European Journal of Human Genetics (2010) 18, 942–947

www.nature.com/ejhg

ARTICLE

# A simple Bayesian mixture model with a hybrid procedure for genome-wide association studies

Yu-Chung Wei[1,2], Shu-Hui Wen[3], Pei-Chun Chen[1,4], Chih-Hao Wang[5] and Chuhsing K Hsiao*[,1,4]

**Genome-wide association studies often face the undesirable result of either failing to detect any influential markers at all because of a stringent level for testing error corrections or encountering difficulty in quantifying the importance of markers by their *P*-values. Advocates of estimation procedures prefer to estimate the proportion of association rather than test significance to avoid overinterpretation. Here, we adopt a Bayesian hierarchical mixture model to estimate directly the proportion of influential markers, and then proceed to a selection procedure based on the Bayes factor (BF). This mixture model is able to accommodate different sources of dependence in the data through only a few parameters. Specifically, we focus on a standardized risk measure of unit variance so that fewer parameters are involved in inference. The expected value of this measure follows a mixture distribution with a mixing probability of association, and it is robust to minor allele frequencies. Furthermore, to select promising markers, we use the magnitude of the BF to represent the strength of evidence in support of the association between markers and disease. We demonstrate this procedure both with simulations and with SNP data from studies on rheumatoid arthritis, coronary artery disease, and Crohn's disease obtained from the Wellcome Trust Case–Control Consortium. This Bayesian procedure outperforms other existing methods in terms of accuracy, power, and computational efficiency. The R code that implements this method is available at http://homepage.ntu.edu.tw/~ckhsiao/Bmix/Bmix.htm.**
*European Journal of Human Genetics* (2010) **18,** 942–947; doi:10.1038/ejhg.2010.51; published online 21 April 2010

**Keywords:** Bayesian inference; GWAS; mixture model; WTCCC

## INTRODUCTION

Statistical analysis for testing simultaneously a large number of hypotheses in genome-wide association studies (GWAS) has received considerable attention. Several algorithms have focused on two- or multi-stage procedures, or on the selection of ordered *P*-values under a controlled error rate.[1–3] Other procedures have suggested a mixture structure on *P*-values, or have adopted the null normal distribution with a dispersed density such as a mixture of beta distributions or a mixture of normals.[4,5] The underlying principle of all these methods is to separate the markers into two groups on the basis of *P*-values, one showing association with the disease trait and the other not. It is thus intuitive to incorporate the mixture approach in genomic analysis, in which only a small proportion, say $\lambda$, among the enormous number of markers shows association. Our rationale for using a mixture model, however, goes beyond the purpose of clustering, as it considers the estimation procedure primarily as a tool of inference and it accommodates the possible dependence among markers directly.

An estimation approach, rather than hypothesis testing, has earlier been advocated for large-scale association studies primarily so as to avoid overinterpretation of statistical significance.[6] By adopting the mixture model, we get an important advantage of incorporation of dependence among observed data, when several sources of dependence arise. For example, linkage disequilibrium (LD) among dense SNPs is one source of correlation. Genotyping data collected from a given individual may produce within-subject dependence. Markers associating with the disease phenotype share the common feature of 'showing association,' and thus may induce yet another type of correlation within a gene set or pathway. Whereas the existence of such dependence in data, and thus among the statistics, may complicate the task of multiple hypotheses testing, a mixture structure on a properly defined statistic may in contrast provide an explanation of dependence in general. For these reasons, several GWAS have adopted the mixture model conceptualization.[4,7]

Bayesian statistical inference has attracted many researchers for its ability to deal with complex models and with uncertainty from different sources.[8] For example, an exploration approach and an empirical Bayes approach with hierarchical modeling have been considered.[9–11] Others have proposed a fully Bayesian approach with a Dirichlet process mixture model for differential gene expression, or a hierarchical mixture model for normalized microarray data.[12] These models are useful when the information about an evolutionary region or linkage is available, though they may be computationally intensive because of a large number of parameters. Other proposals have included use of an asymptotic Bayes factor (BF) as a means of incorporating the information of minor allele frequency (MAF) and sample size,[13] and using a hierarchical mixture model on the logarithm of odds ratios as was briefly mentioned earlier.[14] Nevertheless, a specific mixture distribution accommodating different sample sizes and variances among various markers has not been previously suggested.[14] The large number of parameters remains a challenging

