MODELLING

Bayesian statistical methods for genetic association studies

Matthew Stephens and David J. Balding*[‡,§]

Abstract | Bayesian statistical methods have recently made great inroads into many areas of science, and this advance is now extending to the assessment of association between genetic variants and disease or other phenotypes. We review these methods, focusing on single-SNP tests in genome-wide association studies. We discuss the advantages of the Bayesian approach over classical (frequentist) approaches in this setting and provide a tutorial on basic analysis steps, including practical guidelines for appropriate prior specification. We demonstrate the use of Bayesian methods for fine mapping in candidate regions, discuss meta-analyses and provide guidance for refereeing manuscripts that contain Bayesian analyses.

Frequentist
A statistical school of thought in which inferences about unknowns are justified not with reference to probabilities for the inferred value, but on the basis of measures of performance under imaginary repetitions of the procedure that was used to make the inference.

*Departments of Statistics and Human Genetics, University of Chicago, Chicago, Illinois 60637, USA.
‡Department of Epidemiology and Public Health, Imperial College, St Mary's Campus, Norfolk Place, London W2 1PG, UK.
§Current address: Institute of Genetics, University College London, London WC1E 6BT, UK.
e-mails: mstephens@uchicago.edu; d.balding@ic.ac.uk
doi:10.1038/nrg2615

The usual (frequentist) approach to assessing evidence for a population association between genetic variants and a phenotype of interest is to compute a *p*-value for the null hypothesis (H_0) of no association. Despite their widespread use, *p*-values have a striking and fundamental limitation[1–3]: from a *p*-value alone it is difficult to quantify how confident one should be that a given SNP is truly associated with a phenotype. Indeed, the same *p*-value computed at different SNPs or in different studies can have different implications for the plausibility of a true association depending on the factors that affect the power of the test, such as the minor allele frequency (MAF) of the SNP and the size of the study. This is because the probability that a SNP with a given *p*-value is truly associated with the phenotype depends not only on how unlikely that *p*-value is under H_0 (which is the same for all tests) but also on how unlikely it is under the alternative hypothesis H_1 (which differs from test to test). For example, a *p*-value of 10^{-8} may seem to offer strong evidence against H_0, but if a test has very low power then such a *p*-value may be almost as unlikely under H_1 as under H_0 and therefore provide little evidence against H_0. One response to such concerns is to avoid performing low-powered tests, for example by discarding low-MAF SNPs. However, this approach is inadequate for solving the problem and carries the risk that detectable associations might be discarded.

Bayesian methods provide an alternative approach to assessing associations that alleviates the limitations of *p*-values at the cost of some additional modelling assumptions. For example, a Bayesian analysis requires explicit assumptions about effect sizes at truly associated SNPs. Because of computational constraints, Bayesian approaches were not widely used until about 15 years ago, since when they have become more prevalent in many areas of science, including genetics[4–9]. This advance is now extending to genetic association studies, as recent papers have shown practical and theoretical advantages of using Bayesian approaches for the assessment of association[10–20]. Several software packages (for example, SNPTEST[11] and BIMBAM[12,20]) now allow simple genome-wide Bayesian analyses to be performed easily and quickly on a standard desktop computer.

Many genetics researchers are currently unfamiliar with Bayesian methods, and some may be reluctant to adopt them because they fear that editors and reviewers will also be unfamiliar with them. However, we believe that the benefits of Bayesian methods will lead to their widespread use in future genetic association analyses. Bayesian methods compute measures of evidence that can be directly compared among SNPs within and across studies. In addition, they provide a rational and quantitative way to incorporate biological information, and they can allow for a range of possible genetic models in a single analysis. Moreover, Bayesian approaches allow a coherent approach to combining results across studies (meta-analysis), across SNPs in genes and across gene pathways, which will be increasingly important as we move from single-SNP analyses towards more integrative approaches.

In this Review, we present a guide for newcomers to understanding and implementing a Bayesian analysis in some of the most common settings. We focus particularly on the additional modelling assumptions

Population association
Also known as true association. An association between a SNP and a phenotype that is present in the population from which a sample is taken. A population association can arise owing to population structure, but for simplicity we assume here that this possibility has been eliminated (for example, by covariate adjustment) and hence that population associations are caused by a functional SNP, either directly or through linkage disequilibrium.

p-value
The probability, if the null hypothesis were true, that an imaginary future repetition of the study would generate stronger evidence for association than that actually observed. A *p*-value is conventionally interpreted as measuring the strength of evidence for association, but there is no simple relationship between a *p*-value and the probability that the association is genuine.

Power
For a given population association, the power of a statistical test is the probability that the null hypothesis is rejected under imaginary repetitions of the study.

Is it more common to assume the same pi for all SNPs? - see exchangeability

that are required by Bayesian approaches compared with standard frequentist analyses, and we attempt to provide practical guidelines for the most important of these assumptions. We hope that this will facilitate more interpretable analyses and more robust conclusions from future genetic association studies.

Calculating probabilities of association

We consider first the problem of computing, for each SNP in a genome-wide association (GWA) study, the probability that it is truly associated with phenotype. This posterior probability of association (PPA) can be thought of as the Bayesian analogue of a *p*-value obtained, for example, by using the Armitage trend test (ATT) or the Fisher exact test (for example, REF. 21). The calculation can be split into the three steps below.

Choose a value for π, the prior probability of H_1. The value of π at a SNP quantifies our belief, given current knowledge, that the SNP is associated with the phenotype in question. One can allow π to vary across SNPs — for example, it could vary depending on the MAF, proximity to genes of interest or conservation across species. However, if π is assumed to be the same for all SNPs, it can be interpreted as a prior estimate of the overall proportion of SNPs that are truly associated with a phenotype. Typically, only a minority of SNPs is expected to be truly associated with a given phenotype: the range 10^{-4} to 10^{-6} has been suggested[13] for π. The probability of H_0 is taken to be $1 - \pi$, which implicitly assumes that either H_0 or H_1 is true. Therefore, we should specify H_0 and H_1 so that they exhaust all realistic possibilities or else keep in mind that any results are conditional on this assumption.

Compute a Bayes factor for each SNP. A Bayes factor (BF) is the ratio between the probabilities of the data under H_1 and under H_0. The BF is similar to a likelihood ratio, but it compares two different models rather than two parameter values in a model. The observed data are BF times more likely under H_1 than under H_0, and so the larger the BF, the stronger the support in the data for H_1 compared with H_0. A BF of one indicates that the data are equally probable under the two hypotheses and hence offers no help in discriminating between them.

Calculate the posterior odds on H_1. The BF and π can be used to compute the posterior odds (PO) on H_1:

$$PO = BF \times \pi/(1 - \pi) \quad (1)$$

This can be used to calculate the PPA:

$$PPA = PO/(1 + PO) \quad (2)$$

See TABLE 1 for some numerical values of BF and π and the resulting PPA.

The PPA can be interpreted directly as a probability, irrespective of power, sample size or how many other SNPs were tested. It can be used in a further analysis involving the relative costs of false positives and false negatives to make an explicitly reasoned decision about which SNPs to pursue further[14].

Intuitively, the PPA combines the evidence in the observed association data (the BF) with the prior probability (π) that a SNP is truly associated with phenotype. Because π is typically so small, the BF has to be large (for example, $>10^4 - 10^6$) to provide convincing evidence for an association (that is, to give a PPA close to 1). This contrasts with many other scientific applications in which a BF of 10 can be considered as strong evidence[22] against H_0. The requirement for a large BF is analogous to setting a stringent threshold for genome-wide significance in a frequentist approach. However, in a Bayesian analysis the reason for requiring a large BF is the small number of SNPs that is expected to be truly associated and not the large number of tests that is actually or potentially performed (which is the usual argument for requiring low *p*-values). See BOX 1 for further discussion of these multiple-testing issues.

