A Bayesian Partitioning Model for Detection of Multilocus Effects in Case-Control Studies

Debashree Ray, Xiang Li, Wei Pan, James S Pankow, Saonli Basu

Supplementary Materials

1 Details of dimension moves of the RJMCMC

To implement our Bayesian Partitioning Model (BPM) approach, we constructed a Markov Chain using reversible jump (RJMCMC) with "dimension" moves, and "allocation" & "coefficient" moves within a fixed dimension. The "dimension" moves include 'death' and 'birth' steps to increase or decrease the dimension by one. The dimension parameter K can take 4 values : 0, 1, 2, and 3, which refers to the case that the model has parameter(s) α; α and β_1; α and β_2; and all three parameters α, β_1, β_2 in our logistic regression model

$$\log\left(\frac{P(y_i=1|\boldsymbol{X}_i,\boldsymbol{\mathcal{A}})}{1-P(y_i=1|\boldsymbol{X}_i,\boldsymbol{\mathcal{A}})}\middle|\alpha,\beta_1,\beta_2\right)=\alpha+\beta_1 Z_{1i}+\beta_2 Z_{2i}, \tag{1}$$

where $i=1,2,...,n$, $\beta_1<0$ and $\beta_2>0$ respectively defines the fixed effects of the LR and the HR group of SNPs, $\boldsymbol{\mathcal{A}}$ is the $p\times 3$ allocation matrix of the p SNPs, and the predictors Z_{1i}, Z_{2i} are respectively the values of scores for the LR and HR groups of a specific individual i $(i=1,2,...,n)$. It is to be noted that a dimension move from $K^{(t-1)}=1$ to $K^{(t)}=3$ involves increasing the dimension by 1 although the dimension parameter K is increased by 2.

The acceptance probability for the dimension moves (from step $\overline{t-1}$ to step t) is $min\left(1, a(K^{(t-1)},K^{(t)})\right)$, where $a\left(K^{(t-1)},K^{(t)}\right)$ is given by

$$\begin{aligned}a(K^{(t-1)},K^{(t)})=&\frac{P[K^{(t)}].P[\boldsymbol{\beta}^{(t)},\boldsymbol{\mathcal{A}}^{(t)}|K^{(t)}].P[\boldsymbol{y}|\boldsymbol{\mathcal{A}}^{(t)},\boldsymbol{\beta}^{(t)},K^{(t)}].P[K^{(t-1)}|K^{(t)}]}{P[K^{(t-1)}].P[\boldsymbol{\beta}^{(t-1)},\boldsymbol{\mathcal{A}}^{(t-1)}|K^{(t-1)}].P[\boldsymbol{y}|\boldsymbol{\mathcal{A}}^{(t-1)},\boldsymbol{\beta}^{(t-1)},K^{(t-1)}].P[K^{(t)}|K^{(t-1)}]}\\&\times\frac{Q[D^{(t-1)}|D^{(t)}]}{Q[D^{(t)}|D^{(t-1)}]}\times|1|\end{aligned} \tag{2}$$

Let us now look at the possible forms of proposal $Q[D^{(t)}|D^{(t-1)}]$ depending upon the move from $K^{(t-1)}$ to $K^{(t)}$. Suppose $K^{(t-1)}=0$ and $K^{(t)}=1$. The difference between $\boldsymbol{\beta}^{(t-1)}$

and $\boldsymbol{\beta}^{(t)}$ is due to the extra random component β_1. The proposal density $Q[D^{(t)}|D^{(t-1)}]$ is then the density from which β_1 is drawn, which in this case is a truncated standard normal density $N(0,1) \times I(\beta_1 < 0)$. The other proposal density $Q[D^{(t-1)}|D^{(t)}]$ is just 1 since the move from $\boldsymbol{\beta}^{(t)} = (\alpha, \beta_1)$ to $\beta^{(t-1)} = \alpha$ requires dropping the parameter β_1. For β_2, the truncated density $N(0,1) \times I(\beta_2 > 0)$ is used. The possible moves along with the corresponding acceptance probabilities are listed below.

Case 1 : $(K^{(t-1)}, K^{(t)}) = (0,1)$ or $(2,3)$: β_1 is added with increase in dimension parameter. Thus, $a(K^{(t-1)}, K^{(t)})$ equals

$$\frac{P[\boldsymbol{\beta}^{(t)}|\boldsymbol{\mathcal{A}}^{(t)}, K^{(t)}]P[\boldsymbol{\mathcal{A}}^{(t)}|K^{(t)}]P[\boldsymbol{y}|\boldsymbol{\mathcal{A}}^{(t)}, \boldsymbol{\beta}^{(t)}, K^{(t)}]}{P[\boldsymbol{\beta}^{(t-1)}|\boldsymbol{\mathcal{A}}^{(t-1)}, K^{(t-1)}]P[\boldsymbol{\mathcal{A}}^{(t-1)}|K^{(t-1)}]P[\boldsymbol{y}|\boldsymbol{\mathcal{A}}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, K^{(t-1)}]} \times \frac{1}{N(0,1) \times I(\beta_1 < 0)}$$

Case 2 : $(K^{(t-1)}, K^{(t)}) = (0,2)$ or $(1,3)$: β_2 is the extra parameter, so $a(K^{(t-1)}, K^{(t)})$ equals

$$\frac{P[\boldsymbol{\beta}^{(t)}|\boldsymbol{\mathcal{A}}^{(t)}, K^{(t)}]P[\boldsymbol{\mathcal{A}}^{(t)}|K^{(t)}]P[\boldsymbol{y}|\boldsymbol{\mathcal{A}}^{(t)}, \boldsymbol{\beta}^{(t)}, K^{(t)}]}{P[\boldsymbol{\beta}^{(t-1)}|\boldsymbol{\mathcal{A}}^{(t-1)}, K^{(t-1)}]P[\boldsymbol{\mathcal{A}}^{(t-1)}|K^{(t-1)}]P[\boldsymbol{y}|\boldsymbol{\mathcal{A}}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, K^{(t-1)}]} \times \frac{1}{N(0,1) \times I(\beta_2 > 0)}$$

