RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

SNP Prioritization Using a Bayesian Probability of Association

John R. Thompson,[1]* Martin Gögele,[2] Christian X. Weichenberger,[2] Mirko Modenese,[2] John Attia,[3,4] Jennifer H. Barrett,[5] Michael Boehnke,[6] Alessandro De Grandi,[2] Francisco S. Domingues,[2] Andrew A. Hicks,[2] Fabio Marroni,[7] Cristian Pattaro,[2] Fabrizio Ruggeri,[8] Giuseppe Borsani,[9] Giorgio Casari,[10] Giovanni Parmigiani,[11,12] Andrea Pastore,[13] Arne Pfeufer,[14] Christine Schwienbacher,[2] Daniel Taliun,[2] CKDGen Consortium,[1] Caroline S. Fox,[15,16] Peter P. Pramstaller,[2,17,18] and Cosetta Minelli[2]

[1]Department of Health Sciences, University of Leicester, Leicester, United Kingdom; [2]Center for Biomedicine, European Academy Bozen/Bolzano (EURAC), Bolzano, Italy; [3]Centre for Clinical Epidemiology and Biostatistics, Hunter Medical Research Institute, The University of Newcastle, Newcastle, NSW, Australia; [4]Department of General Medicine, John Hunter Hospital, Newcastle, NSW, Australia; [5]Section of Epidemiology and Biostatistics, Leeds Institute of Molecular Medicine, University of Leeds, Leeds, United Kingdom; [6]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan; [7]Institute of Applied Genomics, Udine, Italy; [8]National Research Council, Institute of Applied Mathematics and Information Technology, Milan, Italy; [9]Department of Biomedical Sciences and Biotechnology, University of Brescia, Brescia, Italy; [10]Vita-Salute San Raffaele University and Center for Translational Genomics and Bioinformatics, Milan, Italy; [11]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts; [12]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts; [13]Department of Economics, Ca' Foscari Venezia University, Venezia, Italy; [14]Department of Bioinformatics and Systems Biology IBIS, Helmholtz Zentrum Munich, German Research Center for Environmental Health (GmbH), Neuherberg, Germany; [15]The National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, Massachusetts; [16]Center for Population Studies, Framingham, Massachusetts; [17]Department of Neurology, Central Hospital, Bolzano, Italy; [18]Department of Neurology, University of Lübeck, Lübeck, Germany

Received 10 August 2012; Revised 30 October 2012; accepted revised manuscript 22 November 2012.
Published online 26 December 2012 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21704

ABSTRACT: **Prioritization is the process whereby a set of possible candidate genes or SNPs is ranked so that the most promising can be taken forward into further studies. In a genome-wide association study, prioritization is usually based on the *P*-values alone, but researchers sometimes take account of external annotation information about the SNPs such as whether the SNP lies close to a good candidate gene. Using external information in this way is inherently subjective and is often not formalized, making the analysis difficult to reproduce. Building on previous work that has identified 14 important types of external information, we present an approximate Bayesian analysis that produces an estimate of the probability of association. The calculation combines four sources of information: the genome-wide data, SNP information derived from bioinformatics databases, empirical SNP weights, and the researchers' subjective prior opinions. The calculation is fast enough that it can be applied to millions of SNPS and although it does rely on subjective judgments, those judgments are made explicit so that the final SNP selection can be reproduced. We show that the resulting probability of association is intuitively more appealing than the *P*-value because it is easier to interpret and it makes allowance for the power of the study. We illustrate the use of the probability of association for SNP prioritization by applying it to a meta-analysis of kidney function genome-wide association studies and demonstrate that SNP selection performs better using the probability of association compared with *P*-values alone.**
Genet Epidemiol 37:214–221, 2013. © 2012 Wiley Periodicals, Inc.

KEY WORDS: **replication; prior knowledge; genome-wide studies**

Introduction

Prioritization is the process whereby a set of possible candidate genes or SNPs is ranked so that the most promising can be taken forward into further studies. Usually prioritization is based exclusively on *P*-values, but it is possible to supplement *P*-values with external information extracted from bioinformatics databases on features thought to identify those genes or SNPs that are likely to be associated with the disease or trait under study. Recently, Gögele et al. surveyed the methods used by researchers meta-analyzing genome-wide association studies (GWAS) and found that just over three-quarters of the meta-analyses selected the SNPs to be sent for replication based entirely on their *P*-values [Gögele et al., 2012]. The remaining quarter of the studies also took into account biological factors that the researchers believed would improve the chance that the SNP would turn out to

Supporting Information is available in the online issue at wileyonlinelibrary.com.

*Correspondence to: John R. Thompson, Professor of Genetic Epidemiology, Department of Health Sciences, Room 214f, Adrian Building, University of Leicester, Leicester LE1 7RH, UK. E-mail: john.thompson@le.ac.uk

have a genuine effect; for instance, some used information on whether or not the SNP is close to a good candidate gene. However, this external information was used in diverse and largely unstructured ways.

