Experimental Designs for Reliable Detection of Linkage Disequilibrium in Unstructured Random Population Association Studies

Roderick D. Ball¹

New Zealand Forest Research Institute, Rotorua, New Zealand

Manuscript received November 19, 2003

Accepted for publication February 9, 2005

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 that can be linked to the value of an F -statistic 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 reevaluated 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.

THE advent of dense maps of single-nucleotide polymorphisms (SNPs) covering the genome with 300,000 or more markers offers new opportunities to 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 article we develop experimental design techniques necessary to find the sample size needed to reliably detect a given level of linkage disequilibrium between a biallelic QTL and marker.

This article 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. Further information including relationships between Bayes factors, posterior probabilities, type I error rates, P -values, power, and false discovery rates as evidence for gene effects is given in APPENDIX B. Then, we review and correct the deterministic power calculation from LUO (1998) (with technical details in APPENDIX A) and the classical approach to power of experiments and then introduce a generic Bayes factor (SPIEGELHALTER 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 $\frac{1}{20}$ 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 38,400 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 $\frac{1}{20}$ 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 3 kb, 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,

¹Address for correspondence: New Zealand Forest Research Institute, P.B. 3020, Rotorua, New Zealand.
E-mail: rod.ball@forestresearch.co.nz

to locate SNPs near the genes and hence to sequence regions containing genes. Results in the literature on the range of usable linkage disequilibrium are variable. DUNNING *et al.* (2000) report a similarly low range of disequilibrium in humans, while TAILLON-MILLER *et al.* (2000; reviewed in BOEHNKE 2000) report strong LD at distances as high as 1 Mb. 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 a number of genes contribute 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 "postgenome" 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 and 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; BOEHNKE and LANGEFELD 1998; SPIELMAN and EWENS 1998; LONG and LANGLEY 1999). 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 that where there may be population substructure we recommend the method of PRITCHARD *et al.* (2000a,b) 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 on the basis of a combination of marker values associated with the disease or quantitative trait. In this article, we consider the single-marker

approach, with data from association studies, but note that 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 make it essential to have soundly based statistical evidence for putative associations. This may not be the case in general in published literature, *i.e.*, ALTSHULER *et al.* (2000), discussed in GURA (2001, p. 594):

... a team lead by Altshuler tried a slightly different approach. The investigators 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 ~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 led 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 article is to design experiments for testing associations with the required quantifiably 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 nonallelic heterogeneity, genetic-by-environment interactions, intragene 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).

METHODS

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 is the probability of observing a test statistic more extreme than the observed test statistic under repeated sampling assuming the null hypothesis of no effect holds. The problem is that there is no valid interpretation of P -values as evidence for a real effect independent of sample size and test setup. One reason is that the P -value is a probability under H_0 ; the corresponding probability under H_1 , which is not controlled by the P -value alone, may also be comparably small (APPENDIX B, Equations B10 and B11), in which case the evidence for H_1 will be weak. For a given sample size and test setup, smaller and smaller P -values correspond to stronger and stronger evidence. However, we cannot say how small a P -value is required for a given strength of evidence or even quantify strength of evidence, except by reference to Bayesian methods for specific sample sizes and test setups. The larger the sample size, the smaller the P -values need to be to correspond to a given strength of evidence. We shall see that, for the large sample sizes needed for detecting linkage disequilibrium, reasonable evidence corresponds to very small P -values (Table 3), much smaller than those generally used, explaining the unreliability of published results. For further discussion of P -values and their relationship with Bayes factors and posterior probabilities see APPENDIX B.

Overestimation of effects—selection bias: A related problem is overestimation 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 reusing the same data to estimate the size of an effect as was used to select a marker results in an upward bias in the size of effects that can be large, in percentage terms. The solution is either to 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. 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.

Other problems: The interpretation of P -values as representing good evidence for real effects is one, although 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 “stable” across different environments or subpopulations 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 and then obtain good evidence for an interaction.

This article quantifies the level of replication required for association studies, for reliably detecting linkage disequilibrium 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 *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 biallelic marker and a biallelic 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 biallelic 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 assume a biallelic marker with alleles M, m and a biallelic QTL with alleles A, a . Following LUO (1998), let q, p be the probability of A, M , respectively. In the biallelic case, linkage disequilibrium is specified by a single coefficient, D , such that the joint probabilities of alleles are given by

$$\Pr(A, M) = \Pr(A)\Pr(M) + D = qp + D, \quad (1)$$

$$\Pr(A, m) = \Pr(A)\Pr(m) - D = q(1 - p) - D \quad (2)$$

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

$$Q = \Pr(A|M) = q + D/p, \quad (3)$$

$$R = \Pr(A|m) = q - D/(1 - p). \quad (4)$$

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

Statistical models: The statistical model for observed phenotypes is

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

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 are analyzed with a one-way analysis-of-variance model, with marker genotypes as groups, *i.e.*,

