# Bayesian mixture model for clustering rare-variant effects in human genetic studies

Guhan Ram Venkataraman[1], Yosuke Tanigawa[1], Matti Pirinen[2,3,4*], Manuel A. Rivas[1†]

**Rare-variant aggregate analysis from exome and whole genome sequencing data typically summarizes with a single statistic the signal for a gene or the unit that is being aggregated. However, when doing so, the effect profile within the unit may not be easily characterized across one or multiple phenotypes. Here, we present an approach we call Multiple Rare-Variants and Phenotypes Mixture Model (MRPMM), which clusters rare variants into groups based on their effects on the multivariate phenotype and makes statistical inferences about the properties of the underlying mixture of genetic effects. Using summary statistic data from a meta-analysis of exome sequencing data of 184,698 individuals in the UK Biobank across 6 populations, we demonstrate that our mixture model can identify clusters of variants responsible for significantly disparate effects across a multivariate phenotype; we study three lipid and three renal traits separately. The method is able to estimate (1) the proportion of non-null variants, (2) whether variants with the same predicted consequence in one gene behave similarly, (3) whether variants across genes share effect profiles across the multivariate phenotype, and (4) whether different annotations differ in the magnitude of their effects. As rare-variant data and aggregation techniques become more common, this method can be used to ascribe further meaning to association results.**

[1]*Department of Biomedical Data Science, Stanford University, Stanford, CA, USA*

[2]*Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland*

[3]*Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland*

[4]*Department of Public Health, Faculty of Medicine, University of Helsinki, Helsinki, Finland*

[*]matti.pirinen@helsinki.fi

[†]mrivas@stanford.edu

## 1 Introduction

Population-scale sequencing studies are becoming pervasive[1–4]. As a result, analyses considering the joint contribution of rare variants to disease susceptibility and phenotypic variation are also becoming pervasive. Commonly used aggregation approaches for rare-variant association studies include the sequence kernel association test, the burden test, and more-general Bayesian model comparison methods[5–14]. However, aggregation as performed in these methods also tends to lose information within the blocks (typically, genes) specified; that is, they fail to indicate whether certain variants (or certain types of variants) within the same block may have different effects on the multivariate phenotype. In particular, they do not pinpoint which variants are driving the association signal. It is thus critical that we develop methods that can "trace back" variants' effects, clustering the variants used within the blocks of interest into groups that have distinct per-phenotype aggregate effects while adequately accounting for the uncertainty that rare-variant studies exhibit.

Clustering methods fall into several categories, the most prevalent of which are distance-based methods (such as K-means and self-organizing maps) that are sensitive to the noise that typically plagues biological data[8]. Model-based methods such as mixture models are a simple but elegant alternative that assume that data are generated from multiple source distributions, which are then learned and set as the "clusters". Algorithms such as Expectation Maximization can estimate latent variables that underlie these clusters. The finite-mixture model, one such model-based method, assumes a finite number of clusters and asserts that this number can be estimated using goodness-of-fit criteria like the Bayesian Information Criterion (BIC) or the Akaike Information

Criterion (AIC)[15,16].

In this study, we propose to use a Bayesian hierarchical mixture model where a hierarchical structure is introduced to allow the sharing of information among related clusters (in our case, genes) and where the number of clusters are pre-specified. Our approach, the Multiple Rare-Variants and Phenotypes Mixture Model (MRPMM), considers several factors when estimating parameters of interest underlying the mixture of effects driving an association signal. We calculate matrices of genetic correlations among phenotypes of interest using both common or null variants and rarer or significant variants. The method also estimates the spread of effects across predicted consequences of the genomic variants (variant annotations), which is a critical aspect of interpreting genetic findings. The annotations represent expected severity of impact on phenotypes (for example, protein-truncating variants, or PTVs[17,18], are predicted to truncate the protein product, and are purported to be much more deleterious than other variants). In MRPMM, we use summary statistics (for single-variant single-phenotype GWAS, these are per-variant estimates of marginal univariate effect size and corresponding standard errors). In practice, sharing of individual genotype and phenotype data across groups in large genetic consortia is difficult to achieve due to privacy concerns and consent issues; using summary statistics can bypass these issues while also increasing computational efficiency without reducing accuracy. Insights from Liu et al.[19] and Cichonska et al.[20] also suggest that the use of additional summary statistics, like covariance estimates across variants and studies, respectively, enable a lossless ability to detect gene-based association signals using summary statistics alone.

## 2 Methods

**Algorithm:** Our goal is to cluster the variants into groups based on their effects on the multivariate phenotype; that is, we are interested in the joint posterior distribution that indicates cluster memberships per-variant and per-gene. For multi-parameter models such as these, the joint posterior may be difficult to sample from directly. Often, it is easier to sample sequentially from the full conditional distribution of each parameter [21] using a Gibbs sampler, a Markov Chain Monte Carlo (MCMC) algorithm that constructs a dependent sequence of parameter values whose distribution approximates the joint posterior [21–23]. We implement the Gibbs sampler in MRPMM as follows:

- We index genes by $j = 1, \ldots, J$, denoting variant $m$ in gene $j$ by $v_{jm}$. Within gene $j$, the variants $v_{jm}$ are assigned to clusters $c = 1, ..., C$, each of which is represented by an unscaled effect size parameter $\boldsymbol{b}_c$.

- The prior for $\boldsymbol{b}_c$ is $\mathcal{N}(0, \boldsymbol{\Theta}_0)$, where $\boldsymbol{\Theta}_0$ is an estimate of genetic correlation across the traits.

- Consider the model with $C > 1$ clusters. In order to model the sharing of clusters across the genes, we first draw a $C$-dimensional probability vector $\boldsymbol{\pi}_0 \sim \text{Dirichlet}(1, 1, 1, \ldots, 1)$.

- Next, for each gene $j$, we draw a probability vector $\boldsymbol{\pi}_j | \boldsymbol{\pi}_0 \sim \text{Dirichlet}(\alpha \boldsymbol{\pi}_0)$ to determine the mixture proportions $\pi_{jc}$ that dictate how the variants in gene $j$ are distributed across the clusters $1, \ldots, C$.

- The parameter $\alpha$ governs how similar the cluster proportions are across genes; it is drawn from a prior $\alpha \sim \text{Inv-Gamma}(1, 1)$.

- The algorithm also takes into account the functional annotation, via $\sigma_a^2$, a variance parameter for annotation $a$ across the clusters. It follows an inverse-gamma prior $\sigma_a^2 \sim$ Inv-Gamma $(sh_a, sp_a)$ , with hyperparameter values $sh_a$ (shape) and $sp_a$ (spread).

- Under the above assumptions of the model, the phenotype of individual $i$ with a rare allele of variant $m$ with annotation $a$ in gene $j$ is $\boldsymbol{y}_i|(\boldsymbol{\pi}_j, \boldsymbol{b}, \sigma_a^2) \sim \sum_{c=1}^{C} \pi_{jc} \mathcal{N}\left(\sigma_a \boldsymbol{b}_c, \widehat{\mathrm{V}}_Y\right)$, where $\widehat{\mathrm{V}}_Y$ is the estimated residual variance-covariance matrix of phenotypes after the effects of the variants have been regressed out.

- If we have access only to summary statistics of estimated effect sizes $\widehat{\boldsymbol{\beta}}_{jm}$, then we estimate variance-covariance matrices $\widehat{\mathrm{V}}_{jm}$ for each variant $m$ in gene $j$ as $\widehat{\mathrm{V}}_{jm} \approx \mathrm{diag}(SE_m) \times \Omega \times \mathrm{diag}(SE_m)$, where $SE_m$ denotes the $K$-dimensional vector of standard errors across $K$ phenotypes for variant $m$, and $\Omega$ is the $K \times K$ matrix of correlation of errors estimated from null variants. Then, our sampling model for the data is $\widehat{\boldsymbol{\beta}}_{jm}|(\boldsymbol{\pi}_j, \boldsymbol{b}, \sigma_a^2) \sim \sum_{c=1}^{C} \pi_{jc} \mathcal{N}\left(\sigma_a \boldsymbol{b}_c, \widehat{\mathrm{V}}_{jm}\right)$ .

