

Linkage Disequilibrium Mapping of Quantitative Trait Loci under Truncation Selection

Momiao Xiong^a Ruzong Fan^b Li Jin^c

^aHuman Genetics Center, University of Texas-Houston, Houston, Tex., ^bDepartment of Statistics, Texas A&M University, College Station, Tex., ^cDepartment of Environmental Health, University of Cincinnati, Cincinnati, Ohio, USA

Key Words

Linkage disequilibrium · QTL · Selection · SNP marker · Regression

Abstract

As a dense map of single nucleotide polymorphism (SNP) markers are available, population-based linkage disequilibrium (LD) mapping or association study is becoming one of the major tools for identifying quantitative trait loci (QTL) and for fine gene mapping. However, in many cases, LD between the marker and trait locus is not very strong. Approaches that maximize the potential of detecting LD will be essential for the success of LD mapping of QTL. In this paper, we propose two strategies for increasing the probability of detecting LD: (1) phenotypic selection and (2) haplotype LD mapping. To provide the foundations for LD mapping of QTL under selection, we develop analytic tools for assessing the impact of phenotypic selection on allele and haplotype frequencies, and LD under three trait models: single trait locus, two unlinked trait loci, and two linked trait loci with or without epistasis. In addition to a traditional χ^2 test, which compares the difference in allele or haplotype frequencies in the selected sample and population sample, we present multiple regression methods for LD mapping of QTL, and investigate which methods are effective in employing phenotypic selection for QTL mapping. We also develop a statisti-

cal framework for investigating and comparing the power of the single marker and multilocus haplotype test for LD mapping of QTL. Finally, the proposed methods are applied to mapping QTL influencing variation in systolic blood pressure in an isolated Chinese population.

Copyright © 2002 S. Karger AG, Basel

Introduction

Understanding the genetic basis of quantitative trait variation within and between populations entails identifying the number of trait loci and their locations in the genome, estimating the genetic effects of the individual trait locus, and linking genotypes to phenotypes. The most widely used method for mapping quantitative trait loci (QTL) in humans is linkage analysis [Haseman and Elston, 1972; Amos et al., 1989; Schork, 1993; Blangero and Almasy, 1997; Blangero et al., 2000; Wu et al., 1994]. Linkage analysis, the most reliable of gene mapping methods when applied to Mendelian traits or to quantitative trait loci with high heritability, has proven to be a much less reliable tool for mapping QTL with small effects [Risch, 2000]. The lack of success of linkage analysis for mapping complex traits, including quantitative trait loci, with small effects coupled with great progress in the development of dense single nucleotide polymorphism (SNP) maps of the human genome [Gray et al., 2000; Horikawa

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2002 S. Karger AG, Basel
0001-5652/02/0533-0158\$18.50/0

Accessible online at:
www.karger.com/journals/hhe

Dr. Momiao Xiong
Human Genetics Center
University of Texas-Houston
PO Box 20334, Houston, TX 77225 (USA)
Fax +1 713 500 0900, E-Mail mxiong@sph.uth.tmc.edu

et al., 2000; The International SNP Map Working Group, 2001; Wang et al., 1998] has led to a novel approach of linkage disequilibrium (LD) mapping [Kaplan and Morris, 2001]. Genome-wide direct association studies [Risch and Merikangas, 1996] and indirect association studies [Collins et al., 1997] have been proposed.

In contrast to linkage analysis, which requires pedigrees, association studies or LD mapping of genetic traits use both pedigree and population data. Population-based case-control designs compare allele frequencies in cases with those in controls, and are relatively straightforward to implement. One way of adapting population-based case-control studies for quantitative traits is to use selective genotyping [Slatkin, 1999]. The selected sample corresponds to the cases, and the population sample corresponds to the controls. Simple χ^2 and t tests were proposed to detect a rare allele of large effect at a QTL by contrasting allele frequencies in the selected sample versus population sample or comparing the mean of individuals with two different genotypes in the selected sample. Selected genotyping has also been applied to detection of QTL for multiple traits [Bovenhuis and Spelman, 2000] and to mapping QTLs with epistasis in rats with hypertension [Ohno et al., 2000].

Maximizing the potential of detecting LD is fundamental to the success of LD mapping of QTL. In this paper, we propose two strategies for increasing the observed LD signal: (1) phenotypic selection and (2) haplotype LD mapping. Selective genotyping is an efficient way to increase LD. To facilitate LD mapping of QTL, we first develop analytic tools for investigating the change in allele and haplotype frequencies and LD under truncation selection. Most of the previous work has focused on one locus (a trait locus or a marker locus) [Narain, 1990; Lynch and Walsh, 1998; Slatkin, 1999]. Theoretic analysis for selection on multiple loci was investigated by Turelli and Barton [1994], which involved sophisticated mathematics. In this paper, we consider both one locus and two linked or unlinked locus models with or without epistasis. These analytic formulas will reveal the dependency of allele and haplotype frequencies, and LD on the selection intensity, the allele and haplotype frequencies before selection, the genetic additive and dominance effects, additive \times additive, additive \times dominance, dominance \times additive, and dominance \times dominance epistatic effects. These theoretical results will provide the foundation for evaluating the power of statistical methods for mapping QTL in the selected sample. We demonstrate that the regression model for mapping QTL has higher power than the most commonly used χ^2 test, which sug-

gests that the regression model is a powerful population-based method for QTL analysis. Finally, the proposed methods and strategies are applied to the mapping of a QTL influencing variation in systolic blood pressure in an isolated Chinese population.

LD under Truncation Selection

Single Locus

We consider a trait locus with two alleles Q and q and a marker locus with two alleles M and m . Let P_Q and P'_Q (or P_q and P'_q) be the frequency of trait allele Q (or allele q) in the population and the selected sample, respectively. Let P_M and P'_M (or P_m and P'_m) be the frequency of allele M (or allele m) in the population and selected sample, respectively. Let D' and D be the measure of LD between the marker and disease loci in the selected sample and population, respectively. Suppose that a phenotypic value y in a population is normally distributed, and is influenced by the trait locus Q and environmental factors. We assume that the phenotypic means of genotypes QQ , Qq , and qq are $\mu + a$, $\mu + d$, and $\mu - a$, respectively. The environmental effect is normally distributed $N(0, \sigma_e^2)$. Truncation selection is defined as including individuals in the sample if their phenotypic values exceed a truncation point T . Truncation selection produces a selected sample with different gene frequencies at the trait locus from the overall population. Let I be a measure of the intensity of selection and α_Q be the average effect of the gene substitution.

It can be shown that the difference in the frequency of the trait allele Q , in the population and the selected sample is given by

$$P'_Q - P_Q = P_Q P_q \alpha_Q I / \sigma_p,$$

where σ_p^2 is the total variance of the phenotypic value y [Bulmer, 1980; Falconer and Mackay, 1996; Hartl and Clark, 1989, p. 456; Turelli and Barton, 1994]. We show that the difference in the marker allele frequency is given by (Appendix A)

$$P'_M - P_M = \frac{D \alpha_Q I}{\sigma_p}. \quad (1)$$

The above equation demonstrates that the change in the marker allele frequency due to truncation selection depends on the intensity of selection, the strength of LD between the marker and trait loci, the average effect of the gene substitution, and the standard deviation of the phenotypic values. In the absence of LD, selection will not cause a change in the marker allele frequencies. Let P'_{MQ} and P_{MQ} be the frequency of haplotype MQ , in the select-

ed sample and population, respectively. Similarly, we can show that (Appendix A)

$$P'_{MQ} - P_{MQ} = \frac{P_{MQ}P_q\alpha_Q I}{\sigma_p} \quad (2)$$

Let the heritability be denoted by h^2 . Using equations (1) and (2), we can obtain

$$D' = D \left[1 + \frac{(P_q - P_Q)\alpha_Q I}{\sigma_p} - \frac{h^2}{2} I^2 \right] \quad (3)$$

It is interesting to note that when $P_q < P_Q$, the level of the linkage disequilibrium may be reduced in the selected sample.

Two Trait Loci

Consider two trait loci with alleles A and a at the first locus, and alleles B and b at the second locus. Let D_{AB} be the measure of the LD between these two loci in the population. The additive \times additive, additive \times dominance, dominance \times additive, and dominance \times dominance effects are denoted by e_{AA} , e_{AD} , e_{DA} , and e_{DD} , respectively. Let P_a , P_A , P_b , and P_B be the frequencies of the alleles a , A , b and B , respectively. Denote the average effects of the gene substitution at the first and second locus by α_A and α_B , respectively. Let P'_{AB} , P'_{Ab} , P'_{aB} and P'_{ab} , and P_{AB} , P_{Ab} , P_{aB} and P_{ab} be the frequencies of the haplotypes AB , Ab , aB and ab in the selected sample and population sample, respectively. We obtain the following equations in Appendix B:

$$\begin{aligned} P'_{AB} - P_{AB} &= \frac{P_{AB}I}{\sigma_p} [P_a\alpha_A + P_b\alpha_B + e_{AB}], \\ P'_{Ab} - P_{Ab} &= \frac{P_{Ab}I}{\sigma_p} [P_a\alpha_A - P_B\alpha_B + e_{Ab}], \\ P'_{aB} - P_{aB} &= \frac{P_{aB}I}{\sigma_p} [-P_A\alpha_A + P_b\alpha_B + e_{aB}], \\ P'_{ab} - P_{ab} &= \frac{P_{ab}I}{\sigma_p} [-P_A\alpha_A - P_B\alpha_B + e_{ab}], \end{aligned} \quad (4)$$

where the formulas for e_{AB} , e_{Ab} , e_{aB} and e_{ab} are given in Appendix B. The above equations show that the change in haplotype frequencies depends upon the average effects of the gene substitution at the individual trait locus and epistatic effects between the trait loci. From equation (4), we can obtain the change in allele frequency for trait A in selected and population samples:

$$\begin{aligned} P'_A - P_A &= \frac{I}{\sigma_p} \{P_A P_a \alpha_A + D_{AB} \alpha_B + P_A P_a (P_B - P_b) e_{AA} \\ &+ 2(P_A P_a P_b P_B - D_{AB}^2) e_{AD} \\ &+ (P_A - P_a) [P_a P_a (P_B - P_b) + D_{AB} (P_A - P_a)] e_{DA} \\ &+ (P_a - P_A) [2P_a P_A P_b P_B + (P_B - P_b) (P_A - P_a) D_{AB} + 2D_{AB}^2] e_{DD}\}. \end{aligned}$$

Statistical Models

Frequency-Based Test Statistics

In the preceding section we demonstrated that truncation selection, or selective genotyping, will cause a change in the trait allele frequencies or marker allele frequencies in the presence of LD between the marker and trait loci. Intuitively, the differences in allele frequencies in the selected sample and population can be used for testing the existence of QTL. For single locus analyses, Slatkin [1999] assumed that the frequency of the trait allele is small and used a χ^2 test based on the differences in the numbers of individuals with a heterozygous genotype in the selected sample and population for QTL mapping. In this section, we consider a χ^2 test based on the differences in allele frequency between the selected sample and population sample, and extend this χ^2 test to multiple loci.

