

# Bayesian Meta-Analysis of Genetic Association Studies with Different Sets of Markers

Claudio Verzilli,<sup>2,10</sup> Tina Shah,<sup>1,10</sup> Juan P. Casas,<sup>2</sup> Juliet Chapman,<sup>2</sup> Manjinder Sandhu,<sup>3</sup> Sally L. Debenham,<sup>3</sup> Matthijs S. Boekholdt,<sup>4</sup> Kay Tee Khaw,<sup>3</sup> Nicholas J. Wareham,<sup>5</sup> Richard Judson,<sup>6</sup> Emelia J. Benjamin,<sup>7</sup> Sekar Kathiresan,<sup>7</sup> Martin G. Larson,<sup>7</sup> Jian Rong,<sup>7</sup> Reecha Sofat,<sup>1</sup> Steve E. Humphries,<sup>8</sup> Liam Smeeth,<sup>2</sup> Gianpiero Cavalleri,<sup>9</sup> John C. Whittaker,<sup>2,\*</sup> and Aroon D. Hingorani<sup>1</sup>

Robust assessment of genetic effects on quantitative traits or complex-disease risk requires synthesis of evidence from multiple studies. Frequently, studies have genotyped partially overlapping sets of SNPs within a gene or region of interest, hampering attempts to combine all the available data. By using the example of C-reactive protein (CRP) as a quantitative trait, we show how linkage disequilibrium in and around its gene facilitates use of Bayesian hierarchical models to integrate informative data from all available genetic association studies of this trait, irrespective of the SNP typed. A variable selection scheme, followed by contextualization of SNPs exhibiting independent associations within the haplotype structure of the gene, enhanced our ability to infer likely causal variants in this region with population-scale data. This strategy, based on data from a literature based systematic review and substantial new genotyping, facilitated the most comprehensive evaluation to date of the role of variants governing CRP levels, providing important information on the minimal subset of SNPs necessary for comprehensive evaluation of the likely causal relevance of elevated CRP levels for coronary-heart-disease risk by Mendelian randomization. The same method could be applied to evidence synthesis of other quantitative traits, whenever the typed SNPs vary among studies, and to assist fine mapping of causal variants.

## Introduction

Genetic effects underlying complex traits and disorders are small, and their detection requires comprehensive typing of single nucleotide polymorphisms (SNPs) in large samples.<sup>1,2</sup> Many previous genetic association studies have been underpowered,<sup>3,4</sup> and even very large biobanks<sup>5</sup> may not individually provide conclusive results for certain outcomes. Quantitative synthesis of evidence from available studies remains vital,<sup>6–8</sup> even in the era of genome-wide analyses.<sup>9–11</sup> However, a major obstacle is that studies of the same gene, region, or even the genome as a whole may type a different repertoire of SNPs, thereby yielding partially overlapping genotypic data. Moreover, often only single SNP summary data, for instance genotype means at each SNP, is reported.

The meta-analysis of results from each marker in isolation would exclude those studies that did not type the marker in question, with a potential loss of power; moreover, multiple single-SNP analyses are difficult to interpret. Instead, it would be useful to be able to combine data with information from all sites, adjusting any association at each site for the possible correlation with the remaining variants. One could then disentangle effects at causal sites

from those at sites that are in LD with a causal variant(s) and also borrow information across studies. With focus on a quantitative trait, we develop a Bayesian hierarchical linear regression that models linear transformations of the study-specific genotype-group-specific phenotypic means and that uses pairwise LD measurements between markers to make posterior inference on adjusted effects. Information on pairwise marker LD is often provided by the individual studies as part of the results reported. Alternatively, for markers that are not considered jointly in any of the study at hand, it can often be obtained from public databases. This information is then used to specify informative priors in our Bayesian framework. Specifically, the between-marker correlations are modeled by introduction of spatially correlated random effects having a conditional autoregressive distribution (CAR).<sup>12,13</sup> The between-study variability is then accommodated with a random intercept term across studies.

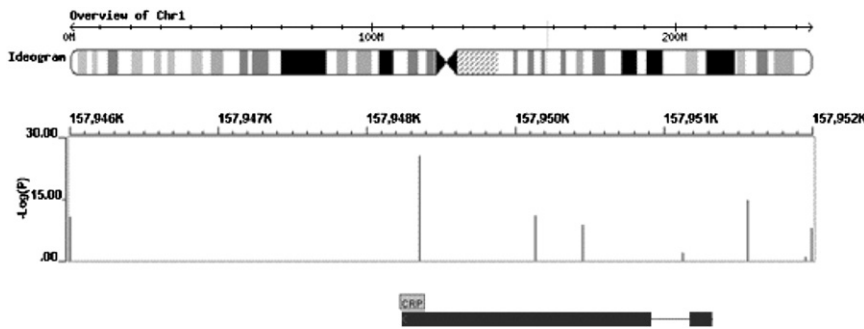
Our approach is motivated by the meta-analysis of studies assessing the effect of variants in the C-reactive protein (CRP [MIM 123260]) gene region on plasma CRP levels. CRP is a circulating monomorphic hepatic acute-phase protein that indexes and may mediate aspects of the inflammatory response.<sup>14</sup> Aside from acute-phase elevations,

<sup>1</sup>Centre for Clinical Pharmacology, University College London, London WC1E 6JF, UK; <sup>2</sup>Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK; <sup>3</sup>Department of Public Health and Primary Care, University of Cambridge, Cambridge CB1 8RN, UK; <sup>4</sup>Department of Cardiology, Academic Medical Center, Amsterdam 1100 DD, Netherlands; <sup>5</sup>MRC Epidemiology Unit, University of Cambridge, Cambridge CB2 0QQ, UK; <sup>6</sup>Genaissance Pharmaceuticals, New Haven, CT 06511, USA; <sup>7</sup>Framingham Heart Study, Framingham, MA 01702-5827, USA; <sup>8</sup>Centre for Cardiovascular Genetics, University College London, London WC1E 6JF, UK; <sup>9</sup>Molecular and Cellular Therapeutics, RCSI Research Institute Royal College of Surgeons in Ireland, Dublin 2, Ireland

<sup>10</sup>These authors contributed equally to this work.

\*Correspondence: [john.whittaker@shtm.ac.uk](mailto:john.whittaker@shtm.ac.uk)

DOI 10.1016/j.ajhg.2008.01.016. ©2008 by The American Society of Human Genetics. All rights reserved.



**Figure 1. Location of the Eight CRP SNPs Typed Directly in the 26 Data Sets Included in This Study**

The upper track shows chromosomal location; the middle track shows SNP location and Log(P) for the per-allele random-effect meta-analysis (from left to right, the SNPs are ordered as follows: rs3093077, rs1205, rs1130864, rs1800947, rs1417938, rs3091244, rs2794521, and rs3093059); and the lower track shows the intron/exon structure of the *CRP* gene.

blood concentrations of CRP show similar within-individual variability to serum cholesterol, and like cholesterol, CRP has been shown to be associated with future coronary heart disease (CHD) risk in observational studies.<sup>15</sup> However, the etiological relevance of this potentially important and highly studied link with CHD is uncertain because CRP may simply be a marker for established risk factors or for subclinical atheroma.<sup>16,17</sup> Common SNPs that are in the gene encoding CRP and that influence its level may help provide insight on the link because, unlike CRP itself, genotype is fixed and unaffected by subclinical disease and the naturally randomized allocation of alleles at conception balances the distribution of potential confounding factors among genotypic classes. Genetic associations are therefore less prone to biases that limit causal inference from observational studies, and genetic studies possess properties of a randomized intervention trial.<sup>16–18</sup> Therefore, identification of *CRP*-gene variants (HGNC: 2637; 1q21–q23) that influence its concentration is fundamental to evaluating the causal relevance of *CRP* with the principle of Mendelian randomization.<sup>19</sup>

In the absence of hepatic stores of CRP, and given its constant rate of clearance, gene transcription provides the major point of regulation.<sup>14</sup> Transcription may be modified by regulatory SNPs because concentrations of CRP show strong concordance among monozygotic twins and family studies suggest substantial heritability.<sup>20</sup> In populations of European descent, there are 11 common SNPs with minor allele frequency >5% within 6 kb of the *CRP* gene, but extensive linkage disequilibrium (LD) means that four major haplotypes account for 94% of chromosomes (see [Web Resources](#)).<sup>21,22</sup> Individual reports evaluating associations of *CRP* SNPs with CRP concentration have either typed single SNPs or a subset of SNPs (sometimes tag SNPs) in this region (see [Table S1](#) available online). However, the SNPs have varied across studies, thereby limiting the ability to pool all available data. We therefore developed a new integrative approach to evidence synthesis of genetic association studies that allows for this complexity.

Methods for combining data from genome-wide scans with nonoverlapping sets of SNPs with individual-level genotyping data have been recently proposed by Marchini et al.<sup>23</sup> Here, because individual-level data are not available for most of the studies on *CRP*, we develop

a method that allows the synthesis of studies providing only summary data. Also, we are mainly concerned with the synthesis of SNP data in regions of interest for fine mapping, where the number of markers typed is small and interest is on disentangling independent effects using a variable selection scheme, for which the Marchini approach is not suitable.

## Material and Methods

We first conducted a literature-based systematic review of all relevant studies (irrespective of the SNP typed). A total of 23 published data sets identified by systematic review evaluated associations of eight SNPs (rs3093059; rs2794521; rs3091244; rs1417938; rs1800947; rs1130864; rs1205; and rs3093077) in the *CRP* gene with CRP concentration ([Figure 1](#)). With data from SeattleSNPs, a combination of three SNPs (rs1130864; rs1205; and rs3093077) was identified as haplotype tag SNPs with the haplotype  $r^2$  method in European subjects. These tag SNPs were typed in three additional population-based studies, thereby giving an aggregate of 26 studies including 32,802 subjects. No SNP was typed in every study, but there was partial overlap of SNP typing across several studies (see [Appendix A](#) and [Table S1](#)).