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

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

Classical power calculation: The classical statistical experimental design seeks to determine designs with a given “power” $\mathcal{P}$ at a given significance level α ; *i.e.*, the probability of obtaining a result with P -value less than α is $\mathcal{P}$, assuming the effect to be detected is at least a

TABLE 1
Expected genotypic frequencies and phenotypic values from LUO (1998)
for QTL/marker genotype combinations

QTL	Marker:	Frequencies (f_{ij})			Expected value
		MM	Mm	mm	
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		$p^2(1-Q)^2$	$2p(1-p)(1-Q)(1-R)$	$(1-p)^2(1-R)^2$	$-d$

Marker genotypes are MM , Mm , mm ; QTL genotypes are AA , Aa , aa .

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,

$$F > F_{1-\alpha;v_1,v_2}, \quad (7)$$

where $F_{1-\alpha;v_1,v_2}$ is the $1 - \alpha$ point of the F -distribution on v_1 , v_2 d.f.

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 biallelic QTL in linkage disequilibrium with a marker with parameters D , d , h , p , and q defined above.

Under the alternative hypothesis F is distributed as a noncentral F -distribution with noncentrality parameter δ given by

$$\delta = \frac{\text{EMS}_b}{\text{EMS}_w} \frac{v_1(v_2 - 1)}{v_2} - v_1, \quad (8)$$

where EMS_b and EMS_w are the expected values of the mean squares between and within marker classes (*cf.* Table 2; JOHNSON and KOTZ 1970, p. 189; LUO 1998, Equation 6.2). The power is given by

$$\Pr(F_{v_1,v_2}(\delta) > F_{1-\alpha;v_1,v_2}), \quad (9)$$

where $F_{v_1,v_2}(\delta)$ is a random variable with the F -distribution with v_1 , v_2 d.f., and noncentrality parameter δ .

To determine δ and complete the calculation, it remains to determine the expected mean squares EMS_b and 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 A.

Experimental design with power to obtain a given Bayes factor: Bayes' theorem follows from the basic law of conditional probability:

$$\Pr(A|B)\Pr(B) = \Pr(A \cap B) = \Pr(B|A)\Pr(A) \quad (10)$$

or

$$\Pr(A|B) = \Pr(B|A)\Pr(A)/\Pr(B). \quad (11)$$

If we have observed B the probability of A being true is $\Pr(A|B)$. 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

$$g(\theta|y)f(y) = f(y|\theta)\pi(\theta). \quad (12)$$

It follows (by integrating both sides over θ) that

$$f(y) = \int f(y|\theta)\pi(\theta)d\theta \quad (13)$$

$$g(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta)d\theta}. \quad (14)$$

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* $g(\theta|y)$ 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 nonzero 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 $\pi_i(\theta_i)$, $f_i(y|\theta_i)$, $g_i(\theta_i|y)$, and $f_i(y)$ be

TABLE 2
ANOVA table for single-marker analysis

	d.f.	SS	MS	F
Between marker classes	$v_1 = 2$	SS_b	$\text{MS}_b = \text{SS}_b/v_1$	$F = \text{MS}_b/\text{MS}_w$
Within marker classes	$v_2 = n - 3$	SS_w	$\text{MS}_w = \text{SS}_w/v_2$	

TABLE 3
***P*-values corresponding to various Bayes factors, for testing for linkage disequilibrium**
between a biallelic marker and QTL

<i>n</i>	Bayes factor (<i>B</i>)						
	$\frac{1}{20}$	$\frac{1}{10}$	$\frac{1}{5}$	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}

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 $\Pr(y|H_i)$. Additionally let $\pi(H_i)$ denote the prior probabilities for each model to be the “true” 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 nonzero 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.,

$$B = \frac{\Pr(y|H_1)}{\Pr(y|H_0)}. \quad (15)$$

Higher Bayes factors (>1) are stronger evidence for H_1 , while lower Bayes factors (<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

$$\Pr(H_1|y)/\Pr(H_0|y) = B \times \pi(H_1)/\pi(H_0), \quad (16)$$

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

$$\text{posterior odds} = \text{Bayes factor} \times \text{prior odds}. \quad (17)$$

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 ($100 \times \frac{1}{10} = 10$). This may not sound impressive to readers accustomed to *P*-values <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 noninformative prior distribution

on the effect and sample size >100 , a *P*-value of 0.05 can correspond to a Bayes factor <1 , i.e., little or no evidence against H_0 or even evidence *for* H_0 . (cf. BERGER and BERRY 1988 and Table 3).

