

Submitted to the Annals of Applied Statistics

SEMI-PARAMETRIC COVARIATE-MODULATED LOCAL FALSE DISCOVERY RATE FOR GENOME-WIDE ASSOCIATION STUDIES

BY RONG W. ZABLOCKI^{†,‡}, RICHARD A. LEVINE[†] ANDREW J. SCHORK[§] SHUJING XU[§] YUNPENG WANG[¶] CHUN C. FAN[§] AND WESLEY K. THOMPSON^{§,||,**,*}

*San Diego State University[†], Claremont Graduate University[‡], University of California at San Diego[§], University of Oslo, Norway[¶], Institute of Biological Psychiatry^{||}, and The Lundbeck Foundation Initiative for Integrative Psychiatric Research^{**}*

While genome-wide association studies (GWAS) have discovered thousands of risk loci for heritable disorders, so far even very large meta-analyses have recovered only a fraction of the heritability of most complex traits. Recent work utilizing variance components models has demonstrated that a larger fraction of the heritability of complex phenotypes is captured by the additive effects of SNPs than is evident only in loci surpassing genome-wide significance thresholds, typically set at a Bonferroni-inspired $p \leq 5 \times 10^{-8}$. Procedures that control false discovery rate can be more powerful, yet these are still under-powered to detect the majority of non-null effects from GWAS. The current work proposes a novel Bayesian semi-parametric two-group mixture model and develops a Markov Chain Monte Carlo (MCMC) algorithm for a covariate-modulated local false discovery rate (*cmfdr*). The probability of being non-null depends on a set of covariates via a logistic function, and the non-null distribution is approximated as a linear combination of B-spline densities, where the weight of each B-spline density depends on a multinomial function of the covariates. The proposed methods were motivated by work on a large meta-analysis of schizophrenia GWAS performed by the Psychiatric Genetics Consortium (PGC). We show that the new *cmfdr* model fits the PGC schizophrenia GWAS test statistics well, performing better than our previously proposed parametric gamma model for estimating the non-null density and substantially improving power over usual *fd*r. Using loci declared significant at *cmfdr* ≤ 0.20 , we perform follow-up pathway analyses using the Kyoto Encyclopedia of Genes and Genomes (KEGG) *homo sapiens* pathways database. We demonstrate that the increased yield from the *cmfdr* model results in an improved ability to test for pathways associated with schizophrenia compared to using those SNPs selected according to usual *fd*r.

*To whom correspondence should be addressed.

Keywords and phrases: Bayesian mixture model, B-spline densities, Genome-wide association study, multiple-comparison procedures, mixture of experts

1. Introduction. While genome-wide association studies (GWAS) have discovered thousands of risk loci for heritable disorders, so far even large meta-analyses have recovered only a fraction of the heritability of most complex traits. Recent work utilizing variance components models [Purcell et al. \(2009\)](#); [Yang et al. \(2010\)](#); [Davies et al. \(2011\)](#); [Yang et al. \(2015\)](#) has demonstrated that a much larger fraction of the heritability of complex phenotypes is captured by the additive effects of common variants than is evident only in loci surpassing genome-wide significance thresholds. Thus, the emerging picture is that traits such as these are highly polygenic, and that a large fraction of the heritability is accounted for by numerous loci each with a very small effect ([Glazier, Nadeau and Aitman, 2002](#)).

An example is given by the motivating application of this paper, a large meta-analysis of schizophrenia GWAS performed by the Psychiatric Genetics Consortium (PGC, www.med.unc.edu/pgc). Schizophrenia is a complex disorder with a heritability (total variability in liability of disease due to variability in genetic factors) estimated from family studies as high as 80%. The latest PGC analyses ([Psychiatric-Genomics-Consortium, 2014](#)) combined 82,315 subjects from 52 sub-studies to identify 108 independent regions (128 significant variants) that explained 3% of risk variability. Predictive models using liberally selected collections of thousands of variants not reaching the accepted significance in the PGC study explained as much as 18% of the variability in an independent sample ([Psychiatric-Genomics-Consortium, 2014](#)). Further, mixed models used to estimate the total variability in schizophrenia risk explained by all SNP variants tested in the PGC GWAS suggest that as much as 43% of the variability could, in theory, be explained by the collection of variants used for these studies ([Psychiatric-GWAS-Consortium, 2011](#)). Taken together these findings suggest that schizophrenia is highly polygenic, with many tiny genetic effects yet to be discovered by conventional statistical approaches and significance criteria, even using more liberal thresholds based on false discovery rate methods ([Benjamini and Hochberg, 1995](#); [Efron and Tibshirani, 2002](#)).

Methods for estimating and controlling false discovery rates typically treat all hypothesis tests as exchangeable, ignoring any auxiliary covariates that may influence the distribution of test statistics ([Benjamini and Hochberg, 1995](#); [Efron and Tibshirani, 2002](#)). For example, the *local false discovery rate* (fdr) ([Efron and Tibshirani, 2002](#)) rests on a simple two-groups mixture model for test statistic Z . Letting f_0 and f_1 be the probability density functions corresponding to null and non-null tests, respectively, the marginal pdf of Z is given by

$$(1.1) \quad f(z) = (1 - \pi_1)f_0(z) + \pi_1f_1(z),$$

where π_1 is the non-null proportion. The *fdr* is then defined as the posterior probability the test is null given the observed test statistic $Z = z$.

Covariate-modulated fdr (*cmfdr*) attempts to incorporate the effects of auxiliary covariates into *fdr* estimation. Ferkingstad et al. (2008) proposed a uniform-beta mixture model for f , first stratifying on levels of a scalar covariate x and then estimating the parameters of the mixture model within each stratum separately. Lewinger et al. (2007) proposed a noncentral χ^2 distribution for f_1 , where the prior proportion π_1 and the non-centrality parameter are linear combinations of the covariates, passed through non-linear link functions. Zablocki et al. (2014) proposed a gamma distribution for f_1 where covariates contribute not only to f_1 , but also to the prior probability of being non-null. Scott et al. (2015) developed f_1 as a location mixture of null (normal) density and only the prior probability depended on covariates.

These parametric approaches can be efficient if the model fit is adequate. However, the assumed parametric distributions may not always provide an adequate fit to the underlying true non-null distribution, in which case a more flexible nonparametric alternative is desirable to avoid biases in estimating the *cmfdr*. For example, we found that the gamma distribution underestimated the tails of f_1 in the PGC Schizophrenia GWAS test statistics, leading to elevated estimates of the *cmfdr*, and hence a loss of power for some loci. The current paper is an extension of Zablocki et al. (2014) to incorporate a more flexible model for the non-null density. We take a semi-parametric approach, modeling the mixture density f as a weighted combination of a normal null distribution with B-spline densities bounded away from zero. These non-negative weights are smooth functions of a vector of locus-specific covariates \mathbf{x} , and normalized to sum to unity. From this mixture model for the density f , we can compute a semi-parametric *cmfdr*, or posterior probability that a test is null given the observed test score z and vector of covariates \mathbf{x} . Model inference is performed via a Markov Chain Monte Carlo (MCMC) sampling algorithm.

Section 2 presents a two-group semi-parametric model for *cmfdr* incorporating covariates into the estimation of the non-null proportion and density. We describe the MCMC sampling algorithm in Supplementary 1. Section 3 presents Monte Carlo simulations and an application to the PGC Schizophrenia GWAS data. Here, we show large increases in power utilizing functional genomic annotations in the *cmfdr* model, compared with standard *fdr* and previous *cmfdr* methods. The increased yield of SNPs allows for a more powerful pathway analysis of SNPs surpassing a significance threshold of *cmfdr* ≤ 0.20 . Section 4 concludes with a brief discussion and future directions.