[1]Department of Public Health, Institute of Epidemiology and Research Center for Gene, Environment, and Human Health, National Taiwan University, Taipei, Taiwan, ROC; [2]Institute of Statistics, National Chiao Tung University, Hsinchu, Taiwan, ROC; [3]Department of Public Health, College of Medicine, Tzu-Chi University, Hualien, Taiwan, ROC; [4]Bioinformatics and Biostatistics Core, Research Center for Medical Excellence, National Taiwan University, Taipei, Taiwan, ROC; [5]Department of Cardiology, Cardinal Tien Hospital, College of Medicine, Fu-Jen Catholic University, Taiwan, ROC
*Correspondence: Dr CK Hsiao, Department of Public Health, Institute of Epidemiology and Research Center for Gene, Environment, and Human Health, National Taiwan University, No. 17, Xu-Zhou Road, Room 523, Taipei 100, Taiwan, ROC. Tel: +886 2 3366 8032; Fax: +886 2 2341 8562; E-mail: ckhsiao@ntu.edu.tw
Received 13 November 2009; revised 9 March 2010; accepted 10 March 2010; published online 21 April 2010

task. Other proposals have included the false positive report probability, a combined measure of frequentist significance and Bayesian subjective probability.[15] Criticism arises, however, with its Bayesian interpretation.[16] Therefore, for GWAS, a fully Bayesian approach with a simpler model may provide a better tool for inference and for wider applicability.

We propose here a hybrid procedure based on a Bayesian hierarchical mixture model for biallelic markers. This Bayesian mixture model considers not only the grouping of markers, but also the general dependence among data. This model is able to capture the existing dependence with only a few parameters, and thus is simpler than other Bayesian models. There are two important features of our proposed model of particular note. First, we consider a statistic $y_i$, standardized to be of unit variance, and use a mixture structure on its parameter $\gamma_i$. In other words, this $y_i$ represents the 'standardized' difference in allele frequencies between case and control groups, and is drawn from a distribution with a mean $\sqrt{n_i}\times\gamma_i$ and variance 1, where $n_i$ is the sample size and the index $i$ ranges from 1 to $M$ markers. The use of such $y_i$ has a major function in our procedure and its contribution will be explained in later sections. The mixture prior on $\gamma_i$ is an important element to account for the general dependence among statistics. In contrast to other mixture structures on *P*-values or on their transformations, such as test statistics, the adopted model is more intuitive and simpler to implement. Second, the information contained in the mean parameter $\gamma_i$ is not affected by the sample size. Without worrying about the influence from the sample size, a *unit information* prior for the $i$-th marker becomes easier to formulate. In addition, most procedures consider the small proportion of association markers, $\lambda$, a nuisance parameter and often fail to provide a stable estimate for it, not to mention failing to use its estimate at the stage of selecting significant components. We will show here that the posterior inference of $\lambda$ improves greatly the performance of the association test. After estimation of $\lambda$, a test with the BF at the marker-specific level is carried out to rank and identify susceptible genes.[17] Several GWAS from the Wellcome Trust Case–Control Consortium (WTCCC) as well as simulations are used to evaluate performance of the model.

## MATERIALS AND METHODS

### Bayesian mixture structure

Suppose there are $M$ markers under study, and let $\lambda$ denote the small proportion of association markers among them. The mixture setting of $M\times\lambda$ and $M\times(1-\lambda)$ markers within a Bayesian hierarchical model then explains the dependence among markers through their corresponding 'standardized' statistics $y_i$'s as follows. For the $i$-th marker, $y_i$ is the standardized difference in genetic measurements, say gene expression levels, between two types of tissues or experimental conditions:

$$y_i = \frac{\hat{g}_{\mathrm{cs},i} - \hat{g}_{\mathrm{cn},i}}{\sqrt{\mathrm{Var}(\hat{g}_{\mathrm{cs},i})+\mathrm{Var}(\hat{g}_{\mathrm{cn},i})}} = \frac{\hat{g}_{\mathrm{cs},i} - \hat{g}_{\mathrm{cn},i}}{\sqrt{\frac{\hat{\sigma}^2_{\mathrm{cs}}}{n_{\mathrm{cs}}}+\frac{\hat{\sigma}^2_{\mathrm{cn}}}{n_{\mathrm{cn}}}}}$$

where $\hat{g}_{\mathrm{cs},i}$ and $\hat{g}_{\mathrm{cn},i}$ are the observed average levels for the case group and for the control group, respectively; $n_{\mathrm{cs}}$ and $n_{\mathrm{cn}}$ are the numbers of tissues in each group; and $\hat{\sigma}^2_{\mathrm{cs}}$ and $\hat{\sigma}^2_{\mathrm{cn}}$ are estimates of variance in each group.

Each $y_i$ is of unit variance and represents a standardized effect size. For biallelic markers such as SNPs, $y_i$ is the 'standardized' difference in allele frequencies between two groups,

$$y_i = \frac{\hat{p}_{\mathrm{cs},i} - \hat{p}_{\mathrm{cn},i}}{\sqrt{\frac{\hat{p}_{\mathrm{cs},i}(1-\hat{p}_{\mathrm{cs},i})}{n_{\mathrm{cs}}}+\frac{\hat{p}_{\mathrm{cn},i}(1-\hat{p}_{\mathrm{cn},i})}{n_{\mathrm{cn}}}}}$$

where $\hat{p}_{\mathrm{cs},i}$ and $\hat{p}_{\mathrm{cn},i}$ are the observed minor allele frequencies (MAFs) of the $i$-th marker for the case and control groups, respectively. Note that for biallelic markers, this standardization uses the empirical variance estimates without loss of degrees of freedom in estimation. Therefore, this procedure is able to avoid the formulation of a prior distribution on nuisance variance parameters. This model is more flexible than one that assumes equal variance for all markers,[9] and is easier to work with than a highly parameterized model with unequal variances for all markers.