The use of external biological information for selecting good candidate SNPs is intuitively sensible but it is not obvious how best to combine this information with the statistical evidence from the GWAS itself. This leaves us with difficult questions, such as, is there a better chance of replicating a SNP with a *P*-value of 10^{-6} that lies in a gene desert, or one with a *P*-value of 10^{-5} that lies in or near a candidate gene? A full answer to this question would require a Bayesian analysis in which the researchers carefully quantified the biological information in the form of prior distributions for each SNP. This ideal Bayesian analysis would be hard to implement on a large scale, partly because of the difficulty of specifying all of the priors and partly because of the time needed to perform the Bayesian calculations on large numbers of variants.

In a companion paper, Minelli et al. describe the use of experts' opinions and empirical evidence to estimate the relative importance of 14 types of SNP or gene information that can be extracted from bioinformatics databases [Minelli et al., 2013]. This work provides weights that can be used to combine the various types of external information into a single measure of the relative probability that a SNP is associated with the disease or trait. Here we consider a fast, approximate Bayesian analysis that combines those prior relative probabilities with the data from a GWAS to calculate the posterior probabilities of association.

The proposed method for calculating the probability of association allows us to combine GWAS data with information from bioinformatics databases in a coherent and reproducible way. We show that the probability of association has many advantages when compared with the *P*-value; it is more easily understood, it takes account of the power of the study and it can incorporate expert knowledge and nonexperimental information about the variant. Our work is illustrated by an application to actual data from a meta-analysis of GWAS of kidney function in which the probability of association based on prior information extracted from bioinformatics databases outperforms the *P*-value when it is used to rank SNPs for future investigation.

Methods

We used an approximate Bayesian analysis, closely related to a previously reported modified form of the False Positive Report Probability (FPRP), to combine prior knowledge with GWA statistical evidence. The FPRP was suggested by Wacholder et al. as an aid to interpreting *P*-values in studies that involve a large number of significance tests [Wacholder et al., 2004]. The motivation for this approach lays in the common observation that many statistically significant findings in epidemiological studies are never replicated and eventually become accepted as false positives. Wacholder et al. argued that the FPRP measures the probability that a significant finding will eventually turn out to have been due to chance and described how the FPRP depends on three quantities: the *P*-value, the power of the test, and the fraction of tested null hypotheses that are actually null. Tests with an FPRP less than some preset threshold, such as 50%, are considered noteworthy and thus are candidates for further investigation.

The method for calculating the FPRP assumes that the underlying size of the effect is either zero (null hypothesis, H_0) or some specified alternative H_A with effect size T_A, and that variation about those values is due entirely to sampling. If z_α is the critical value for a one-tailed test, then, for a test statistic, T, the significance level is $\alpha = P(T > z_\alpha | H_0)$, and $1 - \beta = P(T > z_\alpha | H_A)$ represents the power. Assuming that a proportion π of tests are performed for which the alternative is true, the probability that H_0 is true given that the test is significant is,

$$\text{FPRP} = P(H_0 \mid T > z_\alpha) = \alpha(1-\pi)/[\alpha(1-\pi) + (1-\beta)\pi]$$

For the purposes of a replication study, it is more natural to talk in terms of the probability that a significant SNP is truly positive, or, put another way, the chance that it would replicate in an ideal (very large) replication study with a strict criterion for declaring replication. This probability is $P(H_A | T > z_\alpha)$ or 1-FPRP.

The FPRP has been criticized for two main reasons, [Lucke, 2009 and Thomas and Clayton, 2004]. First, the formula for the FPRP was thought to be an oversimplification because it is based on a choice between two completely specified hypotheses. However, Thomas and Clayton acknowledged that in practice the inaccuracies caused by this simplification are likely to be small. Second, when evaluating the results of a study, we know not only whether a test is significant or not, but also the actual *P*-value, and once the *P*-value is known the calculation of the FPRP is no longer appropriate. Care is needed in moving from statements about the class of all tests with, say, a *P*-value below 0.01, to statements about a specific test that has a *P*-value of exactly 0.01. Power, and hence the FPRP, is directly relevant to the former, but not the latter.

A joint analysis of the Wellcome Trust Case-Control Consortium study of Coronary Artery Disease and the German Myocardial Infarction Study addressed these objections to the FPRP by modifying it for use with a specific known *P*-value [Samani et al., 2007]. That modification produced an approximate Bayesian analysis against a fixed alternative hypothesis, but, as we will see, further modifications can be used to investigate more flexible alternatives. The modified version of the FPRP depends on a prior assessment of the probability, π_i, that the i^{th} SNP is associated with the outcome. Assuming that under the null hypothesis of no association, the distribution of the effect size estimate is approximately normal, the probability of association, A_i, associated with a one-tailed test is,

$$\begin{aligned} A_i &= P(H_A \mid P - value = \alpha) = P(H_A \mid T_i = z_\alpha) \\ &= \pi_i \varphi(T_i; T_A, \sigma_i)/[(1-\pi_i)\varphi(T_i; 0, \sigma_i) + \pi_i \varphi(T_i; T_A, \sigma_i)] \end{aligned}$$

where T_i is the observed effect size and $\varphi(y;m,s)$ represents the density function of the normal distribution with mean m and SD s evaluated at y, and σ_i is the SE of the estimate of

effect size for that SNP. Effectively, this formula replaces the tail areas in the FPRP with likelihood ratios. Importantly, the interpretation also changes and A_i represents the probability of association given the actual experimental data on that SNP.