Because the PPA can easily be computed from the BF for any given π, in practice the BF is often used as the primary summary of the evidence for association at a SNP, which leaves open the choice of π. If π is chosen to be constant over SNPs, the BF gives the same ranking of SNPs as the PPA. In TABLE 1 the BFs also give the same ranking as the trend test *p*-values. In general BF and *p*-value rankings are similar except for low-MAF SNPs, and in BOX 2 we discuss assumptions under which the two rankings are almost identical, including low-MAF SNPs. It is also possible to translate a BF into a *p*-value by treating it as a classical test statistic and computing *p*-values by permutation[12] — this approach is sometimes referred to as a 'Bayes/non-Bayes compromise' (REF. 23). This can generate test statistics with good frequentist properties but does not solve the problem that the interpretation of the resulting *p*-value depends on factors that affect power.

Table 1 | **Frequentist and Bayesian summaries of evidence for association**

Trait	SNP*	*p*-values‡		$\log_{10}$(BF)§	PPA	
		Trend test	General test		$\pi = 10^{-4}$	$\pi = 10^{-5}$
BD	rs420259	2.2×10^{-4}	6.3×10^{-8}	4.1	0.56	0.11
CD	rs9858542	7.7×10^{-7}	3.6×10^{-8}	4.7	0.83	0.33
T2D	rs9939609	5.2×10^{-8}	1.9×10^{-7}	5.3	0.95	0.67
CD	rs17221417	9.4×10^{-12}	4.0×10^{-11}	8.9	0.99999	0.99987
T1D	rs17696736	2.2×10^{-15}	1.5×10^{-14}	12.5	1.00000	1.00000

The table shows *p*-values under two tests, a Bayes factor (BF) and corresponding posterior probability of association (PPA) for two values of the prior probability, π, for certain SNPs reported by the Wellcome Trust Case Control Consortium[13]. When BF ≈ $1/\pi$ (rows 1–3), the PPA, can depend sensitively on the researcher's prior level of scepticism (as shown by the difference in PPA depending on whether $\pi = 10^{-4}$ or $\pi = 10^{-5}$), and researchers may differ in their conclusions. When BF>>$1/\pi$, then PPA ≈ 1 (rows 4–5) and the change in π has little effect. BD, bipolar disorder; CD, Crohn's disease; T1D, type 1 diabetes; T2D, type 2 diabetes. * SNPs have been selected to illustrate particular points and not to be representative of the study results. ‡ Taken from REF. 13. § $BF = 0.8\ BF^a + 0.2\ BF^g$, in which BF^a and BF^g are the additive and general model BFs from REF. 13.

Calculating a Bayes factor

Calculating a BF requires assumptions that are similar to those required for performing a frequentist power analysis. These assumptions concern the effects on phenotype

Box 1 | **Multiple testing**

The usual frequentist rationale for setting the significance level for testing an individual SNP is based on controlling the probability of wrongly rejecting H_0 for at least one SNP, assuming that all SNPs follow H_0. This results in procedures (for example, the Bonferroni correction) that require more stringent significance thresholds as more tests are performed. These approaches can be criticized because the assumption of no true association at any SNP in the genome is highly implausible. Moreover, they can lead to undesirable consequences: investigators may refrain from performing additional analyses that could reveal interesting associations (for example, tests involving non-additive effects) because they fear the 'multiple testing' penalty that these additional analyses will confer on all tests. This contributes to a waste of scientific effort because expensive data are not thoroughly interrogated.

By contrast, the Bayesian analyses we outline here do not depend on the number of tests performed. It is helpful to distinguish two types of multiple testing.

The first type is multiple tests of the same null hypothesis against different alternative models (such as additive or dominant genetic models). In a Bayesian approach, a single Bayes factor (BF), and hence a single posterior probability of association (PPA), is obtained through a weighted average over the alternative models. This takes account of differences in power among tests (within the BF calculation) and of differences in plausibility among alternative models (through model weights). Frequentist approaches, such as a Bonferroni correction, have difficulty in taking account of either of these factors.

The second type is multiple tests of different null hypotheses, such as when testing many SNPs for association. In the Bayesian approach, the strength of the evidence for each SNP being associated with the phenotype (the BF) is weighed against its prior probability to compute the PPA for each SNP, without reference to the number of SNPs tested.

Those familiar with the Bonferroni correction may worry that, as more tests are performed, the expected number of false positive associations will increase. Although this is true, under reasonable assumptions the expected number of true positive associations will also increase and the ratio of true positives to false positives will remain roughly constant. Informally, if one cares about the false discovery rate (FDR), rather than the probability of making even one false discovery, then the number of tests is effectively irrelevant: what is relevant is the proportion of tests that are null.

As the above paragraph suggests, the Bayesian approach of reporting a PPA, which depends on the proportion of tests that are null (π) but not on the number of tests, has close connections with frequentist approaches that control the FDR[40]. Indeed, some frequentist approaches to controlling the FDR, which are popular in microarray analyses, do not involve explicit consideration of the number of tests but do involve treating π as a parameter to be estimated[41]. Estimating π in the context of genome-wide association (GWA) studies is usually harder than for microarray experiments, because π is usually much smaller for a GWA study than for a typical microarray experiment that is looking for differential expression between two conditions. However, the difficulty in assigning π does not change the logic that both the FDR and the PPA depend on its value.

Bayesian
A statistical school of thought that holds that inferences about any unknown parameter or hypothesis should be encapsulated in a probability distribution, given the observed data. Computing this posterior probability distribution usually proceeds by specifying a prior distribution that summarizes knowledge about the unknown before the observed data are considered, and then using Bayes' theorem to transform the prior distribution into a posterior distribution.

of different genotypes at a SNP under H_1. Note that the apparent effect sizes at a tested SNP, which we model here, could also be due to an ungenotyped functional SNP with different effect sizes. We focus here on binary (for example, case–control) phenotypes, for which effect sizes can be expressed in terms of odds ratios (ORs). For quantitative (continuous) phenotypes, the effect size parameters are usually differences between the genotype-specific phenotype means.

Let θ_{het} denote the logarithm (base e) of the OR between the heterozygote and the common homozygote. Let θ_{hom} denote the logarithm of the OR between rare and common homozygotes. The null hypothesis is:

$$H_0: \theta_{het} = \theta_{hom} = 0 \quad (3)$$

The general alternative, H_1, is that at least one of θ_{het} and θ_{hom} is non-zero. If we consider a precise alternative, for example:

$$H_1: \theta_{het} = t_1, \theta_{hom} = t_2 \quad (4)$$

in which t_1 and t_2 are known, then:

$$BF = \frac{P(data \mid \theta_{het} = t_1, \theta_{hom} = t_2)}{P(data \mid \theta_{het} = 0, \theta_{hom} = 0)} \quad (5)$$

However, in practice the values of θ_{het} and θ_{hom} under H_1 are unknown, and computing the numerator of the BF requires averaging (mathematically: integrating) over possible values for t_1 and t_2, which are weighted by their plausibilities before the association data were observed. These weights are referred to as the prior distribution for θ_{het} and θ_{hom} under H_1, and specifying this distribution is a crucial part of any Bayesian analysis, which we now consider in detail. In addition, there are usually nuisance parameters, such as the intercept or covariate parameters in logistic regression, that must have prior distributions assigned to them. However, we do not discuss these in detail because the choice of prior distribution for these parameters is generally not crucial (essentially because they are common to H_0 and H_1).

Genetic models. Specifying a prior distribution for θ_{het} and θ_{hom} under H_1 usually proceeds by first selecting one or more genetic models and then specifying effect size prior distributions under those models. We now review four types of genetic model — and their associated effect size prior distributions — for which the BF is easily calculated using software or simple formulae. We then offer suggestions on how these could be used and combined in practice. Note that the models discussed here can be used to design a study as well as to analyse it[14]. Before the study is conducted the PPA can be regarded as a random quantity, and the researcher can investigate its distribution under a chosen model for various choices of design parameters, such as sample size.

Model 1: Strict additive model. In this model, $\theta_{hom} = 2\theta_{het}$ and so there is only a single effect size parameter. The most common choice is a normal (Gaussian) prior distribution for θ ($= \theta_{het}$) with a mean of zero and a standard deviation of σ, which we denote as $N(0,\sigma)$. SNPTEST and BIMBAM can both compute BFs under this model (in BIMBAM, this is done by setting $\sigma_d = 0$). Both programs use a Laplace approximation (for example, REF. 11) and produce similar results. Alternatively, Wakefield[14] has derived a simple formula that generally agrees closely with SNPTEST and BIMBAM. The Wakefield approximate Bayes factor (WABF) against H_0 can be written:

$$WABF = \sqrt{\{V^2/(V^2+\sigma^2)\}} \exp(\sigma^2 Z^2/2(V^2 + \sigma^2)) \quad (6)$$

in which $Z = \hat{\theta}/V$, $\hat{\theta}$ is the maximum-likelihood estimate of θ and V is the standard deviation of $\hat{\theta}$. Z is the statistic for the Wald test, which is similar to the ATT and

Meta-analysis
The combination of the results of multiple scientific studies that address the same, or similar, hypotheses.