To satisfy the purpose of association studies, we specify the mixture structure on the mean standardized risk difference $\gamma_i$. For ease of notation, let $n_i=n_{\mathrm{cs},i}=n_{\mathrm{cn},i}$; then the mean of $y_i$ can be factored as $\gamma_i\times\sqrt{n_i}$, where $\gamma_i$ is a function of the MAFs. This $\gamma_i$ follows a mixture distribution with a mixing parameter $\lambda$,

$$\gamma_i|\lambda\sim (1-\lambda)\times f_0(\gamma_i)+\lambda\times f_1(\gamma_i|t,u)$$

That is, with probability $1-\lambda$, $\gamma_i$ follows a distribution $f_0$ of no association, whereas with probability $\lambda$, $\gamma_i$ follows $f_1$, a uniform distribution over $(-u,-t)$ and $(t,u)$ of association. The mixing weight $(1-\lambda)$ denotes the majority of non-associated markers, and $f_0(\gamma_i)$ can be a conservative indicator function, $I(0)$, degenerating at the zero mass, or a distribution over a pre-specified interval for 'no association.' Under $f_1$, $\gamma_i$ ranges in the set $(-u,-t)\cup(t,u)$ (see Supplementary Materials for discussion about their values and the complete model specification).

### Effect size and rare MAF

The standardized risk difference $\gamma_i$ here can be considered as a measure of effect size. As compared with the odds ratios, it is much less sensitive to rare MAFs. The influence from rare MAFs is reduced to a minimum because $y_i$ has been standardized by its SD. In Figure 1a and b, the solid lines show the behavior of $\gamma_i$ and the odds ratio, respectively, versus the rare MAFs of the control group. It is apparent that $\gamma_i$ remains stable in the range of rare MAFs, while the odds ratio, even divided by 10 (OR/10), contains much more variation. The same pattern can be observed when the difference in allele frequencies between two groups is fixed (dashed lines in Figure 1). Therefore, a mixture structure on $\gamma_i$ is more suitable in terms of variability than an odds ratio.

### Formulation of the priors

For the two components $f_0$ and $f_1$ in the mixture distribution of $\gamma_i$, we consider an indicator function for $f_0(\gamma_i)$ and a uniform distribution for $f_1$. The latter implies that all values are equally likely over the intervals $(t,u)$ and $(-u,-t)$. When the case and control groups are of the same size (ie, $n_{\mathrm{cs}}=n_{\mathrm{cn}}$), the range can be derived analytically as ($t$=0.071, $u$=0.825) for biallelic markers with MAFs in (0.05, 0.50). When sample size differs, the range depends on the ratio of $n_{\mathrm{cs}}$ and $n_{\mathrm{cn}}$ (see Supplementary Material). The above finite values are recommended from a conservative viewpoint, because a larger $u$ would imply stronger evidence against the null *a priori*, and may lead to an improper posterior for inference.

### Hybrid method of the global estimate and the BF

With the likelihood, mixture structure, and prior distribution established, the posterior distribution of $\lambda$ can then be derived (Supplementary Material). Next, based on this posterior distribution $f(\lambda|y_1,\ldots,y_M)$, one can estimate $\lambda$, the proportion of influential markers, and proceed to the statistical inference. We term the estimated posterior mode, $\hat{\lambda}$, the *global estimate* of the association proportion. In other words, $M\times\hat{\lambda}$ is the estimated posterior mode of $M\times\lambda$, the number of association markers. This estimate borrows strength from all marker information, and, as the results will show later, it is accurate and stable. If the estimated proportion $\hat{\lambda}$ is non-zero, then a total of $M\times\hat{\lambda}$ markers are considered candidates and will be selected by the rankings of the BFs[17] for the $M$ markers. The smaller the BF, the stronger the evidence supporting the association with the disease under study.

To compute the BF for the $i$-th marker, we test $\lambda_i=0$ indicating no association versus $\lambda_i=1$ indicating an association. Next, we select the leading $M\times\hat{\lambda}$ markers with the largest values of 1/BF, the inverse of BF. The magnitude indeed implies directly the strength of evidence for association.[17] In general, when BF is less than 1/100, the strength of evidence against the null is considered decisive;[17] that is, the data provide strong evidence supporting the association. Using the threshold 1/100 for BF usually results in more signals than using the global estimate, and is considered more conservative.

