European Journal of Human Genetics (2010) 18, 746–750

www.nature.com/ejhg

REVIEW

Bayes factors in complex genetics

Stephen Sawcer*,[1]

The past few years have seen tremendous progress in our understanding of the genetics underlying complex disease, with associated variants being identified in dozens of traits. Despite the fact that this growing body of empirical evidence unequivocally shows the necessity for extreme levels of significance and large samples sizes, the reasoning behind these requirements is not always appreciated. As genome-wide association studies reach the limits of their resolution in the search for rarer and weaker effects, the need for appropriate design and interpretation will become ever more important. If the genetic analysis of complex disease is to avoid accumulating false positive claims, as it has in the past, then researchers will need to allow for less tangible variables such as power and prior odds rather than relying exclusively on significance when assessing the results of these studies. In this review, the basic foundations of association testing are explained from a Bayesian perspective and the potential benefits of Bayes factors as a means of measuring the weight of evidence in support of an association are described.
European Journal of Human Genetics (2010) **18,** 746–750; doi:10.1038/ejhg.2010.17; published online 24 February 2010

Keywords: association; complex genetics; significance; power; prior odds; Bayes factors

Testing for association is one of the most frequently used paradigms in biomedical research. Identifying differences between cases and controls can shed invaluable light on the aetiology of a disease. Although, in principle, any potentially relevant 'exposure' could be tested for association, measuring exposure to environmental factors is frequently complex, imprecise and subject to bias. Even where established assessment tools exist, it can be difficult to meaningfully measure an environmental exposure. If the effect of an exposure is large, such as the effect of smoking on the risk of developing lung cancer, then crude measures of the exposure can be sufficient.[1] Otherwise, the inaccuracies inherent in measuring the exposure may swamp any systematic difference. One of the main advantages of genetics is that an individual's exposure to any given allele can generally be measured with extremely high accuracy. Genotyping data is highly reproducible, both within and across laboratories. It is the accuracy with which an exposure can be measured that ultimately limits the size of effects that can be detected, the more accurate the measurement, the smaller the effect that can be reliably shown.

WHY DO WE NEED STATISTICS?

As it is never possible to test an exposure in an entire population, we inevitably have to base our assessment of any potentially relevant aetiological factor on its appearance in a sample of cases and a sample of controls. Even when unbiased and truly random, this sampling process can generate an apparent difference between cases and controls regardless of whether there is a difference at the population level. Faced with this unavoidable variation, we need a means to assess the extent to which any observed difference is indicative of a genuine difference at the population level, as opposed to just being a consequence of random variation in the sampling and/or measurement process; that is, we need to be able to infer to what extent we can be sure that any observed association is true as opposed to false positive.

Statistical analysis provides a means to judge the degree of confidence with which we can distinguish between these two opposing positions (hypotheses); genuine association, in which there really is an exposure difference at the population level, and the null hypothesis in which no such difference exists. Assuming that all sources of variation in the estimates of exposure are random and free from bias, the more extreme the case–control difference, the more likely it is that the tested exposure is indeed genuinely associated. The probability of observing any given level of difference, or something more extreme, is defined as the significance (*P*-value) when it is calculated under the null hypothesis and as the power when it is calculated under the alternative hypothesis of genuine association. Before performing any test for association, it is traditional to set some arbitrary significance cut off value, with the intention to declare results as 'positive' if they are more extreme than this cut off and 'negative' if they are less extreme. This thinking gives rise to the familiar 'two by two' table (see Figure 1).

WHAT SIGNIFICANCE THRESHOLD SHOULD BE SELECTED?

