Experimental designs for reliable detection of linkage disequilibrium in unstructured random population association studies.

Roderick D. Ball, Forest Research.

Draft: November 18, 2003

Running head: Reliable detection of linkage disequilibrium

Corresponding author:

Dr Roderick D. Ball, New Zealand Forest Research Institute,

P.B. 3020, Rotorua, New Zealand.

Telephone: 64-7-347-5899

Facsimilie: 64-7-347-9380

Email : *rod.ball@forestresearch.co.nz*

ABSTRACT

A method is given for design of experiments to detect associations (linkage disequilibrium) in a random population between a marker and a quantitative trait locus (QTL), or gene, with a given strength of evidence, as defined by the Bayes factor. Using a version of the Bayes factor which can be linked to the value of an F-statistic together with an existing deterministic power calculation, makes it possible to rapidly evaluate a comprehensive range of scenarios, demonstrating the feasibility, or otherwise, of detecting genes of small effect. The Bayes factor is advocated for use in determining optimal strategies for selecting candidate genes for further testing or applications. The prospects for fine scale mapping of QTL are re-evaluated in this framework. We show that large sample sizes are needed to detect small effect genes with a respectable sized Bayes factor, and to have good power to detect a QTL allele at low frequency it is necessary to have a marker with similar allele frequency near the gene.

KEYWORDS : Linkage disequilibrium, Association tests, QTL, QTN, Genome scan, Candidate genes, Experimental design, Bayes factor.

The advent of dense maps of single nucleotide polymorphisms (SNPs) covering the genome with 300,000 or more markers offers new opportunities find and identify genes, by testing for population level associations between the SNP and disease or other trait of interest. Associations occur because of linkage disequilibrium between the marker and trait. 'Linkage disequilibrium (LD) mapping' aims to detect and locate genes relative to a map of existing genetic markers. Location information is obtained because the distance between the gene and a marker on a chromosome is one factor influencing the closeness of association between the gene and marker. In a population, recombinations affecting the association between a gene and marker may occur over many generations. This potentially gives a much finer resolution than pedigrees used for quantitative trait loci (QTL) mapping. Finer resolution comes at a cost, however. More genotyping is needed per individual and as we shall show, larger sample sizes are needed to take advantage of a finer level of resolution. In this paper we develop experimental design techniques necessary to find the sample size needed reliably detect a given level of linkage disequilibrium between a bi-allelic QTL and marker.

This paper is structured as follows. First we review prospects for LD mapping with SNP markers, possible strategies and results to date, including problems with reliability of detected QTL, illustrating the need for a sound measure of statistical evidence, and experimental design. Then we discuss measures of statistical evidence, or criteria for 'detecting' associations. We argue that there are problems with commonly used p-values as a measure of evidence and advocate the *Bayes Factor* (see e.g. Spiegelhalter and Smith 1982) as a replacement measure. Then, we review and correct the deterministic power calculation from Luo (1998), and the classical approach to power of experiments, then introduce a generic Bayes factor (Spiegel-

HALTER and SMITH 1982) for comparing linear models, in the absence of prior information. This is linked to the classical power calculation, to give designs with a given probability to detect an effect with a given Bayes factor. Results are presented for sample sizes ranging from 600 to 4800 and Bayes factors ranging from 1/20 to 20, and for a range of QTL and marker allele frequencies ranging from 0.01 to 0.5 for sample sizes ranging from 600 to 38400 with a Bayes factor of 20. Applications to genome scans and testing large sets of candidate genes are discussed, with results for posterior odds ranging from 1/20 to 20 (Bayes factors ranging from 250 to 1 000 000). The discussion section gives wider applications and implications of the method for strategies for gene discovery.

KRUGLYAK (1999), reviewing prospects for whole genome linkage disequilibrium mapping, suggests that the useful range of linkage disequilibrium in the human population is around $3kb$, corresponding to a map with 500,000 SNP markers. The high density of SNP markers, combined with the short range of linkage disequilibrium in a population means that it is possible, in principle, to locate SNPs near the genes and hence to sequence regions containing genes. Results in the literature on the range of useable linkage disequilibrium are variable. DUNNING *et al* (2000) report a similarly low range of disequilibrium in humans, while TAILLON-MILLER *et al* (reviewed in BOEHNKE (2000)) report strong LD at distances as high as 1Mb. At the other extreme LD extending over a wide range has been reported in domesticated species by e.g. DUNNER *et al* (1997), and FARNIR *et al* (2000) for cattle and by MCRAE *et al* (2002) for sheep, who reported LD even between unlinked markers.

The relatively short range of linkage disequilibrium in some populations makes it possible, in principle, to study complex diseases with non-Mendelian inheritance (i.e. disease is not caused by a single gene, but there are a number of genes contributing to disease susceptibility). If other populations with LD extending over a somewhat greater range can be found this reduces the amount of marker genotyping to more affordable levels, at a cost of lower potential resolution. See RISCH (2000) for a discussion of the issues and optimal 'post-genome' strategies.

Similar considerations apply to quantitative traits. Colleagues in the Forest Research Cell Wall Biotechnology Centre are interested in economic traits for forest trees, such as growth rate or wood density. Results from QTL mapping studies (WILCOX et al 1997; KUMAR et al 2000; BALL 2001) suggest that these traits may be influenced by a number of smaller effect genes.

ROSES (2000) discusses a strategy for determining an individual's response to medicine(s), based on selecting a subset of SNP markers associated with the responses to medicine(s) of interest, then using the values of these SNPs for an individual to predict the individual's response.

There are two general strategies for detecting associations: testing for random associations in a population or using family based tests such as the transmission disequilibrium test (ALLISON 1997; LONG and LANGLEY 1999; SPIELMAN and EWENS, 1998; BOEHNKE and LANGEFELD 1998). See

Nielsen and Zaykin (2001) for a review.

The random population association test has the disadvantage of being confounded with allele frequency differences in subpopulations, if there is any population substructure, while the transmission disequilibrium test requires availability of many small families.

Note: Where there may be population substructure we recommend the method of Pritchard et al (2000a, 2000b) for estimating and allowing for population substructure. This requires additional markers and may reduce power due to the additional number of parameters to be estimated.

With either of these strategies there are currently two common approaches: testing for single marker associations or testing haplotypes based on a combinations of marker values associated with the disease or quantitative trait. In this paper, we consider the single marker approach, with data from association studies, but note the method could be applied to haplotype data where two classes of haplotypes are considered.

The large number of possible SNPs, and the even larger number of possible haplotype combinations makes it essential to have soundly based statistical evidence for putative associations. This may not be the case in general in published literature: Altshuler *et al* (2000), discussed in Gura (2001, p594):

> " ...a team lead by Altshuler tried a slightly different approach. The invesigators went back to published studies linking other SNPs to diabetes ...and retested 16 reported SNPs in a new

> group. . . 13 SNPs were common enough in the population to study . . . only one SNP held up in the target populations . . . the SNP association had been reported in 1998, but four out of five subsequent analyses with only a few hundred patients each could not confirm the linkage."

These results are summed up by Altshuler (Hampton 2000):

> "The lack of replication of the others points to the need for larger samples, controls for population differences, *and stronger statistical evidence prior to claiming an association*." (emphasis added)

Terwilliger and Weiss (1998, Figure 4) show the distribution of around 260 reported p-values from association studies in two journals, and note that there is no evidence of departure from the uniform distribution (i.e. no evidence of any real effect):

> ". . . investigators are too frequently gambling on and publishing results in situations where the evidence is not at all compelling."

These results point to lack of *reliability* of published results. Clearly, the statistical criteria used have not lead to *reliability* of 'detected' SNPs in these examples. Experience, and the results below, suggest this problem is likely to be widespread. The objective of this paper is to design experiments for testing associations with the required quantifyably stronger statistical evidence.