Case 3 : $(K^{(t-1)}, K^{(t)}) = (1,0)$ or $(3,2)$: Since β_1 is dropped, $a(K^{(t-1)}, K^{(t)})$ equals

$$\frac{P[\boldsymbol{\beta}^{(t)}|\boldsymbol{\mathcal{A}}^{(t)}, K^{(t)}]P[\boldsymbol{\mathcal{A}}^{(t)}|K^{(t)}]P[\boldsymbol{y}|\boldsymbol{\mathcal{A}}^{(t)}, \boldsymbol{\beta}^{(t)}, K^{(t)}]}{P[\boldsymbol{\beta}^{(t-1)}|\boldsymbol{\mathcal{A}}^{(t-1)}, K^{(t-1)}]P[\boldsymbol{\mathcal{A}}^{(t-1)}|K^{(t-1)}]P[\boldsymbol{y}|\boldsymbol{\mathcal{A}}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, K^{(t-1)}]} \times \Big(N(0,1) \times I(\beta_1 < 0)\Big)$$

Case 4 : $(K^{(t-1)}, K^{(t)}) = (2,0)$ or $(3,1)$: Since β_2 is dropped, then $a(K^{(t-1)}, K^{(t)})$ equals

$$\frac{P[\boldsymbol{\beta}^{(t)}|\boldsymbol{\mathcal{A}}^{(t)}, K^{(t)}]P[\boldsymbol{\mathcal{A}}^{(t)}|K^{(t)}]P[\boldsymbol{y}|\boldsymbol{\mathcal{A}}^{(t)}, \boldsymbol{\beta}^{(t)}, K^{(t)}]}{P[\boldsymbol{\beta}^{(t-1)}|\boldsymbol{\mathcal{A}}^{(t-1)}, K^{(t-1)}]P[\boldsymbol{\mathcal{A}}^{(t-1)}|K^{(t-1)}]P[\boldsymbol{y}|\boldsymbol{\mathcal{A}}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, K^{(t-1)}]} \times \Big(N(0,1) \times I(\beta_2 > 0)\Big)$$

2 Implementation of the RJMCMC

To implement our RJMCMC steps, we start with a starting allocation $\boldsymbol{\mathcal{A}}^{(0)}$ at dimension $K = 3$ and a starting $\boldsymbol{\beta}^{(0)}$. A random choice is made about whether to stay in the same dimension ($K = 3$) or reduce the dimension by 1 ($K = 1$ or $K = 2$). If a move is made from $K^{(0)} = 3$ to $K^{(1)} = 2$, β_1 is dropped, else β_2 is dropped. If β_1 is dropped, there should be only HR SNPs or HR and NA SNPs. So, all the LR SNPs of $\boldsymbol{\mathcal{A}}^{(0)}$ are assigned to the NA group in the candidate allocation $\boldsymbol{\mathcal{A}}^{*}$. If the dimension move selects staying in the same dimension, a SNP is then randomly selected. Using the probabilities

$$p_{js}^{(t)} = \frac{P[\boldsymbol{y}|\boldsymbol{\beta}^{(t-1)}, \boldsymbol{\mathcal{A}}_{(-j)}^{(t-1)}, \boldsymbol{\mathcal{A}}_{j}^{(t)} = \boldsymbol{a}_s] P[\boldsymbol{\mathcal{A}}_{j}^{(t)} = \boldsymbol{a}_s]}{\sum_{k=1}^{3} P[\boldsymbol{y}|\boldsymbol{\beta}^{(t-1)}, \boldsymbol{\mathcal{A}}_{(-j)}^{(t-1)}, \boldsymbol{\mathcal{A}}_{j}^{(t)} = \boldsymbol{a}_k] P[\boldsymbol{\mathcal{A}}_{j}^{(t)} = \boldsymbol{a}_k]}, \tag{3}$$

where $s = 1, 2, 3$, $\boldsymbol{a}_s \in \{(1,0,0)', (0,1,0)', (0,0,1)'\}$, a new allocation is chosen for the selected SNP, which gives us a candidate $\boldsymbol{\mathcal{A}}^{*}$. Under this allocation $\boldsymbol{\mathcal{A}}^{*}$, $\boldsymbol{\beta}$ is updated using Metropolis-Hastings (MH) Algorithm. Within a particular $\boldsymbol{\mathcal{A}}^{*}$, we iteratively updated $\boldsymbol{\beta}$ 10 times via the MH algorithm. We thus obtain a candidate $\boldsymbol{\beta}^{*}$ for a candidate allocation $\boldsymbol{\mathcal{A}}^{*}$. Using MH Algorithm, we then decide whether the parameters should be updated ($\boldsymbol{\mathcal{A}}^{(1)} = \boldsymbol{\mathcal{A}}^{*}$, $\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{*}$) or remain the same ($\boldsymbol{\mathcal{A}}^{(1)} = \boldsymbol{\mathcal{A}}^{(0)}$, $\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(0)}$). In our simulation studies and the real data analysis, we repeated the above process $10,000$ and $500,000$ times respectively to generate MCMC samples from the posterior distribution $P[\boldsymbol{\mathcal{A}}, \boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}]$ and estimated the marginal posterior probabilities of LR, HR and NA group for each SNP by averaging over the different $\boldsymbol{\beta}$ values for each risk allocation $\boldsymbol{\mathcal{A}}$. It is to be noted that $10,000$ MCMC iterations really correspond to $100,000$ iterations in our case since within each iteration, $\boldsymbol{\beta}$ is iterated 10 times.

3 Comparison of M-score and P-score of BPM

To illustrate the advantage of the proposed pair-wise-score modeling (P-score) over the main effect modeling (M-score) in presence of interaction, we did a three-loci simulation study. We simulated data on 3 loci with epistatic interactions in 1000 cases and 1000 controls using the following logistic regression model:

$$\text{logit(p)} = -5 + X_1 + 0.5X_3 + 3X_1X_2X_3,$$

where X_j is the number of minor alleles for j-th SNP, $j = 1, 2, 3$. The simulation was repeated 200 times, with 10,000 MCMC iterations in each simulation. For each MCMC iteration of a simulated data, we sampled a risk-allocation $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{\beta}$ and implemented our pair-wise scoring (P-score) algorithm and M-score. We then used our MCMC scheme to move between the different risk-allocations and calculated the average posterior probability of each SNP belonging to LR, HR or NA category from 5,000 sampled risk-allocations (first 5,000 iterations were discarded as burn-in). So, for example, for a given score and a given dataset, the posterior probability of SNP1 to belong to LR category is the proportion of times SNP1 is allocated to the LR group in the 5,000 non-discarded MCMC iterations.