- Similar to the individual level data model above, this formulation assumes independence between the variants, which is approximately true when each individual carries at most one of the rare variants considered. A connection between the two data types is that $\widehat{\boldsymbol{\beta}}_{jm} \approx \frac{1}{1-2f_{jm}} \overline{\boldsymbol{y}}_{jm}$ and $\widehat{\mathrm{V}}_{jm} \approx \frac{1}{2Nf_{jm}(1-2f_{jm})} \widehat{\mathrm{V}}_Y$, where $f_{jm}$ is the frequency of the rare allele at variant $m$ of gene $j$ among the $2N$ haplotypes, and $\overline{\boldsymbol{y}}_{jm}$ is the mean phenotype of the carrier individuals of that allele.

- To compare models with different numbers of clusters, we use the Bayesian Information

Criterion (BIC)[15], defined for the model $\mathcal{M}_C$ with $C$ clusters as

$$BIC_C = -2\log\left(p\left(D|\widehat{\theta_C}, \mathcal{M}_C\right)\right) + \nu_C \log(n),$$

where $D$ denotes the observed data, $\widehat{\theta_C}$ is the vector of maximum likelihood estimates of the parameters of $\mathcal{M}_C$, $\nu_c$ is the number of independent parameters of $\mathcal{M}_c$, and $n$ is the number of data points. The difference in BIC values between two models approximates twice the logarithm of the Bayes Factor between the models, with lower BIC value corresponding to the model preferred.

MRPMM utilizes Metropolis-Hastings (MH) steps to update the parameters of interest. The algorithm is run for $n_{\text{burn}} + n_{\text{iter}}$ iterations, of which the first $n_{\text{burn}}$ are discarded as an initial "burn-in". With superscripts in parentheses denoting iteration, the steps of the algorithm are as follows.

1. **Initialize parameters.**

   (a) Draw $\alpha^{(0)} \sim \text{Inv-Gamma}(1, 1)$.

   (b) Draw $\boldsymbol{\pi}_0^{(0)} \sim \text{Dirichlet}(1, 1, 1, \ldots, 1)$.

   (c) Draw for each gene $j$: $\boldsymbol{\pi}_j^{(0)} \sim \text{Dirichlet}\left(\alpha^{(0)}\boldsymbol{\pi}_0^{(0)}\right)$;

   (d) $\boldsymbol{b}_1^{(0)} = \mathbf{0}$.

   (e) For each $c$ in $2, \ldots, C$ draw $\boldsymbol{b}_c^{(0)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Theta}_0)$.

   (f) $(\sigma_a^2)^{(0)} = 0.2^2$ for all $a$.

2. **Repeat the following steps for iterations** $t = 1, 2, \ldots, n_{\text{burn}} + n_{\text{iter}}$:

   (a) Update $\boldsymbol{\pi}_0$ (probability of cluster assignment independent of gene). We use a Metropolis-Hastings sub-step [24] using a proposal centred around the current value, drawing $\boldsymbol{\pi}_0^{'} \sim \text{Dirichlet}\left(\gamma\boldsymbol{\pi}_0^{(t-1)}\right)$. To calculate the acceptance probability, we define the normalizing constant for the C-dimensional Dirichlet distribution with parameters $\boldsymbol{z}$ as

$$D(\boldsymbol{z}) = \frac{\Gamma\left(\sum_{c=1}^{C} z_c\right)}{\prod_{c=1}^{C}\Gamma\left(z_c\right)}$$

   and the density at point $\boldsymbol{x}$ as

$$p_{Dir}(\boldsymbol{x}|\boldsymbol{z}) = D(\boldsymbol{z})\prod_{c=1}^{C} x_c^{z_c-1}.$$

   With this notation, the Metropolis-Hastings transition probability from $\boldsymbol{\pi}_0^{(t-1)}$ to $\boldsymbol{\pi}_0^{'}$ is $p_{Dir}\left(\boldsymbol{\pi}_0^{'}|\gamma\boldsymbol{\pi}_0^{(t-1)}\right)$, and the density of the observed data depends on $\boldsymbol{\pi}_0^{(t-1)}$ only through the product $\prod_{j=1}^{J} p_{Dir}\left(\boldsymbol{\pi}_j|\alpha\boldsymbol{\pi}_0^{(t-1)}\right)$. Hence, the proposal acceptance probability is

$$\lambda = \min\left(1, \frac{p_{Dir}\left(\boldsymbol{\pi}_0^{(t-1)}|\gamma\boldsymbol{\pi}_0^{'}\right)\prod_{j=1}^{J} p_{Dir}\left(\boldsymbol{\pi}_j|\alpha\boldsymbol{\pi}_0^{'}\right)}{p_{Dir}\left(\boldsymbol{\pi}_0^{'}|\gamma\boldsymbol{\pi}_0^{(t-1)}\right)\prod_{j=1}^{J} p_{Dir}\left(\boldsymbol{\pi}_j|\alpha\boldsymbol{\pi}_0^{(t-1)}\right)}\right)$$

   Ergo, with probability $\lambda$, we set $\boldsymbol{\pi}_0^{(t)} = \boldsymbol{\pi}_0^{'}$, and with probability $1-\lambda$, $\boldsymbol{\pi}_0^{(t)} = \boldsymbol{\pi}_0^{(t-1)}$.

   (b) For each gene $j = 1, \ldots, J$ we update $\boldsymbol{\pi}_j$ (per-gene cluster parameter) to be

$$\boldsymbol{\pi}_j \sim \text{Dirichlet}\left(\alpha\boldsymbol{\pi}_0 + \left(\sum_{m=1}^{M_j} I(\delta_{jm} = 1), \sum_{m=1}^{M_j} I(\delta_{jm} = 2), \ldots, \sum_{m=1}^{M_j} I(\delta_{jm} = C)\right)\right),$$

where $\delta_{jm}$ is the index of the cluster to which the variant $m$ of gene $j$ belongs to and $I(\cdot)$ is the indicator function.

(c) Update $\delta_{jm}$ for all $j = 1, \ldots, J$ and $m = 1, \ldots, M_j$. For all $c$, compute

$$p^{'}_{jmc} = p\left(\hat{\boldsymbol{\beta}}_{jm}; \boldsymbol{b}_c, \sigma_a^2\right)\pi_{jc},$$

and renormalize in such a way that $p_{jmc} \propto p^{'}_{jmc}$ sums to 1 over all $c$. Then, we sample

$$\delta_{jm} \sim \text{Discrete}(p_{jmc}).$$

(d) Update $\boldsymbol{b}_c$ (cluster effect profile) using a Gibbs update from a Gaussian distribution:

$$\begin{aligned}
\text{mean} &= \left(\boldsymbol{\Theta}_0^{-1} + \sum_{v_{jm}\in c} \sigma^2_{a_{jm}} \boldsymbol{\Sigma}_c^{-1}\right)^{-1} \left(\sum_{v_{jm}\in c} \sigma_{a_{jm}} \boldsymbol{\Sigma}_c^{-1} \hat{\boldsymbol{\beta}}_{jm}\right) \\
\text{var} &= \left(\boldsymbol{\Theta}_0^{-1} + \sum_{v_{jm}\in c} \sigma^2_{a_{jm}} \boldsymbol{\Sigma}_c^{-1}\right)^{-1}.
\end{aligned}$$

(e) Update $\sigma_a^2$ (annotation spread parameter). We use a Metropolis-Hastings sub-step using a random-walk proposal. That is, sequentially for each annotation $a$, we sample a proposal value

$$\sigma^{'}_a = |\eta|, \text{ where } \eta \sim \mathcal{N}\left(\sigma_a^{(t-1)}, \xi_0\right),$$

where $\xi_0$ is a hyperparameter controlling the spread of the proposals. In the examples,

we have used $\xi_0 = 1$. Then we calculate the acceptance probability

$$\lambda = \min\left(1, \frac{p\left(\sigma_a^{'}|sh_a, sc_a\right)\prod_{\text{anno}(v_{jm})=a}\mathcal{N}\left(\widehat{\boldsymbol{\beta}}_{jm}|\sigma_a^{'}\boldsymbol{b}_{c_{jm}}, \widehat{\boldsymbol{V}}_{jm}\right)}{p\left(\sigma_a^{(t-1)}|sh_a, sc_a\right)\prod_{\text{anno}(v_{jm})=a}\mathcal{N}\left(\widehat{\boldsymbol{\beta}}_{jm}|\sigma_a^{(t-1)}\boldsymbol{b}_{c_{jm}}, \widehat{\boldsymbol{V}}_{jm}\right)}\right),$$

where the products are over those variants $v_{jm}$ whose annotation is $a$ and $c_{jm}$ is the cluster of variant $v_{jm}$. With probability $\lambda$, we set $\sigma_a^{(t)} = \sigma_a^{'}$, and with probability $1-\lambda$, $\sigma_a^{(t)} = \sigma_a^{(t-1)}$.

