doi: 10.1046/j.1529-8817.2003.00058.x

Power of Linkage Disequilibrium Mapping to Detect a Quantitative Trait Locus (QTL) in Selected Samples of Unrelated Individuals

A. Tenesa[1,*], S. A. Knott[1], A. D. Carothers[2] and P. M. Visscher[1]

[1] *Institute of Cell, Animal and Population Biology, University of Edinburgh, Edinburgh EH9 3JT, Scotland, UK*

[2] *MRC Human Genetics Unit, Western General Hospital, Crewe Road, Edinburgh, Scotland, UK*

Summary

We considered a strategy to map quantitative trait loci (QTLs) using linkage disequilibrium (LD) when the QTL and marker locus were multiallelic. The strategy involved phenotyping a large number of unrelated individuals and genotyping only selected individuals from the two tails of the trait distribution. Power to detect trait-marker association was assessed as a function of the number of QTL and marker alleles. Two patterns of LD were used to study their influence on power. When the frequency of the QTL allele with the largest effect and that of the marker allele linked in coupling were equal, power was maximum. In this case, increasing the number of QTL alleles reduced the power. The maximum difference in power between the two LD patterns studied was ∼30%. For low QTL heritabilities ($h^2_{\mathrm{QTL}} < 0.1$) and single trait studies we recommend selecting around 5% of the upper and lower tails of the trait distribution.

Introduction

Quantitative traits are those measured on a continuous scale. They are complex because there is not a simple relationship between genotype and phenotype, and may be of interest to human geneticists because they may be easier to collect than binary disease traits and are correlated with disease status. For example, a patient with ischaemic heart disease is generally treated and controlled by his/her blood pressure or cholesterol level, but rarely directly for the heart condition.

Linkage disequilibrium (LD) is defined as the non-random association of population allele frequencies at two or more loci (Ayres & Balding, 2001), and is used at the population level to map trait loci. If a marker and trait locus are in LD, then the marker locus will be associated with the phenotype controlled by the trait locus. However, the ability to detect an association between a given allele at a marker locus and a trait depends on the amount of LD between the two loci. Although theoretically one could predict the amount of LD between two loci as a simple function of the physical distance between them (Hartl & Clark, 1997), empirical studies show that this relationship is not simple (Daly *et al.* 2001; Jeffreys *et al.* 2001). This suggests that the distribution of LD in the region of interest must be carefully studied before a statistically significant or non-significant association is reported, because the former does not always imply close linkage (e.g. significant LD can arise between non-syntenic loci) and the latter does not always imply a lack of it. Population stratification can generate significant LD between non-syntenic loci and, hence, false positives. Using family data, rather than unrelated cases and controls, overcomes the problem of population stratification because case and control samples are obtained from the same genetic background and contrasts are done within families and not across families. However, family-based designs are not always possible, especially for late onset traits in which parental data are often unavailable.

[*] Corresponding author: Albert Tenesa. Institute of Cell, Animal and Population Biology, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, Scotland, UK. E-mail: albert.tenesa@ed.ac.uk

In the absence of parental data, the use of unrelated cases and controls is an appealing alternative, provided that the possibility of population stratification can be ruled out or the effects of population structure can be eliminated. Pritchard *et al.* (2000a,b) showed that population structure can be inferred using a set of unlinked markers and individuals assigned to different subpopulations. Testing within subpopulations or taking into account the average level of association observed throughout the genome, e.g. by using a genomic control (Devlin & Roeder, 1999), would make it possible to allow for false positives due to population stratification.

Selective genotyping is the term used when individuals only from the upper and lower tail of the trait phenotypic distribution are genotyped (Lander & Botstein, 1989; Darvasi & Soller, 1992). This strategy is efficient and powerful under some circumstances (Allison *et al.* 1998) because most of the information resides in individuals with extreme phenotypes (Carey & Williamson, 1991). It is especially useful when the cost of genotyping is much greater than the cost of collecting phenotypes, and when a single phenotype is studied.

Schork *et al.* (2000) studied the power to detect a trait-marker association using individuals sampled from the upper and lower tails of the quantitative trait phenotypic distribution. Both marker and QTL were assumed to be biallelic. The aim of the present study is to investigate and predict the power of LD mapping when the QTL and marker loci are multiallelic. In particular, we studied:

1. The effect on power when the QTL is assumed to be multiallelic as opposed to biallelic.
2. Two different and simple patterns of LD, to investigate their influence on power.
3. The economically optimal proportion of the quantitative trait (QT) distribution selected for a given power, depending on the relative cost of genotyping and phenotyping.

Methods

Individuals sampled from the tails of the trait distribution are classified as upper or lower tail depending on whether their trait value is respectively greater or less than a given threshold. The study design for a practical case would be: (1) to phenotype a number of individuals for a quantitative trait; (2) to select individuals with extreme phenotypes (e.g. the 10% upper and lower values for the quantitative trait) to be genotyped; (3) to compare the counts of the different alleles at a locus in the upper and lower tails.