## Bayesian Hierarchical Model

We indicate with  $Y_i^s$  the continuous trait of interest for subject  $i \in \{1, \dots, n^s\}$  and study  $s \in \{1, \dots, S\}$ . If all studies have genotyped individuals at all  $m$  marker locations, and these data are available for all individuals (individual patient data [IPD]), a sensible approach to pool information across studies would be the random-effect model

$$\mathbf{Y}^s \sim \mathbf{N}(\mathbf{C}^s \boldsymbol{\beta} + \mathbf{1}_m \mu_s, \sigma^2 \mathbf{I}_m) \quad (1)$$

where  $\mathbf{C}^s$  is the  $n^s \times (m + 1)$  design matrix coding for the chosen genetic model (e.g., for an additive models, 0, 1, and 2 for homozygous wild-type, heterozygous, or homozygous mutant genotypes, respectively) and the intercept term,  $\mu_s \sim \mathbf{N}(0, \sigma_s^2)$  is a study-specific random intercept term,  $\mathbf{1}_m$  is the  $n^s \times 1$  vector of ones,  $\mathbf{I}_m$  is the  $n^s \times n^s$  identity matrix, and  $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_m)'$  is the  $(m + 1) \times 1$  vector of regression coefficients of interest measuring the effect of genotype group on  $Y$ . One could then assess the relative importance of each marker by using a variable selection scheme; we use a reversible jump algorithm on the space of possible models as part of the MCMC scheme as described later in the text.<sup>24,25</sup>

However, studies will rarely consider all  $m$  markers together; rather,  $m_s \leq m$  will have been typed in study  $s$  corresponding to

a subset  $L_s$  of columns of the complete design matrix  $\mathbf{C}^s$ ,  $\mathbf{X}^s$  say of size  $n^s \times (m_s + 1)$ . Also, complete individual patient data for all studies are rarely available. Instead, we have the summary statistics reported in each study as in the case of the CRP studies. Typically, data will consist of means, variances, and numbers of individuals for each genotype groups and each marker. These are denoted by  $\bar{y}_{gj}^s$ ,  $v_{gj}^s$ , and  $n_{gj}^s$ , respectively, for genotype group  $g = 1, \dots, G_j$  of marker  $j \in \{L_s\}$  in study  $s$ . The notation allows for marker-specific numbers of genotype groups and thus the possibility of having a mixture of biallelic and triallelic markers, as in the application to the CRP data, or different genetic models.

Our approach uses Equation (1) as the building block but models the linear transformations  $\mathbf{T}^s = \mathbf{X}^s \mathbf{Y}^s$  as multivariate normally distributed across studies

$$\mathbf{T}^s \sim \text{MVN}_{m_s} \left( \mathbf{X}^s \mathbf{C}^s \beta + \mathbf{X}^s \mathbf{1}_n \mu_s, \sigma^2 \mathbf{X}^s \mathbf{X}^s \right) \quad (2)$$

where  $\mathbf{X}^s$  indicates the transpose of  $\mathbf{X}^s$ . All entries of the vector  $\mathbf{T}^s$  can be obtained from the available data summaries. For instance, the first element corresponding to the intercept term is the overall sum of the  $y$  values, and any other entry can be obtained similarly from the genotype-group-specific phenotype means and counts  $\bar{y}_{gj}^s$ ,  $n_{gj}^s$ .

However, the new design matrix  $\mathbf{W}^s = \mathbf{X}^s \mathbf{C}^s$  is only partially observed. In particular, only the dot products involving the columns of  $\mathbf{X}^s$  with themselves or the intercept term can be derived from the observed genotype-group counts. The remaining entries are replaced by their expected values under Hardy-Weinberg equilibrium (HWE) and the known pairwise LD patterns. Specifically, indicating with  $w_{hl}$ ,  $h \neq l$ , a generic such entry, we first obtain an estimate of the joint bivariate genotype distribution from the known marginal allele frequencies and the pairwise measure of LD.<sup>26</sup> For example, if both markers are biallelic, this involves estimation of the  $3 \times 3$  matrix of the genotype distribution, and this estimation is then multiplied by the study size to give expected counts. Finally, we obtained  $w_{hl}$  by summing the appropriate entries of the resulting matrix of expected counts multiplied by the values used to code the genotype groups in the design matrices. Notice that the vector of coefficients  $\beta$  retains the same interpretation and scale of the original model in Equation (1) (in the example below, additive effect of variants on log CRP plasma levels) because it is derived from a linear transformation of the variables therein.

As well as in the derivation of the new design matrix  $\mathbf{W}^s$ , prior information of between-marker LD patterns is also incorporated in the specification of the (partially unobserved) variance-covariance matrix in Equation (2). Specifically, we partition  $\sigma^2 \mathbf{X}^s \mathbf{X}^s$  into a spatially structured component and a residual, unstructured, component. We obtained the former by introducing marker-specific random effects having a zero-mean conditional autoregressive distribution

$$\mathbf{U} \sim \text{CAR} \left( \mathbf{0}, \sigma_u^2 (\mathbf{I} - \gamma \mathbf{R})^{-1} \mathbf{M} \right) \quad (3)$$

where  $\mathbf{U}$  is a vector of size  $m$ , the number of unique markers across studies,  $\mathbf{R}$  is a matrix of weights reflecting spatial associations between the elements of  $\mathbf{T}^s$ , and  $\mathbf{M}$  is a diagonal matrix.<sup>12,27</sup> Thus the covariance matrix in Equation (2) becomes  $\sigma_u^2 (\mathbf{I} - \gamma \mathbf{R})^{-1} \mathbf{M} + \sigma_\epsilon^2 \text{diag}(\mathbf{X}^s \mathbf{X}^s)$  where we set  $\gamma \equiv 1$  and  $(\mathbf{I} - \gamma \mathbf{R})^{-1} \mathbf{M} \equiv (\mathbf{X}^s \mathbf{X}^s - \text{diag}(\mathbf{X}^s \mathbf{X}^s))$ , where with  $\mathbf{X}^s \mathbf{X}^s$  we indicate the weighted average of the study-specific cross-products  $\mathbf{X}^s \mathbf{X}^s$  with weights given by the number of subjects in each study. The latter equivalence then reflects our prior information on pairwise LD, as the off-diagonal

elements of the matrix  $\mathbf{X}^s \mathbf{X}^s$  are replaced by their expected values given the LD patterns.

Conditional on the study and marker-specific random effects  $\mu_s$  and  $U_j$ , the elements of  $\mathbf{T}^s$  are independent and we can rewrite Equation (2) as

$$T_j^s | \mu_s, U_j \sim \mathcal{N} \left( \mathbf{W}_{j, \cdot} \beta + \sum_{i=1}^{n_s} x_{i,j}^s \mu_s + U_j, \sigma_\epsilon^2 \text{diag} \left( \mathbf{X}^s \mathbf{X}^s \right)_{jj} \right) \quad (4)$$

$$j \in \{L_s\}.$$

We further assume that the marker and study-specific pooled variances

$$v_j^s = \frac{\sum_g (n_{gj}^s - 1) v_{gj}^s}{\sum_g (n_{gj}^s - 1)}$$

have a scaled chi-square distribution

$$v_j^s = \frac{\sigma_\epsilon^2}{\sum_g n_{gj}^s - G_j} \chi_{\sum_g n_{gj}^s - G_j}^2.$$

Albeit an approximation, this assumption is likely to hold when the individual SNP associations are modest, as it is reasonable to expect in this setting. Also, by modeling the observed variances, we are able to impute any missing values from their full conditional distribution as part of the MCMC scheme.

The hierarchical specification is completed by assuming a distribution for the between-studies random effects  $\mu_s$ . In order to accommodate outliers and heavy tails, we assume the mixture of normals

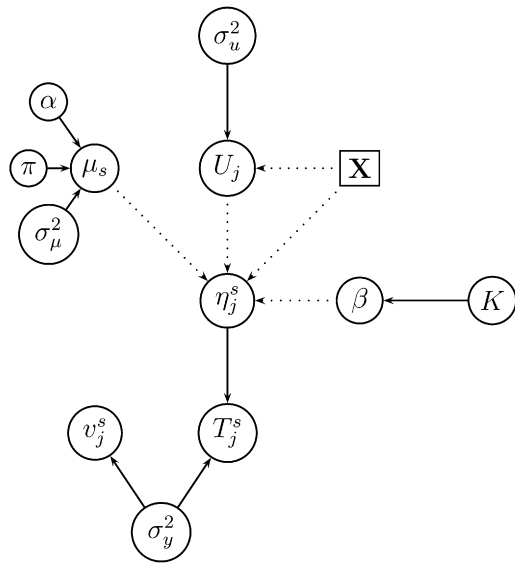
$$\mu_s \sim \pi \mathcal{N}(\alpha_1, \sigma_\mu^2) + (1 - \pi) \mathcal{N}(\alpha_2, \sigma_\mu^2) \quad (5)$$

with  $\alpha_2 = -\pi \alpha_1 / (1 - \pi)$ .<sup>28</sup>

To select important marker-phenotype associations, we use a reversible jump algorithm on the space of models in the MCMC scheme.<sup>24,25</sup> In brief, denoting with  $k$  the number of SNP currently included in a model, we made a proposal to change the current model by adding a marker, deleting a marker, or swapping a marker currently in the model with one from the remaining SNPs. The new model is accepted with probability proportional to its likelihood. Conditional on the accepted model, new values for the appropriate subset of parameters in  $\beta$  are then sampled from their full conditional distribution. By monitoring the different models visited, we readily obtained posterior model probabilities. The algorithm is not guaranteed to visit all possible models, but in many cases, if the number of available predictors is not large as here, an acceptable qualitative assessment of the support received from the data by the different models is possible. Finally, prior distributions for all remaining unknown parameters are as follows

$$\begin{aligned} \beta_j &\sim \mathcal{N}(0, \sigma_\beta^2), \quad j = 1, \dots, m \\ \alpha_1 &\sim \mathcal{N}(0, 1e - 6) \\ \pi &\sim \text{Beta}(1, 1) \\ \sigma_u^2 &\sim \text{Gamma}(0.001, 0.001) \\ \sigma_\beta^2 &\sim \text{Gamma}(0.001, 0.001) \\ \sigma_U^2 &\sim \text{Gamma}(0.5, 0.0005) \\ \sigma_y^2 &\sim \text{Gamma}(0.001, 0.001) \end{aligned}$$

where the prior for the precision of the spatial effects is that suggested by Kensall and Wakefield.<sup>29</sup> The reversible jump algorithm requires specification of a prior on model size. Typical choices include a uniform prior on the model space or a Poisson or geometric



**Figure 2. Graphical Representation of Equation (4)**  
Solid and dotted lines represent stochastic and deterministic dependencies, respectively.

distribution on the number of regression terms  $k$  included in each model.<sup>25,30,31</sup> The simulation study in the next section includes a sensitivity analysis of this choice. A graphical representation of the hierarchical model (4) is given in Figure 2.

### Single-Marker Random-Effect Meta-Analysis

Results from the multilocus model are compared to those obtained from a more traditional single-locus random-effects meta-analysis in both simulation studies and with real data from the *CRP*-gene region. For the latter, a per-allele effect (95% CI) of individual SNPs on CRP concentration was derived from each individual study. The individual-study linear trend (additive effect) per category increase in genotype with mean data was calculated by simple linear regression, with genotypes coded as 0, 1, and 2 for homozygous common allele, heterozygous, and homozygous rare-allele, respectively, with the least-square linear-trend-coefficient formula, which only depends on the mean values and its standard deviations. A sensitivity analysis restricted to studies with more than 500 subjects, healthy at time of blood sampling, or to studies that reported all the required standard deviations was also conducted (Table S2). Subsequently, the study-specific linear trend and its standard error were pooled with random-effect models. Subsidiary analyses included pairwise comparisons within each polymorphism. The DerSimonian and Laird Q test, and the  $I^2$  test,<sup>32</sup> were used for evaluating the degree of heterogeneity between studies.