Posterior probability of association
The probability that a SNP is truly associated with a phenotype. The posterior probability of association depends on modelling assumptions that should be made explicit in a careful analysis.

Likelihood ratio
The ratio of the probabilities of the observed data for two different values of the unknown parameter(s) under a given statistical model.

also to the likelihood ratio test in logistic regression. A convenient feature of the WABF is that $\hat{\theta}$ and V are usually available from the output of a standard frequentist analysis, which allows a classical analysis to be easily converted into a Bayesian analysis under this model (see the *SLCO1B1* example below).

Model 2: Strict dominant or recessive models. These are both single-parameter models that are obtained by setting $\theta_{het} = \theta_{hom}$ for dominant models and $\theta_{het} = 0$ for recessive models. The methods for strict additive models can also be used to compute BFs for strict dominant or recessive models by using a $N(0,\sigma)$ prior distribution on θ_{hom}. For example, to fit a dominant model, one can simply code all the rare homozygotes as heterozygotes and then apply the methods for additive models.

Model 3: General models centred on additivity. Both SNPTEST and BIMBAM include options to relax the additivity assumption and still place the most weight on parameter values that are close to additive; a user-specified value controls the concentration of these prior distributions around additive models. It would be straightforward to similarly consider models centred on dominance or recessivity, but this does not yet seem to have been implemented. Unlike the additive model, dominant and recessive models imply that only one genotype has a different effect from the other two. Because of this, apparent recessive and dominant effects may be more likely to reflect a genotyping anomaly than apparent additive effects, which may partly explain the focus on additive models in current practice. However, as genotyping quality has improved, this rationale has become less compelling.

What would a prior look like for these last two models? Still normal?

Model 4: General models not centred on additivity. For binary data, REF. 21 proposed a general model that is not based on logistic regression but directly models the case–control counts for each genotype; we use pBF (prospective BF) to denote the corresponding BF. Under H_1 the case–control proportions for each of the three genotypes are assumed to be independent, whereas under H_0 these proportions are equal. For purposes of illustration, only uniform prior distributions for the genotype proportions were previously reported[21], and in Supplementary information S1 (box) we introduce more realistic and efficient prior distributions based on the beta distribution. Adjusting the beta prior parameters allows one to control the effect size distribution under H_1 but does not allow near-additive effects to be weighted more than far-from-additive effects. It is not easy to include covariates in this model, which is straightforward under regression models.

The models described above are all prospective because they treat the genotypes as fixed and the phenotypes as random observations. For case–control studies, retrospective models (which fix phenotypes and treat genotypes as random) should be used in principle. However, in many practical settings the two approaches give the same results[24], and the more convenient prospective model is often used for case–control data (for example, REF. 13). A retrospective BF (rBF), which is analogous to the pBF, is described in Supplementary information S1 (box).

Each of the above models and associated prior distributions has its limitations, and it may be advantageous to combine two or more of them. For example, although the additive model is attractive in its simplicity, we know that not all functional SNPs act additively, and an analysis based on an additive model alone risks missing truly associated SNPs with non-additive effects. Fortunately, combining models in a Bayesian analysis is straightforward: the overall BF is the weighted average of the BFs computed under each model. (As the pBF and rBF are not based on logistic regression, they do not use exactly the same H_0 as the other models and so averaging, for example, the WABF and the pBF is not strictly valid; however, the two H_0 can be chosen to be similar and so this is unlikely to cause a problem in practice.) Below we discuss both averaging over genetic models and averaging over effect sizes in each model.

Box 2 | Connections between Bayes factors and *p*-values

Given the ubiquity of *p*-values, it is natural to seek relationships between them and the Bayes factor (BF). For example, given a *p*-value from a published study, is it possible to compute a corresponding BF? We have emphasized that an advantage of BFs over *p*-values is that the strength of the evidence that the BF conveys does not vary with factors that affect power, such as sample size or minor allele frequency (MAF). As this is not true of *p*-values, it follows that any translation from *p*-values to BFs has to depend on further assumptions. Nevertheless, some interesting connections exist between BFs and *p*-values.

Optimistic BFs

Under general assumptions[1], the following result holds for *p*-values that satisfy $p < 1/e$, in which $e \approx 2.72$:

$$BF > -1/(e\, p \log(p)) \quad (7)$$

For example, a SNP with $p = 10^{-6}$ has $BF > 2.7 \times 10^4$, and hence if $\pi = 10^{-4}$ the SNP has a substantial probability of being unassociated (posterior probability of association (PPA) < 0.72). By contrast, if $p = 10^{-7}$ then $BF > 2.3 \times 10^5$ and the PPA could be >0.95. Note also that when $p = 0.05$, we obtain $BF > 2.5$, which is at best only modest evidence against H_0. This is striking given the widespread adoption of a significance level of 0.05.

As equation 7 provides only an upper bound on the BF, it can be thought of as providing an 'optimistic' BF for a given *p*-value. It seems unlikely that any useful lower bound exists, and so there is no corresponding 'pessimistic' BF.

The implied prior of *p*-values

The BF under a strictly additive model, with a $N(0,\sigma)$ distribution for the effect size under H_1, produces approximately the same ranking as *p*-values from standard additive-model tests, providing that σ^2 is chosen to be proportional to a factor that depends on sample size[42] but is asymptotically proportional to 1/MAF(1 – MAF). A similar result holds for non-additive models[42].

Therefore, for a given sample size, ranking SNPs by their *p*-values is equivalent to a Bayesian analysis that makes some very specific assumptions. In particular, it assumes that truly associated low-MAF SNPs tend to have larger effect sizes than SNPs with a larger MAF. Broadly speaking, this assumption may be reasonable[43,44], but there is no apparent justification[45] for the mathematical form 1/MAF(1 – MAF). Bayesians are free to choose the dependence of σ on MAF according to whatever formula they believe best fits the available background information, and hence they can in principle develop better ways to prioritize SNPs for follow up. Furthermore, and perhaps more importantly, this example shows that frequentist analyses can make implicit assumptions of which the user is unaware — see the 'Imputation' subsection in the main text for another example.

Table 2 | **Tail probabilities under various prior distributions**

	$\pi = 10^{-4}$, $\theta \sim N(0,0.2)$	$\pi = 10^{-4}$, $\theta \sim N(0,0.4)$	$\pi = 10^{-4}$, mixture of normals*	NEG shape 1.0, scale 0.0012	NEG shape 1.4, scale 0.006	NEG shape 1.8, scale 0.015
$P[\|\theta\|>0.05]$	8.0×10^{-5}	9.0×10^{-5}	8.1×10^{-5}	5.8×10^{-4}	3.5×10^{-3}	1.7×10^{-2}
$P[\|\theta\|>0.1]$	6.2×10^{-5}	8.0×10^{-5}	6.4×10^{-5}	1.4×10^{-4}	5.3×10^{-4}	2.1×10^{-3}
$P[\|\theta\|>0.2]$	3.2×10^{-5}	6.2×10^{-5}	3.6×10^{-5}	3.6×10^{-5}	7.8×10^{-5}	2.0×10^{-4}
$P[\|\theta\|>0.4]$	4.5×10^{-6}	3.2×10^{-5}	8.8×10^{-6}	9.0×10^{-6}	1.1×10^{-5}	1.7×10^{-5}
$P[\|\theta\|>1]$	5.7×10^{-11}	1.2×10^{-6}	1.1×10^{-6}	1.4×10^{-6}	8.5×10^{-7}	6.2×10^{-7}

Each entry gives the probability for a SNP to have an effect size θ above a given threshold under a given prior distribution. In the first 3 columns the SNP is assumed to have probability 10^{-4} of a non-zero effect. Given a non-zero effect, the distribution for θ is either a single normal (columns 1 and 2) or a weighted average of 3 normal distributions (column 3), each with a different standard deviation. The final value of column 1 illustrates the 'thin tails' property of the normal distribution, which markedly reduces the weighting of very large effect sizes. In the final three columns all SNPs have a non-zero effect, but the normal–exponential–gamma (NEG)[29,46] prior distribution has a sharp peak around zero, so most effect sizes are negligible. The NEG has 'fat tails' that decrease only slowly for large effect sizes, which indicates agnosticism about the magnitude of non-negligible effects. From the middle row, the expected numbers of SNPs with $|\theta|>0.2$ (so that the odds ratio is >1.22 or <0.82) per 100,000 tested range from 3.2 (column 1) to 20 (column 5). Note that many of these SNPs will not be detected in a typical genome-wide association study because of limited power, and also that these numbers can include multiple SNPs in high linkage disequilibrium with the same functional variant. * Under H_1, $\theta \sim 0.9 \times N(0,0.2) + 0.05 \times N(0,0.4) + 0.05 \times N(0,0.8)$.