Terwilliger and Weiss (1998) suggest as a cause of the problem, that the simple model underlying the association tests does not reflect the genetic complexity, citing allelic and non-allelic heterogeneity, genetics by environment interactions, intra gene interactions (epistasis) etc. We suggest that, while this may be true, resulting in complex statistical interactions, there is no hope of detecting the interactions if one cannot first detect and isolate the main effects, which would imply a model similar to that underlying association tests. We suggest the problem is not with the model but with the basic statistical experimental design (insufficient power) and statistical criteria for detection (p-values).

Evidence for a real effect: p-value and posterior probability for H_1. We argue that the p-value, the commonly used measure of statistical significance, is not a sound measure of evidence for a real effect, and that the solution of the problems is to adopt a sound framework for statistical inference, based on probability theory, first given in Bayes (1763). The p-value measures only the proportion of times a more extreme value of the test statistic result is obtained among repeated experiments assuming the null hypothesis of no effect holds. The p-value is not the same as the *reliability*, which is the proportion of times the 'detected' effects are real.

A low p-value shows that the *null hypothesis* (H_0) (e.g. a model with no linkage disequilibrium or no QTL effect, plus various distributional and independence assumptions), is unlikely to be the true model, and the null hypothesis is said to be 'rejected'. This is often interpreted as evidence for a real effect by end users. This interpretation is flawed, however, because the p-value is not the same as the probability that the null hypothesis is true. The correct interpretation is that the significance level represents a hurdle for the experimenter to get over, often a necessary condition for publication, nothing more. A given p-value may or may not correspond to strong evidence for a real effect.

The p-value measures differences between the true model and H_0. Since any model is only an approximation to the process generating the data, statistical 'significance' ($p < \alpha$, for any given α) will be obtained for sufficiently large sample sizes. There is no guarantee, however, that the alternative hypothesis (e.g. a model with non-zero linkage disequilibrium between a marker and QTL plus similar distributional and independence assumptions) which we wish to establish, is any more likely. In other words, there is no guarantee that any given p-value represents positive evidence for the effect being tested. This problem gets worse as sample sizes increase—for a given p-value, the relative likelihood of the alternative hypothesis gets smaller and smaller as the sample size gets larger and larger.

In summary, there is no good relationship between p-values and evidence for an effect which is independent of sample size. In the Bayesian framework for statistical analysis, reliability is given by the posterior probability for the alternative hypothesis (H_1). Proper use of Bayesian methodology would immediately show up weak evidence to both the reader and the experimenter, pointing to the need for further replication.

Over-estimation of effects: selection bias. A related problem is over-estimation of effects caused by selection bias. If effects of selected markers are real but not reliably detected, i.e. the power of detection is not high, then re-using the same data to estimate the size of an effect as was used to select a marker results in an upwards bias in the size of effects which can be large, in percentage terms. The solution is to either use an independent population to obtain unbiased estimates of effects (which is inefficient) or, in the Bayesian framework, to use model averaging. See BALL (2001) for a Bayesian model averaging approach for obtaining unbiased estimates in the QTL mapping context. [1]

Other problems. The interpretation of p-values as representing good evidence for real effects is one, though possibly not the only problem with reliability of published data. Other problems may include spurious associations arising from neglecting population structure or inadequate experimental design. Or, it may be that the genetic effects are not be 'stable' across different environments or sub-populations sampled by various trials. Before claiming that there is instability, however, we need to be sure there really was good evidence for an effect in the first place, then obtain good evidence for an interaction.

This paper quantifies the level of replication required for association studies, for reliably detecting linkage equilibrium between a DNA marker (or gene) and a trait, with particular interest in detecting small effect genes. To do this we apply the Bayesian approach to assessing evidence, using the

[1] The model averaging approach applies ideally to sets of multiple markers, but can also be applied when considering only a single marker, in which case the models averaged over correspond to the null hypothesis H_0 with no effect and the alternative hypothesis H_1 with a real effect.

Bayes factor. Sample sizes needed to detect effects of various sizes with a given strength of evidence as defined by the Bayes factor, with a given probability (power), for a given level of linkage disequilibrium between a bi-allelic marker and a bi-allelic quantitative trait locus, are determined.

We make use of an existing deterministic power calculation (Luo 1998), for the power of detecting linkage disequilibrium between a bi-allelic quantitative trait locus (QTL) and a marker. The type I error rate (or p-value) corresponding to the desired Bayes factor is determined, for each set of parameter values and plugged into the deterministic power calculation.

Genetic model: We will assume a bi-allelic marker with alleles M, m and a bi-allelic QTL with alleles A, a. Following LUO (1998), let q, p be the probability of A, M respectively. In the bi-allelic case, linkage disequilibrium is specified by a single coefficient, D, such that the joint probabilities of alleles are given by:-

$$\begin{aligned} Pr(A,M) &= Pr(A)Pr(M) + D = qp + D, \qquad (1) \\ Pr(A,m) &= Pr(A)Pr(m) - D = q(1-p) - D, \qquad (2) \end{aligned}$$

(Cf WEIR 1996). It follows that the conditional probabilities are given by

$$\begin{aligned} Q &= Pr(A \mid M) = q + D/p, \qquad (3) \\ R &= Pr(A \mid m) = q - D/(1-p). \qquad (4) \end{aligned}$$

The genotypic frequencies of pairwise combinations of QTL and marker genotypes are determined from these quantities. (Table 1.)

Table 1: Expected genotypic frequencies and phenotypic values from LUO (1998) for QTL/marker genotype combinations. Marker genotypes are MM, Mm, mm; QTL genotypes are AA,Aa,aa.

	frequencies (f_{ij})			expected
marker	MM	Mm	mm	value
QTL				
AA	p^2Q^2	$2p(1-p)QR$	$(1-p)^2R^2$	d
Aa	$2p^2Q(1-Q)$	$2p(1-p)(Q+R-2QR)$	$2(1-p)^2R(1-R)$	h
aa	$p^2(1-Q)^2$	$2p(1-p)(1-Q)(1-R)$	$(1-p)^2(1-R)^2$	$-d$

Statistical models: The statistical model for observed phenotypes is

$$y_{ijk} = \mu + \beta_i + \omega_{ij} + e_{ijk}, \quad e_{ijk} \sim N(0, \sigma^2), \qquad (5)$$

where i, j, k index marker genotype, QTL genotype and observations within marker and QTL genotype, respectively. Since the QTL genotypes are unobserved, the data is analysed with a one-way analysis of variance model, with marker genotypes as groups i.e.

$$y_{ijk} = \mu + \beta_i + \epsilon_{ijk}, \quad \epsilon_{ijk} \sim N(0, \sigma_w^2), \tag{6}$$

The ANOVA table for this analysis is shown in Table 2.

Table 2: ANOVA table for single marker analysis

	df	SS	MS	F
between marker classes	$\nu_1 = 2$	SS_b	$MS_b = SS_b/\nu_1$	$F = MS_b/MS_w$
within marker classes	$\nu_2 = n - 3$	SS_w	$MS_w = SS_w/\nu_2$	

Classical power calculation: Classical statistical experimental design seeks to determine designs with a given 'power' P at a given significance level α i.e. the probability of obtaining a result with p-value less than α, is P assuming the effect to be detected is at least a certain size. The effect is said to be 'detected' if this significance level is obtained.

The null hypothesis is 'rejected' at level α if the observed value F statistic is greater than the $1 - \alpha$ point of its distribution under the null hypothesis:

$$F > F_{1-\alpha:\nu_1,\nu_2} \tag{7}$$

where $F_{1-\alpha:\nu_1,\nu_2}$ is the $1 - \alpha$ point of the F distribution on ν_1, ν_2 degrees of freedom.

The probability of rejecting the null hypothesis or *power* is the probability that the F statistic exceeds the critical value. To find this probability we need to know the actual distribution of the F statistic under the alternative hypothesis, which in this case means there is a bi-allelic QTL in linkage disequilibrium with a marker with parameters D, d, h, p, q defined above.