The estimated posterior probabilities of different risk-allocations of 3 SNPs for each simulated dataset were summarized in Figure 1. For Figure 1, we ignored the distinctiveness of the 3 SNPs and only considered 10 possible categories 'iH-jL-kN', where i SNPs are categorized as HR (high-risk or *bad*), j as LR (low-risk or *good*) and k as NA (nost-associated or *null*), $i + j + k = 3$. For a given score, the corresponding plot in Figure 1 has 200 points (corresponding to 200 datasets) at each of the 10 categories. For example, in case of M-score, all 200 points at 0 for category 0H-0L-3N means in each of the 200 datasets, none of the non-discarded iterations had an allocation where all 3 SNPs were categorized as NA. On the other hand, for category 2H-0L-1N, the 200 points were distributed between 0 and 1; each point corresponds to the proportion of times 2 SNPs were allocated in HR and 1 SNP in NA group for that particular dataset. In our simu-

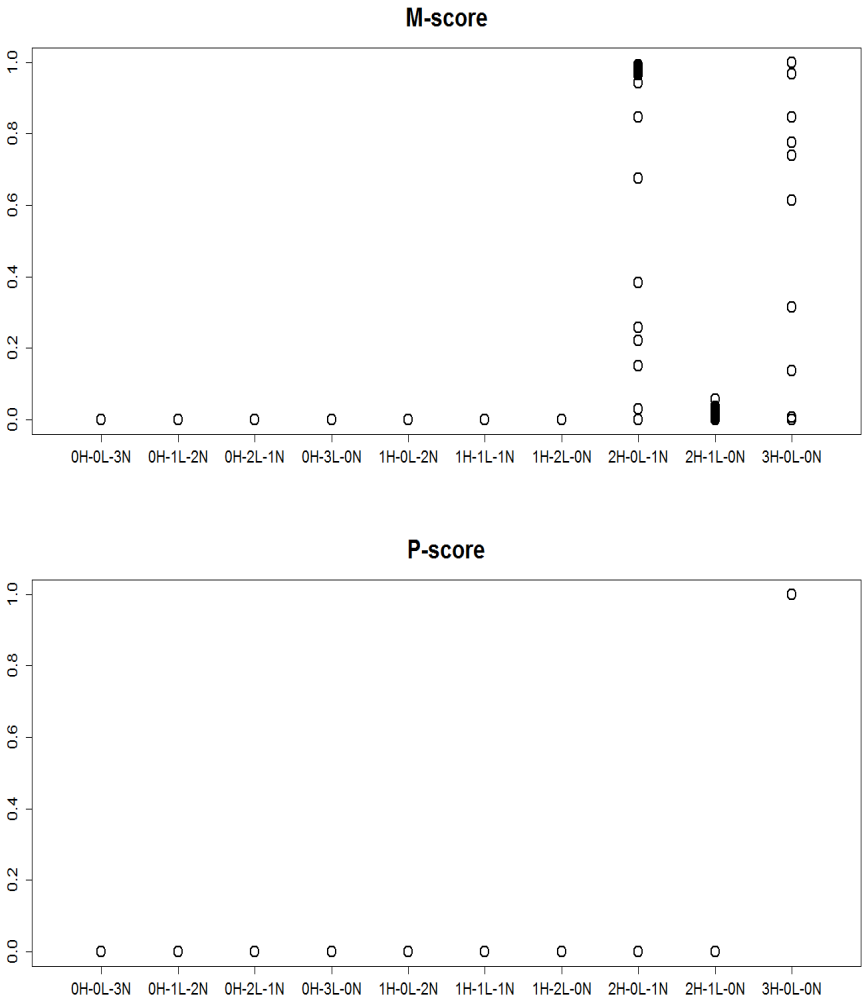


Figure 1: Figure shows the proportion of times each of the 10 categories 'iH-jL-kN' (where i SNPs in HR, j in LR and k in NA groups, $i + j + k = 3$) are observed in each of the 200 simulated datasets. For M-score, although all iterations of 154 datasets allocate the SNPs to 3H-0L-0N category, 38 out of 200 datasets did not have a single 3H-0L-0N allocation among 5000 iterations. On the other hand, for P-score, all iterations of all datasets had 3H-0L-0N allocation.

lated data, two loci (SNP1 and SNP3) had main effects, and the third one (SNP2) only entered the model through a three-way interaction with the other two SNPs. According to Figure 1, SNP2 was mis-categorized as LR or NA SNP by M-score in many datasets, whereas P-score categorized all 3 SNPs as HR with probability 1.

Table 1 compares the marginal posterior probabilities of each SNP being categorized as LR, NA or HR by the two scoring algorithms. The estimated proportions in Table 1 are based on the allocations of the three SNPs in 5000×200 iterations (5000 MCMC iterations of each of 200 datasets). When the BPM M-score approach (equivalent to

Table 1: Estimated marginal posterior probabilities of categorizing each of the 3 SNPs by the 2 scoring algorithms of BPM. Here all 3 SNPs are HR; SNP1 and SNP3 contribute through main effect while SNP2 enters the model through a 3-way interaction with the other 2 SNPs. M-score captures the true risk of SNP2 only 79% of the times while P-score captures it all the time.

	M-score			P-score		
Locus	LR	NA	HR	LR	NA	HR
SNP1	0	0	1	0	0	1
SNP2	0.00447	0.20353	0.792	0	0	1
SNP3	0	0	1	0	0	1

main effect logistic regression) was used, the estimated posterior probability of correctly allocating all three loci into the 'high-risk' group was at most 0.792, where SNP1 and SNP3 were correctly allocated to HR group with probability 1. When the pair-wise scoring algorithm was used, the estimated posterior probability of correctly allocating all three loci into the high-risk group was 1. This simulation study demonstrates the increase in power to detect a three-way interaction using the P-score as compared to the M-score for our BPM approach.

4 Simulation 1: Comparison of BPM, BEAM and LKM

For our simulation study, we compared the power of our BPM approach, BEAM (Bayesian Epistasis Association Mapping) and LKM (Logistic Kernel Machine) methods to detect multilocus association.