Inspection of Figure 1 shows that for any randomly selected potentially relevant factor, before any experiment has been performed, the odds that this factor is genuinely associated with the disease are R/S (the so-called prior odds). After testing if the result is positive (ie if the observed *P*-value is equal to or is more extreme than the selected significance threshold), then the odds that the tested factor is associated becomes a/b (the posterior odds). Simple algebra confirms that

$$\text{PosteriorOdds} = \left(\frac{\text{Power}}{\text{Significance}}\right) * \text{PriorOdds} \qquad (1)$$

This equation shows that the confidence we can place in any positive result is determined by three variables: the prior odds, the significance

[1]Department of Clinical Neuroscience, University of Cambridge, Addenbrooke's, Cambridge, UK
*Correspondence: Dr S Sawcer, Department of Clinical Neuroscience, University of Cambridge, Addenbrooke's Hospital, Hills Road, Cambridge, CB2 2QQ, UK.
Tel: + 44 1223 216073; Fax: + 44 1223 336941; E-mail: sjs1016@mole.bio.cam.ac.uk
Received 2 September 2009; revised 12 January 2010; accepted 13 January 2010; published online 24 February 2010

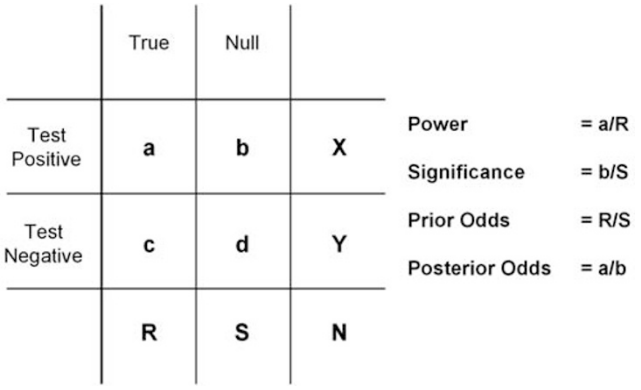

Figure 1 This figure shows the familiar two by two table, the null and alternative hypotheses in the columns and the two alternate test outcomes in the rows. If we consider a particular class of potentially relevant exposures (for example, genetic variant that are common in the human population), there might be N of these in total of which R are genuinely associated (the 'True' hypothesis holds) and S are not associated (the 'null' hypothesis holds). For any given sample and selected significance threshold, if all N factors were tested, then a certain number from each class would give positive test results and the remainder would test negative. If we imagine averaging these counts over all possible studies with the same design, we can complete the expected numbers in each cell of the two by two table. Among the genuinely associated factors, on average 'a' will exceed the selected significance threshold and 'c' will not (a+c=R). Likewise, among the unassociated factors on average 'b' will exceed the selected significance threshold and 'd' will not (b+d=S). Respectively a, b, c and d are the true positives, false positives, false negatives and true negatives. As significance is defined as the probability of seeing data this extreme or more extreme if an unassociated factor is tested, then by definition significance= b/S. Similarly, as power is defined as the probability of seeing data this extreme or more extreme if an associated factors is tested (ie under the alternative hypothesis), then again by definition power=a/R.

threshold and the power. The ratio between power and significance indicates how much more likely one is to see data at or exceeding the selected threshold if a tested factor is indeed associated as opposed to if it is unassociated. A significance threshold of 5% (P=0.05) is traditionally used in biomedical research. If power is high (c100%) and the prior odds are even, that is if the null and alternative hypotheses are equally likely before testing, then the odds that a positive result is true (the posterior odds) will be 20:1. In short, when these underlying assumptions are valid, we can expect almost all results that are positive at the 5% level to be true. However, confidence in the 5% threshold must be lowered if the power and/or prior odds are reduced (see below).

Analyzing Eq. (1), it is important to remember that no matter how large the sample size or how strong the effect sought, the power can never be >1. In this 'best case' scenario (Eq. (2)), it is clear that the Prior Odds are the primary determinant of what significance threshold needs to be set if the Posterior Odds are to be meaningful. If one wishes to be confident that a 'positive' result is more likely to be true than false, then one has to set a significance threshold commensurate with the Prior Odds. If the prior odds are low, as they are in the genetic analysis of complex disease (see below), then it is essential to set a correspondingly extreme significance threshold. At less extreme significance thresholds, the Posterior Odds will remain $<<1$ and, therefore, most of the 'positive' results will inevitably be false.

