Supplementary Material to:

Exploration of Empirical-Bayes Hierarchical Modelling for the Analysis of Genome-Wide Association Study Data

Elizabeth A. Heron[†1] Colm O'Dushlaine[†] Ricardo Segurado[†] Louise Gallagher[†] Michael Gill[‡]

† Neuropsychiatric Genetics Research Group,
Dept. of Psychiatry, Trinity College Dublin,
Trinity Centre for Health Sciences, James's St.,
Dublin 8, Ireland.

1 Mixture Model and Bayesian Model Interpretations

Combining Equations 2.1 and 2.2 (Main Text), we can consider the empirical-Bayes hierarchical model as a mixed model with both fixed and random effects in the frequentist setting as given by:

$$\text{logit}(p_{mi}) = \alpha_m + Z_m\gamma X_{mi} + X_{mi}\tau^2 t_{mm},$$

where $Z_m\gamma X_{mi}$ would be considered a fixed effect and $X_{mi}\tau^2 t_{mm}$ a random effect. Alternatively, this model may be viewed in the Bayesian framework, where the posterior distribution for the unknown parameters given the data is given by:

$$\begin{aligned} P(\{\alpha_m\},\{\beta_m\},\{\gamma_k\},\tau|\{y_i\}) \quad \propto \quad & \prod_{m=1}^{M}\prod_{i=1}^{N}[p_{mi}^{y_i}(1-p_{mi})^{1-y_i}] \\ & \times \prod_{m=1}^{M}\prod_{k=1}^{K}\left[\frac{1}{\sqrt{2\pi\tau^2 t_{mm}}}\exp\left(\frac{-(\beta_m - Z_{mk}\gamma_k)^2}{2\tau^2 t_{mm}}\right)\right] \\ & \times\pi(\tau^2)\prod_{m=1}^{M}\pi(\alpha_m)\prod_{k=1}^{K}\pi(\gamma_k). \end{aligned}$$

The term $\prod_{m=1}^{M}\prod_{k=1}^{K}\left[\sqrt{2\pi\tau t_{mm}}^{1/2}\exp\left(-(\beta_m - Z_{mk}\gamma_k)^2(2\tau^2 t_{mm})^{-1}\right)\right]$, which is the second-level regression (see Equation 2.2 (Main Text)) plays the role of an informative prior distribution on the unknown effect size parameters β. The term $\prod_{m=1}^{M}\prod_{i=1}^{N}[p_{mi}^{y_i}(1-p_{mi})^{1-y_i}]$ is the

[1]Corresponding author. E-mail: eaheron@tcd.ie
Tel: +353(1)8962424

likelihood with the remaining terms, $\pi(.)$'s, being prior distributions on the other unknown parameters. In this context τ^2 and $\{\gamma_k\}$ also have prior distributions as they are also unknown parameters. A fully Bayesian approach would allow for posterior distributional estimates of all the unknown parameters of the model. For large numbers of markers, estimating the unknown parameters of the model would not be feasible due to the high dimensionality and complexity of the resulting model and hence empirical-Bayes approaches have been developed. Presenting the full Bayesian model here helps to clarify the role played by the various components of the model.

2 Empirical-Bayes Shrinkage Factors

Starting with γ, a weighted least squares estimator is given by:

$$\hat{\gamma} = (Z^T D Z)^{-1} Z^T D \hat{\beta}, \tag{1}$$

where D is the diagonal matrix of weights $D = \text{Diag}(w_1, \ldots, w_M)$, $w_m = (V_m + \hat{\tau}^2)^{-1}$, $V_m = \text{SE}(\hat{\beta}_m)^2$. An approximately unbiased estimator of τ^2, the second-level variance, is given by:

$$\hat{\tau}^2 = \frac{\sum_{m=1}^{M} w_m \{(M/(M-K))(\hat{\beta}_m - Z_m^T \hat{\gamma})^2 - V_m\}}{\sum_{m=1}^{M} w_m}. \tag{2}$$