## Results

### Simulation Studies

We considered various scenarios differing in the number of studies and, for the multilocus approach, in the priors on the model space. Data were obtained as follows: We first simulated a pool of 4000 haplotypes at seven biallelic markers. Pairwise LD measures ( $r$ ) between the seven

|          |   | Linkage Disequilibrium |        |         |         |         |         |
|----------|---|------------------------|--------|---------|---------|---------|---------|
|          |   | 2                      | 3      | 4       | 5       | 6       | 7       |
| Marker 1 | 1 | 0.0418                 | 0.0055 | -0.0300 | 0.0124  | 0.0260  | 0.0144  |
|          | 2 |                        | 0.0394 | 0.0056  | 0.0148  | 0.0154  | 0.0141  |
|          | 3 |                        |        | -0.1941 | -0.0784 | -0.0567 | -0.0471 |
|          | 4 |                        |        |         | 0.3697  | 0.2737  | 0.2443  |
|          | 5 |                        |        |         |         | 0.7388  | 0.6181  |
|          | 6 | $r$                    |        |         |         |         | 0.8361  |
|          |   | Marker 2               |        |         |         |         |         |

**Figure 3. Pairwise LD Measures between Markers Used in the Simulation Study**

Pairwise LD Measures are  $r$  values.

SNPs are shown in Figure 3, with high LD only between the last three markers. SNP 6 is assumed to be the single causal site in the region and is retained in all subsequent analyses. Given the high LD between SNP 5, 6, and 7, we expect the results from the univariate analyses to be less conclusive than those from the multiple marker approach that adjusts for the between-marker correlations. The study size  $n_s$  was drawn from a normal distribution with mean 600 and variance 100, rounded to the nearest integer. Then, for subject  $i \in \{1, \dots, n^s\}$  and study  $s$ , a continuous phenotype  $y_i^s$  is simulated as

$$y_i^s = \beta_0 + \beta_6 g_{i6} + \mu_s + \varepsilon_i \quad (6)$$

where  $g_{i6}$  denotes the genotype of subject  $i$  at marker site 6 (0, 1, or 2 for homozygous wild-type, heterozygous, or homozygous mutant, respectively),  $(\beta_0, \beta_6) = (1, 2)$ ,  $\mu_s \sim N(0, 1)$ , and  $\varepsilon \sim N(0, 1)$ . To reflect the fact that not all markers are typed in every study, we select at random  $m_s$  markers out of the possible seven for each study. Thus, in most cases the univariate analyses are based on fewer than the maximum total of  $S$  studies. For each simulated data set, we also estimated the unadjusted univariate additive effects and their standard errors at each SNP site; the additive effects are then combined in the univariate random-effect analyses.<sup>33,34</sup> Tables 1 and 2 present the results from the multiple-marker meta-analyses. The number of studies considered was 10, 20, or 40. In each case, the tables report the results obtained with Poisson priors on the model size in the reversible jump algorithm with different means (1 or 2 for priors a and b, respectively) or a uniform prior on the model space (prior c). Notice that the Poisson priors give more weight to the null model and may in general be a more reasonable choice in this setting. For example,

**Table 1. Bayesian Multilocus Meta-Analysis**

| Number of Studies | Prior | Parameter              | $\beta_1$       | $\beta_2$        | $\beta_3$       | $\beta_4$        | $\beta_5$        | $\beta_6$       | $\beta_7$       | $\sigma_y^2$   | $\sigma_\mu^2$ |                |
|-------------------|-------|------------------------|-----------------|------------------|-----------------|------------------|------------------|-----------------|-----------------|----------------|----------------|----------------|
|                   |       | True                   | 0               | 0                | 0               | 0                | 0                | 2               | 0               | 1              | 1              |                |
| 10                | a     | Post prob <sup>a</sup> | 0.01<br>(0.01)  | 0.01<br>(0.004)  | 0.01<br>(0.004) | 0.005<br>(0.002) | 0.01<br>(0.01)   | 1.00<br>(0.00)  | 0.03<br>(0.02)  |                |                |                |
|                   |       | Mean <sup>a</sup>      | -0.01<br>(0.05) | -0.004<br>(0.04) | 0.03<br>(0.03)  | 0.01<br>(0.03)   | -0.01<br>(0.07)  | 2.00<br>(0.04)  | 0.16<br>(0.14)  | 1.19<br>(0.01) | 1.25<br>(0.61) |                |
|                   |       | BCI length             | 0.19            | 0.20             | 0.15            | 0.14             | 0.35             | 0.19            | 0.76            |                |                |                |
|                   | b     | Post prob              | 0.02<br>(0.01)  | 0.01<br>(0.004)  | 0.01<br>(0.01)  | 0.01<br>(0.004)  | 0.02<br>(0.01)   | 1.00<br>(0.001) | 0.07<br>(0.08)  |                |                |                |
|                   |       | Mean                   | -0.03<br>(0.04) | -0.01<br>(0.03)  | 0.03<br>(0.03)  | 0.02<br>(0.03)   | -0.01<br>(0.08)  | 2.00<br>(0.05)  | 0.15<br>(0.16)  | 1.08<br>(0.02) | 1.13<br>(0.34) |                |
|                   |       | BCI length             | 0.24            | 0.22             | 0.16            | 0.15             | 0.36             | 0.23            | 0.76            |                |                |                |
|                   | c     | Post prob              | 0.05<br>(0.02)  | 0.04<br>(0.02)   | 0.04<br>(0.02)  | 0.03<br>(0.01)   | 0.07<br>(0.03)   | 1.00<br>(0.001) | 0.15<br>(0.05)  |                |                |                |
|                   |       | Mean                   | -0.03<br>(0.04) | -0.01<br>(0.04)  | 0.02<br>(0.03)  | 0.01<br>(0.03)   | -0.01<br>(0.07)  | 1.99<br>(0.04)  | 0.13<br>(0.12)  | 1.12<br>(0.01) | 1.42<br>(0.46) |                |
|                   |       | BCI length             | 0.24            | 0.21             | 0.17            | 0.15             | 0.35             | 0.33            | 0.76            |                |                |                |
|                   | 20    | a                      | Post prob       | 0.01<br>(0.01)   | 0.01<br>(0.002) | 0.01<br>(0.003)  | 0.004<br>(0.002) | 0.01<br>(0.003) | 1.00<br>(0.00)  | 0.03<br>(0.08) |                |                |
|                   |       |                        | Mean            | -0.04<br>(0.03)  | -0.02<br>(0.03) | 0.03<br>(0.02)   | 0.01<br>(0.02)   | -0.01<br>(0.05) | 1.99<br>(0.04)  | 0.09<br>(0.10) | 1.11<br>(0.01) | 1.10<br>(0.26) |
|                   |       |                        | BCI length      | 0.16             | 0.14            | 0.11             | 0.10             | 0.23            | 0.12            | 0.56           |                |                |
| b                 |       | Post prob              | 0.02<br>(0.02)  | 0.01<br>(0.01)   | 0.01<br>(0.01)  | 0.01<br>(0.003)  | 0.02<br>(0.01)   | 1.00<br>(0.001) | 0.05<br>(0.04)  |                |                |                |
|                   |       | Mean                   | -0.04<br>(0.03) | -0.02<br>(0.03)  | 0.02<br>(0.03)  | 0.01<br>(0.02)   | -0.02<br>(0.05)  | 1.99<br>(0.03)  | 0.13<br>(0.112) | 1.06<br>(0.01) | 1.08<br>(0.17) |                |
|                   |       | BCI length             | 0.16            | 0.15             | 0.12            | 0.10             | 0.24             | 0.17            | 0.57            |                |                |                |
| c                 |       | Post prob              | 0.06<br>(0.05)  | 0.03<br>(0.01)   | 0.03<br>(0.01)  | 0.02<br>(0.01)   | 0.06<br>(0.05)   | 1.00<br>(0.002) | 0.13<br>(0.08)  |                |                |                |
|                   |       | Mean                   | -0.05<br>(0.03) | -0.02<br>(0.02)  | 0.02<br>(0.02)  | 0.01<br>(0.02)   | -0.01<br>(0.06)  | 1.99<br>(0.03)  | 0.10<br>(0.108) | 1.09<br>(0.01) | 0.99<br>(0.16) |                |
|                   |       | BCI length             | 0.18            | 0.15             | 0.12            | 0.11             | 0.26             | 0.23            | 0.59            |                |                |                |

<sup>a</sup> Results are averages (std) over 100 replicated data sets. Mean posterior estimates and credible intervals are conditional on the SNP being included in a model.

prior a is Poisson(1) and assigns a probability of ~0.26 of having more than one associated site (~0.59 for b). The values shown are averages over 100 replicates. For each scenario, we report the marginal posterior probability of selecting each SNP and the mean and 95% credible intervals

of the posterior distributions of each additive effect, conditional on the SNP being selected.<sup>35,36</sup> Note that posterior distributions can be reliably estimated only for markers with relatively high posterior probability of inclusion (e.g., >0.5), and results in the table should be interpreted

**Table 2. Bayesian Multilocus Meta-Analysis**

| Number of Studies | Prior | Parameter  | $\beta_1$       | $\beta_2$       | $\beta_3$        | $\beta_4$        | $\beta_5$       | $\beta_6$      | $\beta_7$       | $\sigma_y^2$   | $\sigma_\mu^2$ |
|-------------------|-------|------------|-----------------|-----------------|------------------|------------------|-----------------|----------------|-----------------|----------------|----------------|
|                   |       | True       | 0               | 0               | 0                | 0                | 0               | 2              | 0               | 1              | 1              |
| 40                | a     | Post prob  | 0.01<br>(0.15)  | 0.01<br>(0.003) | 0.004<br>(0.002) | 0.003<br>(0.001) | 0.01<br>(0.004) | 1 (0)          | 0.025<br>(0.28) |                |                |
|                   |       | Mean       | -0.05<br>(0.02) | -0.03<br>(0.02) | 0.02<br>(0.02)   | 0.01<br>(0.01)   | -0.02<br>(0.04) | 1.99<br>(0.02) | 0.10<br>(0.08)  | 1.04<br>(0.01) | 1.09<br>(0.14) |
|                   |       | BCI length | 0.12            | 0.10            | 0.08             | 0.07             | 0.17            | 0.10           | 0.46            |                |                |
|                   | b     | Post prob  | 0.03<br>(0.03)  | 0.01<br>(0.01)  | 0.01<br>(0.00)   | 0.01<br>(0.002)  | 0.02<br>(0.01)  | 1 (0)          | 0.04<br>(0.03)  |                |                |
|                   |       | Mean       | -0.05<br>(0.02) | -0.02<br>(0.02) | 0.02<br>(0.02)   | 0.01<br>(0.02)   | -0.03<br>(0.03) | 1.99<br>(0.02) | 0.12<br>(0.07)  | 1.03<br>(0.01) | 0.97<br>(0.13) |
|                   |       | BCI length | 0.12            | 0.11            | 0.08             | 0.07             | 0.19            | 0.12           | 0.46            |                |                |
|                   | c     | Post prob  | 0.07<br>(0.06)  | 0.03<br>(0.01)  | 0.02<br>(0.01)   | 0.01<br>(0.01)   | 0.04<br>(0.02)  | 1 (0)          | 0.11<br>(0.06)  |                |                |
|                   |       | Mean       | -0.05<br>(0.02) | -0.03<br>(0.02) | 0.02<br>(0.02)   | 0.01<br>(0.01)   | -0.01<br>(0.04) | 1.99<br>(0.02) | 0.10<br>(0.10)  | 1.03<br>(0.01) | 1.01<br>(0.11) |
|                   |       | BCI length | 0.12            | 0.11            | 0.10             | 0.077            | 0.19            | 0.18           | 0.48            |                |                |

Results are averages (std) over 100 replicated data sets. Mean posterior estimates and credible intervals are conditional on the SNP being included in a model.