$$\text{PosteriorOdds} = \left(\frac{1}{\text{Significance}}\right) * \text{PriorOdds} \qquad (2)$$

Although in any given situation we cannot know the prior odds with certainty, in the genetics of complex disease it has been possible to determine very realistic estimates for this critical parameter, at least as it relates to common variants (genetic variants in which both alleles have a frequency of more than a few percent). The Human Genome Project has shown that there are some 10 million common variants in the human population.[2,3] In comparison, segregation analysis of recurrence risks in complex diseases such as multiple sclerosis (OMIM 126200) suggest that only a modest number of these variants are likely to be relevant in any particular disease.[4,5] Estimating this number is difficult as segregation analysis has little ability to distinguishing between a restricted number of modest effects (odds ratio, OR: 1.2–1.3) and a larger number of small effects (OR: <1.1).[5] Furthermore, linkage disequilibrium (LD, the correlation between closely linked variants) means that association may be detectable at flanking variants as well as causal ones; indeed, current genome-wide association screening strategies rely on this indirect testing. On the other hand, as power is inversely related to effect size, the enhanced prior odds applicable if smaller effects prevail would be offset by correspondingly reduced power. The inflation in prior odds resulting from LD is likewise limited by the corresponding reduction in power at indirectly associated variants. Combining these data suggests that a figure of 100 is a reasonable estimate for the effective number of modest effect loci (OR: 1.2–1.3) that are likely to be relevant. These data thus indicate that the prior odds (ie the odds that any randomly selected common variant is relevant) are approximately 100 000 to 1 against.[6] Others have used alternate logic to come to the same figure +/− an order of magnitude.[7] To secure the same level of confidence in 'positive' results that we traditionally associate with the 5% significance threshold we must, therefore, set a significance threshold of approximately 5×10^{-7}. It is only at this extreme P-value that we can adequately compensate for the very low prior probability that any randomly selected variant is in fact genuinely associated.[7] One way to improve the prior odds is to use existing knowledge to guide the selection of variants to study, the so-called candidate gene approach. This ideology has been the cornerstone of the genetic analysis of complex disease for several decades. However, even if all available sources of additional information are used in an exercise called genomic convergence,[8] it is unlikely that prior odds can be improved much beyond 1000 to 1 against.[9] Assuming the logic used to judge a variant as a candidate is sound, then for strong candidate variants, we might be able to relax the significance threshold to 10^{-4}. It is important to draw a distinction between selecting a variant for study on the basis of some preconceived logic regarding its candidature and inventing an apparent explanation for why a variant identified as part of a screening process might be thought of as a candidate. It seems inescapable that the later will have less effect on the prior odds and will, therefore, provide lower posterior odds.

The need to use an extreme significance threshold in the genetic analysis of complex disease is a consequence of the size of the genome and is uninfluenced by the strength of the effects sought. Working on isolated populations or in clinically more refined sub-groups, in which more favourable allele frequencies and/or effect sizes might be hoped for, does not negate the need for an extreme significance threshold. Even if the theorized advantages of these study designs are correct, and the increase in power is able to offset any accompanying reduction in sample size, the required significance threshold cannot be relaxed. Indeed, as most of these strategies are only likely to improve the power to find some of the relevant risk alleles, it could be argued that they effectively reduce the prior odds and, therefore, require even more extreme significance.

In the context of this absolute requirement for an extreme significance threshold, two questions immediately spring to mind.

WHAT SAMPLE SIZE SHOULD BE USED?

