S1 Supplemental Methods

S1.1 UK Biobank data

The UK Biobank data contains the whole genome sequencing for about 500,000 participants, with genotyping of 800,000 variants genome-wide and imputation to 90 million variants (Sudlow et al., 2015). Our study is conducted under an approved project from UK Biobank. We used genotyping data of 22 autosomes from individuals of European ancestry after applying a series of quality control (QC) procedures. We used two quantitative traits in our study, the direct low-density lipoproteins cholesterol (LDL, Data-Field 30780) and the high-density lipoprotein cholesterol (HDL, Data-Field 30760), as well as the mean corpuscular volume (MCV, Data-Field 30040) and the mean platelet volume (MPV, Data-Field 30100) for data transformation. `PLINK2` (Chang et al., 2015) was used to perform all QC steps, GWA analyses and LD calculation.

Quality control. For QC, we filtered out variants with minor allele frequency below 0.01, missing call rates exceeding 0.01, Hardy-Weinberg equilibrium exact test p-value below 1e-6, and with one or more multi-character allele codes or single-character allele codes outside of 'ACGT' (`--maf 0.01 --hwe 1e-6 --geno 0.01 --snps-only just-acgt`). We also excluded ambiguous SNPs, i.e., those with complementary alleles, either 'C/G' or 'A/T' SNPs. We excluded samples with missing call rates exceeding 0.05 (`--mind 0.05`), samples with putative aneuploidy, samples with excess third-degree relatives, outliers based on heterozygosity and missing rates, and samples that withdrawal from the study. We restricted our analysis to samples with genetic ethnic reported as "White British" from the UKB data (Data-Field 21000) that corresponds to the commonly labeled "European ancestry". After QC, a total of 489,953 variants remained for a total of 334,851 people (155,402 males, 179,449 females). The total genotyping rate is 0.954831. In calculations that involve the genotype matrix, any remaining missing genotype was filled by the average value of the non-missing ones.

GWA and LD calculation. In GWA studies, we included sex, age, body mass index (BMI), and the top 20 PCs from principal component analysis on the genotype data representing population structure as the covariates (a total of 23 covariates). PCA is performed using `flashpca` (Abraham et al., 2017) on the variants pruned using `PLINK2` command `--indep-pairwise 100 10 0.1`. In-sample LD was calculated on UKB data using command `--r square`.

S1.2 Input to *ML-MAGES*: GWA summary statistics and LD

Suppose the individual-level GWA data contains genotypes of m variants (SNPs) from a set of n diploid individuals, where each sample contains two sets of chromosomes. SNP genotypes are typically coded as 0, 1 or 2 to represent the number of non-reference alleles at each position. Let the genotype matrix be $X : n \times m$, where $X_{ij} \in \{0, 1, 2\}$, and let x_j be the j-th column of the matrix denoting the genotypes of variant j. Let $y \in \mathbb{R}^n$ be the phenotype, and assume it is standardized.

Let $\beta \in \mathbb{R}^m$ denote some unknown true effect sizes of the variants. Assuming that effects of each copy of an allele on the trait is additive (i.e., the "additive model") (Cantor et al., 2010; Uffelmann et al., 2021),

$$y = X\beta + \epsilon, \tag{1}$$

where $\epsilon \in \mathbb{R}^n$ is noise following $\epsilon_i \sim \mathcal{N}(0, \eta^2)$. In GWA studies, simple linear regression $y = x_j\beta_j + \epsilon$ is performed on each genetic variant across the genome on y, the trait of interest. The estimated effect is denoted as $\hat{\beta}_j$, with standard error se_j. The regression commonly includes covariates like age, sex, and top principal components of the genotype matrix which incorporate population structure.

The $m \times m$ in-sample LD matrix is the genetic correlation matrix. Each entry $R_{jj'}$ is the linkage disequilibrium (LD) between a pair of SNPs j and j', calculated as the correlation based on genotypic allele counts. The LD score of a SNP j, which quantifies the amount of genetic variation tagged by the variant, is the sum of its squared correlations with all other SNPs and can be calculated as $\ell_j = \sum_{j'=1, j' \neq j}^{m} R_{jj'}$ (Bulik-Sullivan et al., 2015).

Due to privacy concerns, the individual-level genotype data X is often not publicly accessible. However, there are abundant summary-level data of GWA summary statistics in public database like GWAS catalog (Buniello et al., 2018; Sollis et al., 2022). Methods only requiring summary-level data are therefore preferable to ensure their general applicability (Pasaniuc and Price, 2016). Similar to many existing methods (Thompson et al., 2015; Stephens, 2016; Zhu and Stephens, 2017; Zhang et al., 2018), we intend our method to use only summary-level data and LD information as input. LD may be obtained on out-of-sample reference population of the same ancestry (e.g., from 1000 Genomes Project (Fairley et al., 2019)) when individual-level data is not available.

S1.3 Related methods for effect size shrinkage

Shrinkage via mixture models. A group of methods use Gaussian mixture models to introduce sparsity (i.e., zeros) into effect sizes, essentially shrinking the inflated effects (Logsdon et al., 2010; Guan and Stephens, 2011; Zhou et al., 2013; Thompson et al., 2015; Stephens, 2016; Zhu and Stephens, 2017; Zhang et al., 2018; Holland et al., 2020). Many methods use a mixture of two components, usually zero-mean Gaussian distributions, with one of them having a point mass at zero to model the near-zero effects from null SNPs (Logsdon et al., 2010; Guan and Stephens, 2011; Zhou et al., 2013; Thompson et al., 2015; Zhu and Stephens, 2017; Zhang et al., 2018; Holland et al., 2020). Zhang et al. (2018) also extends this approach to three mixture components, allowing some non-null variants to have distinctly larger effects. The adaptive shrinkage method *ash* by Stephens (2016) uses a mixture of K components for the true effect size distribution with a point mass for truly null SNPs. Urbut et al. (2018) introduces the multivariate adaptive shrinkage method *mash* which uses multivariate Gaussian components, where the covariance matrices of the mixture components are designed to captures patterns of multiple groups of effects. In both *ash* and *mash*, the variance-covariance matrices of the K mixture components are generated and fixed for a pre-specified number K, while the mixture weights are estimated to show the distributions of effects. Both methods work with summary-level data and do not explicitly account for LD.

Fine-mapping. The goal of fine-mapping is similar to the goal of shrinkage, but on a much smaller scale: to pinpoint, within some small trait-associated region from a GWA study, a few variants that very likely contain a causal one. The focal regions are usually small chromosomal segments that show a lots of significant hits in GWA, where the aggregation of hits is likely due to high LD between the SNPs. Statistical methods for fine-mapping include identifying the lead variant in the associated region based on p-values and Bayesian approaches which assign posterior probabilities of causality to each variant (Spain and Barrett, 2015; Pasaniuc and Price, 2016; Schaid et al., 2018; Hormozdiari et al., 2014; Benner et al., 2016; Yang et al., 2023; Ghosal et al., 2024). *SuSiE* (Wang et al., 2020) formulates the task as a variable selection problem and uses a Bayesian model for sparse multiple regression to identify variants with non-zero effect; its extension *SuSiE-RSS* (Zou et al., 2022) allows the use of only summary data as input. However, in fine-mapping, the true effects are highly sparse; fine-mapping techniques rarely scale to the genome-wide shrinkage task, especially if the trait is highly polygenic. Additionally, Bayesian inference, often performed through MCMC, can become computationally prohibitive genome-wide.