**Table 3. Single-Locus Random-Effects Meta-Analysis**

|                   |    | SNP ID          | 1                 | 2                 | 3                | 4                | 5                | 6                | 7                |
|-------------------|----|-----------------|-------------------|-------------------|------------------|------------------|------------------|------------------|------------------|
| Number of Studies |    | True            | 0                 | 0                 | 0                | 0                | 0                | 2                | 0                |
|                   | 10 | Mean (std)      | -0.033<br>(0.028) | -0.023<br>(0.024) | 0.069<br>(0.026) | 0.312<br>(0.026) | 1.158<br>(0.039) | 1.990<br>(0.040) | 1.945<br>(0.036) |
|                   |    | Mean BCI length | 0.385             | 0.379             | 0.367            | 0.365            | 0.414            | 0.434            | 0.482            |
|                   |    | Mean (std)      | -0.03<br>(0.018)  | -0.019<br>(0.018) | 0.072<br>(0.017) | 0.314<br>(0.017) | 1.169<br>(0.025) | 1.999<br>(0.026) | 1.942<br>(0.03)  |
|                   | 20 | Mean BCI length | 0.258             | 0.253             | 0.245            | 0.244            | 0.269            | 0.286            | 0.327            |
|                   |    | Mean (std)      | -0.032<br>(0.014) | -0.021<br>(0.012) | 0.067<br>(0.013) | 0.311<br>(0.015) | 1.165<br>(0.02)  | 1.998<br>(0.017) | 1.94<br>(0.021)  |
|                   | 40 | Mean BCI length | 0.183             | 0.179             | 0.175            | 0.175            | 0.194            | 0.203            | 0.23             |

Results are averages over 100 replicated data sets.

with this in mind. The marginal probability of selecting the causal site is 1 independently of the prior used and even when considering as few as ten studies, with almost no variability across replicates. Notably, all other markers have posterior inclusion probabilities close to zero and would therefore not be selected if we were to use the traditional threshold of 0.5. All conditional mean additive effects are very close to the true values with a minor bias only for the effect at SNP 7, which is the SNP in highest LD with the causal site. The choice of prior distribution on model space does not have a large effect on the results, with possibly narrower credible intervals and slightly larger posterior probability of including SNP 7 under prior c compared to priors a and b. This is to be expected because the Poisson priors favor models with few terms, whereas the uniform prior gives equal weight to all models. The tables also report the results for the variance terms  $\sigma_{\mu}^2$  and  $\sigma_{\nu}^2$ , which have posterior estimates close to the true values in both cases. Increasing the number of studies has little or no effect on both marginal posterior probabilities and posterior estimate bias but does lead to narrower credible intervals as expected.

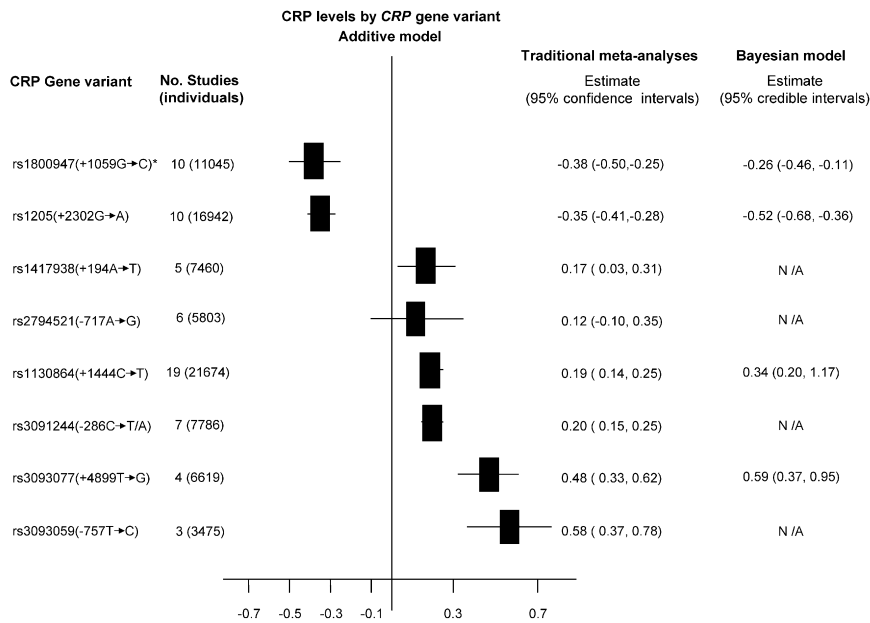
The univariate analyses on the other hand fail to unambiguously identify the causal site at position 6 (Table 3). On the basis of results reported therein, although SNP 6 shows the highest association with the phenotype, SNP 7 could still be considered causal if no prior information is available to discriminate between the two. Even markers 4 and 5 would be selected on the basis of posterior credible intervals; paradoxically, increasing the number of studies only exacerbates the problem because credible intervals become narrower.

The previous simulation study assumed the same LD pattern across studies because study-specific genotype data are simulated from a common haplotype pool. To mimic a more realistic scenario, we further considered study-specific LD patterns by simulating genotype counts from study-specific haplotype pools characterized by slightly different LD structures. The multilocus analysis then uses the average LD table shown in Figure D1 (in which we also report the standard deviations of the pairwise  $r^2$  values across studies in brackets). Results are reported in Table 4 for replicates with 20 studies. The method

**Table 4. Bayesian Multilocus Meta-Analysis when the LD Structure Is Allowed to Vary across Studies**

| Number of studies | Prior | Parameter  | $\beta_1$       | $\beta_2$       | $\beta_3$      | $\beta_4$       | $\beta_5$       | $\beta_6$      | $\beta_7$      | $\sigma_{\nu}^2$ | $\sigma_{\mu}^2$ |   |
|-------------------|-------|------------|-----------------|-----------------|----------------|-----------------|-----------------|----------------|----------------|------------------|------------------|---|
|                   |       | True       | 0               | 0               | 0              | 0               | 0               | 0              | 2              | 0                | 1                | 1 |
| 20                | a     | Post prob  | 0.01<br>(0.01)  | 0.01<br>(0.01)  | 0.03<br>(0.02) | 0.03<br>(0.06)  | 0.01<br>(0.01)  | 1.00<br>(0.00) | 0.03<br>(0.03) |                  |                  |   |
|                   |       | Mean       | -0.02<br>(0.04) | -0.02<br>(0.03) | 0.07<br>(0.01) | 0.04<br>(0.04)  | -0.01<br>(0.06) | 1.97<br>(0.01) | 0.13<br>(0.11) | 1.06<br>(0.02)   | 1.07<br>(0.23)   |   |
|                   |       | BCI length | 0.16            | 0.15            | 0.10           | 0.10            | 0.23            | 0.14           | 0.58           |                  |                  |   |
|                   | b     | Post prob  | 0.01<br>(0.01)  | 0.01<br>(0.01)  | 0.02<br>(0.01) | 0.01<br>(0.01)  | 0.02<br>(0.01)  | 1.00<br>(0.00) | 0.08<br>(0.05) |                  |                  |   |
|                   |       | Mean       | -0.01<br>(0.04) | 0.01<br>(0.04)  | 0.06<br>(0.02) | -0.02<br>(0.04) | -0.01<br>(0.07) | 1.98<br>(0.04) | 0.14<br>(0.28) | 1.08<br>(0.02)   | 1.03<br>(0.13)   |   |
|                   |       | BCI length | 0.2             | 0.12            | 0.15           | 0.15            | 0.32            | 0.29           | 0.67           |                  |                  |   |
|                   | c     | Post prob  | 0.03<br>(0.01)  | 0.03<br>(0.01)  | 0.11<br>(0.13) | 0.10<br>(0.08)  | 0.04<br>(0.05)  | 1.00<br>(0.00) | 0.12<br>(0.04) |                  |                  |   |
|                   |       | Mean       | -0.02<br>(0.03) | -0.02<br>(0.02) | 0.05<br>(0.03) | 0.04<br>(0.04)  | 0.01<br>(0.04)  | 1.99<br>(0.03) | 0.15<br>(0.05) | 1.04<br>(0.01)   | 1.07<br>(0.12)   |   |
|                   |       | BCI length | 0.17            | 0.15            | 0.12           | 0.10            | 0.22            | 0.24           | 0.58           |                  |                  |   |

Results are averages (std) over 100 replicated data sets. Mean posterior estimates and credible intervals are conditional on the SNP being included in a model. See Figure D1.



**Figure 4. Summary Effect from Traditional Meta-Analysis and Bayesian Multiple-SNP Hierarchical Linear Model of the Eight SNPs in the CRP Gene**

Values shown are additive genetic effects on (log) CRP levels with 95% confidence intervals or credible intervals for traditional and Bayesian analyses, respectively. For the Bayesian analysis, results are shown only for those markers that appear to be strongly associated after variable selection (see Figure 5). N/A refers to SNPs excluded from the model. The asterisk indicates the dominant model. Negative values indicate the variant allele is associated with a lower CRP concentration.

appears to be fairly robust to minor deviations in LD patterns across studies (similar to those observed for the real data in the next section); large differences in LD structures across studies would necessarily invalidate the meta-analytical approach because there would be little information to borrow for variable selection.