The sample size needed to ensure that there is adequate power to identify association at the required significance threshold depends on the strength of the effects being sought. Again, although we cannot know with any certainty what effect sizes will be relevant in a complex disease, whole genome linkage analysis provides some important information, which sets a crude upper limit on these effects (this limit is less restrictive for rarer alleles). After more than a decade of whole genome linkage screening, it is evident that very few common risk alleles are detectable by linkage. In multiple sclerosis, for example, the high-density single-nucleotide polymorphism (SNP)-based whole genome linkage screen performed by the International Multiple Sclerosis Genetics Consortium only found one region of linkage, that due to the well-established association with the *1501 allele of the DRB1 gene from the major histocompatibility complex.[10] No other significant linkage was identified in this well-powered screen. The results from this screen and similar studies in other complex diseases indicate that apart from the few loci, such as those identified by linkage, common variants influencing the risk of complex traits are extremely unlikely to increase risk by more than a factor of 2.0, and most likely by <1.5. At this level, at least 2000 cases and 2000 controls are required to provide power to identify association with a common variant at a significance threshold of 5×10^{-7}.[7] See Supplementary data file for more information.[6]

HOW SHOULD WE INTERPRET INTERMEDIATE RESULTS?

Comparing and contrasting the results from association studies is not always straightforward, as the strength of evidence for association is a complex reflection of both the observed *P*-value and the power of the study. This issue is especially relevant for results falling in the intermediate range, that is in which the *P*-value has more extreme significance than the familiar 5%, but does not quite reach the 5×10^{-7} level. Fortunately, Bayes factors (BFs) provide a single measure of the strength of evidence for association, which appropriately integrates the influences of the observed *P*-value and the power of the study, enabling meaningful ranking of results within and across the studies.

BAYES FACTORS

For a given set of observed data, Eq. (3) shows the relationship between the posterior odds and the prior odds

$$\text{PosteriorOdds} = \left(\frac{P_1}{P_0}\right) * \text{PriorOdds} \qquad (3)$$

where P_1 is the probability of observing this particular set of data if the tested variant is genuinely associated at the population level and P_0 is the probability of seeing the same data if the tested variant is not associated (ie under the null hypothesis). This ratio is known as a Bayes Factor (BF) and is akin to the ratio of power and significance in Eq. (1). The difference here is that the probabilities P_1 and P_0 relate to the particular set of data that has been observed rather than the probability of seeing data at or more extreme than a selected threshold. In many respects, Log_{10}(BF) might be thought of as the association study equivalent of a LOD score in a linkage analysis. Both Log_{10}(BF) and LOD scores are Log_{10} measures of how much more likely it is to see the observed data if the tested variant is genuinely relevant as opposed to the null. Empirical data from linkage analysis has confirmed the theoretical prediction that LOD scores need to be >3.6 if they are to compensate for the low prior odds of linkage and have a high posterior odds of being true.[11] Likewise, we can see that Log_{10}(BF) must be $\gg5$ if a result is to adequately compensate for the even lower prior odds of association.

The difficulty of course is calculating the values of P_1 and P_0, especially the former. As we cannot know for certain what effect is attributable to any given locus, we can only calculate the BF for a given set of data by making assumptions about the underlying effects. If we have tested a bi-allelic variant such as a common SNP with a minor allele frequency of 10%, then if we assume a particular heterozygote OR (eg 1.2) and genetic model, the calculated BF will tell us something about the extent to which the observed data supports this particular possibility. If the Log_{10}(BF) value calculated in this way is $\gg5$, then we can be confident that the observation is likely to be true positive; the more extreme the BF value, the more likely it is to be true. For less extreme Log_{10}(BF) values (ie those ≤5), although the posterior odds will be <1, the results can at least be ranked against other tests in terms of strength of evidence. The *P*-value alone does not always allow this clarity (see Box 1). Further mathematical and practical detail concerning the calculation of BFs is provided in Supplementary data file.