Under the alternative hypothesis F is distributed as a non-central F distribution with non-centrality parameter δ given by:

$$\delta = \frac{EMS_b}{EMS_w} \frac{\nu_1(\nu_2 - 1)}{\nu_2} - \nu_1, \tag{8}$$

where EMS_b, EMS_w are the expected values of the mean squares between and within marker classes (Cf Table 2; LUO equation 6.2; JOHNSON AND KOTZ 1972, p 189). The power is given by:

$$Pr(F_{\nu_1,\nu_2}(\delta) > F_{1-\alpha:\nu_1,\nu_2}), \tag{9}$$

where $F_{\nu_1,\nu_2}(\delta)$ is a random variable with the F distribution with ν_1, ν_2 degrees of freedom and non centrality parameter δ.

To determine δ and complete the calculation, it remains to determine the expected mean squares EMS_b, EMS_w for the problem. Details of the calculation including derivation of modified values for ω_{23}, EMS_b, and EMS_w are given in Appendix 1.

Experimental design with power to obtain a given Bayes factor:

Bayes' theorem follows from the basic law of conditional probability:

$$Pr(A \mid B)Pr(B) = Pr(A \cap B) = Pr(B \mid A)Pr(A) \tag{10}$$

or

$$Pr(A \mid B) = Pr(B \mid A)Pr(A)/Pr(B). \tag{11}$$

If we have observed B the probability of A being true is $Pr(A \mid B)$. Bayesian statistics applies this with B as the data (y) and A as an unknown parameter (θ). Replacing A by θ and B by y and representing the probability functions by f, π and g gives

$$g(\theta \mid y)f(y) = f(y \mid \theta)\pi(\theta) \tag{12}$$

It follows (by integrating both sides over θ) that

$$f(y) = \int f(y \mid \theta)\pi(\theta) \tag{13}$$

$$g(\theta \mid y) = \frac{f(y \mid \theta)\pi(\theta)}{\int f(y \mid \theta)\pi(\theta)}. \tag{14}$$

This has the following interpretation: the *prior distribution* $\pi(\theta)$, represents our knowledge of the unknown θ before observing the data y, and the *posterior distribution* $g(\theta \mid y)$ represents our knowledge of the unknown θ *after* observing the data. In other words π has been *updated* to g. $f(y)$ is the probability of the data.

Now suppose we have two hypotheses (or models) H_0 (e.g. the hypothesis of no linkage disequilibrium), and H_1 (e.g. the hypothesis of a non-zero QTL effect in linkage disequilibrium with a marker). Each hypothesis represents a model with unknown parameters and probability density functions. Let θ_i be the parameters under $H_i\,(i = 0, 1)$ and let $\pi_i(\theta_i)$, $f_i(y \mid \theta_i)$, $g_i(\theta_i \mid y)$, $f_i(y)$ be the probability functions. Now $f_i(y)$ is the probability of the data under hypothesis H_i, so we write it as a conditional probability $Pr(y \mid H_i)$. Additionally let $\pi_i(H_i)$ denote the prior probabilities for each model.

The *Bayes factor*, B, measures the strength of evidence in the data in support of one hypothesis (or model) H_1 (e.g. the hypothesis of a non-zero QTL effect in linkage disequilibrium with a marker) over another H_0 (e.g. the hypothesis of no linkage disequilibrium), and is defined as the ratio of the probability of the data under H_1 to the probability of the data under H_0 i.e.

$$B = \frac{Pr(y \mid H_1)}{Pr(y \mid H_0)}. \tag{15}$$

Higher Bayes factors (greater than 1) are stronger evidence for H_1, while lower Bayes factors (less than 1) are evidence for H_0. A Bayes factor of 1 means the data are equally likely under H_0 and H_1, so there is no evidence either way. Bayes' theorem in this case can be written as

$$Pr(H_1 \mid y)/Pr(H_0 \mid y) = B \times \pi(H_1)/\pi(H_0) \tag{16}$$

where $\pi(H_i), Pr(H_i \mid y)$ are the prior and posterior probabilities of H_i i.e.

$$\text{posterior odds} = \text{Bayes factor} \times \text{prior odds}. \tag{17}$$

The Bayes factor has a natural interpretation in terms of betting odds. Prior odds are the odds we would be prepared to bet on prior to seeing the data;

posterior odds are the odds we would be prepared to bet on after seeing the data. Equation (17) has the interpretation that the Bayes factor is the factor by which we multiply our prior odds after seeing the data. For example if we had prior odds of $1 : 10$ (i.e. odds *against* a QTL in a given region), and a Bayes factor of 100 our posterior odds would be $10 : 1$ ($100 \times 1/10 = 10$). This may not sound impressive to readers accustomed to p-values less than 0.01 or even 0.001 but we shall see that it is not easy to obtain evidence this strong. For example in a t-test, with a non-informative or nearly non-informative prior distribution on the effect and sample size greater than 100, a p-value of 0.05 can correspond to a Bayes factor less than 1, i.e. little or no evidence against H_0 or even evidence *for* H_0. (Cf Berger and Berry 1998, and Table 3 below).

Note: The reader may notice some similarity between the likelihood ratio and the Bayes factor: in fact the Bayes factor is the same as the likelihood ratio—if there are no unknown parameters. However use of the Bayes factor is not the same as the use of the likelihood ratio test (Cf the discussion).

Prior odds are part of prior knowledge, but do not affect the Bayes factor. The Bayes factor is, however, affected by the prior distribution of parameters ($\pi_i(\theta_i)$ above), particularly the parameters being tested. In general the more prior information we have on the parameter values under H_1, the higher the Bayes factor we will obtain.

In a particular situation, where there is prior information one should choose somewhat conservative prior distributions for these parameters, so the evidence for an effect is not exaggerated. A vague prior which puts most prior mass on very large numbers, should not be used naively either–in the limit as prior information tends to zero, the Bayes factor also tends to zero. To develop a generic method, or where there is no prior information, one way to proceed is to start with priors with little or no prior information and update these priors using a small 'training sample' (y_a), (i.e. replace the priors by the posteriors after observing y_a) and use the rest of the data to estimate the Bayes factor with the updated priors. It can be shown that this is equivalent to defining

$$B(y) = B_0(y)/B_0(y_a) \tag{18}$$

where $B_0(y)$ denotes the Bayes factor with data y and the original priors and $B(y)$ is the Bayes factor with data y and the updated priors.

Note that now, (setting $y = y_a$ in (18)), the Bayes factor is calibrated to be 1 for the small training sample. This is reasonable because the training sample is small, so contains little or no evidence either way, which is consistent with the interpretation of $B(y) = 1$. This is the motivation for the approach of Spiegelhalter and Smith (1982) who obtain, for a one-way analysis of variance model :-

$$B = \left[\frac{1}{2}\frac{(m+1)}{n}\prod_{i=1}^{m} n_i\right]^{1/2}\left[1 + \frac{(m-1)}{(n-m)}F\right]^{-n/2}, \tag{19}$$

where m is the number of groups, n_i the number in each group, n is the total sample size, and F is the classical F-value.

This form of the Bayes factor is particularly convenient, because it links directly to the F-statistic used in existing power calculations. To detect an association with power P to achieve a given Bayes factor B we solve for F in (19), and select a design using the existing deterministic power calculation.

Equation (19) applies to testing for linkage disequilibrium, (Cf the ANOVA table, Table 2), where $m = 3$ is the number of marker classes, and n_1, n_2, n_3 are the number of observations in each marker class, and F is the F-value. To obtain a deterministic formula we set n_1, n_2, n_3 to their expected values based on the frequencies of marker classes MM, Mm, mm in Table 1, and the total sample size:

$$n_1 = np^2, n_2 = 2np(1-p), n_3 = n(1-p)^2. \tag{20}$$

Substituting in (19) gives

$$B \approx \left[4n^2p^3(1-p)^3\right]^{1/2}\left[1 + \frac{2}{(n-3)}F\right]^{-n/2}. \tag{21}$$

This should be a good approximation for the large sample sizes we need: variations in observed proportions in each marker class will be small, and hence will have little effect on the relationship between the F value and Bayes factor in (19).