The process continues by iterating over Equations 1 and 2 until convergence is reached. An arbitrary initial guess is chosen for $\hat{\tau}^2$. If $\hat{\tau}^2$ becomes negative over the course of the iterations it is forced to be non-negative ($\hat{\tau}^2 \geq 0$). The shrinkage factor for each marker is then given by:

$$\hat{B}_m = ((M-K-2)/(M-K)) V_m w_m.$$

For the approximate empirical-Bayes confidence intervals the square of the standard error is given by (Morris (1983)):

$$s_m^2 = V_m[1 - ((M - \hat{k}_m)/M)\hat{B}_m] + v_m(\hat{\beta}_m - Z_m^T \hat{\gamma})^2, \tag{3}$$

where $\hat{k}_m = M w_m [Z(Z^T D Z)^{-1} Z^T]_{mm}$ and $v_m = (2/(M-K-2))\hat{B}_m^2(\bar{V} + \hat{\tau}^2)/(V_m + \hat{\tau}^2)$, which is an approximation for the variance of $\hat{B}_m$ and the average variance is given by: $\bar{V} = (\sum w_m V_m)/\sum w_m$. Thus a 95% approximate confidence interval for the empirical-Bayes estimate of the logarithm of the odds ratio is given by:

$$\hat{\beta}_{m;EB} \pm 1.96 s_m. \tag{4}$$

For further details see Morris (1983).

3 Simulation Strategy

To simulate case-control data for the empirical-Bayes hierarchical model we begin by fixing the number of cases and controls to be N_1 and N_0, respectively. Also the number of markers (SNPs), M, is set. We then generate data by starting at the top level of the hierarchical model specified in Equation 2.2 (Main Text). The additional genotypic covariate information summarised in the matrix Z is simulated as either independent binary or multinomial variates. We note that in real experimental covariate data, many of the entries in Z would be zero and we try to mimic this with the choice of parameters for either the Binomial or Multinomial distributions from which we simulate. We choose one of the columns of Z to contain the most informative additional covariate information and hence the most influential as regards a SNP being causative or associated. This is a reasonable assumption as it will often be the case that one or more of the additional types of genotypic data will be more informative than the others (a rationale used by Chen and Witte (2007) in the development of the weighting function in their empirical estimators).

At this stage a number of SNPs (at least 1) are chosen to be associated, by this we mean that in our simulation studies the odds ratios for these particular SNPs will be large relative to all other SNPs simulated. The coefficients γ of the second-level regression, given in Equation 2.2 (Main Text), are fixed as constants, with the coefficient of the covariate that was chosen to be most informative set at a higher value with respect to all others. The mean of the second-level regression is then given by $Z\gamma$. A variance-covariance matrix is constructed: $\tau^2 T$. This matrix represents the residual variation that remains in the first-level coefficients (the β_m's) after the second-level covariate information has been incorporated. For computational ease we do not simulate any correlations between the SNPs and so T can be set as the identity matrix. τ^2 is not chosen to vary across SNPs but rather it is constant and thus we are, in this case, modelling equal residual variability across the SNPs. Given the mean and variance we can now simulate β_m's from the appropriate Normal distributions. This completes the first part (second-level) of the simulation.

The second part of the simulation study involves simulating realistic genotypes for the N individuals for each of the M SNPs. We start with the MAFs, which can be either fixed at constant values, or they can be randomly drawn from a Uniform distribution on $(0, 0.5)$. In order to simulate the genotypes it is also necessary to consider the prevalence of the disease, which is defined to be the probability of disease, and which we fix at a constant value. Simulation of the genotypes for each marker is considered for cases and controls

separately. Specifically, for each marker, for each individual, we simulate the genotypes using a multinomial distribution with parameters $P(aa_m|\text{case})$, $P(Aa_m|\text{case})$ and $P(AA_m|\text{case})$, if the individual is a case and $P(aa_m|\text{control})$, $P(Aa_m|\text{control})$ and $P(AA_m|\text{control})$, if the individual is a control, for the three genotypes aa_m, Aa_m and AA_m. In order to calculate these six probabilities for each marker it is necessary to specify the genetic model, e.g, dominant, additive, etc., and to calculate the baseline risk of disease for each marker, α_m, as all other parameters have been set. Further details of how these are calculated are given below. Thus we now have a strategy for simulating a set of cases and controls with each of their genotypes for the M markers. All data simulation and analysis is carried out in the statistical computing language: R (2009).

