SHORT REPORT

The power of genome-wide sib pair linkage scans for quantitative trait loci using the new Haseman–Elston regression method

P. C. Sham and J. H. Zhao

Institute of Psychiatry, De Crespigny Park, London SE5 8AF, U.K.

Abstract

Power calculations for linkage analysis are typically conducted on the assumption of a single locus that affects the trait. Here we report a simple procedure for conducting a power analysis for a genome-wide linkage scan of a quantitative trait under the influence of multiple loci. This procedure is designed for sib pair data analysed by the new Haseman–Elston regression method. The results show that samples as large as 10 000 sib pairs will often not allow quantitative trait loci (QTLs) to be clearly identified. Instead, linkage genome scan using sib pairs must be regarded as a blunt screening tool that will help to focus attention to 10%, or more, of the genome.

Keywords genome scan, linkage, QTL, Haseman–Elston, sib pairs, power.

Introduction

Linkage-based genetic mapping, hugely successful for single-gene Mendelian diseases, is now being applied to common disorders and quantitative phenotypes. However, genome-wide linkage scans conducted on common disorders have so far produced inconsistent patterns of results. Typically, initial linkage findings are only moderately strong, and have been replicated in only a proportion of subsequent studies. Theoretical calculations have shown that the power of linkage analysis decreases rapidly with decreasing effect size.[1] If the effect size for a quantitative trait is defined as half the difference in trait means between the two homozygous genotypes,[2] then halving the effect size will increase the sample size necessary for a certain power level by a factor of $2^2 = 4$ in an association analysis, but $2^4 = 16$ in a linkage analysis.[3] However, although the power to detect an individual quantitative trait locus (QTL) is low, complex traits are typically under the influence of multiple QTLs, and the power of detecting at least some QTLs may be much greater.[4] Here, we report simulation studies of the performance of genome-wide linkage scans for QTLs under a range of underlying multilocus models.

Genetic model

We assume that the quantitative trait has sib correlation r (assumed to be 0.25) and narrow-sense heritability h^2 (assumed to be 0.4). Under random mating, the maximum value of h^2 is $2r$; with any discrepancy being due to shared family environment, dominance, epistasis, or gene–environment interactions. The heritability can be partitioned into the contributions from individual quantitative trait loci (QTL). We consider a number of different scenarios (shown in Table 1) of how h^2 is divided among the QTLs. The QTLs are assumed to be located at random throughout the genome.

New Haseman–Elston regression

The classical Haseman–Elston method of QTL linkage analysis regresses the squared trait difference $(y_1 - y_2)^2$ onto the estimated proportion of alleles identical by descent (IBD), π, at the test locus. The new Haseman–Elston method[5] is the same as the classical method, except that the squared trait difference is replaced by the cross-product of mean-adjusted trait values $(y_1 - \mu)(y_2 - \mu)$, denoted by w. If the test locus is

Correspondence to: P. C. Sham, Institute of Psychiatry, De Crespigny Park, London SE5 8AF, U.K.
E-mail: spjupcs@iop.kcl.ac.uk

Table 1 Underlying genetic models

1. Four QTLs each accounting for 10% of phenotypic variance
2. Eight QTLs each accounting for 5% of phenotypic variance
3. Sixteen QTLs each accounting for 2.5% of phenotypic variance
4. Thirty-two QTLs each accounting for 1.25% of the phenotypic variance
5. A geometric progression in effect size, with the largest contribution being 10% and the smallest contribution being negligibly over 0%, assuming that the effect of each successive QTL decreases by a factor 0.75. Only the first 9 QTLs will have a heritability contribution of over 1%. We will simulate only these 9 QTLs.
6. A geometric progression in effect size, with the largest contribution being 7.5% and the smallest contribution being negligibly over 0%, assuming that the effect of each successive QTL decreases by a factor 0.8125. Only the first 10 QTLs will have a heritability contribution of over 1%. We will simulate only these 10 QTLs.
7. A geometric progression in effect size, with the largest contribution being 5% and the smallest contribution being negligibly over 0%, assuming that the effect of each successive QTL decreases by a factor 0.875. Only the first 13 QTLs will have a heritability contribution of over 1%. We will simulate only these 13 QTLs.

at recombination fraction θ from a QTL that contributes an additive variance of V_A to the trait, then the regression coefficient is $(1 - 2\theta)^2 V_A$. It is convenient to standardize the trait y to have mean 0 and variance 1, so that the variance of the cross-product w is $1 + r^2$. It is also convenient to subtract 0.5 from π, and to assume that linkage information is complete, in which case the variance of π is 1/8. The covariance between w and π is $(1 - 2\theta)^2 V_A/8$.