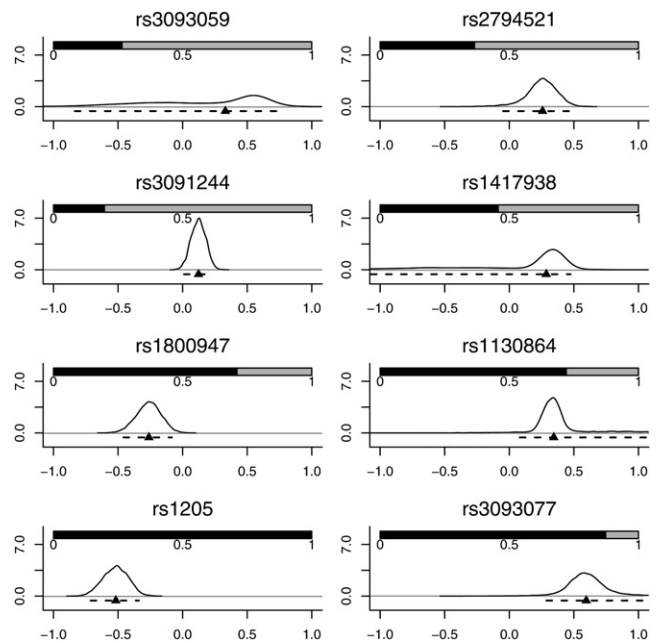
Finally, we considered reducing the effect at the causal site to 1.5 or placing it at marker position 2, which is in linkage equilibrium with the other sites: In both cases, the causal site is selected with high posterior probability ( $>0.8$ , results not shown).

The WinBUGS code used to fit the model is given in Appendix C.

### A Meta-Analysis of CRP Studies

The traditional single-locus meta-analyses require that the available data be partitioned into groups of studies in which the same SNP was typed directly. In these analyses, seven SNPs were associated with a codominant effect on CRP concentration (Figure 4) with the per-allele effect in the range of 0.19–0.58 mg/L (absolute p values: rs1800947 =  $4.35 \times 10^{-9}$ ; rs1205 =  $7.76 \times 10^{-26}$ ; rs1417938 =  $1.77 \times 10^{-2}$ ; rs1130864 =  $2.73 \times 10^{-11}$ ; rs3091244 =  $4.50 \times 10^{-15}$ ; rs3093077 =  $5.03 \times 10^{-11}$ ; and rs3093059 =  $2.27 \times 10^{-8}$ ), corresponding to  $\sim 0.3$ – $0.8$  SD of the population distribution of CRP 37. The main effect estimates were robust to analyses limited to studies of  $>500$  subjects (Table S2), providing strong evidence for an association at this locus. However, because pooled analyses of this type are limited to individual SNPs, it is unclear which of these SNPs have independent effects and which are associated because of correlation with other observed or unobserved SNPs, including the true causal variant(s). This can be overcome by incorporating available information on pairwise LD in the region (Table S3) within a Bayesian multilocus model as described

above. Bayesian model selection can then facilitate identification of variants showing the strongest independent association with CRP concentration (Figure 5 and Table 5). The approach yields posterior model probabilities



**Figure 5. Results from the Multiple-SNP Meta-Analysis using the Bayesian Hierarchical Linear Model**

The shaded bars show the posterior probability that each SNP is included in a model, calculated from the posterior sample of models. The x axis indicates the additive effects of each SNP on log CRP plasma levels, conditional on that SNP being included in the model, and the y axis indicates the corresponding posterior density. The curves can thus be interpreted as smoothed histograms representing the probability that the SNP effects take the values on the x axis. Also shown are the densities, medians ( $\blacktriangle$ ), and 95% credible intervals (---) for the additive effects of each SNP on log CRP levels.

**Table 5. Application to the Meta-Analysis of CRP Studies**

| Prob | SNP included |           |           |           |           |           |        |           |
|------|--------------|-----------|-----------|-----------|-----------|-----------|--------|-----------|
|      | rs3093059    | rs2794521 | rs3091244 | rs1417938 | rs1800947 | rs1130864 | rs1205 | rs3093077 |
| 0.22 |              |           |           |           | •         | •         | •      | •         |
| 0.12 |              |           |           |           | •         | •         | •      | •         |
| 0.10 |              |           | •         |           | •         |           | •      | •         |
| 0.07 |              |           |           |           | •         |           | •      | •         |
| 0.06 | •            |           |           |           | •         | •         | •      | •         |

Models with more than 2% posterior probability are shown. Results assume a Poisson(2) prior on model size in the reversible jump algorithm.

conditional on the observed data from which marginal probabilities of association for each SNP can be readily obtained. Of the markers considered, SNPs rs1130864, rs1205, and rs3093077, all in the 3 UTR, retain the strongest independent association with CRP concentration. An additional synonymous SNP in exon 2 (rs1800947) appears to be important, although its posterior probability of association is sensitive to the prior on the model space, and becomes unimportant if a more restrictive prior on the number of associated markers in the region is used (results not shown). These four SNPs yield the model with the highest posterior probability (Figures 4 and 5 and Table 5). Again, the models were not materially altered when analyses were limited to studies of >500 subjects (results not shown).

Notably, SNPs rs1130864, rs1205, and rs3093077 formed the trio of tag SNPs. Because each tag SNP marks a different haplotype, the Bayesian model implies the presence of at least three functional SNPs regulating CRP level (Figure 6). Using HapMap, we found that there were 11 SNPs in strong LD with rs1205, (five with pairwise  $r^2 = 1$ ) within an associated interval of ~100 kb. There were 11 SNPs in strong LD with rs3093077 (nine with pairwise  $r^2 = 1$ ), within a larger associated interval of ~300 kb. A total of 22 SNPs lay in an associated interval of 100 kb encompassing rs1130864 (nine with pairwise  $r^2 = 1$ ) (Figure 7). Because tightly linked SNPs were identified in the associated intervals,

a careful assessment of potential functionality for each of these SNPs is now required.

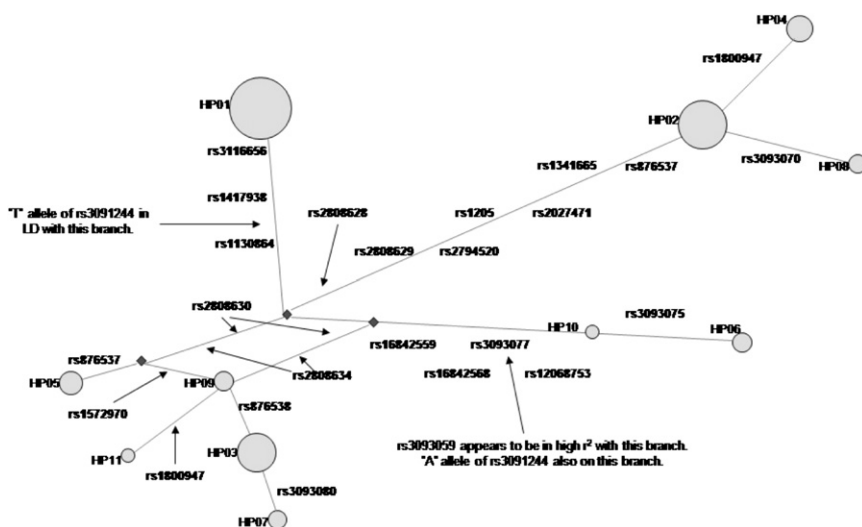
As mentioned in the previous section, in order to accommodate outliers and heavy tails, we assumed the distribution of the between-studies random effects  $\mu_s$  to be a mixture of normals. In particular, inspection of the residuals from a model fitted without the between-studies random effect appears to suggest the use of a two-component mixture, see Figure 8. The graph plots a sample of the quantities

$$r_j^s = T_j^s - \mathbf{W}_{j,\cdot}^{(t)} \beta^{(t)} - U_j^{(t)} \quad (7)$$

for current values of the spatial random effects  $U$  and model at iteration  $t$ .<sup>28</sup> The posterior distribution of  $\alpha_1$  and  $\alpha_2$  had means of  $-0.014$  and  $3.356$ , respectively (Figure 8), whereas  $\pi$  had posterior median estimate of  $0.879$ . By monitoring the mixture component assignments of each study, we found that outlying studies were mostly assigned to the second component as expected (results not shown).

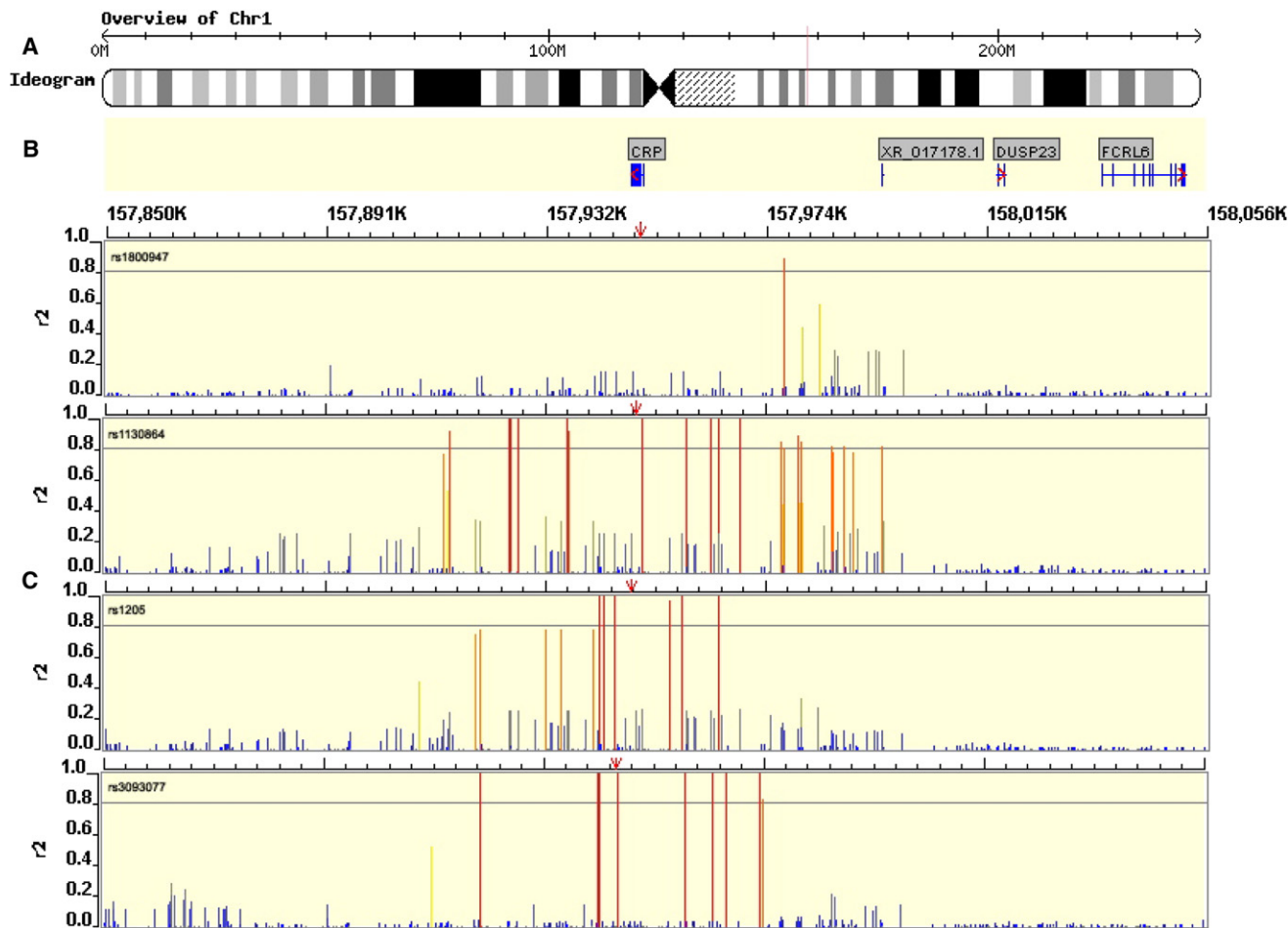
### Discussion

With only small genetic effects expected to contribute to most complex diseases, the meta-analysis of studies that consider variants in the same genetic region is a promising



**Figure 6. A Reduced Median Network Constructed with HapMap CEPH Data for a 20 kb Region Containing the CRP Gene** Yellow circles indicate haplotypes. The size of each circle is proportional to the frequency of that haplotype in the HapMap CEPH population. Non-HapMap SNPs (indicated in italics) were placed on the network with information from other CEPH populations.





