

Supplementary Materials for **Integrative Bayesian Analysis of Neuroimaging-Genetic Data with Application to Cocaine Dependence**

Shabnam Azadeh, Brian P. Hobbs, Liangsuo Ma, David A. Nielsen, Frederick G. Moeller, and Veerabhadran Baladandayuthapan

S1 Markov Chain Monte Carlo Performance

In this section, we report the the details of Markov Chain Monte Carlo (MCMC) performance in computing the posterior model-specific probabilities and Bayesian model averaging for (i) entire brain and (ii) the voxels that were flagged as significant using our FDR approach for *any* genetic covariate – which resulted in the total number of significant voxels being 23,293. We used *BMS* package of *R* software to produce the results. The *bms* function in the R package "BMS" was used to consider a burn-in of 10000, and 20000 iterations. The best models are used for convergence analysis between the likelihoods and MCMC frequencies, as well for as likelihood-based inference. Our MCMC method was based on a birthdeath algorithm.

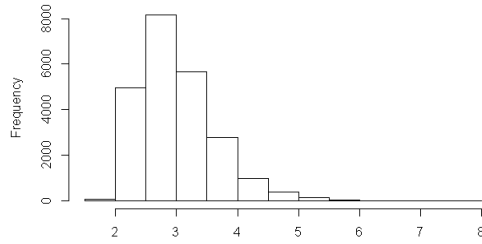
Average number of regressors: Figure S1(a) and Figure S2(a) present the posterior mean of model size using BMA for significant voxels and entire brain, respectively. The average of model size is 3.007 and 2.081 with standard deviation of 0.6272 and 0.5620 for significant voxel and whole brain, respectively. See Table S2 for all the summary measures.

Acceptance rates of MCMC sampler: For each voxel, we calculated the MCMC acceptance rate as the ratio of the number of times that a model was accepted to the number of MCMC iterations. Figure S1(b) and Figure S2(b) depict the acceptance rate of MCMC method while deriving the posterior probabilities of significant voxels and whole brain, respectively. The mean acceptance rates are 0.25 and 0.24 for for significant voxels and entire brain, respectively, which indicate good mixing behaviors of our sampler.

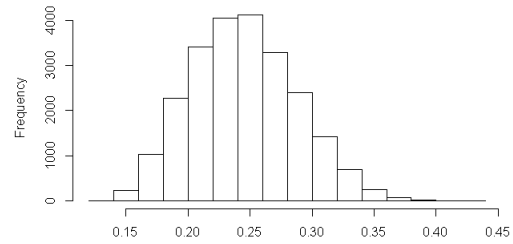
Computation times: The average MCMC elapsed time is 4.856 and 4.154 seconds with stander deviation of 0.36 and 0.44 for significant voxels and entire brain, respectively. ¹

Correlation between MCMC frequencies and marginal likelihoods: Considering the size of \mathcal{M} , $2^{24} = 16777216$, needs a large of the MC^3 sampler to determine the high posterior probability models. We report a correlation coefficient between visit frequencies and posterior probabilities based on equation (3.1.6), main paper, from a run of 20000 recorded drawings after a burn-in of 10000 drawings. Figure S1(c) and Figure S2(c) represents the correlation between the MCMC frequencies and their marginal likelihoods of significant voxels and whole brain. The high values of correlation show the set of iterations draw models with high posterior probabilities which indicate the good behavior of the sampler (Frenández et al 2001).

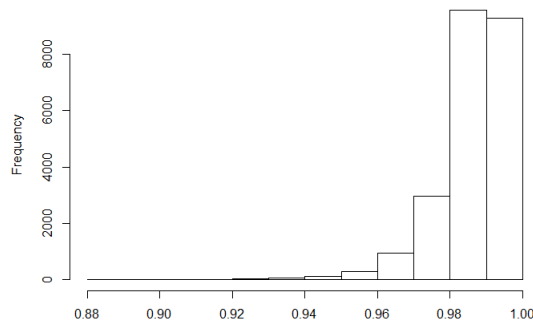
¹The computation system consists of 220 processor cores in 16 compute nodes. 8 nodes have 12 processor cores (2.67 GHz) and 96GB of RAM per node. 8 nodes have 16 processor cores (2.00 GHz) and 128GB of RAM per node. All of the nodes are connected via 10Gbit Ethernet both to each other and to the PVFS fast scratch storage system.



(a)



(b)



(c)

Figure S1: Significant voxels: (a) Top left represents histogram of posterior mean of model size. (b) Top right shows histogram of MCMC acceptance rate. (c) Bottom is a histogram of correlation between the MCMC frequencies and their marginal likelihood for significant voxels.

Table S1 list the descriptive statistics of average number of regressors, acceptance rates of MCMC sampler, Computation times, and Correlation between MCMC frequencies and marginal likelihoods for both significant voxels and entire brain.

We also plotted the average number of regressors across the entire brain obtained by our BMA procedure. Figure S3 displays the coronal, sagittal, and axial views of average number of regressors map for the whole white matter of brain.