Genetic Model

Consider a locus with an arbitrary number of alleles that contributes to the genetic component of a quantitative trait. Alleles at the locus are labelled as Q_i. With n alleles at a locus there are $n(n+1)/2$ possible genotypes and the same number of genotypic values. The population frequency of allele Q_i is labelled q_i. The genotypic value (G_{ij}) for genotype Q_iQ_j is parameterised as:

$$G_{ij} = G_{ji} = G_{ii} + k_{ij} \times (G_{jj} - G_{ii}); \quad i < j;$$
$$i \epsilon [1, n-1]; \quad j \epsilon [2, n]; \quad k_{ij} \epsilon [0, 1] \tag{1}$$

where k_{ij} provides a measure of dominance between alleles Q_j and Q_i. If $k_{ij} = 0$, Q_i is dominant to Q_j; if $k_{ij} = 0.5$, Q_i and Q_j act additively; and if $k_{ij} = 1$, Q_i is recessive to Q_j. The difference between the genotypic value of the Q_jQ_j and Q_1Q_1 genotypes is represented as $2a_j$ (where $G_{11} = 0 = 2a_1$; $j \epsilon [2, n]$) and is expressed in residual standard deviations.

Mixture Model

Assuming there are $n(n+1)/2$ genotypes with normally distributed phenotypes, the observed joint phenotypic distribution is a weighted average of the underlying normal distributions. The probability density function for a mixture of normals is:

$$\rho(x) = \sum_{i=1}^{n} \sum_{j=1; j \geq i}^{n} f_{ij} \varphi\left(x \mid \mu_{ij}, \sigma_{ij}^2\right) \tag{2}$$

where f_{ij} is the frequency of genotype Q_iQ_j, μ_{ij} is the mean value for genotype Q_iQ_j, σ_{ij}^2 is the variance in trait values for individuals with genotype Q_iQ_j (within genotype variance) and $\varphi(x \mid \mu, \sigma^2)$ is the normal probability density function with mean μ and variance σ^2. The locus is assumed to be in Hardy-Weinberg equilibrium with frequencies q_i^2 for homozygous Q_iQ_i genotypes and $2q_iq_j$ for heterozygous Q_iQ_j genotypes. Without loss of generality, the within-genotype variance (σ_E^2) is assumed to be 1.

When the QTL effect is small, then the observed joint distribution can be approximated by a single normal distribution with mean and variance equal to:

$$\mu_{Pop} = \sum_{i=1}^{n} \sum_{j=1;j\geq i}^{n} \mu_{ij} f_{ij}; \quad \sigma_{Pop}^2 = \sigma_G^2 + \sigma_E^2$$

where σ_G^2 is the genetic variance due to the QTL and $\sigma_E^2 = 1$ as above. Although all results shown in this work were performed assuming a mixture distribution, the approximation to a normal gave practically the same results for the range of QTL effects considered.

Selecting Individuals from the Upper and Lower Tails

It is assumed that we are interested in the QTL allele that is associated with the highest genotypic value and that $2a_1 < 2a_2 < 2a_3 < \cdots < 2a_n$. This seems reasonable when carrying out selective genotyping, because selection of individuals in opposite tails will produce an enrichment of the QTL allele frequencies that cause lower or higher trait values relative to a random sample of individuals. The selected fractions in the upper and lower tails are α_U and α_L respectively, with corresponding truncation points τ_U and τ_L. The latter were obtained by solving the following non-linear equations using Newton's method as described in Ducrocq & Quaas (1988):

$$\alpha_U = \sum_{i=1}^{n} \sum_{j=1;j\geq i}^{n} f_{ij}[1 - \Phi\{(\tau_U - \mu_{ij})/\sigma_{ij}\}] \tag{3}$$

$$\alpha_L = \sum_{i=1}^{n} \sum_{j=1;j\geq i}^{n} f_{ij}[\Phi\{(\tau_L - \mu_{ij})/\sigma_{ij}\}] \tag{4}$$

where $\phi(\tau)$ is the cumulative standard normal distribution.

Using Bayes' theorem, the conditional probability of sampling a Q_i allele given that individuals have been sampled from the upper α_U percentile of the trait distribution can be written as:

$$\begin{aligned} P(Q_i \mid x > \tau_U) &= \frac{P(x > \tau_U \mid Q_i)\, P(Q_i)}{P(x > \tau_U)} \\ &= \frac{\sum_{j=1}^{n} P(x > \tau_U \mid Q_i Q_j)\, P(Q_i Q_j \mid Q_i)\, P(Q_i)}{P(x > \tau_U)} \\ &= \frac{q_i \sum_{j=1}^{n} q_j \left(1 - \Phi\left(\frac{\tau_U - \mu_{ij}}{\sigma_{ij}}\right)\right)}{\alpha_U} \end{aligned} \tag{5}$$

Equivalent probabilities can be computed for samples from the lower tail. Note that in equations (5) and (6) we are no longer assuming that $i \leq j$, as in equation (2).

$$P(Q_i \mid x \leq \tau_L) = \frac{q_i \sum_{j=1}^{n} q_j \left(\Phi\left(\frac{\tau_L - \mu_{ij}}{\sigma_{ij}}\right)\right)}{\alpha_L} \tag{6}$$

LD Between Trait and Marker Loci

In most cases genotypic information is obtained on marker loci rather than on the trait locus itself. For instance, one could genotype individuals for a number of marker loci scattered across the whole genome and test for an association between marker status at each locus and phenotype. A statistically significant association between marker status and phenotype would suggest that there is statistically significant LD between marker and trait loci at the population level. This does not always imply linkage between the loci (e.g. significant LD can be found between non-syntenic loci due to stratification, drift, etc.), but it will be assumed in what follows that close linkage is the cause of the LD.