**Figure 7. Genomic Context for *CRP* Gene**

(A) Ideogram depicting the chromosome and region in which the *CRP* gene lies (red line).  
 (B) Gene diagram with introns and exons depicted as horizontal and vertical blue lines, respectively.  
 (C) Pairwise  $r^2$  LD values between independently associating SNPs from Bayesian analysis (identified in top left of window, position indicated by red arrow) and all other HapMap SNPs in the region (release 20, build 35, red =  $r^2 > 0.8$ , yellow =  $0.5 < r^2 < 0.8$ , gray =  $0.3 < r^2 < 0.5$ , blue =  $0.2 < r^2 < 0.3$ , and dark gray = missing data).

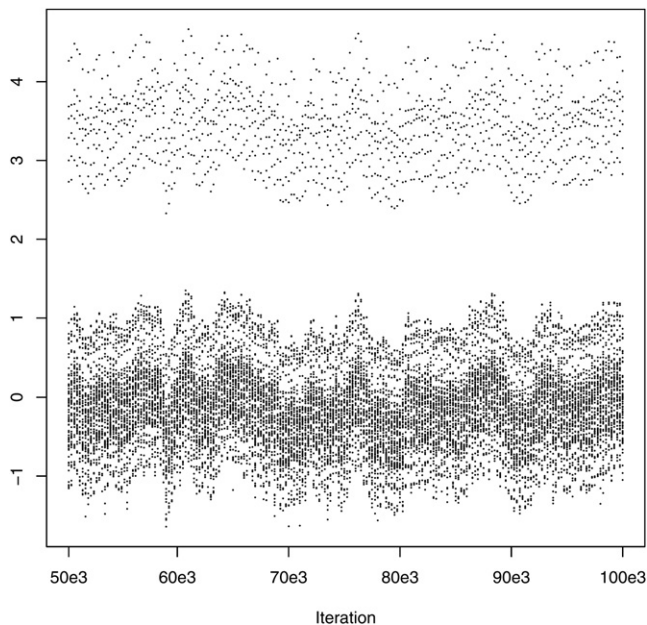
strategy to increase our chances of finding any associations. Recognizing the importance of this approach, several coordinated efforts have been initiated to ensure that results from the individual studies follow agreed guidelines and can be combined more easily.<sup>7</sup>

Most of the meta-analyses conducted so far have considered each marker in isolation, ignoring the possible correlation between markers due to linkage disequilibrium that reduces efficiency and that compromises the identification of any causal site. In this paper, we have presented a multi-marker approach that yields estimates of effect at each site adjusted for the effects of other variants, as in multiple regression. In both the simulation study and the application to the CRP data, we assumed an additive genetic model. Other choices are possible and would only involve changes in the entries of the matrices  $\mathbf{W}$  and  $\mathbf{X}'\mathbf{X}$ .

The methods borrow from the spatial data literature and incorporate the prior knowledge of marker pairwise LD in a fully Bayesian framework. For example, similar hierarchical models with spatial random effects are used extensively

in the analysis of spatial epidemiological data. A convenient feature of the joint specification (Equation [3]) is that it allows incorporation of the required correlation structure as prior information in an explicit way.<sup>13,37</sup> In addition, a reversible jump algorithm on the space of possible model structures enables the selection of the most promising associations. The proposed approach assumes data on a continuous phenotype. However, it could be extended to the case of discrete outcomes, say case-control status, by introducing a further set of continuous latent variables related to the discrete outcome as in probit regression. Extensions to include metaregression are straightforward and only involve introduction of a further hierarchy for the vector of coefficients  $\beta$  in Equation (4) with means that would then depend on study-specific covariates. Work on these extensions is currently in progress.

When applied to the meta-analysis of studies in the *CRP*-gene region, results provide evidence for three CRP modifying alleles distributed over three of four common haplotypes in Europeans. These alleles could account for the



**Figure 8. Posterior Sample of Residuals from the Hierarchical Model of Material and Methods Fitted without the Between-Study Random-Effect Term  $\mu_s$**

strong association with CRP of each of the three SNPs that are chosen for their ability to tag others and that mark the different haplotypes. The associated interval for each independently associating SNP extended at least 100 kb from either side of the open reading frame with a very sharp boundary of LD for at least two of these. Within each interval were a number of additional candidate causal SNPs in complete LD with the index SNP from the Bayesian analysis, any of which could, in theory, regulate CRP. Although the A and T alleles of the triallelic SNP rs3091244 appeared to exhibit functionality in previous reporter-gene studies in vitro,<sup>21</sup> this SNP was not retained within the Bayesian model. Experimental studies of this type may be biased toward the study of potential regulatory SNPs in the immediate vicinity against those located remotely from the gene of interest because of size constraints on reporter-gene constructs. This might explain why results of such reporter studies are, at times, discordant with the findings of association analyses in populations<sup>38</sup> or alternative experimental approaches to assessing functionality.<sup>39</sup> Irrespective of the true causal sites, the three tag SNPs adequately capture functional variation at this locus for large-scale gene-disease association studies. Although the naive expectation would be of narrower limits of error around the point estimates of SNP effects  $\mu_s$  with a Bayesian approach that includes all studies simultaneously, this was not observed. This is because unlike the traditional meta-analyses, the Bayesian analyses were corrected for the effect of other SNPs; that is, uncertainty about which SNPs are directly associated with the trait was properly incorporated in the analyses. However, the simultaneous use of all data strengthens evidence for an association at the gene level; the null model does not appear at all in the posterior

sample of models, reflecting virtual certainty of an effect on CRP at this gene.

Our approach facilitates the integration of data from studies that have genotyped different SNPs across the same gene or region utilizing prior information on LD. It has a number of favorable attributes and potential applications. By increasing the available data set of information on any SNP, the efficiency of evidence synthesis is enhanced and the reliability of any identified associations is increased. Further, the variable selection procedure allows inference on the relative magnitude of any marker-phenotype association and identifies those SNPs that show the strongest association with the phenotype, either because they are the functional site(s) or because they exhibit the strongest allelic association with (unobserved) functional sites. IPD (where available) can also be incorporated readily into the analysis because the regression parameters measuring the effect of variants retain the same interpretation when considering aggregate data (i.e., phenotype means by genotype groups as with CRP studies) or IPD (see [Material and Methods](#)). Moreover, where a robust evidence based on genetic association with a quantitative trait already exists (as it does for many blood measures, e.g., HDL cholesterol, triglycerides, and others), the methods described could be used to add and integrate partially overlapping SNP data from new genome-wide analyses, thereby harnessing existing data for both replication, and to gain insight into likely causal sites in a gene or region. The methods we describe, which use the freely available software WinBUGS, are likely to be of substantial value both to the emerging networks of investigators engaged in synthesis of evidence on genetic associations of complex quantitative traits and disorders<sup>7</sup> and to those applying and extending findings from genome-wide association studies.

## Appendix A

### Systematic Review

Two electronic databases (PubMed Medline and EMBASE) were searched with the text words, which were also MeSH terms, polymorphism(s), mutation(s), gene(s), genetic, variant(s), and SNP(s) in combination with C-reactive protein and CRP. The literature search was limited to human and to the English language. Any additional studies in the references of all identified publications were also searched. For inclusion, studies had to have an analytical design (case control, prospective, or cross sectional) and examine the association between any polymorphisms in the *CRP* gene and low-chronic CRP concentrations in individuals of European descent. Studies measuring *CRP* only during acute phase of an inflammatory response (e.g., acute ischemia or infection stimuli) were excluded. In areas where more than one polymorphism had been studied, information about the LD between them was extracted where available. If relevant information was not reported

(mean CRP levels, standard deviations, genotype numbers, or linkage disequilibrium data), or it was not reported stratified by ethnicity, the authors were contacted in several occasions to obtain the information. A total of four potential studies ( $n = 2614$ ) in European subjects were excluded because of unavailability of data in the appropriate form (Flex 2004,  $n = 471$ ; Obisesan 2004,  $n = 63$ ; Zee 2004,  $n = 260$ ; Carlson 2005,  $n = 1820$ ) (see Table S1).

### New Data Sets

NPHS II is a prospective study of 3012 healthy white European middle-aged men, of which a total of 2479 with CRP genetic data and CRP concentrations were included in this report. Recruitment in the study commenced in 1989<sup>40</sup> in nine general practices. None of the participants had a clinical history of unstable angina, myocardial infarction (including silent infarction), coronary surgery, other cardiovascular diseases, aspirin or anticoagulant use, or malignant disease (except skin cancer other than melanoma) at the time of recruitment. The Ely Study is a prospective population-based cohort study of the etiology and pathogenesis of type 2 diabetes and related metabolic disorders in 1122 individuals recruited in 1990 in Ely, Cambridge-shire.<sup>41</sup> Complete data on biochemical and anthropometric variables were available in 839 participants, and a total of 548 individuals with data on the CRP genotypes and CRP levels were included in this analysis. The EPIC-Norfolk study is a population-based cohort study, recruiting participants from general practices in Norfolk.<sup>42</sup> For the present report, only control participants from a nested case-control study in coronary heart disease were included, providing a total of 2196 participants with both data on CRP genetic variants and CRP concentrations.

### New Genotyping

Polymorphisms in the human CRP gene (HGNC: 2637; 1q21-q23) were identified by reference to public-domain databases of human sequence variation. We used this information to generate a consensus map of polymorphic sites. By using validated genotype data (minor allele frequency >5%) from subjects of European descent from the SeattleSNPs database and the human HapMap database (see Web Resources), we examined the pattern of linkage disequilibrium across the CRP gene. We then used the haplotype LD  $r^2$  method to select a set of tagging (t)SNPs capable of capturing maximum haplotype diversity among subjects of European descent by using the program TagIT (see Web Resources).

### LD

Public domain databases (see Web Resources) and individual publications were examined for information on the LD structure in the CRP gene. Both  $D'$  and  $r^2$  values were recorded, but  $r^2$  values were utilized in Bayesian modeling. If more than one  $r^2$  value for a given pairwise was reported, a weighted mean  $r^2$  was obtained.