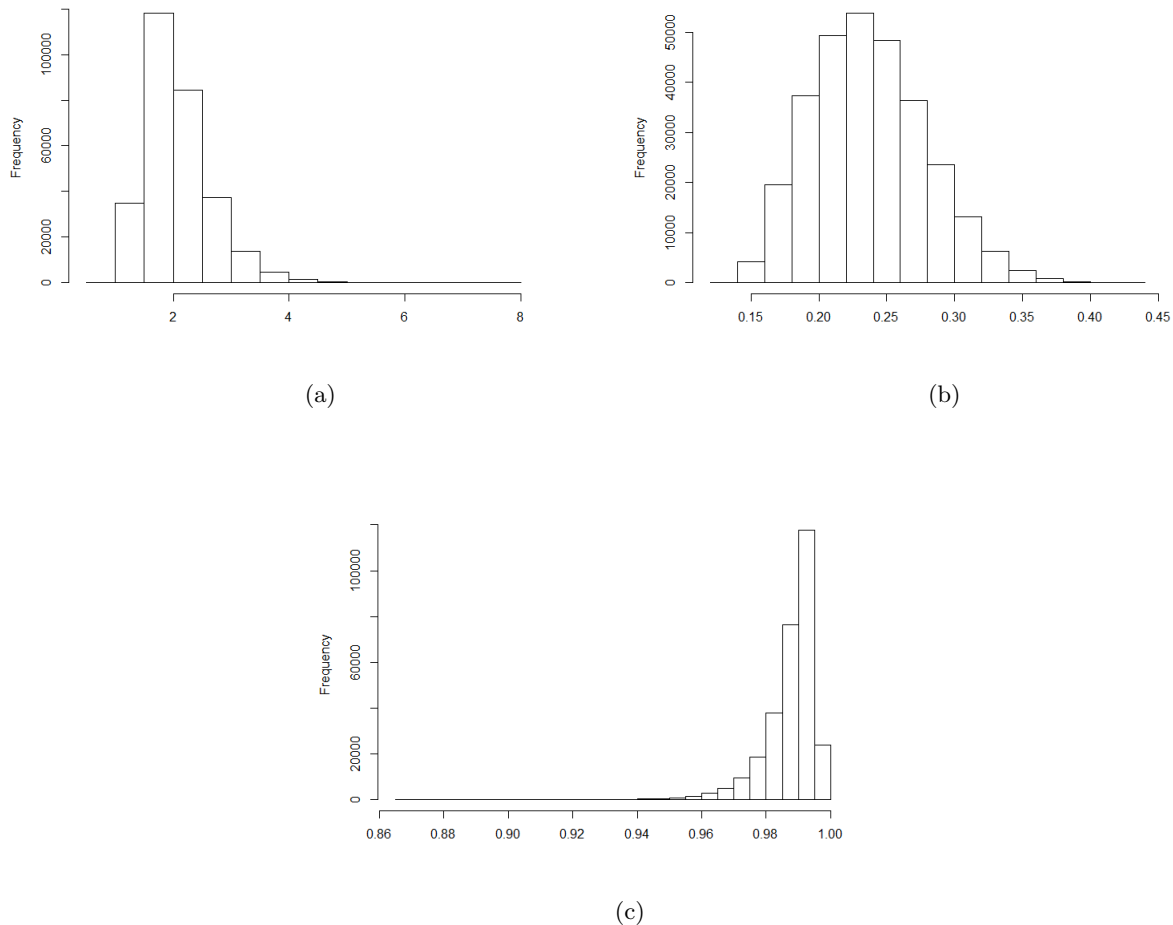


Figure S2: Entire brain: (a) Top left represents histogram of posterior mean of model size. (b) Top right shows histogram of MCMC acceptance rate. (c) Bottom is a histogram of correlation between the MCMC frequencies and their marginal likelihood.

S2 Model performance

We evaluated the performance of our Bayesian model averaging (BMA) based model and full Bayesian model (Full) with no model averaging. We computed two model selection metrics as follows:

1. approximate Deviance information criterion (aDIC): We used a variant of the Deviance information criterion (DIC) which is a hierarchical modeling generalization of the AIC (Akaike information criterion) and BIC (Bayesian information criterion).

For voxel ν , DIC is calculated as:

$$DIC(\nu) = D(\bar{\boldsymbol{\theta}}) + 2p_D, \quad (\text{S2.0.1})$$

where $D(\bar{\boldsymbol{\theta}})$ is a classical estimate of fit, and p_D presents the effective number of parameters

Table S1: Summary of Markov Chain Monte Carlo Performance: (a) top panel shows the summary of MCMC performance for significant voxels. (b) bottom panel indicate the summary of MCMC performance for entire brain.