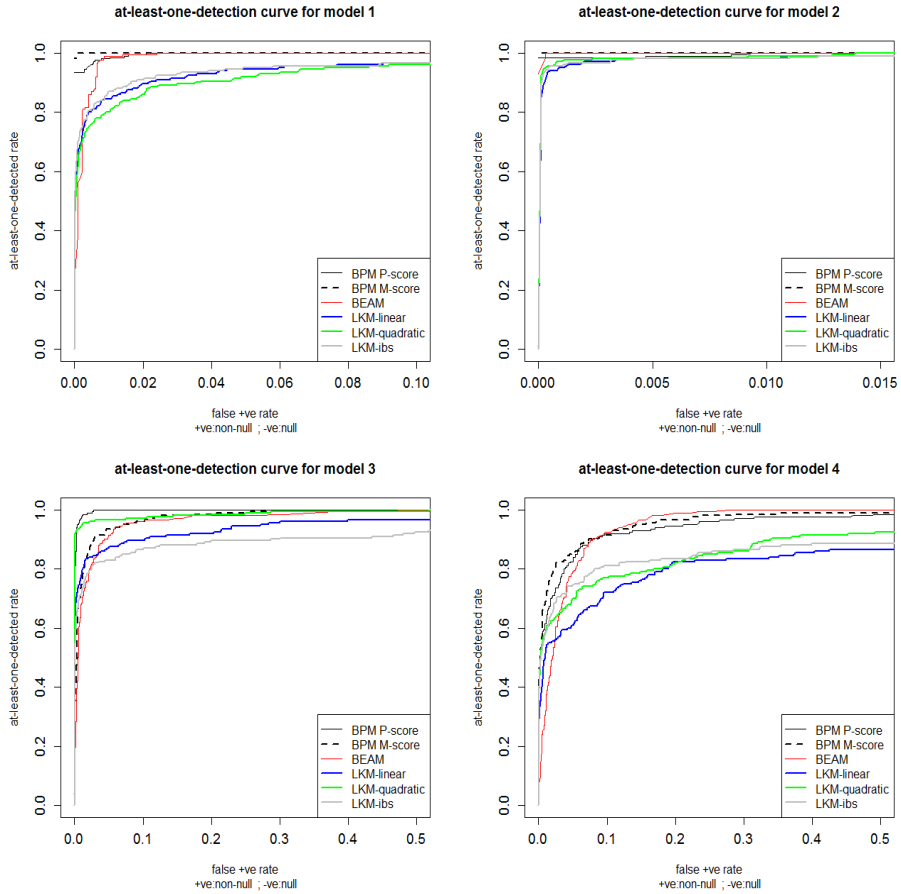


Figure 2: The power curves for methods BPM (M- and P-scores), BEAM and LKM (linear, quadratic and ibs kernels) for detecting at least one causal SNP in a set of 20 independent SNPs, out of which first 4 were causal and 16 were null. The power was calculated as the proportion of times (out of the 200 simulated datasets) at least one associated/causal SNP was detected. The heavy black curve is for BPM M-score, light black for P-score, red for BEAM, blue for LKM linear kernel, green for LKM quadratic kernel and gray for LKM ibs kernel.

5 Simulation 2: Details

Here we performed a simulation study with correlated SNPs to see the impact of linkage disequilibrium (LD) on our BPM approach.

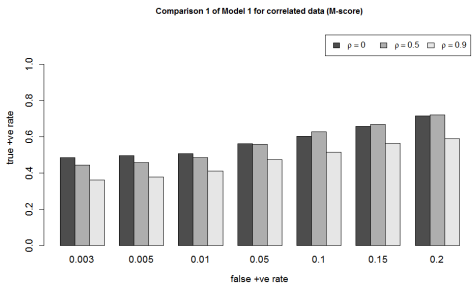
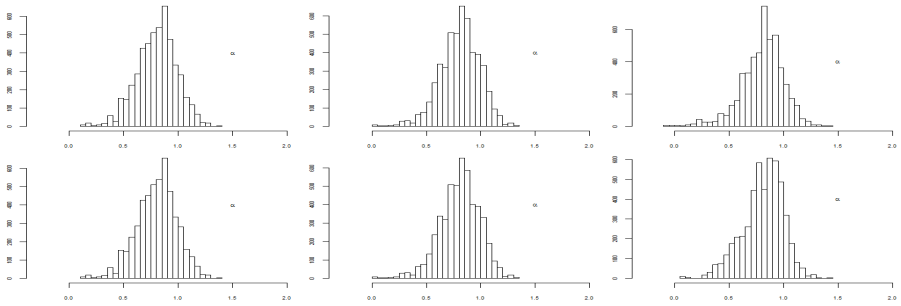


Figure 3: The barplot shows the power of BPM M-score approach for various type-1 error levels. We simulated data on 20 SNPs. The 4 causal SNPs and the rest 16 null SNPs had AR1(ρ) correlation structures. The causal SNP set was not correlated with the non-null SNP set. 3 different values of correlation parameter were taken: $\rho = 0, 0.5, 0.9$. Moderate correlation among SNPs do not seem to affect the performance of BPM much but for high SNP-SNP correlation, BPM may have low power for low error levels.

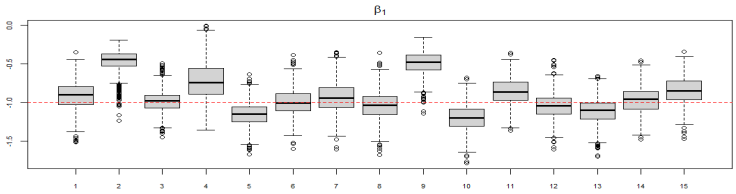
6 Convergence checks for Simulation study

6.1 Convergence of $\boldsymbol{\beta}$ parameter

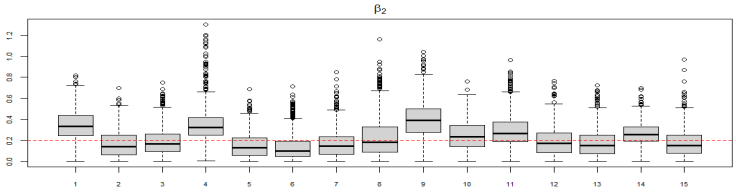
For our BPM approach, convergence of the RJMCMC was checked using only the common parameter α since the parameters change as the RJMCMC moves from one dimension to another (Sisson, 2005).