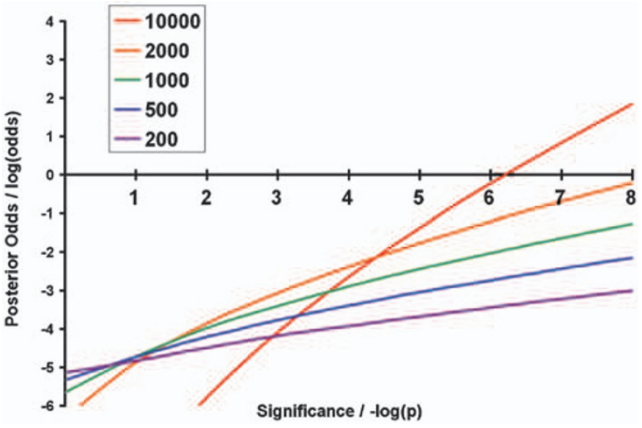

Figure 2 The figure shows the relationship between the Posterior Odds that a result is true (plotted on a Log_{10} scale on the *y* axis) and the observed *P*-value (plotted on a Log_{10} scale on the *x* axis). Five sample sizes are listed in the legend; in each, the number of cases and controls is equal; the 200 line thus indicates the posterior odds for a study involving 200 cases and 200 controls and so on. BFs were calculated using the SNPTEST program [7] assuming that the risk allele has a frequency of 10%, an odds ratio (OR) of 1.2 and follow multiplicative model. The Prior Odds are assumed to be 10^5 against. These curves are calculated assuming that studies are free from imperfections such as genotyping error, population stratification and differential missingness. In real studies, these issues may further confound interpretation and bias results.[6]

If we make the simplifying assumption that all the risk alleles in a complex disease have the same OR and risk allele frequency, then we can produce Figure 2, indicating the posterior odds that would be conferred by any observed *P*-value in this simplified scenario. This figure illustrates the futility of small (under powered) studies. The curves for small studies are close to horizontal, indicating that whatever the result may be, there is little change in the odds in favour of association. If the *P*-value from such a study fails to reach nominal significance, then nothing has been excluded. Likewise, even if the *P*-value exceeds the nominal significance threshold, it is highly likely to be false positive. If the *P*-value is very extreme (eg exceeds the 5×10^{-7} threshold), then one should be highly suspicious of the study methodology. As there is little power to see this level of significance in a study of this size, the result is most likely to reflect some unappreciated bias. The alternative view, that the study has by 'good luck' identified a common allele with a very large effect, is inconsistent with available linkage data and should, therefore, be viewed with considerable suspicion.

Very larger sample sizes, on the other hand, not only provide substantial power to identify levels of significance associated with

Box 1

The hypothetical studies summarized in Table 1 illustrate the value of BFs. In each case, the observed *P*-value is 1% and, therefore, on its own provides no guidance as to which of these studies is the more likely to be a true positive association. The BFs, on the other hand, are substantially different and enable the studies to be ranked in terms of the strength of evidence each indicates.

We can understand why the BFs are different in these hypothetical studies by considering the power to identify the particular effect assumed in calculating these factors. In study 1, the sample size is appropriate for identifying the effect considered, whereas in study 2, the sample size is inadequate and in study 3, the sample size is considerably more than is necessary. These studies are thus, respectively, appropriately powered, underpowered and 'over' powered (at this level of significance and for an effect of this size). When a study is underpowered, the probability of seeing the observed data, if the locus does exert the tested effect (P_1), is little different from the probability of seeing the data by chance alone (P_0). Thus, for underpowered studies, the BF will be close to 1 and Log_{10}(BF) close to 0. Interpreting the result from study 3 is somewhat less intuitively obvious. In this study, the sample size is such that we would expect the *P*-value to be very much more extreme than 1% if this locus really did have an OR of 1.2. For this particular effect, the probability of seeing data resulting in a *P*-value of only 1% (P_1) is actually smaller than the probability of seeing these data by chance alone (P_0). With $P_1 < P_0$, the BF will be <1 and Log_{10}(BF) will be negative. In other words, observing a *P*-value rather less extreme than we would expect, given the assumed effect, actually provides evidence against the locus being truly associated. Study 4 illustrates the important influence that allele frequency has on power; for a variant with an allele frequency of 1%, a study involving 2000 cases and 2000 controls provides very little power so again the Log_{10}(BF) is close to 0.