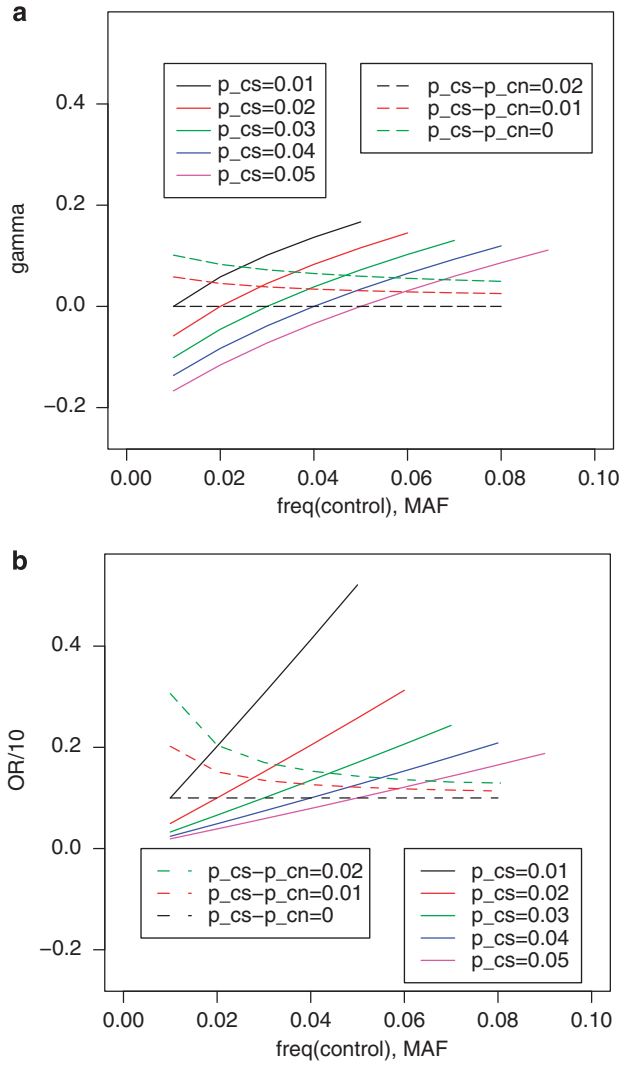


**Figure 1** The solid lines are values of (**a**) $\gamma_i$ (OR/10 in (**b**)) versus $p_{cn}$ (small MAFs of the control group) at different fixed values of $p_{cs}$. The dashed lines are values of (**a**)$\gamma_i$ (OR/10 in (**b**)) versus $p_{cn}$ at fixed values of $p_{cs}-p_{cn}$ ($p_{cs}-p_{cn}$=0.02, 0.01, and 0 from top to bottom, respectively).

## RESULTS

### WTCCC association studies

We consider genotyping data from the WTCCC[18] to evaluate the performance of the hybrid Bayesian procedure and to compare it with other existing methods. From the WTCCC archive, we obtained genotyping data originally from 1999 rheumatoid arthritis (RA) patients, 1988 patients with coronary artery disease (CAD), 2004 with Crohn's disease (CD), and 3004 common controls. Exclusion criteria for SNPs were (1) MAF <0.05, (2) call rate <95%, and (3) failure to meet the quality control criteria of the WTCCC.[18]

### Rheumatoid arthritis

After passing the quality control filters of the WTCCC, a total of 366 037 SNPs were selected for 1860 RA patients and 2938 controls for our analysis. Earlier studies have established an association region, the major histocompatibility complex, also called the HLA complex, on 6p21, and a gene *PTPN22* on 1p13.2.[17–21] Here, we examine whether these findings can be replicated by the proposed method.

**Table 1 Numbers of influential markers for three studies (RA, CAD, CD) from WTCCC under different methods**

| | *Bayes* | P-*value threshold*[a] | *Bon* | q-*value* | *BH* | *wBH* |
|---|---|---|---|---|---|---|
| *Rheumatoid arthritis (1860 cases; 366 037 SNPs)* | | | | | | |
| $M\times\hat{\lambda}$ | 232 | 234 | 200 | 372 | 372 | 372 |
| *Coronary artery disease (1926 cases; 365 984 SNPs)* | | | | | | |
| $M\times\hat{\lambda}$ | 26 | 31 | 29 | 42 | 42 | 42 |
| *Crohn's disease (1748 cases; 366 251 SNPs)* | | | | | | |
| $M\times\hat{\lambda}$ | 74 | 86 | 72 | 178 | 167 | 167 |

[a]Number of significant SNPs identified by the *P*-value threshold of $5\times10^{-7}$ used by WTCCC.