For simplicity of presentation, we begin with a single locus. Suppose that the selected sample corresponds to the case, and the population sample corresponds to the control. Clearly, this fits the design of case-control studies. We may consider a marker locus M which is linked to trait locus Q . Let n_{sM} and n_{sm} be the numbers of marker alleles M and m in the selected sample, respectively, and n_{pM} and n_{pm} be the numbers of alleles M and m in the population sample, respectively. Denote $n_M = n_{sM} + n_{pM}$, $n_m = n_{sm} + n_{pm}$, and $n_s = n_{sM} + n_{sm}$, $n_p = n_{pM} + n_{pm}$. Let $n = n_M + n_m$ be the total number of alleles in the total samples (selected sample and population sample), and $E_{sM} = n_s n_M / n$, $E_{sm} = n_s n_m / n$, $E_{pM} = n_p n_M / n$, and $E_{pm} = n_p n_m / n$. We assume that sample sizes are large enough to allow asymptotic theory to apply. The statistic

$$T_M = \frac{(n_{sM} - E_{sM})^2}{E_{sM}} + \frac{(n_{sm} - E_{sm})^2}{E_{sm}} + \frac{(n_{pM} - E_{pM})^2}{E_{pM}} + \frac{(n_{pm} - E_{pm})^2}{E_{pm}}$$

is asymptotically distributed as a $\chi^2_{(1)}$ distribution under the null hypothesis of no LD between the marker and trait loci. If the marker locus is the trait locus itself, the above test statistic, in which the number of marker alleles is replaced by the corresponding number of trait alleles, can be used for testing the null hypothesis of the absence of QTLs in the region being tested and is denoted by T_Q .

The rationale for extending a single locus χ^2 test to multiple loci is twofold. First, a quantitative trait is controlled by multiple loci, which may be linked or unlinked and interact. Therefore, simultaneously considering multiple loci may increase power. Second, when the number of SNP markers rapidly increases, there may exist a number of SNP markers within a gene influencing the phenotype of interest. The joint use of several SNP markers is

efficient for QTL analysis. Therefore, it is clear that we need to simultaneously consider multiple loci. A straightforward extension of the classic χ^2 test of equality of allele frequencies for a single locus is to consider the difference in haplotype frequencies in the selected sample and population sample for multiple loci.

Suppose that there are t haplotypes among k loci. The number of haplotypes in the selected sample and population sample can be arranged into a $2 \times t$ contingency table where 2 is the number of rows (selected sample and population sample) and t is the number of columns (haplotypes). The expected count E_i for each cell in the table is calculated as the row total multiplied by the column total divided by the table total. Then the test statistic is given by

$$T_H = \sum_{i=1}^{2t} (O_i - E_i)^2 / E_i,$$

where O_i is the observed number of the haplotypes in the cell. Equation (4) shows that for two loci the null hypothesis of no difference in haplotype frequencies between the selected sample and the population sample is equivalent to the null hypothesis of the absence of any QTLs (or functional sites in the region of interests). Standard statistical theory shows that under the null hypothesis T_H follows a χ^2 distribution with $t - 1$ degrees of freedom.

Selection of markers in an analysis is an important issue in the application of T_H statistics. On one hand, using more markers will provide more information. On the other hand, it will also increase the degrees of freedom of the test statistic. In practice, the balance between the gain in more information and increase of the degrees of freedom will determine the appropriate number of markers to be considered.

Regression Methods

Frequency-based tests ignore phenotypic value information of individuals. To employ phenotypic value information of individuals we present regression methods. Regression methods have been widely applied to mapping QTLs in plants and animals. In human genetics, the popular methods for mapping QTLs are to use family structure and pedigree data to perform linkage analysis or association studies. In the presence of linkage between the marker and trait loci, population-based LD methods are useful for fine mapping. In the following, we present regression methods that can be applied to a selected sample or a non-selected population sample for QTL analysis in humans.

Here, we consider multiple loci with epistasis. Assume k loci, M_1, M_2, \dots and M_k , in which there are two alleles M_j and m_j , each with frequencies P_{M_j} and P_{m_j} , respectively. Let y_i be the trait value of the i -th individual in the selected sample ($i = 1, \dots, n$). Let G_{ij} be the genotype of the i -th individual at the j -th marker locus M_j . Let x_{ij} and z_{ij} be the indicator variables at the j -th locus, depending on the genotype of the i -th individual as follows

$$x_{ij} = \begin{cases} 1 & \text{if } G_{ij} = M_j M_j \\ 0 & \text{if } G_{ij} = M_j m_j \\ -1 & \text{if } G_{ij} = m_j m_j \end{cases}$$

and

$$z_{ij} = \begin{cases} 0 & \text{if } G_{ij} = M_j M_j \\ 1 & \text{if } G_{ij} = M_j m_j \\ 0 & \text{if } G_{ij} = m_j m_j \end{cases}$$

If only digenic epistasis is considered, the relation between the trait value of individual i , and the genetic parameters can be expressed by the following equation

$$\begin{aligned} y_i = & \mu + w_i \gamma + \sum_{j=1}^k x_{ij} \alpha_j + \sum_{j=1}^k z_{ij} \delta_j + \sum_{j=1}^k \sum_{l=j+1}^k (x_{ij} x_{il}) e_{AA}^{jl} \\ & + \sum_{j=1}^k \sum_{l=j+1}^k (x_{ij} z_{il}) e_{AD}^{jl} + \sum_{j=1}^k \sum_{l=j+1}^k (z_{ij} x_{il}) e_{DA}^{jl} \\ & + \sum_{j=1}^k \sum_{l=j+1}^k (z_{ij} z_{il}) e_{DD}^{jl} + e_i, \end{aligned}$$

where μ is the overall mean, w_i is a p -dimensional vector of covariate observations such as age, sex, or other observations, γ is a p -dimensional vector of regression coefficients associated with w_i , α_j is the additive genotypic value at the j -th locus, δ_j is the dominance genotype value at the j -th locus, e_{AA}^{jl} , e_{AD}^{jl} , e_{DA}^{jl} and e_{DD}^{jl} are additive \times additive, additive \times dominance, dominance \times additive, dominance \times dominance epistatic values between the j -th locus and the l -th locus, and e_i is an error term with $E[e_i] = 0$ and $\text{Var}(e_i) = \sigma_e^2$. To test the existence of the j -th QTL, one simply fits the model with and without the terms corresponding to the j -th QTL in the model. The improvement in the residual sum of squares (the extra sum of squares) is compared to the residual sum of squares for the full model via the F -ratio.

Power Calculation and Comparison

To evaluate the performance of the proposed test statistics for mapping QTL using selected samples we calculate the power of these statistics to detect QTL. For the

convenience of presentation, throughout this section we assume that the sample size is large enough that asymptotic theory can be applied. To apply large sample theory, we assume that the ratio of population sample size over selected sample size is a constant, denoted as $r = n_p/n_s$. Recall that $n = n_s + n_p$. Consider a marker locus. It can be shown from standard statistical theory that T_M is asymptotically distributed as a noncentral $\chi^2_{(1)}$ distribution under the alternative hypothesis that LD exists between the marker and trait loci. Some calculations show that its noncentrality parameter is given by

$$\lambda_M = \frac{nr}{1+r} \left[\frac{(P'_M - P_M)^2}{P'_M + rP_M} + \frac{(P'_m - P_m)^2}{P'_m - rP_m} \right]$$

$$= \frac{nr}{1+r} \frac{D^2 h^2 I^2}{2P_Q P_q} \left[\frac{1}{(1+r)P_M + \frac{D\alpha_Q I}{\sigma_P}} + \frac{1}{(1+r)P_m - \frac{D\alpha_Q I}{\sigma_P}} \right],$$

where h^2 is the heritability. This demonstrates that at the marker locus the noncentrality parameter λ_M depends on the square of the measure of the LD, D^2 . In the absence of LD, or the presence of weak LD between the marker and trait loci, selective genotyping design has little power to detect QTL at the marker locus. If the marker locus is the trait locus itself, then $D = P_Q P_q$ and λ_M is reduced to

$$\lambda_Q = \frac{nr}{1+r} \frac{h^2 I^2}{2} \left[\frac{P_q}{1+r + \frac{P_q \alpha_Q I}{\sigma_P}} + \frac{P_Q}{1+r - \frac{P_Q \alpha_Q I}{\sigma_P}} \right].$$

Next we discuss the test statistic T_H . For the simplicity of presentation, we consider two trait loci. The extension to multiple loci is straightforward, but the notation is more complicated. We assume that each trait locus has two alleles. The first locus has alleles A and a and the second locus has alleles B and b . The total sample size, n , and the ratio of the sample sizes, r , are defined as before. Following the arguments similar to that for the single locus, we can show that the statistic T_H follows a noncentral $\chi^2_{(3)}$ distribution under the alternative hypothesis of the presence of QTLs and that the noncentrality parameter T_H is given by

$$\lambda_H = \frac{nr}{1+r} \left[\frac{(P'_{AB} - P_{AB})^2}{P'_{AB} - rP_{AB}} + \frac{(P'_{Ab} - P_{Ab})^2}{P'_{Ab} + rP_{Ab}} + \frac{(P'_{aB} - P_{aB})^2}{P'_{aB} + rP_{aB}} + \frac{(P'_{ab} - P_{ab})^2}{P'_{ab} + rP_{ab}} \right]$$

$$= \frac{nr I^2}{(1+r)\sigma_P^2} \left[\frac{P_{AB}(P_a \alpha_A + P_b \alpha_B + e_{AB})^2}{1+r + I(P_a \alpha_A + P_b \alpha_B + e_{AB})/\sigma_P} \right.$$

$$\left. + \frac{P_{Ab}(P_a \alpha_A + P_B \alpha_B + e_{Ab})^2}{1+r + I(P_a \alpha_A + P_B \alpha_B + e_{Ab})/\sigma_P} \right]$$

$$+ \frac{P_{aB}(P_a \alpha_A + P_b \alpha_B + e_{aB})^2}{1+r + I(P_a \alpha_A + P_b \alpha_B + e_{aB})/\sigma_P}$$

$$+ \frac{P_{ab}(-P_A \alpha_A + P_b \alpha_B + e_{ab})^2}{1+r + I(-P_A \alpha_A + P_b \alpha_B + e_{ab})/\sigma_P} \Big].$$

The above equation demonstrates that the noncentrality parameter (and hence power to detect QTL) depends on the intensity of selection, the additive and dominance effects, and epistasis effects as well.

