will increase the homozygosity at the expense of heterozygotes. Also, further meiosis cycles tend to

accumulate crossovers so that the proportion of recombinants will increase and linkage disequilibrium between genes will reduce in the populations. These features in the advanced populations can be very useful and may allow for more productive investigations. For example, the RI populations can increase the homozygosity to assist the detection of additive effects in QTL mapping (20). Further, the AI populations can harbor more recombination events in a short chromosome segment for genetic fine mapping and may provide better power in the separation of closely linked QTL (20-21).

3.2. Genotypic distributions

As will be shown and explained later, an important key to successfully compute the threshold values for these populations is to well devise the genotypic distributions of one, two, three and four genes of the populations into the formulations, so that the specific genome structures of the populations can be well considered under our proposed framework. We now explain briefly how these genotypic distributions are obtained in different populations. Consider an F_2, AI or RI population used for the QTL mapping studies. In general, for m genes, there are 2^m different gametic genotypes and $2^{(2m-1)}+2^m/2$ zygotic genotypes. Therefore, for m=1, 2, 3 and 4, there are 2, 4, 8 and 16 gametic genotypes and 3, 10, 36 and 136 zygotic genotypes. As the different populations undergo different numbers of meiosis cycle, they will have different distributions of gametic and zygotic genotypes in the genomic constitutions. For one gene, there are three possible genotypes, P_1 homozygote, heterozygote and P_2 homozygote. The frequencies of these three genotypes are expected to be 1/4, 1/2 and 1/4, respectively, in the AI populations. In the RI populations, the frequency of heterozygote is halved in each selfing cycle. For m=2, 3 or 4, the genotypic frequencies in different AI and RI populations can be obtained by using the transition equations of Haldane and Waddington (22), Geiringer (23), and Kao and Zeng (20, 24). These transition equations, which are derived under the assumptions of Haldane map function and equal crossover value in two sexes, aim to obatin the genotypic distributions of the populations in different generations when individuals are subject to random mating or selfing process. In RI populations, the 5, 20 and 72 transition equations in Haldane and Waddington (22) and Kao and Zeng (20, 24) can be used to compute the frequencies of the 10, 36 and 136 genotypes of two, three and four genes. In AI populations, the gametic frequencies are first computed, and then the genotypic frequencies can be obtained from the product of gametic frequencies. To obtain the frequencies of the 4, 8 and 16 gametic frequencies for m=2, 3 and 4, Geiringer's approach (23) can be used to formulate the transition equations of gametic frequencies or the sets of transition equations provided by Kao and Zeng (20, 24) can be directly used. Conceptually, it is also possible to obtain the genotypic distributions for any number of genes in any AI and RI populations by extending their approaches. However, there may be too many equations, and each equation contains numerous terms.

4. STATISTCAL MOEL OF INTERVAL MAPPING FO QTL

4.1. Statistical model

The data structure of QTL mapping generally consists of two parts, y_j ($j=1,\cdots,n$) for the quantitative trait value and X_j ($j=1,\cdots,n$) for the genetic markers. An interval mapping statistical model for testing a QTL, Q, at the position x in an interval, I, flanked by markers, M (with alleles M and m) and N (with alleles N and n), is proposed as

$$y_i = \mu + ax_i^* + dz_i^* + \varepsilon_i \quad (1)$$

where y_i is the quantitative trait value of the i th individual, a and d are the additive and dominance effects of Q, x_i^* and z_i^* defined as

$$x_i^* \begin{cases} 1 & \text{if the genotype of Q is QQ,} \\ 0 & \text{if the genotype of Q is Qq,} \\ -1 & \text{if the genotype of Q is qq,} \end{cases} \quad \text{and}$$

$$z_i^* = \begin{cases} \frac{1}{2} & \text{if the genotype of Q is Qq,} \\ -\frac{1}{2} & \text{otherwise,} \end{cases}$$

are the coded variables of genotypes of Q, and ε_i is a random error. We assume ε_i follows N (0, σ^2). In general, as Q may not be coincident with a marker, its genotype is not observable and can be QQ ($x_i^*=1$ and $z_i^*=-1/2$), Qq ($x_i^*=0$ and $z_i^*=1/2$) or qq ($x_i^*=-1$ and $z_i^*=-1/2$) for an individual i .

4.2. Conditional probabilities

Although Q is not observable, its genotypic distribution can be inferred from its flanking marker genotype according to the principle of conditional probability as

$$P(\mathrm{Q} \mid \mathrm{M,N}) = \frac{P(\mathrm{MQN})}{P(\mathrm{MN})}. \quad (2)$$

Therefore, obtaining the above probability involves in the use of the genotypic distributions of two and three genes in the experimental populations. For any advanced population under consideration, the two flanking markers can have nine different genotypes, MN/MN, MN/Mn, Mn/Mn, MN/mN, MNmn (MN/mn or Mn/mN), Mn/mn, mN/mN, mN/mn and mn/mn. For each one of the nine marker genotypes, the genotype of Q can be QQ, Qq or qq . When M, N and Q are considered together, there are 27 different genotypes and 27 corresponding conditional probabilities. In the following, we denote these 27 conditional probabilities by p_{ij}'s, $i=1,2,\cdots,9$ indexing the marker genotypes and $j=1,2,3$ indexing the QTL genotype. It should be pointed out that, in the F_2

Table 1. General formulations for the conditional probabilities (mixing proportions) of a putative QTL, Q, flanked by two markers, M and N, in the advanced populations from two inbred lines

marker genotype	Trait mean	Expected frequency	Conditional probabilities of	Conditional probabilities of	Conditional probabilities of
			QQ	Qq	qq
$\frac{MN}{MN}$	$\overline{y_1}$	C	$\frac{A_1}{C}$	$\frac{A_2}{C}$	$\frac{A_3}{C}$
$\frac{MN}{Mn}$	$\overline{y_2}$	E	$\frac{A_4}{E}$	$\frac{A_5+A_6}{E}$	$\frac{A_7}{E}$
$\frac{Mn}{Mn}$	$\overline{y_3}$	D	$\frac{A_8}{D}$	$\frac{A_9}{D}$	$\frac{A_{10}}{D}$
$\frac{MN}{mN}$	$\overline{y_4}$	E	$\frac{A_{11}}{E}$	$\frac{A_{12}+A_{13}}{E}$	$\frac{A_{14}}{E}$
MmNn	$\overline{y_5}$	$F+G$	$\frac{A_{15}+A_{16}}{F+G}$	$\frac{A_{17}+A_{18}+A_{19}+A_{20}}{F+G}$	$\frac{A_{15}+A_{16}}{F+G}$
$\frac{Mn}{mn}$	$\overline{y_6}$	E	$\frac{A_{14}}{E}$	$\frac{A_{12}+A_{13}}{E}$	$\frac{A_{11}}{E}$
$\frac{mN}{mN}$	$\overline{y_7}$	D	$\frac{A_{10}}{D}$	$\frac{A_9}{D}$	$\frac{A_8}{D}$
$\frac{mN}{mn}$	$\overline{y_8}$	E	$\frac{A_7}{E}$	$\frac{A_5+A_6}{E}$	$\frac{A_4}{E}$
$\frac{mn}{mn}$	$\overline{y_9}$	C	$\frac{A_3}{C}$	$\frac{A_2}{C}$	$\frac{A_1}{C}$

The alleles of M, N and Q are denoted as (M,m), (N,n) and (Q,q), respectively.

$A_1 = P\left(\frac{MQN}{MQN}\right) = P\left(\frac{mqn}{mqn}\right), A_2 = P\left(\frac{MQN}{MqN}\right) = P\left(\frac{mqn}{mQn}\right), A_3 = P\left(\frac{MqN}{MqN}\right) = P\left(\frac{mQn}{mQn}\right), A_4 = P\left(\frac{MQN}{MQn}\right) = P\left(\frac{mqn}{mqN}\right), A_5 = P\left(\frac{MQN}{Mqn}\right) = P\left(\frac{mqn}{mQN}\right), A_6 = P\left(\frac{MqN}{MQn}\right) = P\left(\frac{mQn}{mqN}\right), A_7 = P\left(\frac{MqN}{Mqn}\right) = P\left(\frac{mQn}{mQN}\right), A_8 = P\left(\frac{MQn}{MQn}\right) = P\left(\frac{mqN}{mqN}\right), A_9 = P\left(\frac{MQn}{Mqn}\right) = P\left(\frac{mqN}{mQN}\right), A_{10} = P\left(\frac{Mqn}{Mqn}\right) = P\left(\frac{mQN}{mQN}\right), A_{11} = P\left(\frac{MQN}{mQN}\right) = P\left(\frac{mqn}{Mqn}\right), A_{12} = P\left(\frac{MQN}{mqN}\right) = P\left(\frac{mqn}{MQn}\right), A_{13} = P\left(\frac{MqN}{mQN}\right) = P\left(\frac{mQn}{Mqn}\right), A_{14} = P\left(\frac{MqN}{mqN}\right) = P\left(\frac{mQn}{MQn}\right), A_{15} = P\left(\frac{MQN}{mQn}\right) = P\left(\frac{mqn}{MqN}\right), A_{16} = P\left(\frac{MQn}{mQN}\right) = P\left(\frac{mqN}{Mqn}\right), A_{17} = P\left(\frac{MQN}{mqn}\right), A_{18} = P\left(\frac{MqN}{mQn}\right), A_{19} = P\left(\frac{mqN}{MQn}\right), A_{20} = P\left(\frac{Mqn}{mQN}\right).$

population, it is straightforward to obtain p_{ij}'s, as the genotypic distributions of two and three genes have a simple relationship with the recombination fractions between M, Q and N. For example, the frequencies of digenic gametes $\underline{Mq}$ and $\underline{qN}$ are $P(\underline{Mq}) = r_1/2$ and $P(\underline{qN}) = r_2/2$, where r_1 and r_2 are the recombination fractions between Q and M and between Q and N. The frequency of trigenic gamete $\underline{MqN}$ can be easily obtained by the product of the two digenic frequencies as $P(\underline{MqN}) = 2 \times P(\underline{Mq}) \times P(\underline{qN}) = r_1 r_2/2$, and the conditional probability of qq genotype given the flanking marker genotype MN/MN is simply as

$$P(qq/MMNN) = [P(\underline{MqN})/P(\underline{MN})]^2 = [r_1 r_2/(1-r)]^2,$$

where r is the recombination fraction between M and N. Nevertheless, such a simple relationship for straightforwardly computing the conditional probabilities does not hold in the more advanced populations. Here, we implement the sets of transition equations proposed by Haldane and Waddington (22), Geiringer (23) and Kao and Zeng (20, 24) to obtain the genotypic frequencies of two and three genes as have been mentioned in the previous section. With these frequencies, the 27 conditional probabilities of QTL genotypes given the flanking marker genotypes in any populations can be obtained (see Table 1). These conditional probabilities play very important roles in constructing the statistical QTL mapping model for precise QTL mapping.