Consider a marker locus with an arbitrary number of alleles, linked to the trait locus and in LD with it. We assume that under the null hypothesis the marker locus is in Hardy-Weinberg equilibrium. We require this because we are assuming that each of the two marker alleles that constitute the genotype is sampled independently. Under this design, one would sample alleles in pairs, i.e. the pair of alleles that form a genotype. Hence, the sampling of the two alleles could only be considered independent if the assortment of alleles at the marker locus was random, i.e. the marker locus was in Hardy-Weinberg equilibrium. The marker alleles are represented as M_h, with population frequency m_h. The disequilibrium parameter (δ_{hi}) between marker allele h and QTL allele i is defined as $\delta_{hi} = f_{hi} - m_h q_i$, where f_{hi} is the population frequency of the haplotype $M_h Q_i$. Note also that the following conditions must be fulfilled:

$$\sum_{h=1}^{m} m_h = 1 \tag{7}$$

$$\sum_{i=1}^{n} q_i = 1 \tag{8}$$

$$\sum_{h=1}^{m} \delta_{h1} = \sum_{h=1}^{m} \delta_{h2} = \sum_{h=1}^{m} \delta_{h3} = \cdots = \sum_{h=1}^{m} \delta_{hn} = 0 \tag{9}$$

$$\sum_{i=1}^{n} \delta_{1i} = \sum_{i=1}^{n} \delta_{2i} = \sum_{i=1}^{n} \delta_{3i} = \cdots = \sum_{i=1}^{n} \delta_{mi} = 0 \tag{10}$$

The probability that a haplotype from an individual sampled from the upper tail (α_U) has an allele M_h is given by:

$$P(M_h \mid x > \tau_U) = \sum_{i=1}^{n} P(M_h \mid Q_i) P(Q_i \mid x > \tau_U)$$
$$= \sum_{i=1}^{n} (m_h + \delta_{hi}/q_i) P(Q_i \mid x > \tau_U)$$

since

$$P(M_h \mid Q_i) = P(M_h Q_i)/P(Q_i)$$
$$= (m_h q_i + \delta_{hi})/q_i = m_h + \delta_{hi}/q_i$$

using $P(Q_i \mid x > \tau_U)$ from (5). This reduces to:

$$P(M_h \mid x > \tau_U) = m_h + \sum_{i=1}^{n} (\delta_{hi}/q_i) P(Q_i \mid x > \tau_U) \tag{11}$$

and similarly

$$P(M_h \mid x \leq \tau_L) = m_h + \sum_{i=1}^{n} (\delta_{hi}/q_i) P(Q_i \mid x < \tau_L) \tag{12}$$

Linkage Disequilibrium Distribution Patterns

Disequilibrium between the QTL allele with the greatest effect (Q_n) and the marker allele (M_m) is assumed to be positive ($\delta_{mn} > 0$). For convenience, we assume that this marker is the one with the highest suffix (value of m). The disequilibrium parameter is expressed as a fraction of the maximum disequilibrium possible between the two alleles ($D'_{mn} = \delta_{mn}/\delta_{mn}^{\max}$) (Lewontin, 1964) where $\delta_{mn}^{\max}$ is:

$$\delta_{mn}^{\max} =$$
$$\begin{bmatrix} \min\{m_m q_n, & (1 - m_m)(1 - q_n)\}; & \delta_{mn} < 0 \\ \min\{m_m(1 - q_n), & (1 - m_m)q_n\}; & \delta_{mn} > 0 \end{bmatrix} \tag{13}$$

In order to explore how the LD distribution affects the power to detect an association between a marker allele and trait status, two different ways of fulfilling conditions (9) and (10) were studied as examples. The disequilibrium parameter was first computed as described above for element (m, n) and represented as δ_{mn}. For the first pattern studied, elements in column n were set equal to $-\delta_{mn}/(m - 1)$ and elements in row m were all set equal to $-\delta_{mn}/\ (n - 1)$. All other elements were set equal to $\delta_{mn}/(m - 1)(n - 1)$. This is the model assumed unless otherwise stated. For the second LD pattern, the first element δ_{mn} was computed as above and an element δ_{hi} was selected to be equal to δ_{mn}. Then all elements except δ_{mi} and δ_{hn} were assumed in equilibrium (i.e. $\delta_{hi} = -\delta_{mi} = -\delta_{hn} = \delta_{mn}$).