Note that the reader may find 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 that puts most prior mass on very large numbers should also not be used naively—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

$$B(y) = B_0(y)/B_0(y_a), \quad (18)$$

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 and so contains little or no evidence either way, which is

TABLE 4
Comparison with results from LUO (1998)

Population	n	p	q	D	h_Q^2	ϕ	$\mathcal{P}_{0.05}$	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

Results are shown for the 12 example populations (*cf.* LUO 1998, Tables 2 and 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 . $\mathcal{P}_{0.05}$ 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.

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,

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

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

Note that our definition of B is the reciprocal of that used by Spiegelhalter and Smith, so high Bayes factors correspond to positive evidence for H_1 .

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 $\mathcal{P}$ to achieve a given Bayes factor B we solve for F in (19) and select a design using the existing deterministic power calculation.

Note that in general the Bayes factor is sensitive to the prior for the variable(s) being tested (here marker effects). Use of (19) does not require specification of a prior, which is especially useful in a general experimental design context where there are no specific data and hence no prior. Intuitively, it seems reasonable that the above calibration is equivalent to using a nearly non-informative prior with information comparable to that in the small training sample. This is consistent with our experience—when we have used priors with information equivalent to one data point, the resulting Bayes factors were similar to (19).

Equation 19 applies to testing for linkage disequilibrium (*cf.* the ANOVA table, Table 2), where $m = 3$ is

the number of marker classes; n_1 , n_2 , and 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 , and n_3 to their expected values on the basis of the frequencies of marker classes MM , Mm , and mm in Table 1 and the total sample size:

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

Substituting in (19) gives

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

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 , and q , sample size n , and Bayes factor B , we can solve for $F = F_{1-\alpha, 2, n-2}$ in (21) and then look up the value of α from the F -distribution. Then given α and n we determine the power, $\mathcal{P}$, from the classical power calculations. To determine the sample size n for given B and $\mathcal{P}$ we use interpolation.

An R package (ldDesign) containing code used for the calculations is available on the Comprehensive R archive network (CRAN), from <http://cran.r-project.org/src/contrib/Descriptions/ldDesign.html>.

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 size 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 < 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 4 is a comparison with results for 12 sample populations from LUO (1998). $\mathcal{P}_{0.05}$ is the power to detect an effect with comparisonwise significance level $\alpha = 0.05$, as shown in LUO's (1998) 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 < 1 , *i.e.*, not corresponding to positive evidence for a real effect. The sample sizes $n_{B_{20}}$ are the sample

FIGURE 1.—Power *vs.* disequilibrium for QTL with $h_Q^2 = 0.01, 0.05$.

sizes required to have good power to detect an effect with fairly strong evidence and are substantially (up to 13 times) larger.

Note that our calculations for the power, $\mathcal{P}_{0.05}$, to detect an effect with significance level $\alpha = 0.05$ agree with those from Luo's (1998) Table 3 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 (1998). Luo obtained powers of 0.65, 0.60, which are similar to the power obtained when $\phi = 0$. It may be that he used $\phi = 0$ for the power calculations for these populations.

Figure 1 shows graphs of power *vs.* linkage disequilibrium. Graphs correspond to sample sizes $n = 600, 1200, 2400, 4800$ and QTL heritabilities (or proportions of phenotypic variance explained) of $h_Q^2 = 0.01, 0.05$. Within each graph, the solid lines are graphs of power *vs.* linkage disequilibrium for Bayes factors of $\frac{1}{20}, \frac{1}{10}, \frac{1}{5}, 1, 5, 10$, and 20. The lines, in order of increasing power, correspond to the Bayes factors in decreasing order.

Graphs of power *vs.* 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* do not 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.* top row, middle graph 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 $p = 0.5, q = 0.01$).

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 that 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) and 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 to 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 and marker of up to 3 kb, corresponding to a map with 6 kb spacing and 500,000 markers.

For example, if it is desired to localize an effect to the nearest of 500,000 SNP markers and ~ 10 genes were expected, the prior probability per marker would be $\sim 1/50,000$. 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 localizing 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 ± 3 kb of a given marker. Note that, in relative terms, the sample sizes to obtain the higher Bayes factors do not increase much at the larger sample sizes.

Use with candidate genes: If a large number of "candidate" genes were being tested, at an initial screening stage of experimentation, one may be content with posterior odds of < 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 ($h_Q^2 = 0.01$). From Figure 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

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 $n = 600, 1200, 4800, 9600, 19200, 38400$. Power curves: (—) $q = 0.01$, (---) $q = 0.05$, ($\cdots$) $q = 0.1$, ($- \cdot -$) $q = 0.5$. The curves shown are for the power to detect a QTL *vs.* D' , with $h^2_Q = 0.05$ and a Bayes factor of 20.