(a) Posterior Distribution of α for 6 independent chains for a randomly chosen dataset



(b) Boxplots for posterior β_1 (LR group) for 15 independent datasets



(c) Boxplots for posterior β_2 (HR group) for 15 independent datasets

Figure 4: (a) The posterior distribution of the common parameter α for 6 independent chains under Model 1 with independent-SNP data using BPM M-score approach. For this purpose, a dataset was chosen randomly from the 200 simulated datasets. For all the 6 chains, the starting $\boldsymbol{\beta}$ and $\boldsymbol{\mathcal{A}}$ parameters were different. Each chain had $10,000$ MCMC iterations, where the first $5,000$ were discarded; (b) and (c) Boxplots of posterior β_1 and β_2 values using BPM M-score for 15 randomly chosen datasets (out of 200) of Model 1. The broken red horizontal lines denote the true effect sizes for the respective risk group (-1 for LR and 0.2 for HR group of SNPs).

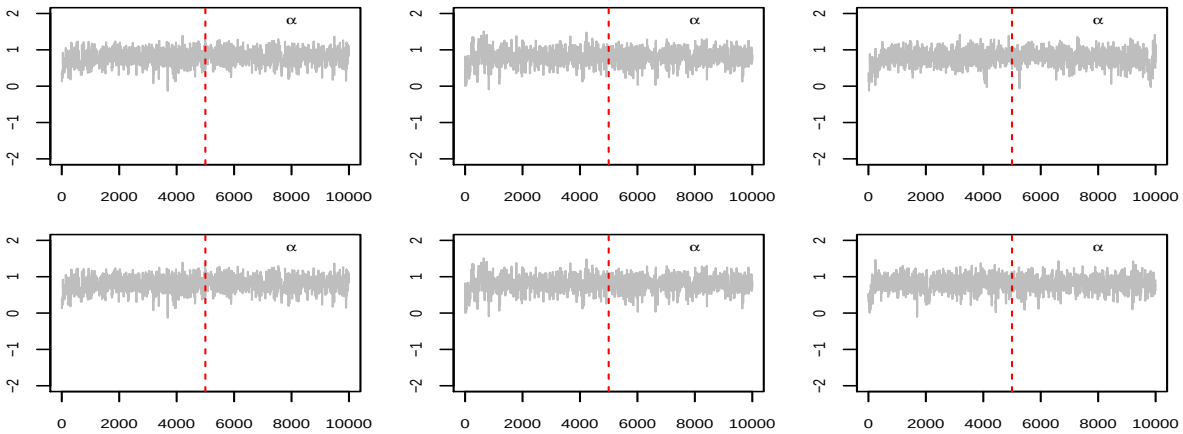
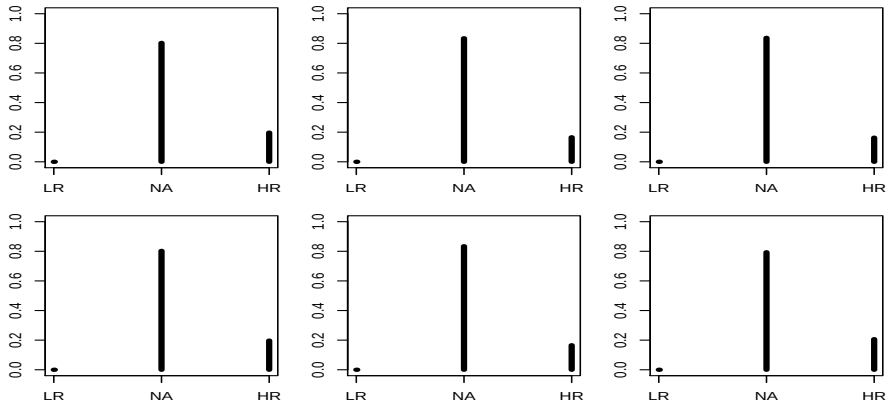


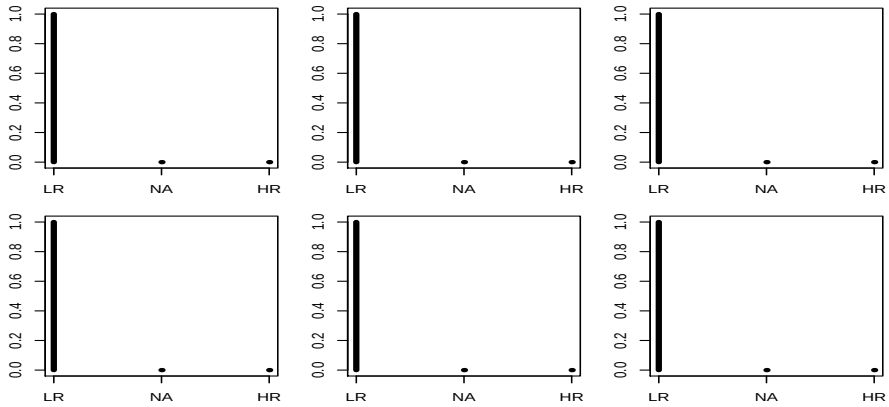
Figure 5: The trace plots of the common parameter α for 6 independent chains under Model 1 with independent-SNP data using BPM M-score approach. For this purpose, a dataset was chosen randomly from the 200 simulated datasets. For all the 6 chains, the starting $\boldsymbol{\beta}$ and $\boldsymbol{\mathcal{A}}$ parameters were different. Each chain had $10,000$ MCMC iterations, where the first $5,000$ were considered as burn-in. The regions left of the broken red vertical lines denote the burn-in periods for each chain.