Now we discuss the asymptotic distribution of the test statistic in regression methods under the alternative hypothesis. We first consider a single trait locus and a marker locus. For the convenience of presentation, throughout this section, we assume no covariates in the model. The null hypothesis $H_0: \alpha = \delta = 0$ can be expressed in matrix form $H\beta = 0$, where $H = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ and $\beta = (\mu, \alpha, \delta)^T$. Let S_F^2 and S_R^2 be the residual variances estimated from the full model and the reduced model ($\alpha = 0$ and $\delta = 0$), respectively. The test statistic is given by $T_R = \frac{n-p-3}{2} (S_R^2 - S_F^2)/S_F^2$. Standard linear model theory shows that [Graybill, 1976] T_R follows a noncentral $F_{2, n-3}$ distribution. At the trait locus, $\lambda_R = \lambda_{RQ}$ and $\lambda'_R = \lambda'_{RQ}$ are given by (Appendix C)

$$\lambda_{RQ} \approx \frac{n}{\sigma_e^2} [2P_Q P_q \alpha_Q^2 + (2dP_Q P_q)^2],$$

$$\lambda'_{RQ} \approx \frac{n}{\sigma_e^2} [2P'_Q P'_q \alpha_Q^2 + (2dP'_Q P'_q)^2].$$

At the marker locus, $\lambda_R = \lambda_{RM}$ and $\lambda'_R = \lambda'_{RM}$ are given by (Appendix C)

$$\lambda_{RM} \approx \frac{n}{\sigma_e^2} \left[2P_Q P_q \alpha_Q^2 \frac{D^2}{P_Q P_q P_M P_m} + (2dP_Q P_q)^2 \frac{D^4}{P_Q^2 P_q^2 P_M^2 P_m^2} \right]$$

$$\lambda'_{RM} \approx \frac{n}{\sigma_e^2} \left[2P'_Q P'_q \alpha_Q^2 \frac{D^2}{P'_Q P'_q P'_M P'_m} + (2dP'_Q P'_q)^2 \frac{D^4}{P_Q'^2 P_q'^2 P_M'^2 P_m'^2} \right], \quad (5)$$

where $\alpha'_Q = a + (P'_q - P_Q)d$. Next we consider two trait loci with or without epistasis. The formula for calculating the noncentrality parameter is complicated and summarized in Appendix D.

To evaluate the performance of the test statistics, we compare their power in the population and selected samples. Figures 1 and 2 show the power of the test statistic T_Q (comparing difference in allele frequencies between the selected sample and population sample) and T_R (based on regression) for recessive and dominant traits, respectively as a function of allele frequency. Figures 1 and 2 demonstrate that, in general, T_R has higher power compared to that of T_Q . For recessive disease, in most cases T_R in the selected sample has less power than that of T_R in the population. For dominant diseases, T_R in the

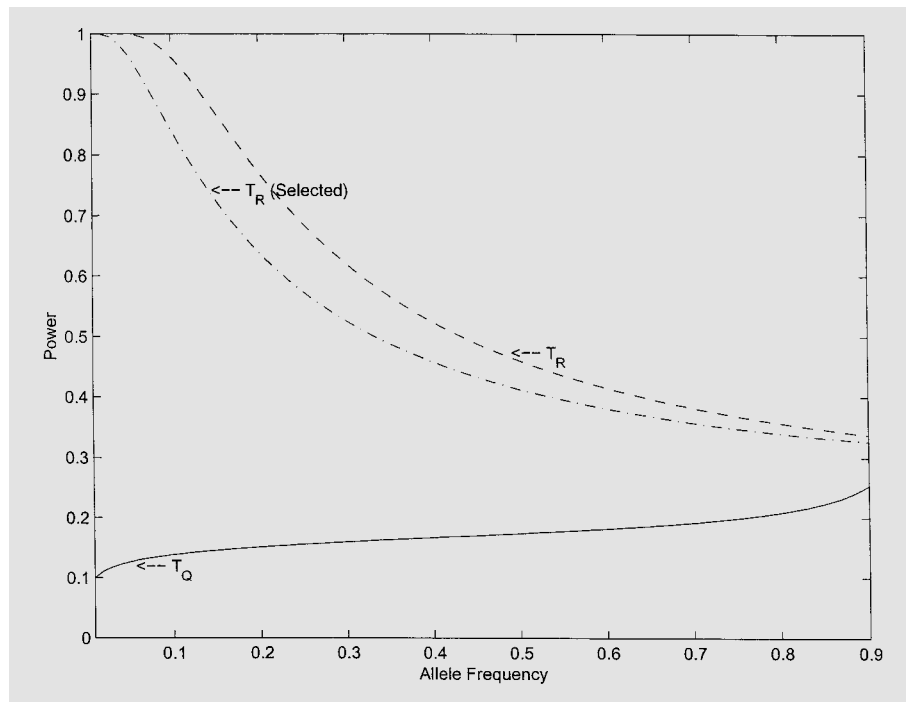


Fig. 1. Power of the χ^2 test T_Q and regression based test T_R as a function of frequency of a trait allele in the population sample and selected sample, assuming $a = 0.5$, $d = -0.5$, $h^2 = 0.05$, $I^2 = 1.74$, $r = 0.5$ and $n = 100$.

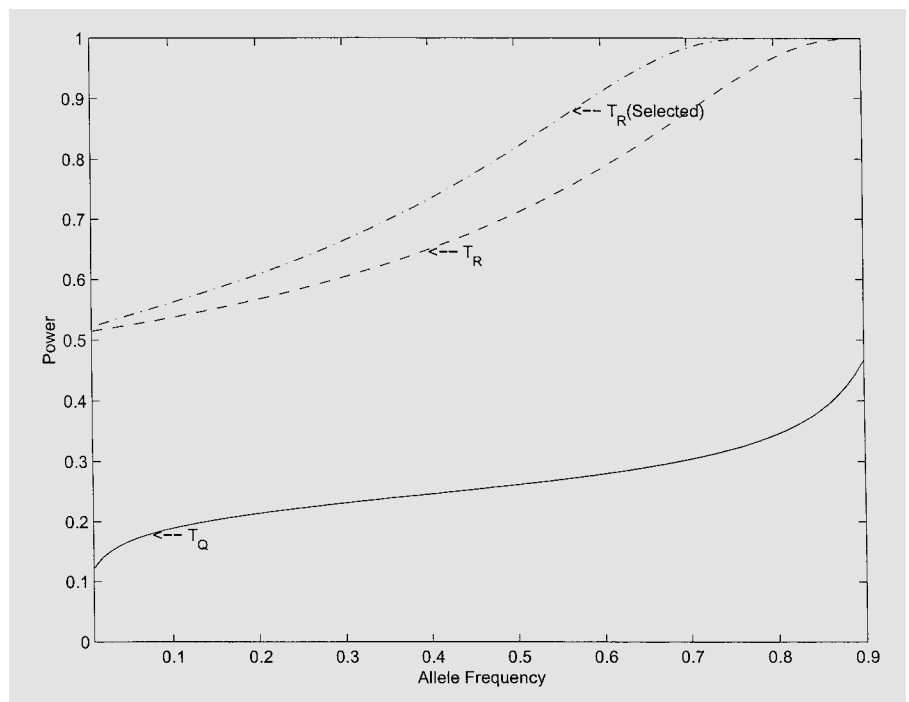


Fig. 2. Power of the χ^2 test T_Q and regression based test T_R with a significance level $\alpha = 0.05$ as a function of frequency of a trait allele in the population sample and selected sample, assuming $a = 0.5$, $d = 0.5$. Other parameters are assumed the same as that of figure 1.

selected sample has higher power than that of T_R in the population. These results tell us that the power of T_R in the selected sample may not always be higher than that in the population samples. Further investigation shows that the power of T_R in both selected and population samples

depends on d , the dominance quantity. Thus, d determines whether power of the T_R in the selected samples is higher than that in population (data not shown).

When the markers are far away from the trait locus, the impact of the selection intensity on the power of the

Fig. 3. Power of the χ^2 test T_M and regression based test T_R with a significance level $\alpha = 0.05$ as a function of genetic distance between the marker and trait loci in the population sample and selected sample, assuming $a = 0.5$, $d = -0.5$, $h^2 = 0.5$, $I^2 = 1.74$, $r = 0.5$, $D_0 = 0.25$, $P_M = 0.1$, $t = 20$ and $n = 100$.