The calculation of the probability of association, A_i, depends on the researcher specifying a guess at T_A, the effect size if the SNP is associated. Specifying this quantity can be difficult and giving the wrong value could potentially cause the analysis to overlook an important SNP. It is both easier and more realistic to say that the anticipated effect size under the alternative hypothesis is in a range, such as, T_A is normally distributed with mean μ_A and SD σ_A, in which case the one-tailed A_i becomes,

$$\pi_i\varphi\left(T_i;\mu_A,\sqrt{\left(\sigma_i^2+\sigma_A^2\right)}\right)\Big/ \times\left[(1-\pi_i)\varphi(T_i;0,\sigma_i)+\pi_i\varphi\left(T_i;\mu_A,\sqrt{\left(\sigma_i^2+\sigma_A^2\right)}\right)\right]$$

When we do not know in advance whether the chosen effect allele will increase or decrease the outcome measure, it would be natural to use a two-tailed procedure. Provided that we believe in advance of seeing the data that the effect of a particular SNP is equally likely to be positive or negative, then the formula for calculating A_i becomes,

$$\left\{\begin{array}{l}\left[\pi_i/2\ \varphi\left(T_i;-\mu_A,\sqrt{\left(\sigma_i^2+\sigma_A^2\right)}\right)+\pi_i/2\ \varphi\left(T_i;\mu_A,\sqrt{\left(\sigma_i^2+\sigma_A^2\right)}\right)\right]\Big/ \\ \left[(1-\pi_i)\ \varphi(T_i;0,\sigma_i)+\pi_i/2\ \varphi\left(T_i;-\mu_A,\sqrt{\left(\sigma_i^2+\sigma_A^2\right)}\right)\right. \\ \left.\quad+\pi_i/2\ \varphi\left(T_i;\mu_A,\sqrt{(\sigma_i^2+\sigma_A^2)}\right)\right]\end{array}\right\}$$

This measure of the probability of association is a very close approximation to the posterior mean of a Bayesian analysis of the effect size using a mixture prior in which a large proportion of SNPs show no association with the disease or trait under study, although the remainder are equally likely to show a harmful effect as a protective effect. The association probability is similar to Wakefield's Bayesian False Discovery Probability (BFDP), which is also an extension of the FPRP that approximates a Bayesian analysis, although Wakefield used different priors [Wakefield, 2007]. We demonstrate the equivalence of our approximation and its Bayesian equivalent in the Supporting Materials and show how extra information can be obtained if the full Bayesian analysis is performed. However, Bayesian analyses are often time consuming to perform so we recommend that A_i be calculated for all SNPs and that the full Bayesian analysis be reserved for the top ranked SNPs.

Results

Hypothetical Data

To illustrate the use of the probability of association, we will first consider a hypothetical GWAS for a binary trait that recruits 2,000 cases and 2,000 controls. In such a study, a SNP with a minor allele frequency (MAF) of 0.25 in controls would have an SE for the allelic log odds ratio of about $\sigma = 0.052$. This SE is not much affected by small changes in the allele frequency. The bottom part of Table 1 shows the probability of association for such a SNP over a range of *P*-values and prior beliefs, π. In this example, our prior assessment of the likely effect size was that the log odds ratio for the associated SNPs can be described by a normal distribution with mean $\mu_A = 0.15$ and SD $\sigma_A = 0.05$. This implies that we anticipate that 95% of associated SNPs will have odds ratios, or their inverses, between 1.05 and 1.28. The variation in *P*-values arises because observed effect sizes will vary even if SNPs are drawn from a common prior.

Table 1. Probabilities of association expressed as percentage for a range of *P*-values and prior probabilities of association. The calculations are for a SNP with an MAF of 0.25 and an anticipated log odds ratio of about 0.15 (OR ≈ 1.16)

	P-value			
π_i	10^{-5}	10^{-6}	10^{-7}	5×10^{-8}
1,000 cases and 1,000 controls				
1/10,000	7.0	21.5	48.7	57.8
1/5,000	27.3	57.8	82.6	87.3
1/1,000	42.9	73.2	90.5	93.2
1/500	79.1	93.2	97.9	98.6
1/100	88.4	96.5	99.0	99.3
2,000 cases and 2,000 controls				
1/10,000	22.2	61.6	89.5	93.4
1/5,000	58.8	88.9	97.7	98.6
1/1,000	74.1	94.1	98.8	99.3
1/500	93.5	98.8	99.9	99.9
1/100	96.7	99.4	99.9	99.9

Often SNPs are sent to replication with *P*-values of 10^{-5} or even less significant and we can see from the results in Table 1 that, were there little prior biological reason to make us suppose that this SNP is associated with the outcome, and, say, we believed in advance that only 1/10,000 of such SNPs is associated, then the probability of association in a very large replication study would only be 22%. However, if our biological knowledge were such that the SNP fell into a class for which we believed that 1/1,000 were associated, then for the same *P*-value we would expect that 74% of such SNPs would eventually replicate. Table 1 also shows that under these conditions we would expect that a SNP with a prior probability of 1/1,000 and a *P*-value of 10^{-5} would be slightly more likely to be associated than a SNP with a *P*-value of 10^{-6} and a prior probability of 1/10,000.