A BF indicates the degree to which an observed set of data is consistent with an assumed underlying effect. An alternative question is to ask what underlying effect is most consistent with the observed data. Finding the maximum likelihood solution to Eq. (3) to determine the underlying effect that is most consistent with the observed data (and its confidence interval) is a familiar exercise and has the appeal that it requires fewer assumptions than are needed to calculate a BF. However, the results generated in this way will only be meaningful if the posterior odds that result is true are $\gg 1$. If the BF is insufficient to compensate for the low prior odds, then any apparent association is likely to be a false positive and any calculated estimate of the 'most likely' effect size will be virtually meaningless. Considering the final column in the table illustrates this point. Inspection of these calculated values shows a counterintuitive inverse relationship between power (sample size) and estimated effect size; the less powerful the study the greater the estimated effect. In other words, studies that have the least chance to identify any real effect could be interpreted as having 'identified' the most interesting effects!

Unfortunately, the practice of reporting estimated effect sizes for results of intermediate significance (ie those with *P*-values of $<5\%$ but $>5\times10^{-7}$) rather than only calculating these estimates for results with high posterior odds of being true further confounds the interpretation of association studies. It is unfortunate that journals and authors have a tendency to report results such as those from study 2 as 'a significant association identifying a risk allele exerting a large effect', an interpretation that can be misleading.

Table 1 Hypothetical case-control studies showing association with the same significance but different BF

Study	N *cases*	N *controls*	*MAF/%*	P-*value*	*Log_{10}(BF)*[a]	*OR (CI)*
1	2000	2000	10	0.01	1.18	1.20 (1.05–1.39)
2	100	100	10	0.01	0.35	2.11 (1.18–3.78)
3	10 000	10 000	10	0.01	−0.78	1.09 (1.02–1.16)
4	2000	2000	1	0.01	0.58	1.69 (1.14–2.50)

MAF, minor allele frequency; OR, odds ratio; the ratio of the odds of having the disease if you are a heterozygote (ie carry one copy of the risk allele) as compared with the odds of having the disease if you are a homozygote for the wild-type allele (ie carry no copies of the risk allele). For modest effects like these OR ≈ GRR. CI=95% (confidence interval in the OR).
[a]The BFs were calculated assuming that in each study the minor allele is the risk allele (RAF=MAF), the OR=1.2 and a multiplicative model applied.

high posterior odds, but also enable variants, which fail to reach nominal significance to be excluded. For sample sizes in the 10 000 case range, variants that have *P*-values of $>5\%$ have Log_{10}(BF) values that are < -2, indicating that the odds that this variant exerts on the tested effect, have been reduced by more than a factor of 100. The slope and position of these curves are critically dependent on the underlying model assumed in calculating the BFs (see Supplementary data). If we consider smaller underlying effect sizes, then even the 10 000 case line will start to lean over towards the horizontal, indicating that for a study to be discriminating in identifying much smaller effects, even larger samples sizes will be necessary. As we do not actually know the underlying effect sizes, one way to deal with this uncertainty is to calculate P_0 and P_1 for each possible effect size and then integrate these values, weighting each by the probability of that effect size. This process requires us to make some prior assumption about the probability of each effect size. A normal distribution of effects sizes has been suggested such that 30% of the effects have an OR of >1.2, but only 2% have an OR of >1.5 etc.[7] The problem with the BFs calculated in this manner is that most of the underlying effect sizes considered are very small and, therefore, for sample sizes such as those considered in Figure 2, there is little power to identify most of the presumed underlying effects. As a result, such BFs are little different between these studies.