(a) Significant voxels	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Average number of regressors	1.841	2.546	2.892	3.007	3.344	7.704
Acceptance rates of MCMC sampler	0.1366	0.2140	0.2432	0.2449	0.2738	0.4284
Computation times	4.027	4.590	4.828	4.856	5.095	6.237
Correlation between MCMC frequencies and marginal likelihoods	0.8840	0.9824	0.9884	0.9859	0.9921	0.9980
(b) Entire brain						
Average number of regressors	0.9896	1.6700	1.9750	2.0810	2.3740	7.6420
Acceptance rates of MCMC sampler	0.1283	0.2056	0.2338	0.2367	0.2643	0.4362
Computation times (seconds)	2.885	3.819	4.136	4.154	4.451	6.498
Correlation between MCMC frequencies and marginal likelihoods	0.8676	0.9847	0.9898	0.9876	0.9929	0.9984

of the model (Spiegelhalter et al (2002)). In BMA, we consider the marginal likelihoods of models. So $DIC_j(\nu)$ in model space \mathcal{M} is calculated as:

$$DIC_j(\nu) = 2 \{k_j(\nu) + 1\} - 2 \{\mathcal{L}_j(\nu)\}, \quad (\text{S2.0.2})$$

where $k_j(\nu)$ presents the total number of regressors in model M_j , and $\mathcal{L}_j(\nu)$ is the marginal likelihood of model M_j . Using a proportion of MCMC frequencies as a weight, we can extend DIC to BMA settings as a weighted average of the model specific DICs. Thus, ‘‘approximate’’ deviance information criterion (aDIC) for voxel ν , $aDIC_j(\nu)$, in model space \mathcal{M} is calculated as

$$aDIC(\nu) = \sum_{j=1} DIC_j(\nu) \times w_j, \quad (\text{S2.0.3})$$

where w_j is a weight determined by the MCMC sampling frequency of model j .

To formally compare the aDIC of BMA, $aDIC_{BMA}$, versus the DIC of full model, DIC_{Full} , we used a paired t-test where the null hypothesis $H_0 : aDIC_{BMA} - DIC_{Full} = 0$ versus a (one-sided) alternative hypothesis of $H_a : aDIC_{BMA} < DIC_{Full}$. The p-values are very close to zero ($p < 2.2 \times 10^{-16}$) – thus rejecting the null hypothesis in favor of the alternative hypothesis, $aDIC_{BMA} < DIC_{Full}$, which shows BMA performs better in terms of the aDIC model selection criterion.

2. Bayesian information criterion (BIC): Bayesian information criterion (BIC) is another type of the model selection criterion that can be used to choose among several set of models. It is closely related to AIC. Similarly, for voxel ν , $BIC_j(\nu)$ of the models in model space \mathcal{M} is derived by:

$$BIC_j(\nu) = -2(\mathcal{L}_j(\nu)) + (k_j(\nu) + 1)\log(n), \quad (\text{S2.0.4})$$

where $\mathcal{L}_j(\nu)$ is the marginal likelihood of model M_j , $k_j(\nu)$ is the total number of regressors in model M_j , and n is the sample size. In BMA, $BIC(\nu)$ is calculated by:

$$BIC(\nu) = \sum_{j=1} BIC_j(\nu) \times w_j, \quad (\text{S2.0.5})$$

where w_j is a weight based on the number of MCMC sampling frequency of model j . Since only have one model for Bayesian full model, so BIC (and DIC) are calculated without considering the weights.



Figure S3: Brain map depicting numbers of regressors of the entire white matter. Top left: coronal view, Top right: sagittal view, Bottom left: axial view. The colorbar indicates number of regressor for each voxel. The multi-slice sagittal views were generated using MRICroN software.

To formally compare the BIC of BMA, BIC_{BMA} , versus the BIC of full model, BIC_{Full} , we used a paired t-test where the null hypothesis $H_0 : BIC_{BMA} - BIC_{Full} = 0$ versus a (one-sided) alternative hypothesis of the $H_a : BIC_{BMA} < BIC_{Full}$. The p-values are very close to zero ($p < 2.2 \times e^{-16}$) – thus rejecting the null hypothesis in favor of the alternative hypothesis, $BIC_{BMA} < BIC_{Full}$, which shows BMA performs better in terms of BIC model selection criterion.

As in previous section, we focused our attention on the voxels that were flagged as significant using our FDR approach for *any* genetic covariate – which resulted in the total number of significant voxels being 23,293. Figure S4 depicts the boxplots of DIC and BIC respectively for the significant voxels. In both scenario BMA have smaller DIC and BIC which indicate the better performance in terms of goodness-of-fit. We further calculated the (absolute) difference between DIC and BIC for BMA and full model. The difference is always negative which indicates BMA is always yielded improved performance when compared to the full model. Figure S5 shows the boxplots of differences. Table S2 list the summary of descriptive statistics for DICs and BICs in BMA and full Bayesian models.