Multi-trait analysis of GWAS (*MTAG*). *MTAG* is a method to jointly analyze summary statistics from GWA studies of multiple traits (Turley et al., 2018). It aims to improve the power to detect associations for each trait separately by leveraging genetic correlation across traits. Taking in GWA summary statistics from multiple traits, it provides trait-specific effect estimates for each SNP, as well as the association statistics for each trait. Its key assumption is that all SNPs share the same variance–covariance structure, which contrasts with the assumptions underlying our association clustering. While *MTAG* (Turley et al., 2018) provides separate effect estimates for each trait, which allows us to incorporate these effects into our analysis framework and compare, it neither performs shrinkage nor assumes variation in variances for SNPs with different association patterns.

Shrinkage as regularized regression: the *gene-ϵ* approach. The observed GWA effects ($\hat{\beta}$) are sometimes called marginal effects, and the true effects in the additive model (β) are sometimes called joint effects. A simple relationship between marginal and joint effects has been derived and used in many studies (Yang et al., 2012; Zhu and Stephens, 2017; Zhang et al., 2018; Holland et al., 2020):

$$\mathbb{E}[\hat{\beta}_j] = \sum_{j'=1}^{m} R_{jj'} \beta_{j'}, \tag{2}$$

or, in matrix form, $\mathbb{E}[\hat{\beta}] = R\beta$. Additional scaling factors are sometimes included for each variant in the linear summation according to their SNP heterozygosity (Holland et al., 2020) or GWA standard errors (Zhu and Stephens, 2017). The observed effect of a variant is a weighted sum of true effects of all variants, weighted by their genetic correlations (captured in $R_{jj'}$) to the focal variant.

Based on this linear relationship, *gene-ϵ* (Cheng et al., 2020) performed shrinkage through a regularized regression. Elastic net (Zou and Hastie, 2005) was shown to work best after comparing against LASSO (Tibshirani, 1996) and ridge regression (Hoerl and Kennard, 1970). The regularized $\tilde{\beta}$ was obtained by

$$\arg\min_{\tilde{\beta}} \frac{1}{m} \|\hat{\beta} - R\tilde{\beta}\|_2^2 + \lambda(\tilde{\beta}), \tag{3}$$

where $\|\cdot\|_1$ is the L1-norm, $\|\cdot\|_2$ is the L2-norm, and $\lambda(\cdot)$ is a regularization term that penalizes for the non-zero values in $\tilde{\beta}$. For elastic net, this is

$$\lambda(\tilde{\beta}) = \lambda_1 \left(\lambda_2 \|\tilde{\beta}\|_1 + (1 - \lambda_2)\|\tilde{\beta}\|_2^2/2 \right), \tag{4}$$

where λ_1 is a multiplier controlling the overall strength of the penalty, and λ_2 is the mixing parameter controlling the combination of L1 and L2. In Cheng et al. (2020), λ_2 was set to 0.5.

S1.4 Synthetic data generation and training data preparation

Generating realistic genome-wide genotype simulations is a complex task, as it involves producing both realistic allele frequency distributions and LD patterns. Simulating genotypes from scratch using a genetic data simulator (see National Cancer Institute (2024) for a list of resources) is beyond the scope of this study. Simulation based on real genotype data is a common practice in genetic studies (Zhu and Stephens, 2017; Zhang et al., 2018; Wang et al., 2020; Cheng et al., 2020; Zou et al., 2022). Here, we first sub-sampled genotypes from real chromosomal data, and then generated synthetic true effect sizes and trait values following the additive model in Eq. 1. The observed effects were obtained using GWA-like regressions on the synthetic trait. The real chromosomal data used in our simulations are from UK Biobank.

We took two approaches to simulate the true effects, one based on just SNP-level data, and one based on gene-level information as well. We refer to these as “SNP-only simulation” and “gene-level simulation”, respectively. For the SNP-only simulation, we subsampled regions of variants in a chromosome from a subset of individuals to get genotype matrices. We then simulated the true effects and noise terms and obtained synthetic traits. For the gene-level simulation, instead of sampling random regions from the chromosomes, an entire chromosome was used to simulate the data. When two or more traits were simulated to evaluate our multi-trait analysis, from all genes in the chromosome, some are chosen to be truly associated to one of the traits, and some are chosen to be associated to multiple. A fraction of variants in those chosen genes were set to be associated, i.e., with non-zero true effects. We vary parameters in the simulations such as the narrow-sense heritability, which is the ratio of additive genetic variance to the total phenotypic variance, the proportion of associated variants, and the influence of variant heterozygosity and LD-score on the variance of effect. Simulation steps are detailed below.

Next, GWA studies are performed on the simulated traits to generate simulated univariate summary-level data. By simulation, we generate both the ground truth effects, β^*, and the GWA summary statistics, $\hat{\beta}^*$

and se^*. SNP-only simulation data is then used as the training and validation data for training the NNs. We also use NP-only simulations to benchmark the performance of different methods for shrinkage. We use different chromosomes for simulating training data and validation data to ensure they do not overlap. Gene-level simulation is not used for training. We use it solely for performance comparison end-to-end, including shrinkage, clustering, and gene-level analysis (Fig. 2).

An issue we encountered was that simulated observed effects do not closely follow the distribution of real GWA effects. Despite the widely used additive model (Eq. 1), the true generative processes underling our genome is much more complicated than those suggested by the theoretical models. However, supervised learning models rely on a good alignment between training and test data to perform well. We bridged this gap between simulated data and real data by applying transformations on the simulated values, both $\hat{\beta}^*$ and se^*, to better approximate the distributions of empirical ones.

We noticed that the distribution of real GWA effects ($\hat{\beta}$) follows a Laplace distribution more closely than a normal one, so we fit a zero-centered Laplace distribution on $\hat{\beta}$ of the traits of interest. We also included an option to fit a non-symmetric Laplace instead. We then transformed simulated $\hat{\beta}^*$ values so that the empirical cumulative distribution follows this fitted Laplace distribution. For the standard errors, we simply re-scaled them so that the scales of synthetic se^* match the real ones. In this way, the simulated values, after transformation, distributed similarly to the real ones, and were ready to be used for model training.