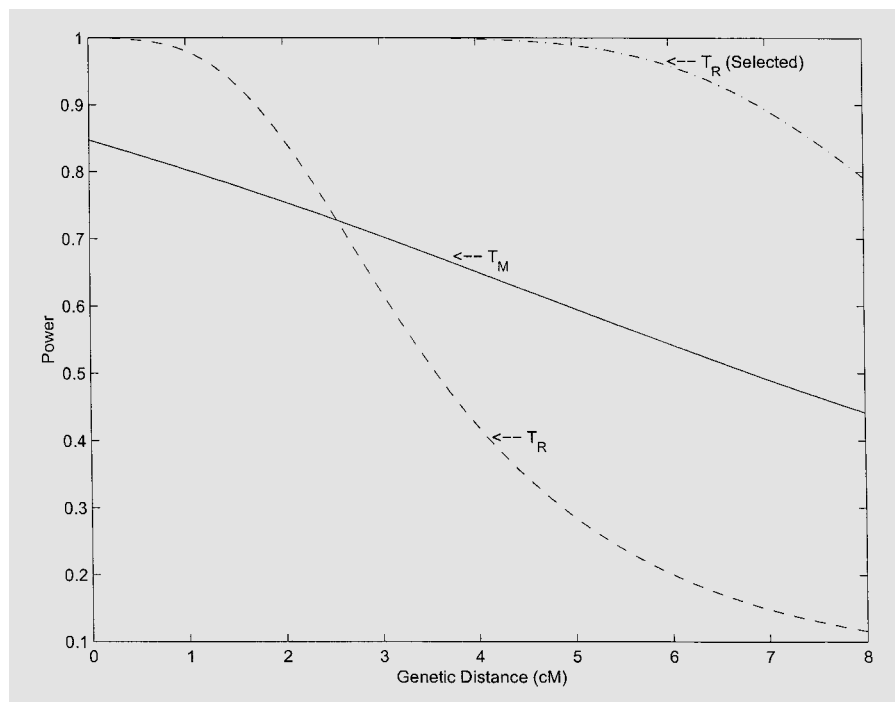
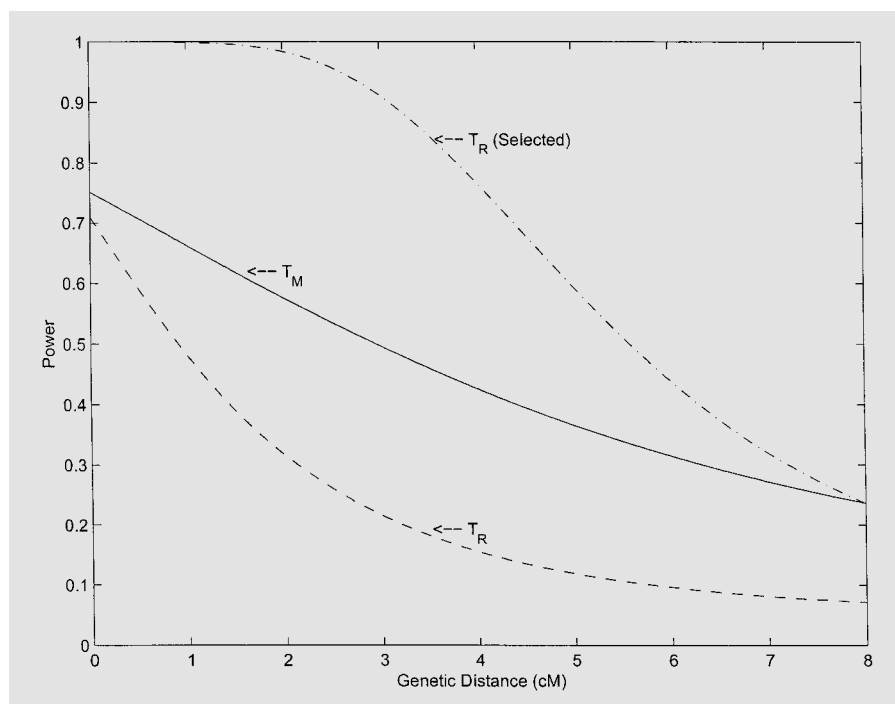


Fig. 4. Power of the χ^2 test T_M and regression based test T_R with a significance level $\alpha = 0.05$ as a function of genetic distance between the marker and trait loci in the population sample and selected sample, the parameters are assumed the same as that of figure 3 except that $P_M = 0.5$.



regression based test T_R is dramatically reduced. For the convenience of presentation, the average level of the LD between the marker and trait loci, $D(t)$ is expressed as $D(t) = D_0 e^{-\theta t}$, where D_0 is the level of the initial LD when the trait mutation is introduced into the population, θ is

the recombination fraction between the marker and trait loci, and t is the time in elapsed generations since the introduction of the trait mutation into the population [Ewens, 1979]. Figures 3 and 4 show the power of the χ^2 test T_M and regression based test T_R as a function of the

Fig. 5. Power of the χ^2 test T_H and regression based test for two unlinked trait loci with a significance level $\alpha = 0.05$ as a function of frequency of trait allele in the population sample and selected sample, assuming $a_A = a_B = 0.5$, $d_A = d_B = -0.5$, $e_{AA} = e_{AD} = e_{DA} = e_{DD} = 0$, $h^2 = 0.05$, $r = 0.5$, $D_{AB} = 0$ and $n = 100$. The frequencies of the trait alleles at two loci are assumed to be equal.

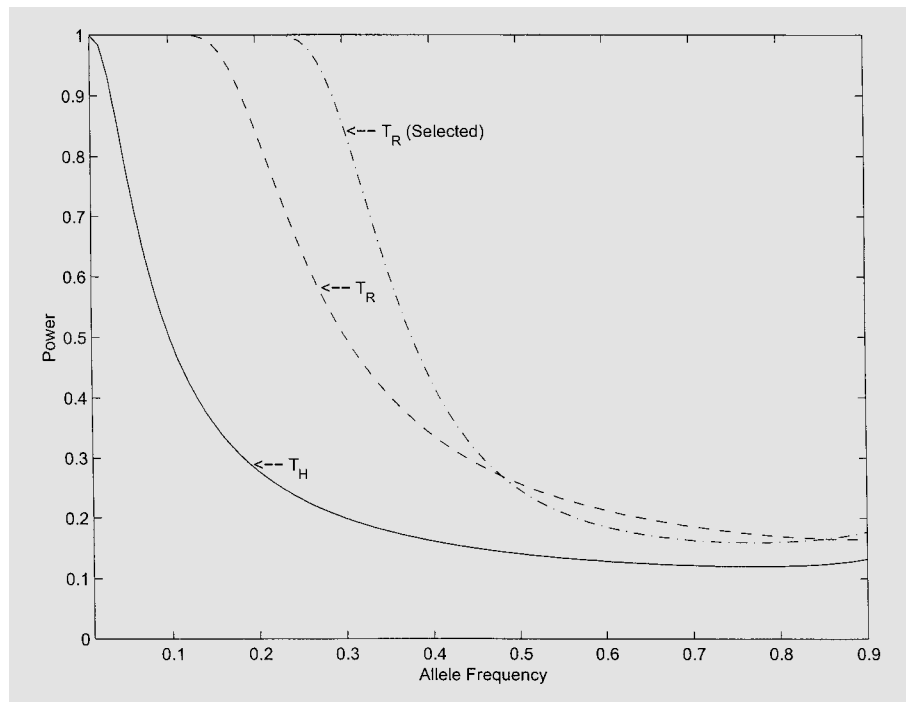
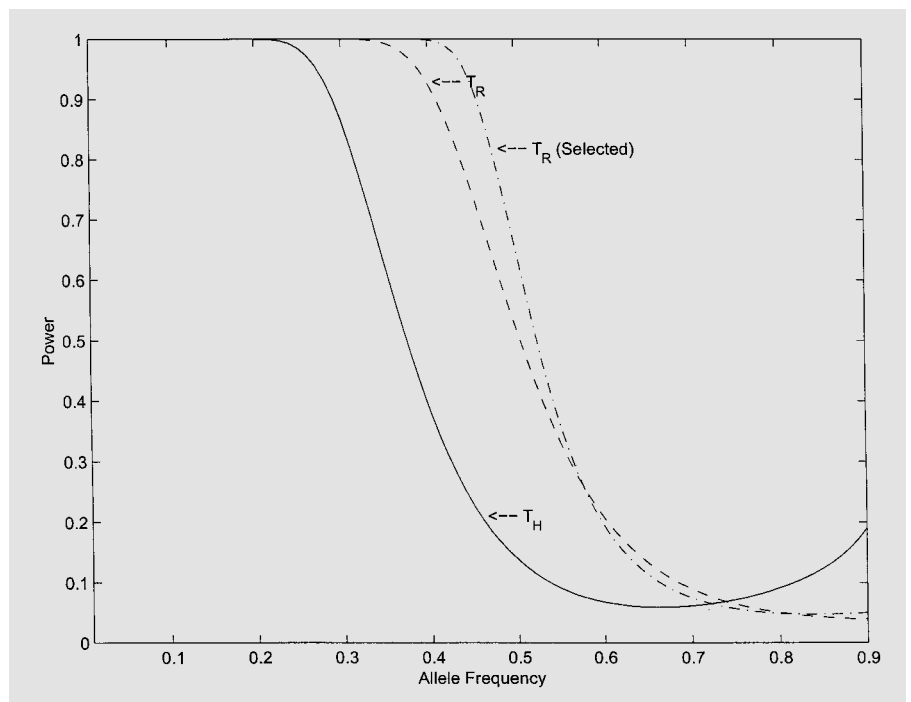


Fig. 6. Power of the χ^2 test T_H and regression based test for two unlinked loci with a significance level $\alpha = 0.05$ as a function of frequencies of trait alleles in the population sample and selected sample, assuming equal allele frequencies at two trait loci, $a_A = 0.5$, $a_B = 0.5$, $d_A = d_B = -0.5$, $e_{AA} = -0.53$, $e_{AD} = 0.64$, $e_{DA} = -0.53$, $e_{DD} = 1.18$, $h^2 = 0.05$, $r = 0.5$, $D_{AB} = 0$ and $n = 100$.



genetic distance between the marker and trait loci. The genetic distance is transformed to recombination fraction via Haldane's mapping function.