The top half of Table 1 shows the corresponding association probabilities for a smaller study of 1,000 cases and 1,000 controls. In such a study, a SNP with an MAF of 0.25 would have an SE of $\sigma = 0.075$. A similar SE would be found for a SNP in the larger study if it had an MAF of about 0.10. So the upper part of the table applies equally to a SNP with an MAF of 0.25 in the smaller study or a SNP with an MAF of 0.10 in the larger study. Here we see that, in lower powered studies, findings are less likely to replicate even when they have the same *P*-value as in better powered studies. In the lower powered study, a SNP with little supporting biological justification and a prior probability of 1/10,000 has only a 58% chance of replicating even when it reaches genome-wide significance (5×10^{-8}). This illustrates the danger of overinterpreting a genome-wide

significant *P*-value when the study is small and prior belief is weak.

An important consideration when interpreting the results in Table 1 is that we have not changed the expected effect size of the SNP, that is, throughout the table we assume an anticipated log odds ratio of around 0.15 (OR $\approx$ 1.16). When a low-frequency variant, or a common variant in a smaller study, reaches genome-wide significance, it must be that the observed log odds ratio was quite large, probably larger than the anticipated value used to create Table 1. This analysis treats such large observed effects with suspicion and asks only if they are more likely to have been generated by an unassociated SNP or one with an effect size of about 0.15. Had we, for instance, anticipated that rare SNPs will have larger effects [Manolio et al., 2009], then the association probabilities would have to be recalculated.

Setting Prior Probabilities Using SNP Characteristics

Minelli et al. [2013] created a training set of SNPs confirmed by replication as being related to one of seven diseases together with matched sets of 1,000 randomly selected control SNPs. Each of the 223 disease related and 7,000 control SNPs was then classified for the presence or absence of the 14 SNP characteristics listed in Table 2. Question 10 was omitted from the original list of 15 questions because of the lack of availability of usable evidence concerning linkage studies. A logistic regression model was used to represent the log odds of being a disease-associated SNP in terms of disease group and all of the characteristics, Q_k, $k = 1 \ldots 14$.

$$\text{Log odds of association in the training set} = \sum_{j=1}^{7} \alpha_j \, disease_j + \sum_{k=1}^{14} \beta_k Q_k \qquad (1)$$

The estimated log odds ratios, $\hat{\beta}$, and the corresponding odds ratios are given in Table 2. For the training data set, the model has a pseudo-R^2 of 21%. The coefficients are estimated under the assumption that the importance of a SNP characteristic does not depend on the disease under study. The log odds ratios provide a set of weights for predicting disease association from SNP characteristics. These log odds ratios are similar to the independent empirical estimates for the characteristics given by Minelli et al., except that those estimates were obtained by examining each SNP characteristic separately.

Using our weights, β, we can estimate the relative probability that a SNP is disease associated with the outcome given its SNP characteristics, but as the training set has a case-control design, it cannot give the constant in the linear predictor needed to give the absolute probability of disease association for a given SNP. Suppose that we wish to assess the prior probability of association for each member of a test set of SNPs. That is we need to estimate μ in the linear predictor,

$$\text{Estimated log odds of association for test SNP i} = \hat{\mu} + \sum_{k=1}^{14} \hat{\beta}_k Q_{ik}$$

To estimate μ, we make a subjective estimate of the proportion of all SNPs in the test set that are related to the disease and then use the characteristics of the test SNPs in a simple search algorithm to find an estimate of μ such that it correctly predicts the subjectively anticipated proportion of associated SNPs.

The method for predicting a SNP's prior probability of association can be used with any set of characteristics and weights, β. In particular, we investigated in sensitivity analyses the use of weights based on either independent empirical estimates or expert opinion as reported by Minelli et al., as well as weights based on empirical estimates for a reduced set of the seven most informative questions, obtained through a stepwise procedure to eliminate SNP characteristics that do not contribute to the modeling of the training set (Table 2).

Illustrative Example: A GWAS of Kidney Function

If we have prior biological evidence that is actually predictive of SNPs that are associated with the outcome, then it must be better to use that information than to ignore it. The open

Table 2. Log odds ratio (OR) and SE used when calculating the relative probability of true association based on SNP characteristics calculated from the training set constructed by Minelli et al. [2013]. Stepwise selection was performed using a forward criterion that added characteristics that were significant at the 5% level

External information	Log OR (SE)	OR	Stepwise log OR (SE)
Q1: SNP in transcribed but not translated region	−0.60 (0.22)	0.55	−0.60 (0.19)
Q2: SNP in a translated region that does not change the amino acid	−1.19 (0.87)	0.31	
Q3: SNP changes the amino acid but not in a functional domain	0.30 (0.62)	1.35	
Q4: SNP in a functional protein domain	0.53 (0.68)	1.70	
Q5: SNP in a regulatory region that is not transcribed	−0.06 (0.42)	0.94	
Q6: SNP in a transcribed regulatory region	0.52 (0.26)	1.67	0.50 (0.25)
Q7: SNP in a genomic region conserved in vertebrates	0.90 (0.39)	2.46	0.87 (0.32)
Q8: SNP in a gene (±5 kb) investigated with this phenotype in >1 study	3.44 (0.24)	31.15	3.38 (0.22)
Q9: SNP in a gene (±5 kb) investigated with this phenotype in one study	1.51 (0.30)	4.54	1.51 (0.30)
Q11: SNP in a locus close to other SNPs investigated with this phenotype	0.96 (0.22)	2.61	0.97 (0.22)
Q12: SNP in a gene (±5 kb) associated with the phenotype in mouse models	0.04 (1.11)	1.04	
Q13: SNP in a gene (±5 kb) expressed in tissue relevant to the phenotype	0.86 (0.26)	2.35	0.87 (0.26)
Q14: SNP in a gene (±5 kb) encoding a protein in a pathway relevant to the phenotype	−0.40 (0.22)	0.67	
Q15: SNP in a gene (±5 kb) with protein–protein interactions relevant to the phenotype	0.23 (0.19)	1.25	