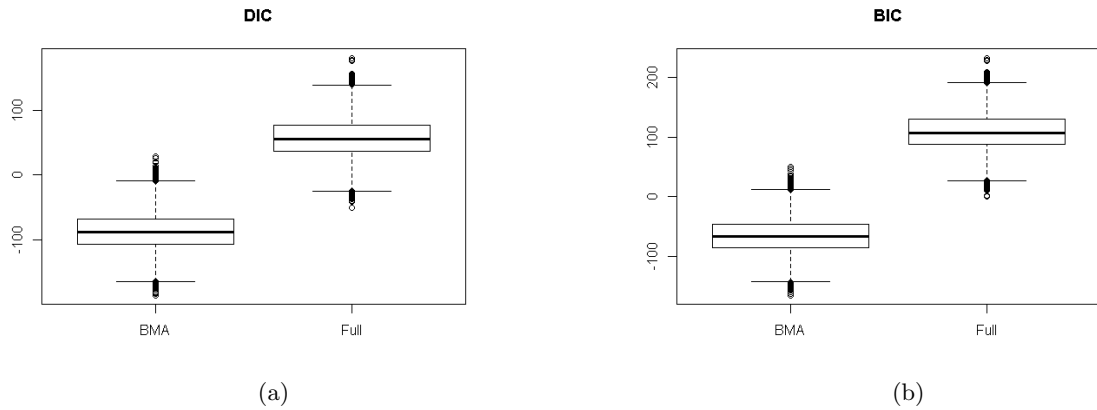


Figure S4: Left (a) is a boxplot of approximate deviance information criterion. Right (b) is a boxplot of Bayesian information criterion

Table S2: Summary of approximate deviance information criterion and Bayesian information criterion for both Bayesian model averaging and full Bayesian model.

Model Selection Criteria	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
aDIC for BMA	-185.10	-106.00	-87.76	-86.40	-67.20	28.43
DIC for Full	-50.95	36.68	55.76	57.26	77.69	180.40
BIC for BMA	-165.00	-84.93	-66.70	-65.35	-46.26	49.31
BIC for Full	0.55	88.19	107.30	108.80	129.20	231.90

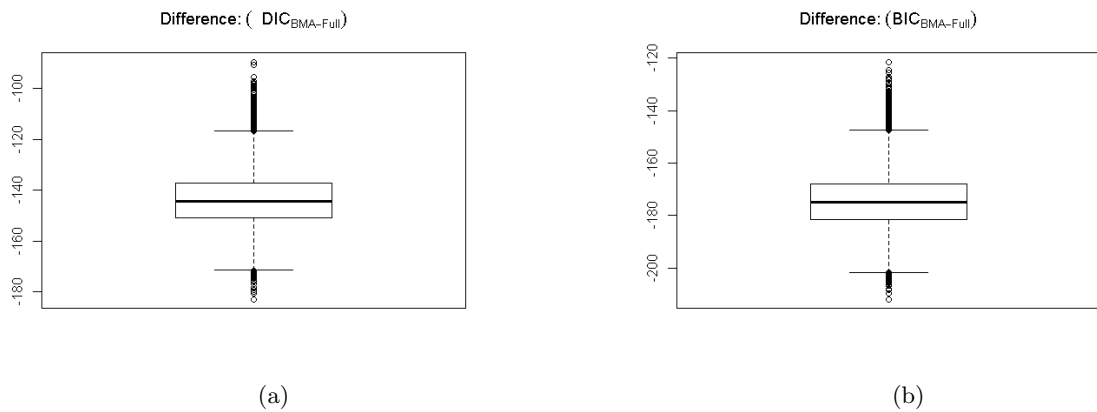


Figure S5: Left (a) is a boxplot of difference of deviance information criterion. Right (b) is a boxplot of difference of Bayesian of information criterion

S3 Predictive performance

We evaluate the predictive performance of our models as a measure of goodness of fit. Assuming we want to predict the observable $\mathbf{Y}_f(\nu)$ given the corresponding regressor X_f , where f is a forecast. We can calculate the predictive distribution of $\mathbf{Y}_f(\nu)$ as follows (Frenández et al 2001):

$$p(\mathbf{Y}_f(\nu) | Y(\nu)) = \sum_{j=1} f_t(\mathbf{Y}_f(\nu) | n-1, \overline{\mathbf{Y}}(\nu) + \frac{1}{g+1} X'_{f,j} \beta_j^*, \frac{n-1}{d_j^*} \left\{ 1 + \frac{1}{n} + \frac{1}{g+1} X'_{f,j} (X'_{f,j} X_j)^{-1} X_{f,j} \right\}^{-1}) P(M_j | \mathbf{Y}(\nu)), \quad (\text{S3.0.1})$$

where $f_t(x | d, b, a)$ presents the p.d.f of a univariate t-student distribution with degree of freedom of d , location b , and precision of a . The $\overline{\mathbf{Y}}(\nu)$ is a average of response variables. The $X_{f,j}$ shows the j elements of X_f corresponding to the model M_j regressor. $\beta_j^* = (X'_{f,j} X_j)^{-1} X_{f,j} \mathbf{Y}(\nu)$, and

$$d_j^* = \frac{1}{g+1} \mathbf{Y}'(\nu) M_{X_j} \mathbf{Y}(\nu) + \frac{g}{g+1} (\mathbf{Y}(\nu) - \overline{\mathbf{Y}}(\nu))^{-1} (\mathbf{Y}(\nu) - \overline{\mathbf{Y}}(\nu)). \quad (\text{S3.0.2})$$