(f) Update $\alpha$ (parameter determining sharing of clusters across genes). We again use a Metropolis-Hastings sub-step using a random-walk proposal. That is, we sample a proposal value $\alpha^{'} = |\eta|$, where $\eta \sim \mathcal{N}(\alpha^{(t-1)}, \xi_\alpha)$ where $\xi_\alpha$ is a fixed value controlling the variance of the proposal distribution. Then, we compute the acceptance probability

$$\lambda = \min\left(1, \frac{p\left(\alpha^{'}\right)\prod_{j=1}^{J} p_{Dir}\left(\boldsymbol{\pi}_j^{(t)}|\alpha^{'}\boldsymbol{\pi}_0^{(t)}\right)}{p\left(\alpha^{(t-1)}\right)\prod_{j=1}^{J} p_{Dir}\left(\boldsymbol{\pi}_j^{(t)}|\alpha^{(t-1)}\boldsymbol{\pi}_0^{(t)}\right)}\right),$$

where the prior for $\alpha$ is Inv-Gamma(1,1) (i.e., $p(\alpha) = \alpha^{-2}\exp\left(-\alpha^{-1}\right)$), and $p_{Dir}$ is the density of the Dirichlet distribution as defined above in part (a). With probability $\lambda$, we set $\alpha^{(t)} = \alpha^{'}$, and with probability $1-\lambda$, $\alpha^{(t)} = \alpha^{(t-1)}$.

**Data:** We used a combination of self-reported ancestry (UK Biobank field ID 21000), principal component analysis on genotype data, and the relatedness matrix to identify six subpopulations in the study: white British, African, South Asian, non-British white, semi-related, and an admixed population. To determine the first four populations, which contain samples not related closer than

the third degree, we first used the principal components of the genotyped variants from the UK Biobank and defined thresholds on principal component 1 and principal component 2 and further refined the population definition as described elsewhere[25] Semi-related individuals were grouped as individuals whose genetic data (after passing UK Biobank QC filters; sufficiently low missingness rates; and genetically inferred sex matching reported sex), using a King's relationship table, were between conditional third and conditional second degrees of relatedness. Admixed individuals were grouped as unrelated individuals who were flagged as "used_in_pca_calculation" by the UK Biobank and were not assigned to any of the other populations[6].

We performed genome-wide association analysis on individuals with whole-exome sequencing data for three lipid-related phenotypes (high-density lipoproteins [UK Biobank Field 30760], triglycerides [UK Biobank Field 30870], and low-density lipoproteins [UK Biobank Field 30780]) and three renal-related phenotypes (creatinine [UK Biobank Field 30700], cystatin C [UK Biobank Field 30720], and effective glomerular filtration rate [derived from UK Biobank Field 30700]). The analyses were performed for each of the six population subgroups as defined above using PLINK v2.00a (20 October 2020). The quantitative trait values were rank normalized using the –pheno-quantile-normalize flag. We used age, sex, and the first ten genetic principal components as covariates in the analyses. The analysis was performed for 5,850,789 rare (minor allele frequency $\leq 0.01$) protein-truncating (492,151) and protein-altering (5,358,638) variants.

For the admixed population, we conducted local ancestry-corrected GWAS. We first assembled a reference panel from 1,380 single-ancestry samples in the 1000 Genomes Project[26], the

Human Genome Diversity Project[27] and the Simons Genome Diversity Project[28] choosing appropriate ancestry clusters by running ADMIXTURE[29] with the unsupervised setting. Using cross-validation, eight well-supported ancestral population clusters were identified: African, African Hunter-Gatherer, East Asian, European, Native American, Oceanian, South Asian, and West Asian. We then used RFMix v2.03[30] to assign each of the 20,727 windows across the phased genomes to one of these eight ancestry clusters (for all individuals in the UK Biobank). These local ancestry assignments were subsequently used with PLINK2 as local covariates in the GWAS for the admixed individuals for SNPs within those respective windows. PLINK2 allows for the direct input of the RFMix output (the MSP file, which contains the most likely subpopulation assignment per conditional random field [CRF] point) as local covariates using the "local-cov", "local-psam", and "local-haps" flags, the "local-cats0=n" flag (where n is the number of assignments), and the "local-pos-cols=2,1,2,7" flag (for a typical RFMix MSP output file - see "Association Analysis" page on PLINK website).

Subsequently, we used METAL[31] to perform inverse-variance weighted meta-analysis to generate a single summary statistic file per phenotype.

For the remainder, we used Variant Effect Predictor (VEP)[32] to annotate the most severe consequence, the gene symbol, and HGVSp of each variant in the UK Biobank exome data. We calculated minor allele frequencies using PLINK. We provide these metadata, which are necessary for MRPMM, in exome and array tables, available for direct download via the Global Biobank Engine[33] (Code Availability).

## 3 Results

To assess MRPMM's ability to estimate the underlying mixture of effects from summary statistics, we chose two sets of related phenotypes: one set of lipid phenotypes (high-density lipoprotein cholesterol levels [HDL], low-density lipoprotein cholesterol levels [LDL], and triglycerides [TG]), and another set of renal phenotypes (creatinine [CRE], cystatin C [CSTC] and effective glomerular filtration rate [eGFR]). We then identified genes with a significant burden of associated, rare, protein-truncating variants (PTVs) in a meta-analysis comprising 184,698 UK Biobank individuals amongst six cohorts of different ancestries (white British [137,920], non-British white [10,432], African [2,716], South Asian [3,569], semi-related [18,100], and admixed [11,961]). After using a Bayesian model comparison approach[6] to consider GWAS evidence across each set of phenotypes and across each gene, we chose 13 genes that had a $\log_{10}$ Bayes Factor (BF) $\geq 5$ for lipid traits, and 5 genes that met that criterion for renal traits. We then ran MRPMM across PTVs in these genes, increasing the number of hypothesized clusters until there was an increase in BIC, a goodness-of-fit measure which is minimized under ideal conditions. We found that four clusters (including the "null cluster" with a constant effect of 0 across phenotypes) was favored for both sets of phenotypes (Figure 1), and that each cluster had a distinct effect size profile on the multivariate phenotype (Figures 2, 3).

Specifically, we see that there are some *APOB* variants (Figure 4a - left side) that have a strong negative effect on TG and LDL levels (Figure 2D) through exclusive membership to Cluster 3. To the right of those variants, other variants are definitively placed into Cluster 2 and Cluster

1 respectively. The variants that are definitively placed into Cluster 2 feature the *PCSK9* and *ANGPTL3* genes; MRPMM shows that these variants down-regulate TG and LDL levels (Figure 2C). The variants which belong exclusively to Cluster 1 feature the *PDE3B*, *APOC3*, *CETP*, and *ANGPTL8* genes and have positive effects on HDL levels, negative effects on TG levels, and mild negative effects on LDL levels (Figure 2B). We can perform a similar visual analysis with the renal multivariate phenotype results (Figure 4b). Variants from the *CGNL1*, *RNF186*, *SLC22A2*, and *SLC34A3* genes largely seem to fall under Cluster 2, whereas variants from *CST3* clearly fall into Cluster 1 in an isolated manner. Cluster 3 seems to be populated partially by several variants in *CGNL1*. These variant-specific breakdowns and their gene-aggregate counterparts (Figures 5a, 5b) help trace back the variants' effect profile on the multivariate phenotype. Unlike in aggregation approaches, where a single statistic captures which genes are associated with the trait without providing much context as to the nature of the association, MRPMM has the useful characteristic of being able to cluster not only variants but also genes into different effect profiles. We provide single-trait MRPMM results for all traits across the UK Biobank on the Global Biobank Engine[33] (Code Availability).