questions are whether the model that underlies our calculation of the probability of association is sufficient to capture the way that *P*-values and biological knowledge should be combined and specifically whether the priors could be so misleading that it would be more sensible to rely on the *P*-values alone. To investigate this, we performed a study using data from a GWA investigation of kidney function from the CKDGen consortium [Köttgen et al., 2010], combining their data with relative probabilities based on the work of Minelli et al. [2013].

The glomerular filtration rate (GFR) is a measure of kidney function and has been studied in many GWA and linkage studies. The CKDGen data used here relate to GFR estimated from serum creatinine (eGFR-crea) and are based on a meta-analysis of 67,093 subjects from 20 studies [Köttgen et al., 2010]. This meta-analysis confirmed the top 28 regions as being truly associated, although presumably there are other associated regions with smaller effects that were not replicated. We compared the published results from the final meta-analysis with the findings of the 10 smallest studies of eGFR-crea from the same consortium that together had a total sample size of 9,434 subjects. Our interest was in seeing how well this discovery sample would predict the final results of the full meta-analysis and whether this performance would be improved by the inclusion of information about the SNP characteristics.

We took the top 100,000 SNPs from the meta-analysis of 9,434 subjects and used a bioinformatics tool to obtain answers to the 14 questions suggested by Minelli et al. [2013]. These questions and the weights used to combine them are summarized in Table 2. The weights enable us to calculate the relative probabilities of association, but to use them in the calculations, they need to be turned into SNP-specific prior probabilities of association and this requires us to specify our beliefs about the proportion of the 100,000 SNPs that are truly associated with kidney function. On top of this, we must specify the average and SD of the anticipated effect size among the associated SNPs. We based these subjective priors on information taken from an earlier publication and assessed their impact on the final SNP selection in a sensitivity analysis. In setting these priors, it is important to allow for the potential impact of the winner's curse that will tend to exaggerate the effect sizes in discovery samples from previous studies [Zollner and Pritchard, 2007].

Köttgen et al. report summary data for eGFR-crea in six populations and these all had means of about 80 ml/min/1.73 m^2 and SDs of about 20 ml/min/1.73 m^2 [Köttgen et al., 2009]. The analysis of eGFR-crea is usually conducted on a log scale in order to remove the effects of skewness, and on this scale, the variance will be about 0.0625 ($20^2/80^2$). They identified four loci that together explained about 0.7% of the phenotypic variance. So we might anticipate that typical associated loci will each explain say 0.1% to 0.2% of the variance and 0.1% would correspond to a variance of 6.25×10^{-5}. A SNP with an allele frequency, f, that is scored 0,1,2 will have a variance of $2f(1-f)$ and the variance explained by that SNP will be approximately $2\ \beta^2\ f(1-f)$, where β is the regression slope. Assuming an allele frequency of 0.25 and equating this variance to the anticipated figure of 6.25×10^{-5} allows us to calculate the expected slope, which is 0.013, corresponding to a 1.3% rise in eGFR-crea per risk allele. After making a range of similar calculations based on slightly different assumptions, we decided to adopt a normal prior for the regression slopes among associated SNPs that had a mean of 0.015 and an SD of 0.005.

We chose to suppose that 1/1,000 of the top 100,000 SNPs is truly associated. This would suggest 100 associated SNPs, corresponding to a smaller number of associated loci because of linkage disequilibrium. As with all priors, the anticipated proportion of associated SNPs needs to be selected without reference to the experimental data that is to be analyzed.

In order to make the findings clearer by limiting the impact of linkage disequilibrium, the 100,000 SNPs were pruned by dividing them into regions separated by at least 50 k base-pairs. This gave just under 10,000 regions containing an average of ten SNPs, and we assumed that our priors apply equally to these pruned data. From within each region, the best SNP (according to either *P*-value or probability of association) was selected and these top SNPs were ranked. The rank in the final meta-analysis was based on the best final *P*-values over the same regions.

Table 3 shows the result of selecting the best region according to discovery *P*-value alone and contrasts them with results for selection based on the probability of association. The shaded cells of Table 3 show that had the top 50 regions from the discovery sample been sent to replication, using the *P*-value alone, only one of the eventual top ten regions would have been included, although basing the selection on the probability of association would have included six of the eventual top ten regions. In general, it is noticeable that in small studies of weak effects, *P*-values alone do not provide a particularly reliable guide to which SNPs will eventually replicate, but that using prior information does improve performance. Results concerning the association of the individual SNPs (rather than regions) can be found in the Supporting Materials, Table S1.