Calculation of Power

Under the null hypothesis (H_0) of no association between a marker locus and a trait, the distributions of the marker alleles in the upper and lower tails are identical. We test this using a contingency table with m rows and 2 columns, where the entries in the h^{th} row correspond to the number of M_h alleles ($h = 1, \ldots, m$), and those in the 1st and 2nd columns correspond to the numbers of alleles in the lower and upper tails respectively. The conventional statistic X^2, based on this table, is asymptotically distributed under H_0 as chi-squared with $m - 1$ degrees of freedom. Under the general alternative hypothesis (H_1), X^2 is asymptotically distributed as non-central chi-squared with $m - 1$ degrees of freedom and non-centrality parameter, λ, given by

$$\lambda = N_L \sum_{h=1}^{m} \frac{(p_{Lh1} - p_{Lh0})^2}{p_{Lh0}} + N_U \sum_{h=1}^{m} \frac{(p_{Uh1} - p_{Uh0})^2}{p_{Uh0}} \tag{14}$$

where N_L, N_U denote the numbers of alleles sampled from the lower and upper tails respectively,

$$p_{Lh0} = Pr(M_h \mid x \leq \tau_L, H_0) \tag{15a}$$

$$p_{Uh0} = Pr(M_h \mid x > \tau_U, H_0) \tag{15b}$$

$$p_{Lh1} = Pr(M_h \mid x \leq \tau_L, H_1) \tag{15c}$$

$$p_{Uh1} = Pr(M_h \mid x > \tau_U, H_1) \tag{15d}$$

and the expressions on the right of equations (15a-d) are obtained by substituting appropriate values of δ_{hi} in equations (11) and (12) (Kendall & Stuart, 1961). Power is then defined as the probability that a non-central χ^2 with $m - 1$ degrees of freedom and non-centrality parameter λ is greater than the critical value defined by a central χ^2 with $m - 1$ degrees of freedom and significance level α.

Optimal Proportion Genotyped

The total cost depends on the numbers of individuals phenotyped (S_f) and genotyped ($S_g = (N_U + N_L)/2$), as well as the costs of phenotyping (K_f) and genotyping (K_g) per individual (Darvasi & Soller, 1992). Therefore, for a given power, the ratio S_g/S_f ($=p$, say) that minimises the cost can be determined. The total proportion selected to genotype (p) is equal to $\alpha_U + \alpha_L$. We assume in what follows that $\alpha_U = \alpha_L$ (i.e. that $p = 2\alpha_U = 2\alpha_L$). Although it may not always be optimal to set $\alpha_U = \alpha_L$, it is justified by the absence of prior knowledge concerning the model parameters. If *F(p)* denotes the total cost, then

$$F(p) = K_g S_g + K_f S_f = K_f S_g \left(K + \frac{1}{p} \right) \tag{16}$$

where

$$K = \frac{K_g}{K_f} \tag{17}$$

Note that in (16) S_g is also a function of p. The value of p that minimizes the cost function for a wide range of values of K was obtained numerically for values of p between 0.0001 and 1.

Results

Effect of the Number of Individuals Genotyped when the Number of Individuals Phenotyped is Fixed

Figure 1 shows how power increases as a larger proportion of the 2000 individuals phenotyped is genotyped until a maximum is reached (vertical dashed line in Figure 1) at 55% for markers with $m = 2$, 4, 6 and 10 alleles. The frequency of the m^{th} allele at the marker was kept constant in all situations considered; all other alleles were at equal frequencies, $m_h = (1 - m_m)/(m - 1)$. This assumption leads to the same amount of disequilibrium between M_m and Q_2, regardless of the number of marker alleles (Terwilliger, 1995).

Genotyping more than 55% of the individuals phenotyped leads to a decrease in power when using this type of test and we therefore restrict our investigations

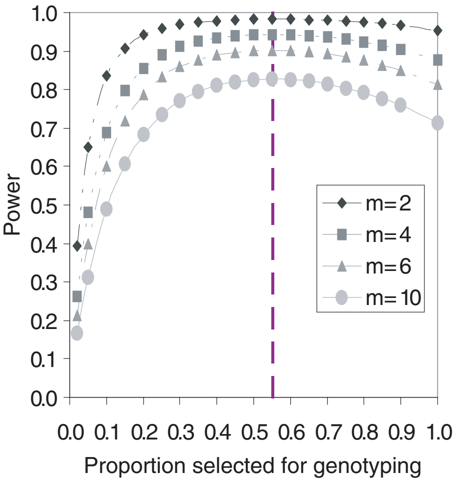

Figure 1 Effect of the proportion of individuals genotyped when the number of individuals phenotyped is fixed to 2000. Assumptions: additive model ($k_{12} = 0.5$), biallelic QTL, equal proportions of individuals selected for genotyping from the upper and lower tail, significance level (α) 0.05, $q_2 = m_m = 0.1$ and $m_h = (1 - m_m)/(m - 1)$, where m is the number of marker alleles and $h \in [1, m - 1]$, $a_2 = 0.5$, $h^2_{\text{QTL}} = 0.043$. The vertical dashed line represents the proportion selected that gives the highest power.

to moderate to high intensities of selection. This reduction in power is due to the increased amount of noise added by individuals in the middle of the quantitative trait distribution.

Effect of the Number of Marker Alleles and Proportion Selected on Power when the Number of Individuals Genotyped is Fixed

Figure 2 shows, for a biallelic QTL, how the number of marker alleles influenced power as a function of the proportion of individuals selected to be genotyped and the amount of disequilibrium. Note that in this case (unlike the previous section) the total number of phenotypes measured increases as the proportion of the QT distribution decreases. With the total number of individuals genotyped fixed at $S_g = 500$, power decreased with increasing number of alleles at the marker locus, as a result of the increase in the number of degrees of freedom for the X^2 test (d.f. $= m - 1$). Power also decreased as the proportion of the QT distribution genotyped increased. Power was similar (close to 100%) in all cases when the proportion selected was low and the amount of disequilibrium was high.