Simulating effect sizes. When simulating effect sizes of causal variants, we used a flexible variance model that incorporates allele frequency and linkage disequilibrium (LD). For each variant j,

$$\text{Var}(\beta_j) \propto (2p_j(1-p_j))^s \, \ell_j^w, \tag{5}$$

where p_j is the minor allele frequency (MAF) of the variant, and ℓ_j is its LD-score; s and w are two parameters controlling the influence of variant heterozygosity ($2p_j(1-p_j)$) and LD-score on the variance. We considered various parameter settings, with $s \in \{-0.25, 0\}$ and $w \in \{-1, 0\}$. The models used for our downstream analyses were trained using $s = -0.25, w = 0$, and evaluated on all settings. $s = 0$ corresponds to no dependence on allele frequency, while $s = -0.25$ induces larger effects for rarer variants, reflecting empirical observations that rare variants can contribute disproportionately to trait variation. $w = 0$ corresponds to no LD dependence, whereas $w = -1$ down-weights variants in high-LD regions, consistent with the LDAK model (Speed et al., 2012, 2017) which assigns higher weights to SNPs located in low-LD regions. Together, these choices represent a range of various genetic architectures, from uniform models that ignore frequency and LD to architectures that favor rare variants in low-LD regions. For each causal variant, the effect size β_j is then drawn from a normal distribution, $\beta_j \sim \mathcal{N}\left(0, \text{Var}(\beta_j)\right)$.

SNP-only simulation. We subsampled several regions of $m^* = 1000$ variants in a chromosome from a subset of $N^* = 10000$ individuals, giving a genotype data matrix X^* of size $m^* \times N^*$. The SNP-only simulation procedure for each sampled X^* is as follows:

1. Set the simulation parameters: allele-frequency factor s, LD-score factor w, fraction of causal variants p_{causal}, and narrow-sense heritability h^2.
2. Randomly pick a set of causal variants, C, with the number of causal variants $|C| = \max(2, p_{\text{causal}} \times m^*)$.
3. Simulate true effects as $\beta^*_j \sim \mathcal{N}(0, \text{Var}(\beta_j))$ for $j \in C$ following Eq. 5 and $\beta^*_j = 0$ otherwise.
4. Re-scale all simulated effects β^* so that $\text{Var}(X^*\beta^*) = h^2$.
5. Simulate a noise vector ϵ^* of size N^*, with $\epsilon^*_i \sim \mathcal{N}(0, 1 - h^2)$.
6. Simulate synthetic traits as $y^* = X^*\beta^* + \epsilon^*$.
7. Perform simple regressions on m^* variants separately:

   $$y^* = x^*_j \beta^*_j + \epsilon^*,$$

   giving an observed effect $\hat{\beta}^*_j$ and its corresponding standard error se^*_j for each variant.
8. Subset the full LD matrix of the chromosome calculated using the real genotype data to corresponding variants to get LD of the simulated data, R^*, as the $m^* \times m^*$ matrix block.
9. Transform the simulated $\hat{\beta}^*$ and rescale se^* so that they distribute similarly to the real ones.
10. Construct features Ω^* from $\hat{\beta}^*$, se^*, and R^* for supervised learning following Eq. 2.

The transformation of the simulated betas $\hat{\beta}^*$ was done by first fitting a Laplace distribution to the real GWA effects (here we use the trait MCV), and then transforming the empirical distribution of simulated data to match that of the fitted Laplace. The simulated standard errors se^* were scaled by $\frac{s^*-a^*}{b^*-a^*}(b-a)+a$, where (a,b), and (a^*,b^*) are the (1%,99%)-quantile values of the real se and the simulated se^*. The quantile values instead of the minimum and maximum were used to make the scaling more robust to outliers.

For the training data, we generated 100 simulations for each set of parameter settings ($s \in \{-0.25, 0\}$, $w \in \{-1, 0\}$, $p_{\text{causal}} \in \{0.01, 0.05\}$, and $h^2 \in \{0.3, 0.7\}$, a total of 16 combinations), using Chromosomes 18, 19, 21, and 22, resulting in a total of $100 \times 16 \times 4 \times 1000 = 6,400,000$ data points. For the validation data, we generated 100 simulations using Chromosome 20, resulting in a total of $100 \times 16 \times 1000 = 1,600,000$ data points. The validation data was also used in SNP-level performance comparisons.

Gene-level simulation. Gene-level simulation was based on 15,250 genotyped variants on Chromosome 15 from UK Biobank, with 433 genes with at least two SNPs (we used the same gene list as in Cheng et al. (2020)). We generated the causality patterns of two simulated traits. For simulation, we randomly sampled 5% from all genes with at least 6 SNPs to be truly associated to either one or both of the traits. Either 30% of all the SNPs in each causal gene or two SNPs, whichever comes larger, were chosen to be causal. For those associated to both, true effects were drawn from a zero-mean bivariate normal that result in a high correlation—either positive or negative—between them. For variants only associated to one trait, we sampled the true effects from a zero-mean univariate normal. The narrow-sense heritability of both traits are fixed at $h^2 = 0.7$. The rest of the simulation procedure follows the same steps 4 to 10 as in the SNP-only simulation. We generated 100 gene-level simulations, each containing the effects of 15,250 variants for two traits. Altogether there were 200 single-trait simulations, and they were used for performance comparisons of shrinkage methods (Fig. 2).

S1.5 LD block decomposition

Both elastic net (Cheng et al., 2020) and *SuSiE-RSS* (Wang et al., 2020; Zou et al., 2022) are computationally expensive when shrinking many variants, e.g., all genotyped variants from an entire chromosome—even for the shortest autosome. Performing shrinkage using these methods on the entire LD matrix of a chromosome is therefore inefficient.

Variants far apart on the chromosome have relatively low LD, making the LD matrix close to a blockdiagonal matrix. It is therefore common to split the genome into nearly independent blocks of LD. While operations on the full LD can be computationally expensive, applying them to smaller matrices can make these operations faster and parallelizable (Berisa and Pickrell, 2015; Kim et al., 2017; Privé, 2021). We decomposed the LD matrix of each chromosome into multiple LD blocks and performed shrinkage on each block separately. This allowed us both to greatly reduce the problem size. The results obtained on each LD block were then aggregated together.

We adopted a method from Berisa and Pickrell (2015) to get approximately independent LD blocks. Each antidiagonal term of the LD is represented by the sum of its elements, where a low sum value indicates that variants on two sides of the antidiagonal are weakly linked. Next, a low-pass filter is passed on the antidiagonal sums to reduce noise and several minima are chosen to be the candidate segment boundaries, followed by a local search in the proximity of each minimum for fine-tuning. Number of blocks in each chromosome was chosen so that each block contains around 1000 variants.

S1.6 Neural network architectures and training settings

In our network models, each fully-connected hidden layer is followed by a batch normalization layer and a ReLU activation function. The final layer, which is also a fully-connected linear layer, outputs the desired regularized effect. To prevent over-fitting, drop-out layers with drop out rate of 0.2 were included. A graphical representation of the architectures is in Fig. S1.