## 4 Future Directions and Discussion

In this study, we used a Bayesian hierarchical mixture model to estimate the underlying mixture of components driving association signals between various protein-truncating variants and a set of genetically related phenotypes. By explicitly modeling the sharing of effects across genes, we are able to use Gibbs sampling to approximately infer the joint posterior distribution and thereby

assign variants to clusters. In both applications, we see that clusters have vastly different effect size profiles on the sets of phenotypes chosen; this shows that aggregating rare-variant signal in blocks (e.g. genes) may not fully encapsulate the information that is available in summary statistics. Coupling a Bayesian model comparison approach as described by Venkataraman et al.[6] with MRPMM may be a way to (1) systematically screen for genes associated with the set of phenotypes of interest and then (2) cluster the effect size profiles of rare variants within these genes, thereby providing a window into the underlying biology. For example, aligning the results from the MRPMM analysis and other function-elucidating analyses like protein domain models or 3D structure analysis could potentially lead to the identification of promising therapeutic targets. Going forward, it is also essential that this analysis be performed using whole genome sequencing data; as MRPMM is able to use any type of annotation, how variant effects translate across epigenomic profiles and/or conservation patterns may become relevant and useful to analyze in these settings. Overall, MRPMM provides interpretability at the level of individual variants, in contrast to typical rare-variant techniques that work only at the level of aggregated variants.

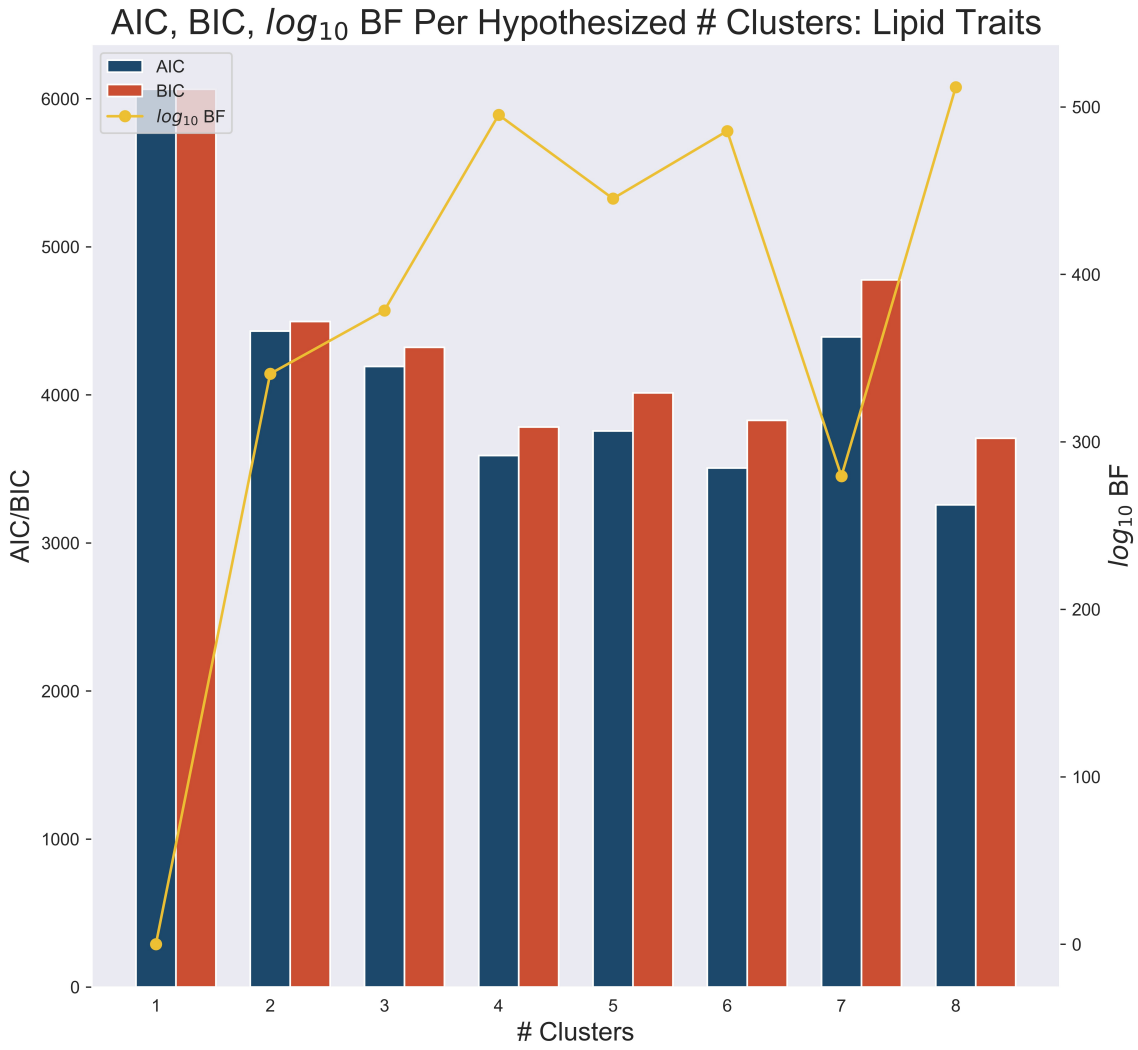


**Figure 1a:** AIC, BIC (goodness-of-fit), and $\log_{10}$ BF as compared to the null cluster for number of clusters of effects on a lipid-related multivariate phenotype of high-density lipoprotein cholesterol (HDL), triglyceride (TG) levels, and low-density lipoprotein cholesterol (LDL). The algorithm was stopped when these AIC or BIC values trended upwards and $\log_{10}$ BF was maximized. In this case, we stopped at 5 and chose 4 clusters.

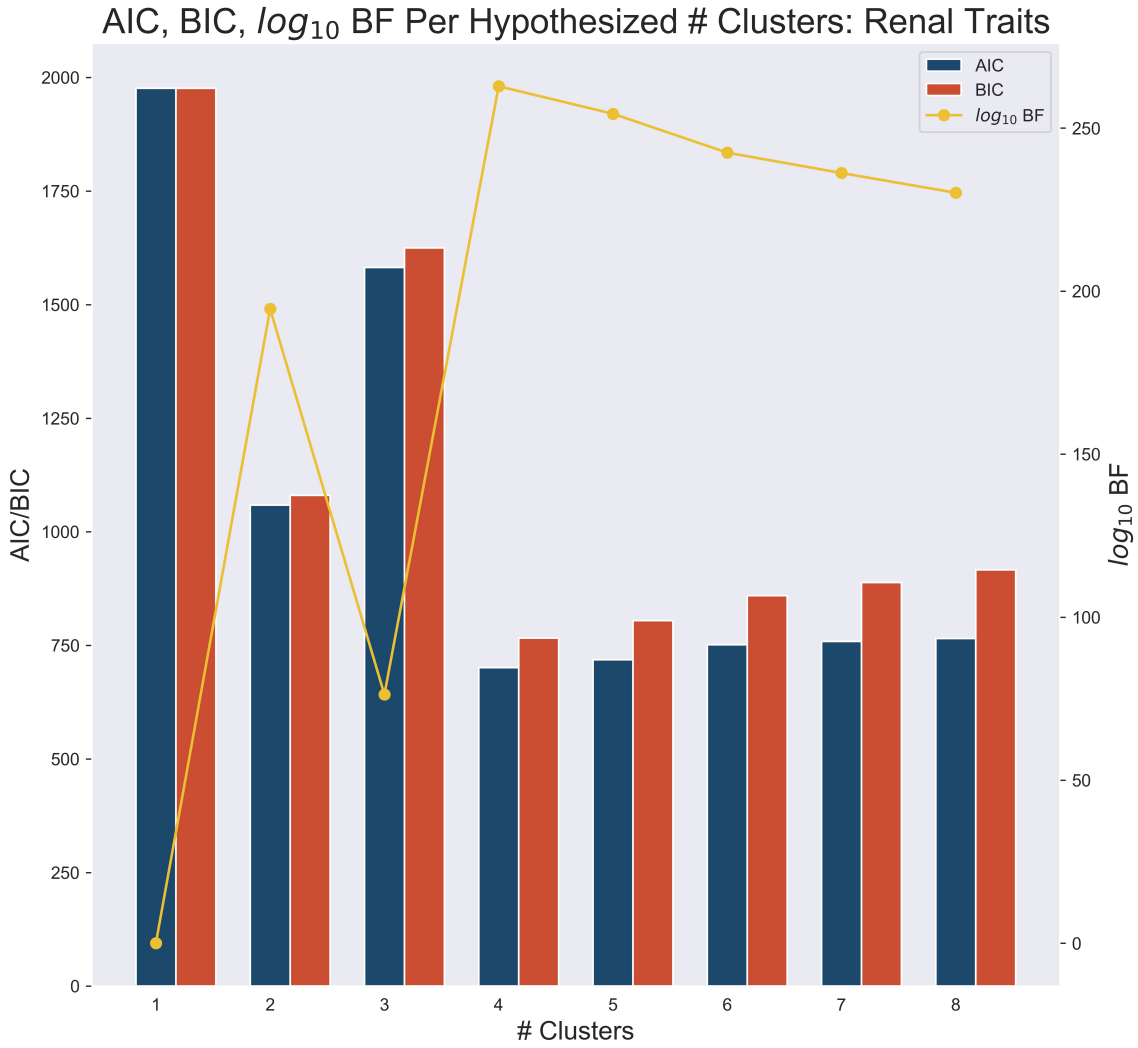


**Figure 1b:** AIC, BIC (goodness-of-fit), and $\log_{10}$ BF as compared to the null cluster for number of clusters of effects on a renal-related multivariate phenotype of creatinine (CRE), cystatin C (CSTC) levels, and effective glomerular filtration rate (eGFR). The algorithm was stopped when these AIC or BIC values trended upwards and $\log_{10}$ BF was maximized. In this case, we stopped at 5 and chose 4 clusters.