For any valid choice of values of QTL/marker allele frequencies, effects and linkage disequilibrium parameters D, d, h, p, q, sample size n and Bayes factor B we can solve for $F = F_{1-\alpha:2,n-2}$ in (21), then lookup the value of α from the F-distribution. Then given α and n determine the power, P, from the classical power calculations. To determine the sample size n, for given B and P we use interpolation.

R code used for the calculations of this paper will be submitted to the

Comprehensive R archive network (CRAN), (http://lib.stat.cmu.edu/R/CRAN).

RESULTS

Table 3 shows the type I error rates (or p-values) corresponding to various Bayes factors. For a desired Bayes factor, (e.g. $B = 10$ which implies the data are 10 times more likely under the hypothesis of a real effect, than under the hypothesis of no effect), look up the type I error rate in the table for the sample size desired. For example, if the sample size is 1728, we need a type I error rate of $\alpha < 2.35 \times 10^{-4}$ to get $B = 10$. If the sample sizes is 300, we need a type I error rate of $\alpha < 1.42 \times 10^{-3}$ to get $B = 10$. For these sample sizes, $\alpha = 0.05$ and even $\alpha = 0.01$ are clearly not a good option, corresponding mostly to Bayes factors less than 1. For example with $n = 864$, $B = 1/10$, we have $P \approx 0.05$— showing that a p-value of 0.05 can correspond to evidence *against* an effect.

Table 3: Type I error rates (p-values) corresponding to various Bayes factors, for testing for linkage disequilibrium between a bi-allelic marker and QTL.

	Bayes factor (B)						
n	1/20	1/10	1/5	1	5	10	20
300	0.270	0.136	0.069	0.0139	2.83×10^{-3}	1.42×10^{-3}	7.18×10^{-4}
432	0.188	0.094	0.047	0.0096	1.94×10^{-3}	9.73×10^{-4}	4.89×10^{-4}
600	0.135	0.068	0.034	0.0068	1.38×10^{-3}	6.92×10^{-4}	3.47×10^{-4}
864	0.093	0.047	0.023	0.0047	9.49×10^{-4}	4.76×10^{-4}	2.39×10^{-4}
1200	0.067	0.034	0.017	0.0034	6.84×10^{-4}	3.40×10^{-4}	1.71×10^{-4}
1728	0.047	0.023	0.012	0.0023	4.69×10^{-4}	2.35×10^{-4}	1.18×10^{-4}
2400	0.033	0.017	0.008	0.0017	3.37×10^{-4}	1.69×10^{-4}	8.44×10^{-5}
3756	0.021	0.011	0.005	0.0010	2.15×10^{-4}	1.07×10^{-4}	5.37×10^{-5}
4800	0.017	0.008	0.004	0.0008	1.68×10^{-4}	8.38×10^{-5}	4.19×10^{-5}

Table 4 is a comparison with results for 12 sample populations from Luo

(1998). $P_{0.05}$ is the power to detect an effect with comparison-wise significance level $\alpha = 0.05$, as shown in LUO Table 3. Also shown are the equivalent Bayes factors, B, and the sample size $n_{B_{20}}$ required to obtain a Bayes factor of 20 with power 0.9. Note that the Bayes factors for the original sample sizes are all less than 1, i.e. not coresponding to positive evidence for a real effect. The sample sizes $n_{B_{20}}$ are the sample sizes required to have good power to detect an effect with fairly strong evidence, and are substantially (up to 13 times) larger.

Table 4: Comparison with results from LUO (1998). Results are shown for the 12 example populations (Cf LUO TABLES 2, 3.) with sample size n, marker and QTL allele frequencies p, and q, linkage disequilibrium D, QTL heritability h_Q^2 and dominance ratio ϕ. $P_{0.05}$ is the power to detect an effect with $\alpha = 0.05$, B is the corresponding Bayes factor, and $n_{B_{20}}$ is the sample size required to achieve a Bayes factor of 20 with power 0.9.

pop.	n	p	q	D	h_Q^2	ϕ	$P_{0.05}$	B	$n_{B_{20}}$
1	100	0.5	0.5	0.1	0.1	0.0	0.18	0.88	1837
2	200	0.5	0.5	0.1	0.1	0.0	0.34	0.42	1837
3	200	0.5	0.5	0.2	0.1	0.0	0.91	0.42	381
4	200	0.5	0.5	0.1	0.2	0.0	0.62	0.42	849
5	200	0.5	0.5	0.1	0.1	0.5	0.31	0.42	2047
6	200	0.5	0.5	0.1	0.1	1.0	0.25	0.42	2640
7	200	0.3	0.3	0.1	0.1	0.0	0.46	0.54	1211
8	200	0.7	0.7	0.1	0.1	0.0	0.46	0.54	1211
9	200	0.3	0.5	0.1	0.1	0.0	0.39	0.54	1476
10	200	0.5	0.3	0.1	0.1	0.0	0.39	0.42	1513
11	200	0.4	0.6	0.1	0.2	1.0	0.45	0.45	1259
12	200	0.6	0.4	0.1	0.2	1.0	0.54	0.45	995

Note: Our calculations for for the power $P_{0.05}$ to detect an effect with significance level $\alpha = 0.05$ agree with those from Table 3 of LUO, and with our stochastic simulations, except for populations 11 and 12, which agree with our simulations but do not agree with the results from LUO. LUO obtained powers of 0.65, 0.60, which are similar to the power obtained when $\phi = 0$. It may be that he has used $\phi = 0$ for the power calculations for these populations.

Figure 1 shows graphs of power versus linkage disequilibrium. Panels correspond to a sample sizes $n = 600, 1200, 2400, 4800$, and QTL heritabilities (or proportions of phenotypic variance explained) of $h^2_Q = 0.01, 0.05$. Within each panel, the solid lines are graphs of power versus linkage disequilibrium for Bayes factors of 1/20,1/10,1/5,1,5,10,20. The lines, in order of increasing power correspond to the Bayes factors in decreasing order.

Graphs of power versus disequilibrium for various values of marker and QTL allele frequencies are shown in Figure 2. To facilitate comparisons across the range of allele frequencies disequilibrium has been represented as $D' = D/D_{max}$, which varies from 0 to 1, with 1 representing the maximum possible linkage disequilibrium for the given values of marker and allele frequencies.

Note that the low allele frequencies *per se* don't have a major effect on the power obtained when disequilibrium reaches its maximum, although this may happen for a lower D when q is lower (Cf the middle panel of the top row with $n = 4800, p = 0.2$). However, very low power occurs when there is a poor match between marker and QTL frequencies (e.g. $p = 0.01, q = 0.5$ or $p = 0.5, q = 0.01$).