5. LIKELIHOOD OF THE STATISTICAL MODEL

For an individual i, the unobservable Q at position x can be QQ, Qq or qq with certain probabilities depending on the flanking marker genotypes. Now let q_{i1}, q_{i2} and q_{i3} denote the three probabilities of Q being QQ, Qq or qq for the individual i, respectively, and these probabilities at position x can be obtained from the 27 conditional probabilities in Table 1. That is q_{ij}'s, $i=1,2,\cdots,n$, can be found from p_{kj}'s, $k=1,2,\cdots,9$. If Q is QQ (with probability q_{i1}), the distribution follows $N(\mu_1,\sigma^2)$, where $\mu_1=\mu-a+d/2$. Similarly, Q can be Qq or qq (with probability q_{i2} or q_{i3}), and the distribution follows $N(\mu_2,\sigma^2)$ or $N(\mu_3,\sigma^2)$, respectively, where $\mu_2=\mu-d/2$ and $\mu_3=\mu+a+d/2$. Therefore, the likelihood of an individual is a mixture of three normals with different means and mixing proportions, μ_j's and q_{ij}'s, $j=1,2,3$. For a sample with n individuals, the log likelihood function for $\theta=(\mu,a,d,\sigma^2)$ at position x is the sum of the log likelihood of the n individuals as

$$l(a,d,\mu,\sigma^2)=-\frac{n}{2}\log(2\pi\sigma^2)+\sum_{i=1}^{n}\log[\sum_{j=1}^{3}q_{ij}\times\exp(-\frac{(y_i-\mu_j)^2}{2\sigma^2})] \quad (3)$$

Note that q_{ij}'s can be determined by the position x and need not to be estimated here. To assist the following derivations of score statistics, we classify the n individuals into nine categories according to their marker genotypes and reformulate equation (3) as

$$l(a,d,\mu,\sigma^2)=-\frac{n}{2}\log(2\pi\sigma^2)+\sum_{i=1}^{9}\{\sum_{j=1}^{n_i}\log[\sum_{k=1}^{3}p_{ik}\times\exp(-\frac{(y_j-\mu_k)^2}{2\sigma^2})]\} \quad (4)$$

where n_1, n_2, n_3, n_4, n_5, n_6, n_7, n_8 and n_9 are the numbers of individuals with the nine marker genotypes MN/MN, MN/Mn, Mn/Mn, MN/mN, MNmn, Mn/mn, mN/mN, mN/mn and mn/mn, respectively. Note that the log likelihoods for the individuals with the same marker genotype have the same mixing proportions, p_{ik}'s, in the reformulated equation.

6. SCORE TEST STATISTICS

6.1. Score functions

The score functions for the additive and dominance effects are the derivatives of the log likelihood (Equation (4)) with respective to the parameters, a and d, and using the MLEs of μ and σ^2, $\hat{\mu}=\sum y_i/n$ and $\hat{\sigma}^2=\sum(y_i-\hat{\mu})^2/n$, evaluated under $H_0: a=0$ and $d=0$ at position x. Let $u_1(x)$ and $u_2(x)$ denote the score functions of a and d, respectively. The two score functions can be found as

$$u_1(x)=\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}[(q_{i1}-q_{i3})\times(y_i-\bar{y})] \quad and \quad u_2(x)=\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}[(q_{i2}-q_{i1}-q_{i3})\times(y_i-\bar{y})].$$

If the n individuals are classified into nine categories according to their marker genotypes, the score functions become

$$u_1(x)=\frac{n}{\hat{\sigma}^2}\sum_{i=1}^{9}\{(p_{i1}-p_{i3})-[\sum_{i=1}^{9}(p_{i1}-p_{i3})\times f_i]\}\times f_i\times\bar{y}_i \quad (5)$$

and

$$u_2(x)=\frac{n}{2\hat{\sigma}^2}\sum_{i=1}^{9}\{(p_{i2}-p_{i1}-p_{i3})-[\sum_{i=1}^{9}(p_{i2}-p_{i1}-p_{i3})\times f_i]\}\times f_i\times\bar{y}_i, \quad (6)$$

where f_i's ($f_i=n_i/n$) and $\bar{y}_i$'s, $i=1,2,\cdots,9$, are the frequencies of individuals and trait means in the nine flanking marker categories. To simplify the following derivations, now let k_i's and l_i's be

$$k_i=(p_{i1}-p_{i3})-[\sum_{i=1}^{9}(p_{i1}-p_{i3})\times f_i] \quad (7)$$

and

$$l_i=(p_{i2}-p_{i1}-p_{i3})-[\sum_{i=1}^{9}(p_{i2}-p_{i1}-p_{i3})\times f_i], \quad (8)$$

$i=1,2,\cdots,9$. Note that k_i's and l_i's are closely related to the genotypic distributions of two and three genes, as p_i's and f_i's are functions of the genotypic frequencies of two and three genes in the population. Then, under the null, the variances of $u_1(x)$ and $u_2(x)$ are

$$var(u_1(x))=\frac{n}{\hat{\sigma}^2}\sum_{i=1}^{9}k_i^2\times f_i \quad and \quad var(u_2(x))=\frac{n}{4\hat{\sigma}^2}\sum_{i=1}^{9}l_i^2\times f_i, \quad (9)$$

and the covariance between $u_1(x)$ and $u_2(x)$ is

$$cov(u_1(x),u_2(x))=\frac{n}{2\hat{\sigma}^2}\sum_{i=1}^{9}k_i\times l_i\times f_i. \quad (10)$$

6.2. Score test statistics

If only additive or dominance effect under the null, $H_0: a=0$ or $H_0: d=0$, is considered, the score test statistic, $U_1(x)$ or $U_2(x)$, is $u_1(x)$ or $u_2(x)$ divided by its standard deviation as

$$U_1(x)=\frac{u_1(x)}{\sqrt{var(u_1(x))}} \quad or \quad U_2(x)=\frac{u_2(x)}{\sqrt{var(u_2(x))}}.$$

If both additive and dominance effects are considered at a time, we may define the score function as $u(x)=(u_1(x)\ u_2(x))'$. The score test statistic, $U^2(x)$, for $H_0: a=0$ and $d=0$ against $H_1: a\neq 0$ and $d\neq 0$ at location x takes the form

$$U^2(x)=\begin{pmatrix}u_1(x) & u_2(x)\end{pmatrix}V^{-1}\begin{pmatrix}u_1(x)\\ u_2(x)\end{pmatrix}, \quad (11)$$

where V is the variance-covariance matrix of $u(x)$. The

elements in V are the variances and covariance of $u_1(x)$ and $u_2(x)$ in equations (9) and (10). The above derivations for score test statistics are relatively simpler as compared to the derivations for the MLE, as it avoids the parameter estimation of the normal mixture likelihood, which usually involves in the use of the iterated EM algorithm for obtaining the MLE (25). Also, as given by Cox and Hinkley (26) and Chang *et al.* (4), the maximum of the $U^2(x)$ is asymptotically equivalent to the maximum of LRT. Therefore, the maximum of $U^2(x)$ under the null hypothesis can be used to assess the threshold value of the maximum likelihood approach in QTL mapping.