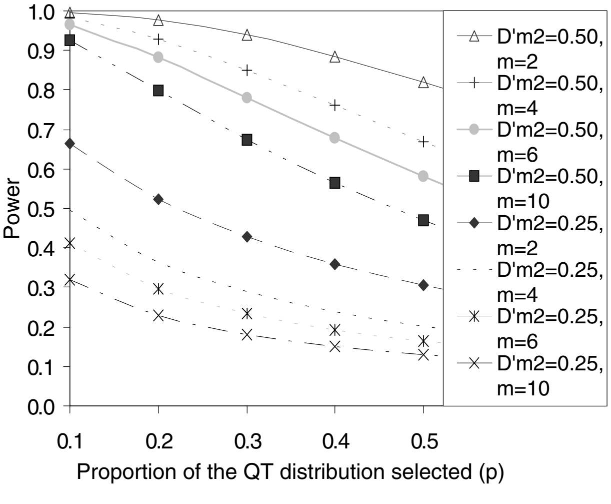

Figure 2 Effect of the proportion of individuals selected, amount of disequilibrium (D') and the number of marker alleles (m) on power. Assumptions: additive model ($k_{12} = 0.5$), biallelic QTL, equal proportions of individuals selected for genotyping from the upper and lower tail, total sample size $S_g = 500$ individuals, significance level (α) 0.05, $q_2 = m_m = 0.1$ and $m_h = (1 - m_m)/(m - 1)$, where m is the number of marker alleles and $h \epsilon [1, m - 1]$, $a_2 = 0.5$, $h^2_{\mathrm{QTL}} = 0.043$.

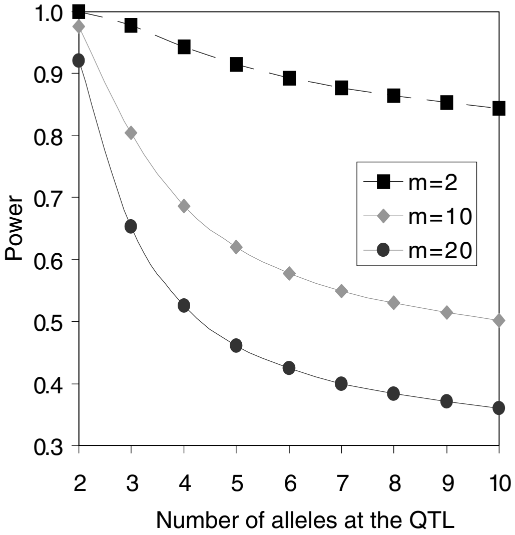

Figure 3 Effect of the number of QTL alleles on power. a_i is defined as $(i - 1)a_n/(n - 1)$, q_i is defined as $(1 - q_n)/(n - 1)$ where n is the number of QTL alleles, $i \epsilon [1, n - 1]$, $a_n = 0.5$ and $q_n = 0.2$. Marker allele frequencies were set to $m_m = 0.2$ and $m_h = (1 - m_m)/(m - 1)$ where m(=2, 10 or 20) is the number of marker alleles and $h \epsilon [1, m - 1]$. A total of 500 individuals (S_g) were selected for genotyping from the upper and lower 10% QT distribution ($N_U = N_L$). The genetic model was assumed additive ($k_{ij} = 0.5$), $D'_{mn} = 0.5$, with the significance level (α) $= 0.05$.

Effect of the Number of QTL Alleles on Power

Figure 3 shows how the number of QTL alleles influenced power for a fixed number of individuals genotyped. The number of QTL alleles assumed varied from 2 to 10, and the difference in genotypic values between the two extreme homozygotes (that is, between Q_1Q_1 and Q_nQ_n) remained constant. For the other alleles, the increase in genotypic value of Q_iQ_i with respect to $Q_{i-1}Q_{i-1}$ was equal to $2^*a_n/(n - 1)$ for $i \epsilon [2, n]$. Note that there are infinite combinations of genotypic values that would lead to the same QTL heritability for fixed allele frequencies when the QTL is multiallelic. Results expressed in this way (as the difference between the two more extreme homozygous genotypes) have greater generality than a sample of all the possible genotypic combinations with the same QTL heritability. The marker locus was assumed to have 2, 10 or 20 alleles. In all cases considered, when the number of alleles at the QTL increased the power decreased.

This reduction in power was larger with a higher number of marker alleles. Note that for the 20-allele marker, almost half of the reduction in power occurred when the number of QTL alleles increased from 2 to 3.

The QTL heritability under the conditions assumed in Figure 3 varied with the number of alleles at the QTL. It showed a slight decrease with the increase in the number of alleles ($h^2 = 0.074$ for a biallelic QTL and $h^2 = 0.054$ for a 10-allele QTL). In order to check whether the difference in power was due to the increasing number of alleles or to the reduction in the locus heritability, the case of a QTL locus with two versus 3-10 alleles (keeping the same heritability as for the biallelic locus) was studied. Figure 4 shows how power varied with heritability. The continuous line shows how power varied with heritability when the QTL was assumed biallelic. For a given heritability, the individual dots represent the power obtained for a QTL with different number of alleles. In all cases the marker was assumed biallelic, with $m_2 = q_n = 0.2$. The reduction in power with heritability was much larger when accompanied by an increase in the number of alleles at the QTL, showing that it was mainly due to the increase in the number of QTL alleles rather than to the reduction in heritability.