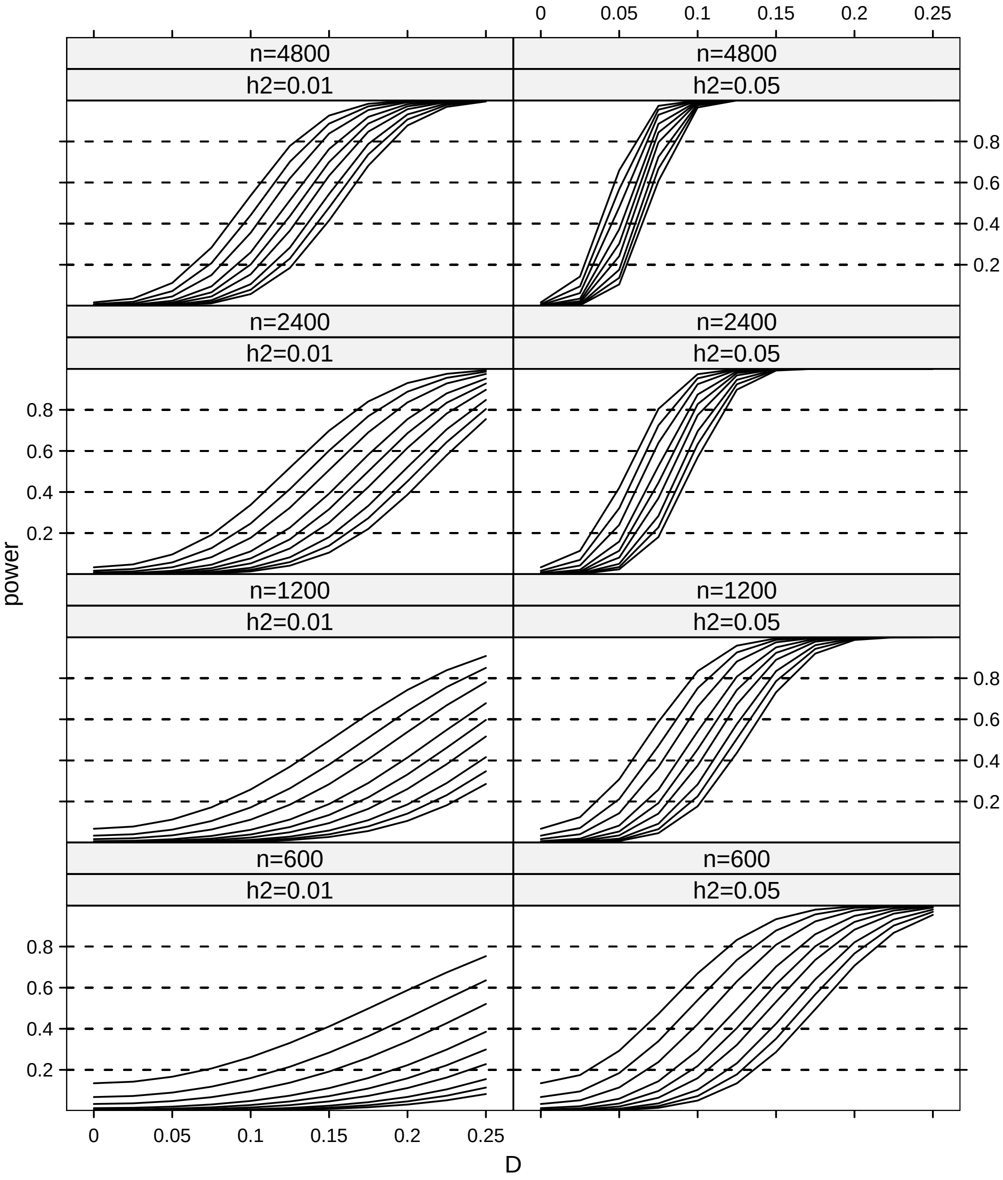

Figure 1: Power versus disequilibrium for QTL with $h_Q^2 = 0.01, 0.05$.

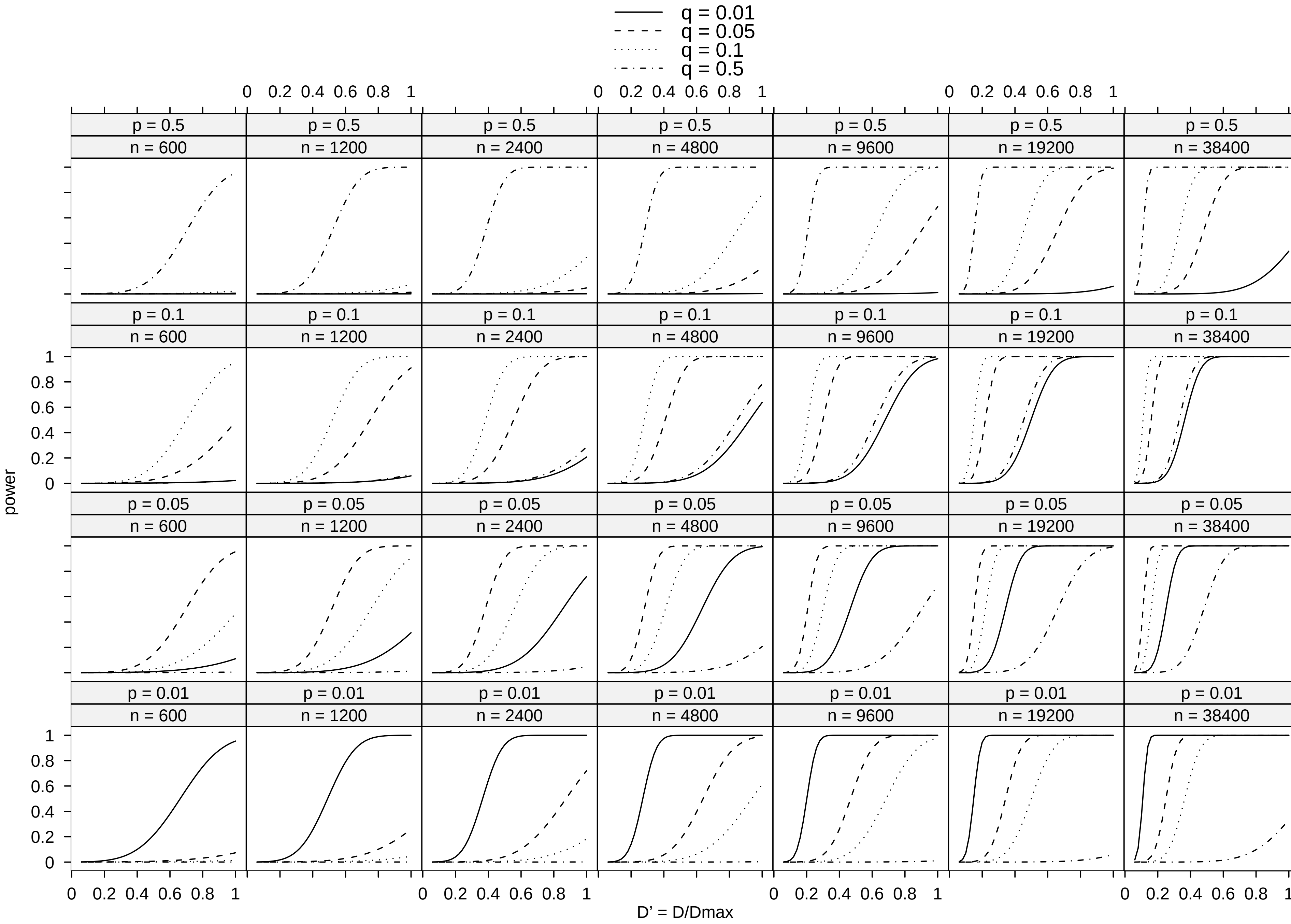
q = 0.01
q = 0.05
q = 0.1
q = 0.5
p = 0.5
p = 0.1
p = 0.05
p = 0.01
n = 600
n = 1200
n = 2400
n = 4800
n = 9600
n = 19200
n = 38400
power
0
0.2
0.4
0.6
0.8
1
D' = D/Dmax

Figure 2: Effects of marker allele frequencies (p) and QTL allele frequencies (q) on power to detect linkage disequilibrium. Each panel corresponds to a combination of $p = 0.01, 0.05, 0.1, 0.5$ and $n = 600, 1200, 4800, 9600, 19200, 38400$. Within each panel, power curves for each of $q = 0.01, 0.05, 0.1, 0.5$ are shown by different line types. The curves shown are for the power to detect a QTL versus D, with $h_Q^2 = 0.05$, and a Bayes factor of 20.