6.3. Asymptotic forms of score test statistics

To understand the null distributions of the score test statistics, the asymptotically equivalent forms of $U_1(x)$, $U_2(x)$ and $U^2(x)$ are derived below. In derivations, we follow Haldane and Waddington (22) to use the oblique letters, C, D, E, E and G, to denote the genotypic frequencies. Let C be the frequency of MN/MN or mn/mn genotype, D be the frequency of Mn/Mn or mN/mN genotype, E be the frequency of MN/Mn, MN/mN, Mn/mn or mN/mn genotype, F be the frequency of MN/mn genotype, and G be the frequency of Mn/nM genotype, respectively. Note that the exact values of C, D, E, F and G in the different advanced populations can be computed using sets of transition equations as mentioned in the previous sections. Therefore, for large n in a population, we have $f_1 = f_9 = C$, $f_2 = f_4 = f_6 = f_8 = E$, $f_3 = f_7 = D$ and $f_5 = F + G$ asymptotically and can formulate the asymptotic forms of $u_1(x)$, $u_2(x)$, $var(u_1(x))$, $var(u_1(x))$ and $cov(u_1(x), u_2(x))$ in terms of C, D, E, F and G in any populations. We then have

$$U_1(x) \approx Z_1(x) = \frac{\sqrt{C} \times k_1 \times W_1 + \sqrt{E} \times k_2 \times W_2 + \sqrt{D} \times k_3 \times W_3 + \sqrt{E} \times k_4 \times W_4}{\sqrt{C \times k_1^2 + E \times k_2^2 + D \times k_3^2 + E \times k_4^2}} \quad (12)$$

where

$$W_1 = \frac{\bar{y}_{1\cdot} - \bar{y}_{9\cdot}}{\sqrt{2\sigma^2/(C \times n)}}, W_2 = \frac{\bar{y}_{2\cdot} - \bar{y}_{8\cdot}}{\sqrt{2\sigma^2/(E \times n)}}, W_3 = \frac{\bar{y}_{3\cdot} - \bar{y}_{7\cdot}}{\sqrt{2\sigma^2/(D \times n)}}, W_4 = \frac{\bar{y}_{4\cdot} - \bar{y}_{6\cdot}}{\sqrt{2\sigma^2/(E \times n)}}.$$

Also, due to symmetry, the term $\sum_{i=1}^{9}(p_{i1} - p_{i3}) \times f_i$ in k_i (equation (7)) will be zero and $k_i = p_{i1} - p_{i3}$ in a large population. Further, it can be shown that W_1, W_2, W_3 and W_4 all follow $N(0,1)$ asymptotically. Consequently, $Z_1(x)$ follows $N(0,1)$ asymptotically. Similarly, we have the score test statistic of the dominance effect as

$$U_2(x) \approx Z_2(x) = \frac{\sqrt{C} \times l_1 \times W_1^* + \sqrt{E} \times l_2 \times W_2^* + \sqrt{D} \times l_3 \times W_3^* + \sqrt{E} \times l_4 \times W_4^* + \sqrt{F+G} \times l_5 \times W_5^*}{\sqrt{C \times l_1^2 + E \times l_2^2 + D \times l_3^2 + E \times l_4^2 + (F+G) \times l_5^2}}, \quad (13)$$

where

$$W_1^* = \frac{\bar{y}_1 + \bar{y}_9}{\sqrt{2\sigma^2/(C \times n)}}, \; W_2^* = \frac{\bar{y}_2 + \bar{y}_8}{\sqrt{2\sigma^2/(E \times n)}}, \; W_3^* = \frac{\bar{y}_3 + \bar{y}_7}{\sqrt{2\sigma^2/(D \times n)}},$$

$$W_4^* = \frac{\bar{y}_4 + \bar{y}_6}{\sqrt{2\sigma^2/(E \times n)}}, \; W_5^* = \frac{\bar{y}_5}{\sqrt{2\sigma^2/[(F+G) \times n]}}.$$

Note that the term $\sum_{i=1}^{9}(p_{i2} - p_{i1} - p_{i3}) \times f_i$ in l_i (equation (8)) will be zero in the F_2 and AI populations, and it is $(0.5)^{t-2} - 1$ in the RI F_t populations. As W_1^*, W_2^*, W_3^*, W_4^* and W_5^* follow $N(0,1)$ asymptotically, it is also straightforward to show that $Z_2(x)$ follows $N(0,1)$ asymptotically. If both effects are considered at a time, the asymptotic forms of $U^2(x)$, denoted by $Z^2(x)$, can be obtained from equations (11) by using the asymptotic forms of $u_1(x)$, $u_2(x)$, $var(u_1(x))$, $var(u_2(x))$ and $cov(u_1(x), u_2(x))$. As $cov(Z_1(x), Z_2(x)) = 0$, it implies $U^2(x) \approx Z^2(x) = Z_1^2(x) + Z_2^2(x)$ in the populations.

7. GAUSSIAN STOCHASTIC PROCESS

7.1. Covariance between test statistics

The previous discussions mainly focus on deriving the score test statistics at a fixed position x. As presented in Chang *et al.* (4), the score test statistics along the genomes can be described by Gaussian stochastic process asymptotically. To obtain Gaussian processes for the different populations, we now derive the relationship between the test statistics at two different positions by considering the changes in the genomic structure between populations. Now consider $Z(x')$ and $Z(x'')$ at two different positions x' and x'' in two distinct intervals, (A,B) and (C,D), respectively (note that A, B, C and D denote markers with alleles (A,a), (B,b), (C,c) and (D,d), and oblique letters C, D, E, F and G denote the genotypic frequencies). To compute their covariances, we need to first obtain the covariances between their components at the two different positions. To obtain these component covariances, we reformulate $Z_1(x)$ and $Z_2(x)$ (Equations (12) and (13)) in much simpler expressions as

$$Z_1(x) = \sum_{i=1}^{4} s_i \times W_i \quad \textit{and} \quad Z_2(x) = \sum_{i=1}^{5} t_i \times W_i^*,$$

where s_i's and t_i's are the associated coefficients of W_i and W_i^* in $Z_1(x)$ and $Z_2(x)$. For example,

$$s_1 = (\sqrt{C} \times k_1) / \sqrt{C \times k_1^2 + E \times k_2^2 + D \times k_3^2 + E \times k_4^2}$$

and

$$t_1 = (\sqrt{C} \times l_1) / \sqrt{C \times l_1^2 + E \times l_2^2 + D \times l_3^2 + E \times l_4^2 + (F+G) \times l_5^2}.$$

Similarly, the remaining s_i's and t_i's are well defined and can be found in Equations (12) and (13). Then,

the covariance between $Z_1(x')$ and $Z_1(x'')$ can be expressed more succinctly as

$$cov(Z_1(x'),Z_1(x'')) = cov(\sum_{i=1}^{4} s_i' \times W_i', \sum_{i=1}^{4} s_i'' \times W_i''), \quad (14)$$

and the covariance between $Z_2(x')$ and $Z_2(x'')$ is

$$cov(Z_2(x'),Z_2(x'')) = cov(\sum_{i=1}^{5} t_i' \times W_i^{*\prime}, \sum_{i=1}^{5} t_i'' \times W_i^{*\prime\prime}). \quad (15)$$

As t_i's and s_i's are constants in a population, the covariances in Equations (14) and (15) will depend on the covariances between W_i' and W_i'' and between $W_i^{*\prime}$ and $W_i^{*\prime\prime}$ in different positions, i.e. $cov(W_i',W_j'')$'s and $cov(W_i^{*\prime},W_j^{*\prime\prime})$'s. To obtain $cov(W_i',W_j'')$'s and $cov(W_i^{*\prime},W_j^{*\prime\prime})$'s, note that the two different positions x' and x'' can be in the two neighboring or non-neighboring intervals. If they are in the neighboring intervals (flanking markers B and C are the same), evaluating $cov(W_i',W_j'')$'s and $cov(W_i^{*\prime},W_j^{*\prime\prime})$'s needs to consider the genotypic distributions of three loci. If they are in the non-neighboring intervals, the evaluation needs to take the genotypic distributions of four loci into account. For example,

$$cov(W_1',W_1'') = cov(\frac{\bar{y}_{1\cdot}' - \bar{y}_{9\cdot}'}{\sqrt{2\sigma^2/(C' \times n)}}, \frac{\bar{y}_{1\cdot}'' - \bar{y}_{9\cdot}''}{\sqrt{2\sigma^2/(C'' \times n)}})$$

$$= \frac{n \times C' \times C''}{2\sigma^2} \times cov(\bar{y}_{1\cdot}' - \bar{y}_{9\cdot}', \bar{y}_{1\cdot}'' - \bar{y}_{9\cdot}''), \quad (16)$$

where $\bar{y}_{1\cdot}'$'s and $\bar{y}_{9\cdot}'$'s ($\bar{y}_{1\cdot}''$'s and $\bar{y}_{9\cdot}''$s) denote the trait means of marker genotypes, AB/AB and ab/ab (CD/CD and cd/cd), and C (C'') is the frequency of individuals with AB/AB or ab/ab genotype (CD/CD and cd/cd). This covariance needs to evaluate the four covariances, $cov(\bar{y}_{1\cdot}', \bar{y}_{1\cdot}'')$, $cov(\bar{y}_{1\cdot}', \bar{y}_{9\cdot}'')$, $cov(\bar{y}_{9\cdot}', \bar{y}_{1\cdot}'')$ and $cov(\bar{y}_{9\cdot}', \bar{y}_{9\cdot}'')$, between trait means in different intervals.