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 favorable: 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 $<1/1000$ 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 ± 3 kb 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 5

Sample sizes required for power of 0.9 of detection of linkage disequilibrium between a biallelic QTL and a biallelic 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

Posterior odds	Bayes factor	Sample size	
		$h^2_Q = 0.05$	$h^2_Q = 0.01$
$1/20$	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

Prior probability per marker is assumed to be $1/50,000$.

DISCUSSION

Detection of markers in linkage disequilibrium with a trait is a complex statistical problem. Classical single-marker test methods have not led 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 article 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.

TABLE 6

Sample sizes required for power of 0.9 of detection of linkage disequilibrium between a biallelic QTL and a biallelic 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

Posterior odds	Bayes factor	Sample size	
		$h_Q^2 = 0.05$	$h_Q^2 = 0.01$
$1/_{20}$	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

Prior probability per marker is assumed to be $1/5000$.

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 disequilibrium 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 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 neighborhood of a gene generally vary wildly, in a nonsmooth fashion. To detect a gene, particularly one with a low-frequency allele, it is important not only 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 and 40,000, to detect and locate QTL that 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 neighboring markers, justifying higher prior probabilities when estimating the marginal probability for a set of markers in a larger region (*cf.* BALL 2001). This cannot 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 usable 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; BOEHNEKE and LANGEFELD 1998; SPIELMAN and EWENS 1998; LONG and LANGLEY 1999). 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 can be rederived. We would expect that a transmission test of equivalent power (in the classical sense) to the designs considered here would 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 biallelic marker for which the methods of this article 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 were 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 that avoids this prior dependence, giving a generic result that can be applied in the absence of further prior information. In specific situations it may be possible to do better when 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 data set; 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 does not 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 maximizes over the unknown parameters,

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

where $\hat{\theta}_1$ and $\hat{\theta}_0$ are chosen to maximize their respective likelihood functions $f_1(y|\theta_1)$, $f_0(y|\theta_0)$. This strongly favors the alternative hypothesis, which usually has one or more additional parameters to maximize 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 comparisonwise or experimentwise *P*-values or *P*-values from a permutation test. A permutation test is just a nonparametric way of obtaining a *P*-value. Of course the experimentwise *P*-value usually corresponds to a much lower comparisonwise *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 larger 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 of 0.05, which may represent very weak evidence—"you get (only) 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 underpowered, 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.

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 and an NSOF grant from the New Zealand Forest Research Institute.

LITERATURE CITED

- ALLISON, D. B., 1997 Transmission disequilibrium tests for quantitative traits. *Am. J. Hum. Genet.* **60**: 676–690.
- ALTSHULER, D., J. N. HIRSCHHORN, M. KLANNEMARK, C. M. LINDGREN, M. C. VOHL *et al.*, 2000 The common PPAR γ Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat. Genet.* **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. *Philos. Trans. R. Soc.* **53**: 370–418.
- BENJAMINI, Y., and Y. HOCHBERG, 1995 Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**: 289–300.
- BERGER, J., and D. BERRY, 1988 Statistical analysis and the illusion of objectivity. *Am. Sci.* **76**: 159–165.
- BERGER, J. O., and T. SELLKE, 1987 Testing a point null hypothesis: the irreconcilability of *P* values and evidence (with discussion). *J. Am. Stat. Assoc.* **82**: 112–139.
- 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.
- EDWARDS, W., H. LINDMAN and L. J. SAVAGE, 1963 Bayesian statistical inference for psychological research. *Psych. Rev.* **70**: 193–242.
- 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.
- 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.
- 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., 2001 Can SNPs deliver on susceptibility genes? *Science* **293**: 593–595.
- HAMPTON, T., 2000 Research brief. Focus, September 29, 2000. Harvard University, Cambridge, MA.
- JOHNSON, N. L., and N. KOTZ, 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. *Nat. Genet.* **22**: 139–144.
- KUMAR, S., R. J. SPELMAN, D. J. GARRICK, T. E. RICHARDSON, M. LAUSBERG *et al.*, 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 Res.* **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 *et al.*, 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**: 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. *Am. J. Hum. Genet.* **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 & Sons, New York.
- SELLKE, T., M. J. BAYARRI and J. BERGER, 2001 Calibration of P -values for testing precise null hypotheses. *Am. Stat.* **55**: 62–71.
- SPIEGELHALTER, D., and A. F. M. SMITH, 1982 Bayes factors for linear and log-linear models with vague prior information. *J. R. Stat. 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.
- STOREY, J. D., 2003 The positive false discovery rate: a Bayesian interpretation and the q -value. *Ann. Stat.* **31**: 2013–2035.
- TAILLON-MILLER, P., I. BAUER-SARDINA, N. L. SACCONI, 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*, edited by R. D. BURDON and J. M. MOORE. FRI Bulletin No. 203, Rotorua, New Zealand.