We trained the model using the Adam optimizer with a learning rate of 10^{-4}. The objective function was the Mean Squared Error (MSE) loss. To place greater importance on accurately predicting non-zero effects, we also included an option to apply a higher weight to these values during training. We trained each model with a batch size of 100 for a maximum of 500 epochs and implemented early-stopping to prevent over-fitting. The training always terminated early.

Neural-network shrinkage performed consistently across network architectures with different number of layers and input sizes, and no specific architecture excels at all tasks. In this study, we compared a few architectures and chose the 2-layer and 3-layer architectures using top 15 variants in demonstrations. We did not delve into the exploitation of model parameters and training hyperparameters, as our focus lies primarily on demonstrating the practical application of such deep supervised learning methods for shrinkage. In general, while more complex models may possess greater computational power, they are also subject to a higher computational burden and an increased risk of overfitting to the specific training data. Therefore, we recommend selecting an architecture that maintains a good balance between complexity and generalization, ideally validated on small subsets of data.

S1.7 Probability density functions of distributions

Below are the probability density functions of the distributions used in the inference of zero-mean infinite-mixture model. A categorical distribution parameterized by π of size K is denoted as Cat(π), with pdf

$$f(z|\pi) = \text{Cat}(z|\pi) = \prod_{k=1}^{K} [\pi_k]^{z_k} . \tag{6}$$

A Beta distribution with shape parameters a and b is denoted as Beta(a, b), with pdf

$$f(v|a,b) = \text{Beta}(v|a,b) = \frac{1}{B(a,b)} v^{a-1}(1-v)^{b-1}, \tag{7}$$

where $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ is the beta function and $\Gamma(\cdot)$ is the gamma function.

A Wishart distribution with ν degrees of freedom and a $p \times p$ symmetric positive-definite scale matrix L is denoted as $\mathcal{W}(L,\nu)$, with pdf

$$f(A|L,\nu) = \mathcal{W}(A|L,\nu) = W(L,\nu)|A|^{\frac{\nu-p-1}{2}} \exp\{-\frac{1}{2}\text{Tr}(L^{-1}A)\}, \tag{8}$$

where $W(L,\nu) = \frac{1}{2^{\nu p/2}|L|^{\nu/2}\Gamma_p(\nu/2)} = 2^{-\nu p/2}\,|L|^{-\nu/2}\left(\pi^{\frac{p(p-1)}{4}} \prod_{j=1}^{p} \Gamma(\frac{\nu+1-i}{2})\right)^{-1}$.

S1.8 Variational inference

To reduce the problem size of clustering, we first removed all variants with zero and near-zero regularized effects. Let $\tilde{\beta} \in \mathbb{R}^{m \times p}$ be the effects of m variants on p traits. The clustering input is $\gamma \in \mathbb{R}^{J \times p}$, where

$$\gamma_i \in \{\tilde{\beta}_j \in \mathbb{R}^p | \tilde{\beta}_j \not\approx \mathbf{0} \,\text{for}\, j = 1, \ldots, m\} \tag{9}$$

are the J variants with non-zero effects for any of the p traits. We cluster the J variants according to their associations represented by distributions of their effect sizes. The latent indicator variables $\{z_i\}_{i=1}^{J}$ are what we want the clustering algorithm to output, where

$$z_{ik} = \begin{cases} 1 & \text{if variant } i \text{ belongs to cluster } k \\ 0 & \text{otherwise} \end{cases}, \tag{10}$$

for a total of K clusters to be determined by the model.

Under the zero-mean infinite-mixture model,

$$\gamma_i \sim \sum_{k=1}^{\infty} \pi_k \mathcal{N}(0, A_k^{-1}) \tag{11}$$

$$z_i \sim \text{Cat}(\pi), \tag{12}$$

$$\pi_k = v_k \prod_{\ell=1}^{k-1} (1 - v_\ell). \tag{13}$$

Note that v deterministically gives the values of π. We use the priors

$$v_k \sim \text{Beta}(a_0, b_0) = \text{Beta}(1, \alpha) \tag{14}$$

and

$$A_k \sim \mathcal{W}(L_0, \nu_0), \tag{15}$$

for v and A, where Beta(a, b) denotes a Beta distribution and $\mathcal{W}(L, \nu)$ denotes a Wishart distribution, and a_0, b_0, L_0, ν_0 are the hyperparameters.

We need to approximate the intractable posterior $p(z, v, A|\gamma)$ of our model. Variational inference (VI) (Jordan et al., 1999; Wainwright and Jordan, 2008) is a popular technique in Bayesian statistics and machine learning to approximate complex probability distributions by a "closest" simpler variational distribution, typically from a predefined family of simpler distributions, which is parameterized by variational parameters.

We approximated the posterior distribution of our model using VI. The variational variables are $\{z, v, A\}$: the latent indicators z, the parameters for mixture weights v, and the precision matrices of the Gaussian distributions A. The observed data are the non-zero regularized effects $\{\gamma\}$. A graphical representation of the model is included in Fig. S2. Specifically, we use the technique of mean-field inference and coordinate-ascent optimization to perform inference. Under mean-field assumptions, the variational distribution factorize over partitions of mutually independent latent variables. Coordinate ascent variational inference (CAVI) is a typical method to optimize the evidence lower bound (ELBO), in which the variational approximation of each partition of the latent variables is optimized in turn while holding the others fixed. This is sometimes also called variational expectation maximization, where coordinate updates on local variables correspond to the E-step and the updates on global variables correspond to the M-step in the classic EM. Optimization steps are detailed below.

VI for the zero-mean infinite Gaussian mixture model. The likelihood of data given the model is

$$p(\gamma|z, A) = \prod_{i=1}^{J} \prod_{k=1}^{\infty} \left[\mathcal{N}(0, A_k^{-1})\right]^{z_{ik}} . \tag{16}$$

The conjugate prior on variational variables is

$$\begin{aligned} p(z, \pi, A) &= p(z|\pi)\, p(v)\, p(A) \\ &= \prod_{i=1}^{J} \text{Cat}(z_i|\pi) \prod_{k=1}^{\infty} \text{Beta}(v_k|a_0, b_0) \prod_{k=1}^{\infty} \mathcal{W}(A_k|L_0, \nu_0) \\ &= \prod_{i=1}^{J} \prod_{k=1}^{\infty} \left[v_k \prod_{\ell=1}^{k-1} (1 - v_\ell) \right]^{z_{ik}} \times \prod_{k=1}^{\infty} \frac{v_k^{a_0-1}(1 - v_k)^{b_0-1}}{B(a_0, b_0)} \\ &\quad \times \prod_{k=1}^{\infty} W(L_0, \nu_0) |A_k|^{\frac{\nu_0-p-1}{2}} \exp\{-\frac{1}{2}\text{Tr}(L_0^{-1} A_k)\}. \end{aligned} \tag{17}$$

The target distribution is the posterior $p(z, v, A|x)$, and the variational distribution is $q(z, v, A)$. We use mean-field approximation and coordinate-ascent optimization to perform the inference. That is, we assume that q can factorize over parameters $\theta = \{z, v, A\}$. The mean field approximation of the posterior $p(z, v, A|\gamma)$ is

$$\begin{aligned} p(z, v, A|\gamma) \approx q(z, v, A) &= q(z)q(v)q(A) \\ &= \prod_{i=1}^{J} \text{Cat}(z_i|r_i) \prod_{k=1}^{K-1} \text{Beta}(v_k|a_k, b_k) \prod_{k=1}^{K} \mathcal{W}(A_k|L_v, \nu_k). \end{aligned} \tag{18}$$

In the approximation distribution, K is set to an appropriately large value, e.g., $K = 20$. The variational posterior truncated at this upper limit K can be a reasonable approximation to the infinite mixture (Blei and

Jordan, 2006).