7.2. Covariance between trait means

For n large, we have

$$cov(\bar{y}_{1\cdot}', \bar{y}_{1\cdot}'') = cov(\sum_{j=1}^{n_1'} \frac{y_{1j}'}{n_1'}, \sum_{j=1}^{n_1''} \frac{y_{1j}''}{n_1''}) = \frac{n \times P(\frac{\text{ABCD}}{\text{ABCD}}) \times \sigma^2}{[n \times P(\frac{\text{AB}}{\text{AB}})] \times [n \times P(\frac{\text{CD}}{\text{CD}})]}$$

$$= \frac{P(\frac{\text{ABCD}}{\text{ABCD}})}{P(\frac{\text{AB}}{\text{AB}}) \times P(\frac{\text{CD}}{\text{CD}})} \times \frac{\sigma^2}{n},$$

where n_1' (n_1'') is the number of individuals with AB/AB or ab/ab (CD/CD or cd/cd) marker genotype. Note that $n \times P(\frac{\text{AB}}{\text{AB}}) = C'$ and $n \times P(\frac{\text{CD}}{\text{CD}}) = C''$ in a population. Therefore, we have

$$cov(\bar{y}_{1\cdot}', \bar{y}_{1\cdot}'') = \frac{P(\frac{\text{ABCD}}{\text{ABCD}})}{C' \times C''} \times \frac{\sigma^2}{n}. \quad (17)$$

Similarly, we can obtain the other components of covariances as

$$cov(\bar{y}_{1\cdot}', \bar{y}_{9\cdot}'') = \frac{P(\frac{\text{ABcd}}{\text{ABcd}})}{P(\frac{\text{AB}}{\text{AB}}) \times P(\frac{\text{cd}}{\text{cd}})} \times \frac{\sigma^2}{n} = \frac{P(\frac{\text{ABcd}}{\text{ABcd}})}{C' \times C''} \times \frac{\sigma^2}{n},$$

$$cov(\bar{y}_{9\cdot}', \bar{y}_{1\cdot}'') = \frac{P(\frac{\text{abCD}}{\text{abCD}})}{P(\frac{\text{ab}}{\text{ab}}) \times P(\frac{\text{CD}}{\text{CD}})} \times \frac{\sigma^2}{n} = \frac{P(\frac{\text{abCD}}{\text{abCD}})}{C' \times C''} \times \frac{\sigma^2}{n},$$

$$cov(\bar{y}_{9\cdot}', \bar{y}_{9\cdot}'') = \frac{P(\frac{\text{abcd}}{\text{abcd}})}{P(\frac{\text{ab}}{\text{ab}}) \times P(\frac{\text{cd}}{\text{cd}})} \times \frac{\sigma^2}{n} = \frac{P(\frac{\text{abcd}}{\text{abcd}})}{C' \times C''} \times \frac{\sigma^2}{n}.$$

Since $P(\frac{\text{AB}}{\text{AB}}) = P(\frac{\text{ab}}{\text{ab}}) = C'$ and $P(\frac{\text{CD}}{\text{CD}}) = P(\frac{\text{cd}}{\text{cd}}) = C''$. The four covariances between trait means in different intervals depend on the genotypic frequencies of two and four loci. Consequently, the covariance between W_1' and W_1'' for non-neighboring intervals in Equation (16) can be obtained, and it is

$$cov(W_1',W_1'') = \frac{1}{2} \times [P(\frac{\text{ABCD}}{\text{ABCD}}) - P(\frac{\text{ABcd}}{\text{ABcd}}) - P(\frac{\text{abCD}}{\text{abCD}}) + P(\frac{\text{abcd}}{\text{abcd}})],$$

which depends on the genotypic frequencies of four loci. Similarly, for neighboring intervals, the covariance between W_1' and W_1'' is a function of the genotypic frequencies of three loci, and it is

$$cov(W_1',W_1'') = \frac{1}{2} \times [P(\frac{\text{ABC}}{\text{ABC}}) + P(\frac{\text{abc}}{\text{abc}})].$$

The other covariances, such as $cov(W_i',W_i'')$'s, $cov(W_j^{*\prime},W_j^{*\prime\prime})$'s, $cov(W_i',W_j^{*\prime})$'s, $cov(W_i'',W_j^{*\prime\prime})$'s, $cov(W_i',W_j^{*\prime\prime})$'s and $cov(W_j^{*\prime},W_i'')$'s, $i = 1,2,3,4$ and

$j=1,2,3,4,5$, can be obtained in a similar way by first deriving their component covariances, i.e. $cov(\bar{y}_{i.}', \bar{y}_{j.}'')$'s, $i, j=1,2,\cdots,9$. In general, obtaining these covariances between different W_i' and W_i'' needs to first evaluate the covariances between different $\bar{y}_{i.}'$ and $\bar{y}_{i.}''$ in the different intervals (see Appendix), which will involve in evaluating 36 genotypic frequencies of three genes for neighboring intervals and 136 genotypic frequencies of for genes for non-neighboring intervals (20, 24). With these covariances, the covariances between the asymptotic score test statistics, $Z_1(x')$, $Z_2(x'')$, $Z_2(x')$ and $Z_1(x'')$, in different intervals can be computed. It is found that the covariances between $Z_1(x)$ and $Z_2(x)$ either in the same or different intervals are zeros, i.e. $cov(Z_1(x'), Z_2(x'))=0$ ($cov(Z_2(x''), Z_1(x''))=0$), $cov(Z_1(x'), Z_2(x''))=0$ and $cov(Z_2(x'), Z_1(x''))=0$, asymptotically. Essentially, by incorporating these covariances into the variance-covariance matrix of Gaussian process, it can consider the differences in genome structure between different populations in computing the threshold values. The above derivations allow us to explore the behaviors of threshold values across populations.

8. SIMULATING THE NULL DISTRIBUTION

The scheme of simulating the null distributions of $\sup Z_1^2(x)$, $\sup Z_2^2(x)$ and $\sup Z^2(x)$ is outlined in this section. If only additive (dominance) effect is considered, we may simulate the components of $Z_1(x)$' s ($Z_2(x)$'s), i.e. W_i ' s (W_i^*' s) in our case, and then to compute $Z_1(x)$'s ($Z_2(x)$'s) and $\sup Z_1^2(x)$ ($\sup Z_2^2(x)$) throughout the genomes as suggested by Chang *et al.* (4). If both effects are considered at a time, we suggest to simulate their subcomponents, $\bar{y}_i$'s, $i=1,2,\cdots,9$, (components of W_i's and W_i^*'s) to obtain the asymptotic forms of $u_1(x)$, $u_2(x)$, $var(u_1(x))$, $var(u_2(x))$ and $cov(u_1(x), u_2(x))$ and then to compute $Z^2(x)$ along the genomes using Equation (11), and finally to obtain $\sup Z^2(x)$. Note that it is also feasible to obtain $\sup Z_1^2(x)$ ($\sup Z_2^2(x)$) by simulating $\bar{y}_i$'s.

When simulating $\bar{y}_i$'s for every intervals, note that there are constraints on $\bar{y}_i$'s between the current l th interval and the next $(l+1)$ th interval due to sharing a common flanking marker. These constraints are

$$C^{(l+1)}\times\bar{y}_1^{(l+1)}+E^{(l+1)}\times\bar{y}_2^{(l+1)}+D^{(l+1)}\times\bar{y}_3^{(l+1)}$$

$$=C^{(l)}\times\bar{y}_1^{(l)}+E^{(l)}\times\bar{y}_4^{(l)}+D^{(l)}\times\bar{y}_7^{(l)} \quad (18)$$

$$E^{(l+1)}\times\bar{y}_4^{(l+1)}+(F^{(l+1)}+G^{(l+1)})\times\bar{y}_5^{(l+1)}+E^{(l+1)}\times\bar{y}_6^{(l+1)}$$

$$=E^{(l)}\times\bar{y}_2^{(l)}+(F^{(l)}+G^{(l)})\times\bar{y}_5^{(l)}+E^{(l)}\times\bar{y}_8^{(l)} \quad (19)$$

$$D^{(l+1)}\times\bar{y}_7^{(l+1)}+E^{(l+1)}\times\bar{y}_8^{(l+1)}+C^{(l+1)}\times\bar{y}_9^{(l+1)}$$

$$=D^{(l)}\times\bar{y}_3^{(l)}+E^{(l)}\times\bar{y}_6^{(l)}+C^{(l)}\times\bar{y}_9^{(l)}. \quad (20)$$

To simulate the null distribution of $\sup Z^2(x)$ for a genome with k intervals, we suggest the following steps:

1.Generate ($\underline{\bar{y}}^{(1)}$, $\underline{\bar{y}}^{(2)}$, $\cdots$, $\underline{\bar{y}}^{(k)}$) from $N(\underline{0}, \sum)$, where $\underline{\bar{y}}^{(1)}$ is a vector containg the nine trait means in the first interval, and $\underline{\bar{y}}^{(l)}$, $l=2,3,\cdots,k$, is a vector containing six of the nine trait means, e.g., $\bar{y}_1$, $\bar{y}_2$, $\bar{y}_4$, $\bar{y}_5$, $\bar{y}_7$ and $\bar{y}_8$, in the l th intervals, and $\sum$ is the variance-covariance matrix of the trait means. The construction of $\sum$ for the normal distribution needs to evaluate the covariances between different $\bar{y}_i$'s by using the genotypic distributions of three and four genes (see APPENDIX).

2.Compute the remaining three trait means, e.g., $\bar{y}_3$, $\bar{y}_6$ and $\bar{y}_9$, for the 2nd to k th intervals using the three constraints (equations (18), (19) and (20)). The dimension of $\sum$ is usually large and is $(6k+3)\times(6k+3)$.

3. Compute the score test statistics, $Z^2(x)$'s, at all positions in the k intervals along the genome, and find and record their maximum.

The above steps are repeated many times, say 10,000 times, to obtain the approximate distribution of $\sup Z^2(x)$. The threshold value at significant level α can be determined accordingly. The R program of our approach is available on http://www.stat.sinica.edu.tw/~chkao/.