6.2 Convergence of $\boldsymbol{\mathcal{A}}$ parameter

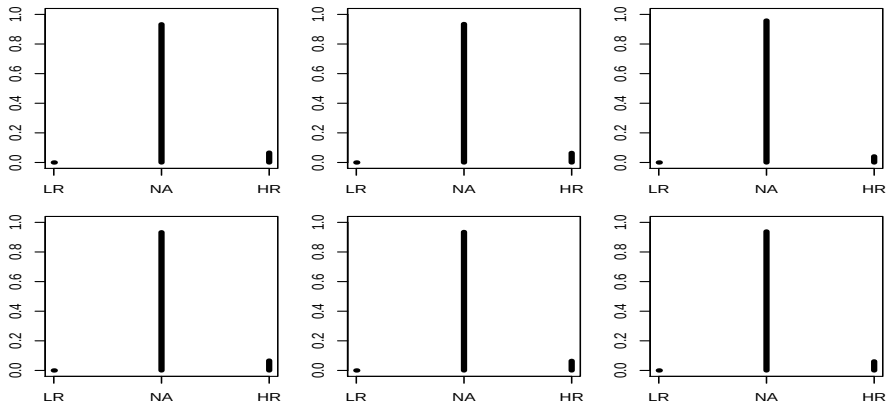
The parameter $\boldsymbol{\mathcal{A}}$ is a $p \times 3$ risk allocation matrix, where j-th row $\boldsymbol{\mathcal{A}}'_j$ corresponds to the allocation of j-th SNP, $j = 1, 2, ..., p$. The possible allocations of a SNP are $(1, 0, 0)$, $(0, 1, 0)$ or $(0, 0, 1)$ which respectively indicates that the SNP is in LR, NA or HR category. $\boldsymbol{\mathcal{A}}$ lies on a high-dimensional discrete space $\{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}^p$. For checking the convergence of $\boldsymbol{\mathcal{A}}$, we chose SNPs 2 ($\in$ HR), 3 ($\in$ LR) and 5 ($\in$ NA) from a randomly chosen dataset under Model 1 and checked the estimated marginal posteriors of $\boldsymbol{\mathcal{A}}_2$, $\boldsymbol{\mathcal{A}}_3$ and $\boldsymbol{\mathcal{A}}_5$ across 6 chains. Each chain had different starting parameters $\boldsymbol{\beta}$ and $\boldsymbol{\mathcal{A}}$. Figure 6 shows consistent estimate of marginal probabilities of each of the SNPs to belong to the 3 categories across all 6 chains. This shows convergence of the $\boldsymbol{\mathcal{A}}$ parameter.



(a) Marginal Posterior Distribution of SNP 2 (HR group) for 6 chains



(b) Marginal Posterior Distribution of SNP 3 (LR group) for 6 chains



(c) Marginal Posterior Distribution of SNP 5 (NA group) for 6 chains

Figure 6: (a), (b) and (c) The estimated marginal posterior distributions of 3 randomly chosen SNPs from the 3 categories are plotted for 6 independent chains under Model 1. Simulated data assumed 20 independent SNPs with the first 4 SNPs causal. BPM M-score approach was used for all the 6 chains. For this purpose, a dataset was chosen randomly from the 200 simulated datasets. For all the 6 chains, the starting $\boldsymbol{\beta}$ and $\boldsymbol{\mathcal{A}}$ parameters were different. Each chain had $10,000$ MCMC iterations, where the first $5,000$ were discarded.

7 Convergence checks for Real data analysis: gene MMRN1

We first analyzed gene MMRN1 from Chromosome 4 consisting of 57 SNPs after screening. For BPM, the M-score chain for α passed HW stationarity test and half-width mean test at 5% level without any burn-in while a burn-in of $100,000$ was needed for P-score. Figure 7 shows the posterior distributions and the trace plots of α for both the chains. We observe symmetric, unimodal distributions for α. Also, the trace plots are suggesting good mixing for α for both scores of BPM.

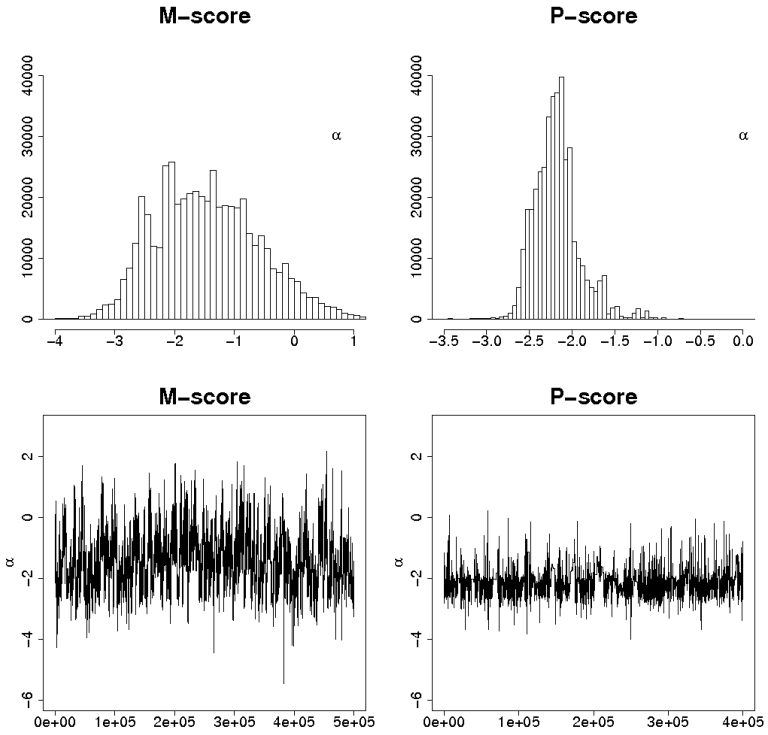


Figure 7: The first two plots give the posterior distribution of α for M- and P-scores respectively, while the last two plots give the trace plot of α over the $500,000$ RJMCMC iterations using MMRN1 gene dataset (with 57 SNPs after screening).

8 Prior specification of the SNPs

For the construction of our RJMCMC, we made three simplifying assumptions. One of the assumptions include equal prior probabilities for a SNP to be in each of the 3 categories LR (low-risk), NA (null) and HR (high-risk). This choice of prior gave a simplified form (equation (4) of main paper) of the acceptance probability in equation (3) of main paper. For real data analysis, we applied our BPM method to the top two genes identified by another gene-based association analysis (which provided only gene-based p-values). So, our decision to assign such high prior probability for each SNP to be in non-null category is reasonable. However, if applied to a genome-wide data, a more informative prior would be to assign much higher probability for each SNP to be in the NA (null) group. For example, BEAM has default prior probabilities of 0.01 for each SNP to belong to marginal or interaction groups. We performed a small simulation experiment to study how change in the prior specification affects the performance of BPM.

Let π be the prior probability of a SNP to be non-null. The prior probabilities of the SNP to be in LR, NA and HR categories are $\pi/2$, $1-\pi$ and $\pi/2$ respectively. For this simulation experiment, we considered 3 different prior settings.

Prior 1: $\pi = 2/3$ (this is the simplifying prior we used in our paper)

Prior 2: $\pi = 0.05$

Prior 3: $\pi = 0.01$

We simulated data for Models 1 and 3 as described in Section 3.1 of our main paper.