Two features emerge from figures 3 and 4. First, the power of the regression-based test T_R in the selected sam-

ple is highest among the three tests. Second, whether the power of T_M is higher than that of T_R in the population samples depends on the frequency of the marker allele. When the marker allele frequency is equal to $P_M = 0.5$, the power of T_M is always higher than that of T_R in the popu-

Fig. 7. Power of the χ^2 test T_H and regression based test for two loci with a significance level $\alpha = 0.05$ as a function of trait allele frequency in the population sample and selected sample. The parameters are assumed to be the same as that in figure 5 except that the measure of LD D_{AB} is equal to 0.3 and $a_A = a_B = 1.5$.

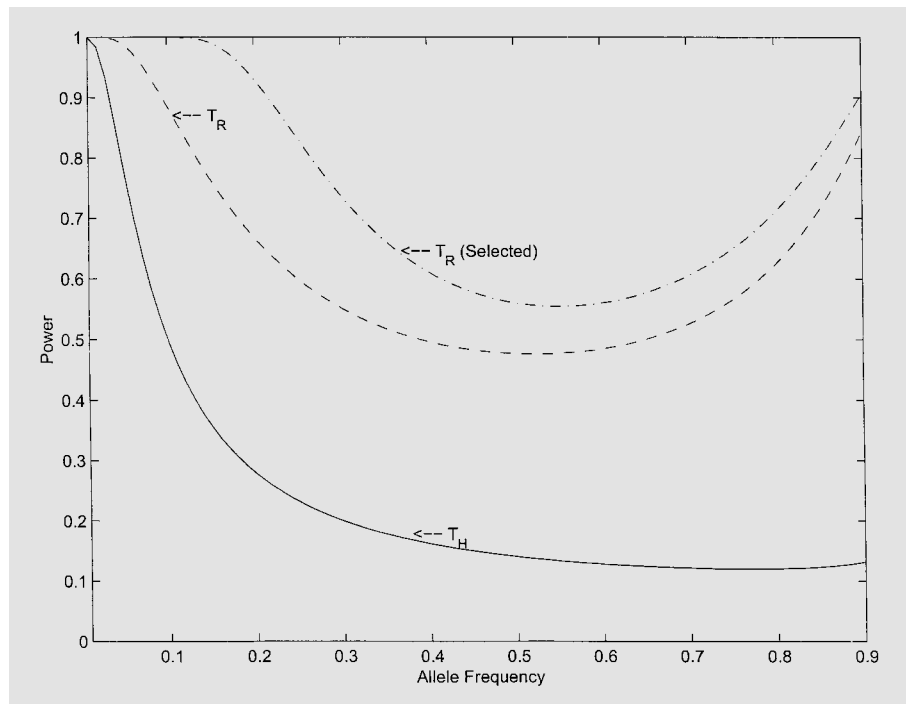
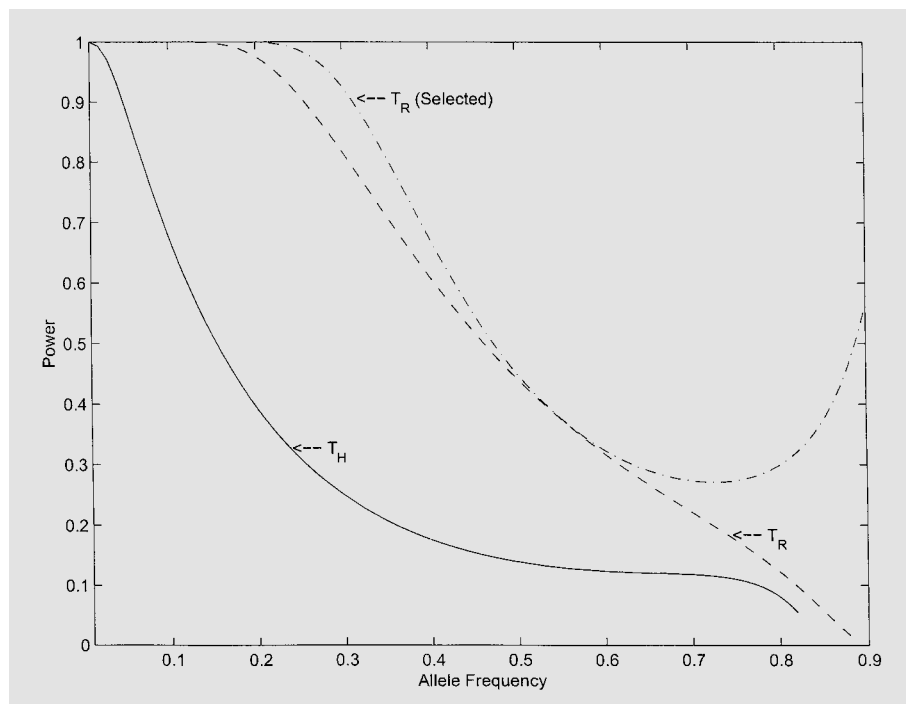


Fig. 8. Power of the χ^2 test T_H and regression based test for two loci with a significance level $\alpha = 0.05$ as a function of trait allele frequency in the population sample and selected sample. The parameters are assumed to be the same as that in figure 6 except that $D_{AB} = 0.3$ and $a_A = a_B = 1.5$.



lation samples. However, when the marker allele frequency is equal to $P_M = 0.1$, then the power of T_M is higher than that of T_R in the population samples for high trait allele frequencies, but lower than that of T_R in the population for low trait allele frequencies.

Now we discuss two trait loci. First, we consider two trait loci without LD. Figure 5 shows the power of the T_H test and the regression based tests for two loci as a function of allele frequency, assuming equal allele frequencies at two loci. It demonstrates that the power of the regres-

Table 1. Results of association study and regression analysis performed for three SNPs: GH1-4886, GH1-5107 and GH1-5158 in GH1 in an isolated Chinese population

a Association studies

	GH1-5158		GH1-4886		GH1-5107	
	χ^2	p value	χ^2	p value	χ^2	p value
Case >160 (46, 21.9%)	0.0807	0.7764	0.3609	0.5480	0.8012	0.3737
Control <120 (41, 19.5%)						

b Regression analysis

	GH1-5158		GH1-4886		GH1-5107	
	F	p value	F	p value	F	p value
Population sample (211)	2.5940	0.0772	3.5310	0.034	4.2425	0.0157
Selected population >120 (145, 69%)	10.0878	0.0001	8.2076	0.0004	6.6392	0.0018

F = F statistic.

sion based test with and without selection is higher than that of the T_H test. The power of T_R in selected samples is higher than that in the population samples for low trait allele frequencies, but somewhat lower than that in the population samples for high trait allele frequencies. Figure 6 shows the power of these three tests in the presence of epistasis. In general, the pattern of figure 6 is similar to that of figure 5. However, when the trait allele frequencies are higher than 0.7, the power of T_R in both selected samples and population samples are a little lower than that of T_H .

Finally, we investigate two linked loci. Figures 7 and 8 show the power of the χ^2 test T_H and the regression based tests for two linked trait loci with $D_{AB} = 0.3$ (D_{AB} is a measure of the LD between two linked trait loci). It is clear that in the presence of LD the power of the T_R in both selected samples and population samples is higher than that of T_H test.

Examples

To illustrate the performance of the proposed methods for identification of QTL, we present applications to detect genes influencing variation in systolic blood pressure (SBP) in an isolated Chinese population, Xiangchang. Two groups, 210 and 244 unrelated individuals who were over 50 years old, were randomly sampled from Xiangchang, Yuexi County, Anhui province in China.

Each subject was visited at his/her home and systolic blood pressure was measured twice at least 20 min apart using a random zero sphygmomanometer. The average of the two readings was used for the analyses reported here. Individuals who were presently taking or had taken antihypertensive medications were not included in the study. Growth hormone 1 (GH1), which provides the signal to express insulin-like growth factor 1 (IGF1), was investigated as a can-

didate gene influencing systolic blood pressure. Previously, it was reported that IGF1 could stimulate cell proliferation in vascular smooth muscle [Andronico et al., 1996]. Three SNPs, GH1 5158, GH1 4886 and GH1 5107 in the promoter region of GH1 were genotyped for both 210 and 297 samples [Jiang et al., 2001]. To perform an association study, individuals with systolic blood pressure greater than a prespecified value were taken as a group of cases and individuals with systolic blood pressure less than a prespecified value were taken as a group of controls. Individuals with systolic blood pressure >120 mm Hg were taken as selected samples. The regression was performed on both population samples (210 samples from Xiangchang) and the selected samples, and the results are summarized in table 1. It can be seen from table 1 that although association studies showed no evidence of association of GH1 with blood pressure, regression analysis suggested some relationship between these three SNP markers in GH1 and blood pressure variation. Specifically, table 1 demonstrates that the regression analysis in the selected sample has higher power to identify trait loci than association studies and regression analysis in the population sample. Recent studies have suggested that IGF1 plays a role in the cardiovascular complication of hypertension [Andronico et al., 1996].

Furthermore, we studied the β_2 -adrenergic receptor (ADRB2) gene and endothelin-1 (EDN1) genes, which have been reported to be associated with blood pressure variation [Rosmond et al., 2000; Morita et al., 1999]. Table 2 shows the results of regression analysis testing the hypothesis of interaction between ADRB2 and EDN1 in affecting systolic blood pressure variation in the sample of 244 unrelated individuals. Regression analysis has shown no evidence that both genes ADRB2 and EDN1 influence systolic blood pressure variation in the population sample. However, regression analysis did show significant evidence that ADRB2 plays a role in explaining systolic blood pressure variation in the selected sample. In the population sample, the regression analysis did not detect interaction between gene ADRB2 and gene EDN1, but in the selected sample, it did reveal an interactive effect between ADRB2 and EDN1 in influencing systolic blood pressure variation.