Odds
The probability of the occurrence of a particular event (for example, the onset of disease) divided by the probability of the event not occurring. It is often mathematically convenient to transform a probability, which must lie between zero and one, to odds, which can take any positive value.

Bonferroni correction
When multiple hypotheses are tested, the Bonferroni correction to the overall desired significance level (α) is obtained by dividing it by the number of tests (k), so that each hypothesis is rejected if p-value < α/k.

False discovery rate
For a sequence of hypothesis tests, the false discovery rate is the proportion of times H_0 is true among those tests for which H_0 is rejected.

Odds ratio
The odds ratio comparing, for example, two genotypes is the odds for individuals with the first genotype divided by the odds for individuals with the second genotype.

Logistic regression
A regression model for binary outcomes (such as case and control) in which the logarithm of the odds is related linearly to one or more predictors, such as SNP minor allele count(s).

Laplace approximation
A method for approximating the integral of a (possibly multidimensional) probability density based on replacing that density by a Gaussian probability density with the same mean and variance–covariance matrix.

Maximum-likelihood estimate
The maximum-likelihood estimate of an unknown parameter in a statistical model is the value of the parameter that maximizes the probability under the model of the observed data.

Averaging over genetic models. We recommend allowing for both additive and non-additive effects by, for example, using a weighted combination of models 2 and 3 above or a combination of models 1, 2 and 4. Weights should be chosen to reflect the investigator's belief about the plausibility of each type of effect given the current theory and previous data. For example, if an investigator believes that most SNPs will act in a near-additive manner, with a minority acting either dominantly or recessively, then they might put 80% weight on the general model centred on additivity (model 3) and 10% weight on each of the dominant and recessive models (model 2). In this case, the overall BF would be computed as $0.8\,BF^a + 0.1\,BF^d + 0.1\,BF^r$, in which BF^a, BF^d and BF^r are the BFs under the near-additive, dominant and recessive models, respectively.

Even small weights on non-additive models can allow the identification of large non-additive effects without substantial reduction in the ability to detect near-additive effects. Although the choice of weights may differ from one researcher to another, typically the effect on the PPA is small unless some weights are chosen to be at, or close to, zero (see further discussion below). An example is provided in TABLE 1, in which BFs were computed under a weighted average of additive and general (near-additive) models. For bipolar disorder, SNP rs420259 has a PPA of 0.56 when $\pi = 10^{-4}$, which is inconclusive but indicates that the SNP merits further investigation. Under a strictly additive model the PPA would be 0.01 and the SNP would probably have been ignored.

Combining models in an interpretable way is much harder in the frequentist framework. For example, for realistic effect and sample sizes the ATT tends to have more power than the Fisher exact test for associations that are close to additive and, conversely, the Fisher exact test has better power to detect far-from-additive effects. The two tests can be combined by, for example, regarding the smaller of the two p-values as a test statistic — a method that is similar to the MAX test[25]. However, this treats the ATT and Fisher exact test equally, which may not be optimal. An attractive Bayesian alternative is to form a weighted combination of the WABF and rBF — which can be viewed as Bayesian analogues of the ATT and Fisher exact test (see Supplementary information S1 (box)) — in accordance with the researcher's prior probability of additive and non-additive effects.

Averaging over effect sizes. The prior distributions associated with models 1–3 above involve an $N(0,\sigma)$ distribution for the effect size, which decays quickly in the tails. Consequently, no single value of σ allows for most effects to be small while also making sufficient allowance for occasional large effects. For example, under an additive model in which $\pi = 10^{-4}$ and the prior distribution on θ is N(0,0.2) (TABLE 2, column 1), ~5 SNPs per million are predicted to have $|\theta|>0.4$, but the probability for a SNP to have $|\theta|>1$ is minuscule — <1 in 10 billion (recall that $\theta = 1$ means OR = 2.72). Although most investigators would accept that $|\theta|>1$ rarely occurs, they may be reluctant to assume it to be this unlikely for a phenotype that lacks previous GWA studies. Increasing σ to 0.4 (TABLE 2, column 2) increases $P[|\theta|>1]$, but it also implies that 32 SNPs per million have $|\theta|>0.4$, which may be regarded as unrealistically high.

A simple solution to this problem is to replace the $N(0,\sigma)$ prior distribution with a mixture of normal distributions[12] that gives an increased, although still small, probability to very large ORs and only slightly affects the probabilities that are assigned to more moderate effects (TABLE 2, column 3). The BF under the mixture prior distribution is simply the weighted average of the BFs under each normal prior distribution, so the WABF and other methods based on the normal prior distribution are also applicable to the mixture prior distribution. In addition to the possible dependence of π on the MAF, as mentioned above, it is plausible that σ varies with the MAF. One way to incorporate this is to choose mixture weights that give more weight to larger values of σ as the MAF decreases.

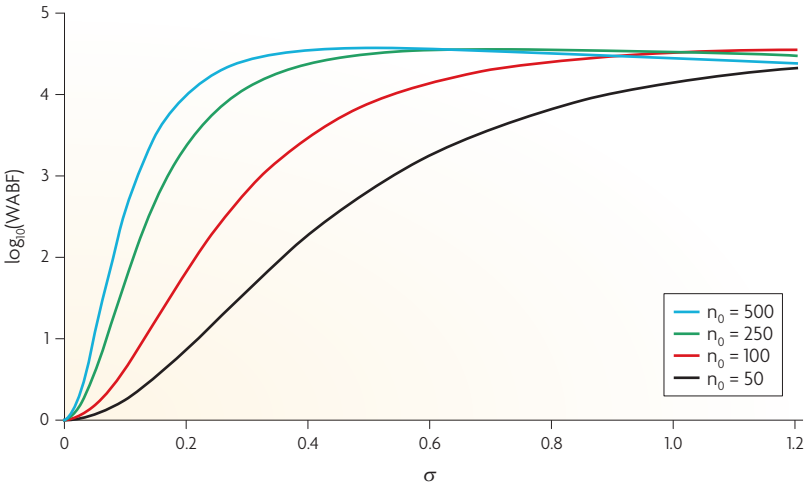


Figure 1 | **Dependence of the Bayes factor on minor allele count and on the prior standard deviation of the effect size.** The curves show the Wakefield approximate Bayes factor (WABF; equation 6) for a SNP with a *p*-value $\approx 5 \times 10^{-7}$ using 4 values of n_0, which is the minor allele count among cases and controls combined. There are n_0 cases and n_0 controls, so the minor allele fraction remains constant at 0.25. As σ (the standard deviation of the effect size) increases from 0, the $\log_{10}$(WABF) for each SNP rises from 0 to a maximum value of 4.57 before gradually decreasing as σ continues to increase. If $n_0 \geq 250$, the Bayes factors (BFs) vary by roughly one order of magnitude for $0.2 < \sigma < 1$, but when $n_0 = 50$, the BF varies more markedly, by several orders of magnitude for σ in this range. If $\pi = 10^{-4}$, then $\log_{10}$(WABF)<4.57 implies PPA<0.79. Therefore, under our assumptions, a SNP just reaching the *p*-value threshold of 5×10^{-7} still has a substantial chance of being a false discovery.