Behaviour of regression coefficient across a chromosome

In a genome scan using the new Haseman–Elston method, a regression coefficient is calculated at regular intervals across all 23 chromosomes. At each point, the estimated regression coefficient consists of a systematic part due to the influence of any neighbouring QTLs, and a random part which is due to sampling fluctuations. Let the estimated regression coefficient at position i be b_i, then we can write $b_i = s_i + r_i$, where s_i and r_i represent the systematic and random components, respectively. The systematic part is determined by the sum of the influences of all the QTLs on the same chromosome, i.e.

$$s_i = \sum_j (1 - 2\theta_{ij})^2 V_{Aj}$$

where the summation is over all QTLs on the same chromosome, θ_{ij} is the recombination fraction between position i and QTL j, and V_{aj} is the additive variance due to QTL j.

The random part at position i has mean 0; its variance is equal to the variance of the regression coefficient b_i, which is

$$v_i = \frac{\mathrm{Var}(w) - s_i^2 \mathrm{Var}(\pi)}{n \mathrm{Var}(\pi)}$$

$$= \frac{8(1 + r^2) - s_i^2}{n}$$

where n is the sample size measured by the number of sib-pairs and s_i is the magnitude of the systematic part as defined above. The correlation between the random parts of the regression coefficient of two positions separated by recombination fraction θ is $(1 - 2\theta)^\theta$.

Simulation of regression coefficients in a genome scan

The simulation of regression coefficient estimates across an arbitrary number of positions on the genome consists of the following steps:

1 Generate QTL positions.

2 Calculate systematic components for all genome positions, denoted s_i.

3 Calculate variance of random components for all genome positions, denoted v_i.

4 Generate an independent normal (0,1) deviate for every position, denoted z_i.

5 For the first position of a chromosome, the random part is given by

$$R = \sqrt{v_1}\, z_1$$

6 For other positions of a chromosome, the random part is given by

$$R_i = cR_{i-1} + dz_i$$

$$c^2 = (1 - 2\theta)^2 \sqrt{\frac{v_i}{v_{i-1}}}$$

$$d^2 = v_i - (1 - 2\theta)^2 \sqrt{v_i v_{i-1}}$$

where θ is the recombination fraction between positions $i - 1$ and i.

Calculation of lod score equivalent

Under the null hypothesis of no QTL near a position, the regression coefficient estimate b_i has mean 0 and variance v_i. If $b_i < 0$, then the lod score equivalent is 0, otherwise it is

$$\begin{aligned}\text{lod} &= \frac{b_i^2}{2\ln(10)\text{v}_i}\\ &= \frac{nb_i^2}{16(1+r^2)\ln(10)}\end{aligned}$$

Assessment of genome scan strategies

For QTL mapping, the ultimate aim for a genome-wide linkage scan is to rank the different positions of the genome in increasing priority for further studies. This ranking can be done according to the regression coefficients or to lod scores. To evaluate the ‘QTL yield’ in a certain percentile of lod scores under a set of assumptions, we simulated the regression coefficients for multiple genome scans under those assumptions. For each genome scan, we sort the genome positions according to the size of the lod score at that position. We then note the percentile into which each QTL falls (the QTLs being arranged in order of decreasing variance).

Multi-stage genome scans

We wish to investigate the properties of a two-stage genome scan, in which stage one has $n = 6000$ sib pairs and stage two has $n = 4000$ sib pairs. The simulation of each genome scan consists of the following steps.

1 Generate two genome scans according to the sample sizes of the two stages. These will yield two sets of systematic components and two sets of random components.