The important influence of allele frequency on power, and, therefore, the BF associated with any given *P*-value, is illustrated in Figure 3. As we would anticipate, there is very little difference in these curves for

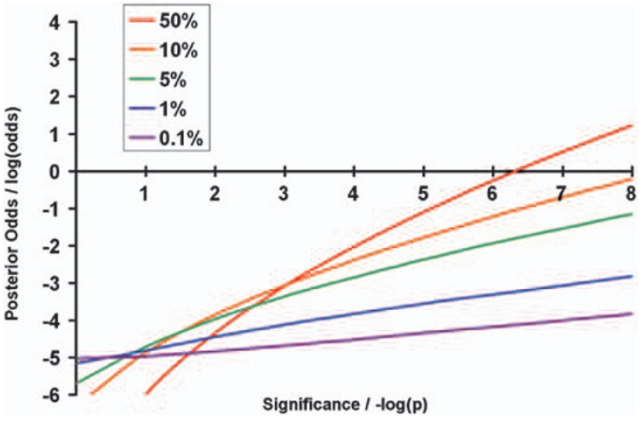

Figure 3 The figure shows the relationship between the Posterior Odds that a result is true (plotted on a Log_{10} scale on the *y* axis) and the observed *P*-value (plotted on a Log_{10} scale on the *x* axis). The curves are based on a study using 2000 cases and 2000 controls and assume differing risk allele frequencies, as shown in the legend. BFs were calculated using the SNPTEST program[7] assuming an odds ratio (OR) of 1.2 and a multiplicative model. The Prior Odds are assumed to be 10^5 against. These curves are calculated assuming that studies are free from imperfections such as genotyping error, population stratification and differential missingness. In real studies, these issues may further confound interpretation and bias results.[6]

common alleles, but as the power drops off significantly for variants with minor allele frequencies of less than a few percent, the curves for these variants are substantially more horizontal.

CONCLUSION

For many years, researchers in complex genetics have naively relied on the traditional interpretation of association studies and assumed that *P*-values of $<5\%$ indicate true positive findings regardless of the sample size considered. It has taken the field some time to realize that two inescapable issues undermine this position and demand a more stringent analysis. First, the extremely low prior odds that any given common variant is relevant (c100 000:1 against) means that a correspondingly more extreme significance threshold must be used before the posterior odds can reliably be assumed to be >1. The fact that complex genetics requires *P*-values of $<5\times10^{-7}$ to produce the same confidence that we traditional associate with the 5% threshold has been a bitter pill to swallow. The second and equally difficult issue is that of effect size. The fact that with very few exceptions, whole genome linkage analysis has failed to identify genes of relevance in complex disease places an upper limit on the size of effects that can be attributable to common variants. These modest effect sizes, especially when combined with the requirement for extreme significance, mean that sample sizes have to be large. For many years, we have based our association studies on a few hundred cases and controls in the belief that the effects being sort would more than double the risk. In reality, very few such loci exist in any given complex trait, and it is now clear that most relevant common variants have OR of <1.3. For effects of this size, studies must involve thousands rather than hundreds of samples.

The fact that extreme levels of significance are necessary to compensate for the low prior odds and that very large sample sizes are needed to provide sufficient power to identify modest effect sizes at these high levels of significance has set a new standard, but has also left a gap in which interpretation of results is less clear. What should we make of the studies that generate intermediate *P*-values ($<5\%$ but $>5\times10^{-7}$)? Interpretation requires consideration of both the *P*-value and the power, which in turn is influenced by sample size and allele frequency. Fortunately, BFs provide a single measure, which integrates these various influences and provide a meaningful single measure regarding the strength of evidence provided by any observed data. Considering the BFs in relation to any particular signal strength allows one to infer to what extent that particular effects has been supported or even excluded by the observed data. The fact that BFs provide a clearer measure for the weight of evidence in favour of association means that they will also help interpretation of whether or not candidate biological pathways are enriched for modest associations. It should be noted that the account presented here relates to case–control studies and that subtle, but potentially important, differences might apply in calculating the BFs for studies with alternate designs.