In coordinate-ascent variational inference (CAVI), we iterate over each of the variables z, v, and A. For each variable set t, we set

$$q^*(t) \propto \exp \mathbb{E}_{q_{\theta_{\neq t}}} \left[\log p(t|\theta_{\neq t}, \mathbf{b})\right], \tag{19}$$

while keep all other $\{q_s(\cdot)\}_{s\neq t}$ fixed, where $\theta_{\neq t}$ denotes the sets of variables excluding t. The iteration stops when the algorithm converges to a local optimum of the non-convex ELBO objective. This can be thought of as a variational EM algorithms, where the variational E-step involves updating $q(z)$, and the variational M-step involves updating $q(v)$ and $q(A)$. The derivation of the optimization steps is similar to that of a non-zero multivariate mixture in Nickl (2020).

Variational E-step. Following CAVI, in each iteration, $q(z)$ is updated as

$$\begin{aligned}
\log q(z) &= \mathbb{E}_{q(v,A)}[\log p(\gamma, z, v, A)] + C \\
&= \mathbb{E}\left[\log \prod_{i=1}^{J} \prod_{k=1}^{\infty} \left[\mathcal{N}(\gamma_i|0, A_k^{-1})\right]^{z_{ik}}\right] + \mathbb{E}\left[\log \prod_{i=1}^{J} \text{Cat}(z_i|\pi)\right] + C \\
&= \sum_{i=1}^{J} \sum_{k=1}^{K} z_{ik} \log \rho_{ik} + C.
\end{aligned} \tag{20}$$

where $\mathbb{E}[\cdot]$ is with respect to $q(v, A)$, C denotes some constant, and

$$\log \rho_{ik} = \frac{1}{2}\mathbb{E}\left[\log|A_k|\right] - \frac{p}{2}\log(2\pi) - \frac{1}{2}\mathbb{E}[\gamma_i^T A_k \gamma_i] + \mathbb{E}[\log v_k] + \sum_{j=1}^{k-1} \mathbb{E}[\log(1 - v_j)]. \tag{21}$$

The expectations in Eq. 21 evaluate to

$$\begin{aligned}
\mathbb{E}_{q(v,A)}[\log|A_k|] &= \langle \log|A_k| \rangle = \sum_{j=1}^{p} \psi(\frac{\nu_k + 1 - j}{2}) + p\log 2 + \log|A_k| \\
\mathbb{E}_{q(v,A)}[\gamma_i^T A_k \gamma_i] &= \nu_k \text{Tr}(\gamma_i \gamma_i^T L_k) \\
\mathbb{E}_{q(v,A)}[\log v_k] &= \langle \log v_k \rangle = \psi(a_k) - \psi(a_k + b_k) \\
\mathbb{E}_{q(v,A)}[\log(1 - v_k)] &= \langle \log(1 - v_k) \rangle = \psi(b_k) - \psi(a_k + b_k).
\end{aligned} \tag{22}$$

where $\psi(x) = \frac{d}{dx}\log(\Gamma(x)) = \frac{\Gamma'(x)}{\Gamma(x)}$ is the digamma function.

The updated $q(z)$ follows a Categorical distribution

$$q(z) = \prod_{i=1}^{J} \prod_{k=1}^{K} r_{ik}^{z_{ik}}, \quad \text{with } r_{ik} = \frac{\rho_{ik}}{\sum_j \rho_{ij}}. \tag{23}$$

Variational M-step. Next, we need to update $q(v)$ and $q(A)$ in each iteration.

$$\begin{aligned}
\log q(v, A) &= \mathbb{E}_{q(z)}[\log p(\gamma, z, v, A)] + C \\
&= \mathbb{E}\left[\log \prod_{i=1}^{J} \prod_{k=1}^{\infty} \left[\mathcal{N}(\gamma_i | 0, A_k^{-1})\right]^{z_{ik}}\right] + \mathbb{E}\left[\log \prod_{i=1}^{J} \text{Cat}(z_i | \pi)\right] \\
&\quad + \mathbb{E}\left[\log \prod_{k=1}^{\infty} \text{Beta}(v_k | a_0, b_0)\right] + \mathbb{E}\left[\log \prod_{k=1}^{\infty} \mathcal{W}(A_k | L_0, \nu_0)\right] + C \\
&= \sum_{i=1}^{J} \sum_{k=1}^{K} r_{ik} \left(\frac{1}{2} \log |A_k| - \frac{1}{2} \gamma_i^T A_k \gamma_i\right) + \sum_{i=1}^{J} \sum_{k=1}^{K} \mathbb{E}\left[z_{ik} \log \pi_k\right] \\
&\quad + \sum_{k=1}^{K} \left[(a_0 - 1) \log v_k + (b_0 - 1) \log(1 - v_k)\right] \\
&\quad + \sum_{k=1}^{K} \left[\frac{\nu_0 - p - 1}{2} \log |A_k| - \frac{1}{2} \text{Tr}(L_0^{-1} A_k)\right] + C,
\end{aligned} \tag{24}$$

where $\sum_{k=1}^{K} \mathbb{E}\left[z_{ik} \log \pi_k\right] = \sum_{k=1}^{K} \left[\sum_{j=k+1}^{K} r_{ij} \log(1 - v_k)\right] + r_{ik} \log v_k$.

Separating the expression into terms that depend only on v and those that only depend on A, the approximate posterior becomes

$$\log q(v, A) = \log q(v) + \log q(A). \tag{25}$$

For v terms:

$$\begin{aligned}
\log q(v) &= \sum_{i=1}^{J} \sum_{k=1}^{K} \left[\left(\sum_{j=k+1}^{K} r_{ij} \log(1 - v_k)\right) + r_{ik} \log v_k\right] x \\
&\quad + \sum_{k=1}^{K} \left[(a_0 - 1) \log v_k + (b_0 - 1) \log(1 - v_k)\right] + C \\
&= \sum_{k=1}^{K} \left(a_0 - 1 + \sum_{i=1}^{J} r_{ik}\right) \log v_k + \left(b_0 - 1 + \sum_{i=1}^{J} \sum_{j=k+1}^{K} r_{ij}\right) \log(1 - v_k) + C.
\end{aligned} \tag{26}$$

Therefore, $q(v)$ is a product of Beta distributions $q(v) \propto \prod_{k=1}^{K-1} \text{Beta}(v_k | a_k, b_k)$, where the product stops at

$K-1$ as $q(v_k=1)=1$ by the truncated stick-breaking construction, and

$$a_k = a_0 + \sum_{i=1}^{J} r_{ik},\ b_k = b_0 + \sum_{i=1}^{J}\sum_{j=k+1}^{K} r_{ij}. \tag{27}$$

For A terms:

$$\begin{aligned}\log q(A) &= \sum_{i=1}^{J}\sum_{k=1}^{K} r_{ik}\left(\frac{1}{2}\log|A_k| - \frac{1}{2}\gamma_i^T A_k \gamma_i\right)\\ &\quad + \sum_{k=1}^{K}\left[\frac{\nu_0 - p - 1}{2}\log|A_k| - \frac{1}{2}\text{Tr}(L_0^{-1}A_k)\right] + C.\end{aligned} \tag{28}$$

Match the terms with the product of Wishart distributions

$$\begin{aligned}\log q(A) &= \sum_{k=1}^{K}\log\mathcal{W}(A_k|L_k,\nu_k)\\ &= \sum_{k=1}^{K}\log W(L_v,\nu_k) + \frac{\nu_k - p - 1}{2}\log|A_k| - \frac{1}{2}\text{Tr}(L_k^{-1}A_k),\end{aligned} \tag{29}$$

and use the property

$$x^T A x = \text{Tr}(xx^T A),$$

we can write $q(A)$ as a product of Wishart distributions $q(A) \propto \prod_{k=1}^{K}\mathcal{W}(A_k|L_k,\nu_k)$ with

$$L_k = \left(L_0^{-1} + \sum_{i=1}^{J} r_{ik}\gamma_i\gamma_i^T\right)^{-1},\ \nu_k = \nu_0 + \sum_{i=1}^{J} r_{ik}. \tag{30}$$

Convergence. The convergence of the optimization can be checked by evaluating the ELBO after each iteration of the E and M-steps:

$$\begin{aligned}\text{ELBO(q)} &= \sum_{z}\int_{v}\int_{A} q(z,v,A)\log\frac{p(\gamma,z,v,A)}{q(z,v,A)}\,dv\,dA\\ &= \mathbb{E}[\log p(\gamma|z,A)] + \mathbb{E}[\log p(z|\pi)] + \mathbb{E}[\log p(v)] + \mathbb{E}[\log p(A)]\\ &\quad - \mathbb{E}[\log q(z)] - \mathbb{E}[\log q(v)] - \mathbb{E}[\log q(A)],\end{aligned} \tag{31}$$

where $\mathbb{E}[\cdot]$ is with respect to $q(z,v,A)$. Each term in Eq. 31 can be evaluated separately; here we do not exhaustively enumerate the evaluations.

Local optimum. VI is known to converge to a local maximum. We therefore chose to run the inference procedure multiple times, each starting from a different initialization, and aggregated the multiple runs to reduce the chance of being trapped in sub-optimal clustering. This is done by using the majority K^* value from the runs and choosing the run that yields a model with lowest Bayesian information criterion (BIC) across all runs with the K^* value.

S1.9 Univariate gene enrichment test

The (univariate) gene-level test statistic is

$$\tilde{Q}_g = \tilde{\beta}_g^T \tilde{\beta}_g, \tag{32}$$

where the subscript $\cdot_g$ indicates all variants in the gene g. According to Cheng et al. (2020), such a quadratic form of SNP effects is commonly used in gene enrichment methods. Based on the SNP-level null hypothesis $H_0 : \mathbb{E}[\beta_i^2] \leq \sigma_\epsilon^2$, $\tilde{Q}_g$ is tested against the gene-level enrichment null hypothesis $H_0 : Q_g = 0$ that is dependent on σ_ϵ^2. The normality assumption for true effects allow a linear combination of chi-squared test statistics to be used to test the significance of the gene. Cheng et al. (2020) uses Imhof's method (IMHOF, 1961) implemented in `R`. Alternatively, tests like Davies (1980) and Liu et al. (2009) serve similar purpose. In this study, to compare the univariate gene-level results using different shrinkage methods, we implemented the enrichment test in `Python` using Liu et al. (2009) available through the package `chiscore`. We corrected the output p-values for multiple testing by controlling the controls the false discovery rates (FDR) at level $\alpha = 0.05$ (Benjamini and Hochberg, 1995).

Verification with biological processes of genes. We used the gene annotation of SNPs from NCBI's Reference Sequence (RefSeq) database (Pruitt, 2004) in the UCSC Genome Browser, same as in Cheng et al. (2020). The Gene Ontology (GO) knowledge base (Ashburner et al., 2000; Aleksander et al., 2023) provides a good resource to assess the gene enrichment test results. It contains knowledge about the biological processes that gene products may carry out, with which we can investigate the functions of genes that are identified as associated, possibly to varying degrees, to the traits of interest. We used the gene set enrichment analysis tool *Enrichr* (Chen et al., 2013) to verify identified genes against their known GO biological processes (Fig. 3C-D).

S1.10 Categorization of the association clusters

Variants in each clusters can roughly be grouped into three types: trait-specific, shared-association, and non-prioritized association. In the bivariate clustering case, each cluster has a variance-covariance matrix Σ_k. We computed the ratio between the largest and smallest eigenvalues of Σ_k. We also looked the angle between x- or y-axis and the vector represented by the eigenvector corresponding to the largest eigenvalue, which represents the direction of the major axis of the covariance ellipse of the Gaussian. If the ratio is large enough (e.g., > 5) and the angle is small (e.g., $< 15^\circ$ to either axis), then we designated the cluster to be trait-specific. If the cluster is not trait-specific, then we categorize it to have shared association. However, this type of clusters may sometimes contain variants that have large trait-specific effects, but get mixed together with those having large effects in both traits. We grouped them as "shared association", although finer-scaled categorization of association types is possible. For clusters that have Σ_k with both eigenvalues small, i.e., having a small $\text{Tr}(\Sigma_k)$, we treated them as non-prioritized. The threshold for $\text{Tr}(\Sigma_k)$ can be adapted based on observed patterns in the output.

When there are more than traits, Gaussians mixtures reside in higher-dimensional spaces, making geometric visualization more challenging. Nevertheless, by examining the angles between the axes of the hyper-ellipses and the trait axes, as well as the lengths of the hyper-ellipse axes projected onto the trait axes, we can establish criteria to identify clusters associated with specific trait(s).

S1.11 Comparison to a linear NN model

One advantage of using neural-network-based effect size shrinkage is that, unlike elastic net which assumes linear relationships between true and inflated effects, network architecture design like the *ML-MAGES 2L and 3L* can easily capture non-linearities among correlated effects of variants.