Table 1 lists the estimated number of association markers based on the Bayesian hybrid procedure (Bayes), the Bonferroni procedure (Bon), *q*-values, the Benjamini and Hochberg procedure (BH), and a false discovery control with a *P*-value weighting scheme (wBH).[22] Among the 232 SNPs identified by the Bayesian procedure, 9 are near *PTPN22* on 1p13.2, and 200 are in 6p21.32–6p21.33. Both regions generated strong signals. In fact, 217 of the SNPs are on either chromosome 1 or 6. For comparison, the number of SNPs identified by all other procedures (except Bon) is >1.5 times the number (232) obtained with our procedure, and may result in more false positive regions, while the Bon provides the smallest number 200. Figure 2a displays the distributions of these 232 SNPs, and Figure 2b shows the Manhattan plot of the negative log10 transformation of BFs. Supplementary Table A.1 lists their chromosome locations. For those SNPs whose individual BFs imply decisive evidence, details are in Supplementary Table A.2. The Bayesian inference further provides a measure of association strength for each marker (ie, BF< 1/100). The identified SNPs merit further functional or pathway analysis, and can serve as the next candidate regions for further analysis (Supplementary Table A.2).

### Coronary artery disease

On the basis of the quality control criteria, a total of 365 984 SNPs of 1926 subjects with CAD were derived. The Bayes identified 26 SNPs, the Bon selected 29 SNPs, and the others 42. Among the 26 SNPs, 16 are in 9p21.3. This region contains two cyclin-dependent kinase inhibitors, *CDKN2A* and *CDKN2B*, both of which have been reported to associate with CAD.[23] The locations of identified SNPs are displayed in Figure 3 and the details are in Supplementary Table A.1. It is worth noting that, although the earlier identified *APOE* gene on 19q13 did not show a signal in the WTCCC reports and was not selected by $\hat{\lambda}$ either, the Bayesian test did provide decisive evidence of association (ie, BF<1/100, Supplementary Table A.3). This again shows the advantage of the hybrid procedure in providing candidate markers based on the strength of evidence.

### Crohn's disease

A total of 366 251 SNPs and 1748 subjects with CD passed the quality control filters, and were included in the analysis. Seventy-four SNPs were identified by the Bayesian global estimate, 72 were determined by the Bon, 178 by *q*-values, and 167 by both BH and wBH. The locations of identified SNPs are presented in Supplementary Figure A.1 with details in Supplementary Table A.1.

Among the identified 74 SNPs, 22 were in 5p13.1, 15 in 1p31.3, 12 in 16q12.1, 7 in 10q24.2, and 5 in 2q37.1, respectively. These regions

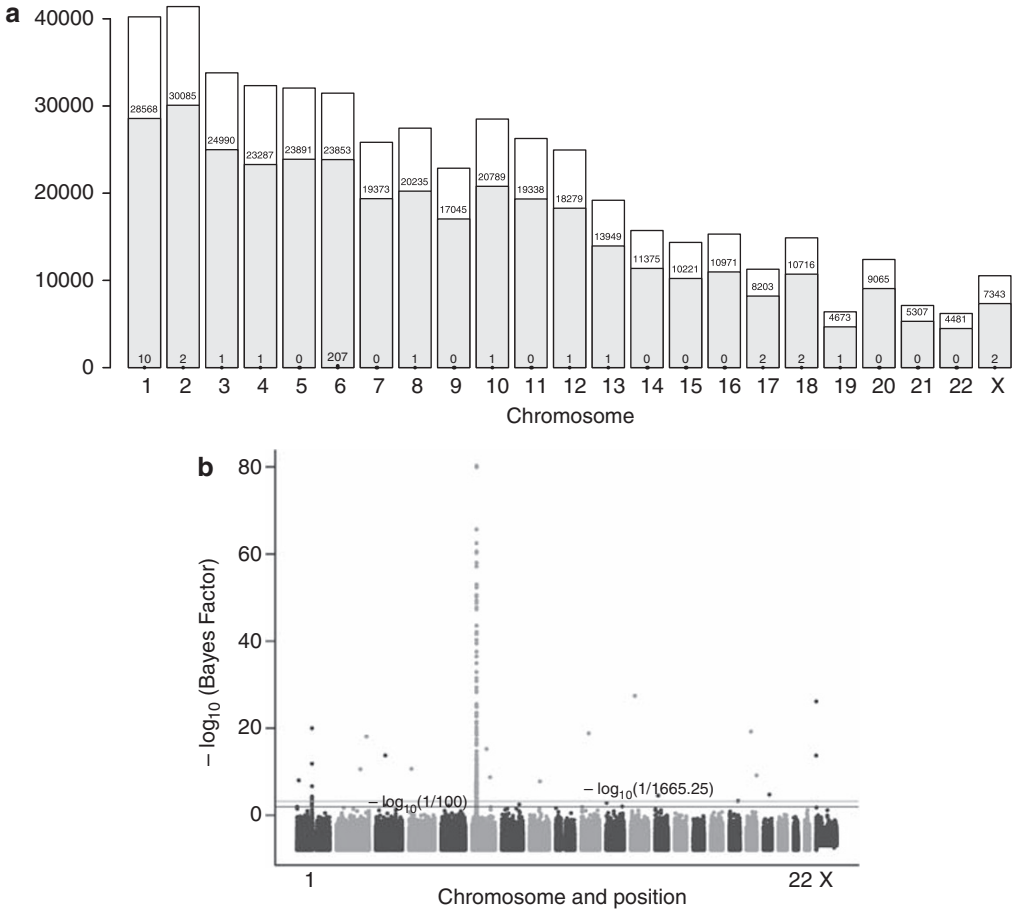


**Figure 2** (**a**) Grey bars represent the numbers of SNPs passing the quality control filters for the corresponding chromosomes. The values at the bottom of the bars indicate the numbers of SNPs selected by $\hat{\lambda}$ for RA. (**b**) Plot of −log10(BF) by chromosome locations.