The Bayesian framework described above is not the only way to interpret the data emerging from complex genetics and is by no means definitive. The method used to estimate the prior odds is crude and the power calculations are based on mathematically convenient models, which have no obvious biological counter part.[12] The frequentist framework provides an alternate way to interpret these data in this approach the significance threshold is adjusted to correct for multiple testing (using methods such as the Bonferroni correction[13] or the false discovery rate[14]) and thereby control the family-wise error rate. Although a frequentist interpretation has the advantage that it avoids the need to estimate prior odds, it turns out to be no more robust, as estimating multiplicity is predictably as crude as estimating prior odds.[15–17] Furthermore, it turns out that the recommended significance thresholds emerging from frequentist analysis are virtually identical to those provided by Bayesian analysis.[18,19] The convergence of these various interpretations is unsurprising as ultimately each is simply trying to adequately compensate for the enormous size of the human genome. Which ever framework of interpretation is preferred it is clear from the available empirical evidence[20] that the recommended thresholds are valid and ignored at a researcher's peril.

CONFLICT OF INTEREST

The author declares no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by the Wellcome Trust (084702/Z/08/Z), the Medical Research Council (G0700061), the National Institute of Health (RO1 NS049477) and the Cambridge NIHR Biomedical Research Centre. I thank all my colleagues in the International Multiple Sclerosis Genetics Consortium (IMSGC) and the Wellcome Trust Case Control Consortium (WTCCC) for their support and tireless efforts to move the genetics of multiple sclerosis forward. I would especially thank Hywel Jones, An Goris, David Clayton and Maria Ban for their careful scrutiny of the manuscript and their helpful comments.

1 Doll R, Hill AB: Smoking and carcinoma of the lung; preliminary report. *Br Med J* 1950; **2**: 739–748.
2 Kruglyak L, Nickerson DA: Variation is the spice of life. *Nat Genet* 2001; **27**: 234–236.
3 International HapMap Consortium: The International HapMap Project. *Nature* 2003; **426**: 789–796.
4 Yang Q, Khoury MJ, Friedman J, Little J, Flanders WD: How many genes underlie the occurrence of common complex diseases in the population? *Int J Epidemiol* 2005; **34**: 1129–1137.
5 Lindsey JW: Familial recurrence rates and genetic models of multiple sclerosis. *Am J Med Genet A* 2005; **135**: 53–58.
6 Sawcer S: The complex genetics of multiple sclerosis: pitfalls and prospects. *Brain* 2008; **131**: 3118–3131.
7 Wellcome Trust Case Control Consortium (WTCCC): Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007; **447**: 661–678.
8 Hauser MA, Li YJ, Takeuchi S *et al*: Genomic convergence: identifying candidate genes for Parkinson's disease by combining serial analysis of gene expression and genetic linkage. *Hum Mol Genet* 2003; **12**: 671–677.
9 Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N: Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst* 2004; **96**: 434–442.
10 International Multiple Sclerosis Genetics Consortium (IMSGC): A high-density screen for linkage in multiple sclerosis. *Am J Hum Genet* 2005; **77**: 454–467.
11 Lander E, Kruglyak L: Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* 1995; **11**: 241–247.
12 Clayton DG: Prediction and interaction in complex disease genetics: experience in type 1 diabetes. *PLoS Genet* 2009; **5**: e1000540.
13 Bonferroni CE: Teoria statistica delle classi e calcolo delle probabilita. Pubblicazioni del R. *Istituto Superiore di Scienze Economiche e Commerciali di Firenze* 1936.
14 Benjamini Y, Hochberg Y: Controlling the false discovery rate – a practical and powerful approach to multiple testing. *J R Stat Soc B* 1995; **57**: 289–300.
15 Risch N, Merikangas K: The future of genetic studies of complex human diseases. *Science* 1996; **273**: 1516–1517.
16 Thomas DC, Clayton DG: Betting odds and genetic associations. *J Natl Cancer Inst* 2004; **96**: 421–423.
17 Pe'er I, Yelensky R, Altshuler D, Daly MJ: Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol* 2008; **32**: 381–385.
18 Dudbridge F, Gusnanto A: Estimation of significance thresholds for genomewide association scans. *Genet Epidemiol* 2008; **32**: 227–234.
19 Wakefield J: Bayes factors for genome-wide association studies: comparison with P-values. *Genet Epidemiol* 2009; **33**: 79–86.
20 Hindorff LA, Sethupathy P, Junkins HA *et al*: Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 2009; **106**: 9362–9367.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)