Communicating editor: J. B. WALSH

APPENDIX A: DERIVATION OF MODIFIED VALUES FOR EMS_b AND EMS_w

First we calculate μ , β_2 , and then ω_{23} . Values for β_1 , β_3 , and other ω_{ij} values are obtained similarly. Taking expectations in Equation 5,

$$\mu = E(y) = E(y|AA)Pr(AA) + E(y|Aa)Pr(Aa) + E(y|aa)Pr(aa). \quad (A1)$$

Substituting QTL genotype probabilities and values and simplifying gives

$$\mu = d \times q^2 + h \times 2q(1 - q) - d \times (1 - q)^2 = (2q - 1)d + 2q(1 - q)h. \quad (A2)$$

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

$$\mu + \beta_2 = E(y|Mm) = E(y|Mm, AA)Pr(AA|Mm) + E(y|Mm, Aa)Pr(Aa|Mm) + E(y|Mm, aa)Pr(aa|Mm). \quad (A3)$$

The genotype conditional probabilities are given by

$$Pr(AA|Mm) = Pr(A|M)Pr(A|m) = QR, \quad (A4)$$

$$Pr(Aa|Mm) = Pr(A|M)Pr(a|m) + Pr(A|m)Pr(a|M) = Q(1 - R) + (1 - Q)R = Q + R - 2QR, \quad (A5)$$

$$Pr(aa|Mm) = Pr(a|M)Pr(a|m) = (1 - Q)(1 - R). \quad (A6)$$

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

$$\beta_2 = \frac{[D(1 - 2p)]d + d[2D + (1 - 2p)(1 - 2q)]h}{p(1 - p)}, \quad (A7)$$

which agrees with Equation 3.2 of LUO (1998).

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

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

Solving for ω_{23} gives

$$\begin{aligned} \omega_{23} &= -d - \mu - \beta_2 \\ &= -d - (2q - 1)d - 2q(1 - q)h - \frac{D(1 - 2p)d + [2D^2 + D(1 - 2p)(1 - 2q)]h}{p(1 - p)} \\ &= -\frac{[D(1 - 2p) - 2pq(1 - p)]d + [2D^2 + (1 - 2p)(1 - 2q)D + 2pq(1 - p)(1 - q)]h}{p(1 - p)}. \end{aligned} \quad (A8)$$

This differs from the expression of LUO (1998, p. 207) by a factor -1 .

LUO (1998, Equations 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 proportion of the i th marker genotype and f_{ij} is the proportion of the i th marker genotype/ j th QTL genotype combination.

The formula given for EMS_w is

$$EMS_w = \frac{1}{n - G_m} \left\{ \sum_{i=1}^{G_m} \sum_{j=1}^{G_q} f_{ij} n \omega_{ij}^2 - \sum_{i=1}^{G_m} \frac{1}{f_i} \left[\sum_{j=1}^{G_q} f_{ij} (1 + (n - 1)f_{ij}) \omega_{ij}^2 + 2(n - 1) \sum_{j < k \leq G_q} f_{ij} f_{ik} \omega_{ij} \omega_{ik} \right] \right\} + \sigma^2, \quad (A9)$$

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

Use of the formulas for EMS_b and EMS_w from Luo gave results that 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 rederive 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),

$$\text{SS}_T = \text{SS}_b + \text{SS}_w = (G_m - 1)\text{MS}_b + (n - G_m) \text{MS}_w.$$

Taking expectations,

$$V_p = E(\text{SS}_T) = (G_m - 1)\text{EMS}_b + (n - G_m)\text{EMS}_w,$$

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

$$\text{EMS}_b = \frac{1}{G_m - 1} [(n - 1)V_p - (n - G_m)\text{EMS}_w]. \quad (\text{A10})$$

Finally, we need an expression for V_p . Substituting from Equation 5 and taking variances gives

$$\begin{aligned} V_p &= \text{var}(y) = \text{var}(\mu + \beta_i + \omega_{ij} + e_{ijk}) = \text{var}(\mu + \beta_i + \omega_{ij}) + \text{var}(e_{ijk}) \\ &= E(\beta_i + \omega_{ij})^2 + \sigma^2 = \sum_{i=1}^{G_m} \sum_{j=1}^{G_q} f_{ij}(\beta_i + \omega_{ij})^2 + \sigma^2. \end{aligned} \quad (\text{A11})$$