Sensitivity. As there is always some flexibility in the choice of weights and the choice of values for σ, one should consider the sensitivity of the results to these choices. Briefly, sensitivity tends to be greatest in situations with less information, such as small studies, or when testing SNPs with a low MAF. FIGURE 1 shows how the WABF varies with σ for different minor allele counts when the *p*-value $\approx 5 \times 10^{-7}$. For common SNPs that are typed in large studies, small changes in σ will typically have little effect on the BF, but σ can have a big impact if the minor allele count is not large.

Example. Variants in *SLCO1B1* have been reported[26] to be associated with statin-induced myopathy. The most significant SNP from the GWA study was rs4363657, which had a reported *p*-value of 4.1×10^{-9}, which is conventionally regarded as highly significant. Based on the reported summary data[26], $\sigma = 0.2$, 0.4 and 0.8 give $\log_{10}$(WABF) = 2.2, 4.1 and 5.2, respectively. Therefore, the WABF depends strongly on σ. If one assumes $\sigma = 0.2$ and $\pi = 10^{-4}$ (TABLE 2, column 1), the PO is approximately $10^{2.2}/10^4$, which gives a PPA of 0.02, so the SNP would be dismissed as almost certainly a false discovery. By contrast, using the mixture-of-normals prior distribution (TABLE 2, column 3), one obtains $\log_{10}$(BF) = 3.9 and PPA $\approx$ 0.44.

The reason that different prior distributions produce such different conclusions here is that the data suggest a large effect size ($\hat{\theta} = 1.46$, OR = 4.3) in a small study (only about 25 copies of the risk allele were observed among the 192 controls[26]). Under the N(0,0.2) prior distribution the observed effect size is almost impossible *a priori*, and therefore the analysis leads to the conclusion that it is almost certainly due to sampling error. By contrast, the mixture prior distribution puts greater weight on large θ, and so leads to the conclusion that the observed association merits further investigation. An alternative Bayesian analysis based on estimation rather than testing (BOX 3) reaches similar conclusions. However, replication data and functional results for a non-coding variant in high linkage disequilibrium (LD) with rs4363657 were also reported[26]. These bolster the case for association but are not taken into account in our reanalyses.

Beyond simple analyses

Incorporating external biological information. In the PPA calculations above we have ignored (as do most analyses) the fact that some SNPs may be good candidates for influencing a phenotype — for example, SNPs that lie in or near a gene that has a known biological function and is plausibly related to the phenotype. The MAF of a SNP, known copy number variation, conservation across species and evidence of selection may also be relevant to both the prior plausibility of association and the effect size under H_1. Indeed, many investigators use such information informally when interpreting the results of an association study. A Bayesian approach allows such reasoning to be quantified through the specification of π, which facilitates scrutiny and rational debate.

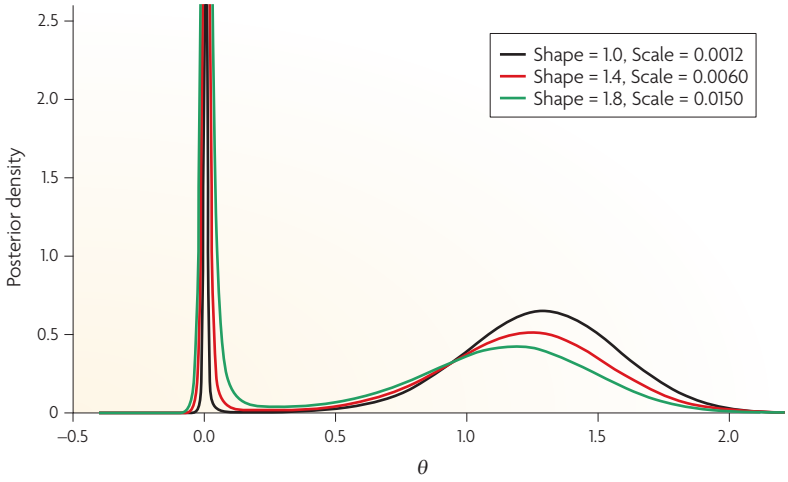
Hypothetical example. In an association study for C-reactive protein (CRP) levels, a SNP in the hepatocyte nuclear factor 1 homeobox A (*HNF1A*) gene has a BF of 3×10^3, whereas a SNP in a gene desert has a BF of 10^4. Investigator 1 assigns a prior probability $\pi = 10^{-4}$ to each SNP and obtains PPAs of 0.23 for the *HNF1A* SNP and 0.50 for the gene-desert SNP. He therefore argues that the gene-desert SNP is a higher priority for follow up. Investigator 2 judges that SNPs near *HNF1A* are good candidates for affecting CRP levels, and she adopts a prior probability of $\pi = 5 \times 10^{-4}$ for the SNP near *HNF1A* and chooses $\pi = 0.9 \times 10^{-4}$ for the SNP in the gene desert. She obtains PPAs of 0.60 for the *HNF1A* SNP and 0.47 for the for the gene-desert SNP and argues that the *HNF1A* SNP is a higher priority for follow up. Both of these investigators would agree that both SNPs should be followed up if resources allow this and that the case is not yet convincing for either association being genuine. Investigator 3 argues that *HNF1A* is an excellent candidate gene for affecting CRP levels and elects to use a prior probability of $\pi = 10^{-2}$ for the SNP in this gene. He then obtains a PPA of 0.97 and concludes that the association is genuine.

As this example shows, the choice of π for each SNP can and should be informed by evidence, but inevitably it has a subjective element. There is a widespread desire to avoid subjectivity in scientific reasoning, but we would argue that the real problem is hidden subjectivity; openness about subjective assumptions can be helpful in clarifying the roles of important but difficult-to-assess factors, such as the genomic context of a SNP. Although the prior probability adopted by Investigator 3 may seem

Statin
A class of drugs that is used to lower cholesterol levels in people with, or at risk of, cardiovascular disease.

Box 3 | Do we need a null hypothesis?

Both *p*-values and the Bayes factors (BFs) we have focused on here require a null hypothesis H_0. The dichotomization of effects into zero and non-zero values can be criticized as being artificial and unrealistic because there may be many SNPs that are weakly associated with a complex phenotype with effect sizes that are too small to be either detectable or of practical interest. Recognizing this dichotomy as artificial makes π, the proportion of SNPs following H_0, difficult to interpret and hence difficult to assign a meaningful value to.



An alternative Bayesian approach is to use a prior distribution for the effect-size parameter θ that does not categorize effects into zero and non-zero, and then focus on estimating θ instead of testing whether $\theta = 0$. A suitable prior distribution should assign high probability densities to values of θ near zero, but the distributions should have 'fat tails' that allow for a few SNPs to have substantial effect sizes. This can be accomplished using the normal-exponential-gamma (NEG)[29,46] prior distribution (TABLE 2). Next, Bayes' theorem is used to compute the posterior distribution for θ, which can be presented visually for direct interpretation. Posterior probabilities of the form $P(|\theta|>t)$, in which t denotes some effect-size threshold of interest, are comparable to the posterior probability of association (PPA) in the hypothesis-testing paradigm. An advantage of the estimation approach is that the user can explicitly control t, which may vary for different phenotypes and perhaps decline over time as knowledge advances and study design and analysis improve. To demonstrate this approach, we applied it to the data for SNP rs4363657 from the genome-wide study in REF. 26 (see the figure). We assumed an additive model and considered each of the NEG distributions from TABLE 2 as prior distributions for θ.

The figure shows the posterior densities for θ, which were obtained by numerical approximation. In each case the posterior density is bimodal, featuring a large spike around zero (truncated in the figure) that corresponds to values that are strongly supported by the prior distribution and a broad mode above one that indicates effects that are supported by the data but that are 'shrunk' towards zero by an amount that depends on the prior shape parameter. The posterior probabilities of $|\theta|>0.1$ under the 3 prior distributions are approximately 0.47, 0.39 and 0.35, which are similar to the PPA of 0.44 that we computed under the hypothesis-testing paradigm using a mixture-of-normals prior distribution (see main text). Despite the fact that the genome-wide study data generated a highly significant *p*-value of 4.1×10^{-9}, our Bayesian analyses indicate that, under a wide range of different assumptions, these data alone are not sufficient to ascertain whether the association is genuine. A further reinforcement of the difficulty of interpreting *p*-values comes from noting that, under a range of prior assumptions, the evidence for this SNP being truly associated is not as strong as for SNPs with less significant *p*-values in TABLE 1.