Application recommendations: The Bayes factor gives a well defined measure of strength of evidence, independent of the experimental design or sample size used. The optimal value to use depends on the costs of further experimentation, and possible benefits. A Bayes factor of 20 or more represents good evidence for an effect, however one needs to factor in the prior odds. An effect which is *a priori* unlikely needs a high Bayes factor to obtain respectable posterior odds. A more formal cost-benefit analysis is possible, in the Bayesian decision theory framework, see e.g. DeGroot (1970), Lindley (1985).

Use with genome scans. In a genome scan the prior odds for a gene to be in the *vicinity* of a particular marker, defined as the region closer to that marker than any other, are proportional to the number of genes expected and inversely proportional to the number of markers (Cf Ball 2001). In a genome scan with many markers the prior probability would be small, therefore a high Bayes factor would be required. Kruglyak (1999) suggests that $D = 0.1$ may be obtainable for the human population, at distances between QTL an marker of up to $3kb$, corresponding to a map with $6kb$

Table 5: Sample sizes required for power of 0.9 of detection of linkage disequilibrium between a bi-allelic QTL and a bi-allelic marker with given posterior odds for linkage disequilibrium with $D = 0.1$, $p = 0.5$, $q = 0.5$ in a genome scan with 500 000 SNP markers. Prior probability per marker is assumed to be 1/50 000.

		sample size	
posterior odds	Bayes factor	$h_Q^2 = 0.05$	$h_Q^2 = 0.01$
1/20	2 500	5 572	30 640
1/5	10 000	6 008	32 792
1	50 000	6 524	35 397
5	250 000	7 031	37 949
20	1 000 000	7 465	40 089

spacing and 500 000 markers.

For example, if it is desired to localise an effect to the nearest of 500 000 SNP markers and there were about 10 genes expected, the prior probability per marker would be around 1/50 000. To obtain respectable posterior odds of say 20:1 in (16), we would require a Bayes factor of 1 000 000.

Table 5 shows sample sizes required for localising QTL in a genome scan with 500 000 SNP markers, assuming there are 10 QTL. A sample size of $\approx 40\,000$ is needed for a power of 0.9 to obtain posterior odds of 20:1 for a QTL explaining 1% of the phenotypic variance with $D = 0.1$ to be within $\pm 3kb$ of a given marker. Note that, in relative terms, the sample sizes to obtain the higher Bayes factors don't increase much at the larger sample sizes.

Use with candidate genes. If a large number of 'candidate' genes was being tested, at an initial screening stage of experimentation, one may be content with posterior odds of less than 1. In this case one would want to be sure the effects of interest had a high probability (power) of being accepted, while the genes not affecting the trait have a low probability. For example suppose 50 000 candidate genes were tested with a sample size of $n = 4800$, and we are looking for QTL explaining 1% of the phenotypic variance ($h^2_Q = 0.01$). From Fig 1 with $n = 4800$, $B = 1$ we have a power of 0.5 provided $D \geq 0.13$, and a power of 0.001 provided $D = 0$. Thus with this design we would, on average end up with 50 false positives and half of the genes with $D \geq 0.13$. On the other hand requiring $B = 20$ would eliminate 80% of the genes with $D \geq 0.13$, which is hardly satisfactory. For QTL with $h^2_Q = 0.05$, the situation is more favourable: with $n = 4800$ and $D \geq 0.1$ there is a 95% power to detect genes with $D \geq 0.1$ with a Bayes factor of 20, with a rate of less than 1/1000 false positives.

Table 6 shows sample sizes required for selecting genes from a set of 50 000 candidate genes, assuming there are 10 true genes. A sample size of $\approx 36\,000$ is needed for a power of 0.9 to obtain posterior odds of 20:1 for a gene explaining 1% of the genetic variance, with $D = 0.1$ to be within $\pm 3kb$ of a given marker. As with Table 5, the sample sizes needed to obtain the higher Bayes factors are not much higher in relative terms.

Table 6: Sample sizes required for power of 0.9 of detection of linkage disequilibrium between a bi-allelic QTL and a bi-allelic marker with given posterior odds for linkage disequilibrium with $D = 0.1$, $p = 0.5$, $q = 0.5$ in a set of 50 000 markers representing candidate genes. Prior probability per marker is assumed to be 1/5000.

		sample size	
posterior odds	Bayes factor	$h_Q^2 = 0.05$	$h_Q^2 = 0.01$
1/20	250	4 826	26 808
1/5	1 000	5 288	29 093
1	5 000	5 808	31 658
5	25 000	6 322	34 223
20	100 000	6 762	36 406

DISCUSSION

Detection of markers in linkage disequilibrium with a trait is a complex statistical problem. Classical single marker test methods have not lead to reliable SNPs in published data, and the need for more rigorous statistical evidence for detection of SNPs has been noted. A major problem is with the interpretation of p-values as evidence. The results of this paper confirm the wide range in p-values required to obtain a given Bayes factor, as sample size varies.

The results of Table 4 show that using a threshold of $P = 0.05$ does not correspond to positive evidence for a real effect, and up to 13 times larger sample sizes are needed to obtain good evidence (a Bayes factor of 20) with power 0.9.

The results of Figure 1 show that good power is achieved with a sample size of $n = 2400$, for a 5% QTL with $D = 0.15$, to detect linkage with a

Bayes factor of 20. However at $n = 4800$ and $D = 0.1$ power is low even for a Bayes factor of 1.

The results of Figure 2 show that power may be very low if low QTL allele frequency is is not matched by a low marker frequency. In this case the linkage disequilibrium is much lower than the maximum possible for the given QTL allele frequency. This sensitivity of power to allele frequency is one reason why plots of the LD test statistics in the neighbourhood of a gene generally vary wildly, in a non-smooth fashion. To detect a gene, particularly one with a low frequency allele, it is not only important to have a marker close to the gene but also to have a marker with similar allele frequency. For a randomly chosen marker in the vicinity of a gene it seems that this would occur with low probability for a low frequency QTL allele. This further increases the number of markers needed for adequate genome coverage and/or the population size needed to detect such genes.

The results of Tables 5 and 6 show that it is feasible, albeit with large sample sizes of the order of 7500, 40 000 to detect and locate QTL which explain 5%, 1% or more of the genetic variance respectively; with a linkage disequilibrium of $D = 0.1$, by testing sets of up to 50 000 candidate genes or 500 000 SNP markers. For genome scans, it is possible to borrow strength from neighbouring markers, justifying higher prior probabilities when estimating the marginal probability for a set of markers in a larger region (Cf Ball 2001). This can't be pushed too far, however: because of the short-range of usable linkage disequilibrium indicated in Kruglyak (1999), only a small number of the 500 000 markers would be close enough to a given gene.

An alternative, in view of the large sample sizes indicated for detecting small QTL with the genome scan or brute force candidate gene approaches, is a hybrid approach whereby candidate genes or genomic regions are preselected e.g. from differential gene expression or QTL mapping studies. Another alternative is to find populations with a somewhat larger range of useable linkage disequilibrium, giving less precise information but with a lower genotyping cost.

Various alternative experimental designs are possible, such as the transmission disequilibrium test, which avoid potential spurious associations due to population substructure (ALLISON 1997; LONG and LANGLEY 1999; SPIELMAN and EWENS, 1998; BOEHNKE and LANGEFELD 1998). We do not consider these directly but note that a transmission test can be viewed as a test of association between transmission of an allele and a trait, and the power calculations re-derived. We would expect a transmission test of equivalent power (in the classical sense) to the designs considered here, would be expected to have a similar power to achieve a given Bayes factor, i.e. existing comparisons between the power of these tests and those of association studies would apply, however this is yet to be confirmed.

An alternative to considering single markers separately is to study haplotypes. For example, if one is interested in a particular haplotype or class of haplotypes from a subset of SNPs being associated with a higher level of disease, compared with all other haplotypes from the subset, this can be considered a bi-allelic marker for which the methods of this paper apply. If more than two haplotype classes are considered the equivalent power calculations would need to be derived. Use of p-values is problematic in this situation because p-values are affected by considerations, such as which other haplotypes or haplotype classes are tested. The Bayesian approach is well suited to haplotypes—the Bayes factor or posterior probability does not depend on such considerations. Of course if there was a very large number of combinations or partitions of haplotypes the prior probability for a randomly chosen haplotype would be low, and a high Bayes factor would be required.