To evaluate the predictive performance, we split the sample into train and test sets as follows: (i) n observations are used for posterior inference (training) and (ii) q observations are retained (test-set) to examine the precision of the predictive performance. Then, for each voxel ν in the tests set, we compute the log predictive score (LPS) for $f = n+1, \dots, n+q$ as follows:

$$LPS(\nu) = -\frac{1}{q} \sum_{f=n+1}^{n+q} \log\{p(\mathbf{Y}_f(\nu) | \mathbf{Y}(\nu))\} \quad (\text{S3.0.3})$$

A smaller $LPS(\nu)$ indicates the better prediction. Assuming *i.i.d* sampling the LPS approximates an integral which equivalent to the sum of Kullback-Leibler divergence between the actual sampling density and the predictive density which is provided in equation (S3.0.6), and the entropy of the sampling distribution. Thus, LPS capture uncertainty through two sources: (i) the lack of fit and (ii) inherent sampling uncertainty (see Frenández et al 2001 for more details). Under a Normal sampling model with fixed variance σ_* this entropy equals to $\ln(\sigma_* \sqrt{2\pi e})$. So, a known normal Normal distribution with fixed σ_* would have the same inherent predictive uncertainty. Hence, by choosing $\sigma_* = \frac{\exp(LPS)}{\sqrt{2\pi e}}$, we could calculate the inherent uncertainty through the LPS scores.

We evaluated the predictive performance of LPS over 4 different training and tests splits: {33%, 25%, 20%, 10%}. We compared the inherent model uncertainty through LPS for Bayesian model averaging (BMA) and Full Bayesian model (Full). Figure S6 show the contour (scatter) plot of LPS for BMA versus Full model for all voxels. In all scenarios, the BMA-based LPS scores are lower than the Full model, thus indicating better predictive performance. To formally compare the LPS of BMA, LPS_{BMA} , versus the LPS of full model, LPS_{Full} , we used a paired t-test where the null hypothesis is $H_0 : LPS_{BMA} - LPS_{Full} = 0$ versus a (one-sided) alternative hypothesis of $H_a : LPS_{BMA} < LPS_{Full}$. For all four scenarios, the p-values are very close to zero ($p < 2.2 \times e^{-12}$) – thus rejecting the null hypothesis in favor of the alternative hypothesis, $LPS_{BMA} < LPS_{Full}$, which shows BMA performs better in terms of inherent uncertainty and prediction.