Table 3. Number of regions falling in each category of rank comparing (A) *P*-value alone; (B) probability of association of a discovery sample of 9,434 subjects against the same regions in a meta-analysis of 67,093 subjects (final rank)

	Discovery rank (*P*-values alone)			
Final rank	1–10	11–50	51–100	100+
(A)				
1–10	0	1	0	9
11–50	1	1	0	37
51–100	0	1	2	47
100+	9	37	48	9,279
	Discovery rank (probability of association)			
Final rank	1–10	11–50	51–100	100+
(B)				
1–10	2	4	1	3
11–50	2	0	1	36
51–100	0	2	2	46
100+	6	34	46	9,287

In Table 3, it can be seen that two regions from the eventual top ten are not identified by *P*-value alone but are in the two ten regions when ranked by the probability of association. These correspond to 15q21.1 and 16p12.3 both of which were reported by Köttgen as known loci [Köttgen et al., 2010]. These regions would only have been ranked as the 30th and 172nd most promising based on the best regional *P*-values from the smaller discovery study of 9,434 subjects. However, within 15q21.1, there was a SNP with characteristics 8, 11, 13, 14, and 15 (defined in Table 2), this gave it a prior probability of association of 0.07 that when combined with its *P*-value of 2.7×10^{-5} produced an association probability of 0.97. Similarly, in 16p12.3 region, there was a SNP with characteristics 8, 11, and 13 that gave it a prior of 0.08 and combining this with its *P*-value of 2.2×10^{-4} produced an association probability of 0.96. It is clear from the weight given to characteristic 8, which identifies SNPs close to genes that had previously been investigated in association with this phenotype in more than one study, that the priors will be heavily influenced by known or well-investigated genes. Although this example illustrates that the association probability works in principle, the usefulness of the technique for discovering new regions will depend on the ability to identify SNP characteristics that are less dependent on previously confirmed associations.

In the Supporting Materials, we show the effect of the discovery sample size by repeating this experiment with smaller and then larger discovery samples. The results (Supporting Materials Tables S2 and S3) are broadly similar, with larger discovery samples identifying more replicating regions but benefiting relatively less from the information on SNP characteristics. Minelli et al. suggest, that for outcomes that have not been studied thoroughly in the past, independent weights might also be considered, that is weights based on each characteristic considered separately. In a further sensitivity analysis, we looked at the effects of using such independent weights (Supporting Materials Tables S4 and S5). Although the results were quite similar, neither the independent empirical weights, nor the independent weights based on expert opinion, performed quite as well as the weights from the joint logistic regression given in Table 2. The weights based on stepwise selection of the most important seven questions gave results very similar to those for the full set of 14 questions.

Other researchers might take a different view of the priors, which would lead them to different beliefs in the probabilities of association, so it is important to conduct a sensitivity analysis for one's choices of prior and such an analysis is included in the Supporting Materials as Table S7. The results are not changed greatly by making π five times larger or five times smaller, by increasing or decreasing the mean anticipated effect size by a third or by increasing or decreasing the SD by 50%.

Finally, we were concerned that our example might artificially exaggerate the impact of the prior evidence. Questions 8 and 9 from Minelli et al. assess whether a SNP is in a previously identified gene [Minelli et al., 2013]. In our GFR example, it is likely that some of the data that went into the final meta-analysis had previously been published elsewhere, so it might have been picked up in the database searches for this question [Minelli et al., 2013]. We repeated the analysis excluding papers based on studies subsequently included in the final meta-analysis, but found that the results were identical to those in Table 3.

Full Bayesian Analysis

The association probabilities in Table 1 are very close to those that would be obtained from a full Bayesian analysis in which the observed effect size for a given SNP is assumed to have been drawn from a mixture of three distributions centered at 0 and $\pm\mu$ with probabilities $1 - \pi_i$, $\pi_i/2$, and $\pi_i/2$, and for which the value μ is itself drawn from a normal distribution with known mean μ_A and SD σ_A. The OpenBUGS (http://www.openbugs.info/w/FrontPage) code for fitting this Bayesian model is given in the Supporting Materials along with the results corresponding to those in Table 1 (Supporting Materials Table S8). The full Bayesian analysis for this table took 8 min to complete in OpenBUGS, meaning that the full analysis would be extremely time consuming if applied to an entire GWAS. The posterior means from the full Bayesian analysis are almost identical to the estimates shown in Table 1, so the probability of association using the formula given at the start of this section can be thought of as an approximation to the posterior mean from a full Bayesian analysis.

The type of prior used for the effect size is a form of the so-called spike-and-slab prior that has a long history in Bayesian variable selection [Mitchell and Beauchamp, 1988], although the prior probability, $\pi_{i,}$ that the SNP is associated can be derived from the SNP characteristics in the same way that it was for the approximation.

It is possible to generalize the Bayesian model. For instance, if we were uncertain about the prior assessment of being associated, we might decide to replace the fixed value π_i with a distribution. If the priors come from the SNP characteristics, then the prior distribution might be chosen to reflect the uncertainty in the prediction resulting from the training set and from the subjective assessment of the overall proportion of associated SNP is the test set.