The R code for implementing the methods proposed in this paper may be found at https://github.com/rongw16/cmfd_r_semi-parametric_model.

2. Method.

2.1. *Covariate-modulated local false discovery rate.* We use as our starting point the simple two-group mixture model as specified by Eq. (1.1). Let Z_i be random variables, $i = 1, \dots, N$, where Z_i denotes the test statistic for the i^{th} test. We consider the scenario where for each Z_i we also have an $(M + 1)$ -dimensional vector of covariates (including intercept) denoted by $\mathbf{x}_i = (1, x_{1i}, x_{2i}, \dots, x_{Mi})^T$. The test statistics Z_i are assumed independent, with marginal density f conditional on \mathbf{x} given by

$$(2.1) \quad f(z|\mathbf{x}) = \pi_0(\mathbf{x})f_0(z|0, \sigma_0^2) + \pi_1(\mathbf{x})f_1(z|\mathbf{x}),$$

where $f_0(\cdot|0, \sigma_0^2)$ denotes a normal density with mean 0 and variance σ_0^2 and $\pi_0(\mathbf{x}) = 1 - \pi_1(\mathbf{x})$. The non-null prior probability π_1 and density f_1 depend on the auxiliary covariates \mathbf{x} as specified in Section 2.2.

We define the cmfdr as the posterior probability that the test is null given $Z = z$ and \mathbf{x} , which by Bayes' Rule is given by

$$(2.2) \quad \begin{aligned} \text{cmfdr}(z|\mathbf{x}) &= \frac{\pi_0(\mathbf{x})f_0(z)}{f(z|\mathbf{x})} \\ &= \frac{\pi_0(\mathbf{x})f_0(z)}{\pi_0(\mathbf{x})f_0(z) + \pi_1(\mathbf{x})f_1(z|\mathbf{x})}. \end{aligned}$$

The “zero assumption” of Efron (2007) states that tests with z -scores close to zero are primarily of null cases. This is required to ensure the non-null distribution is identifiable. As in Efron (2007), the default assumption in our applications is that any test with $|z| \leq 0.68$ (corresponding to the middle 50% of the standard normal distribution) is considered a null test, i.e., the non-null density $f_1(z) = 0$ for $|z| \leq 0.68$. Martin and Tokdar (2012) note that identifiability is not guaranteed for a two-group model with an empirical null involving an unknown variance parameter; however, since a theoretical (standard normal) null poorly describes the behavior of the null in many applications, an empirical null is often required (Scott et al., 2015; Efron, 2004). To solve the problem, Martin and Tokdar (2012) and Scott et al. (2015) impose a “tail assumption” on their models such that f_1 has heavier tails than f_0 , where f_0 is a normal distribution with unknown mean and variance and f_1 is a location mixture of f_0 . We show that our model is identifiable under the zero assumption and other mild conditions (Supplementary 3). In our application of the model to the PGC schizophrenia data, we run multiple chains (each with 23000 iterations) with different random initial values. Figure 1 to 6 in Supplementary 4 depicts convergence of the parameter estimates.

2.2. *Covariate-modulated mixture density.* We first introduce a global latent indicator vector $\boldsymbol{\delta} = (\delta_1, \dots, \delta_N)^T$, where $\delta_i = 1$ if the i^{th} test is non-null and zero otherwise, and N is the total number of tests. It is assumed that $\delta_i \stackrel{\text{ind}}{\sim} \text{Bernoulli}\{\pi_1(\mathbf{x}_i)\}$, where

$$(2.3) \quad \pi_1(\mathbf{x}_i) = P(\delta_i = 1 | \boldsymbol{\gamma}, \mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\gamma})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\gamma})},$$

and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2, \dots, \gamma_M)^T$ is an $(M + 1)$ -vector of unknown parameters. Let \mathbf{x} denote the $(M + 1) \times N$ covariate matrix with columns \mathbf{x}_i . Then the joint density of $\boldsymbol{\delta}$ given $\boldsymbol{\gamma}$ and annotations \mathbf{x} is given by

$$(2.4) \quad f_{\boldsymbol{\delta}}(\boldsymbol{\delta} | \mathbf{x}, \boldsymbol{\gamma}) = \prod_{i=1}^N \left[\left\{ \frac{\exp(\mathbf{x}_i^T \boldsymbol{\gamma})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\gamma})} \right\}^{\delta_i} \left\{ \frac{1}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\gamma})} \right\}^{1 - \delta_i} \right].$$

The marginal density of Z given by Eq. (2.1) is a mixture of a null density f_0 and a non-null density f_1 , each symmetric around zero. Note, the assumption that f_0 and f_1 are symmetric around zero is appropriate for the GWAS example presented here, but could easily be relaxed for other applications. We also assume that z scores from null tests are independent and normally distributed with mean zero, that is $Z_i | \delta_i = 0 \sim N(0, \sigma_0^2)$. Thus, the likelihood of the null tests is given by

$$(2.5) \quad f_0(\mathbf{z}_0 | \boldsymbol{\delta}, \sigma_0^2) = (2\pi\sigma_0^2)^{-\frac{N_0}{2}} \exp \left\{ \frac{-\mathbf{z}_0^T \mathbf{z}_0}{2\sigma_0^2} \right\},$$

where $N_0 = N - \boldsymbol{\delta}^T \boldsymbol{\delta}$ is the number of tests for which $\delta_i = 0$ and \mathbf{z}_0 denotes the corresponding N_0 -dimensional vector of z -scores. The parameter σ_0^2 is unknown and estimated from the data (the “empirical null”).

The non-null density f_1 is approximated by a finite mixture of B-spline densities (B-splines normalized to integrate to unity, [Lopes and Dias \(2012\)](#)) with weights that vary smoothly as a function of covariates. B-splines are basis functions having compact support, defined by their polynomial degree and the number and placement of knots ([Eilers and Marx, 1996](#)). In the remainder of the paper, we use cubic B-spline densities with knots of multiplicity one fixed by the user, leading to piecewise cubic models with continuous first and second derivatives. Rather than focus on knot selection, the strategy here is to include enough knots to allow a flexible fit and to estimate variance parameters that control the smoothness of the fit ([Ruppert, 2002](#); [Thompson and Rosen, 2008](#)).

Specifically, the likelihood of the non-null cases is given by

$$(2.6) \quad f_1(\mathbf{z}_1 | \mathbf{x}_1, \boldsymbol{\delta}, \boldsymbol{\alpha}) = \prod_{i:\delta_i=1} \left\{ \sum_{k=1}^K c_{ki} g_k(z_i) \right\},$$

where \mathbf{z}_1 is the vector of z -scores corresponding to non-null tests of dimension; let $N_1 = \boldsymbol{\delta}^T \boldsymbol{\delta}$ and \mathbf{x}_1 is the corresponding $(M+1) \times N_1$ matrix of annotations. The g_k are cubic B-spline densities and the

$$c_{ki} = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\alpha}_k)}{\sum_{l=1}^K \exp(\mathbf{x}_i^T \boldsymbol{\alpha}_l)}$$

are non-negative weights so that $\sum_{k=1}^K c_{ki} = 1$. Coefficient c_{ki} is the probability that the i^{th} test belongs to the k^{th} B-spline component, given $\delta_i = 1$ and covariates \mathbf{x}_i . These coefficients depend on an $(M+1) \times K$ unknown parameter matrix

$$\boldsymbol{\alpha}_{(M+1) \times K} = \begin{bmatrix} \alpha_{01} = 0 & \alpha_{02} & \alpha_{03} & \cdots & \alpha_{0K} \\ \alpha_{11} = 0 & \alpha_{12} & \alpha_{13} & \cdots & \alpha_{1K} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \alpha_{M1} = 0 & \alpha_{M2} & \alpha_{M3} & \cdots & \alpha_{MK} \end{bmatrix},$$

where the k^{th} column $\boldsymbol{\alpha}_k$ corresponds to the k^{th} B-spline component and $\boldsymbol{\alpha}_m$ denotes the row corresponding to the m^{th} covariate (including intercept), $m = 0, 1, 2, \dots, M$. For identifiability, the first column $\boldsymbol{\alpha}_1 = \mathbf{0}$.