Model 1: $\text{logit(p)} = -4 + \frac{1}{5}X_1 + \frac{1}{5}X_2 - X_3 - X_4$, where $X_j = 0, 1, 2$ denote SNP j with $0, 1, 2$ minor alleles respectively.

Model 3: $\text{logit(p)} = -4 + 2X_1X_2X_3X_4$, where $X_j = 0, 1, 2$ denote SNP j with $0, 1, 2$ minor alleles respectively.

For our simulation scenario with 4 non-null SNPs among 20, a prior more stringent than Prior 3 does not make sense. For each model and each prior, we applied both M- and P-scores of BPM. For comparison of BPM performance across different priors, we calculated the true positive rate (tpr) and the false positive rate (fpr) of BPM. Details of calculation of fpr and tpr can be found in Section 3.1 of main paper. Table 2 shows the performance

Table 2: BPM performance for different priors: Power of the BPM approach in detecting the four associated SNPs for Bonferroni corrected error level 0.0025 (= 0.05/20) based on 200 datasets with 200 cases and 200 controls. 3 different prior specifications have been used to study BPM's performance. Default prior is the equal probability prior used throughout in the paper. π is the probability of a SNP to be non-null. For more stringent prior (such as $\pi = 0.001$), the chains were run longer to ensure convergence. N is the length of the chain. The first $N/2$ iterations were discarded as burn-in.

Prior	N	**Model 1** (main effects only)		**Model 3** (interaction effect only)	
		M-score	P-score	M-score	P-score
Prior 1 (default)	$10,000$	0.433	0.439	0.211	0.479
Prior 2 ($\pi = 0.05$)	$10,000$	0.461	0.424	0.188	0.514
Prior 3 ($\pi = 0.01$)	$10,000$	0.443	0.456	0.184	0.460

of BPM for different models for different choices of prior. We see that choice of prior does not strongly affect power of BPM in detecting association from different models. Our choice of prior not only simplifies the acceptance probability in our RJMCMC but also requires a smaller number of MCMC iterations for convergence. More stringent priors would require longer chains and more computation time.

9 Adjustment of additional covariate effects

The BPM model (1) provides the flexibility of adjusting other covariate effects. Suppose, in addition to response vector $\boldsymbol{Y}_{n\times 1}$ and genotype matrix $\boldsymbol{X}_{n\times p}$, all individuals have data on q covariates (such as age, gender, race etc.). Let $\boldsymbol{C} = (\boldsymbol{C}_1, ..., \boldsymbol{C}_q)$ be the $n \times q$ matrix of covariates. Given a specific risk allocation $\boldsymbol{\mathcal{A}}$, we use the logistic regression

$$\log\left(\frac{P(y_i = 1|\boldsymbol{X}_i, \boldsymbol{\mathcal{A}}, \boldsymbol{C}_i)}{1 - P(y_i = 1|\boldsymbol{X}_i, \boldsymbol{\mathcal{A}}, \boldsymbol{C}_i)}\middle|\boldsymbol{\beta}, \boldsymbol{\gamma}\right) = \boldsymbol{\beta}'\boldsymbol{Z}_i + \boldsymbol{\gamma}'\boldsymbol{C}_i \tag{4}$$

where $\boldsymbol{\beta} = (\alpha, \beta_1, \beta_2)'$ is the vector of fixed effects for the 3 groups of SNPs, $\boldsymbol{Z}_i = (1, Z_{1i}, Z_{2i})'$ is the vector of scores for the SNP groups of an individual i ($i = 1, 2, ..., n$), and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)'$ is the vector of q fixed effects corresponding to the q covariates. In the absence of additional covariates (as in equation (1)), we assumed priors for our paramaters $\boldsymbol{\beta}$ and $\boldsymbol{\mathcal{A}}$ (details can be found in the main paper):

$$\boldsymbol{\beta} \sim N_3(\boldsymbol{\mu}, \boldsymbol{V}) \times I(\beta_1 < 0) \times I(\beta_2 > 0)$$

$$P[\boldsymbol{\mathcal{A}}] = \prod_{j=1}^{p} P[\boldsymbol{\mathcal{A}}_j] = constant$$

Currently, due to other covariates, we need to assume a suitable prior for the additional parameter $\boldsymbol{\gamma}$. Assume $\boldsymbol{\gamma} \sim N_q(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\gamma}})$ for some suitable choice of covariance matrix $\boldsymbol{V}_{\boldsymbol{\gamma}}$. Our interest is still in the joint posterior distribution of $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{\beta}$. We use our RJMCMC scheme to study this joint posterior. The acceptance probability for dimension moves (from step $\overline{t-1}$ to step t) is $\min\left(1, a(K^{(t-1)}, K^{(t)})\right)$, with

$$a(K^{(t-1)}, K^{(t)}) = \frac{P[\boldsymbol{y}|\boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)}, \boldsymbol{\mathcal{A}}^{(t)}, K^{(t)}]}{P[\boldsymbol{y}|\boldsymbol{\beta}^{(t-1)}, \boldsymbol{\gamma}^{(t-1)}, \boldsymbol{\mathcal{A}}^{(t-1)}, K^{(t-1)}]} \tag{5}$$

Here K is the dimension parameter. At t-th iteration of the RJMCMC, $K^{(t)}$ indicates whether there is only NA SNPs, or NA & LR SNPs, or NA & HR SNPs, or all 3 groups of SNPs in the model. One can deduce $K^{(t)}$ from the corresponding risk allocation $\boldsymbol{\mathcal{A}}^{(t)}$. We obtain $P[\boldsymbol{y}|\boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)}, \boldsymbol{\mathcal{A}}^{(t)}, K^{(t)}]$ and $P[\boldsymbol{y}|\boldsymbol{\beta}^{(t-1)}, \boldsymbol{\gamma}^{(t-1)}, \boldsymbol{\mathcal{A}}^{(t-1)}, K^{(t-1)}]$ using the model

in (4).