To allow for the effect-size distribution to vary with minor allele frequency (MAF), we could choose NEG parameters as a function of MAF. Further, an independent NEG prior distribution could be assigned to a dominance term to investigate non-additive models.

extreme, it is similar to the prior probabilities that are implicitly assumed in candidate gene analyses that use much less stringent significance thresholds than are typical for a GWA study. By making such assumptions explicit, the Bayesian analysis facilitates clear discussion and honest assessment of the strength of the evidence.

Imputation. Genotype imputation methods[11,12,27] have recently emerged as a powerful approach for testing variants that were not genotyped in a study. Many SNPs are easy to impute accurately, but SNPs that are not well correlated with genotyped SNPs are hard to impute (no genotype is assigned a probability close to one). Tests of hard-to-impute SNPs tend to have lower power than tests of easy-to-impute SNPs, and this power difference will occur whatever testing procedure is used, even if the uncertainty in imputed genotypes is properly accounted for. The lower power at hard-to-impute SNPs should be taken into consideration when assessing the strength of the evidence for an association. Simply ranking imputed SNPs by their *p*-values ignores these differences in power among tests — or, equivalently, it makes the implicit assumption that all tests have the same power. This would require that hard-to-impute SNPs have larger effect sizes than easy-to-impute SNPs, which seems absurd. To avoid this, *p*-values for imputed SNPs need to be weighted according to power. Doing this correctly leads to replacing the *p*-value with a BF, which automatically reduces the weight of the hard-to-impute SNPs. See REF. 20 for further discussion and comparisons.

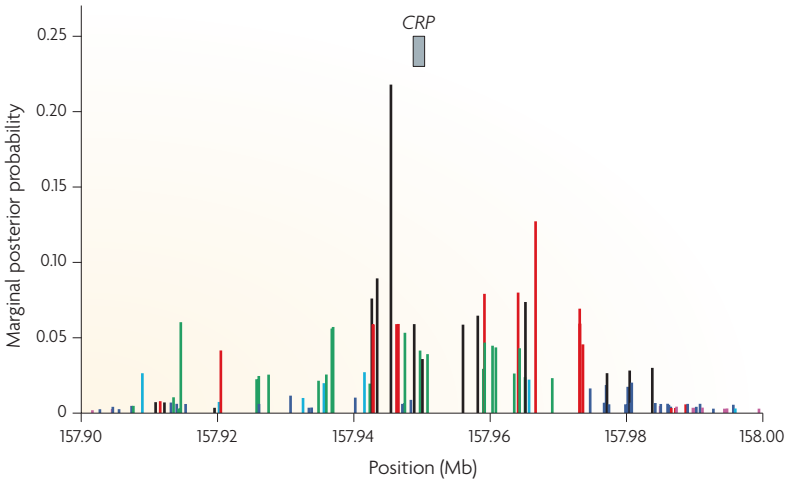
Fine mapping. When, as often arises, many SNPs in a genomic region show association with a phenotype, it is likely that most of these SNPs are not functional, but

Genotype imputation method
A method for estimating ('imputing') the unobserved genotypes of study subjects, both for individuals with missing or unreliable genotypes at a genotyped SNP and for all individuals at an ungenotyped SNP.

their association reflects LD with one or a few causal SNPs that may or may not be genotyped. Important unresolved challenges are: to identify how many distinct causal associations underlie the association data; to detect which of the genotyped SNPs are causal or those that best tag unobserved causal variants; and to quantify the strength of the evidence in each case.

Have some of these/to what end have they been resolved?

Box 4 | **An example of fine mapping using a Bayesian approach**



Using data from an association study for C-reactive protein (CRP)[47], we applied a Bayesian regression approach, implemented using BIMBAM[12], to analyse 103 HapMap SNPs that lie within 50 kb of the *CRP* gene (see the figure). The observed data consisted of 4 SNPs genotyped in 1,910 individuals and a further 10 SNPs genotyped in 980 of those individuals. We used BIMBAM to impute the remaining genotypes in all 1,910 individuals using 60 unrelated HapMap individuals as a panel to learn about patterns of linkage disequilibrium (LD) among SNPs. We assumed that effects add across SNPs, and at each SNP we used the prior S2 from REF. 12, which is a general prior that is centred on additivity (model 3 in the main text). We averaged the results over σ_a = 0.1, 0.2 and 0.4, and used $\sigma_d = \sigma_a/4$ (REF. 12). We used a prior distribution on the number of functional SNPs that gave weights in the ratios 4/2/1/0 to 1-SNP, 2-SNP, 3-SNP and >3-SNP models, respectively. All models with the same number of functional SNPs were assumed to be equally likely *a priori*.

Under these assumptions the posterior probabilities for 1, 2 and 3 functional SNPs were 0.03, 0.32 and 0.64, respectively. Therefore, our analysis supports multiple functional variants that influence plasma CRP levels in the region of the *CRP* gene — additional analysis may reveal support for more than the three SNPs that we considered here. The figure shows the posterior probability of association (PPA) for each SNP (assuming at most three SNPs in each model), and the SNPs are coloured to highlight groups that have high LD with each other, so we expect that at most one of them is a causal SNP.

To summarize the results, the 3 main groups of SNPs (coloured black, red and green) each have marginal probabilities that add up to >0.7. Therefore our results suggest that each of these groups includes or tags one functional SNP, although there is substantial uncertainty about which one (the largest marginal probability of any SNP is 0.22 and the largest probability assigned to any specific combination of SNPs is just 0.0085). These low posterior probabilities can make the Bayesian result seem complicated, but this shows the large uncertainty about which SNPs are functional, which is inevitable in the presence of high LD among SNPs.

Given the complexity of the situation, the limitations of this analysis (for example, it ignores possible interactions among SNPs) and the limitations of the data (the majority of the SNPs are imputed, rather than typed), it is reassuring that the results are broadly consistent with those of another investigation[16] of this gene that used different data and a different Bayesian analysis and identified SNPs rs1130864, rs1205 and rs3093077 as showing independent association with CRP concentration. Although these SNPs are not assigned the highest marginal posterior probabilities in our analysis, one of them is included in each of the black, red and green clusters shown in the figure.

Given a set of SNPs *S* for which genotype data are available, the goal of identifying causal or causal-tagging SNPs can be formulated as a problem of deciding which explanatory variables should be included in a regression model[9,12,16]. In a Bayesian approach, each subset of *S* is assigned a prior probability of being the 'true' model, and posterior probabilities are then computed; both prior and posterior probabilities should add up to one over all models (including the model that includes no SNPs).

Have any methods been developed to avoid this?

Although undesirable, unless *S* is small it is often necessary in practice to assign zero prior probability to many models that are collectively implausible. For example, one might assume that the number of causal variants is at most three and assign zero prior probability to all models with more than three SNPs. If this assumption is wrong, the analysis is still likely to be useful in helping to identify the three most important causal variants. Even with this assumption, there could be many thousands of models to assess, which could make it challenging to summarize their posterior probabilities. One useful summary is the PPA for each individual SNP, which is obtained by adding up the probabilities of all models containing that SNP. One can also summarize the evidence for exactly X causal SNPs in a region by adding up the posterior probabilities that are assigned to models containing X SNPs.

Would be good to develop a rigorous way to quantify the "shared" PPA among high-LD SNPs in certain regions (if not already done)

If several SNPs are in strong LD and each is strongly associated with phenotype, it is likely that at most one is causal, but it is difficult to assess which one. In the Bayesian approach outlined above, the high-LD SNPs are likely to 'share' the PPA among them so that the total PPA over these SNPs is high but each SNP has only a modest PPA. Extracting this information from the full posterior distribution is possible (BOX 4) but not straightforward. By contrast, step-wise selection or penalized likelihood approaches, such as the Lasso approach[28,29], typically select just one or a few SNPs. These approaches provide a superficially simpler solution and can be faster to compute but do not provide probabilities measuring the (often low) level of certainty associated with the resulting models, which is in contrast to the rich information provided in a full Bayesian solution.