## Appendix B

### Recovering the Joint Distribution of Multiallelic Sites from Allele Frequencies and Marginal Diallelic $r^2$ Values

We consider two loci, the first locus having  $G_1 =$  three alleles and the second locus having  $G_2 =$  two alleles. The joint probability of the haplotypes at these two loci can be represented in a  $3 \times 2$  table of the form:

**Table B1. Full  $3 \times 2$  Table of Haplotype Probabilities**

| Allele at Locus 1 | Allele at Locus 2       |   |                   |
|-------------------|-------------------------|---|-------------------|
|                   | 1                       | 2   |                   |
| 1                 | $p_{11}$                | $P_1 - p_{11}$                                | $P_1$             |
| 2                 | $p_{21}$                | $P_2 - p_{21}$                                | $P_2$             |
| 3                 | $Q_1 - p_{11} - p_{21}$ | $(1 - Q_1) - (P_1 - p_{11}) - (P_2 - p_{21})$ | $(1 - P_1 - P_2)$ |
|                   | $Q_1$                   | $(1 - Q_1)$                                   | 1                 |

where  $p_{ij}$  denotes the joint probability of allele  $i$  at locus 1 and allele  $j$  at locus 2,  $P_i$  denotes the probability of allele  $i$  at locus 1, and  $Q_1$  denotes the probability of allele 1 at locus 2.

The internal cells of this table are not observed. Our problem is to derive this table of probabilities on the basis of information from the margins of this table ( $P_1$ ,  $P_2$  and  $Q_1$ ) and pair-wise correlation within two marginal tables of the form:

**Table B2. First Marginal  $2 \times 2$  Haplotype Table**

| Allele at Locus 1 | Allele at Locus 2 |                                 |              |
|-------------------|-------------------|---------------------------------|--------------|
|                   | 1                 | 2                               |              |
| 1                 | $p'_{11}$         | $P'_1 - p'_{11}$                | $P'_1$       |
| 3                 | $Q'_1 - p'_{11}$  | $(1 - Q'_1) - (P'_1 - p'_{11})$ | $(1 - P'_1)$ |
|                   | $Q'_1$            | $(1 - Q'_1)$                    | 1            |

and

**Table B3. Second Marginal  $2 \times 2$  Haplotype Table**

| Allele at Locus 1 | Allele at Locus 2  |                                    |               |
|-------------------|--------------------|------------------------------------|---------------|
|                   | 1                  | 2                                  |               |
| 2                 | $p''_{21}$         | $(P''_2 - p''_{21})$               | $P''_2$       |
| 3                 | $Q''_1 - p''_{21}$ | $(1 - Q''_1) - (P''_2 - p''_{21})$ | $(1 - P''_2)$ |
|                   | $Q''_1$            | $(1 - Q''_1)$                      | 1             |

In the first of these tables,  $p'_{11}$  denotes the joint probability of allele 1 at locus 1 and allele 1 at locus 2, but now this probability is conditional upon the allele at locus 1 having either a 1 or 3 allele. Similarly,  $p_{21}$  denotes the probability of allele 2 at locus 1 and allele 1 at locus 2 conditional upon the allele at locus 1 being either a 2 or a 3.

We do not observe the two tables above but only the appropriate deviations from linkage disequilibrium,  $\delta'_{13}$  and  $\delta''_{23}$ , defined by

$$\delta'_{13} = P'_1 * Q'_1 - p'_{11}$$

and

$$\delta'_{23} = P'_2 * Q'_1 - p'_{21}.$$

We know that the probability of each pair-wise haplotype in Table 2 is equal to the corresponding probability of that pairwise haplotype in Table 1, divided by the probability that the allele at the first locus is either equal to 1 or 3,  $(1 - P_2)$ . This means that:

$$p'_{11} = \frac{p_{11}}{(1 - P_2)}$$

and

$$(Q'_1 - p'_{11}) = \frac{(Q_1 - p_{11} - p_{21})}{(1 - P_2)}.$$

Therefore,

$$\begin{aligned} Q'_1 &= \frac{(Q_1 - p_{11} - p_{21})}{(1 - P_2)} + \frac{p_{11}}{(1 - P_2)} \\ &= \frac{(Q_1 - p_{21})}{(1 - P_2)}, \end{aligned}$$

and

$$p_{11} = p'_{11}(1 - P_2).$$

Following a similar argument, we also find that

$$Q'_1 = \frac{(Q_1 - p_{11})}{(1 - P_1)}$$

and

$$p_{21} = p'_{21}(1 - P_1).$$

By writing  $p'_{11}$  and  $p'_{21}$  in terms of  $\delta'_{13}$  and  $\delta'_{23}$ , we find that:

$$\begin{aligned} p_{11} &= p'_{11}(1 - P_2) \\ &= (P'_1 * Q'_1 - \delta'_{11})(1 - P_2) \\ &= \left( \frac{P_1}{(1 - P_2)} \frac{(Q_1 - p_{21})}{(1 - P_2)} - \delta'_{11} \right) (1 - P_2) \\ &= \frac{1}{(1 - P_2)} (P_1 Q_1 - P_1 p_{21} - \delta'_{11} (1 - P_2)^2) \end{aligned}$$

and

$$\begin{aligned} p_{21} &= p'_{21}(1 - P_1) \\ &= (P'_2 * Q'_1 - \delta'_{21})(1 - P_1) \\ &= \left( \frac{P_2}{(1 - P_1)} \frac{(Q_1 - p_{11})}{(1 - P_1)} - \delta'_{21} \right) (1 - P_1) \\ &= \frac{1}{(1 - P_1)} (P_2 Q_1 - P_2 p_{11} - \delta'_{21} (1 - P_1)^2). \end{aligned}$$

This means that we have two equations in two unknowns,  $p_{11}$  and  $p_{21}$ , so that by substituting the second equation for  $p_{21}$  into the first equation for  $p_{11}$ , we can then solve this equation in terms of  $p_{11}$ . Substituting the expression for  $p_{12}$  into that for  $p_{11}$  gives:

$$\begin{aligned} p_{11} &= \frac{1}{(1 - P_2)} (P_1 Q_1 - P_1 p_{21} - \delta'_{11} (1 - P_2)^2) \\ &= \frac{1}{(1 - P_2)} \left( P_1 Q_1 - \frac{P_1}{(1 - P_1)} (P_2 Q_1 - P_2 p_{11} \right. \\ &\quad \left. - \delta'_{21} (1 - P_1)^2) - \delta'_{11} (1 - P_2)^2 \right), \end{aligned}$$

and rearranging in terms of  $p_{11}$  results in the equation:

$$p_{11} = P_1 Q_1 + \frac{(1 - P_1)}{(1 - P_1 - P_2)} (P_1 (1 - P_1) \delta'_{21} - (1 - P_2)^2 \delta'_{11}).$$

We may then write  $p_{21}$  in the form

$$p_{21} = P_2 Q_1 + \frac{(1 - P_2)}{(1 - P_1 - P_2)} (P_2 (1 - P_2) \delta'_{11} - (1 - P_1)^2 \delta'_{21}).$$

We are now able to calculate the probability of every cell of Table B1 in terms of  $p_{11}$ ,  $p_{12}$ ,  $P_1$ ,  $P_2$ , and  $Q_1$ .

Note that  $\delta'_{11}$  and  $\delta'_{21}$  can be obtained from the relevant  $r^2$  values with the formulae:

$$\delta'_{11} = (+/-) \sqrt{(r'_{11})^2 P'_1 (1 - P'_1) Q'_1 (1 - Q'_1)}$$

and

$$\delta'_{21} = (+/-) \sqrt{(r'_{21})^2 P'_2 (1 - P'_2) Q'_1 (1 - Q'_1)}.$$

Care must be taken when choosing which sign to assign these  $\delta$  values because they must be consistent with the margins of the Tables B2 and B3.

## Appendix C

### WinBUGS Code for the Model Described in Material and Methods

```

model {
  # likelihood
  for(j in 1:Q) {# where Q = sum_s (m_s + 1)
    T[j] ~dnorm(theta[j],tauy[j])
    tauy[j] <- tau.y/XsXs[j] # XsXs[j] in Equation (4)
    theta[j] <- psi[j]+sumXis[j]*mu[study[j]]+U[marker[j]] #
    linear predictor in Equation (4)
  }
  # pooled variances
  for(i in 1:L) {# where L = sum_s m_s
    scale[i] <- tau.y/2
    shape[i] <- pooled[i,2]/2
    pooled[i,1] ~dgamma(shape[i],scale[i]) # uses the gamma
    parameterization
  }
  #reversible jump part as detailed in Lunn et al.25
  psi[1:Q] <- jump.lin.pred(W[1:Q,1:m],K,tau.beta)
  id <- jump.model.id(psi[1:Q])
  pred[1:(m+1)] <- jump.lin.pred(psi[1:Q],X.pred[1:
  (m+1),1:m])
  for(i in 1:m){
    X.pred[i,i] <- 1
    for(j in 1:(i-1)) {X.pred[i,j] <- 0}
    for(j in (i+1):m) {X.pred[i,j] <- 0}
    X.pred[(m+1),i] <- 0
    effect[i] <- pred[i] -pred[m+1]
  }
  # mixture distribution for study effects
  for(s in 1:nstudies) {
    mu[s] ~dnorm(mumu[s],tau.mu)
  }
}

```

```

mumu[s] ← alpha[comp[s]]
comp[s] ~dcat(phi[])
}
phi[2] ← 1-phi[1]
alpha[2] ← (-phi[1]*alpha[1])/(1-phi[1])
# prior distributions
U[1:m] ~car.proper(thetaU[],M[],adj[],num[],m[],prec,1)
# thetaU vector of zeros of length m (number of unique
markers)
# M is weighted average of the XsXs matrices
# details on vectors adj, num and m are given in the
manual for GeoBUGS
prec ~dgamma(0.5,0.0005)
tau.y ~dgamma(0.001,0.001)
tau.mu ~dgamma(0.001,0.001)
tau.beta ~dgamma(0.0001,0.0001)
phi[1] ~dbeta(1,1)
alpha[1] ~dnorm(0.0,1.0E-6)
K ~dpois(1) # scenario (a)
}

```

The MCMC chain was run for 1,000,000 iterations with a burn-in of 500,000 and thinning of 100 iterations, which took ~30 min of CPU time on an Intel Xeon 2.80 GHz with 2 GB of RAM. Convergence was checked by visual inspection of posterior traces and by running chains with different initial values.<sup>36</sup>

## Appendix D

|          |   | Linkage Disequilibrium |                    |                     |                     |                     |                     |
|----------|---|------------------------|--------------------|---------------------|---------------------|---------------------|---------------------|
|          |   | 2                      | 3                  | 4                   | 5                   | 6                   | 7                   |
| Marker 1 | 1 | 0.0684<br>(0.0486)     | 0.0063<br>(0.0161) | 0.0051<br>(0.0143)  | 0.0100<br>(0.0137)  | 0.0089<br>(0.0167)  | -0.0054<br>(0.0147) |
|          | 2 |                        | 0.0287<br>(0.0559) | -0.0021<br>(0.0264) | 0.0090<br>(0.0142)  | 0.0149<br>(0.0144)  | 0.0013<br>(0.0151)  |
|          | 3 |                        |                    | -0.1594<br>(0.271)  | -0.0454<br>(0.1243) | -0.0266<br>(0.0935) | -0.0312<br>(0.0816) |
|          | 4 |                        |                    |                     | 0.4017<br>(0.2558)  | 0.2812<br>(0.1842)  | 0.2396<br>(0.1567)  |
|          | 5 |                        |                    |                     |                     | 0.7078<br>(0.1256)  | 0.6098<br>(0.1423)  |
|          | 6 | r                      |                    |                     |                     |                     | 0.8610<br>(0.1762)  |
|          |   | Marker 2               |                    |                     |                     |                     |                     |

**Figure D1.** Mean Pairwise LD Measures between Markers Used in the Simulation Study when Allowing LD Patterns to Vary across Studies

## Supplemental Data

Three tables are available at <http://www.ajhg.org/>.