We also introduce a local indicator vector $\boldsymbol{\eta} = (\eta_1, \eta_2, \dots, \eta_{N_1})^T$. The element $\eta_i \in \{1, \dots, K\}$ specifies the B-spline component from which the i^{th} non-null test statistic z_i is generated. The $\eta_i \stackrel{\text{ind}}{\sim} \text{Multinomial}(\mathbf{c}_i)$, where $\mathbf{c}_i = (c_{1i}, \dots, c_{Ki})^T$. The joint density of $\boldsymbol{\eta}$ given $\boldsymbol{\delta}$, $\boldsymbol{\alpha}$, and \mathbf{x}_1 is given by

$$(2.7) \quad \begin{aligned} f_{\boldsymbol{\eta}}(\boldsymbol{\eta} | \boldsymbol{\delta}, \boldsymbol{\alpha}, \mathbf{x}_1) &= \prod_{i:\delta_i=1} \prod_{k=1}^K \left\{ P_{\boldsymbol{\eta}}(\eta_i = k | \mathbf{x}_i)^{I(\eta_i=k)} \right\} \\ &= \prod_{i:\delta_i=1} \prod_{k=1}^K \left[\left\{ \frac{\exp(\mathbf{x}_i^T \boldsymbol{\alpha}_k)}{\sum_{l=1}^K \exp(\mathbf{x}_i^T \boldsymbol{\alpha}_l)} \right\}^{I(\eta_i=k)} \right]. \end{aligned}$$

In summary, at the global level, the covariates modulate the probability of the null and non-null status of each test. At the local level (within the non-null distribution), the covariates modulate the B-spline component assignment probability for each non-null test.

2.2.1. *Prior distributions.* We specify prior distributions for parameters σ_0^2 , γ , and α . The rows of α are assumed independent. Based on [Eilers and Marx \(1996\)](#), [Lang and Brezger \(2004\)](#), [Chib and Jeliazkov \(2006\)](#), and [Rosen and Thompson \(2015\)](#) we propose the following prior distribution for rows $\alpha_{m\cdot}$. Let

$$\begin{bmatrix} \alpha_{m2} \\ \alpha_{m3} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} c\tau_m^2 & 0 \\ 0 & c\tau_m^2 \end{bmatrix} \right), \quad m = 0, \dots, M,$$

where c is a fixed constant and $\tau^2 = (\tau_0^2, \tau_1^2, \dots, \tau_M^2)^T$ is a $(M+1)$ -vector hyperparameter. In our test runs, $c=10, 100$, or 1000 give similar results; hence $c=100$ is taken in the implementation. The remaining α_{mk} , $k = 4, 5, \dots, K$, are assumed normally distributed with mean $2\alpha_{m(k-1)} - \alpha_{m(k-2)}$ and variance τ_m^2 . The prior distribution on $\alpha_{m\cdot}$ may be expressed in the more compact form as

$$P(\alpha_{m\cdot} | \tau_m^2) \propto (\tau_m^2)^{-\frac{K-1}{2}} \exp \left\{ -\frac{1}{2\tau_m^2} \alpha_{m(2:K)} \mathbf{\Omega}^* \alpha_{m(2:K)}^T \right\},$$

where $\alpha_{m(2:K)}$ is a $(K-1)$ -vector of B-spline components for the m^{th} covariate and $\mathbf{\Omega}^*$ is a $(K-1) \times (K-1)$ matrix defined as follows. Let

$$\mathbf{D}_{(K-3) \times (K-1)} = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \cdots & 1 & -2 & 1 \end{bmatrix}$$

and $\mathbf{\Omega} = \mathbf{D}^T \mathbf{D}$. We define $\mathbf{\Omega}^* = \mathbf{\Omega}$, except for $\Omega_{(1,1)}^* = \Omega_{(1,1)} + c^{-1}$ and $\Omega_{(2,2)}^* = \Omega_{(2,2)} + c^{-1}$ to ensure that the matrix $\mathbf{\Omega}^*$ is positive definite.

We propose Inverse Gamma prior for each τ_m^2 based on [Wand et al. \(2011\)](#), [Gelman et al. \(2006\)](#) and [Rosen and Thompson \(2015\)](#),

$$\begin{aligned} \tau_m^2 | a_m &\sim \text{Inverse Gamma} \left(\frac{\nu}{2}, \frac{\nu}{a_m} \right), \\ a_m &\sim \text{Inverse Gamma} \left(\frac{1}{2}, \frac{1}{A^2} \right), \\ m &= 0, \dots, M, \end{aligned}$$

where $\mathbf{a} = (a_0, a_1, \dots, a_M)^T$ is a $(M+1)$ -vector hyperparameter and a_m follows an Inverse Gamma distribution. Hyper-parameters ν and A are assumed known; in our experience, values of ν , 10 or 20, and values of A , 10 or 10,000, yield similar results, as observed in [Rosen and Thompson \(2015\)](#). Therefore,

we take $\nu = 10$ and $A = 10$ in the implementation. The kernel probability functions of τ_m^2 and a_m take the following forms:

$$P_{\tau_m^2}(\tau_m^2 | a_m) \propto (\tau_m^2)^{(-\frac{\nu}{2}-1)} a_m^{-\frac{\nu}{2}} \exp\left(-\frac{\nu}{\tau_m^2}\right)$$
$$P_{a_m}(a_m) \propto a_m^{-\frac{1}{2}-1} \exp\left(-\frac{1}{a_m}\right)$$

To complete the model, we assume weakly informative priors on the unknown parameters γ and σ_0^2 :

- $\gamma \sim N(\mathbf{0}, \Sigma_\gamma)$
- $\sigma_0^2 \sim \text{Inverse Gamma}(a_{\sigma_0}, b_{\sigma_0})$

where hyperparameters Σ_γ , a_{σ_0} , and b_{σ_0} are fixed by the user. In the simulations and data application, we set Σ_γ to be diagonal with variance 10,000 and $(a_{\sigma_0}, b_{\sigma_0}) = (0.001, 0.001)$. Conditional posterior distributions and the MCMC sampling algorithm are described in Supplementary 1.

3. Results.

3.1. Simulation study. In these simulation studies, we set the minimum non-null $|z|$ -score at 1.96, 0.68, and 0.25 to represent high, medium, and low power scenarios (corresponding to the central 95th, 50th, and 20th percentiles of a standard normal distribution, respectively). We set $\gamma = (-5.29, 2.5, -1.5)^T$, $\gamma = (-3.74, 1.2, -1)^T$, and $\gamma = (-3.06, 0.5, -0.2)^T$ to represent large, medium, and small effects, respectively. These choices for γ_0 set the true non-null proportion in all simulations around 5%. The variance parameter $\sigma_0^2 = 1.2$. The values for α , τ^2 and \mathbf{a} are drawn from their respective distributions as described in Section 2.2.1.