Table 2. Results of regression analysis for detection of interaction between the genes ADRB2 and EDN1 in an isolated Chinese population

	ADRB2-565		EDN1-2657	
	F	p value	F	p value
Population sample (244)	0.3557	0.7011	0.3539	0.7023
Selected population >140 (153, 62.7%)	7.9542	0.0007	0.7066	0.4061
Selected population >150 (181, 74%)	14.9634	6.5629×10^{-11}	2.1516	0.1252
	ADRB2-565, EDN1-2657			
	F		p value	
Population sample (244)	0.3213		0.5713	
Selected sample >140 (153, 62.7%)	7.4223		0.0078	
Selected sample >150 (181, 74%)	8.1549		0.0060	

F = F statistic.

Discussion

Having the complete sequence of the human genome will have a great impact on genetic studies of complex traits and will lead to new approaches in mapping and identifying complex trait loci. Furthermore, identifying and cataloging sequence variation in populations will provide valuable tools for investigating the effect of functional variants on phenotypes. Positional cloning of a gene showing association with type 2 diabetes is an encouraging example of successful applications of SNPs and LD mapping [Horikawa et al., 2000; Altshuler et al., 2000]. Population-based case-control association studies or LD mapping may emerge as a powerful and efficient approach for gene mapping [Risch, 2000; Chakravarti, 2001] although some pessimism has been voiced [Weiss and Terwilliger, 2000]. In contrast to family-based methods for mapping QTL, which have been the major tools for QTL analysis, in this paper we investigated population-based methods for mapping QTL.

Kaplan and Morris [2201] discussed several issues concerning population-based association studies of qualitative traits by studying the joint behavior of the noncentrality parameters of the test statistic as a function of the frequency of the disease allele. Correspondingly, in this paper, we addressed several issues concerning association studies of quantitative trait under selection by investigating the impact of selection on allele and haplotype frequencies by analyzing the joint behavior of noncentrality parameters of the test statistics as a function of the frequency of the trait allele and genetic distance between the marker and trait locus.

Selective genotyping plays a central role in our analysis. It serves either to generate case samples or to increase the available LD information. To provide the conceptual framework for LD mapping of QTL under selection, we developed analytic tools for assessing the impact of phenotypic selection on LD, allele and haplotype frequencies under three trait models: single trait locus, two unlinked trait loci, and two linked trait loci with or without epistasis. These tools are useful not only for LD mapping of QTL, but also for evolution and population genetic studies.

One of the many benefits of the human genome sequence will be the identification of multiple polymorphic sites within a gene, which occur naturally in the human population. We face a new challenge of simultaneously using linked polymorphic sites. Therefore, in this paper, we considered models for investigating the joint impact of phenotypic selection on allele and haplotype frequencies at linked loci.

To design efficient population-based studies for QTL mapping, we need to evaluate the power of the proposed methods for detecting QTLs and investigating the impact of various factors such as selection intensity, allele and haplotype frequencies before selection (or in the population sample), and the ages of trait mutations on the power to detect QTL. Most existing methods for evaluating the power to detect QTL using selective genotyping are based on simulation. Some analytic methods for evaluating the power of the test statistics for mapping QTL consider simple genetic models [Van Gestel et al., 2000]. In this paper, we developed analytic formulas for calculating the power of population-based association studies for QTL mapping under selection.

The choice of statistical methods for population-based LD mapping of QTL is an important problem in QTL analysis. The developed analytic formulas for power calculation provide tools for comparing the power of different methods. The natural and traditional methods for population-based QTL mapping is to extend the population-based case-control studies used for qualitative traits to quantitative traits. As we showed in this paper, comparing the difference in allele frequencies between the selected sample and population sample is not an efficient method for mapping QTL. As an alternative approach, in this paper we presented multiple regression models for mapping QTL. The multiple regression models for mapping QTL have several advantages. First, the multiple regression model can explicitly model covariates such as age, sex, and epistatic effects. Second, in most cases, multiple regression models using selected samples have higher power to detect QTL than the population-based case-control studies. Third, multiple regression models can be directly applied to the population even with no selection. Our preliminary results showed that selection, in most cases, increased the power of T_R at the marker locus. However, at the trait locus, selection, in some cases, might not always increase the power of T_R . Moreover, we also demonstrated that the impact of selection on the power to detect QTL depends on the parameters of population genetic models such as allele frequencies, genetic variances, and the age of the trait mutation.

To illustrate the performance of the proposed methods for mapping QTL, we presented practical applications to detect genes influencing variation in SBP in an isolated Chinese population. We successfully identified GHI, ADRB2, and EDN1 genes showing evidence of association with SBP variations as well as interaction between ADRB2 and EDN1.

A problem with a population-based case-control design is its sensitivity to population substructure. Population substructure can create a spurious association between markers and disease [Spielman et al., 1993]. To overcome this problem, statistical methods such as genomic control and structured association have recently been proposed to use genomic information in the samples to adjust for population stratification [Devlin et al., 2001]. The impact of selection on these methods will be investigated in the future. The extension of the results to more than two loci is straightforward, but complicated in notation. Further investigation for selection of functional variants which make major contributions to the trait variation is important for QTL analysis and will be pursued in the future.

Acknowledgment

Momiao Xiong is supported by National Institutes of Health grant No. GM56515 and HL5448. The authors thank Dr. Joana Floros and Mr. Joshua M. Akey, and two anonymous reviewers for their helpful comments on this paper, which helped to improve its presentation.

Appendix A

Let $P_{MQ}, P_{Mq}, P_{mQ}, P_{mq}$ and $P'_{MQ}, P'_{Mq}, P'_{mQ}, P'_{mq}$ be the frequencies of the haplotypes $MQ, Mq, mQ,$ and mq before and after selection. The frequencies of the marker genotypes MM and Mm before and after selection are denoted by P_{MM}, P_{Mm}, P'_{MM} and P'_{Mm} , respectively. It is well known that

$$\begin{aligned} P_{MQ} &= P_M P_Q + D, P_{Mq} = P_M P_q - D, P_{mQ} = P_m P_Q - D, \\ P_{mq} &= P_m P_q + D. \end{aligned} \quad (6)$$

The change in genotype frequency at the marker locus is caused by the changes in the trait allele frequencies due to truncation selection. Let $W_{11}, W_{12},$ and W_{22} be the fitness of the genotypes $QQ, Qq,$ and qq , i.e., the proportions of individuals with genotypes $QQ, Qq,$ and qq being selected, respectively. Thus, we have (assuming Hardy-Weinberg equilibrium)

$$\begin{aligned} P'_{MM} &= \frac{P_{MQ}^2 W_{11} + P_{MQ} P_{Mq} W_{12} + P_{Mq}^2 W_{22}}{\bar{W}} \\ P'_{Mm} &= \frac{2P_{MQ} P_{mQ} W_{11} + 2(P_{MQ} P_{mq} + P_{Mq} P_{mQ}) W_{12} + 2P_{Mq} P_{mq} W_{22}}{\bar{W}}, \end{aligned}$$

where $\bar{W} = P_Q^2 W_{11} + 2P_Q P_q W_{12} + P_q^2 W_{22}$ is the average fitness. Using relations in (6), we obtain

$$P'_M = P'_{MM} + P'_{Mm}/2 = P_M + \frac{D(P'_Q - P_Q)}{P_Q P_q}.$$

Since $P'_Q - P_Q = P_Q P_q \alpha_Q I / \sigma_P$, it follows that $P'_M - P_M = D I \alpha_Q / \sigma_P$. From Hartl and Clark [1989, p. 456], we have $W_{11} - W_{12} \approx Z(a - d)$ and $W_{12} - W_{22} \approx Z(a + d)$. Here Z represents the height of the normal density at the point T where one truncates the trait values. Thus

$$\begin{aligned} P'_{MQ} - P_{MQ} &= \frac{P_{MQ} P_Q W_{11} + P_{MQ} P_q W_{12}}{\bar{W}} - P_{MQ} \\ &= \frac{P_{MQ} P_q}{\bar{W}} [P_Q (W_{11} - W_{12}) + P_q (W_{12} - W_{22})] \\ &= \frac{P_{MQ} P_q \alpha_Q Z}{\bar{W}} = \frac{P_{MQ} P_q \alpha_Q I}{\sigma_P}. \end{aligned}$$

In the last equation in the above calculations, we use the facts that the average fitness \bar{W} can be interpreted as the proportion of the population selected, i.e., $\bar{W} = B$, and $Z/B = I / \sigma_P$ [Falconer and Mackay, 1996; Hartl and Clark, 1989].

Appendix B

Consider two trait loci A and B . Let P_{AB} , P_{Ab} , P_{aB} , and P_{ab} be the frequencies of the haplotypes AB , Ab , aB and ab , respectively. Let \bar{W} be the average fitness at two trait loci which is defined by

$$\begin{aligned} \bar{W} = & P_{AB}^2 W_{AABB} + 2P_{AB}P_{Ab}W_{AABb} + P_{Ab}^2 W_{AAbb} + 2P_{AB}P_{aB}W_{AaBB} \\ & + 2P_{AB}P_{ab}W_{AaBb} + 2P_{Ab}P_{aB}W_{AaBb} + 2P_{Ab}P_{ab}W_{Aabb} + P_{ab}^2 W_{aabb} \\ & + 2P_{aB}P_{ab}W_{aaBB} + P_{ab}^2 W_{aaBb}, \end{aligned}$$

where W_{AABB} is the fitness of individuals with the genotype $AABB$ and fitness for the individuals with the other genotypes is similarly defined. Thus,

$$\begin{aligned} P'_{AB} - P_{AB} = & \frac{P_{AB}}{\bar{W}} \{P_{AB}[P_{AB}(W_{AABB} - W_{AABb}) \\ & + P_{aB}(W_{AABB} - W_{AaBB}) + P_{ab}(W_{AABB} - W_{AaBb})] \\ & + P_{Ab}[P_{Ab}(W_{AABb} - W_{AAbb}) + P_{aB}(W_{AABb} - W_{AaBb}) \\ & + P_{ab}(W_{AABb} - W_{Aabb})] + P_{aB}[P_{Ab}(W_{AaBB} - W_{AaBb}) \\ & + P_{aB}(W_{AaBB} - W_{aaBB}) + P_{ab}(W_{AaBB} - W_{aaBb})] \\ & + P_{ab}[P_{Ab}(W_{AaBb} - W_{Aabb}) + P_{aB}(W_{AaBb} - W_{aaBb}) \\ & + P_{ab}(W_{AaBb} - W_{aabb})\}. \end{aligned} \quad (7)$$