To assess the necessity of capturing non-linear relationships, we further evaluated the performance of ML-MAGES framework using a single-layer neural network that directly connects all the input features to the output. The size of this single layer is determined by the number of features. This special network is only capable of capturing linear relationships, thus serves as a baseline comparison for showing how important non-linearity is in addressing effect size inflation. The performance of this special network is labeled as LINEAR across all figures. It is not surprising that *ML-MAGES (2L) and ML-MAGES (3L)* consistently outperformed LINEAR (Fig. 2; Supplemental Figs. S3-S4), which demonstrates the importance of capturing non-linear relationships among correlated effects in effect size shrinkage.

S1.12 Performance in a simulated three-trait scenario

To show that our *ML-MAGES* framework is generally applicable to multi-trait association patterns, in addition to the performance comparison we conducted with two simulated traits, we extended our simulation to a three-trait scenario. The generation of simulation data closely follows that in Section S1.4, and is detailed below.

A total of 100 simulations are generated. In each simulation, using a sub-sampled set of $N^* = 10000$ individuals from the UKB genotype data in chormosome 15, we randomly select 5% of genes to be truly associated to some traits, categorized into five association types: associated with only one trait (three types: trait 1-specific, trait 2-specific, and trait 3-specific, each with 1/7 probability), associated with two traits (traits 1&2-shared, with 2/7 probability), or associated with all three (all traits-shared, with 2/7 probability). To limit scenario complexity, we excluded certain pairwise combinations for shared associations between two traits. Since traits are synthetic, we generalized this scenario by focusing on traits 1 and 2 without loss of generality. Within each such gene, either 30% of variants or 2 variants—whichever is larger—are simulated to be associated, i.e., having non-zero true effects.

For variants with trait-specific associations, their true effects were sampled from zero-mean normal distribution with variance in Eq. 5. For variants with shared associations among p traits, the true effects of the p traits were sampled simultaneously from a zero-mean multivariate distribution. The variance-covariance matrix for variant j is constructed by assigning its effect variance Var(β_j) from the univariate simulation (Eq. 5) to all entries in a square matrix V of size $p \times p$, then multiplying each pair of non-diagonal entries in this matrix (representing the covariance between two traits) by a sampled correlation value ρ so that the matrix is symmetric positive semi-definite, making V a valid variance-covariance matrix. In the three-trait scenario, for simplicity, we let correlations for all pairs to be the same $\rho \in [-0.5, -0.4] \cup [0.6, 1]$, sampled from the negative interval uniformly with 50% probability and from the positive interval uniformly with 50% probability. The narrow-sense heritability of all traits was fixed at $h^2 = 0.7$, and $s = -0.25, w = 0$ (Eq. 5), as in the univariate simulation.

Because we simulated three traits, there were 300 single-trait simulations in total. Supp. Fig. S3 shows the performances of different shrinkage methods. For gene-level multivariate analysis performance, we only included three of the five ground-truth association types as representatives: trait 1-specific, traits 1&2-

shared, and all traits-shared. In both SNP-level shrinkage and gene-level association pattern identification, our method *ML-MAGES* consistently achieved good performance.

S1.13 Additional simulation comparisons when training uses imputation data

To further demonstrate the reliability of the method, we trained the models using simulation based on imputed genotype data of UKB, which contains regions with much denser variants subjecting to much higher correlations than the genotyped data. We applied these models to simulated evaluation data based on genotyped data of Chromosome 15 (15,250 SNPs), same as before. The results are shown in Supp. Fig. S4.

The simulated model training data was generated based on imputation data from Chromosomes 7 to 22, which contains a total of 2,283,694 variants (the other chromosomes were not included for computational burden caused by processing large number of variants in imputation data). Simulated samples based on Chromosomes 10, 13, 17, and 19, all randomly chosen, were used in the validation set, and the rest in the training set. The difference in data density leads to significant variation in the correlations between variants. Highly-correlated markers are particularly subject to inflation, making it difficult to shrink their effect sizes, as these effects can easily surpass true effects due to high LD. As a result, it was hard for models trained on imputed data to fully reduce spurious associations to near-zero values while retaining the effects of putative true associations. To account for non-sparsity caused by marker density, we measured the performance by both PRC and weighted RMSE, where the weights are applied inversely proportional to the fraction of variants with true non-zero and zero effects.

In simulation, we varied the fraction of causal genes (f_{cg}) and the fraction of causal variants in each causal gene (f_{cs}), while fixing the fraction of causal variants not in any causal genes to be 0.1. The simulated effects were then transformed, as described in Supplemental Methods S1.4, to match the distribution of estimated GWA effects of the trait mean corpuscular volume (MCV) in the imputation data. Each simulated data sample was based on a chromosomal segment of 1,000 variants from the imputation data. We generate 150 such data samples for each of the chromosome, resulting in a total of 600,000 data samples from training chromosomes and 200,000 from validation chromosomes. From them, a total of 15,000 were subset and used for training, and 5,000 were used for validation. When effect values are very small, training can suffer from numerical instability of some weights or gradients in the neural networks; therefore, we scaled the simulated effects by 250, which equals one over the absolute value of the original transformed effects rounded to the nearest 10. Note that when applying the trained models, inputs were scaled up by 250 as well, and the corresponding outputs were then divided by 250 afterwards.

We evaluated the model's performance on simulation data generated from genotyped data. The difference between the simulation of training and evaluation data, as well as the difference in simulation settings, allowed us to examine robustness of *ML-MAGES* when the training simulations do not perfectly reflect the true underlying effect size distribution, which is often the case in practice. We also included the performance of elastic net shrinkage using untransformed synthetic effects, as the original elastic net method does not depend on the generation or transformation of training data.

The much higher density of the imputed variants caused significantly larger LD between them, which subsequently introduces large inflation in the effects, making the shrinkage problem much more challenging overall than in the genotyped data. Although the model's training performance was less optimal when using simulations based on imputed data, models trained on imputed data still consistently outperformed the elastic net method in shrinkage when applied to simulations based on genotyped data. This holds across various simulation settings that differ in the proportion of true causal genes and causal variants, as shown in Supp. Fig. S4.

S1.14 Implementation details

We implement the method and perform our analyses in `Python 3.9`. Neural networks are implemented using the package `PyTorch`. Elastic net shrinkage is implemented using the package `scikit-learn`. *SuSiE-RSS* (*susie*) is performed in `R` using the library `susieR` (Wang et al., 2020; Zou et al., 2022). *FINEMAP* (Benner et al., 2016) and *MTAG* (Turley et al., 2018) software are downloaded separately and run from command lines using the `subprocess` package of `Python`. We conducted both our training and non-training tasks on a machine with Linux 5.14 OS, x86-64 architecture, Intel Xeon Processor E5-2675, using 1 CPU node. The experiments in our study require no more than 18 GB of memory per chromosome for genotyped data. All analyses are parallelized and partitioned, for example, by chromosome, or by LD blocks from a chromosome.