Each of the nine combinations of power scenarios and covariate effects includes 100 datasets, each dataset includes $N=50,000$ hypothesis tests where $K = 5$. Two covariates are generated, with x_1 binomial and x_2 standard normal random variables. We compare the proposed cmfdr model to an intercept only model, which is functionally equivalent to the fdr given in Efron (2007). For each setting each dataset, the MCMC algorithm was run for 18,000 iterations with 1,400 retained samples.

Table 1 presents the median values of sensitivity, specificity, false discovery proportion (FDP, defined as the proportion of incorrectly identified non-null nodes) and number of the non-null cases identified, as well as corresponding 95% credible intervals from 100 runs. Significance cutoffs for both fdr and cmfdr are set to 0.05. Specificity is consistently high and FDP is

consistently low across all conditions. Sensitivity and the number of identified non-null cases are consistently higher in cmfdr comparing with fdr (horizontal comparisons) across all conditions. Increased sensitivity is more pronounced with low and medium power regardless of covariate effects. For example, at high power large covariate effect scenario, sensitivity increases 6.9% and 195 more non-null cases are identified by cmfdr comparing to fdr; where as for the medium power/large covariate effect scenario, sensitivity is increased by 14.4% and 400 more non-null cases are identified by cmfdr. These results suggest that in the high power scenario, the null and non-null distributions tend to be naturally separated, the covariate effects may become less important.

3.2. Schizophrenia GWAS Application. For this study we used publicly available (<https://www.med.unc.edu/pgc/downloads>) results from the PGC Schizophrenia GWAS meta-analysis (Psychiatric-Genomics-Consortium, 2014). These data consist of summary statistics for 9,279,485 SNP variants. For each SNP variant independently, a fixed effects meta-analysis was performed across the results of 52 sub-studies. Each sub-study used a logistic regression to test the count of one of the two variant alleles (0, 1 or 2) for association with schizophrenia (as a case-control outcome), adjusted for nuisance covariates.

The allele counts of variants in close proximity on the genome are correlated (termed Linkage Disequilibrium or LD, Reich et al. (2001)) with the dependence falling off approximately exponentially with distance, although at variable rates across the genome. As a result, the test statistics from a GWAS are not independent and have a variable width, approximately block diagonal correlation structure. To obtain an approximately independent subset of test statistics, we compute the pairwise squared correlation coefficient (r^2) between allele counts for all pairs of SNPs within a conservatively large window of 1,000,000 base pairs. Genotype data for the PGC study were not available, so correlations were estimated in an independent, but representative, collection of European individuals sequenced as part of the 1000 genomes project (Genomes-Project-Consortium, 2012). To facilitate follow-up pathway analyses, we assigned SNPs with gene annotations corresponding to genes within 50,000 base pairs for that given SNP. Genes were selected based on the 242 Kyoto Encyclopedia of Genes and Genomes (KEGG) *homo sapiens* pathways (Kanehisa and Goto, 2000; Kanehisa et al., 2016) SNPs within the major histocompatibility complex (MHC) on chromosome 6 were removed due to the extensive and complex correlation structure within the region. The resulting test statistics were then randomly pruned

TABLE 1
Performance comparison between semi-parametric cmfdr and fdr under different conditions.

Simulation ¹			cmfdr	fdr	
N=50000					
100 datasets/setting					
High power ² , Large effect ³	Sensitivity (%)	87.8	[86.3, 88.9]	80.9 [79.1, 82.2]	
	Specificity (%)	99.9	[99.9, 99.9]	99.9 [99.9, 99.9]	
	FDP ⁴ (%)	0.2	[0.1, 0.4]	0.4 [0.3, 0.5]	
	Number of non-null identified	2479	[2376, 2568]	2284 [2193, 2378]	
	Medium effect	Sensitivity (%)	85.9	[84.6, 87.6]	80.7 [79.2, 82.4]
		Specificity (%)	99.9	[99.9, 99.9]	99.9 [99.9, 100]
		FDP (%)	0.2	[0.1, 0.4]	0.4 [0.2, 0.4]
		Number of non-null identified	2397	[2302, 2468]	2256 [2162, 2327]
	Small effect	Sensitivity (%)	86.6	[85.1, 87.8]	83.4 [82.1, 84.6]
		Specificity (%)	99.9	[99.9, 100]	99.9 [99.9, 99.9]
		FDP (%)	0.2	[0.1, 0.4]	0.2 [0.1, 0.3]
		Number of non-null identified	2178	[2113, 2226]	2099 [2035, 2152]
Medium power, Large effect	Sensitivity (%)	64.1	[62.0, 65.7]	49.6 [47.4, 51.4]	
	Specificity (%)	99.9	[99.9, 99.9]	99.9 [99.9, 99.9]	
	FDP (%)	0.6	[0.3, 1.0]	0.7 [0.6, 0.9]	
	Number of non-null identified	1816	[1737, 1878]	1407 [1340, 1482]	
	Medium effect	Sensitivity (%)	61.7	[60.0, 63.1]	51.4 [49.7, 53.8]
		Specificity (%)	99.9	[99.9, 99.9]	99.9 [99.9, 99.9]
		FDP (%)	0.5	[0.3, 0.8]	0.6 [0.4, 0.7]
		Number of non-null identified	1724	[1649, 1802]	1446 [1365, 1532]
	Small effect	Sensitivity (%)	65.3	[63.2, 67.7]	60.8 [58.9, 62.9]
		Specificity (%)	99.9	[99.9, 99.9]	99.9 [99.9, 99.9]
		FDP (%)	0.4	[0.4, 1.2]	0.4 [0.3, 0.5]
		Number of non-null identified	1649	[1580, 1723]	1534 [1471, 1604]
Low power, Large effect	Sensitivity (%)	52.8	[50.9, 54.8]	37.7 [35.8, 40.3]	
	Specificity (%)	99.9	[99.9, 100]	99.9 [99.9, 99.9]	
	FDP (%)	0.7	[0.2, 0.6]	1.0 [0.7, 1.1]	
	Number of non-null identified	1495	[1415, 1583]	1069 [1001, 1154]	
	Medium effect	Sensitivity (%)	51.4	[49.3, 53.3]	41.4 [39.1, 43.4]
		Specificity (%)	99.9	[99.9, 99.9]	99.9 [99.9, 99.9]
		FDP (%)	0.6	[0.3, 1.0]	0.8 [0.6, 0.9]
		Number of non-null identified	1442	[1361, 1511]	1157 [1084, 1223]
	Small effect π_1 (%): 5.0	Sensitivity (%)	57.7	[55.8, 60.2]	52.9 [51.5, 55.1]
		Specificity (%)	99.9	[99.9, 99.9]	99.9 [99.9, 99.9]
		FDP (%)	0.4	[0.1, 0.7]	0.3 [0.2, 0.5]
		Number of non-null identified	1453	[1400, 1501]	1332 [1287, 1392]

¹True non-null proportion in all simulations is controlled around 5% under sample size 50000.

²Simulated data with minimum non-null $|Z|$ score preset at 1.96, 0.68, and 0.25 to represent high, medium, and low power scenarios. ³Covariates modulate the probability of being non-null via parameter γ ; preset $\gamma = (-5.29, 2.5, -1.5)^T$ to represent large effect, $\gamma = (-3.74, 1.2, -1)^T$ to represent medium effect, and $\gamma = (-3.06, 0.5, -0.2)^T$ to represent small effect. ⁴FDP: False discovery proportion. All results presented are the median and [95% credible interval] over 100 runs for each setting. Cutoff for cmfdr and fdr is set to be 0.05.

for approximate independence, such that the estimated squared correlation coefficient r^2 was less than 0.2 for any pair of SNPs. In order to approximate the maximum independent set of those SNPs (to minimize the information loss due to the pruning) our pruning scheme is based on a greedy algorithm which in each step keeps a node with the minimum number of neighbors in a complete graph. The final data are composed of $N = 74,800$ SNP summary statistics (z -scores) on $n = 82,315$ subjects (35,476 cases). The meta-analysis z -scores of the 52 sub-studies are calculated based Willer, Li and Abecasis (2010) and converted to z -test statistics using the inverse (standard normal) probability transform.