2 Using the systematic and random components of the first genome scan, calculate regression coefficients, sort the genome regions according to these coefficients and identify the regions where the coefficients are in the top 25 percentiles.

3 For the regions selected in step 2, calculate the regression coefficients using the systematic and random components of genome scans 1 and 2, and sort these regions according to their regression coefficients.

4 After 1000 simulated genome scans, we obtain a joint distribution of percentiles for the QTLs (if a QTL does not reach stage two, then score its percentile as being >25). We then calculate the average number of QTLs in the top 1 percentile, the top 2 percentiles, the top 3 percentiles, and so on. We also calculated the proportion of simulated genome scans that would fail to detect a single real QTL, if a certain percentile lod score were used to define positive regions.

Results

For the seven scenarios (shown in Table 1), the mean ‘QTL yield’ increases steadily with increasing percentile of lod score (see Table 2). This confirms the relatively low resolving power for linkage analysis for QTLs of minor or modest effects. Alarmingly, there is a high risk of not detecting any real QTL, if only a strict criterion (e.g. top 1 percentile) is used to define positive linkage regions. Even when a lenient criterion (e.g. top 10 percentiles) is used, there remains under some scenarios an almost 10% chance of missing any real QTL (see Table 3). Indeed, a QTL that accounts for as much as 10% of the phenotypic variance will on average be on the 7th percentile in the lod score, and is therefore easily missed.

Table 2 Mean ‘QTL yields’ by percentile lod score in 1000 simulated genome scans

Percentile	Genetic model						
lod	1	2	3	4	5	6	7
1	1.6	1.1	0.6	0.5	1.2	1.0	0.7
2	2.1	1.4	0.9	0.7	1.6	1.3	0.9
3	2.4	1.8	1.2	1.0	1.8	1.5	1.1
4	2.7	2.0	1.4	1.2	2.0	1.7	1.3
5	2.9	2.2	1.6	1.4	2.1	1.8	1.5
6	3.0	2.4	1.8	1.6	2.3	2.0	1.6
7	3.1	2.5	2.0	1.8	2.4	2.1	1.8
8	3.2	2.6	2.1	2.0	2.5	2.2	1.9
9	3.2	2.8	2.3	2.2	2.6	2.3	2.0
10	3.3	2.9	2.4	2.4	2.7	2.5	2.1

Table 3 Number of simulated genome scans (out of 1000) that yielded no QTL, by percentile lod score

Percentile	Genetic model						
lod	1	2	3	4	5	6	7
1	11	190	536	640	69	230	436
5	1	51	192	261	27	77	190
10	0	21	88	92	9	33	90

Discussion

We have presented a method that allows us to rapidly evaluate the likely outcomes of a genome scan for a quantitative trait determined by multiple QTLs of minor or modest effect. The results indicate that linkage analysis using a realistic number of sib pairs will often not allow QTLs to be clearly identified. Instead, QTLs will often have regression coefficient estimates that do not stand out from other positions. A strategy that considers only the one or two small narrow bands in the genome with the largest lod scores is likely to miss many QTLs of small or modest effect. Instead, linkage genome scan using sib pairs must be regarded as a blunt screening tool that will only help to prioritize the different regions of the genome for further studies.

Acknowledgments

This work was supported by the UK Medical Research Council grant G9700821, Wellcome Trust Grant 055379, and US National Institutes of Health grant EY-12562.

References

1 Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996; **273**: 1516–1517.

2 Falconer DS, Mackay TFC. *Introduction to Quantitative Genetics* 4th edn. Longman Group, Harlow, UK, 1996.

3 Sham PC, Cherny SS, Purcell S, Hewitt JK. Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *Am J Human Genet* 2000; **66**: 1616–1630.

4 Suarez BK, Hampe CL, Van Eerdewegh P. Problems of replicating linkage claims in psychiatry. In: *Genetic Approaches to Mental Disorders* (eds ES Gershon, CR Cloninger.) American Psychiatric Press Inc, Washington DC, pp 23–46, 1994.

5 Elston RC, Buxbaurn S, Jacobs KB, Olson JM. Haseman and Elston revisited. *Genetic Epidemiol* 2000; **19**: 1–17.