have all been identified in earlier research as containing genes associated with CD: the gene *PTGER4* on 5p13, *IL23R* on 1p31, *NOD2* on 16q12, *NKX2-3* on 10q24.2, and *ATG16L1* on 2q37.1. Three other SNPs on 5q23.1, 10q21.2, and 18p11.21 were in different regions. For the remaining 10 SNPs, the values of BFs also imply remarkably strong association (Supplementary Table A.4). Further replication studies or pathway analysis would be necessary for confirmation.

### Simulation study

Here, we show the performance of this proposed procedure with simulated data, especially with correlated markers. We consider the independent and correlated settings separately with $\lambda$ between 0.001 and 0.1. The dependence setting uses a spatial structure for correlation for LD, where the correlation between any two markers depends on the physical distance,

$$\rho_{ij} \equiv \mathrm{corr}(\mathrm{marker}_i, \mathrm{marker}_j) = \begin{cases} 1 - \frac{d_{ij}}{\theta}, & \text{if } d_{ij} \le \theta \\ 0, & \text{if } d_{ij} > \theta \end{cases}$$

where $d_{ij}$ is the distance between the $i$-th and $j$-th marker, and $\theta$ is a pre-specified tolerance distance for correlation. The $d_{ij}$ follows a Poisson distribution and $\theta$ depends on the maximum correlation in the generated data. The procedures for generation of $d_{ij}$ and determination of $\theta$ are documented in Supplementary Material. Note that when $M$ is larger than 1000, the correlation between markers at two ends becomes extremely small and thus markers become almost independent. Therefore, it suffices to set $M$ at 1000 in all 100 replications. The global estimation of $\lambda$ is derived, and other existing methods are also performed for comparison.

The results are displayed in Figure 4 and Supplementary Table A.5. For the point estimate of $\lambda$ under independent markers, all procedures provide good estimates, except the conservative Bon. However, when data are simulated with the dependence structure, the Bayesian global estimate performs the best (Figure 4b). Its performance is consistently stable and is not affected by the data structure. The corresponding standard error is small as well ($<0.1\%$ under the independence model and $<1\%$ under the dependence model). The $q$-value (Storey_q) overestimates to a greater degree, confirming the findings in the WTCCC analysis that it tends to overestimate, whereas BH, wBH, and Bon all underestimate (details in Supplementary Materials). We have carried out other simulation studies with settings such as $M$=100 000 and $\lambda$=0.0001, and the results are as good as the case reported here.

## DISCUSSION

On the basis of a simple Bayesian hierarchical mixture model, the hybrid procedure we have used in this article to detect susceptible genes for GWAS has proven useful and robust for extremely small $\lambda$. The major contributions of our approach are the use of the standardized statistic $y_i$, and a further decomposition of its mean to the