Meta-analysis. Once a SNP has been found to be associated with a phenotype in one study, the next step is usually to investigate its effect in further studies, perhaps using different populations and/or study designs. Replication can provide useful confirmation of the original result, but true genetic effects can vary greatly among populations and studies owing to differences in environmental exposures, differences in the LD between tested and functional variants and/or differences in genetic background. Meta-analysis can be used to increase confidence in an apparent association and to investigate the heterogeneity of the effect sizes and genetic models[30].

In 'fixed-effects' meta-analysis, the effect size θ is assumed to be constant across studies and the null hypothesis is $\theta = 0$. In 'random-effects' meta-analysis,

θ is assumed to vary across populations according to an $N(\mu,\tau)$ distribution in which τ is unknown. However, the null hypothesis is usually H_0: $\mu = 0$ instead of the more appropriate H_0': $\mu = \tau = 0$ (see REF. 31 for a discussion in the setting of linkage analyses). Testing of H_0 was developed for applications in which only modest heterogeneity of effects is expected. For genetic associations, a high level of heterogeneity (large τ) is often plausible, but a higher variance in apparent effect sizes makes it less likely that H_0 will be rejected (because an apparent mean effect becomes less significant as the variance increases). By contrast, increasing τ makes the rejection of H_0' more likely. Therefore, heterogeneous genetic effects will often be missed if H_0 is tested instead of H_0'. For example, a fixed-effects meta-analysis gave a p-value of 1.3×10^{-12} for a SNP in the fat mass and obesity-associated (*FTO*) gene and type 2 diabetes but only $p = 0.015$ for a random-effects meta-analysis of the same data[32]. In the random-effects meta-analysis, the high value of τ across studies, which was largely due to the varying body mass indexes of the study subjects, has weakened the evidence for a true association instead of strengthening it.

Broadly speaking, the fixed-effects model is implausible for genetic associations, and researchers are often encouraged to adopt the more conservative random-effects analysis[30,32]. Researchers are understandably reluctant to do this because of the potentially large and typically unwarranted loss of power, and instead they typically test H_0": $\tau = 0$ and perform a fixed-effects analysis if H_0" is accepted. This two-step procedure is unsatisfactory because a false H_0" may be accepted owing to inadequate power or, in large studies, H_0" may be rejected for low heterogeneity, leading to the adoption of a random-effects analysis with reduced power. A Bayesian analysis using, for example, WinBUGS[9,33] for modest numbers of SNPs overcomes these problems by allowing researchers the flexibility to formulate an appropriate null hypothesis and to contrast it with a suitable alternative hypothesis, which allows some heterogeneity (for example, REFS 16,18). Moreover, Bayesian methods can also incorporate variations in genetic models across studies and the effects of deviation from Hardy–Weinberg equilibrium[34].

Guidance for refereeing Bayesian analyses. The Bayesian approach is well suited to the needs of a scientist engaged in the disinterested pursuit of knowledge, but its greater flexibility can make it more difficult for referees and editors to perform a gatekeeper role in maintaining high publication standards. Some critics worry, for example, that a researcher who is desperate to publish at all costs could exaggerate the prior probability of an outcome supported by the data to boost the reported PPA, even though this could subsequently harm the researcher's professional reputation.

What prior assumptions are reasonable in a Bayesian analysis? And what minimum standards can reviewers reasonably impose? A Bayesian analysis requires more assumptions to be made explicitly, and researchers should be required to describe their modelling assumptions in sufficient detail. However, reviewers and editors should keep an open mind towards differing subjective assessments, and Bayesian analyses should not be penalized for openness, particularly when the corresponding frequentist analysis would evade criticism by keeping issues hidden. For example, a Bayesian's choice of effect-size prior distribution is always open to criticism, whereas the frequentist can escape such scrutiny by not making any explicit choice. Yet the problem does not disappear in a frequentist analysis: the performance of frequentist tests depends on their power, and power calculations require assumptions about effect sizes. Similar comments apply to the choice of π. In a frequentist analysis this is hidden in the choice of a 'genome-wide significance' level, which can depend on many factors that are specific to individual studies[35], but this dependence is rarely discussed.

Rather than ask authors to provide a strong rationale for every assumption, we suggest that reviewers focus on whether some assumptions seem to be unreasonable or poorly chosen. If so, they might ask, for example, for results under a different prior distribution to be included. However, it is impossible to present results under all priors, and presenting comprehensive results from one or two defensible priors should, in most cases, suffice. Of course, a thorough analysis — whether Bayesian or frequentist — will assess the sensitivity of key results to assumptions and highlight where a reasonable reader might reach a different conclusion.

Perspective

Even simple Bayesian methods have advantages over standard frequentist analyses, and we believe that the advantages of a Bayesian approach will increase as the current simplistic 'one SNP at a time' testing paradigm is replaced by the testing of more detailed hypotheses in more complex data sets. One common and understandable concern regarding the methods described here is the need to pre-specify subjective values for π and the parameters of the prior distribution of effect sizes under H_1, such as σ and the weights for different genetic models. It is possible and desirable to learn about these quantities from previous association data for the same or related phenotypes. Similarly, it should be possible to quantitatively assess prior probabilities for each SNP on the basis of whether it is a non-synonymous coding SNP or whether it lies in or near a promoter region. Indeed, several recent publications show progress along these lines. For example, a recent Bayesian analysis[36] reported effect-size distributions, the abundance of additive over dominant effects and SNP-specific prior probabilities that affect phenotype in the context of SNP associations with expression traits (see also REF. 37). Furthermore, a positive association has been shown between a gene showing differential expression (across a large experimental database) and it harbouring disease-associated variants[38].

In addition to these potentially exciting new developments, there remain some important extensions to simple Bayesian analysis that would benefit from further development. For example, adjustment for

Hardy–Weinberg equilibrium
This holds at a given locus in a given population when the two alleles of individuals in the population are mutually independent.

population structure does not yet seem to be integrated into available Bayesian analyses, and Bayesian approaches for survival-time data are limited to candidate-gene association studies[39]. Current software has limited functionality for incorporating relevant observed covariates (for example, sex, smoking or covariates measuring population structure), although this is straightforward in principle in a Bayesian regression-based analysis. In Supplementary information S1 (box) we show the performance of several BFs and standard *p*-values in a small simulation study and in a reanalysis of a small GWA study.