9. REAL EXAMPLE AND SIMULATION STUDIES

9.1. Real example

As a real data application, we considered a backcross model to compute the threshold value of QTL mapping in a pseudo-testcross population of Radiata pine. One hundred and twenty markers contributed genotypic information across twelve linkage groups and covered ~1679.3 cM. The traits considered are tree diameter, branch quality scores and brown cone number. The average spacing of the 107 marker intervals was 13.5 cM. The maximum distance between two adjacent markers is 74.8 cM, and the minimum is 1.5 cM. For this practice of QTL detection, Kao *et al.* (7) used Bonferroni argument to choose a value of 12.12 ($\chi^2_{1,0.05/107}$) as a threshold in the

Table 2. Comparison of the proposed and empirical thresholds at α=0.05 for different marker densities in F2, AI Ft and RI Ft populations

		RI	RI	AI	AI
population	□[1]	Proposed[2]	Empirical[3]	Proposed[2]	Empirical[3]
F_2	100	8.103	8.128	8.103	8.128
	50	8.966	8.978	8.966	8.978
	20	9.853	9.998	9.853	9.998
	10	10.574	10.481	10.574	10.481
	5	11.145	11.120	11.145	11.120
	2	11.732	11.750	11.732	11.750
	1	11.959	12.318	11.959	12.318
F_3	100	8.076	9.133	8.183	9.328
	50	9.080	9.206	9.100	9.463
	20	10.099	9.495	10.130	10.266
	10	10.859	10.253	10.887	11.206
	5	11.487	11.096	11.638	11.655
	2	12.212	11.535	12.325	12.384
	1	12.447	12.759	12.580	12.828
F_4	100	8.139	9.072	8.195	9.678
	50	9.059	9.199	9.183	10.113
	20	10.107	10.324	10.343	10.677
	10	11.186	11.099	11.132	11.190
	5	11.860	11.768	11.886	12.043
	2	12.493	12.375	12.729	12.567
	1	12.799	12.729	12.952	12.896
F_6	100	8.273	7.455	8.200	9.972
	50	9.130	9.656	9.292	10.288
	20	10.321	9.732	10.617	11.285
	10	11.315	11.000	11.533	12.299
	5	12.160	11.280	12.293	12.533
	2	12.845	12.250	13.260	13.526
	1	13.257	12.814	13.518	13.973
F_{10}	100	5.562	5.737	8.194	9.879
	50	6.296	6.411	9.410	10.401
	20	6.963	7.307	10.832	11.279
	10	7.928	7.873	11.871	11.831
	5	8.560	8.643	12.710	12.189
	2	9.102	9.210	13.804	12.921
	1	9.327	9.525	14.261	13.504

[1]marker distance (in cM) on a 100-cM chromosome. [2] based on 10,000 simulations. [3] based on 10,000 data sets with 200 individuals from the null distribution. The population considers additive effect only, and the other populations consider both additive and dominance effects.

analysis. By using our approach, a larger threshold value 12.43 is obtained. By using 12.43 as a threshold of QTL mapping, the numbers of detected QTL for the three traits are the same as those by using 12.12 as a threshold in this case.

9.2. Simulation studies

Simulations were performed to evaluate the performance of the proposed method and study the behaviors of the threshold values under various marker densities in several experimental populations. For each population, one chromosome with a total length of 100 cM was simulated. On the chromosome, we assume that there are 101, 51, 21, 11, 6, 3 and 2 evenly spaced markers, i.e. the marker distances are 1, 2, 5, 10, 20, 50 and 100 cM, respectively. The experimental populations considered include F_2, AI (RI) F_3, AI (RI) F_4, AI (RI) F_5, AI (RI) F_6 and AI (RI) F_{10} populations. Except for the RI F_{10} population, both additive and dominance effects are considered. For the RI F_{10} population, only the additive effect is considered, as there are very few heterozygotes due to continuous self. For each case considered, score test statistics are computed every 1 cM, and the number of simulated replicate is 10,000. The 10,000 maximums of the score test statistics along the chromosome are recorded. To validate our method, we also simulate 10,000 sets of the traits and markers from 200 individuals for each marker density and experimental population under the null hypothesis. Each data set is then analyzed by the interval mapping approach, and the LRT statistics were computed every 1 cM and the 10,000 maximums were obtained for comparison. Results of the threshold values at □=0.05 from the maximums of LRT statistics and score test statistics are given in Table 2. It is worth pointing out that the score-statistic approach is several hundred times faster than the LRT-statistic approach to complete the results in Table 2. This advantage of greatly saving computation time has been identified by Chang *et al.* (4).

Table 2 shows that the threshold values obtained from the two different approach are generally very close to each other, especially, in the F_2 population. For example, the values by the score-statistic approach are 8.103, 8.966, 9.853, 10.574, 11.145, 11.732 and 11.959, respectively, for the sever marker densities in the F_2 population, and those from the LRT statistic are 8.128, 8.978, 9.998, 10.481, 11.120, 11.750 and 12.318, respectively. For more advanced AI or RI populations, the differences between the threshold values seem to become relatively larger, especially for the sparse marker density in the more advanced AI populations. For example, in the AI F_{10} population, the threshold values from the score test statistic are 9.410 and 8.194 in the 50- and 100 cM-marker spacing, and those from the LRT statistic are 10.401 and 9.879. Also, the threshold values are increasing for denser marker maps, which is consistent with the findings in previous studies (8, 11). Such an increasing trend becomes more obvious in the more advanced populations as compared with that in the earlier populations. For example, the values for 100- and 1-cM marker spacing are 8.200 and 13.518 (8.273 and 13.257), respectively, in the AI (RI) F_6 population, and they are 8.183 and 12.580 (8.076 and 12.447) in the AI (RI) F_3 population. Besides, the threshold values are higher in the more advanced populations. For example, the values for 10-cM marker spacing are 10.574, 10.887 (10.859), 11.132 (11.186), 11.533 (11.315), 11.871 in the F_2, AI (RI) F_3, AI (RI) F_4, AI (RI) F_5, AI (RI) F_6 and AI F_{10} populations. The threshold values in the AI populations are generally larger as compared to those in the RI populations (except for the case of 100-cM spacing in the F_6 population). For example, the threshold values are 8.075, 9.080, 10.090, 10.859, 11.487, 12.212 and 12.447 for the seven marker densities in the RI F_3 population, and they are 8.183, 9.100, 10.130, 10.887, 11.638, 12.325 and 12.580 in the AI F_3 population. The similar trends can be also observed in the other AI and RI F_t populations.

10. DISCUSSION

When applying the interval mapping procedure to search the whole genomes for QTL, typically, the LRT statistics are constructed over the all possible positions to test for the null hypothesis of no QTL, and the position with the significant maximum of LRT statistic is regarded

as the estimated QTL position. Under such a procedure, the determination of the threshold values for the test statistics to declare significance has been a central issue in QTL mapping. As the maximums of the score test and LRT statistics are asymptotically equivalent (4, 8, 26) and the computation cost of the score test statistics is much cheaper (4, 8), we propose a general score-statistic approach to computing the threshold values of QTL detection for various experimental populations. These experimental populations include F_2, AI and RI populations. In our approach, the score test statistics are formulated in terms of the trait means of marker classes, mixing proportions, and the genotypic distributions of two and three genes of the populations. The asymptotic distribution of the score test statistics along the genomes is characterized by a Gaussian process with mean zero and well-structured variance-covariance matrix. We devise the genotypic distributions of three and four genes into the variance-covariance matrix to take care of the changes in the genomic structure between different populations, so that the threshold values for the different populations can be computed and their behaviors can be explored in various marker densities and genome sizes. The validity of our approach is compared with the LRT statistics by Monte Carlo simulations. In general, the threshold values obtained by our approach are very close to those by the LRT statistics in the various advanced populations as shown in Table 2. Given a significance level and a genome size, the threshold values should be enhanced in denser marker maps and in more advanced populations.

The different advanced populations are subject to different numbers of meiosis cycle either by inbreeding and/or random mating. They will produce different genomic structures, and their genotypic distributions will be different from each other. By recognizing such differences between populations, we incorporate the distributions of two, three and four genes of these populations into the score test statistics and Gaussian processes (equations~(12), (13), (14) and (15)). Therefore, our approach can consider their specific genome structures to well compute their threshold values and investigate their behaviors. More importantly, it has to be pointed out that the genotypic frequencies of three and four genes can be directly obtained by the genotypic frequencies of pairwise genes in the $F2$ populations, as the $F2$ genomes have a Markovian structure under Haldane map function (27). However, for advanced populations, this Markovian property disappears, and obtaining the genotypic frequencies of three and four genes is not straightforward. Here, we use the sets of transition equations proposed by Kao and Zeng (20, 24) to obtain these frequencies. If these frequencies are approximated by using Jiang and Zeng's method (28), which implicitly assumes a Markovian property, and the approximate frequencies are used in the construction of the test statistics and Gaussian processes in the computation of the threshold values. We found that the threshold values obtained by using the approximate genotypic frequencies are generally larger, especially in the denser marker maps and in the more advanced populations (results not shown), as compared to those by using exact frequencies.

The proposed score-statistic approach for computing the threshold value is mainly devised for the interval mapping model, i.e. a single QTL model. Under a single QTL model, the score test statistics for the additive and dominance effects in equations (12) and (13) are not mixtures and are functions of the nine trait means. The nine trait means correspond to the nine flanking marker genotypes whose proportions are adjusted to different population structures. In the setup of the variance-covariance matrices of Gaussian processes, the elements are obtained from the covariances between the test statistics (trait means). Therefore, it needs to consider the distributions of three (for neighboring intervals) and four genes (for non-neighboring intervals) at a time, and the dimension of the variance-covariance matrix is usually large. For example, the dimensions of the matrices are 63×63 (9+(10-1)×6 = 63) and 123×123 (9+(20-1)x6 = 123) for a chromosome with 10 and 20 intervals, respectively. If composite interval mapping (CIM; 6) or multiple interval mapping (MIM; 7) is considered, it will need to evaluate many more genotypic frequencies of more genes to construct a much higher-dimensional variance-covariance matrix in the processes. Taking a two-QTL MIM model, $y_i = \mu + a_1 x_1^* + a_2 x_2^* + \varepsilon_i$, as an example, if searching a chromosome segment with 10 marker intervals for the first QTL (testing H_0: $a_1=0$ and $a_2 \neq 0$) by conditioning on the second QTL in the other region, the score test statistic for a_1 can be classified into 81 different categories according to the 81 marker genotypes of the two flanking intervals ($9^2=81$), and each category contains mixture components. As the score statistic is a mixture, the derivations of its variance and asymptotic form are not simple. Furthermore, the dimension of the variance-covariance matrix in the Gaussian process will increase to 567x567 (81+(10-1)x54=567), and obtaining the covariance elements needs to evaluate the 2080 genotypic frequencies of six genes for this MIM model. If the second QTL is coincident with a marker, the two-QTL MIM model reduces to a CIM model. The corresponding score test statistic can be classified into 27 categories corresponding to 27 marker genotypes. The variance-covariance matrix of the Gaussian process has a 189x189 (27+(10-1)x18=189) dimension, and obtaining the covariance elements needs to evaluate the 528 genotypic frequencies of five genes for this CIM model. Certainly, if more QTL are considered at a time in the model, the obtaining of the score test statistics and Gaussian processes will be even more complicated as they will involve using the distributions of a large number of genes at a time in the populations. Consequently, the issues in determining the threshold values for the multiple-QTL models remain challenging. The approaches to unraveling these issues will not be in a straightforward manner and are worth pursuing in the future.

11. ACKNOWLEDGEMENS

This work was supported by grants NSC99-2118-M-001-011 from the National Science Council, Taiwan, Republic of China.

12. APPENDIX

Evaluation of the covariances between $\bar{y}_i^I$'s and $\bar{y}_j^{II}$'s, $i = 1,2,\ldots,9$, at the two different positions, x^I and x^{II}, in two distinct marker intervals, (A, B) and (C, D), respectively, are presented below. In general, the evaluation involves using the genotypic distributions of two, three and four genes in the populations. There are 81 covariances in total. If the two intervals are nonneighboring, the covariances are functions of the genotypic frequencies of two and four genes. They are listed below.

$$\mathrm{cov}(\bar{y}_1^I, \bar{y}_1^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{ABCD}}{\mathrm{ABCD}}\right)}{C^I C^{II}},\ \mathrm{cov}(\bar{y}_1^I, \bar{y}_2^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{ABCD}}{\mathrm{ABCd}}\right)}{C^I E^{II}},\ \mathrm{cov}(\bar{y}_1^I, \bar{y}_3^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{ABCd}}{\mathrm{ABCd}}\right)}{C^I D^{II}},$$

$$\mathrm{cov}(\bar{y}_1^I, \bar{y}_4^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{ABCD}}{\mathrm{ABcD}}\right)}{C^I E^{II}},\ \mathrm{cov}(\bar{y}_1^I, \bar{y}_5^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{ABCD}}{\mathrm{ABcd}}\right) + P\left(\frac{\mathrm{ABCd}}{\mathrm{ABcD}}\right)}{C^I(F^{II} + G^{II})},$$

$$\mathrm{cov}(\bar{y}_1^I, \bar{y}_6^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{ABCd}}{\mathrm{ABcd}}\right)}{C^I E^{II}},\ \mathrm{cov}(\bar{y}_1^I, \bar{y}_7^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{ABcD}}{\mathrm{ABcD}}\right)}{C^I D^{II}},\ \mathrm{cov}(\bar{y}_1^I, \bar{y}_8^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{ABcD}}{\mathrm{ABcd}}\right)}{C^I E^{II}},$$

$$\mathrm{cov}(\bar{y}_1^I, \bar{y}_9^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{ABcd}}{\mathrm{ABcd}}\right)}{C^I C^{II}},$$

where (A, a), (B, b), (C, c) and (D, d) denote the alleles of markers A, B, C and D, respectively, define C^I (C^{II}) is the frequency of AB/AB (CD/CD) or ab/ab (cd/cd) genotype, $D^I(D^{II})$ is the frequency of Ab/Ab (Cd/Cd) or aB/aB (cD/cD) genotype, $E^I(E^{II})$ is the frequency of AB/Ab (CD/Cd), AB/aB (CD/cD), Ab/ab (Cd/cd) or aB/ab (cD/cd) genotype, $F^I(F^{II})$ is the frequency of AB/ab (CD/cd) genotype, and $G^I(G^{II})$ is the frequency of Ab/aB (Cd/cD) genotype, respectively.

$$\mathrm{cov}(\bar{y}_2^I, \bar{y}_1^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{ABCD}}{\mathrm{AbCD}}\right)}{E^I C^{II}},\ \mathrm{cov}(\bar{y}_2^I, \bar{y}_2^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{ABCd}}{\mathrm{AbCd}}\right)}{E^I D^{II}},\ \mathrm{cov}(\bar{y}_2^I, \bar{y}_3^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{ABcD}}{\mathrm{AbcD}}\right)}{E^I D^I},$$

$$\mathrm{cov}(\bar{y}_2^I, \bar{y}_4^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{ABCD}}{\mathrm{AbCd}}\right) + P\left(\frac{\mathrm{ABCd}}{\mathrm{AbCD}}\right)}{E^I E^{II}},\ \mathrm{cov}(\bar{y}_2^I, \bar{y}_5^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{ABCD}}{\mathrm{Abcd}}\right) + P\left(\frac{\mathrm{ABcd}}{\mathrm{AbCd}}\right) + P\left(\frac{\mathrm{ABCd}}{\mathrm{AbcD}}\right) + P\left(\frac{\mathrm{ABcD}}{\mathrm{AbCd}}\right)}{E^I(F^{II} + G^{II})},$$

$$\mathrm{cov}(\bar{y}_2^I, \bar{y}_6^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{ABCd}}{\mathrm{Abcd}}\right) + P\left(\frac{\mathrm{ABcd}}{\mathrm{AbCd}}\right)}{E^I E^{II}},\ \mathrm{cov}(\bar{y}_2^I, \bar{y}_7^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{ABcD}}{\mathrm{Abcd}}\right) + P\left(\frac{\mathrm{ABcd}}{\mathrm{AbcD}}\right)}{E^I E^{II}},$$

$$\mathrm{cov}(\bar{y}_2^I, \bar{y}_8^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{AbCD}}{\mathrm{AbcD}}\right)}{D^I C^{II}},\ \mathrm{cov}(\bar{y}_2^I, \bar{y}_9^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{AbCd}}{\mathrm{AbcD}}\right)}{D^I E^{II}},\ \mathrm{cov}(\bar{y}_2^I, \bar{y}_9^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{AbcD}}{\mathrm{AbcD}}\right)}{D^I D^{II}},$$

$$\mathrm{cov}(\bar{y}_3^I, \bar{y}_1^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{ABCD}}{\mathrm{AbcD}}\right)}{D^I E^{II}},\ \mathrm{cov}(\bar{y}_3^I, \bar{y}_2^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{ABCD}}{\mathrm{AbCd}}\right) + P\left(\frac{\mathrm{AbCD}}{\mathrm{AbcD}}\right)}{D^I(F^{II} + G^{II})},$$

$$\mathrm{cov}(\bar{y}_3^I, \bar{y}_3^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{ABcD}}{\mathrm{AbcD}}\right)}{D^I D^{II}},\ \mathrm{cov}(\bar{y}_3^I, \bar{y}_4^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{ABCD}}{\mathrm{AbcD}}\right)}{D^I E^{II}},\ \mathrm{cov}(\bar{y}_3^I, \bar{y}_5^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{ABcD}}{\mathrm{Abcd}}\right) + P\left(\frac{\mathrm{AbcD}}{\mathrm{aBcD}}\right)}{D^I(F^{II} + G^{II})},$$

$$\mathrm{cov}(\bar{y}_3^I, \bar{y}_6^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{ABcd}}{\mathrm{Abcd}}\right) + P\left(\frac{\mathrm{AbcD}}{\mathrm{aBcD}}\right)}{D^I E^{II}},\ \mathrm{cov}(\bar{y}_3^I, \bar{y}_7^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{AbcD}}{\mathrm{AbcD}}\right) + P\left(\frac{\mathrm{AbcD}}{\mathrm{aBcD}}\right)}{D^I D^{II}},$$

$$\mathrm{cov}(\bar{y}_3^I, \bar{y}_8^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{AbCD}}{\mathrm{AbcD}}\right)}{D^I C^{II}},\ \mathrm{cov}(\bar{y}_3^I, \bar{y}_9^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{AbCd}}{\mathrm{AbcD}}\right)}{D^I E^{II}},\ \mathrm{cov}(\bar{y}_3^I, \bar{y}_9^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{AbcD}}{\mathrm{AbcD}}\right)}{D^I D^{II}},$$

$$\mathrm{cov}(\bar{y}_4^I, \bar{y}_1^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{AbCD}}{\mathrm{ABCD}}\right)}{F^I C^{II}},\ \mathrm{cov}(\bar{y}_4^I, \bar{y}_2^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{AbCD}}{\mathrm{ABCd}}\right) + P\left(\frac{\mathrm{AbCd}}{\mathrm{ABCD}}\right)}{F^I E^{II}},\ \mathrm{cov}(\bar{y}_4^I, \bar{y}_3^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{AbCD}}{\mathrm{ABcD}}\right) + P\left(\frac{\mathrm{AbcD}}{\mathrm{ABCD}}\right)}{F^I E^{II}},$$

$$\mathrm{cov}(\bar{y}_4^I, \bar{y}_4^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{AbCD}}{\mathrm{ABCd}}\right) + P\left(\frac{\mathrm{AbCd}}{\mathrm{ABCD}}\right)}{F^I E^{II}},\ \mathrm{cov}(\bar{y}_4^I, \bar{y}_5^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{AbCD}}{\mathrm{Abcd}}\right) + P\left(\frac{\mathrm{Abcd}}{\mathrm{AbCD}}\right) + P\left(\frac{\mathrm{AbCd}}{\mathrm{AbcD}}\right) + P\left(\frac{\mathrm{AbcD}}{\mathrm{AbCd}}\right)}{F^I(F^{II} + G^{II})},$$

$$\mathrm{cov}(\bar{y}_4^I, \bar{y}_6^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{AbCd}}{\mathrm{Abcd}}\right) + P\left(\frac{\mathrm{Abcd}}{\mathrm{AbCd}}\right)}{F^I E^{II}},\ \mathrm{cov}(\bar{y}_4^I, \bar{y}_7^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{AbcD}}{\mathrm{AbcD}}\right) + P\left(\frac{\mathrm{AbcD}}{\mathrm{AbCd}}\right)}{F^I E^{II}},$$

$$\mathrm{cov}(\bar{y}_4^I, \bar{y}_8^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{AbCD}}{\mathrm{AbcD}}\right) + P\left(\frac{\mathrm{AbcD}}{\mathrm{AbCD}}\right)}{F^I E^{II}},\ \mathrm{cov}(\bar{y}_4^I, \bar{y}_9^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{AbCd}}{\mathrm{AbcD}}\right) + P\left(\frac{\mathrm{AbcD}}{\mathrm{AbCd}}\right)}{F^I E^{II}},$$

$$\mathrm{cov}(\bar{y}_5^I, \bar{y}_1^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{ABCD}}{\mathrm{aBCD}}\right) + P\left(\frac{\mathrm{AbCD}}{\mathrm{aBCD}}\right)}{(F^I + G^I)C^{II}},\ \mathrm{cov}(\bar{y}_5^I, \bar{y}_2^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{ABCd}}{\mathrm{aBCd}}\right) + P\left(\frac{\mathrm{AbCd}}{\mathrm{aBCd}}\right)}{(F^I + G^I)D^{II}},$$

$$\mathrm{cov}(\bar{y}_5^I, \bar{y}_3^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{ABCD}}{\mathrm{aBcD}}\right) + P\left(\frac{\mathrm{ABcD}}{\mathrm{aBCD}}\right) + P\left(\frac{\mathrm{AbCD}}{\mathrm{aBcD}}\right) + P\left(\frac{\mathrm{AbcD}}{\mathrm{aBCD}}\right)}{(F^I + G^I)E^{II}},$$

$$\mathrm{cov}(\bar{y}_5^I, \bar{y}_4^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{ABCD}}{\mathrm{aBCd}}\right) + P\left(\frac{\mathrm{ABCd}}{\mathrm{aBCD}}\right) + P\left(\frac{\mathrm{AbCD}}{\mathrm{aBCd}}\right) + P\left(\frac{\mathrm{AbCd}}{\mathrm{aBCD}}\right)}{(F^I + G^I)E^{II}},$$

$$\mathrm{cov}(\bar{y}_5^I, \bar{y}_5^{II}) = \frac{\sigma^2}{n}\frac{1}{(F^I + G^I)(F^{II} + G^{II})}\left[P\left(\frac{\mathrm{ABCD}}{\mathrm{abcd}}\right) + P\left(\frac{\mathrm{ABcd}}{\mathrm{abCD}}\right) + P\left(\frac{\mathrm{ABCd}}{\mathrm{abcD}}\right) + P\left(\frac{\mathrm{AbCD}}{\mathrm{aBcd}}\right) + P\left(\frac{\mathrm{Abcd}}{\mathrm{aBCD}}\right) + P\left(\frac{\mathrm{AbCd}}{\mathrm{aBcD}}\right) + P\left(\frac{\mathrm{AbcD}}{\mathrm{aBCd}}\right)\right],$$

$$\mathrm{cov}(\bar{y}_5^I, \bar{y}_6^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{ABCd}}{\mathrm{abcd}}\right) + P\left(\frac{\mathrm{ABcd}}{\mathrm{abCd}}\right) + P\left(\frac{\mathrm{AbCd}}{\mathrm{aBCd}}\right) + P\left(\frac{\mathrm{Abcd}}{\mathrm{aBCd}}\right)}{(F^I + G^I)E^{II}},$$

$$\mathrm{cov}(\bar{y}_5^I, \bar{y}_7^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{ABcD}}{\mathrm{abcD}}\right) + P\left(\frac{\mathrm{AbcD}}{\mathrm{aBcD}}\right)}{(F^I + G^I)D^{II}},\ \mathrm{cov}(\bar{y}_5^I, \bar{y}_8^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{ABcd}}{\mathrm{abcd}}\right) + P\left(\frac{\mathrm{Abcd}}{\mathrm{aBcd}}\right)}{(F^I + G^I)C^{II}},$$

$$\mathrm{cov}(\bar{y}_5^I, \bar{y}_9^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{ABcD}}{\mathrm{abcd}}\right) + P\left(\frac{\mathrm{AbcD}}{\mathrm{aBcd}}\right) + P\left(\frac{\mathrm{AbcD}}{\mathrm{aBcD}}\right) + P\left(\frac{\mathrm{Abcd}}{\mathrm{aBCd}}\right)}{(F^I + G^I)E^{II}},$$

$$\mathrm{cov}(\bar{y}_6^I, \bar{y}_1^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{AbCD}}{\mathrm{abCD}}\right)}{E^I C^{II}},\ \mathrm{cov}(\bar{y}_6^I, \bar{y}_2^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{AbCd}}{\mathrm{abCd}}\right)}{E^I D^{II}},\ \mathrm{cov}(\bar{y}_6^I, \bar{y}_3^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{AbcD}}{\mathrm{abcD}}\right)}{E^I D^I},$$

$$\mathrm{cov}(\bar{y}_6^I, \bar{y}_4^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{AbCD}}{\mathrm{abCd}}\right) + P\left(\frac{\mathrm{AbCd}}{\mathrm{abCD}}\right)}{E^I E^{II}},\ \mathrm{cov}(\bar{y}_6^I, \bar{y}_5^{II}) = \frac{\sigma^2}{n}\frac{P\left(\frac{\mathrm{AbCD}}{\mathrm{abcd}}\right) + P\left(\frac{\mathrm{AbcD}}{\mathrm{abCD}}\right) + P\left(\frac{\mathrm{AbCd}}{\mathrm{abcD}}\right) + P\left(\frac{\mathrm{AbcD}}{\mathrm{abCd}}\right)}{E^I(F^{II} + G^{II})},$$

$$\mathrm{cov}(\bar{y}_6', \bar{y}_6'') = \frac{\sigma^2}{n} \frac{P\left(\frac{AbCd}{abcd}\right) + P\left(\frac{Abcd}{abCd}\right)}{E'E''}, \quad \mathrm{cov}(\bar{y}_6', \bar{y}_7'') = \frac{\sigma^2}{n} \frac{P\left(\frac{AbcD}{abcd}\right) + P\left(\frac{Abcd}{abcD}\right)}{E'E''},$$

$$\mathrm{cov}(\bar{y}_7', \bar{y}_1'') = \frac{\sigma^2}{n} \frac{P\left(\frac{aBCD}{aBCD}\right)}{D'C''}, \quad \mathrm{cov}(\bar{y}_7', \bar{y}_2'') = \frac{\sigma^2}{n} \frac{P\left(\frac{aBCD}{aBCd}\right)}{D'E''}, \quad \mathrm{cov}(\bar{y}_7', \bar{y}_3'') = \frac{\sigma^2}{n} \frac{P\left(\frac{aBCd}{aBCd}\right)}{D'D''},$$

$$\mathrm{cov}(\bar{y}_7', \bar{y}_4'') = \frac{\sigma^2}{n} \frac{P\left(\frac{aBCD}{aBcD}\right)}{D'E''}, \quad \mathrm{cov}(\bar{y}_7', \bar{y}_5'') = \frac{\sigma^2}{n} \frac{P\left(\frac{aBCD}{aBcd}\right) + P\left(\frac{aBCd}{aBcD}\right)}{D'(F'' + G'')},$$

$$\mathrm{cov}(\bar{y}_7', \bar{y}_7'') = \frac{\sigma^2}{n} \frac{P\left(\frac{aBcD}{aBcD}\right)}{D'D''}, \quad \mathrm{cov}(\bar{y}_7', \bar{y}_8'') = \frac{\sigma^2}{n} \frac{P\left(\frac{aBcD}{aBcd}\right)}{D'E''}, \quad \mathrm{cov}(\bar{y}_7', \bar{y}_9'') = \frac{\sigma^2}{n} \frac{P\left(\frac{aBcd}{aBcd}\right)}{D'C''},$$

$$\mathrm{cov}(\bar{y}_8', \bar{y}_1'') = \frac{\sigma^2}{n} \frac{P\left(\frac{aBCD}{abCD}\right)}{E'C''}, \quad \mathrm{cov}(\bar{y}_8', \bar{y}_2'') = \frac{\sigma^2}{n} \frac{P\left(\frac{aBCD}{abCd}\right)}{E'D''}, \quad \mathrm{cov}(\bar{y}_8', \bar{y}_7'') = \frac{\sigma^2}{n} \frac{P\left(\frac{aBcD}{abcD}\right)}{E'D''},$$