APPENDIX B: BAYES FACTORS, POSTERIOR PROBABILITIES, TYPE I ERROR RATES, *P*-VALUES, POWER, AND FALSE DISCOVERY RATES AS MEASURES OF EVIDENCE FOR GENE EFFECTS

To fix notation, suppose H_0 represents the null hypothesis that $\theta = 0$ (a so-called “precise” hypothesis), and H_1 represents the alternative hypothesis that $\theta \neq 0$, where θ represents one or more parameters for a genetic effect to be tested, which are zero under the assumption of no effect. There may be other parameters common to both H_0 and H_1 ; these are suppressed.

Type I error rates and *P*-values: The type I error rate (α) for a test is given by

$$\alpha = \Pr(T \geq T_c | H_0), \quad (\text{B1})$$

where T is a “test statistic,” and T_c is the α -critical level of the test statistic (which we also write as T_α), such that H_0 is rejected if $T \geq T_c$.

Type I error rates are valid if α is chosen preexperimentally. However, α , being chosen preexperimentally, does not make full use of the information in the data, which is not efficient. Fisher proposed reporting the *P*-value, which is the value of α when we set $T_c = T_{\text{obs}}$, the observed value of the test statistic:

$$P = \Pr(T \geq T_{\text{obs}} | H_0). \quad (\text{B2})$$

The *P*-value is the smallest, *i.e.*, most optimistic, value of α under which H_0 *would have been* rejected. However, we cannot set $\alpha = p$ after observing the data—the *P*-value is not a valid type I error rate. Therefore it is problematic to make a decision on the basis of one or more *P*-values. Even if we had chosen α preexperimentally, giving an average type I error rate under repeated sampling, we would not be able to quantify the evidence for a particular sample that may be worse than average.

Bayes factors as evidence and relationships with *P*-values: Bayes factors have a natural interpretation as evidence from Equation 16. Higher Bayes factors correspond to stronger evidence for H_1 .

Intuitively, more extreme values of the test statistics mean that the data are more likely than those for smaller values to have come from H_1 ; *i.e.*, smaller *P*-values correspond to stronger evidence. Experience confirms this: for a given problem (H_0 , H_1 , and sample size) with an intelligently defined test statistic, a plot of *P*-value *vs.* the Bayes factor shows a monotone relationship (cf. Table 3). This means, for example, that *P*-values and Bayes factors would give the same ranking of genes in order of strength of evidence but does not imply Bayes factors and *P*-values are equivalent; how to rank genes in order of expected benefit (considering size of gene effects as well as strength of evidence) or how many putative genes to select for further investigation becomes clear only when Bayes factors or posterior probabilities are calculated.

The relationship between Bayes factors and *P*-values is quantified implicitly in Equation 19 from SPIEGELHALTER and SMITH (1982) for the one-way analysis-of-variance model, where the *F*-value corresponds to the *P*-value through

the F -distribution. This relationship varies in general, depending on sample size (which is apparent from Equation 19) and on problem setup.

More generally SELLKE *et al.* (2001) give “calibrations”; *e.g.*, their “simple calibration” is given as

$$1/B \geq -ep \log(p) \quad (p \leq e), \quad (\text{B3})$$

leading to a fairly general lower bound,

$$\alpha(p) \geq (1 + [-ep \log(p)]^{-1})^{-1} \quad (\text{B4})$$

for a valid postexperimental type I error rate, under a technical assumption, that the distribution of $-\log(p)$ under H_1 has a decreasing failure rate, which is shown to hold for a normal testing example with iid data. Unlike the problem-specific formula from Spiegelhalter and Smith, these bounds are not exact, and so we do not use them for experimental design.

Making a decision on the basis of a P -value amounts to *conditioning on* (*i.e.*, simplifying by replacing the data by) the event

$$E_p = \{x \in S: T(x) \geq T_{\text{obs}}\} \quad (\text{B5})$$

occurring, where S is the sample space, *i.e.*, conditioning on data we did not observe, which is an error (*cf.* BERGER and BERRY 1988). Strictly speaking we did not observe E_p but

$$E_o = \{x \in S: T(x) = T_{\text{obs}}\}. \quad (\text{B6})$$

Similarly, making a decision on the basis of statistical significance for a predetermined value α amounts to conditioning on

$$E_\alpha = \{x \in S: T(x) \geq T_\alpha\}. \quad (\text{B7})$$

We can attempt to calculate a posterior probability for H_0 using only the information $T \geq T_\alpha$, *i.e.*, that the event E_α occurred. We have

$$\Pr(H_0|E_\alpha)\Pr(E_\alpha) = \Pr(E_\alpha|H_0)\Pr(H_0), \quad (\text{B8})$$

by Bayes' rule, where $\Pr(H_0)$ is the prior probability that H_0 is true, and

$$\Pr(E_\alpha) = \Pr(E_\alpha|H_0)\Pr(H_0) + \Pr(E_\alpha|H_1)\Pr(H_1). \quad (\text{B9})$$