3.1 Probability of an Individual Being a Case Given Their Genotype

The probability that an individual has the genotype aa_m at marker m given that they are a case is given by the following:

$$\begin{aligned} P(aa_m|\text{case}) &= \frac{P(\text{case}|aa_m)P(aa_m)}{P(\text{case})} \\ &= \frac{P(\text{case}|aa_m)(1-f_m)^2}{p}, \end{aligned}$$

where f_m is the MAF (allele A is the minor allele) at the marker m and p is the prevalence of the disease. Both of these are fixed in the simulation study. $P(\text{case}|aa_m) = \exp(\alpha_m)/(1+\exp(\alpha_m))$, and it is necessary to solve for α_m, the baseline risk of disease for each marker m. Similarly, expressions for the probabilities of the other genotypes given case-control status can also be derived. Solving for α_m allows for all of these to be calculated.

3.2 Baseline Risk for the Dominant Model

In order to obtain an expression for the baseline risk of disease for each marker, α_m, the genetic model must be taken into account. For a dominant model, using the logistic regression given in Equation 2.1 (Main Text) and assuming Hardy-Weinberg Equilibrium, the prevalence of disease is given by:

$$p = \frac{e^{\alpha_m}}{1+e^{\alpha_m}}(1-f_m)^2 + \frac{e^{\alpha_m+\beta_m}}{1+e^{\alpha_m+\beta_m}}(2f_m - f_m^2), \tag{5}$$

To obtain an expression for α_m we re-write Equation 5 as a quadratic equation in $e^{-\alpha_m}$. To simplify notation we let $x = e^{-\alpha_m}$, $y = e^{-\beta_m}$, and $f = f_m$. The quadratic equation is then

given by:

$$x^2py + xB + p - 1 = 0, \tag{6}$$

where $B = (2f - f^2 - 1)y + (f^2 - 2f) + (p + yp)$. Solving Equation 6 and considering only the positive root, results in

$$\alpha_m = -\log\left(\frac{(-B + \sqrt{(B^2 + 4(1-p)yp))}}{2yp}\right) \tag{7}$$

as an expression for the baseline risk of disease for the dominant model.

3.3 Baseline Risk for the Additive Model

For the additive model, it is necessary to solve a cubic equation given by:

$$\begin{aligned}
&x^3y^3p - x^2[y^3(1 - f^2 - p) + y^2(2f - 2f^2 - p) + y(f^2 - p)] \\
&-x[(y^2 + y)(1 - f^2) + 2(y^2 + 1)(f - f^2) + (y + 1)f^2 - p(y^2 + y + 1)] \\
&-(1 + 2(f - f^2) - p) - p = 0,
\end{aligned}$$

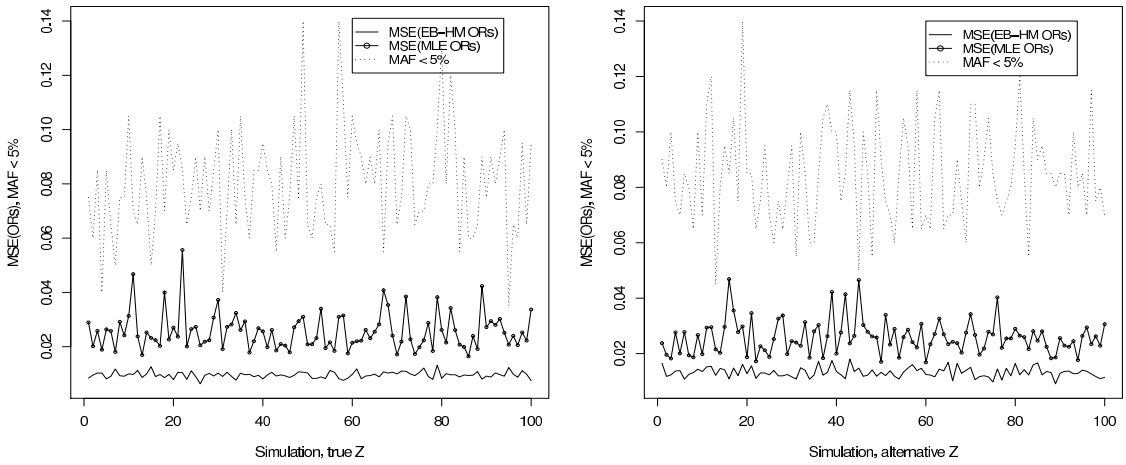
where x, y, f, p are as in Equation 7.

References

Chen, G.K., and Witte, J.S. (2007), "Enriching the Anlaysis of Genomewide Association Studies with Hierarchical Modeling," *The American Journal of Human Genetics*, 81, 397-404.

Morris C.N. (1983), "Parametric Empirical-Bayes Inference: Theory and Applications," *Journal of the American Statistical Association*, 78(381), 47-55.