The Bayes factor depends on the prior for the variable(s) being tested (here marker effects). We have used a generic, or 'default' Bayes factor which avoids this prior dependence, giving a generic result which can be applied in the absence of further prior information. In specific situations it may be possible to do better with prior information on the size of QTL effects is available. For example, the variance due to a QTL can be no more than the genetic variance of the trait, for which estimates are often available. However, in our experience this amount of prior information is not enough to make a major difference.

The approach used here, of determining the power to achieve a given Bayes factor is a hybrid of Bayesian, and non-Bayesian (frequentist) approaches. We are studying the *sampling distribution* (frequentist) of the *Bayes factor* (a Bayesian measure of evidence). This is appropriate because we are interested in possible outcomes of future experiments, *if* there is an association with certain parameter values. However once an experiment has been carried out, in the Bayesian viewpoint, there is only one observed data set, and the Bayes factor is a property of that dataset; we would not normally consider the sampling variability of Bayes factors that might have been obtained.

The problem with p-values remains whether one uses any of the many forms of hypothesis tests: the t-test, F-test, χ^2-test, or likelihood ratio test. One may wonder why the likelihood ratio test doesn't solve the problem, when the Bayes factor is essentially a likelihood ratio. The difference is in how the unknown parameters are treated. In the non-Bayesian likelihood ratio test one maximises over the unknown parameters

$$LR = \frac{f_1(y \mid \hat{\theta}_1)}{f_0(y \mid \hat{\theta}_0)},$$

where $\hat{\theta}_1$, $\hat{\theta}_0$ are chosen to maximise their respective likelihood functions $f_1(y \mid \theta_1)$, $f_0(y \mid \theta_0)$. This strongly favours the alternative hypothesis, which usually has one or more additional parameters to maximise over. To obtain a hypothesis test, the likelihood ratio, LR, is compared to its sampling distribution under the null hypothesis. High values of the likelihood ratio are interpreted as meaning the null hypothesis is rejected. As with any other hypothesis test, there is no guarantee, however, that the alternative hypothesis, is any more likely.

The problem with p-values remains whether one uses comparison-wise or experiment-wise p-values, or p-values from a permutation test. A permutation test is just a non-parametric way of obtaining a p-value. Of course the experiment-wise p-value usually corresponds to a much lower comparison-wise p-value so represents stronger evidence. The interpretation of the p-value, and what threshold to use remains a problem.

Obtaining a Bayes factor of 20 may require much higher sample sizes, than obtaining a p-value of 0.05, say. One may ask: why use Bayes factors if this requires larger sample sizes and hence greater cost, if one has $p = 0.05$? This is not a reflection on the relative efficiency of the methodologies, because the Bayes factor of 20 represents stronger evidence than the p-value 0.05 which may represent very weak evidence— “you (only) get what you pay for”. If a much larger sample size is required for power to obtain the desired Bayes factor this implies that the original experiment was under powered, and the p-value corresponds to a small Bayes factor, i.e. very weak evidence, while creating a misleading impression of evidence for an effect.

ACKNOWLEDGEMENTS

The author thanks Craig Echt and Phillip Wilcox, leaders of the Forest Research Genomics group (now Cell Wall Biotechnology Centre) for supporting this work, and Rowland Burdon for useful comments. This work was funded by the New Zealand Foundation for Research Science and Technology.

LITERATURE CITED

Allison, D. B. 1997 Transmission disequilibrium tests for quantitative traits. Am. J. Hum. Genet. **60:** 676–690.

Altshuler, D., J. N. Hirschhorn, . . . , L. Groop, and E. S. Lander 2000 The common PPARγ Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. Nature Genetics **26:** 76–80.

Ball, R. D. 2001 Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using the Bayesian Information Criterion. Genetics **159:** 1351–1364.

Bayes, T. 1763 An essay towards solving a problem in the doctrine of chances, Phil Trans. Roy. Soc. **53:**370–418.

Berger, J. and D. Berry 1988 Statistical analysis and the illusion of objectivity. Am. Scientist **76:** 159–165.

Boehnke, M. 2000 A look at linkage disequilibrium. Nat. Genet. **25:** 246–247.

Boehnke, M. and C. D. Langefeld 1998 Genetic association mapping based on discordant sib pairs: the discordant-alleles test. Am. J. Hum Genet. **62:** 950–961.

De Groot, M. H. 1970 *Optimal Statistical Decisions.* McGraw-Hill, New York.

Dunning, A. M., F. Durocher, C. S. Healey, M. D. Teare, S. E. McBride *et al.,* 2000 The extent of linkage disequilibrium in four populations with distinct demographic histories. Am. J. Hum. Genet. **67:** 1544–1554.

Dunner, S., C. Charlier, F. Farnir, B.Brouwers, J.Canon *et al* 1997 Towards interbreed IBD fine mapping of the *mh* locus: double-muscling in the *Asturiana de los Valles* breed involves the same locus as in the *Belgian Blue* cattle breed. Mamm. Genome **8:** 430–435

Farnir, F., W. Coppetiers, J.-J. Arranz, P. Berzi, N. Cambisano *et al.* 2000 Extensive genome-wide linkage disequilibrium in cattle. Genome Res. **10:** 220–227.

Gura, T. 2000 Can SNPs deliver on susceptibility genes? Science **293:** 593–595.

Hampton, T. 2000 Research Brief, Focus September 29, 2000. Harvard University.

Johnson, N. L. and Kotz, N. 1970 *Distributions in Statistics: Continuous Univariate Distributions.* Houghton Mifflin, Boston.

Knott, S. A. 1994 Prediction of the power of detection of marker-quantitative trait locus linkage using analysis of variance. Theor. Appl. Genet. **89:**, 318–322.

Kruglyak, L. 1999 Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nature Genetics **22:** 139–144.

Kumar, S., R.J. Spelman, D. J. Garrick, T. E. Richardson, M. Lausberg, and P. L. Wilcox, 2000 Multiple marker mapping of wood density loci in an outbred pedigree of radiata pine. Theor. Appl. Genet. **100:** 926-933.

Lindley, D.V. 1985 *Making Decisions.* John Wiley & Sons, London.

Long, A. D. and C. H. Langley 1999 The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Research **9:** 720–731.

Luo, Z. W. 1998 Linkage disequilibrium in a two-locus model. Heredity **80:** 198–208.

McRae, A. E., J. C. McEwan, K. G. Dodds, T. Wilson, A.M. Crawford and J. Slate 2002 Linkage disequilibrium in domestic sheep. Genetics **160:** 1113–1122.

Nielsen, D. M. and D. Zaykin 2001 Association mapping: where we've been, where we're going. Expert Rev. Mol. Diagn. 1(3) 89–97.

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

Pritchard, J. K., M. Stephens, N. A. Rosenberg, and P. Donnelly 2000b Association mapping in structured populations, American Journal of Human Genetics **67:** 170–181.

Risch, N. J. 2000 Searching for genetic determinants in the new millennium. Nature **405:** 847–856.

Roses, A. D. 2000 Pharmacogenetics and the practice of medicine. Nature **405:** 857–865.

Searle, S. R. 1987 Linear models for unbalanced data. John Wiley and Sons, New York.

Spiegelhalter, D. and A. F. M. Smith 1982 Bayes factors for linear and log-linear models with vague prior information J. Royal Statist Soc. B **44:** 377–387.

Spielman, R. S. and W. J. Ewens 1998 A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am. J. Hum. Genet. **62:** 450–458.

Taillon-Miller, P., I. Bauer-Sardina, N. L. Saccone, J. Putzel, T. Laitinen *et al.* 2000 Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. Nat. Genet. **25:** 324–328.