For each SNP, we also computed three covariates: (1) the Total LD score (TotLD), which is the sum of the squared correlation coefficients between a given SNP and all others within a 1,000,000 base pairs window, again computed in the representative 1000 genomes sample, a measure of the size of the correlation block the SNP resides in; (2) heterozygosity (H), which is the variance of the allele count, or $H = 2(p)(1 - p)$, where p is the frequency of the reference allele; (3) the Total Protein Coding Gene LD score (Protein-Coding), which sums the squared correlation coefficients between a given SNP and all others within a 1,000,000 base pairs window that are in a protein coding gene as annotated on the reference genome (Hsu et al., 2006), a rough measure of the functional DNA within a SNP's correlation block. We have previously shown that these three covariates enrich for non-null SNP associations across a broad range of complex phenotypes (Schork et al., 2013). The distributions of TotLD and ProteinCoding are highly skewed and thus were log-transformed. All three covariates were then standardized to have mean zero and standard deviation one. The MCMC algorithm was applied with $K=5$. Parameter estimates for γ indicate that all three covariates are positively associated with the prior probability of non-null status in semi-parametric cmfdr, where coefficient for TotLD is 0.73, 95% credible interval is [0.61, 0.86]; H: 0.31 [0.24, 0.38] and ProtenCoding: 0.29 [0.22, 0.37]. The positive association are also observed in gamma cmfdr (Zablocki et al., 2014) as well as in FDRreg cmfdr (Scott et al., 2015).

Power to detect non-null SNPs in different models is displayed in Figure 1. This figure compares the number of non-null SNPs rejected under different models as a function of significance threshold. The increase in power for both the gamma and semi-parametric cmfdr approaches compared to fdr, across a range of cut-offs from 0.001 to 0.20, is large. For example, for cut-off 0.20, fdr rejects 175 null hypotheses, semi-parametric cmfdr with all three covariates rejects 588, gamma model cmfdr rejects 368, and FDRreg cmfdr rejects 203. For reference, the commonly-used GWAS threshold of $p \leq 5 \times 10^{-8}$ rejects

12

111 null hypotheses.

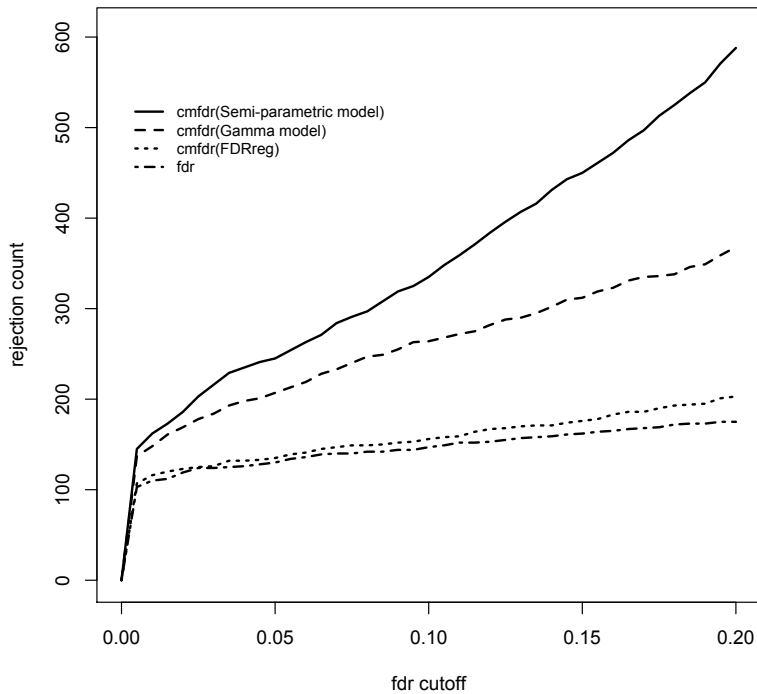


Fig 1: Power curve for fdr, cmfdr (FDRreg), cmfdr (gamma model) and cmfdr (semi-parametric model). The x -axis is the fdr cutoff required to declare a SNP significant. The y -axis is the number of rejected SNPs.

We also investigate the model fits, comparing semi-parametric cmfdr and the parametric gamma cmfdr. Figure 2 presents stratified Q-Q plots by π_0 quantiles. This figure displays the $-\log_{10}$ observed p -values vs. the theoretical $-\log_{10}$ p -values under a standard normal distribution. Each SNP has been assigned to one of three strata based on $\pi_0(\mathbf{x}_i) = 1 - \pi_1(\mathbf{x}_i)$ value by quantiles: $[0.00, 0.33]$, $(0.33, 0.66]$, and $(0.66, 1.00]$. The predicted $-\log_{10}$ p -values estimated from the models are shown with a solid line, dashed line, and dotted line, respectively; the observed $-\log_{10}$ p -values are shown with dots, triangles, and stars. SNPs in the stratum $\pi_0 : [0 - 33]$ have the highest likelihood of being non-null, while SNPs in the stratum $\pi_0 : (66 - 100]$ have the highest probability of being null. The gray dash-dot line indicates where

the Q-Q curve would lie if all SNPs were null under a standard normal distribution. The leftward deflection of the $-\log_{10} p$ -values on the Q-Q plots stratified by π_0 quantiles implies an abundance of non-null SNPs versus the global null hypothesis. The semi-parametric cmfdr displays the best model fit compared to the data. Of the 588 SNPs rejected by the semi-parametric cmfdr at the 0.2 cutoff, 578 are from the stratum $\pi_0 : [0 - 33]$, 9 from the stratum $\pi_0 : (33 - 66]$, and only one from the stratum $\pi_0 : (66 - 100]$. Analogously, of the 368 SNPs rejected by the gamma cmfdr at the 0.2 cutoff, the numbers of SNPs in corresponding strata are 364, 4, and 0, respectively.

Furthermore, we plot the semi-parametric cmfdr (Figure 3a) and gamma cmfdr (Figure 3b) versus the observed absolute z -scores stratified by quantiles of $\pi_0(\mathbf{x}_i)$; fdr is also added as a reference. The gray dotted line is the 0.2 cutoff. For the most enriched sample, the minimum absolute z -scores with semi-parametric cmfdr ≤ 0.2 is 2.25 and with gamma cmfdr ≤ 0.2 is 2.57. For fdr, the minimum absolute z -score under this threshold is 4.46, further demonstrating the increase in power from using cmfdr vs. fdr.

Finally, we compare the non-null densities of semi-parametric (Figure 4a) and gamma (Figure 4b) covariate-modulated mixture models with different values of covariates. The model without covariates is also included (solid lines). Both figures show the non-null densities where all the covariates were set at their corresponding 33 (dash line), 66 (dot line) and 99 (dash-dot line) percentiles. With increasing values for the covariates, the densities show progressively heavier tails. The non-null density of the model without covariates shifts to the right, as compared to the semi-parametric model with covariates in Figure 4a. This shift is probably due to the fact that the variance of the null density (σ_0^2) is larger in the model without covariates (median: 1.31, 95% credible interval: 1.29 - 1.33) than the model with covariates (median: 1.12, 95% credible interval: 1.09 - 1.15). The shift also appears in Figure 4b where the median of σ_0^2 from the gamma model is 1.24 (95% credible interval: 1.22 - 1.26). These results collectively indicate that the enrichment annotation categories we employ here (TotLD, H, and ProteinCoding) provide useful information for selecting “interesting” subsets of SNPs for further analysis.