Combining (B8) and (B9), setting $\pi_0 = \Pr(H_0) = 1 - \Pr(H_1)$, $\alpha = \Pr(E_\alpha|H_0)$, and solving gives

$$\Pr(H_0|E_\alpha) = \left(1 + \frac{\Pr(E_\alpha|H_1)(1 - \pi_0)}{\alpha\pi_0}\right)^{-1}. \quad (\text{B10})$$

Note the dependence on $\Pr(E_\alpha|H_1)$. To obtain small values for $\Pr(H_0)$ and hence good evidence for a real effect we need

$$\Pr(E_\alpha|H_1) = \Pr(T \geq T_\alpha|H_1) \gg \Pr(T \geq T_\alpha|H_0) = \alpha, \quad (\text{B11})$$

which may not be the case, since the left-hand side of the inequality is not controlled solely by α . The left-hand side of (B11) is related to the *power* of the test. Specifically, let $H_1(d)$ represent the hypothesis $\theta = d$ and $\pi(\cdot)$ be the prior for θ in H_1 . Then

$$\mathcal{P}_\alpha(d) = \Pr(T \geq T_\alpha|H_1(d)) \quad (\text{B12})$$

is the power function, giving the power to detect an effect of size d , at significance level α . $\Pr(E_\alpha|H_1)$ can be represented as an integral of the power function,

$$\Pr(E_\alpha|H_1) = \int \mathcal{P}_\alpha(d)\pi(d)dd, \quad (\text{B13})$$

giving the posterior probability of H_0 ,

$$\Pr(H_0|E_\alpha) = \left(1 + \int \mathcal{P}_\alpha(d)\pi(d)dd \times \frac{(1 - \pi_0)}{\alpha\pi_0}\right)^{-1}. \quad (\text{B14})$$

A corresponding Bayes factor is given by

$$B_\alpha = \frac{\Pr(T \geq T_\alpha | H_1)}{\Pr(T \geq T_\alpha | H_0)} = \frac{1}{\alpha} \int \mathcal{P}_\alpha(d) \pi(d) dd. \quad (\text{B15})$$

The left-hand side of (B14) is equivalent to the positive false discovery rate (STOREY 2003; Equation B20 below). Thus (B14) and (B15) show the relationship between the prior and posterior probabilities, a Bayes factor for rejected null hypotheses, type I error rates (α), power curves, and positive false discovery rates. For the same reason that reporting α is not fully efficient, using B_α (or the false discovery rate) instead of the Bayes factor or posterior probability based on the observed data (or $T = T_{\text{obs}}$) is also not fully efficient as a summary of the observed data.

Note that the integral in (B13), *i.e.*, the Bayesian approach involving $\pi(d)$, is needed because d (the parameter being tested) is unknown. Since the power curve (B12) is monotonic, a non-Bayesian approximate upper bound for $\Pr(H_0 | E_\alpha)$ could be obtained if a lower bound, d_1 , for d was known. Suppose we had such a d_1 and $d_1 > 0$ (say); then

$$\Pr(H_0 | E_\alpha) \leq \left(1 + \mathcal{P}_\alpha(d_1) \times \frac{(1 - \pi_0)}{\alpha \pi_0} \right)^{-1}. \quad (\text{B16})$$

Of course we have no such bound *a priori*. A natural approach would be to let d_1 be the lower limit of a confidence interval (*e.g.*, 95 or 99%) for $\hat{d}$, so the bound holds to within a reasonably small error. The bound would not be useful unless the power $\mathcal{P}_\alpha(d_1)$ was reasonably good, so we need d_1 to be at least $2\sigma(\hat{d})$ if $\alpha \leq 0.95$ or $3\sigma(\hat{d})$ if $\alpha \leq 0.99$, etc., where $\sigma(\hat{d})$ is the standard error of the estimate of $\hat{d}$. As an example if $\alpha = 0.05$, $\pi_0 = 0.5$, and $d_1 = 2\sigma(\hat{d})$, so that $\mathcal{P}_\alpha(d_1) \approx 0.5$, then

$$\Pr(H_0 | E_\alpha) \leq \left(1 + 0.5 \times \frac{(1 - 0.5)}{0.5 \times 0.05} \right)^{-1} = \frac{1}{11} \approx 0.09. \quad (\text{B17})$$