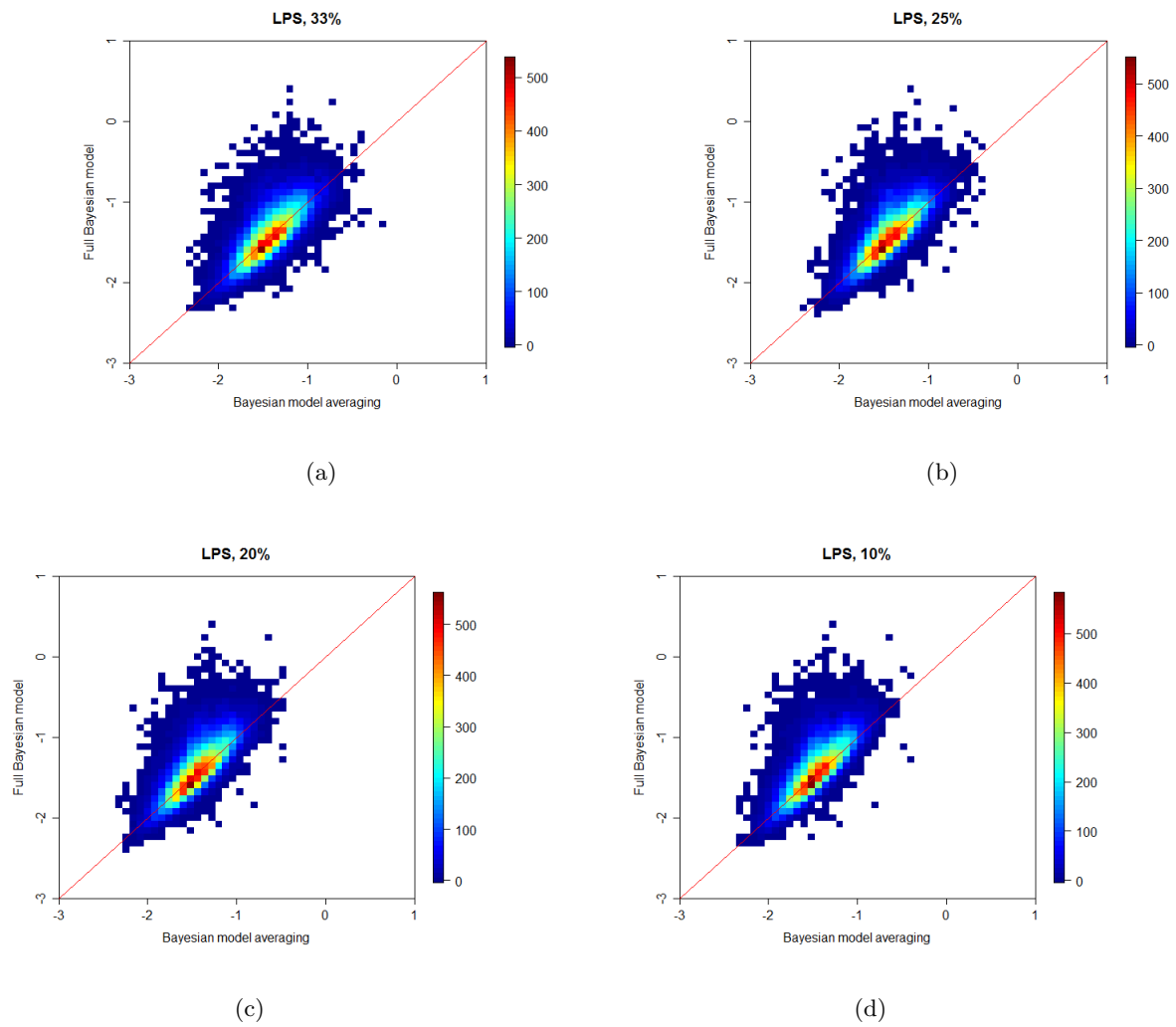


Figure S6: Contour plot of LPS scores for different test splits: (a) 33% , (b) 25% , (c) 20% and (d) 10% of observations. The red-line indicates the 45-degree line through origin.

S4 Comparison of the regression coefficients of full Bayesian model and Bayesian model averaging

In this section, we compare the regression coefficient estimates ($\beta(\nu)$'s) obtained from BMA and full Bayesian model for the voxels mapped to entire white matter of the brain. Figure S7 depicts the contour plot of regression coefficient of BMA versus full model for the covariates: cocaine abuse, $GAD1^a$ and $GAD1^b$. As can be seen there is much larger spread of the β 's obtained from BMA as compared to the full model which typically result in more shrunken (towards zero) effects. Furthermore, we evaluated the number of significant voxels obtained using the full model as follows. Since the posterior probability of BMA and full model are not directly comparable – since the full model does not incorporate model selection – we computed the number of significant voxels by deriving the credible intervals of each regression coefficient, and counting number of times that $\beta = 0$ is not contained in the credible intervals. Figure S8 provides the number of significant voxels of BMA versus Full model for the three covariates: cocaine abuse, $GAD1^a$ and $GAD1^b$. The results show that BMA detects much larger number of significant voxels as compared to full model, which we conjecture is due to the over-shrinkage of the estimates.

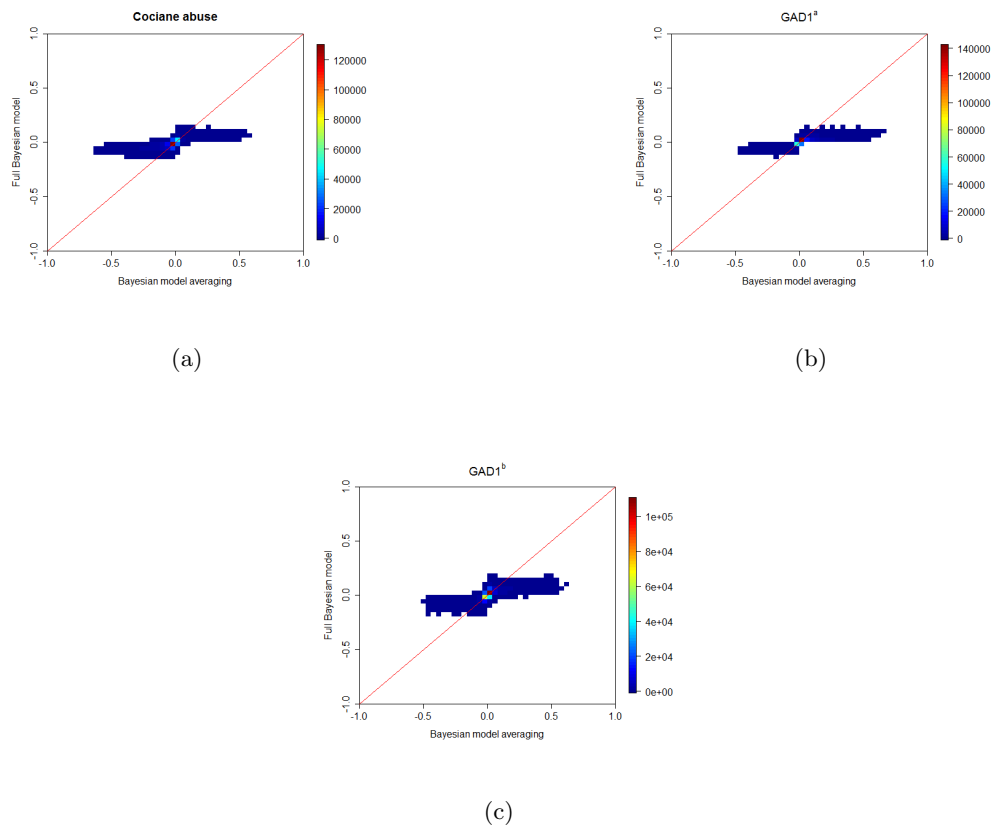
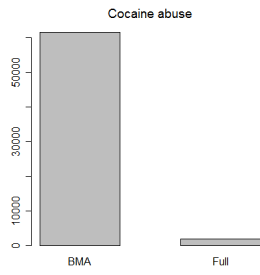
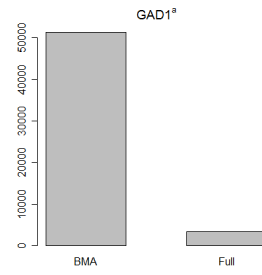