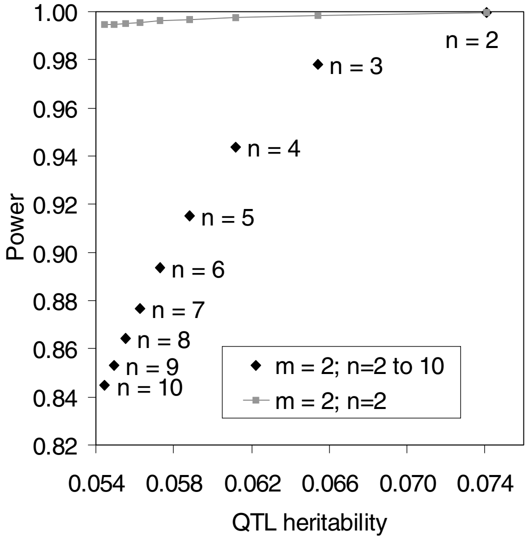

Figure 4 Effect of the number of QTL alleles on power. Comparison between a QTL with 2 alleles and a QTL with n alleles when the locus has the same heritability. a_i is defined as $(i-1)a_n/(n-1)$, q_i is defined as $(1-q_n)/(n-1)$ where n is the number of alleles of the QTL, $i\epsilon[1, n-1]$, $a_n = 0.5$ and $q_n = 0.2$. Marker was assumed biallelic and allele frequency was set to $m_2 = 0.2$. A total of 500 individuals (S_g) was selected for genotyping from the upper and lower 10% QT distribution ($N_U = N_L$). The genetic model was assumed additive ($k_{ij} = 0.5$), $D'_{2n} = 0.5$, with the significance level (α) = 0.05.

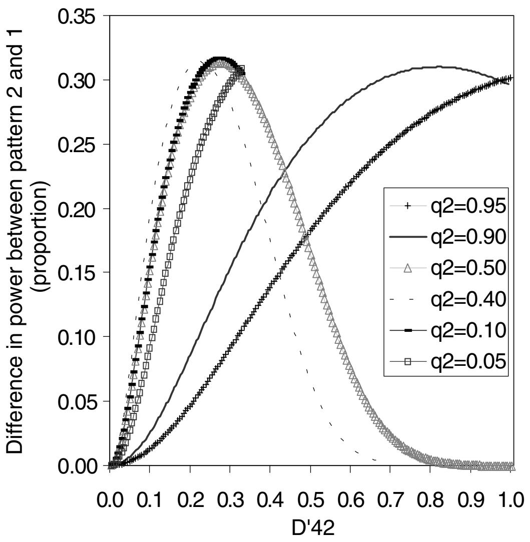

Figure 5 Difference in power between patterns of LD 1 and 2 as a function of the amount of LD. The difference in power is expressed as a proportion of the power obtained for pattern 2 ($\delta_{11} = -\delta_{41} = -\delta_{12} = \delta_{42}$).). A total of 2000 individuals were selected for genotyping (S_g) from the upper and lower 10% QT distribution ($N_U = N_L$). A biallelic QTL was assumed with locus $h^2_{\mathrm{QTL}} = 0.02$. The genetic model was additive ($k_{12} = 0.5$), significance level (α) = 0.05. Marker locus assumed to have four equally frequent alleles.

Difference in Power Between Patterns of LD

Figure 5 shows the difference in power between LD patterns 1 and 2 for a biallelic QTL and a marker with 4 alleles, as a proportion of the power obtained with pattern 2 (the more powerful of the two). In this case both LD patterns had an equal total amount of disequilibrium as measured by Hedrick's D' (Hedrick, 1987). The maximum difference between patterns was about 30%, regardless of the QTL frequency. Differences in power increased with D'_{42} if q_2 was high or low, but if q_2 was intermediate differences in power were maximum when D'_{42} had values that were intermediate for the range of possible values given the allele frequencies (q_2 and m_4).

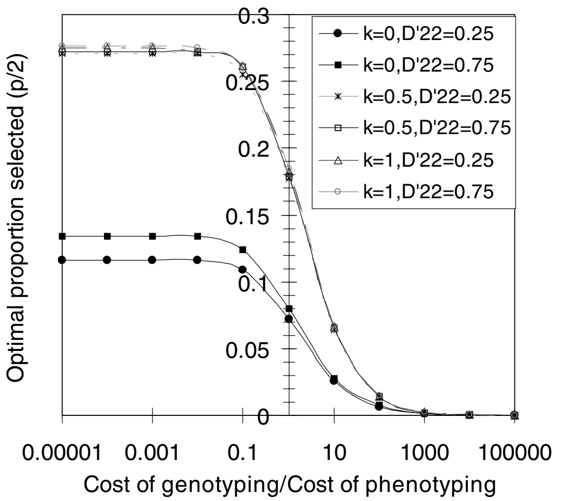

Figure 6 Optimum selection proportion for a power of 80% as a function of the relative costs of genotyping and phenotyping. Different genetic models [recessive ($k_{12} = 0$), additive ($k_{12} = 0.5$) and dominant ($k_{12} = 1$)] and amount of disequilibrium (D'_{22}) were assumed. The same proportions and the same number of individuals were selected for genotyping from the upper and lower tails ($p/2 = \alpha_L = \alpha_U$). The QTL and the marker were assumed biallelic ($q_2 = m_2 = 0.2$), locus $h^2_{\mathrm{QTL}} = 0.02$, significance level ($\alpha$) = 0.05. The horizontal axis is on the logarithmic scale.