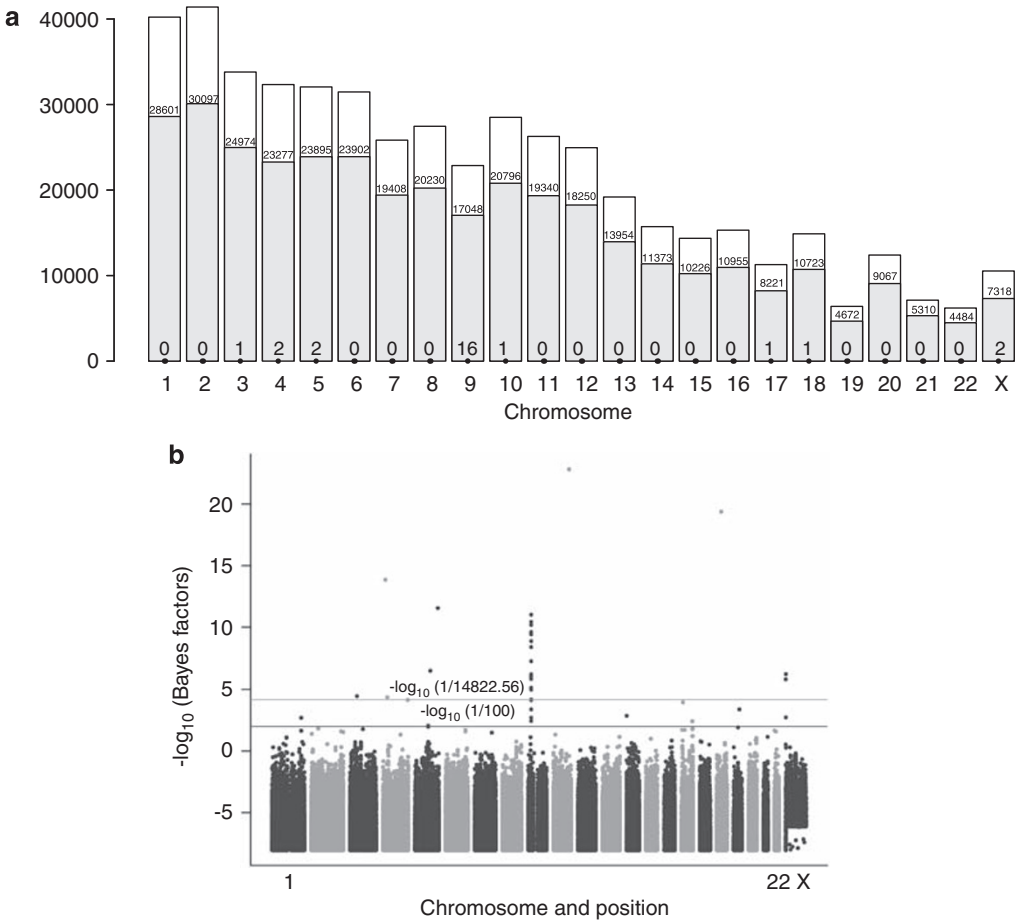


**Figure 3** (**a**) Grey bars represent the numbers of SNPs passing the quality control filters for the corresponding chromosomes. The values at the bottom of the bars indicate the numbers of SNPs selected by $\hat{\lambda}$ for CAD. (**b**) Plot of −log10(BF) by chromosome locations.

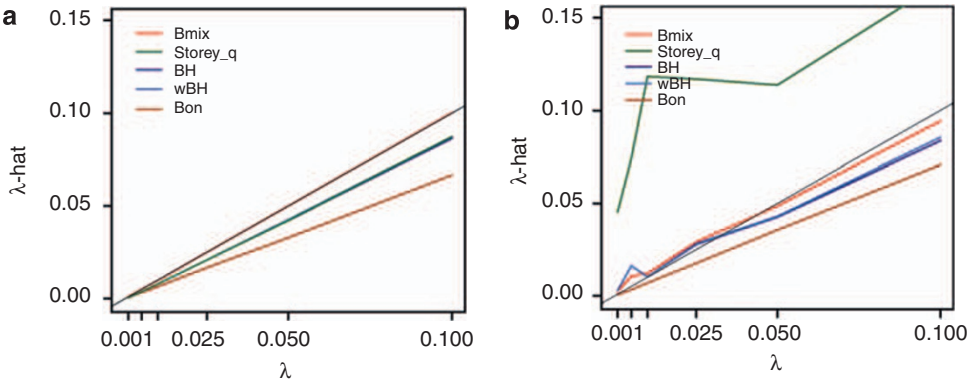


**Figure 4** Estimated proportion $\hat{\lambda}$ versus true $\lambda$ for (**a**) independent and (**b**) dependent markers. Bmix is the global estimate, Storey_q for *q*-values, BH for Benjamini and Hochberg procedure, wBH is false discovery control with *P*-value weighting, and Bon for Bonferroni procedure.

product of $\gamma_i$ and $\sqrt{n_i}$. This novel design and factorization are essential within this setting as they make inference of the proportion of association feasible with a fewer number of parameters – this is a case different from and not explicitly considered by others.[14] The standardized risk difference $\gamma_i$, in addition to the direct interpretation it affords and the easier formulation of priors, is robust to small MAFs and small differences in MAFs, as opposed to when OR or log(OR) is the parameter of interest. The global Bayesian estimate performs well in terms of point accuracy, true positive rate, false positive rate, and proportions of false positive, for both independent and dependent markers (Supplementary Figures A.2–A.4). Furthermore, the second step of the BF evaluation quantifies the strength of evidence for association, and offers a way to measure the 'importance' of markers, thereby making possible allocation of appropriate research resources to the different markers. This quantitative comparison is an advantage that the procedures based on *P*-values cannot provide.