$$\mathrm{cov}(\bar{y}_8', \bar{y}_2'') = \frac{\sigma^2}{n} \frac{P\left(\frac{aBCD}{abCd}\right) + P\left(\frac{aBCd}{abCD}\right)}{E'E''}, \quad \mathrm{cov}(\bar{y}_8', \bar{y}_4'') = \frac{\sigma^2}{n} \frac{P\left(\frac{aBCD}{abcD}\right) + P\left(\frac{aBcD}{abCD}\right)}{E'E''},$$

$$\mathrm{cov}(\bar{y}_8', \bar{y}_5'') = \frac{\sigma^2}{n} \frac{P\left(\frac{aBCD}{abcd}\right) + P\left(\frac{aBcd}{abCD}\right) + P\left(\frac{aBCd}{abcD}\right) + P\left(\frac{aBcD}{abCd}\right)}{E'(F'' + G'')},$$

$$\mathrm{cov}(\bar{y}_8', \bar{y}_8'') = \frac{\sigma^2}{n} \frac{P\left(\frac{aBCd}{abcd}\right) + P\left(\frac{aBcd}{abCd}\right)}{E'E''}, \quad \mathrm{cov}(\bar{y}_8', \bar{y}_9'') = \frac{\sigma^2}{n} \frac{P\left(\frac{aBcD}{abcd}\right) + P\left(\frac{aBcd}{abcD}\right)}{E'E''},$$

$$\mathrm{cov}(\bar{y}_9', \bar{y}_1'') = \frac{\sigma^2}{n} \frac{P\left(\frac{abCD}{abCD}\right)}{C'C''}, \quad \mathrm{cov}(\bar{y}_9', \bar{y}_2'') = \frac{\sigma^2}{n} \frac{P\left(\frac{abCD}{abCd}\right)}{C'E''}, \quad \mathrm{cov}(\bar{y}_9', \bar{y}_4'') = \frac{\sigma^2}{n} \frac{P\left(\frac{abCd}{abcD}\right)}{C'D''},$$

$$\mathrm{cov}(\bar{y}_9', \bar{y}_4'') = \frac{\sigma^2}{n} \frac{P\left(\frac{abCD}{abcD}\right)}{C'E''}, \quad \mathrm{cov}(\bar{y}_9', \bar{y}_5'') = \frac{\sigma^2}{n} \frac{P\left(\frac{abCD}{abcd}\right) + P\left(\frac{abCd}{abcD}\right)}{C'(F'' + G'')},$$

$$\mathrm{cov}(\bar{y}_9', \bar{y}_7'') = \frac{\sigma^2}{n} \frac{P\left(\frac{abcD}{abcD}\right)}{C'D''}, \quad \mathrm{cov}(\bar{y}_9', \bar{y}_8'') = \frac{\sigma^2}{n} \frac{P\left(\frac{abcD}{abcd}\right)}{C'E''}, \quad \mathrm{cov}(\bar{y}_9', \bar{y}_9'') = \frac{\sigma^2}{n} \frac{P\left(\frac{abcd}{abcd}\right)}{C'C''},$$

$$\mathrm{cov}(\bar{y}_1', \bar{y}_8'') = \frac{\sigma^2}{n} \frac{P\left(\frac{ABcd}{Abcd}\right)}{E'C''}, \quad \mathrm{cov}(\bar{y}_3', \bar{y}_8'') = \frac{\sigma^2}{n} \frac{P\left(\frac{AbCd}{Abcd}\right)}{D'E''}, \quad \mathrm{cov}(\bar{y}_4', \bar{y}_8'') = \frac{\sigma^2}{n} \frac{P\left(\frac{ABcd}{aBcd}\right)}{E'C''},$$

$$\mathrm{cov}(\bar{y}_6', \bar{y}_9'') = \frac{\sigma^2}{n} \frac{P\left(\frac{Abcd}{abcd}\right)}{E'C''}, \quad \mathrm{cov}(\bar{y}_7', \bar{y}_6'') = \frac{\sigma^2}{n} \frac{P\left(\frac{aBCd}{aBcd}\right)}{D'E''}, \quad \mathrm{cov}(\bar{y}_8', \bar{y}_9'') = \frac{\sigma^2}{n} \frac{P\left(\frac{aBcd}{abcd}\right)}{E'C''},$$

$$\mathrm{cov}(\bar{y}_9', \bar{y}_6'') = \frac{\sigma^2}{n} \frac{P\left(\frac{abCd}{abcd}\right)}{C'E''}.$$

If the two intervals are neighboring, the covariances are functions of the genotypic frequencies of two and three genes, and they can be obtained in a similar way. The frequencies of the 10, 36 and 136 genotypes for two, three and four genes in the different advanced populations can be obtained by using the transition equations of Haldane and Waddington (1931) and Kao and Zeng (2009, 2010).

13. REFERENCES

1. Lander E, D Botstein: Mapping Mendelian factors underlying quantitative traits using RELP linkage maps. Genetics 121, 185-199 (1989)

2. Dempster A, N Laird, D Rubin: Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. Ser. B* 39, 1-38 (1977)

3. Lander E, N Schork. Genetic dissection of complex traits: Science 265, 2037-2048 (1994)

4. Chang M, RL Wu, S Wu, G Casella: Score statistics for mapping quantitative trait loci. Statistical Applications in Genetics and Molecular Biology 8 Article 16 (2009)

5. Jensen R: Interval mapping of multiple quantitative trait loci. Genetics 135, 205-211 (1113)

6. Zeng ZB: Precision mapping of quantitative trait loci. Genetics 136, 1457-1468 (1994)

7. Kao CH, ZB Zeng, R Teasdale: Multiple interval mapping for Quantitative Trait Loci. Genetics 152, 1203-1216 (1999)

8. Zou F, J Fine, J Hu, D Lin: An Efficient Resampling Method for Assessing Genome-Wide Statistical Significance in Mapping Quantitative Trait Loci Genetics 168, 2307-2316 (2004)

9. Van Ooijen J: Accuracy of mapping quantitative trait loci in autogamous species. Theor Appl Genet 84, 803-811 (1992)

10. Churchill G, R Doerge: Empirical threshold values for quantitative trait mapping. Genetics 138, 963-971 (1994)

11. Rebai A, B Goffinet, B Mangin: Approximate thresholds of interval mapping tests for QTL detection. Genetics 138, 235-240 (1994)

12. Davies R. Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika 64, 247-254 (1977)

13. Dupuis J, D Siegmund: Statistical methods for mapping quantitative trait loci from a dense set of markers. Genetics 151, 373-386 (1999)

14. Piepho HP: A quick method for computing approximate thresholds for quantitative trait loci detection. Genetics 157, 425-432 (2001)

15. Schaid D, C Rowland, D Tines, R Jacobson, G Poland: Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 70, 425-434 (2002)

16. Wang K & J Huang: A score-statistic approach for mapping quantitative trait loci with sibships of arbitrary size. Am J Hum Genet 70, 412-424 (2002)

17. Bai X, L Luo, W Yan, M Rao Kovi, W Zhan, Y Xing: Genetic dissection of rice grain shape using a recombinant inbred line population derived from two contrasting parents and fine mapping a pleiotropic quantitative trait locus qGL7. BMC Genetics, 11:16 (2010)

18. Kelly S, D Nehrenberg, J Peirce, K Hua, B Steffy, T Wiltshire, F Villena, T Garland Jr., Daniel Pomp: Genetic architecture of voluntary exercise in an advanced intercross line of mice. *Physiol Genomics*, 42:190-200 (2010)

19. Weir B: *Genetic Data Analysis II*. Sinauer Associates, Inc. Sunderland, Massachusetts (1996)

20. Kao CH, MH Zeng: An Investigation of the Power for Separating Closely Linked QTL in Experimental Populations. *Genet Res* 92, 283-294 (2010)

21. Darvasi A: Experimental strategies for the genetic dissection of complex traits in animal models. Nat Genet 18, 19 24 (1998)

22. Haldane JBS, C Waddington: Inbreeding and linkage. Genetics 16, 357- 374 (1931).

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

24. Kao CH, MH Zeng: A study on the Mapping of Quantitative Trait Loci in Advanced Populations Derived from Two Inbred Lines. *Genet Res* 91, 85-99 (2009)

25. Kao CH, ZB Zeng: General formulae for obtaining the maximum likelihood estimates and the asymptotic variance-covariance matrix in QTL mapping when using the EM algorithm. Biometrics 53, 653-665 (1997)

26. Cox D, D Hinkley. Theoretical Statistics. Chapman & Hall: London (1974)

27. Haldane JBS: The combination of linkage values and the calculation of distances between the loci of linked factors. *J Genet* 8, 299- 309 (1919)

28. Jiang C, ZB Zeng: Mapping quantitative trait loci with dominant and missing markers in various populations from inbred lines. Genetica 101, 47- 85 (1997)

Abbreviations: QTL: quantitative trait loci, AI: advanced intercrossed, RI: recombinant inbred, MLE: maximum likelihood estimate, LRT: Likelihood ratio test

Key Words Advanced populations, Gaussian process, QTL mapping, Score statistics, Threshold values, Interval mapping, Genotypic distribution, Review

Send correspondence to: Chen-Hung Kao, Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan, ROC, Tel: 886-2-27835611x418, Fax: 886-2-27831523, E-mail:chkao@stat.sinica.edu.tw