Optimum Selected Proportion

Figure 6 shows how the relative cost of genotyping and phenotyping influenced the proportion of individuals selected to be genotyped in order to achieve cost-effectiveness. Two levels of LD and three genetic models were studied for a biallelic QTL and a biallelic marker. The power was fixed at 80% and the QTL heritability was kept constant for all the models. Figure 6 illustrates, how for the cases studied, it would not be worthwhile to genotype more than the upper and lower 27.5% of the individuals phenotyped if the genetic model were additive or dominant, and 12.5% of the individuals when the genetic model were recessive, even when the cost of genotyping these individuals was 100-100000 times less than that of phenotyping ($K < 0.01$). This is because most of the information comes from individuals with extreme phenotypes, so that genotyping less informative individuals produces no increase in power. For example, 80% power for the parameters shown in Figure 6 and D'_{22} equal to 0.75 could be obtained by genotyping and phenotyping 1100 individuals ($p = 1$) or phenotyping 872 individuals and genotyping the upper and lower 218 individuals ($p = 0.5$).

The optimal proportion selected was always largest for the dominant model and smallest for the recessive model when the favourable allele (Q_2) was the less frequent one. This suggests that, for the recessive model and when the frequency of the Q_2 allele is small, the most extreme individuals of the trait distribution must be genotyped in order to increase the frequency of Q_2 alleles in the upper tail. By doing so, the relative frequencies of the individuals Q_1Q_1 and Q_1Q_2 with a positive deviation from their genotypic mean are reduced and the relative frequency of Q_2Q_2 with positive deviations are increased in the upper tail. The level of LD affects the optimal proportion selected. Increasing amounts of LD produced an increase in the optimal proportion selected for all the models of inheritance, and this increase was much more apparent in the recessive model than in the others.

When the less frequent allele was dominant or additive then the optimal proportions selected were relatively insensitive to variations in heritability. For example, for an additive model and $D'_{22} = 0.5$ with h^2_{QTL} values ranging from 0.01 to 0.1, the optimal proportion of individuals to be genotyped varied from $p = 0.538$ to $p = 0.566$ for $K < 0.01$ (results not shown). When the less frequent allele was recessive, then the optimal proportion selected decreased with increasing heritability (results not shown).

Discussion

Quantitative genetics theory is commonly applied under the simplified assumption that loci are biallelic. In this study, power to detect an association between a marker and a trait has been explored and quantified for multiallelic QTL and markers. Although others have previously noted that there may be loss of insight when the assumption that loci are biallelic is made (Nielsen & Weir, 1999), to our knowledge this has not been quantified. We restrict our conclusions to moderate-high intensities of selection because, when selecting individuals from the upper and lower tail of the trait distribution, we are effectively dichotomizing the quantitative trait and therefore ignoring the information within each tail. This loss of information decreases as the selection intensity increases. We have quantified the loss of information in a separate study.

Results shown here are based on the assumption that asymptotic conditions hold, i.e. that sample sizes are sufficiently large. Spurious results can arise if the sample size and/or some of the marker allele frequencies are small. However, we have found that relatively large sample sizes are necessary to obtain reasonable power, and these are expected to be large enough for asymptotic assumptions to hold. Significance thresholds used in this study are insufficient for a whole genome scan, which would require greater stringency. However, this is merely a scaling factor that does not change our general conclusions.

We have shown that, for a given QTL heritability, there is a large difference in power depending on the number of QTL alleles, with the power decreasing with increasing numbers of alleles. This is important because it is usually assumed that the QTL is biallelic, whereas a number of empirical studies have shown that disease loci may have multiple alleles (Hugot *et al.* 2001; Ogura *et al.* 2001; Wright & Hastie, 2001). Therefore, calculations performed assuming a biallelic QTL can seriously overestimate the power.

Two patterns of LD were investigated. Although these were just examples and did not correspond to any particular population genetics model, they illustrate the differences in power that can be seen as a result of the pattern of LD rather than of the amount of disequilibrium as measured by D' (which was identical for the two patterns studied in Figure 5). LD patterns would probably differ from one population to another (and from one pair of markers to another) and depend on the population history. The present approach would not be more or less general than one assuming a given population genetics model.

Bader *et al.* (2001) obtained the optimal proportions selected for DNA pooling when the objective was to minimise the number of individuals to be phenotyped. Their results were similar to ours for the lowest cost ratios (cost genotyping/cost phenotyping). If the cost ratio approximates zero, then what is basically minimised is the amount of phenotyping required.