As before, within a fixed dimension, we first update $\boldsymbol{\mathcal{A}}$ from its full conditionals and then update $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ simultaneously using Metropolis Hastings algorithm. The full conditional of $\boldsymbol{\mathcal{A}}_j$ at step t has the multinomial distribution

$$\left[\boldsymbol{\mathcal{A}}_j^{(t)} \middle| \boldsymbol{\beta}^{(t-1)}, \boldsymbol{\gamma}^{(t-1)}, \boldsymbol{\mathcal{A}}_{(-j)}^{(t-1)}\right] \sim Multinomial\left(m = 1; p_{j1}^{(t)}, p_{j2}^{(t)}, p_{j3}^{(t)}\right)$$

where $p_{j1}^{(t)}, p_{j2}^{(t)}, p_{j3}^{(t)}$ are the posterior probabilities of SNP j to be in the LR, NA and HR group respectively. These posterior probabilities are given by

$$p_{js}^{(t)} = \frac{P[\boldsymbol{y}|\boldsymbol{\beta}^{(t-1)}, \boldsymbol{\gamma}^{(t-1)}, \boldsymbol{\mathcal{A}}_{(-j)}^{(t-1)}, \boldsymbol{\mathcal{A}}_j^{(t)} = \boldsymbol{a}_s] P[\boldsymbol{\mathcal{A}}_j^{(t)} = \boldsymbol{a}_s]}{\sum_{k=1}^{3} P[\boldsymbol{y}|\boldsymbol{\beta}^{(t-1)}, \boldsymbol{\gamma}^{(t-1)}, \boldsymbol{\mathcal{A}}_{(-j)}^{(t-1)}, \boldsymbol{\mathcal{A}}_j^{(t)} = \boldsymbol{a}_k] P[\boldsymbol{\mathcal{A}}_j^{(t)} = \boldsymbol{a}_k]}$$

where $s = 1, 2, 3$, $\boldsymbol{a}_s \in \{(1,0,0)', (0,1,0)', (0,0,1)'\}$. After updating $\boldsymbol{\mathcal{A}}$ from its full conditionals and getting $\boldsymbol{\mathcal{A}}^{(t)}$, we sample $\boldsymbol{\beta}^*$ from the proposal density $N_3\left(\boldsymbol{\beta}^{(t-1)}, \boldsymbol{V}\right) I(\beta_1 < 0) I(\beta_2 > 0)$, and sample $\boldsymbol{\gamma}^*$ from $N_q\left(\boldsymbol{\gamma}^{(t-1)}, \boldsymbol{V}_{\boldsymbol{\gamma}}\right)$. For each draw of $(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)$ from the proposals, we accept $(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)$ as $(\boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)})$ with probability $\min\left(1, a'\left[\begin{pmatrix}\boldsymbol{\beta}^{(t-1)} \\ \boldsymbol{\gamma}^{(t-1)}\end{pmatrix}, \begin{pmatrix}\boldsymbol{\beta}^{*} \\ \boldsymbol{\gamma}^{*}\end{pmatrix}\right]\right)$, where

$$a'\left[\begin{pmatrix}\boldsymbol{\beta}^{(t-1)} \\ \boldsymbol{\gamma}^{(t-1)}\end{pmatrix}, \begin{pmatrix}\boldsymbol{\beta}^{*} \\ \boldsymbol{\gamma}^{*}\end{pmatrix}\right] = \frac{P[\boldsymbol{y}|\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, \boldsymbol{\mathcal{A}}^{(t)}].P[\boldsymbol{\beta}^*|\boldsymbol{\mathcal{A}}^{(t)}].P[\boldsymbol{\gamma}^*]}{P[\boldsymbol{y}|\boldsymbol{\beta}^{(t-1)}, \boldsymbol{\gamma}^{(t-1)}, \boldsymbol{\mathcal{A}}^{(t)}].P[\boldsymbol{\beta}^{(t-1)}|\boldsymbol{\mathcal{A}}^{(t)}].P[\boldsymbol{\gamma}^{(t-1)}]} \times \frac{P[\boldsymbol{\beta}^{(t-1)}|\boldsymbol{\beta}^*]}{P[\boldsymbol{\beta}^*|\boldsymbol{\beta}^{(t-1)}]}$$

10 More about the real data (ARIC Study)

The ARIC study is an ongoing prospective study designed to investigate the etiology and natural history of atherosclerosis and its clinical manifestations, and to measure variation in cardiovascular risk factors, medical care and disease by race, gender, place and time (The ARIC Investigators, 1989). It is a multicenter contract supported by the National Heart, Lung, and Blood Institute. Participants were randomly chosen from four US communities (Forsyth County, NC; Jackson, MS; suburban Minneapolis, MN; Washington County, MD), totaling $15,792$ persons (8710 women, 7082 men) aged 45-64 at baseline (1987-89). The Jackson cohort represents 100% black population while the other three cohorts represent ethnic mix of their communities. Re-examinations of these participants were done in approximate intervals of 3 years. Yearly follow-up interviews are conducted over telephone to assess health status of participants.

ARIC has collected fasting glucose measures from the entire cohort at 4 separate visits over a 9-year period and self-reported physician diagnosis and medication use in up to 14 separate interviews over a 20-year period. Diabetes was classified as 'yes' if any of the following four criteria were met at baseline: fasting glucose $\geq$126 mg/dL, non-fasting glucose $\geq$ 200 mg/dL, self-reported physician diagnosis of diabetes, or self-reported current use of diabetes medications.

The ARIC cohort has been genotyped using the Affymetrix Genome-Wide SNP Array 6.0. Genotyping was completed at the Broad Institute of MIT and Harvard in three batches; the Birdseed algorithm was used for genotype calling. Imputation was performed using MACH 1.0 and HapMap release 21 (Build 35). SNPs with a call rate $< 90\%$, maf $< 1\%$, or deviation from Hardy-Weinberg equilibrium ($p < 10^{-6}$) were excluded for imputation. A total of $869,224$ SNPs were successfully genotyped, and over 2.8 million SNPs were successfully genotyped or imputed. Subjects with call rate $< 95\%$, sex mismatches, inferred 1st degree relatives, extensive mismatches with a non-GWAS reference panel, and genetic outliers based on IBS clustering or EIGENSTRAT (Patterson et al., 2006) were excluded from GWAS analyses.

References

N. J. Patterson, A. L. Price, and D. Reich. Population structure and eigenanalysis. *PLoS Genet*, 2(12):2074–2093, Dec 2006.

S.A. Sisson. Transdimensional markov chains: A decade of progress and future perspectives. *J Am Stat Assoc*, 100:1077–1089, 2005.

The ARIC Investigators. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. *Am J Epidemiol*, 129(4):687–702, 1989.