To examine the biological significance of the SNPs, we performed pathway analyses on the 242 gene sets in the KEGG *homo sapiens* pathways database (<http://www.kegg.jp/>). To perform these pathway analyses, we implemented the ALIGATOR (Holmans et al., 2009) algorithm, which tests for overrepresentation of biological pathways in SNP lists. ALIGATOR corrects for LD between SNPs, variable gene size, and multiple testing of non-independent pathways. Using the 175 SNPs with fdr ≤ 0.20 results in no pathways with p -value ≤ 0.05 (corrected for multiple testing). On the other

hand, there were 10 pathways with p -values ≤ 0.05 using 588 SNPs with semi-parametric cmfdr ≤ 0.20 (Table 2). The p -values using 368 SNPs with gamma cmfdr are also listed for comparison. Axon Guidance is ranked highest in both cmfdr models. The 10 top ranked pathways from semi-parametric cmfdr given in Table 2 provide interesting insight into the pathogenesis of schizophrenia, given that the KEGG database is expertly curated without prior emphasis in terms of disease etiology. The top ranked pathways show abnormal axonal connectivity, lipid metabolizing, and voltage-gated ion channels, as well as comorbid conditions that have been noted among patients with schizophrenia in prior research (Greiner and Nicolson, 1965; Lidow, 2003; Battaglino et al., 2004; Leucht et al., 2007; Putnam, Sun and Zhao, 2011; Maiti et al., 2011; Buckley, Pillai and Howell, 2011; Gardiner et al., 2012; Liu et al., 2013). A complete list of the 242 KEGG *homo sapien* pathways and their ALIGATOR p -values are given in Supplementary 2.

TABLE 2
KEGG PATHWAY with ALIGATOR p-values from three models

Pathway	semi-parametric cmfdr ¹	gamma cmfdr ²	fdr ³
Axon guidance	0.0006	0.002	0.2046
Herpes simplex infection	0.0008	0.027	1
Osteoclast differentiation	0.0062	0.019	1
Pentose phosphate pathway	0.0096	0.521	1
Tuberculosis	0.01	0.0068	0.132
Leishmaniasis	0.0162	0.095	1
Antigen processing and presentation	0.022	0.096	1
Taste transduction	0.033	1	1
Cytokine-cytokine receptor interaction	0.037	0.038	1
Cell adhesion molecules (CAMs)	0.0446	0.131	1

¹ p values from ALIGATOR based on 588 non-nulls identified by the semi-parametric model at cmfdr cutoff 0.2; ² p values from ALIGATOR based on 368 non-nulls identified by the gamma model at cmfdr cutoff 0.2; ³ p values from ALIGATOR based on 175 non-nulls identified without covariates at fdr cutoff 0.2.

4. Discussion. GWAS of highly polygenic traits such as schizophrenia remain underpowered to detect most genetic variants involved in the disorder, even with very large sample sizes. By incorporating auxiliary information, the process of gene discovery can be sped up significantly, along with the assessment of the role of molecular pathways. Moreover, the examination of which auxiliary information is useful for predicting non-null status can be informative of the genetic architecture of polygenic traits.

Using a set of genetic loci (SNPs) pruned for approximate independence, we demonstrate a large increase in power in the PGC schizophrenia data using our semi-parametric cmfdr model compared with fdr, as well as previous

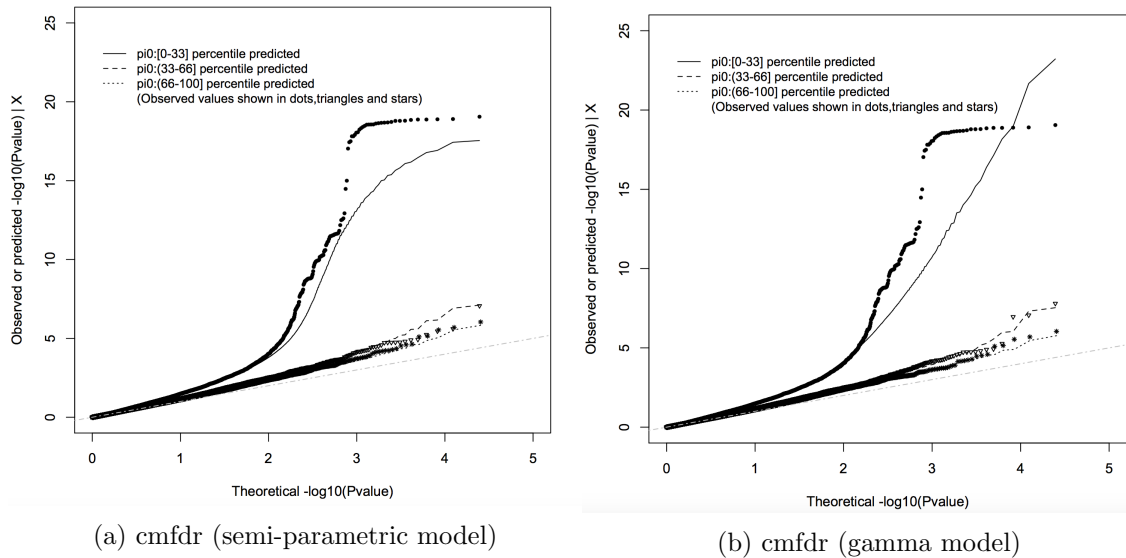


Fig 2: Q-Q plot by π_0 quantile for the PGC Schizophrenia GWAS data. The x -axis is the theoretical $-\log_{10} p$ -values under a standard normal distribution. The y -axis is the $-\log_{10}$ observed or predicted p -value (converted from z -scores). The gray dash-dot line is the reference line indicating where the $-\log_{10} p$ -values would lie if all SNPs were null under a standard normal distribution.

16

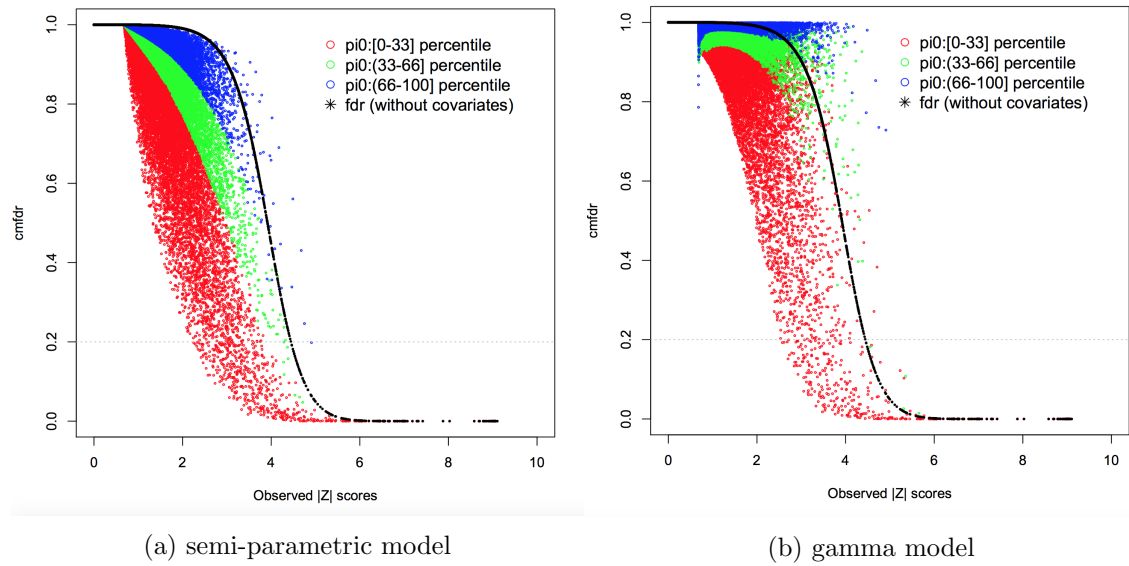


Fig 3: cmfdr and fdr plotted against observed absolute z -scores.