Let G_{AABB} be the genotypic value of the individual with genotypes AA and BB at the loci A and B , respectively. Other genotypic values can similarly be defined. The two-locus genotypic values can be expressed as a function of their additive, dominance, and epistatic genotypic values [Cheverud and Routman, 1996]:

$$\begin{aligned} G_{AABB} &= \mu + a_A + a_B + e_{AA}, \\ G_{AABb} &= \mu + a_A + d_B + e_{AD}, \\ G_{AAbb} &= \mu + a_A - a_B - e_{AA}, \\ G_{AaBB} &= \mu + d_A + a_B + e_{DA}, \\ G_{AaBb} &= \mu + d_A + d_B + e_{DD}, \\ G_{Aabb} &= \mu + d_A - a_B - e_{DA}, \\ G_{aaBB} &= \mu - a_A + a_B - e_{AA}, \\ G_{aaBb} &= \mu - a_A + d_B - e_{AD}, \\ G_{aabb} &= \mu - a_A - a_B + e_{AA}, \end{aligned} \quad (8)$$

where μ is overall mean, a_A and a_B are additive effects at the loci A and B , respectively; d_A and d_B are dominant effects at the loci A and B , respectively; e_{AA} , e_{AD} , e_{DA} and e_{DD} are additive \times additive, additive \times dominance, dominance \times additive and dominance \times dominance epistatic genotypic values, respectively. Following the same argument as that in Hartl and Clark [1989, pp 454–456], we obtain

$$\begin{aligned} W_{AABB} - W_{AABb} &= \frac{Z}{\sigma_P} (a_B - d_B + e_{AA} - e_{AD}), \\ W_{AABB} - W_{AaBB} &= \frac{Z}{\sigma_P} (a_A - d_A + e_{AA} - e_{DA}), \\ W_{AABB} - W_{AaBb} &= \frac{Z}{\sigma_P} (a_A + a_B - d_A - d_B + e_{AA} - e_{DD}), \\ W_{AABb} - W_{AAbb} &= \frac{Z}{\sigma_P} (a_B + d_B + e_{AA} + e_{AD}), \end{aligned}$$

$$\begin{aligned} W_{AABb} - W_{AaBb} &= \frac{Z}{\sigma_P} (a_A - d_A + e_{AD} - e_{DD}), \\ W_{AABb} - W_{Aabb} &= \frac{Z}{\sigma_P} (a_A + a_B - d_A + d_B + e_{AD} + e_{DA}), \\ W_{AaBB} - W_{AaBb} &= \frac{Z}{\sigma_P} (a_B - d_B + e_{DA} - e_{DD}), \\ W_{AaBB} - W_{aaBB} &= \frac{Z}{\sigma_P} (a_A + d_A + e_{DA} + e_{AA}), \\ W_{AaBb} - W_{aaBb} &= \frac{Z}{\sigma_P} (a_A + a_B + d_A - d_B + e_{AD} + e_{DA}), \\ W_{AaBb} - W_{Aabb} &= \frac{Z}{\sigma_P} (a_B + d_B + e_{DA} + e_{DD}), \\ W_{AaBb} - W_{aabb} &= \frac{Z}{\sigma_P} (a_A + d_A + e_{AD} + e_{DD}), \\ W_{AaBb} - W_{aabb} &= \frac{Z}{\sigma_P} (a_A + a_B + d_A + d_B - e_{AA} + e_{DD}). \end{aligned} \quad (9)$$

Substituting the fitness in equations (9) into equation (7) yields

$$\begin{aligned} P'_{AB} - P_{AB} = & \frac{ZP_{AB}}{\bar{W}\sigma_P} \{P_{AB}[P_a a_A + P_b a_B - P_a d_A - P_a d_B + (P_b + P_{aB})e_{AA} \\ & - P_{Ab}e_{AD} - P_{aB}e_{DA} - P_{ab}e_{DD}] + P_{Ab}[P_a a_A + P_b a_B - P_a d_A - P_b d_B \\ & + P_{Ab}e_{AA} + (P_{Ab} + P_a)e_{AD} + P_{ab}e_{DA} - P_{aB}e_{DD}] + P_{aB}[P_a a_A + P_b a_B \\ & + P_a d_A - P_b d_B + P_{aB}e_{AA} + P_{ab}e_{AD} + (P_{Ab} + P_a)e_{AD} - P_{Ab}e_{DD}] \\ & + P_{ab}[P_a a_A + P_b a_B + P_a d_A + P_b d_B - P_{ab}e_{AA} + P_{aB}e_{AD} + P_{Ab}e_{DA} \\ & + (P_{Ab} + P_a)e_{DD}]\} \\ = & \frac{P_{AB}I}{\sigma_P} \{P_a a_A + P_b a_B + [(P_A - P_a)P_b + P_a P_B - D_{AB}]e_{AA} \\ & + [P_b P_A (P_b - P_B) + 2P_a P_b P_B + (P_B - 3P_b)D_{AB}]e_{AD} \\ & + [P_a P_B (P_a - P_A) + 2P_b P_a P_A + (P_A - 3P_a)D_{AB}]e_{DA} \\ & + [P_a P_b (1 - 4P_A P_B) + [1 + 2(P_a - P_A)(P_B - P_b)]D_{AB} - 4D_{AB}^2]e_{DD}\} \\ = & \frac{P_{AB}I}{\sigma_P} [P_a a_A + P_b a_B + e_{AB}]. \end{aligned}$$

Similarly, we can derive other equations in (4).

Appendix C

Let $\beta = (\mu, \alpha, \delta)^\tau$, $R_i = (1, x_i, z_i)^\tau$, $R = (R_1, R_2, \dots, R_n)^\tau$, and $H = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$. The estimator of β is given by $\hat{\beta} = (R^\tau R)^{-1} R^\tau Y$, where $Y = (y_1, \dots, y_n)^\tau$. Assume that there are no covariates in the regression model. Taking expectation on both sides of $R_1 y_1 = R_1 (R_1^\tau y_1 + e_1)$, we have $E(R_1 y_1) = E(R_1 R_1^\tau) \beta$. Therefore, we have $\beta = (E R_1 R_1^\tau)^{-1} E(R_1 y_1)$. Notice that

$$E[R_1^\tau R_1] = \begin{pmatrix} 1 & P_M - P_m & 2P_M P_m \\ P_M - P_m & P_M^2 + P_m^2 & 0 \\ 2P_M P_m & 0 & 2P_M P_m \end{pmatrix}$$

and its inverse

$$[E[R_1^T R_1]]^{-1} = \frac{1}{4P_M^2 P_m^2} \begin{pmatrix} P_M^2 + P_m^2 & -(P_M - P_m) & -P_M^2 - P_m^2 \\ -(P_M - P_m) & 1 - 2P_M P_m & P_M - P_m \\ -P_M^2 - P_m^2 & P_M - P_m & 1 \end{pmatrix}.$$

Moreover, it can be shown that

$$\begin{aligned} E y_1 &= a(P_Q - P_q) + 2dP_Q P_q \\ E(x_1 y_1) &= a[P_{MQ}^2 - P_{mq}^2] + 2d[P_{MQ} P_{Mq} - P_{mQ} P_{mq}] - a[P_{Mq}^2 - P_{mq}^2] \\ &= 2D\alpha_Q + (P_Q - P_q)(P_M - P_m)a + 2d(P_M - P_m)P_Q P_q \\ E(z_1 y_1) &= 2aP_{MQ} P_{mQ} + 2d[P_{MQ} P_{mQ} + P_{Mq} P_{mq}] - 2aP_{Mq} P_{mq} \\ &= -2D(P_M - P_m)\alpha_Q + 2P_M P_m [a(P_Q - P_q) + 2dP_Q P_q] + 4dD^2. \end{aligned}$$

Some calculations show that

$$\alpha \approx \frac{D\alpha_Q}{P_M P_m} + \frac{d(P_M - P_m)D^2}{P_M^2 P_m^2}$$

and

$$\delta \approx \frac{D^2 d}{P_M^2 P_m^2}.$$

Standard statistical theory shows [Graybill, 1976] that under the alternative hypothesis the test statistic T_R follows a noncentral $F_{2, n-3}$ distribution with the noncentrality parameter λ_R given by

$$\lambda_R = (H\beta)^\tau [H(R^\tau R)^{-1} H^\tau]^{-1} H\beta / \sigma_e^2. \quad (10)$$

It can be shown that $\frac{1}{n} R^\tau R \xrightarrow{a.s.} E(R_1 R_1^\tau)$ asymptotically by the strong law of large number. Notice that

$$\begin{aligned} H(R^\tau R)^{-1} H^\tau &\approx H[n E(R_1 R_1^\tau)]^{-1} H^\tau \\ &= \frac{1}{4nP_M^2 P_m^2} \begin{pmatrix} 1 - 2P_M P_m & P_M - P_m \\ P_M - P_m & 1 \end{pmatrix} \\ [H(R^\tau R)^{-1} H^\tau]^{-1} &\approx 2nP_M P_m \begin{pmatrix} 1 & P_m - P_M \\ P_m - P_M & 1 - 2P_M P_m \end{pmatrix}. \end{aligned}$$

Some algebra will show that $\lambda_R = \lambda_{RM}$ is approximated by

$$\lambda_{RM} \approx \frac{n}{\sigma_e^2} \left[\sigma_a^2 \frac{D^2}{P_M P_m P_Q P_q} + \sigma_d^2 \frac{D^4}{P_M^2 P_m^2 P_Q^2 P_q^2} \right].$$

If the marker locus M coincides with the trait locus Q , then $P_M = P_Q$, $P_m = P_q$ and $D = P_Q P_q$. Hence $\lambda_{RQ} \approx \frac{n}{\sigma_e^2} [\sigma_a^2 + \sigma_d^2]$. In the selected sample, plugging the corresponding allele frequencies for the selected sample into above equation, we may get λ'_{RQ} . Using equations (1), (3) and (10), we obtain λ'_{RM} in equation (5).