The association probability calculated by the Bayesian mixture model or by our approximation A_i would apply to the results of an idealized replication study with very high power. However, we could use the results of the full Bayesian analysis to answer questions concerning the power in a study of finite size. First, we could estimate the SE that would be obtained in the replication study of a given size, for instance, a replication sample that is the same size as the discovery would be expected to have the same SE. Then, during the MCMC run, we could take the simulated values of μ when the effect is nonzero and simulate a new observed effect from a normal distribution with mean μ and that SE, although when the effect size is simulated to be zero, we could sample the observed effect from a zero-centered normal distribution with that SE. These simulated replication values could then be compared with whatever criterion for replication the user wished

in order to estimate the power for that sized replication study.

To illustrate the power calculation, we used the example of a SNP with an observed effect size of 1.30 and an SE of 0.0594 (giving a *P*-value of 1×10^{-5}) and $\pi = 0.001$. We used the results of the full Bayesian analysis of length 20,000 based on the code given in the Supplement and simulated a random study result for each item in the chain. That is, the measured study value of the logOR was randomly generated to be either N (μ,0.0594^2) or N (0,0.0594^2). Assuming that replication is declared when a study finds abs(logOR*/0.0594) > 1.96 then for the zero-centered simulations, the power was 4.6% (close to the nominal 5%) and for the nonzero-centered simulations, the power was 84.6%. In order to allow for our uncertainty about whether the SNP is truly associated or not, we can use all 20,000 simulations and find that we satisfy the replication criterion 57.2% of the time. This contrasts with the association probability of 65.5% and reflects the extra error that comes from relying on the results of a finite-sized replication study.

Discussion

A *P*-value summarizes the sampling distribution of the experimental data assuming that the variant is not associated with the disease or trait under study. In contrast, the probability of association is more directly relevant to prioritization because it describes the probability of true association given the experimental data. Perhaps because it is not directly relevant to assessing association, the *P*-value is widely misinterpreted and has frequently been criticized [Goodman, 2007; Thompson, 1998]. On top of its theoretical advantages, the probability of association also has very practical advantages because it can incorporate prior knowledge, and genetics is a field of study in which there is an ever-expanding resource of publicly available data that could be taken into account when analyzing an experiment. We show how the inclusion of prior knowledge using an approximate Bayesian analysis can improve SNP prioritization and increase the success of replication, particularly when the discovery sample is small. Despite current international efforts to pool GWA data across a large number of studies, limited discovery sample size is still a critical issue for rarer disease outcomes or phenotypes that are difficult to measure.

We chose to base our priors on information extracted from bioinformatics databases and combined in the way advocated by Minelli et al., but the method of analysis could be adapted to work with other types of external information provided that they can be converted onto the scale of the relative probability of association [Minelli et al., 2013]. The success of this type of analysis will depend critically on our ability to identify SNP characteristics that are truly informative. Cantor et al. considered three forms of prioritization based on meta-analysis, interaction, and pathways, and each of these sources could also be adapted to inform our priors [Cantor et al., 2010]. We show that when prior information is used wisely, the probability of association will outperform the *P*-value in SNP prioritization. Indeed, the gain from using the probability of association proved to be quite robust to the exact specification of the priors.

The existing literature on prioritization is dominated by algorithms that seek to use the external information separately from the experimental data in a two-stage process, that is the experimental data are used to create a short list of candidates and the data from the bioinformatics databases are then used to prioritize within that list [Moreau and Tranchevent, 2012]. However, some methods have previously been proposed that have sought to combine the two stages. Some, such as Saccone et al. [2008, 2010], used the external information to define ad hoc weights that are then applied to the *P*-values, but others have used the Bayesian approach in order to combine the two types of evidence in a more theoretically grounded way. Fridley et al. [2011] used a latent variable model in which the same SNP-specific unmeasured variable acts as a predictor for the experimental results such as the effect size and the *P*-value and for the SNP characteristics. Several of the Bayesian methods have used a hierarchical structure in which all SNPs are analyzed together as if they were each drawn from the higher level distribution. The full set of SNPs can then be used to learn about the shape of this distribution. In the context of our method, this would be the equivalent of using the full set of SNPs to inform the choice of the average effect size in SNPs that show an association, or to inform the proportion of associated SNPs that show a positive effect as opposed to a negative one, or even to inform the weights used to combine the data from the bioinformatics databases. Although this refinement is perfectly possible within the context of a full Bayesian analysis, it is computationally very demanding. Lewinger et al. used a hierarchical model for the noncentrality parameter of the chi-squared statistic used to test for association in the experimental data [Lewinger et al., 2007], although combining external information via a logistic function in a similar way to Minelli et al. [2013]. Chen and Witte also used a hierarchical model but instead placed the higher level distribution over the effect size in such a way that the higher level distribution could depend on external information [Chen and Witte, 2007]. There are undoubted advantages in these more comprehensive approaches, but perhaps their complexity inhibits their wider adoption and there may be some advantages in the simpler calculations that we propose. It would be possible to adapt our approach in a similar way; if all SNPs were analyzed in the same Bayesian analysis, the results could then be used to update the parameters of the distribution of effects size and even to learn whether SNPs are equally likely to have a positive or negative effect.