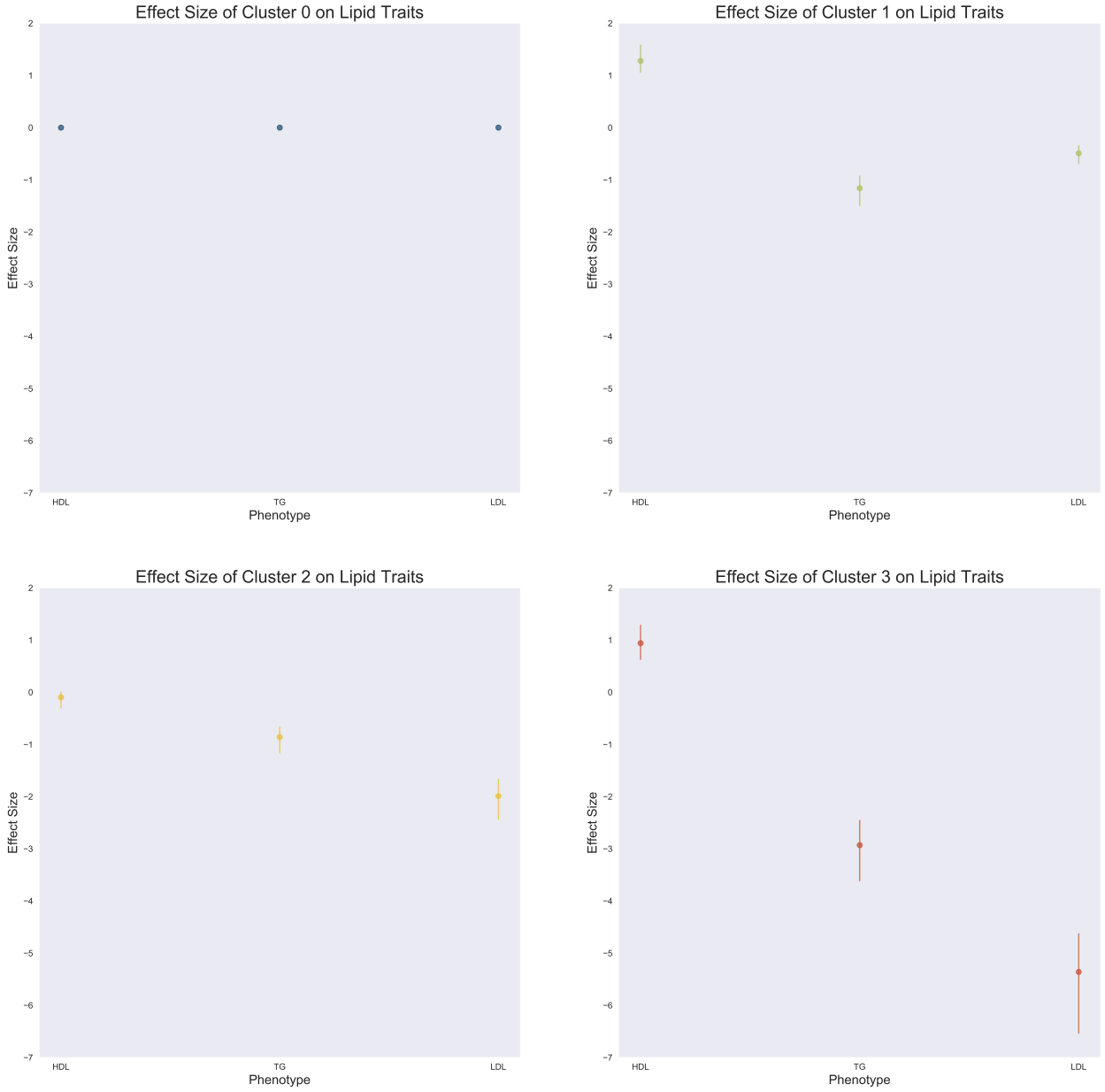


**Figure 2: A (upper left):** Effect sizes of Cluster 0 (with 95% credible intervals) on the lipid-related multivariate phenotype (the "null cluster"). **B (upper right):** Effect sizes of Cluster 1 (with 95% credible intervals) on the lipid-related multivariate phenotype. PTVs in this cluster espouse positive effects on HDL levels and negative effects on TG levels and LDL levels. **C (lower left):** Effect sizes of Cluster 2 (with 95% credible intervals) on the lipid-related multivariate phenotype. PTVs in this cluster espouse null effects on the HDL phenotype and successively negative effects on TG and LDL. **D (lower right):** Effect sizes of Cluster 3 (with 95% credible intervals) on the lipid-related multivariate phenotype. This cluster seems to have disparate effects on HDL (positive effect) as compared to TG (strong negative effect) and LDL (extremely strong negative effect).