Appendix D

Let $\beta = (\mu, \alpha_1, \delta_1, \alpha_2, \delta_2, e_{AA}, e_{AD}, e_{DA}, e_{DD})$,

$$H = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

$R_i = (1, x_{i1}, z_{i1}, x_{i2}, z_{i2}, x_{i1}x_{i2}, x_{i1}z_{i2}, x_{i2}z_{i1}, z_{i1}z_{i2})^\tau$, and $R = (R_1, R_2, \dots, R_n)^\tau$.

Asymptotically, $\frac{1}{n} R^\tau R \rightarrow E(R_1^T R_1)$. Let $W = E[R_1 R_1^\tau]$. W is a square matrix with dimension, 9. Denote the element of the matrix W by W_{ij} . Some calculations yield

$$\begin{aligned} W_{11} &= 1, W_{12} = P_A - P_a, W_{13} = 2P_A P_a, W_{14} = P_B - P_b, \\ W_{15} &= 2P_B P_b, W_{16} = P_A(P_{AB} - P_{Ab}) + P_a(P_{ab} - P_{aB}), \\ W_{17} &= 2(P_{AB} P_{Ab} - P_{aB} P_{ab}), W_{18} = 2(P_{AB} P_{aB} - P_{Ab} P_{ab}), \\ W_{19} &= 2(P_{AB} P_{ab} + P_{Ab} P_{aB}), W_{22} = P_A^2 + P_a^2, W_{23} = 0, \\ W_{24} &= P_A(P_{AB} - P_{Ab}) + P_a(P_{ab} - P_{aB}), W_{25} = W_{17}, \\ W_{26} &= P_A(P_{AB} - P_{Ab}) + P_a(P_{aB} - P_{ab}), W_{27} = 2(P_{AB} P_{Ab} + P_{aB} P_{ab}), \\ W_{28} &= 0, W_{29} = 0, W_{33} = 2P_A P_a, W_{34} = W_{18}, W_{35} = W_{19}, W_{36} = 0, \\ W_{37} &= 0, W_{38} = W_{18}, W_{39} = W_{19}, W_{44} = P_B^2 + P_b^2, W_{45} = 0, \\ W_{46} &= P_B(P_{AB} - P_{aB}) + P_b(P_{Ab} - P_{ab}), W_{47} = 0, \\ W_{48} &= 2(P_{AB} P_{aB} + P_{Ab} P_{ab}), W_{49} = 0, W_{55} = 2P_B P_b, W_{56} = 0, \\ W_{57} &= W_{17}, W_{58} = 0, W_{59} = W_{19}, W_{66} = P_{AB}^2 + P_{Ab}^2 + P_{aB}^2 + P_{ab}^2, \\ W_{67} &= 0, W_{68} = 0, W_{69} = 0, W_{77} = W_{27}, W_{78} = W_{79} = 0, \\ W_{88} &= 2(P_{AB} P_{aB} + P_{Ab} P_{ab}), W_{89} = 0, W_{99} = W_{19}. \end{aligned}$$

The matrix W is a symmetric matrix. Therefore, we have $W_{ij} = W_{ji}$. The statistics for simultaneously testing the existence of two loci in the population (without selection) under the alternative hypothesis of the presence of two loci with epistasis follows a noncentral $F_{8, n-9}$ distribution with noncentrality parameter

$$\lambda_{AB} = \frac{(H\beta)^\tau [H(R^\tau R)^{-1} H^\tau]^{-1} H\beta}{\sigma_e^2}.$$

Asymptotically,

$$\lambda_{AB} \approx \frac{n(H\beta)^\tau [HW^{-1}H^\tau]^{-1} H\beta}{\sigma_e^2}.$$

Substituting the allele and haplotype frequencies by the corresponding frequencies in the selected sample yields the noncentrality parameter of the F distribution of the test statistic in the selected sample.

References

- Altshuler D, Daly M, Kruglyak L: Guilt by association. *Nat Genet* 2000;26:135–137.
- Amos CI, Elston RC, Wilson AF, Bailey-Wilson JE: A more powerful robust sib-pair test of linkage for quantitative traits. *Genet Epidemiol* 1989; 6:435–449.
- Andronico G, Mangano MT, Ferrara L, Lamanna D, Mule G, Cerasola G: Insulin-like growth factor 1 and pressure load in hypertensive patients. *Am J Hypertension* 1996;6:607–609.
- Blangero J, Almasy L: Multipoint oligogenic linkage analysis of quantitative traits. *Genet Epidemiol* 1997;14:959–964.
- Blangero J, Williams JT, Almasy L: Quantitative trait locus mapping using human pedigrees. *Hum Biol* 2000;72:35–62.
- Bovenhuis H, Spelman RJ: Selective genotyping to detect quantitative trait loci for multiple traits in outbred populations. *J Dairy Sci* 2000;83: 173–180.
- Bulmer MG: *The Mathematical Theory of Quantitative Genetics*. Oxford, Clarendon Press, 1980.
- Chakravarti A: Single nucleotide polymorphisms to a future of genetic medicine. *Nature* 2001;409: 822–823.
- Cheverud JM, Routman EJ: Epistasis as a source of increased additive genetic variance at population bottlenecks. *Evolution* 1996;50:1042–1051.
- Collins FS, Ginger MS, Chaakravarti A: Variations of a theme: Cataloging human DNA sequence variation. *Science* 1997;278:1580–1581.
- Devlin B, Roeder K, Bacanu S-A: Unbiased methods for population-based association studies. *Genet Epidemiol* 2001;21:273–284.
- Ewens WJ: *Mathematical Population Genetics*. New York, Springer, 1979.
- Falconer DS, Mackay TFC: *Introduction to quantitative genetics*, ed 4. London, Longman, 1996, pp 15–19.
- Gray IC, Cambell DA, Spurr NK: Single nucleotide polymorphisms as tools in human genetics. *Hum Mol Genet* 2000;9:2403–2408.
- Graybill FA: *Theory and Application of the Linear Model*. Pacific Grove, Wadsworth & Brooks/Cole Advanced Books & Software, 1976.
- Hartl DL, Clark AG: *Principles of Population Genetics*, ed 2. Sunderland, Sinauer, 1989.
- Haseman JK, Elston RC: The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* 1972;2:3–19.
- Horikawa Y, Oda N, Cox NJ, Li X, Orho-Melander M, Hara M, Hinokio Y, Lindner TH, Mashima H, Schwarz PE, del Bosque-Plata L, Horikawa Y, Oda Y, Yoshiuchi I, Colilla S, Polonsky KS, Wei S, Concannon P, Iwasaki N, Schulze J, Baier LJ, Bogardus C, Groop L, Boerwinkle E, Hanis CL, Bell GI: Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nat Genet* 2000;26: 163–175.
- The International SNP Map Working Group: A map of human genome, 2001.
- Jiang J, Akey JM, Shi J, Yiong MM, Wang Y, Shen Y, Yu X, Chen H, Wu H, Xiao J, Lu D, Huang W, Jin L: Association of blood pressure and polymorphisms in the promoter region of catalase in a Chinese population. *Hum Genet* 2001; 109:95–98.
- Kaplan N, Morris R: Issues concerning association studies for fine mapping a susceptibility gene for a complex disease. *Genet Epidemiol* 2001; 20:432–457.
- Lynch M, Walsh B: *Genetics and Analysis of Quantitative Traits*. Sunderland, Sinauer, 1998.
- Morita H, Kurihara H, Kurihara Y, Kuwaki T, Shindo T, Oh-hashii Y, Kumada M, Yazaki Y: Responses of blood pressure and catecholamine metabolism to high salt loading in endothelin-1 knockout mice. *Hypertens Res* 1999; 22:11–16.
- Narain P: *Statistical Genetics*. New York, Wiley, 1990.
- Ohno Y, Tanase H, Nabika T, Otsuka K, Sesaki T, Suzawa T, Morri T, Yamori Y, Saruta T: Selective genotyping with epistasis can be utilized for a major quantitative trait loci mapping in hypertension in rats. *Genetics* 2000;155:785–792.
- Risch NJ: Searching for genetic determinants in the new millennium. *Nature* 2000;405:847–856.
- Risch N, Merikangas K: The future of genetic studies of complex human diseases. *Science* 1996; 273:1516–1617.
- Rosmond R, Ukkola O, Chagnon M, Bouchard C, Bjorntorp P: Polymorphisms of the beta2-adrenergic receptor gene (ADRB2) in relation to cardiovascular risk factors in men. *J Intern Med* 2000;248:239–244.
- Schork NJ: Extended multipoint identity-by-descent analysis of human quantitative traits: Efficiency, power, and modeling considerations. *Am J Hum Genet* 1993;53:1306–1319.
- Slatkin M: Disequilibrium mapping of a quantitative-trait locus in an expanding population. *Am J Hum Genet* 1999;64:1765–1773.
- Spielman RS, McGinnis RE, Ewens WJ: Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 1993;52: 506–561.
- Turelli M, Barton NH: Genetic and statistical analyses of strong selection on polygenic traits: What, me normal? *Genetics* 1994;138:913–941.
- Wang DG, Fan JB, Siao CJ, Bermo A, Yong P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, et al: Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 1998;280:1077–1082.
- Van Gestel S, Houwing-Duistermaat JJ, Adolfsson R, van Duijn C, van Broeckhoven C: Power of selective genotyping in genetic association of quantitative traits. *Behav Genet* 2000;30:141–146.
- Weiss KM, Terwilliger JD: How many diseases does it take to map a gene with SNPs. *Nat Genet* 2000;26:151–157.
- Wu CI, Palopoli MF: Genetics of postmating reproductive isolation in animals. *Annu Rev Genet* 1994;28:283–308.
- Xiong MM, Jin L, Boerwinkle E: Linkage disequilibrium based regression: A method for mapping quantitative trait loci in humans. *Am J Hum Genet* 1998;63:A45.
- Zhao J, Li W, Xiong MM: Population based linkage disequilibrium mapping of QTL: An application to simulated data in an isolated population. *Genet Epidemiol* 2001;21(suppl 1):S655–S659.