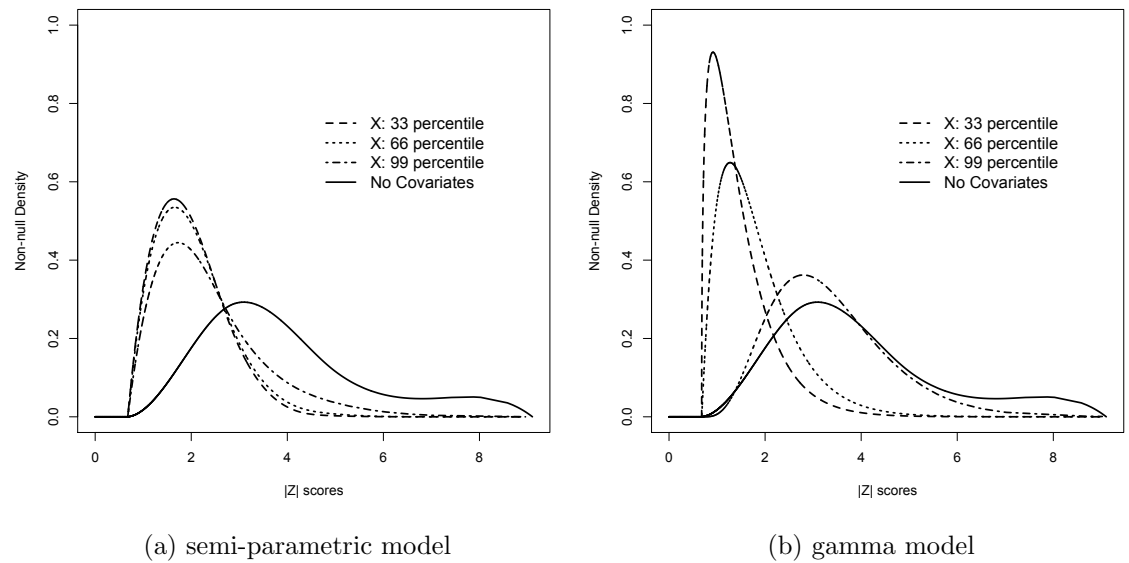


Fig 4: Non-null densities where all three covariates were set at their corresponding 33, 66 and 99 percentiles.

models for cmfdr that either use a parametric model (gamma fdr) or a model that does not incorporate covariate effects (FDRreg) into the estimation of the non-null density. For example, using a 0.20 cut-off, we reject 588 null hypotheses with cmfdr compared with only 175 using fdr, or over 3.4 times as many SNPs as the intercept only model, with a similarly large increase in power vs. FDRreg, and a smaller but still substantial increase in power over gamma fdr. This increase in power appears to be driven by a better-fitting model of the tails of the non-null distribution for highly enriched SNPs.

Our choice of covariates in the PGC schizophrenia application was driven by scientific considerations based on theory and substantial prior evidence that these annotations enriched for non-null associations (Schork et al., 2013). In general, we recommend selection of covariates based on these criteria. However, the model could also be used for exploratory analyses, to examine whether a given annotation significantly enriches for associations. For this use, it would be useful to implement a model-selection metric such as the Watanabe-Akaike Information Criterion (WAIC, Vehtari and Gelman (2014)).

The proposed cmfdr model assumes independence of the z -scores. To ensure this was approximately true in the current data example, we randomly pruned SNPs so that no two SNPs in the sample were correlated at more than $r^2 = 0.20$. We thus need to delete many tests to achieve independence. Our current research considers alternative schemes to explicitly model the effects of the correlation on the values of the z -scores. We are also developing an extension of the cmfdr model that also incorporates biological networks (gene sets with graphical model structure determined by biological interactions).

Acknowledgements. The authors are grateful to the anonymous Associate Editor and two referees for their insightful reviews and comments, which greatly improved the paper. The study is supported by NIH grant R01GM104400.

SUPPLEMENTARY MATERIAL

Supplementary materials for “Semi-parametric covariate-modulated local false discovery rate for genome-wide association studies”:

(). The supplement consists of 4 sections. Section 1 presents conditional posteriors and Gibbs sampling algorithm. Section 2 provides the full list of KEGG *homo sapiens* pathways with ALIGATOR p -values from different models. Section 3 demonstrates identifiability of the mixture model. Section 4 shows convergence diagnosis plots of parameter estimates.

References.

- BATTAGLINO, R., FU, J., SPÄTE, U., ERSOY, U., JOE, M., SEDAGHAT, L. and STASHENKO, P. (2004). Serotonin regulates osteoclast differentiation through its transporter. *Journal of Bone and Mineral Research* **19** 1420–1431.
- BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 289–300.
- BUCKLEY, P. F., PILLAI, A. and HOWELL, K. R. (2011). Brain-derived neurotrophic factor: findings in schizophrenia. *Current opinion in psychiatry* **24** 122–127.
- CHIB, S. and JELIAZKOV, I. (2006). Inference in semiparametric dynamic models for binary longitudinal data. *Journal of the American Statistical Association* **101**.
- DAVIES, G., TENESA, A., PAYTON, A., YANG, J., HARRIS, S. E., LIEWALD, D., KE, X., LE HELLARD, S., CHRISTOFOROU, A., LUCIANO, M. et al. (2011). Genome-wide association studies establish that human intelligence is highly heritable and polygenic. *Molecular psychiatry* **16** 996–1005.
- EFRON, B. (2004). Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis. *Journal of the American Statistical Association* **99** 465.
- EFRON, B. (2007). Size, power and false discovery rates. *The Annals of Statistics* 1351–1377.
- EFRON, B. and TIBSHIRANI, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genetic epidemiology* **23** 70–86.
- EILERS, P. H. and MARX, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical science* 89–102.
- FERKINGSTAD, E., FRIGESSI, A., RUE, H., THORLEIFSSON, G. and KONG, A. (2008). Unsupervised empirical Bayesian multiple testing with external covariates. *The Annals of Applied Statistics* 714–735.
- GARDINER, E., BEVERIDGE, N., WU, J., CARR, V., SCOTT, R., TOONEY, P. and CAIRNS, M. (2012). Imprinted DLK1-DIO3 region of 14q32 defines a schizophrenia-associated miRNA signature in peripheral blood mononuclear cells. *Molecular psychiatry* **17** 827–840.
- GELMAN, A. et al. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis* **1** 515–534.
- GENOMES-PROJECT-CONSORTIUM (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* **491** 56–65.
- GLAZIER, A. M., NADEAU, J. H. and AITMAN, T. J. (2002). Finding genes that underlie complex traits. *Science* **298** 2345–2349.
- GREINER, A. and NICOLSON, G. (1965). Schizophrenia-melanosis. *The Lancet* **286** 1165–1167.
- HOLMANS, P., GREEN, E. K., PAHWA, J. S., FERREIRA, M. A., PURCELL, S. M., SKLAR, P., OWEN, M. J., O'DONOVAN, M. C., CRADDOCK, N., CONSORTIUM, W. T. C.-C. et al. (2009). Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *The American Journal of Human Genetics* **85** 13–24.
- HSU, F., KENT, W. J., CLAWSON, H., KUHN, R. M., DIEKHANS, M. and HAUSSLER, D. (2006). The UCSC known genes. *Bioinformatics* **22** 1036–1046.
- KANEHISA, M. and GOTO, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28** 27–30.
- KANEHISA, M., SATO, Y., KAWASHIMA, M., FURUMICHI, M. and TANABE, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic acids research*