Wakefield [2007] proposed a Bayesian version of the FPRP which he called the Bayesian false-discovery probability (BFDP). The BFDP is defined for binary outcomes in which the prior on the effect size, that is the log odds ratio, is taken to be a zero-centered normal distribution. Under this model, an approximate Bayes factor is derived for comparing the models in which the SNP is either associated or unassociated with the disease, and then from that Bayes factor and the prior probability of association one can derive an expression for

the probability of a false discovery. This approach is similar to our own, differing primarily in our use of a spike-and-slab prior and our use of training data to inform the prior probability of association.

Bayesian methods have become increasingly popular in genetics [Stephens and Balding, 2009]. Often, however, these Bayesian analyses either do not discuss the specification of the priors, or they use noninformative priors so as to allow the results to depend mostly on the data. In selecting SNPs for replication, we have used informative priors and shown that this can improve selection. The specification of such informative priors is an important part of the analysis and must be undertaken with great care. It is also important that a sensitivity analyses is performed to investigate whether the SNP selection would be materially altered by small changes in the priors. Researchers who are cautious of using informative priors can take comfort from the fact that the ultimate judgment about association will be based on the data from the follow-up or replication study, and they might consider that formally stating their prior beliefs is preferable to the informal way that external information is sometimes used when selecting SNPs for replication [Gögele et al., 2012].

Acknowledgments

All researchers from the Center for Biomedicine at EURAC were supported by the Department for Promotion of Educational Policies, Universities and Research of the Autonomous Province of Bolzano, South Tyrol, Italy; Michael Boehnke was supported by NIH grant HG000376. None of the authors declares any conflict of interest.

References

Cantor RM, Lange K, Sinsheimer JS. 2010. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *Am J Hum Genet* 86:6–22.

Chen GK, Witte JS. 2007. Enriching the analysis of genome-wide association studies with hierarchical modeling. *Am J Hum Genet* 81:397–404.

Fridley BL, Iversen E, Tsai YY, Jenkins GD, Goode EL, Sellers TA. 2011. A latent model for prioritization of SNPs for functional studies. *PLoS One* 6(6):e20764.

Gögele M, Minelli C, Thakkinstian A, Yurkiewich A, Pattaro C, Pramstaller P, Little J, Attia J, Thompson JR. 2012. Methods for meta-analysis of genome-wide association studies: critical assessment of empirical evidence. *Am J Epidemiol* 175: 739–749.

Goodman SN. 2007. A comment on replication, p-values and evidence. *Stat Med* 11:875–879.

Köttgen A, Glazer NL, Dehghan A, Hwang SJ, Katz R, Li M, Yang Q, Gudnason V, Launer LJ, Harris TB and others. 2009. Multiple loci associated with indices of renal function and chronic kidney disease. *Nat Genet* 41:712–717.

Köttgen A, Pattaro C, Böger CA, Fuchsberger C, Olden M, Glazer NL, Gao X, Yang Q, Smith AV, O'Connell JR and others. 2010. New loci associated with kidney function and chronic kidney disease. *Nat Genet* 42:376–384.

Lewinger JP, Conti DV, Baurley JW, Triche TJ, Thomas DC. 2007. Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. *Genet Epidemiol* 31:871–882.

Lucke JF. 2009. A critique of the false-positive report probability. *Genet Epidemiol* 33:145–150.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Charavarti A and others. 2009. Finding the missing heritability of complex diseases. *Nature* 461:747–753.

Minelli C, De Grandi A, Weichenberger CX, Gögele M, Modenese M, Attia J, Barrett J, Boehnke M, Borsani G, Casari G and others. 2013. Importance of different types of prior knowledge in selecting genome-wide findings for follow-up. *Genet Epidemiol* 37:205–213.

Mitchell TJ, Beauchamp JJ. 1988. Bayesian variable selection in linear regression. *J Am Stat Assoc* 83:1023–1032.

Moreau Y, Tranchevent L-C. 2012. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat Rev Genet* 13:523–536.

Saccone SF, Saccone NL, Swan GE, Madden PAF, Goate AM, Rice JP, Bierut LJ. 2008. Systemic biological prioritization after a genome-wide association study: an application to nicotine dependence. *Bioinformatics* 24:1805–1811.

Saccone SF, Bolze R, Thomas P, Quan J, Mehta G, Deelman E, Tischfield JA, Rice JP. 2010. SPOT: a web-based tool for using biological databases to prioritize SNPs after a genome-wide association study. *Nucleic Acids Res* 38:W201–W209.

Samani NJ, Erdmann J, Hall A, Hengstenberg C, Mangino M, Mayer B, Dixon RJ, Meitinger T, Braund P, Wichmann HE and others. 2007. Genome-wide association analysis of coronary artery disease. *N Engl J Med* 357:443–453.

Stephens M, Balding DJ. 2009. Bayesian statistical methods for genetic association studies. *Nat Rev Genet* 10:681–690.

Thomas DC, Clayton DG. 2004. Betting odds and genetics associations. *J Natl Cancer Inst* 96:421–423.

Thompson JR. 1998. Re "Multiple comparisons and related issues in the interpretation of epidemiologic data". *Am J Epidemiol* 147:801–806.

Wacholder S, Chanock S, Garcia-Closas M, El Gourmli L, Rothman N. 2004. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst* 96:434–442.

Wakefield J. 2007. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am J Hum Genet* 81:208–227.

Zollner S, Pritchard JK. 2007. Overcoming the winner's curse: estimating penetrance parameters from case-control data. *Am J Hum Genet* 80:605–615.