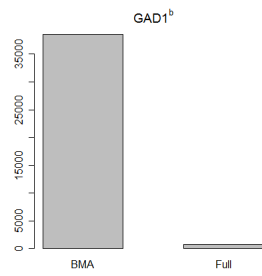
Figure S7: (a) Top left represents contour plot of Bayesian model averaging versus full Bayesian model for cocaine abuse. (b) Top right shows contour plot of Bayesian model averaging versus full Bayesian model for $GAD1^a$. (c) Bottom is a contour plot of Bayesian model averaging versus full Bayesian model for $GAD1^b$. The red-line indicates the 45-degree line through origin.



(a)



(b)



(c)

Figure S8: Barplot of significant voxels for: (a) top left cocaine abuse, (b) top right $GAD1^a$, and bottom $GAD1^b$.

S5 Multi-slice sagittal views

Figures S9, S10, and S11 provide more anatomic locations of Significant regions (SRs) for genetic variants with more than 1000 significant voxels and cocaine consumption abuse, respectively.

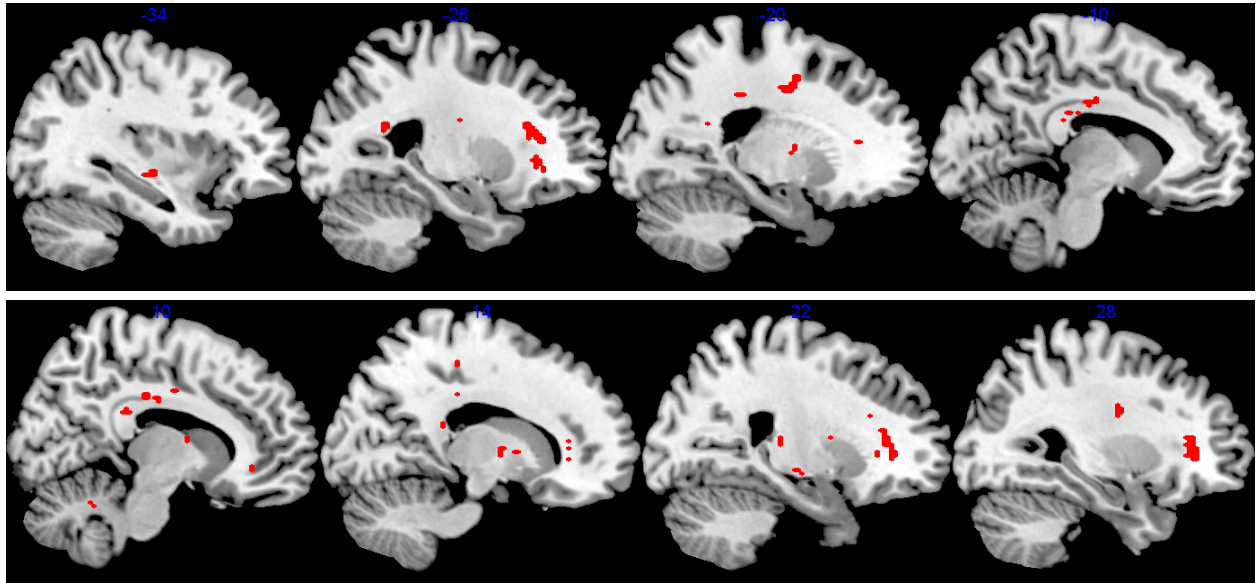


Figure S9: Multi-slice sagittal views displaying significant regions² associated with *GAD1*^a. The SRs, depicted in red, characterize locations for which alteration of mean FA was evident from statistical analysis.

² In the significant voxels, the adjacent voxels equal to or greater than 20 were grouped and named as significant regions (SRs) to limit noisy images while plotting the sagittal views so that all significant voxels might not be displayed. Significant voxels were grouped in SPM software using MarsBaR tool. Then multi-slice sagittal views were generated using MRIcroN software. (Same strategy is used in Figure S10 and Figure S11.)