There are several issues meriting discussion with regard to extension of the model. First, the statistic $y_i$ illustrated here is based on the

standardized difference in allele frequencies. This can be extended to data of genotype counts or microarray data. For instance, for any risk allele under the dominant or recessive model, the difference in frequencies of genotypes under study can be computed and standardized to derive the global estimate. However, one should be cautious in estimating variance. For biallelic markers as discussed here, estimation of the variance does not lose extra degrees of freedom, whereas for other data types, this may not hold true. Second, with microarray data, the range of $\gamma_i$ in the reference prior will differ from the values considered for $(t,u)$ in simulations. In our implementations, we suggest the 10th percentile of $y_i$ for $t$. In fact, our investigations show that any other smaller value will not make any difference in the results. As for $u$, it has to be greater than all $y_i$'s, and hence the maximum of $y_i$ is recommended. Other larger numbers are certainly allowable. Third, for gene expression profiling data, this procedure can be compared with current procedures such as the maxT test. Fourth, for the prior specification on $\gamma_i$, other choices with informative priors are possible. The analytical derivation of the global estimate should be straightforward, or numerical implementations can be adopted. Finally, as already mentioned, this Bayesian hierarchical model is used here to detect association markers based on the hybrid procedure, particularly when the proportion of influential markers is very small, or when quantification of the degree of association is of interest. Further applications include meta-analysis of various GWAS to measure small effects and to assess the degree of heterogeneity among studies. The need to synthesize results across multiple studies indeed opens the era for methodological research and collaborations among different disciplines.

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## ACKNOWLEDGEMENTS

This study makes use of data generated by the Wellcome Trust Case–Control Consortium. A full list of the investigators who contributed to the generation of the data is available from http://www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113. We are grateful to WTCCC for granting the access, and to Dr Lee H Chen for help with plots. We also thank reviewers and the associate editor for their suggestions, which improved this manuscript greatly (NSC 97-2314-B-002-040-MY3).

1 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.
2 Storey JD: A direct approach to false discovery rates. *J R Stat Soc B* 2002; **64**: 479–498.
3 Wen SH, Tzeng JY, Kao JT, Hsiao CK: A two-stage design for multiple testing in large-scale association studies. *J Hum Genet* 2006; **51**: 523–532.
4 Pounds S, Morris S: Estimating the occurrence of false positions and false negatives in microarray studies by approximating and partitioning the empirical distribution of the *P*-values. *Bioinformatics* 2003; **19**: 1236–1242.
5 Pan W, Lin J, Le CT: A mixture model approach to detecting differentially expressed genes with microarray data. *Funct Integr Genomics* 2003; **3**: 117–124.
6 Strömberg U, Björk J, Broberg K, Mertens F, Vineis P: Selection of influential genetic markers among a large number of candidates based on effect estimation rather than hypothesis testing: an approach for genome-wide association studies. *Epidemiology* 2008; **19**: 302–308.
7 McLachlan GJ, Bean RW, Peel D: A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics* 2002; **18**: 413–422.
8 Beaumont MA, Ranala B: The Bayesian revolution in genetics. *Nat Rev Genet* 2004; **5**: 251–261.
9 Scott JG, Berger JO: An exploration of aspects of Bayesian multiple testing. *J Stat Plan Inf* 2006; **136**: 2144–2162.
10 Hung RJ, Brennan P, Malaveille C *et al*: Using hierarchical modeling in genetic association studies with multiple markers: application to case-control study of bladder cancer. *Cancer Epidemiol Biomarkers Prev* 2007; **81**: 397–404.
11 Strömberg U: Empirical Bayes and semi-Bayes adjustments for a vast number of estimations. *Euro J Epidemiol* 2009; **24**: 737–741.
12 Lewinger JP, Conti DV, Baurley JW, Triche TJ, Thomas DC: Hierarchical Bayes prioritization of marker association from a genome-wide association scan for further investigation. *Genet Epidemiol* 2007; **31**: 871–882.
13 Wakefield J: Bayes factors for genome-wide association studies: comparison with *P*-values. *Genet Epidemiol* 2009; **33**: 79–86.
14 Wakefield J: A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am J Hum Genet* 2007; **81**: 208–227.
15 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.
16 Lucke JF: A critique of the false-positive report probability. *Genet Epidemiol* 2009; **33**: 145–150.
17 Kass RE, Raftery AE: Bayes factors. *J Am Stat Assoc* 1995; **90**: 773–795.
18 The Wellcome Trust Case Control Consortium: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007; **447**: 661–678.
19 Horton R, Gibson R, Coggill P *et al*: Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project. *Immunogenetics* 2008; **60**: 1–18.
20 Begovich AB, Carlton VEH, Honigberg LA *et al*: A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (*PTPN22*) is associated with rheumatoid arthritis. *Am J Hum Genet* 2004; **75**: 330–337.
21 Hinks A, Eyre S, Barton A, Thomson W, Worthington J: Investigation of genetic variation across PTPN22 in UK rheumatoid arthritis (RA) patients. *Ann Rheum Dis* 2006; **66**: 683–686.
22 Genovese C, Roeder K, Wasserman L: False discovery control with *P*-value weighting. *Biometrika* 2006; **93**: 509–524.
23 Kronenberg F, Coon H, Avkerich V *et al*: A genome scan for loci influencing anti-atherogenic serum bilirubin levels. *Eur J Hum Genet* 2002; **10**: 539–546.

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