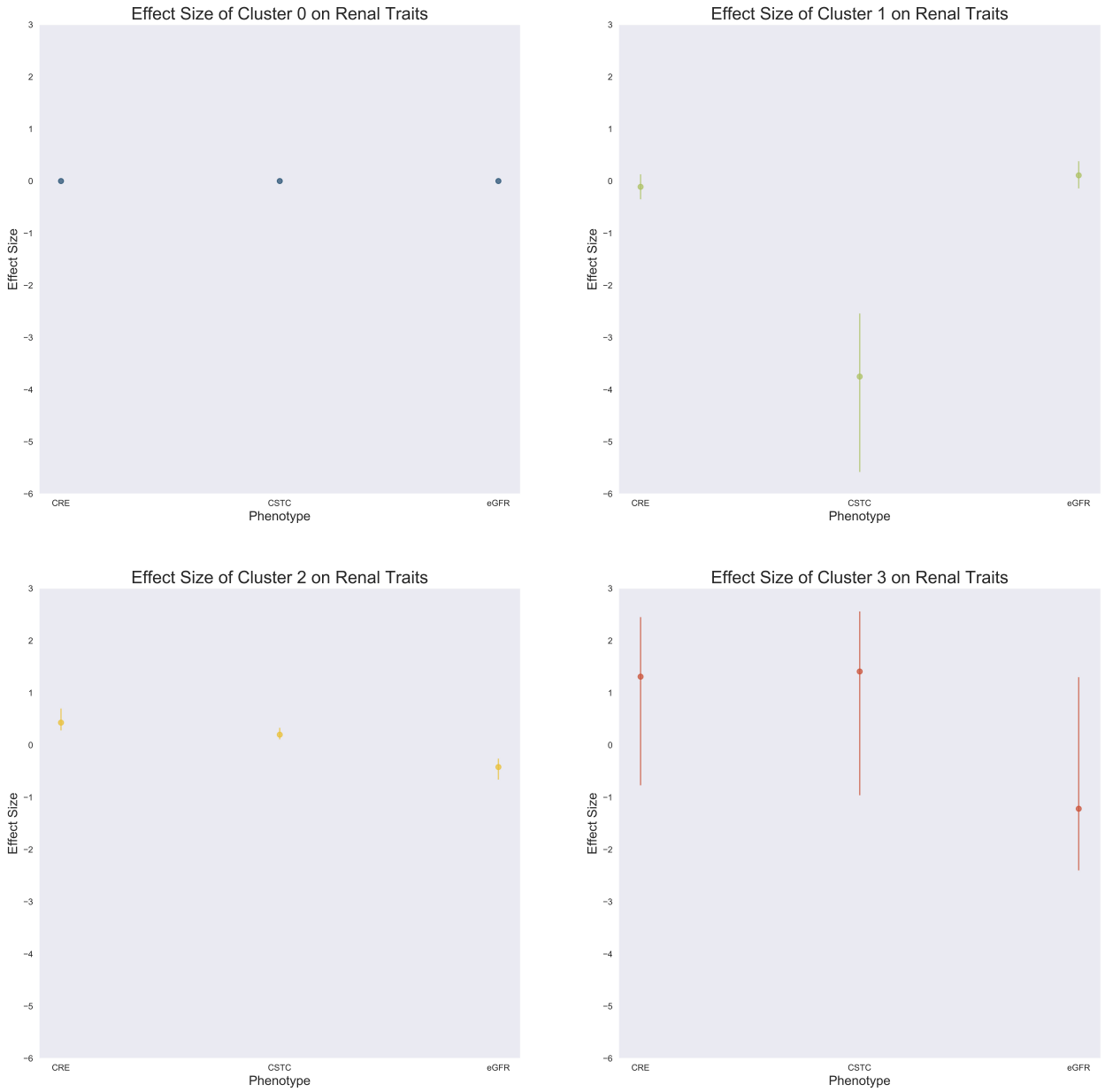


**Figure 3: A (upper left):** Effect sizes of Cluster 0 (with 95% credible intervals) on the renal-related multivariate phenotype (the "null cluster"). **B (upper right):** Effect sizes of Cluster 1 (with 95% credible intervals) on the renal-related multivariate phenotype. PTVs in this cluster espouse null effects on CRE and eGFR levels and strong negative effects on CSTC levels. **C (lower left):** Effect sizes of Cluster 2 (with 95% credible intervals) on the renal-related multivariate phenotype. PTVs in this cluster espouse mild positive effects on CRE and CSTC and mild negative effects on eGFR. **D (lower right):** Effect sizes of Cluster 3 (with 95% credible intervals) on the renal-related multivariate phenotype. This cluster seems to have inconclusive effects on the multivariate phenotype, as the 95% credible intervals for the effects cross 0.

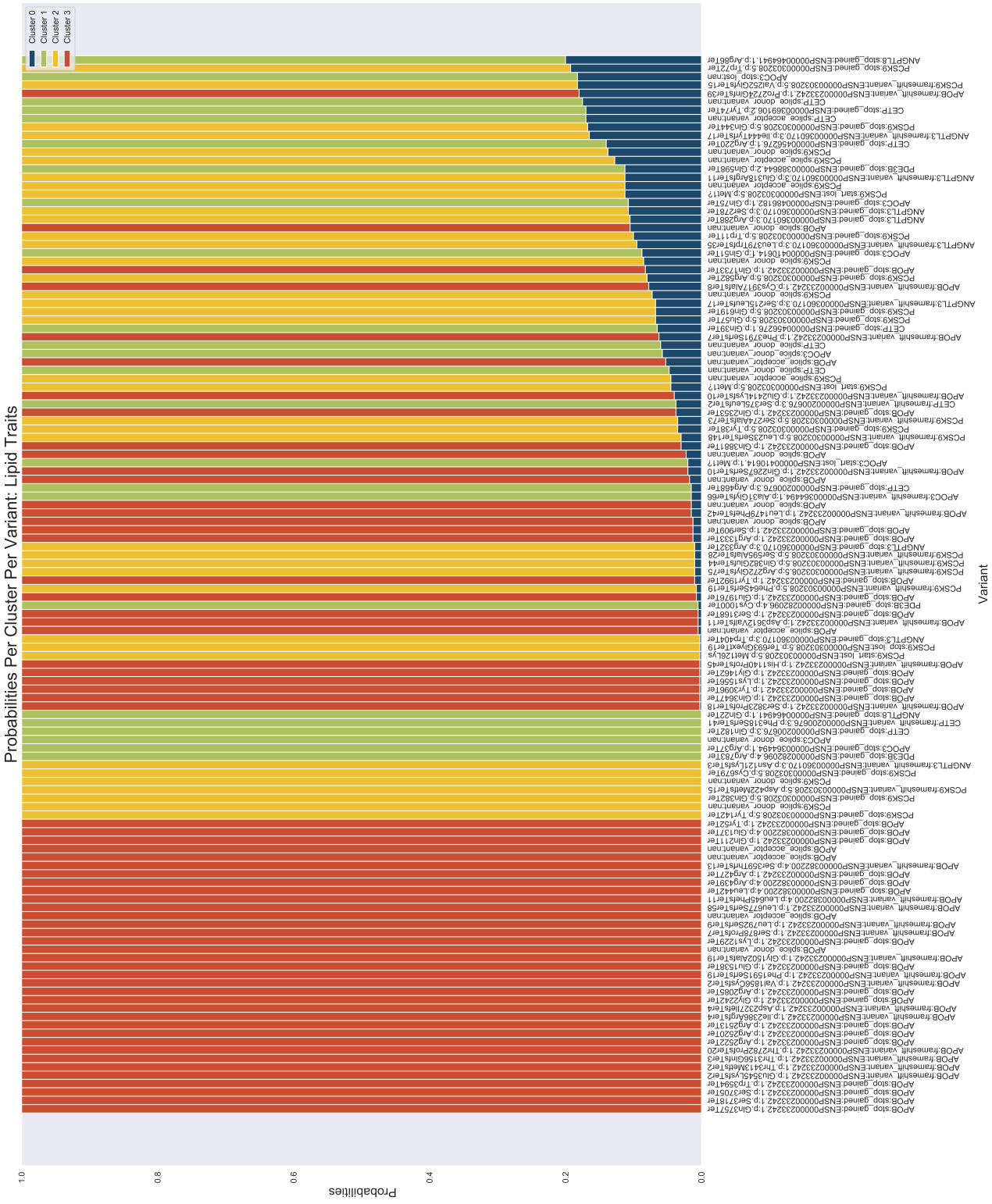


**Figure 4a:** Variant-level posterior probabilities that PTVs in candidate genes in the UK Biobank exome sequencing data set belong to each of four clusters hypothesized with respect to a lipid-related multivariate phenotype of high-density lipoprotein cholesterol (HDL), low-density lipoprotein cholesterol (LDL), and triglyceride (TG) levels. The PTVs in *APOB* that are marked as belonging to Cluster 3 on the left are responsible for coding for a hepatic lipase enzyme, hence their strong effects on triglycerides and LDL cholesterol levels (see Figure 3D). Only PTVs that have posterior probability $\leq 0.2$ of belonging in the null cluster (Cluster 0) are shown. Variants are sorted by the posterior probability of belonging to the null cluster.