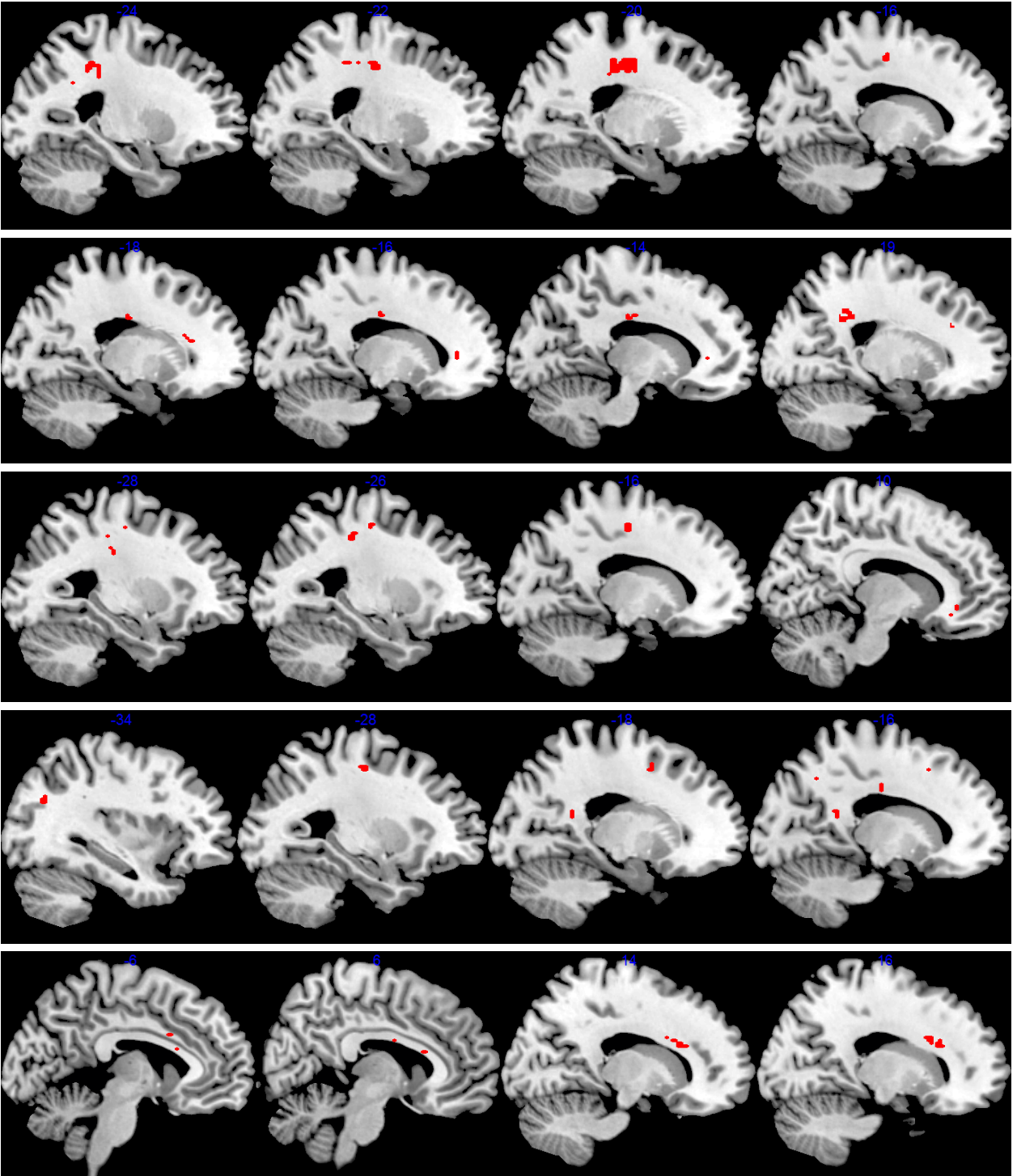


Figure S10: Multi-slice sagittal views displaying significant regions for the following genetic variants: *HTR2A*, *TH*, *SLC6A4^b*, *ADRA1A*, and *SLC6A3^b*, respectively from the top to the bottom panel. The SRs, depicted in red, characterize locations for which alteration of the mean FA was evident from statistical analysis.

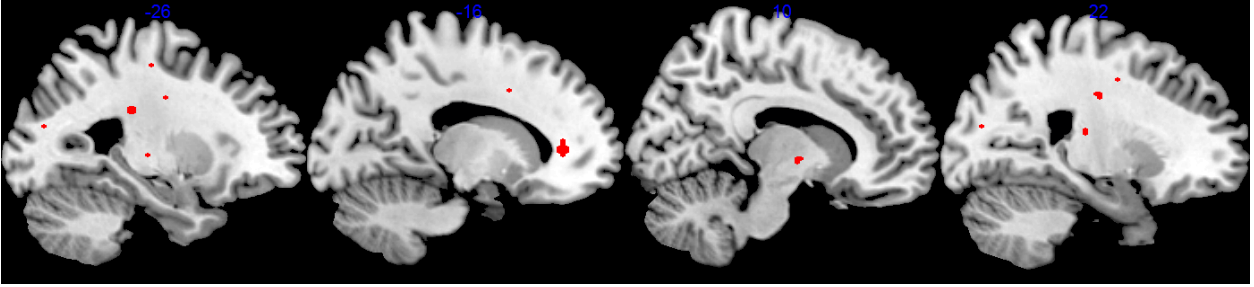


Figure S11: Multi-slice sagittal views displaying significant regions associated with cocaine consumption. The SRs, depicted in red, characterize locations for which alteration of the mean FA was evident from statistical analysis.