We have shown that the optimal proportion of individuals selected to be genotyped decreases to very small values under some circumstances. This is more striking for the most realistic relative costs of genotyping and phenotyping (that is, $K > 1$). As discussed by Lander & Botstein (1989) it is probably unwise to select less than the 5% tails of the trait distribution, because very extreme phenotypes can be the result of inaccurate observation (outliers). For the recessive model, the optimum proportion to select (p) was always lower than this suggested threshold (i.e. $p = 0.1$ in Figure 6) for a locus with $h^2_{\mathrm{QTL}} = 0.02$ when genotyping was 10 times more expensive than phenotyping. For the additive and dominant models the genotyping costs could be up to 10 to 50 times greater than the phenotyping costs for the optimum proportion to be greater than the suggested threshold ($p = 0.1$). Therefore the most cost-effective proportion of individuals genotyped and phenotyped obtained from our study for small QTL heritabilities and realistic cost ratios should be used with caution if the amount of phenotyping carried out is not large. For practical purposes, we recommend selecting about the 5% of both tails of the quantitative trait distribution, which correspond to reasonable genotyping/phenotyping cost ratios for most quantitative traits.

Acknowledgments

We thank Alan Wright and two anonymous referees for useful comments. We are grateful for support from the Biotechnology and Biological Sciences Research Council and the Royal Society. Albert Tenesa thanks Pau Navarro for fruitful discussions.

References

Allison, D.B., Moonseong, H., Schork, N.J., Wong, S.L. & Elston, R.C. (1998) Extreme selection strategies in gene mapping studies of oligogenic quantitative traits do not always increase power. *Hum Hered* **48**, 97–107.

Ayres, K.L. & Balding, D.J. (2001) Measuring gametic disequilibrium from multilocus data. *Genetics* **157**, 413–423.

Bader, J.S., Bansal, A. & Sham, P. (2001) Efficient SNP-based tests of association for quantitative phenotypes using pooled DNA. *Genescreen* **1**,143–150.

Carey, G. & Williamson, J. (1991) Linkage analysis of quantitative traits - increased power by using selected samples. *Am J Hum Genet* **49**, 786–796.

Daly, M.J., Rioux, J.D., Schaffner, S.E., Hudson, T.J. & Lander, E.S. (2001) High-resolution haplotype structure in the human genome. *Nat Genet* **29**, 229–232.

Darvasi, A. & Soller, M. (1992) Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. *Theor Appl Genet* **85**, 353–359.

Devlin, B. & Roeder, K. (1999) Genomic control for association studies. *Am J Hum Genet* **65**, 437.

Ducrocq, V. & Quaas, R.L. (1988) Prediction of genetic response to truncation selection across generations. *J Dairy Sci* **71**, 2543–2553.

Hartl, D.L. & Clark, A.G., (1997) Principles of Population Genetics. Sinauer, 3th edition, Massachusetts.

Hedrick, P.W. (1987) Gametic disequilibrium measures- Proceed with caution. *Genetics* **117**, 331–341.

Hugot, J.P., Chamaillard, M., Zouali, H., Lesage, S., Cezard, J.P., Belaiche, J., Almer, S., Tysk, C., O'morain, C.A., Gassull, M., Binder, V., Finkel, Y., Cortot, A., Modigliani, R., Laurent-Puig, P., Gower-Rousseau, C., Macry, J., Colombel, J.F., Sahbatou, M. & Thomas, G. (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599–603.

Jeffreys, A.J., Kauppi, L. & Neumann, R. (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet* **29**, 217–222.

Kendall, M.G. & Stuart, A. (1961) The Advanced Theory of Statistics. Volume 2. Inference and Relationship. Charles Griffin and Company Limited, 4th edition, Great Britain.

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

Lewontin, R.C. (1964) The interaction of selection and linkage. I. General considerations; Heterotic models. *Genetics* **49**, 49–67.

Nielsen, D.M. & Weir, B.S. (1999) A classical setting for associations between markers and loci affecting quantitative traits. *Genet Res* **74**, 271–277.

Ogura, Y., Bonen, D.K., Inohara, N., Nicolae, D.L., Chen, F.F., Ramos, R., Britton, H., Moran, T., Karaliuskas, R., Duerr, R.H., Achkar, J.P., Brant, S.R., Bayless, T.M., Kirschner, B.S., Hanauer, S.B., Nunez, G. & Cho, J.H. (2001) A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* **411**, 603–606.

Pritchard, J.K., Stephens, M. & Donnelly, P. (2000a) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959.

Pritchard, J.K., Stephens, M., Rosenberg, N.A. & Donnelly, P. (2000b) Association mapping in structured populations. *Am J Hum Genet* **67**, 170–181.

Schork, N. J., Nath, S. K., Fallin, D., & Chakravarti, A. (2000) Linkage disequilibrium analysis of biallelic DNA markers, human quantitative trait loci, and threshold-defined case and control subjects. *Am J Hum Genet* **67**, 1208–1218.

Terwilliger, J.D. (1995) A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *Am J Hum Genet* **56**, 777–787.

Wright, A.F. & Hastie, N.D. (2001) Complex genetic diseases: controversy over the Croesus code. *Genome Biol* **2**, 2007.1–2007.8.

Received: 6 November 2002
Accepted: 11 July 2003