Thus we obtain, approximately, that $\Pr(H_0 | E_\alpha) \leq 0.09$ if

$$\hat{d} \geq 4\sigma(\hat{d}). \quad (\text{B18})$$

Note that this requires an estimated effect of twice the usual $2\sigma(\hat{d})$ recommendation corresponding to $P = 0.05$, to get a posterior probability of just < 0.1 . By comparison, if the assumptions for SELLKE *et al.* (2001), Equations B3 and B4, hold, the P -value corresponding to a valid postexperimental $\alpha = 0.05$ or $\Pr(H_0 | E_\alpha) = 0.05$ is 0.0034, which corresponds to a $2.9\sigma(\hat{d})$ -sized effect.

Problems with the interpretation of P -values as evidence have been pointed out in a number of articles (*e.g.*, EDWARDS *et al.* 1963; BERGER and SELLKE 1987; BERGER and BERRY 1988). Edwards *et al.* showed that the evidence corresponding to a given P -value was not as strong as most people think. The later work by Berger and co-authors shows that the problems with P -values, especially the common choice of $P = 0.05$, are not specific to the choice of prior or to unusual examples but are quite general. For example, Berger and Berry consider the problem of testing for a biased coin. The example has $P = 0.05$, yet the minimum value of $\Pr(H_0)$, assuming $\pi_0 = 0.5$ over a wide class of priors for the probability of heads, is 0.21. Results of this article show similar problems for common analysis-of-variance models (Table 3). For a long time these warnings have largely gone unheeded by the scientific community.

False discovery rates and q -values: Another alternative to p -values is the “false discovery rate” (FDR) (BENJAMINI and HOCHBERG 1995), which has become popular for handling multiple comparisons in genomics. Motivated by the low power obtained when using experimentwise error rates and many spurious significant results obtained when using comparisonwise error rates due to the large number of multiple comparisons being made in genome scans, with the false discovery rate one attempts to control the proportion of false positives among the rejected hypotheses. STOREY (2003) considers a modification of the FDR, namely the positive false discovery rate (pFDR),

$$\text{pFDR} = E(V/R | R > 0), \quad (\text{B19})$$

where V is the number of false positives (H_0 is true, but rejected) and R is the number of rejected null hypotheses (*e.g.*, $T > T_\alpha$). The pFDR avoids a technical problem with V/R when $R = 0$. This is shown to have a Bayesian interpretation,

$$\text{pFDR} = \Pr(H_0 | T \geq T_\alpha). \quad (\text{B20})$$

Note that the right-hand side of (B20) is the same as (B14).

The Bayes error (posterior probability of making a wrong decision for a given T_α) is shown to be a combination of pFDR and the analogous positive false negative discovery rate (pFNR). Storey also defines the “ q -value”: analogous to the p -value being the lowest α under which a hypothesis would be rejected, the q -value is the lowest pFDR consistent with the data. This optimistic choice (analogous to the definition of the p -value) means that the q -value

has some if not all the limitations of the p -value—there is no valid interpretation of the q -value as an error rate (for false discoveries).

From a Bayesian perspective the use of false discovery rates is a step in the right direction—since using the frequentist FDR is approximately similar to using the Bayesian posterior probability. For example, the expected proportion of false positives among selected genes is simply the average posterior probability for the genes. These concepts are, however, superfluous within the Bayesian paradigm and we would not advocate using them unless for convenience, *e.g.*, if they had already been calculated or the Bayesian model was difficult to fit. In the Bayesian paradigm there is no problem with multiple comparisons and no need to control the FDR—*e.g.*, we compute posterior probabilities for each gene, giving the information needed to maximize expected utility directly. The FDR gives the properties of a procedure randomly, choosing a gene from a set of genes, corresponding to rejected null hypotheses, and ignoring the individual gene statistics, while the posterior probability gives the actual probability for each specific gene, which is clearly more informative.

Additionally, to estimate the FDR in the non-Bayesian framework, the prior probability for the proportion of true hypotheses needs to be estimated. This is where the procedure potentially loses efficiency compared to a full Bayesian model.

Rejecting or accepting H_0 is not the only option: Finally, we note that it is not always necessary to “reject” or “accept” H_0 , *i.e.*, to choose only one of H_0 , H_1 and act as if the chosen model is the true model. As in BALL (2001), inference on genetic architecture and avoidance of selection bias in estimates of gene effects can be achieved by considering multiple models according to their posterior probabilities.