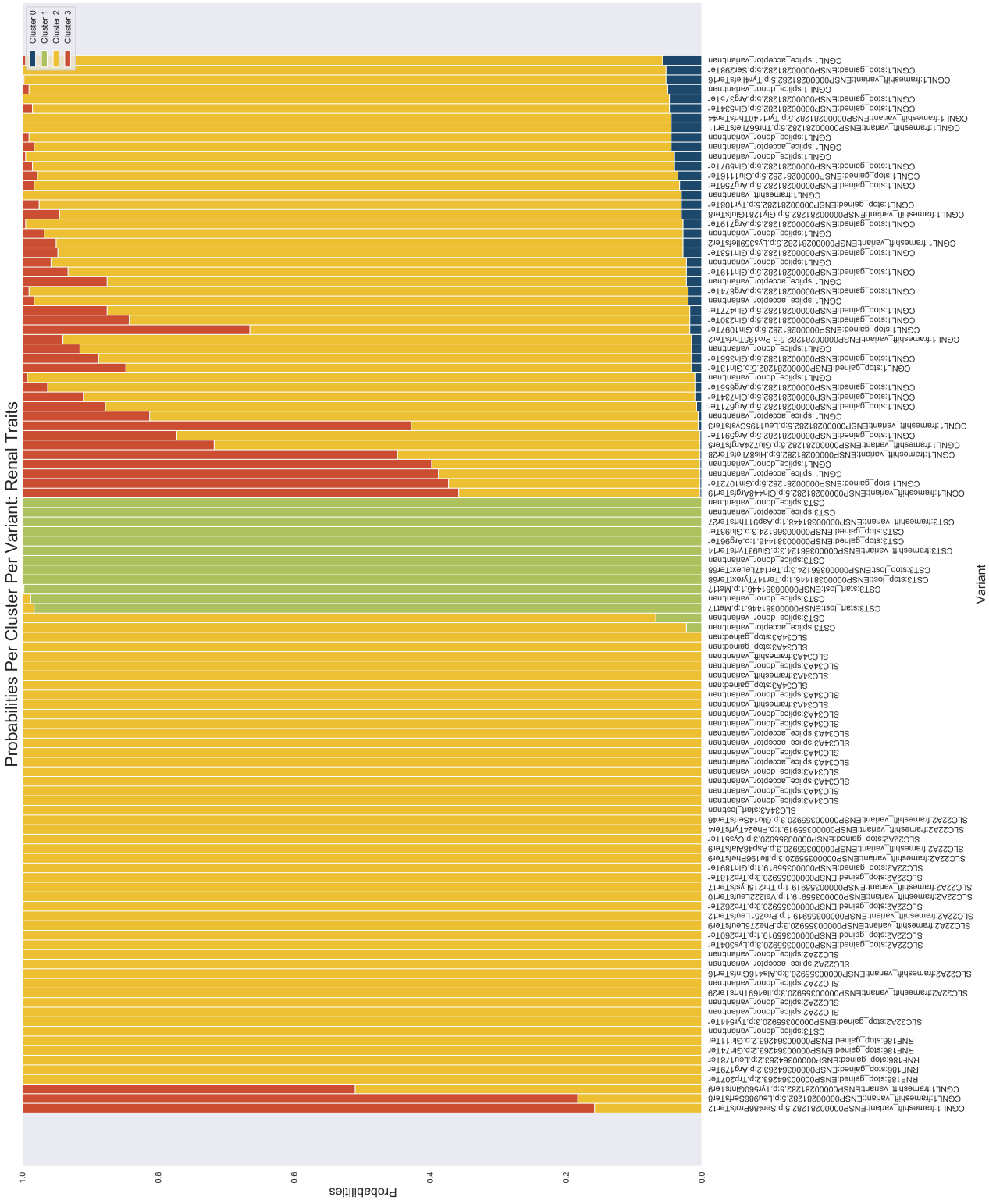


**Figure 4b:** Variant-level posterior probabilities that PTVs in candidate genes in the UK Biobank exome sequencing data set belong to each of four clusters hypothesized with respect to a renal-related multivariate phenotype of creatinine (CRE), cystatin C (CSTC), and effective glomerular filtration rate (eGFR). The PTVs in *CST3* that are marked as belonging to Cluster 1 in the middle are responsible for coding cystatin C directly, hence their strong effects CSTC levels (see Figure 4B). Only PTVs that have posterior probability $\leq 0.2$ of belonging in the null cluster (Cluster 0) are shown. Variants are sorted by the posterior probability of belonging to the null cluster.

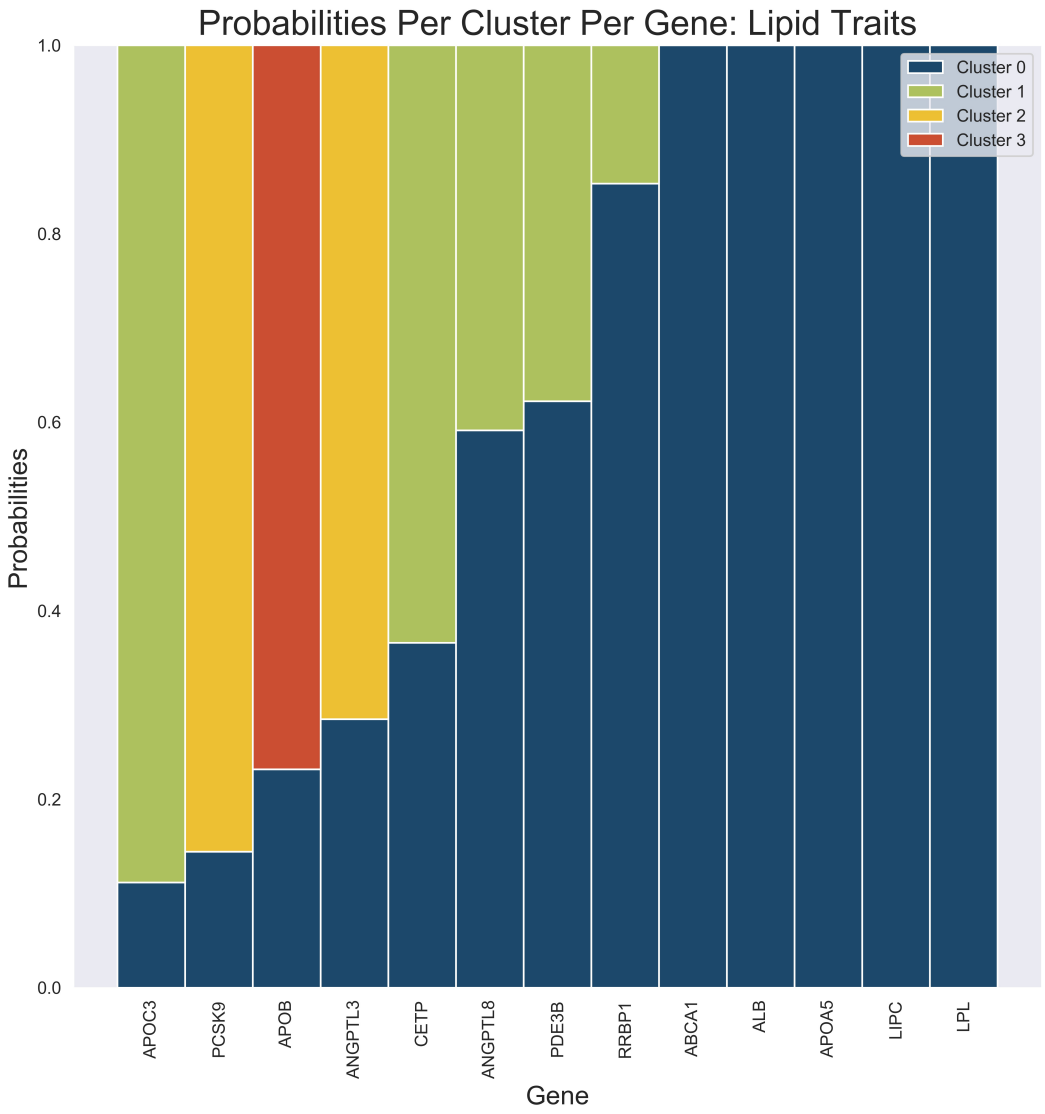


**Figure 5a:** Gene-level posterior probabilities that PTVs in candidate genes in the UK Biobank exome sequencing data set belong to each of four clusters hypothesized with respect to a lipid-related multivariate phenotype of high-density lipoprotein cholesterol (HDL), low-density lipoprotein cholesterol (LDL), and triglyceride (TG) levels. All PTVs in the analysis are accounted for here.