Terwilliger, J. D., and K. M. Weiss 1998 Linkage disequilibrium mapping of complex disease: fantasy or reality? Curr. Opin. Biotechnol. **9:** 578–594.

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

Wilcox, P. L., T. E. Richardson, and S. D. Carson, 1997 Nature of quantitative trait variation in *Pinus radiata*: insights from QTL detection experiments. pp 304-312 In: *Proceedings of IUFRO '97: Genetics of Radiata Pine.* Dec 1-5, 1997. Rotorua, New Zealand. *FRI Bulletin No. 203*, R.D. Burdon and J.M. Moore, Eds.

Appendix 1. Derivation of modified values for EMS_b, EMS_w.

First we calculate μ, β_2 then ω_{23}. Values for β_1, β_3 and other ω_{ij} values are obtained similarly.

Taking expectations in equation (5),

$$\begin{aligned} \mu &= E(y) \\ &= E(y \mid AA)Pr(AA) + E(y \mid Aa)Pr(Aa) + E(y \mid aa)Pr(aa) \qquad (22) \end{aligned}$$

Substituting QTL genotype probabilities and values and simplifying gives

$$\begin{aligned} \mu &= d \times q^2 + h \times 2q(1-q) - d \times (1-q)^2 \\ &= (2q-1)d + 2q(1-q)h. \qquad (23) \end{aligned}$$

Taking expectations conditional on marker class Mm $(i = 2)$

$$\begin{aligned} \mu + \beta_2 &= E(y \mid Mm) \\ &= E(y \mid Mm, AA)Pr(AA \mid Mm) + E(y \mid Mm, Aa)Pr(Aa \mid Mm) \\ &\quad + E(y \mid Mm, AA)Pr(aa \mid Mm) \qquad (24) \end{aligned}$$

The genotype conditional probabilities are given by:

$$\begin{aligned} Pr(AA \mid Mm) &= Pr(A \mid M)Pr(A \mid m) = QR \qquad (25) \\ Pr(Aa \mid Mm) &= Pr(A \mid M)Pr(a \mid m) + Pr(A \mid m)Pr(a \mid M) = Q(1-R) + (1-Q)R \\ &= Q + R - 2QR \qquad (26) \\ Pr(aa \mid Mm) &= Pr(a \mid M)Pr(a \mid m) = (1-Q)(1-R) \qquad (27) \end{aligned}$$

Noting that $E(y \mid Mm, AA) = E(y \mid AA)$ etc, and substituting QTL genotype conditional probabilities and values and solving for β_2 we obtain

$$\beta_2 = \frac{[D(1-2p)]d + d[2D + (1-2p)(1-2q)]h}{p(1-p)} \tag{28}$$

which agrees with equation 3.2 of Luo.

Finally, taking expectations conditional on marker genotype class Mm $(i = 2)$ QTL genotype aa $(j = 3)$ gives:

$$\mu + \beta_2 + \omega_{23} = E(y \mid Mm, aa) = -d$$

Solving for ω_{23} gives:

$$\begin{aligned} \omega_{23} &= -d - \mu - \beta_2 \\ &= -d - (2q-1)d - 2q(1-q)h - \frac{D(1-2p)d + [2D^2 + D(1-2p)(1-2q)]h}{p(1-p)} \\ &= -\frac{[D(1-2p) - 2pq(1-p)]d + [2D^2 + (1-2p)(1-2q)D + 2pq(1-p)(1-q)]h}{p(1-p)} \end{aligned} \tag{29}$$

This differs from the expression of Luo (p 207) by a factor -1.

Luo (Eq 3.1–3.3 and Appendix 1), following Searle (1987) and Knott (1994) gives expressions for β_i, ω_{ij}, and expressions for EMS_b, EMS_w, in terms of β_i, ω_{ij} and f_i, f_{ij}, where f_i, is the relative proportions of the ith marker genotype and f_{ij} the proportions of the ith marker genotype/jth QTL genotype combination.

The formula given for EMS_w is

$$EMS_w = \frac{1}{n-G_m}\left\{\sum_{i=1}^{G_m}\sum_{j=1}^{G_q} f_{ij} n \omega_{ij}^2 - \sum_{i=1}^{G_m}\frac{1}{f_i}\left[\sum_{j=1}^{G_q} f_{ij}(1+(n-1)f_{ij})\omega_{ij}^2 + 2(n-1)\sum_{j<k\leq G_q} f_{ij}f_{ik}\omega_{ij}\omega_{ik}\right]\right\} + \sigma^2 \tag{30}$$

where $G_m = 3$, is the number of marker genotypes and $G_q = 3$ is the number of QTL genotypes.

Use of the formulae for EMS_b and EMS_w, from Luo gave results which did not agree with stochastic simulations. Our calculations of β_i and ω_{ij} agreed with those of Luo except for the equation for ω_{23}, above, which was out by a factor of -1. Using the corrected value for ω_{23} still did not give results that agreed with simulations. The values of EMS_w did agree with simulations, suggesting that there is a further error in EMS_b. Rather than re-derive the expression for EMS_b we take advantage of the fact that once one of EMS_b or EMS_w is known, the other can be obtained by subtraction using the relationships between EMS_w, EMS_b and total sums of squares viz (Cf Table 2):-

$$SS_T = SS_b + SS_w = (G_m - 1)MS_b + (n - G_m)MS_w.$$

Taking expectations

$$V_p = E(SS_T) = (G_m - 1)EMS_b + (n - G_m)EMS_w$$

where V_p is the total, or phenotypic variance. Solving for EMS_b gives

$$EMS_b = \frac{1}{G_m - 1}\left[(n-1)V_p - (n - G_m)EMS_w\right]. \tag{31}$$

Finally, we need an expression for V_p. Substituting from equation (5) and taking variances gives:

$$\begin{aligned} V_p &= \operatorname{var}(y) = \operatorname{var}(\mu + \beta_i + \omega_{ij} + e_{ijk}) \\ &= \operatorname{var}(\mu + \beta_i + \omega_{ij}) + \operatorname{var}(e_{ijk}) \\ &= E(\beta_i + \omega_{ij})^2 + \sigma^2 \\ &= \sum_{i=1}^{G_m} \sum_{j=1}^{G_q} f_{ij}(\beta_i + \omega_{ij})^2 + \sigma^2. \qquad (32) \end{aligned}$$

CHANGELOG

- Expanded introduction with more references and applications of LD. References to Roses, Risch discussing potential of LD and Altshuler et al and Terwilliger and Weiss showing unreliability in published data, and review paper of Nielsen and Zaykin of methods to 2001.

- Introduction to Bayes factors, and the probability distributions involved.

- Paragraph on Bayes factors for vague priors rewritten, with clearer notation.

- More detail on how the Bayes factor formula of Spiegelhalter applies to the problem with an approximate expression in terms of the problem parameters.

- Discussion of the 'random model of QTL effects' (mentioned in Luo) which caused confusion has been removed.

- New table showing comparison of sample sizes for the 12 example populations in Luo, the equivalent Bayes factors and the sample sizes needed to obtain a Bayes factor of 20 with power 0.9.

- Graphs of power versus linkage disequilibrium reduced to a single 4x2 trellis graphics array with n=600,1200,2400,4800 and h2=0.01,0.05. (Figure 1.)

- New Figure 2 showing effects of differing marker and QTL allele frequencies, for the power to detect a QTL with $h_Q^2 = 0.05$ and Bayes factor 20.

- Expanded discussion including various forms of hypothesis test, χ^2, likelihood ratio with comparison-wise, experiment-wise or permutation test thresholds, and why the problems with p-values are common to all of these forms.

- The discussion closes with answer to the question posed by Sarah Otto: "Why use Bayes factors if one already has p=0.05, and it requires more data and cost?"

- New appendix with derivations of ω_{23}, EMS_b and EMS_w to correct the power calculation from Luo.