## Acknowledgments

Dongliang Ge provided help with Figures 1 and 7. Tabular data were kindly provided by Jos E Krieger, Per Tornvall, and Moniek P. M. de Maat. This work was supported by Medical Research Council Research Grant G0600580. T.S. was supported by the British Heart Foundation (PhD Studentship FS/02/086/14760), R.S. was supported by a British Heart Foundation Shillingford Training Fellowship (FS/07/011), S.E.H. was supported by a British Heart Foundation Programme Grant (PG2000/015), A.D.H. was supported by a British Heart Foundation Senior Fellowship (FS/05/125), and L.S. was supported by a Wellcome Trust Senior Clinical fellowship (082178). A.D.H. acknowledges the generous support of the Rosetrees Trust. J.C. acknowledges the support of the Wellcome Trust (GR076024). C.J.V. is supported by a Research Council UK Fellowship. The EPIC-Norfolk study is supported by the Medical Research Council UK, Cancer Research UK, and Stroke Association and Research Into Ageing. The Framingham Heart Study is funded by N01-HC 25195, and Framingham inflammation research is funded by HL076784, AG028321.

Received: September 25, 2007

Revised: November 29, 2007

Accepted: January 22, 2008

Published online: April 3, 2008

## Web Resources

The URLs for data presented herein are as follows:

CRP:C-reactive protein, pentraxin-related, <http://pga.gs.washington.edu/data/crp/>

HapMap homepage, <http://www.hapmap.org/>

Online Mendelian Inheritance in Man (OMIM), <http://www.ncbi.nlm.nih.gov/Omim>

TagIT, <http://popgen.biol.ucl.ac.uk/software.html>

WinBUGS software, <http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml>

## References

- Cardon, L., and Bell, J. (2001). Association study designs for complex diseases. *Nat. Rev. Genet.* 2, 91–99.
- Colhoun, H., McKeigue, P., and Davey Smith, G. (2003). Problems of reporting genetic associations with complex outcomes. *Lancet* 361, 865–872.
- Clayton, D., and McKeigue, P. (2001). Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet* 358, 1356–1360.
- Zeggini, E., Rayner, W., Morris, A.P., Hattersley, A.T., Walker, M., Hitman, G.A., Deloukas, P., Cardon, L.R., and McCarthy, M.I. (2005). An evaluation of hapmap sample size and tagging snp performance in large-scale empirical and simulated data sets. *Nat. Genet.* 37, 1320–1322.
- Cambon-Thomsen, A. (2003). Assessing the impact of biobanks. *Nat. Genet.* 34, 25–26.
- Little, J., Bradley, L., Bray, M., Clyne, M., Dorman, J., Ellsworth, D., Hanson, J., Khoury, M., Lau, J., O'Brien, T., et al. (2002). Reporting, appraising, and integrating data on genotype prevalence and gene-disease associations. *Am. J. Epidemiol.* 156, 300–310.

7. Ioannidis, J.P., Gwinn, M., Little, J., Higgins, J.P., Bernstein, J.L., Boffetta, P., Bondy, M., Bray, M.S., Brenchley, P.E., Buffler, P.A., et al. (2006). A road map for efficient and reliable human genome epidemiology. *Nat. Genet.* 38, 3–5.
8. Seminara, D., Khoury, M., O'Brien, T., Manolio, T., Gwinn, M., Little, J., Higgins, J., Bernstein, J., Boffetta, P., Bondy, M., et al. (2007). The emergence of networks in human genome epidemiology: Challenges and opportunities. *Epidemiology* 18, 1–8.
9. Scott, L., Mohlke, K., Bonnycastle, L., Willer, C., Li, Y., Duren, W., Erdos, M., Stringham, H., Chines, P., Jackson, A., et al. (2007). A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. *Science* 316, 1341–1345.
10. Dina, C., Meyre, D., Gallina, S., Durand, E., Krner, A., Jacobson, P., Carlsson, L., Kiess, W., Vatin, V., Lecoeur, C., et al. (2007). Variation in fto contributes to childhood obesity and severe adult obesity. *Nat. Genet.* 39, 724–726.
11. The Wellcome Trust Case – Control Consortium (2007). Genome – wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447, 661–678.
12. Besag, J., and Kooperberg, C.L. (1995). On conditional and intrinsic autoregressions. *Biometrika* 82, 733–746.
13. Lawson, A.B. (2001). *Statistical Methods in Spatial Epidemiology* (Chichester, UK: John Wiley).
14. Hirschfield, G., and Pepys, M. (2003). C-reactive protein and cardiovascular disease: new insights from an old molecule. *QJM* 96, 793–807.
15. Danesh, J., Wheeler, J., Hirschfield, G., Eda, S., Eiriksdottir, G., Rumley, A., Lowe, G., Pepys, M., and Gudnason, V. (2004). C-reactive protein and other circulating markers of inflammation in the prediction of coronary heart disease. *N. Engl. J. Med.* 350, 1387–1397.
16. Hingorani, A., and Humphries, S. (2005). Nature's randomised trials. *Lancet* 366, 1906–1908.
17. Hingorani, A., Shah, T., and Casas, J. (2006). Linking observational and genetic approaches to determine the role of c-reactive protein in heart disease risk. *Eur. Heart J.* 27, 1261–1263.
18. DaveySmith, G., and Ebrahim, S. (2003). Mendelian randomization: Can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* 32, 1–22.
19. Davey Smith, G., and Ebrahim, S. (2004). Mendelian randomization: Prospects, potentials and limitations. *Int. J. Epidemiol.* 33, 30–42.
20. MacGregor, A., Gallimore, J., Spector, T., and Pepys, M. (2004). Genetic effects on baseline values of c-reactive protein and serum amyloid a protein: A comparison of monozygotic and dizygotic twins. *Clin. Chem.* 50, 130–134.
21. Carlson, C., Aldred, S., Lee, P., Tracy, R., Schwartz, S., Rieder, M., Liu, K., Williams, O., Iribarren, C., Lewis, E., et al. (2005). Polymorphisms within the c-reactive protein (crp) promoter region are associated with plasma crp levels. *Am. J. Hum. Genet.* 77, 64–77.
22. Kardys, I., de Maat, M., Uitterlinden, A., Hofman, A., and Witteman, J. (2006). C-reactive protein gene haplotypes and risk of coronary heart disease: The rotterdam study. *Eur. Heart J.* 27, 1331–1337.
23. Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* 39, 906–913.
24. Green, P. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika* 82, 711–732.
25. Lunn, D., Whittaker, J., and Best, N. (2006). A bayesian toolkit for genetic association studies. *Genet. Epidemiol.* 30, 231–247.
26. Thomas, D. (2004). *Statistical Methods in Genetic Epidemiology* (New York: Oxford University Press).
27. Besag, J., York, J., and Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Ann. Inst. Stat. Math.* 43, 1–59.
28. Dominici, F., Parmigiani, G., Wolpert, R.L., and Hasselblad, V. (1999). Meta-analysis of migraine headache treatments: Combining information from heterogeneous designs. *J. Am. Stat. Assoc.* 94, 16–28.
29. Kelsall, J., and Wakefield, J.C. (1999). Comment on Bayesian models for spatially correlated disease and exposure data. In *Bayesian Statistics 6 Proceedings of the Sixth Valencia International Meeting* (Oxford: Clarendon Press).
30. Denison, D.G.T., Holmes, C.C., Mallick, B.K., and Smith, A.F.M. (2002). *Bayesian Methods for Nonlinear Classification and Regression* (Chichester, UK: John Wiley).
31. Verzilli, C.J., Whittaker, J.C., and Stallard, N. (2005). A hierarchical bayesian model for predicting the functional consequences of amino acid polymorphisms. *J. R. Stat. Soc. Ser. C Appl. Stat.* 54, 191–207.
32. Higgins, J., Thompson, S., Deeks, J., and Altman, D. (2003). Measuring inconsistency in meta-analyses. *BMJ* 327, 557–560.
33. DuMouchel, W.H. (1990). Bayesian meta-analysis. In *Statistical Methodology in the Pharmaceutical Sciences* (New York: Marcel Dekker).
34. Normand, S.L.T. (1999). Meta-analysis: Formulating, evaluating, combining and reporting. *Stat. Med.* 18, 321–359.
35. O'Hagan, A., and Forster, J. (2004). Bayesian Inference. In *Kendall's Advance Theory of Statistics, Vol 2B* (London: Arnold).
36. Carlin, B.P., and Louis, T.A. (2002). *Bayes and Empirical Bayes Methods for Data Analysis* (Boca Raton: Chapman & Hall).
37. Parent, O., and Riou, S. (2005). Bayesian analysis of knowledge spillovers in european regions. *J. Reg. Sci.* 45, 747–775.
38. Ioannidis, J., and Kavvoura, F. (2006). Concordance of functional in vitro data and epidemiological associations in complex disease genetics. *Genet. Med.* 8, 583–593.
39. Cirulli, E., and Goldstein, D. (2007). In vitro assays fail to predict in vivo effects of regulatory polymorphisms. *Hum. Mol. Genet.* 16, 1931–1939.
40. Herbert, A., Lenburg, M., Ulrich, D., Gerry, N., Schlauch, K., and Christman, M. (2007). Open-access database of candidate associations from a genome-wide snp scan of the framingham heart study. *Nat. Genet.* 39, 135–136.
41. Wareham, N., Hennings, S., Prentice, A., and Day, N. (1997). Feasibility of heart-rate monitoring to estimate total level and pattern of energy expenditure in a population-based epidemiological study: The ely young cohort feasibility study. *Br. J. Nutr.* 78, 889–900.
42. Day, N., Oakes, S., Luben, R., Khaw, K., Bingham, S., Welch, A., and Wareham, N. (1999). Epic-norfolk: Study design and characteristics of the cohort. european prospective investigation of cancer. *Br. J. Cancer* 80, 95–103.