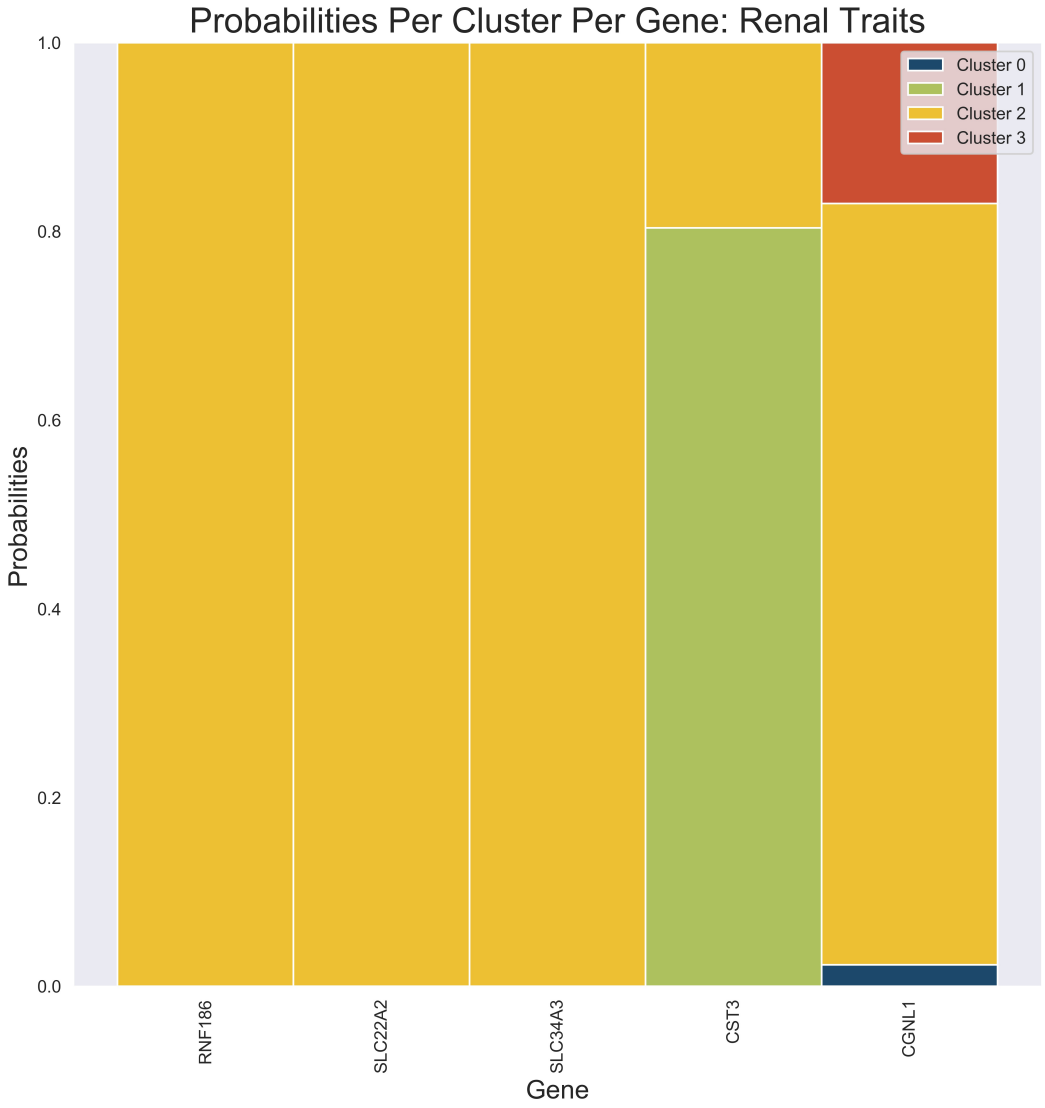


**Figure 5b:** Gene-level posterior probabilities that PTVs in candidate genes in the UK Biobank exome sequencing data set belong to each of four clusters hypothesized with respect to a renal-related multivariate phenotype of creatinine (CRE), cystatin C (CSTC), and effective glomerular filtration rate (eGFR). All PTVs in the analysis are accounted for here.

## 5 Code Availability

We have published the full set of associations ($\log_{10}$ BF $\geq 5$) from an independent effects model amongst PAVs, from a similar effects model amongst PAVs, as well as from a similar effects model amongst PTVs on the Global Biobank Engine[33]. While this study focuses on exome associations (`https://biobankengine.stanford.edu/RIVAS_HG38/mrpgene/all`), we also provide associations for array data (`https://biobankengine.stanford.edu/RIVAS_HG19/mrpgene/all`). For every phenotype, we provide single-phenotype MRPMM results (right-most columns in the tables) for PAVs and PTVs, with cluster effect size estimates as well as cluster assignment probabilities and proportions displayed.

Exome and array metadata tables are available on the Global Biobank Engine for direct download at these links:

`https://biobankengine.stanford.edu/static/ukb_exm_oqfe-consequence_wb_maf_gene_ld_indep_mpc_pli.tsv.gz` - Exome

`https://biobankengine.stanford.edu/static/ukb_cal-consequence_wb_maf_gene_ld_indep_mpc_pli.tsv.gz` - Array

MRPMM was implemented using Python (dependencies: pandas v1.1.5, numpy v1.16.4, sklearn 0.24.0, scipy v1.3.0). The requirements, code, example usages, and interpretation of results files can be found at `https://github.com/rivas-lab/mrpmm`.

## 6 References

1. Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J. A. Rare variants of *IFIH1*, a gene implicated in antiviral responses, protect against type 1 diabetes. Science **324**, 387–389 (2009).

2. 1000 Genomes Project Consortium. A map of human genome variation from population scale sequencing. Nature **467**, 1061–1073 (2010).

3. Rivas, M. A. et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nature Genetics **43**, 1066–1073 (2011).

4. The 1000 Genomes Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature **491**, 56–65 (2012).

5. Wu, M. C. et al. Rare-variant association testing for sequencing data with the sequence kernel association test. American Journal of Human Genetics **89**, 82–93 (2011).

6. Venkataraman, G. R. et al. Bayesian model comparison for rare variant association studies of multiple phenotypes. bioRxiv (2021).

7. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. The American Journal of Human Genetics **95**, 5–23 (2014).

8. Auer, P. L. & Lettre, G. Rare variant association studies: considerations, challenges and opportunities. Genome medicine **7**, 1–11 (2015).

9. Morgenthaler, S. & Thilly, W. G. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (cast). Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis **615**, 28–56 (2007).

10. Price, A. L. et al. Pooled association tests for rare variants in exon-resequencing studies. The American Journal of Human Genetics **86**, 832–838 (2010).

11. Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics **13**, 762–75 (2012).

12. Neale, B. M. et al. Testing for an unusual distribution of rare variants. PLoS genetics **7**, e1001322 (2011).

13. Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics **13**, 762–775 (2012).

14. Derkach, A., Lawless, J. F. & Sun, L. Robust and powerful tests for rare variants using fisher's method to combine evidence of association from two or more complementary tests. Genetic epidemiology **37**, 110–121 (2013).

15. Schwarz, G. Estimating the dimension of a model. The annals of statistics 461–464 (1978).

16. Sakamoto, Y., Ishiguro, M. & Kitagawa, G. Akaike information criterion statistics. Dordrecht, The Netherlands: D. Reidel **81**, 26853 (1986).

17. Rivas, M. A. et al. Assessing association between protein truncating variants and quantitative traits. Bioinformatics **29**, 2419–2426 (2013).

18. Rivas, M. A. et al. Effect of predicted protein-truncating genetic variants on the human transcriptome. Science **348**, 666–669 (2015).

19. Liu, D. J. et al. Meta-analysis of gene-level tests for rare variant association. Nature Genetics **46**, 200–204 (2014).

20. Cichonska, A. et al. metacca: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. Bioinformatics **32**, 1981–1989 (2016).

21. Hoff, P. D. A first course in Bayesian statistical methods (Springer Science & Business Media, 2009).

22. Geman, S. & Geman, D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. Pattern Analysis and Machine Intelligence, IEEE Transactions on 721–741 (1984).

23. Casella, G. & George, E. I. Explaining the Gibbs sampler. The American Statistician **46**, 167–174 (1992).

24. Hastings, W. K. Monte carlo sampling methods using markov chains and their applications. Biometrika **57**, 97–109 (1970).

25. Sinnott-Armstrong, N. et al. Genetics of 35 blood and urine biomarkers in the uk biobank. Nature genetics **53**, 185–194 (2021).

26. 1000_Genomes_Project_Consortium. A global reference for human genetic variation. Nature **526**, 68 (2015).

27. Bergström, A. et al. Insights into human genetic variation and population history from 929 diverse genomes. Science **367** (2020).

28. Mallick, S. et al. The simons genome diversity project: 300 genomes from 142 diverse populations. Nature **538**, 201–206 (2016).

29. Alexander, D. H. & Lange, K. Enhancements to the admixture algorithm for individual ancestry estimation. BMC bioinformatics **12**, 1–6 (2011).

30. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. Rfmix: a discriminative modeling approach for rapid and robust local-ancestry inference. The American Journal of Human Genetics **93**, 278–288 (2013).

31. Willer, C. J., Li, Y. & Abecasis, G. R. Metal: fast and efficient meta-analysis of genomewide association scans. Bioinformatics **26**, 2190–2191 (2010).

32. McLaren, W. et al. The ensembl variant effect predictor. Genome biology **17**, 1–14 (2016).

33. McInnes, G. et al. Global biobank engine: enabling genotype-phenotype browsing for biobank summary statistics. Bioinformatics **35**, 2495–2497 (2019).