44 D457–D462.

- LANG, S. and BREZGER, A. (2004). Bayesian P-splines. *Journal of computational and graphical statistics* **13** 183–212.
- LEUCHT, S., BURKARD, T., HENDERSON, J., MAJ, M. and SARTORIUS, N. (2007). Physical illness and schizophrenia: a review of the literature. *Acta Psychiatrica Scandinavica* **116** 317–333.
- LEWINGER, J. P., CONTI, D. V., BAURLEY, J. W., TRICHE, T. J. and THOMAS, D. C. (2007). Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. *Genetic epidemiology* **31** 871–882.
- LIDOW, M. S. (2003). Calcium signaling dysfunction in schizophrenia: a unifying approach. *Brain research reviews* **43** 70–84.
- LIU, Y., LI, Z., ZHANG, M., DENG, Y., YI, Z. and SHI, T. (2013). Exploring the pathogenetic association between schizophrenia and type 2 diabetes mellitus diseases based on pathway analysis. *BMC medical genomics* **6** 1.
- LOPES, H. F. and DIAS, R. (2012). Bayesian mixture of parametric and nonparametric density estimation: A Misspecification Problem. *Brazilian Review of Econometrics* **31** 19–44.
- MAITI, S., KUMAR, K. H. B. G., CASTELLANI, C. A., O'REILLY, R. and SINGH, S. M. (2011). Ontogenetic de novo copy number variations (CNVs) as a source of genetic individuality: studies on two families with MZD twins for schizophrenia. *PLoS One* **6** e17125.
- MARTIN, R. and TOKDAR, S. T. (2012). A nonparametric empirical Bayes framework for large-scale multiple testing. *Biostatistics* **13** 427–439.
- PSYCHIATRIC-GENOMICS-CONSORTIUM (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511** 421–427.
- PSYCHIATRIC-GWAS-CONSORTIUM (2011). Genome-wide association study identifies five new schizophrenia loci. *Nature genetics* **43** 969–976.
- PURCELL, S. M., WRAY, N. R., STONE, J. L., VISSCHER, P. M., O'DONOVAN, M. C., SULLIVAN, P. F., SKLAR, P., RUDERFER, D. M., MCQUILLIN, A., MORRIS, D. W. et al. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460** 748–752.
- PUTNAM, D. K., SUN, J. and ZHAO, Z. (2011). Exploring schizophrenia drug-gene interactions through molecular network and pathway modeling. In *AMIA Annu Symp Proc* **2011** 1127–1133.
- REICH, D. E., CARGILL, M., BOLK, S., IRELAND, J., SABETI, P. C., RICHTER, D. J., LAVERY, T., KOUYOUJIAN, R., FARHADIAN, S. F., WARD, R. et al. (2001). Linkage disequilibrium in the human genome. *Nature* **411** 199–204.
- ROSEN, O. and THOMPSON, W. K. (2015). Bayesian semiparametric copula estimation with application to psychiatric genetics. *Biometrical Journal* **57** 468–484.
- RUPPERT, D. (2002). Selecting the number of knots for penalized splines. *Journal of computational and graphical statistics* **11**.
- SCHORK, A. J., THOMPSON, W. K., PHAM, P., TORKAMANI, A., RODDEY, J. C., SULLIVAN, P. F., KELSOE, J. R., O'DONOVAN, M. C., FURBERG, H., SCHORK, N. J. et al. (2013). All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. *PLoS genetics* **9** e1003449.
- SCOTT, J. G., KELLY, R. C., SMITH, M. A., ZHOU, P. and KASS, R. E. (2015). False discovery rate regression: an application to neural synchrony detection in primary visual cortex. *Journal of the American Statistical Association* **110** 459–471.
- THOMPSON, W. K. and ROSEN, O. (2008). A Bayesian model for sparse functional data.

Biometrics **64** 54–63.

- VEHTARI, A. and GELMAN, A. (2014). WAIC and cross-validation in Stan. *Submitted*. <http://www.stat.columbia.edu/~gelman/research/unpublished/waic-stan.pdf> Accessed **27** 5.
- WAND, M. P., ORMEROD, J. T., PADOAN, S. A., FUHRWIRTH, R. et al. (2011). Mean field variational Bayes for elaborate distributions. *Bayesian Analysis* **6** 847–900.
- WILLER, C. J., LI, Y. and ABECASIS, G. R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26** 2190–2191.
- YANG, J., BENYAMIN, B., McEVOY, B. P., GORDON, S., HENDERS, A. K., NYHOLT, D. R., MADDEN, P. A., HEATH, A. C., MARTIN, N. G., MONTGOMERY, G. W. et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature genetics* **42** 565–569.
- YANG, J., BAKSHI, A., ZHU, Z., HEMANI, G., VINKHUYZEN, A. A., LEE, S. H., ROBINSON, M. R., PERRY, J. R., NOLTE, I. M., VAN VLIET-OSTAPTCHOUK, J. V. et al. (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature genetics*.
- ZABLOCKI, R. W., SCHORK, A. J., LEVINE, R. A., ANDREASSEN, O. A., DALE, A. M. and THOMPSON, W. K. (2014). Covariate-modulated local false discovery rate for genome-wide association studies. *Bioinformatics* **30** 2098–2104.

RONG W. ZABLOCKI
COMPUTATIONAL SCIENCE RESEARCH CENTER
SAN DIEGO STATE UNIVERSITY
5500 CAMPANILE DRIVE
SAN DIEGO, CA 92182
USA;
INSTITUTE OF MATHEMATICAL SCIENCES
CLAREMONT GRADUATE UNIVERSITY
150 E. 10TH ST.
CLAREMONT, CA 91711
USA

ANDREW J. SCHORK
COGNITIVE SCIENCES GRADUATE PROGRAM
UNIVERSITY OF CALIFORNIA AT SAN DIEGO
9500 GILMAN DRIVE
LA JOLLA, CA 92093
USA

YUNPENG WANG
INSTITUTE OF CLINICAL MEDICINE
UNIVERSITY OF OSLO
OSLO, 0424
NORWAY

RICHARD A. LEVINE
DEPARTMENT OF MATHEMATICS AND STATISTICS
SAN DIEGO STATE UNIVERSITY
5500 CAMPANILE DRIVE
SAN DIEGO, CA 92182
USA

SHUJING XU
DEPARTMENT OF PSYCHIATRY
UNIVERSITY OF CALIFORNIA AT SAN DIEGO
9500 GILMAN DRIVE
LA JOLLA, CA 92093
USA

CHUN C. FAN
COGNITIVE SCIENCES GRADUATE PROGRAM
UNIVERSITY OF CALIFORNIA AT SAN DIEGO
9500 GILMAN DRIVE
LA JOLLA, CA 92093
USA

WESLEY K. THOMPSON
INSTITUTE OF BIOLOGICAL PSYCHIATRY
MENTAL HEALTH CENTRE SCT. HANS
MENTAL HEALTH SERVICES COPENHAGEN
DK-4000
DENMARK
DEPARTMENT OF PSYCHIATRY
UNIVERSITY OF CALIFORNIA AT SAN DIEGO
9500 GILMAN DRIVE
LA JOLLA, CA 92093
USA E-MAIL: wes.stat@gmail.com