R Development Core Team (2009), "R: A Language and Environment for Statistical Computing," *R Foundation for Statistical Computing*, Vienna, Austria, ISBN: 3-900051-07-0, http://www.R-project.org.



(a) (b)

Figure 1: (a) MSE for both the MLEs and the empirical-Bayes estimators (EB-HM) across a range of 100 random simulations, when the same Z matrix is used in the simulation and analysis phases. Also plotted is the proportion of MAFs $< 5\%$ in each simulation. (b) MSE for both the MLEs and the empirical-Bayes estimators across a range of 100 random simulations, when the alternative reduced, noisy Z matrix is used in the analysis phase. Also plotted is the proportion of MAFs $< 5\%$ in each simulation.

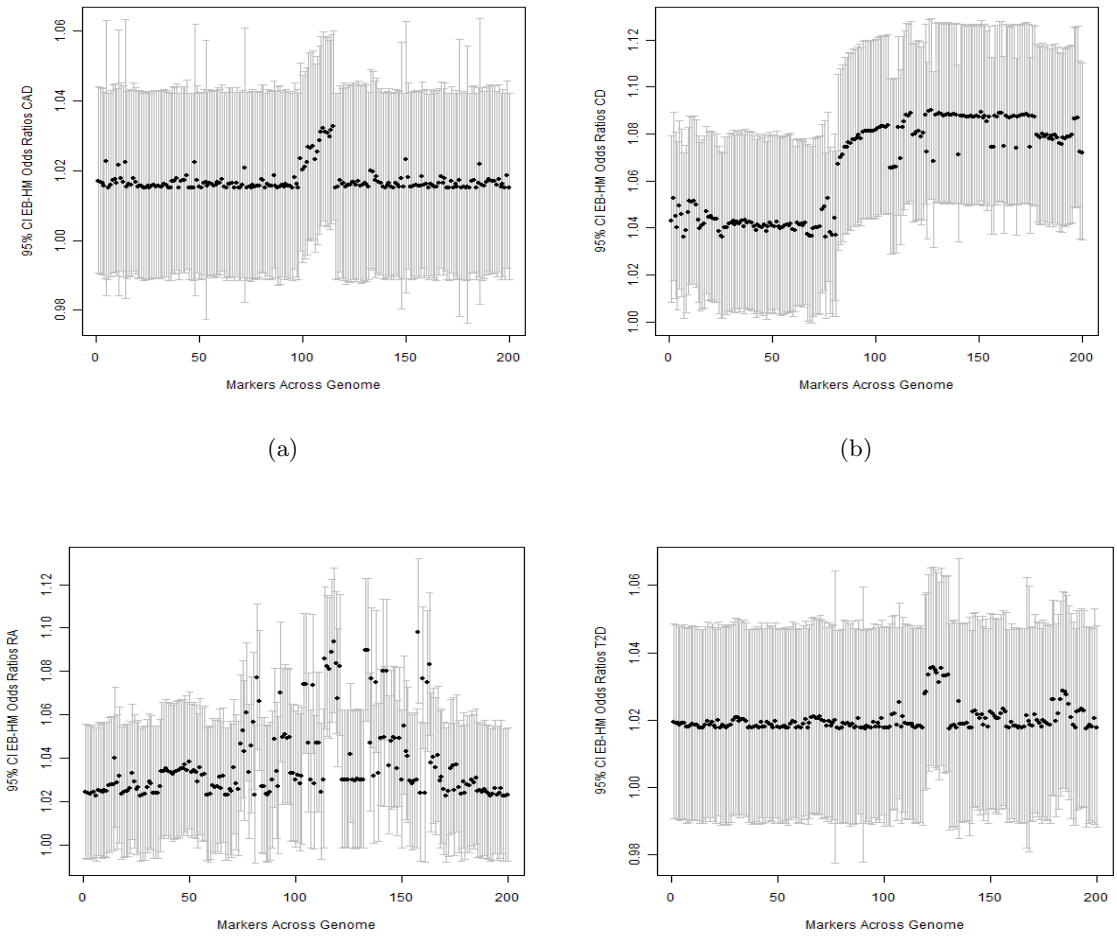


Figure 2: (b) The top 200 markers with the largest EB-HM odds ratios together with their approximate 95% confidence intervals for (a) CAD, (b) CD, (c) RA, and (d) T2D.