References

Abraham G, Qiu Y, Inouye M. 2017. FlashPCA2: principal component analysis of biobank-scale genotype datasets. *Bioinformatics* **33**: 2776–2778.

Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, Ebert D, Feuermann M, Gaudet P, Harris NL, Hill DP, et al. 2023. The Gene Ontology knowledgebase in 2023. *Genetics* **224**: iyad031.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* **25**: 25–29.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **57**: 289–300.

Benner C, Spencer CC, Havulinna AS, Salomaa V, Ripatti S, Pirinen M. 2016. Finemap: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**: 1493–1501.

Berisa T, Pickrell JK. 2015. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32**: 283–285.

Blei DM, Jordan MI. 2006. Variational inference for dirichlet process mixtures. *Bayesian Analysis* **1**: 121–143.

Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Patterson N, Daly MJ, Price AL, Neale BM. 2015. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* **47**: 291–295.

Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, et al. 2018. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research* **47**: D1005–D1012.

Cantor RM, Lange K, Sinsheimer JS. 2010. Prioritizing gwas results: A review of statistical methods and recommendations for their application. *The American Journal of Human Genetics* **86**: 6–22.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**.

Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A. 2013. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**.

Cheng W, Ramachandran S, Crawford L. 2020. Estimation of non-null snp effect size distributions enables the detection of enriched genes underlying complex traits. *PLOS Genetics* **16**: e1008855.

Davies RB. 1980. Algorithm AS 155: The distribution of a linear combination of χ^2 random variables. *Applied Statistics* **29**: 323–333.

Fairley S, Lowy-Gallego E, Perry E, Flicek P. 2019. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Research* **48**: D941–D947.

Ghosal S, Schatz MC, Venkataraman A. 2024. Beatrice: Bayesian fine-mapping from summary data using deep variational inference. *Bioinformatics* **40**: btae590.

Guan Y, Stephens M. 2011. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics* **5**.

Hoerl AE, Kennard RW. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**: 55–67.

Holland D, Frei O, Desikan R, Fan CC, Shadrin AA, Smeland OB, Sundar VS, Thompson P, Andreassen OA, Dale AM. 2020. Beyond snp heritability: Polygenicity and discoverability of phenotypes estimated with a univariate gaussian mixture model. *PLOS Genetics* **16**: e1008612.

Hormozdiari F, Kostem E, Kang EY, Pasaniuc B, Eskin E. 2014. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**: 497–508.

IMHOF JP. 1961. Computing the distribution of quadratic forms in normal variables. *Biometrika* **48**: 419–426.

Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK. 1999. An introduction to variational methods for graphical models. *Machine Learning* **37**: 183–233.

Kim SA, Cho CS, Kim SR, Bull SB, Yoo YJ. 2017. A new haplotype block detection method for dense genome sequencing data based on interval graph modeling of clusters of highly correlated snps. *Bioinformatics* **34**: 388–397.

Liu H, Tang Y, Zhang HH. 2009. A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics & Data Analysis* **53**: 853–856.

Logsdon BA, Hoffman GE, Mezey JG. 2010. A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC Bioinformatics* **11**.

National Cancer Institute. 2024. Genetic simulation resources (GSR). `https://surveillance.cancer.gov/genetic-simulation-resources/`.

Nickl P. 2020. *Bayesian Inference for Regression Models using Nonparametric Infinite Mixtures*. Master's thesis, Technical University of Darmstadt, Darmstadt, Germany.

Pasaniuc B, Price AL. 2016. Dissecting the genetics of complex traits using summary association statistics. *Nature Reviews Genetics* **18**: 117–127.

Privé F. 2021. Optimal linkage disequilibrium splitting. *Bioinformatics* **38**: 255–256.

Pruitt KD. 2004. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* **33**: D501–D504.

Schaid DJ, Chen W, Larson NB. 2018. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics* **19**: 491–504.

Sollis E, Mosaku A, Abid A, Buniello A, Cerezo M, Gil L, Groza T, Güneş O, Hall P, Hayhurst J, et al. 2022. The NHGRI-EBI GWAS catalog: knowledgebase and deposition resource. *Nucleic Acids Research* **51**: D977–D985.

Spain SL, Barrett JC. 2015. Strategies for fine-mapping complex traits. *Human Molecular Genetics* **24**: R111–R119.

Speed D, Cai N, UCLEB Consortium, Johnson MR, Nejentsev S, Balding DJ. 2017. Reevaluation of snp heritability in complex human traits. *Nature Genetics* **49**: 986–992.

Speed D, Hemani G, Johnson MR, Balding DJ. 2012. Improved heritability estimation from genome-wide snps. *The American Journal of Human Genetics* **91**: 1011–1021.

Stephens M. 2016. False discovery rates: a new deal. *Biostatistics* **18**: 275–294.

Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, et al. 2015. Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine* **12**: e1001779.

Thompson WK, Wang Y, Schork AJ, Witoelar A, Zuber V, Xu S, Werge T, Holland D, Andreassen OA, Dale AM. 2015. An empirical bayes mixture model for effect size distributions in genome-wide association studies. *PLOS Genetics* **11**: e1005717.

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **58**: 267–288.

Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, Fontana MA, Nguyen-Viet TA, Wedow R, Zacher M, Furlotte NA, et al. 2018. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature Genetics* **50**: 229–237.

Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, Martin HC, Lappalainen T, Posthuma D. 2021. Genome-wide association studies. *Nature Reviews Methods Primers* **1**.

Urbut SM, Wang G, Carbonetto P, Stephens M. 2018. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nature Genetics* **51**: 187–195.

Wainwright MJ, Jordan MI. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning* **1**: 1–305.

Wang G, Sarkar A, Carbonetto P, Stephens M. 2020. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **82**: 1273–1300.

Yang J, Ferreira T, Morris AP, Medland SE, Madden PAF, Heath AC, Martin NG, Montgomery GW, Weedon MN, Loos RJ, et al. 2012. Conditional and joint multiple-snp analysis of gwas summary statistics identifies additional variants influencing complex traits. *Nature Genetics* **44**: 369–375.

Yang Z, Wang C, Liu L, Khan A, Lee A, Vardarajan B, Mayeux R, Kiryluk K, Ionita-Laza I. 2023. CARMA is a new bayesian model for fine-mapping in genome-wide association meta-analyses. *Nature Genetics* **55**: 1057–1065.

Zhang Y, Qi G, Park JH, Chatterjee N. 2018. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nature Genetics* **50**: 1318–1326.

Zhou X, Carbonetto P, Stephens M. 2013. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genetics* **9**: e1003264.

Zhu X, Stephens M. 2017. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *The Annals of Applied Statistics* **11**: 1561–1592.

Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **67**: 301–320.

Zou Y, Carbonetto P, Wang G, Stephens M. 2022. Fine-mapping from summary data with the "sum of single effects" model. *PLOS Genetics* **18**: e1010299.