1. Sellke, T., Bayarri, M. J. & Berger, J. O. Calibration of *p* values for testing precise null hypotheses. *Am. Stat.* **55**, 62–71 (2001).
2. Sterne, J. A. C. & Davey Smith, G. Sifting the evidence — what's wrong with significance tests? *BMJ* **322**, 226–231 (2001).
3. Ioannidis, J. P. A. Effect of formal statistical significance on the credibility of observational associations. *Am. J. Epidem.* **168**, 374–383 (2008).
4. Ayres, K. L. & Balding, D. J. Measuring departures from Hardy–Weinberg: a Markov chain Monte Carlo method for estimating the inbreeding coefficient. *Heredity* **80**, 769–777 (1998).
5. Shoemaker, J. S., Painter, I. S. & Weir, B. S. Bayesian statistics in genetics — a guide for the uninitiated. *Trends Genet.* **15**, 354–358 (1999).
6. Beaumont, M. A. & Rannala, B. The Bayesian revolution in genetics. *Nature Rev. Genet.* **5**, 251–261 (2004).
7. Marjoram, P. & Tavare, S. Modern computational approaches for analysing molecular genetic variation data. *Nature Rev. Genet.* **7**, 759–770 (2006).
8. O'Hara, R. B., Cano, J. M., Ovaskainen, O., Teplitsky, C. & Alho, J. S. Bayesian approaches in evolutionary quantitative genetics. *J. Evol. Biol.* **21**, 949–957 (2008).
9. Wakefield, J. Bayesian methods for examining Hardy–Weinberg equilibrium. *Biometrics* 13 May 2009 (doi:10.1111/j.1541-0420.2009.01267.x).
10. Lunn, D. J., Whittaker, J. C. & Best, N. A Bayesian toolkit for genetic association studies. *Genet. Epidem.* **30**, 231–247 (2006).
11. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genet.* **39**, 906–913 (2007).
The supplementary material of this article includes a review of frequentist tests and BFs for single-SNP association and a brief review of the Laplace approximation. In particular, it describes the Bayesian analysis methods implemented in the SNPTEST software.
12. Servin, B. & Stephens, M. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet.* **3**, e114 (2007).
This paper includes a description of several of the Bayesian analysis methods that are implemented in the BIMBAM software, including the Bayesian multi-SNP analysis methods that we used in this Review.
13. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
A landmark paper because of the size of the studies, the pioneering use of unphenotyped common controls for a range of diseases and the large number of novel genetic associations reported. The authors also advocate the use of Bayesian approaches for evaluating evidence of association, which was reported alongside traditional *p*-values for the first time in a major study.
14. Wakefield, J. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am. J. Hum. Genet.* **81**, 208–227 (2007).
15. Hosking, F. J., Sterne, J. A. C., Smith, G. D. & Green, P. J. Inference from genome-wide association studies using a novel Markov model. *Genet. Epidem.* **32**, 497–504 (2008).
16. Verzilli, C. *et al.* Bayesian meta-analysis of genetic association studies with different sets of markers. *Am. J. Hum. Genet.* **82**, 859–872 (2008).
17. Fridley, B. L. Bayesian variable and model selection methods for genetic association studies. *Genet. Epidem.* **33**, 27–37 (2009).
18. Newcombe, P. J. *et al.* Multilocus Bayesian meta-analysis of gene–disease associations. *Am. J. Hum. Genet.* **84**, 567–580 (2009).
19. Wakefield, J. Reporting and interpretation in genome-wide association studies. *Intern. J. Epidem.* **37**, 641–653 (2008).
20. Guan, Y. & Stephens, M. Practical issues in imputation-based association mapping. *PLoS Genet.* **4**, e1000279 (2008).
This article includes a detailed discussion of the advantages of Bayesian methods over frequentist methods when assessing associations with imputed SNPs.
21. Balding, D. J. A tutorial on statistical methods for population association studies. *Nature Rev. Genet.* **7**, 781–791 (2006).
This Review covers: preliminary analyses (of Hardy–Weinberg and linkage equilibria, inference of phase and missing genotypes); single-SNP tests of association for binary, continuous and ordinal outcomes; multi-SNP and haplotype analyses; and dealing with population stratification and multiple-testing issues, largely within the frequentist framework.
22. Jeffreys, H. *Theory of Probability* (Oxford Univ. Press, 1961).
23. Good, I. J. The Bayes/non-Bayes compromise: a brief review. *J. Am. Stat. Assoc.* **87**, 597–606 (1992).
24. Seaman, S. R. & Richardson, S. Equivalence of prospective and retrospective models in the Bayesian analysis of case–control studies, *Biometrika* **91**, 15–25 (2004).
25. Freidlin, B., Zheng, G., Li, Z. H. & Gastwirth, J. L. Trend tests for case–control studies of genetic markers: power, sample size and robustness. *Hum. Hered.* **53**, 146–152 (2002).
26. The SEARCH Collaborative Group. *SLCO1B1* variants and statin-induced myopathy — a genomewide study. *N. Engl. J. Med.* **359**, 789–799 (2008).
27. Scott, L. J. *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316**, 1341–1345 (2009).
28. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* **58**, 267–288 (1996).
29. Hoggart, C. J., Whittaker, J. C., De Iorio, M. & Balding, D. J. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet.* **4**, e1000130 (2008).
30. Kavvoura, F. K. & Ioannidis, J. P. A. Methods for meta-analysis in genetic association studies: a review of their potential and pitfalls. *Hum. Genet.* **123**, 1–14 (2008).
31. Van Houwelingen, H. & Lebrec, J. P. in *Meta-analysis and Combining Information in Genetics and Genomics* (eds Guerra, R. *et al.*) 49–66 (CRC Press, 2009).
32. Ioannidis, J. P., Patsopoulos, N. A. & Evangelou, E. Heterogeneity in meta-analyses of genome-wide association investigations. *PLoS ONE* **2**, e841 (2007).
33. Lunn, D. J., Thomas, A., Best, N. & Spiegelhalter, D. WinBUGS — a Bayesian modelling framework: concepts, structure, and extensibility. *Stat. Comput.* **10**, 325–337 (2000).
34. Thompson, J. R., Minelli, C., Abrams, K. R., Thakkinstian, A. & Attia, J. Combining information from related meta-analyses of genetic association studies. *J. R. Stat. Soc. C* **57**, 103–115 (2008).
35. Hoggart, C. J., Clark, T. G., De Iorio, M., Whittaker, J. C. & Balding, D. J. Genome-wide significance for dense SNP and resequencing data. *Genet. Epidem.* **32**, 179–185 (2008).
36. Veyrieras, J.-B. *et al.* High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet.* **4**, e1000214 (2008).
37. Lee, S.-I. *et al.* Learning a prior on regulatory potential from eQTL data. *PLoS Genet.* **5**, e1000358 (2009).
38. Chen, R. *et al.* FitSNPs: highly differentially expressed genes are more likely to have variants associated with disease. *Genome Biol.* **9**, R170 (2008).
39. Tachmazidou, I., Andrew, T., Verzilli, C. J., Johnson, M. R. & De Iorio, M. Bayesian survival analysis in genetic association studies. *Bioinformatics* **24**, 2030–2036 (2008).
40. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate — a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
41. Storey, J. D. A direct approach to false discovery rates. *J. R. Stat. Soc. B* **64**, 479–498 (2002).
42. Wakefield, J. Bayes factors for genome-wide association studies: comparison with *P*-values. *Genet. Epidem.* **33**, 79–86 (2009).
This is the last in a sequence of three single-author papers published by Wakefield in successive years. This paper uses the approximate BF introduced in Reference 14 to highlight what can be regarded as implicit assumptions in the use of standard *p*-values as the primary summaries of evidence for association.
43. Wang, W. Y. S., Barratt, B. J., Clayton, D. G. & Todd, J. A. Genome-wide association studies: theoretical and practical concerns. *Nature Rev. Genet.* **6**, 109–118 (2005).
44. Gorlov, I. P., Gorlova, O. Y., Sunyaev, S. R., Spitz, M. R. & Amos, C. I. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **82**, 100–112 (2008).
45. Greenland, S. Multiple comparisons and association selection in general epidemiology. *Intern. J. Epidem.* **37**, 430–434 (2008).
46. Scheipl, F. & Kneib, T. Locally adaptive Bayesian P-splines with a normal-exponential-gamma prior. *Comput. Stat. Data Anal.* **53**, 3533–3552 (2009).
47. Reiner, A. P. *et al.* Polymorphisms of the *HNF1A* gene encoding hepatocyte nuclear factor-1α are associated with C-reactive protein. *Am. J. Hum. Genet.* **82**, 1193–1201 (2008).

Acknowledgements

We thank C. Hoggart for providing R code to compute the normal-exponential-gamma probability density function and J. Wakefield for helpful discussions and critical reading of an early draft. We thank R. Krauss for access to the CRP genotype and phenotype data that we analysed here. We are also grateful to W. Astle, A. Ramasamy, L. Bottolo, L. Coin, P. O'Reilly and H. Eleftherohorinou for discussions. The authors' work is supported in part by National Institutes of Health grants HL084689 (to M.S.) and EP/C533542 (to D.J.B.).

FURTHER INFORMATION

Matthew Stephens' homepage: http://stephenslab.uchicago.edu
David J. Balding's homepage: http://www.zebfontaine.eclipse.co.uk/djb.htm
BIMBAM: http://stephenslab.uchicago.edu/software.html
SNPTEST: http://www.stats.ox.ac.uk/~marchini/software/gwas/snptest.html
WinBUGS: http://www.mrc-bsu.cam.ac.uk/bugs
***Nature Reviews Genetics* series on Modelling:** http://www.nature.com/nrg/series/modelling/index.html
***Nature Reviews Genetics* series on Genome-wide Association Studies:** http://www.nature.com/nrg/series/gwas/index.html

SUPPLEMENTARY INFORMATION

See online article: S1 (box)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF