Supplementary Note to "Multi-scale Inference of Genetic Trait Architecture using Biologically Annotated Neural Networks"

Pinar Demetci[1,2,*], Wei Cheng[2,3,*], Gregory Darnell[2,4], Xiang Zhou[5,6], Sohini Ramachandran[1-3], and Lorin Crawford[2,7,8,†]

1 Department of Computer Science, Brown University, Providence, RI, USA
2 Center for Computational Molecular Biology, Brown University, Providence, RI, USA
3 Department of Ecology and Evolutionary Biology, Brown University, Providence, RI, USA
4 Institute for Computational and Experimental Research in Mathematics (ICERM), Brown University, Providence, RI, USA
5 Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA
6 Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA
7 Department of Biostatistics, Brown University, Providence, RI, USA
8 Center for Statistical Sciences, Brown University, Providence, RI, USA

*** Authors Contributed Equally**
† Corresponding E-mail: lorin_crawford@brown.edu

Contents

1 Overview of Partially Connected Bayesian Neural Networks

Biologically annotated neural networks (BANNs) are feedforward Bayesian models with have partially connected architectures that are inspired by the hierarchical nature of biological enrichment analyses in GWA studies. The BANNs software takes in one of two data types from genome-wide association (GWA) studies: (*i*) individual-level data $\mathcal{D} = \{\mathbf{X}, \mathbf{y}\}$ where $\mathbf{X}$ is an $N \times J$ matrix of genotypes with J denoting the number of single nucleotide polymorphisms (SNPs) encoded as $\{0, 1, 2\}$ copies of a reference allele at each locus and $\mathbf{y}$ is an N-dimensional vector of quantitative traits (see Figure 1 in the main text); or (*ii*) GWA summary statistics $\mathcal{D} = \{\mathbf{R}, \widehat{\boldsymbol{\theta}}\}$ where $\mathbf{R}$ is a $J \times J$ empirical linkage disequilibrium (LD) matrix of pairwise correlations between SNPs and $\widehat{\boldsymbol{\theta}}$ are marginal effect size estimates for each SNP computed using ordinary least squares (OLS) (see Supplementary Figure 1). In either setting, the BANNs software also requires a predefined list of SNP-set annotations $\{\mathcal{S}_1, \ldots, \mathcal{S}_G\}$ to construct partially connected network layers that represent different scales of genomic units. In this section, we review the hierarchical probabilistic specification of the BANNs framework for individual data; however, note that extensions to summary statistics is straightforward and only requires substituting the genotypes $\mathbf{X}$ for the LD matrix $\mathbf{R}$ and substituting the phenotypes $\mathbf{y}$ for the OLS effect sizes $\widehat{\boldsymbol{\theta}}$.

Without loss of generality, let SNP-set g represent an annotated collection of SNPs $j \in \mathcal{S}_g$ with cardinality $|\mathcal{S}_g|$. The BANNs framework is probabilistically represented as a nonlinear mixed model

$$\mathbf{y} = \sum_{g=1}^{G} h(\mathbf{X}_g \boldsymbol{\theta}_g + \mathbf{1} b_g^{(1)}) w_g + \mathbf{1} b^{(2)}, \tag{1}$$

where $\mathbf{X}_g = [\mathbf{x}_1, \ldots, \mathbf{x}_{|\mathcal{S}_g|}]$ is the subset of SNPs annotated for SNP-set g; $\boldsymbol{\theta}_g = (\theta_1, \ldots, \theta_{|\mathcal{S}_g|})$ are the corresponding inner layer weights; $h(\bullet)$ denotes the nonlinear activations defined for neurons in the hidden layer; $\mathbf{w} = (w_1, \ldots, w_G)$ are the weights for the G-predefined SNP-sets in the hidden layer; $\mathbf{b}^{(1)} = (b_1^{(1)}, \ldots, b_G^{(1)})$ and $b^{(2)}$ are deterministic biases that are produced during the network training phase in the input and hidden layers, respectively; and $\mathbf{1}$ is an N-dimensional vector of ones. For convenience, we assume that the genotype matrix (column-wise) and trait of interest have been mean-centered and standardized. In the main text, $h(\bullet)$ is defined as a Leaky rectified linear unit (Leaky ReLU) activation function [1], where $h(\boldsymbol{x}) = \boldsymbol{x}$ if $\boldsymbol{x} > 0$ and $0.01\boldsymbol{x}$ otherwise. Throughout this Supplementary Note, we will equivalently write Eq. (1) in matrix notation as

$$\mathbf{y} = \mathbf{H}(\boldsymbol{\theta})\mathbf{w} + \mathbf{1} b^{(2)},$$

where $\mathbf{H}(\boldsymbol{\theta}) = [h(\mathbf{X}_1 \boldsymbol{\theta}_1 + \mathbf{1} b_1^{(1)}), \ldots, h(\mathbf{X}_G \boldsymbol{\theta}_G + \mathbf{1} b_G^{(1)})]$ denotes the matrix of nonlinear neurons in the hidden layer which are empirically computed given estimates the input layer weights. The hierarchical structure of the joint likelihood can be seen as a nonlinear take on classical integrative and structural regression models frequently used in GWA analyses [2–8].

As explained in the main text, we treat the weights of the input ($\boldsymbol{\theta}$) and hidden layers ($\mathbf{w}$) as random variables which allows for multi-scale genomic inference on both SNPs and SNP-sets, simultaneously. We assume that SNP-level effects follow a K-mixture of normal distributions

$$\theta_j \sim \sum_{k=1}^{K-1} \pi_{\theta k} \mathcal{N}(0, \sigma_{\theta k}^2) + \pi_{\theta K} \delta_0, \qquad \pi_{\theta K} = 1 - \sum_{k=1}^{K-1} \pi_{\theta k} \tag{2}$$

where δ_0 is a point mass at zero; $\boldsymbol{\sigma}_\theta^2 = (\sigma_{\theta 1}^2, \ldots, \sigma_{\theta K-1}^2)$ are variance of the $K-1$ nonzero mixture components; $\boldsymbol{\pi}_\theta = (\pi_{\theta 1}, \ldots, \pi_{\theta K-1})$ represents the marginal (unconditional) probability that a randomly selected SNP belongs to the k-th mixture component; and $\sum_{k=1}^{K-1} \pi_{\theta k}$ denotes the total proportion of SNPs that have a nonzero effect on the trait of interest. Notice that we write the mixture prior slightly

different from the main text to simplify updates in the algorithm for posterior inference. For reference, one can think of the K-th component as a normal distribution with fixed variance $\sigma^2_{\theta K} = 0$. Intuitively, specifying a larger K allows the neural network to learn general SNP effect size distributions spanning over a diverse class of trait architectures. For example, one can take a nonparametric approach and allow $K \to \infty$ such that Eq. (2) mirrors a Dirichlet process Gaussian mixture [9]. For results in the main text, we follow previous work and fix $K = 3$ [10–12]. This corresponds to the general hypothesis that SNPs can have large, moderate, and small effects on phenotypic variation [13]. For inference on the hidden layer, we assume that enriched SNP-sets contain at least one SNP with a nonzero effect. This simpler criterion is formulated by placing a spike and slab prior on the hidden weights

$$w_g \sim \pi_w \mathcal{N}(0, \sigma^2_w) + (1 - \pi_w)\delta_0. \tag{3}$$

where, due to the integrative form of the likelihood in Eq. (1), the magnitude of association for a SNP-set will be directly influenced by the effect size distribution of the SNPs it contains.

For all hyper-parameters in the model, we assume the following prior distributions

$$\log(\pi_{\theta k}) \sim \mathcal{U}(-\log(J), \log(1)), \qquad \sigma^2_{\theta k} \sim \text{Inv-Gamma}(u_\theta, v_\theta), \tag{4}$$

$$\log(\pi_w) \sim \mathcal{U}(-\log(G), \log(1)), \qquad \sigma^2_w \sim \text{Inv-Gamma}(u_w, v_w). \tag{5}$$

Following previous work [9], we set the shape and scale of the inverse-gamma distributions to be $u_\theta = u_w = 0.1$ and $v_\theta = v_w = 0.1$, respectively. Relatively uninformative uniform priors are placed on $\log \pi_{\theta k}$ and $\log \pi_w$ to reflect our lack of knowledge *a priori* about the proportion on associated SNP and SNP-sets with nonzero weights [14–16]. To facilitate posterior computation and interpretable inference, we also introduce two vectors of binary indicator variables $\boldsymbol{\gamma}_\theta = (\gamma_{\theta 1}, \ldots, \gamma_{\theta J}) \in \{0, 1\}^J$ and $\boldsymbol{\gamma}_w = (\gamma_{w1}, \ldots, \gamma_{wG}) \in \{0, 1\}^G$ where we implicitly assume *a priori* that

$$\Pr[\gamma_{\theta j} = 1] = \Pr[\theta_j \neq 0] = \sum_{k=1}^{K-1} \pi_{\theta k}, \qquad \Pr[\gamma_{wg} = 1] = \Pr[w_g \neq 0] = \pi_w. \tag{6}$$

Alternatively, we say $\gamma_{\theta j}$ and γ_{wg} take values of 1 when weights θ_j and w_g are drawn from the normal "slab", respectively; they take values of 0 otherwise. In the main text, we refer to these indicators as inclusion probabilities [17] and we use the marginal posterior means of these quantities as general summaries of evidence that SNPs and SNP-sets are statistically associated with phenotypic variation.

2 Variational Expectation-Maximization (EM) Algorithm

We modify a previously developed variational expectation-maximization (EM) algorithm to estimate the posterior distribution of parameters in the BANNs framework. The derivations in this section largely follow those developed in previous work [7, 9, 18–20]. As mentioned in the main text, the overall goal of variational inference is to approximate the true posterior distribution for network parameters with a similar distribution from an approximating family [21–25]. The EM algorithm we use aims to minimize the Kullback-Leibler divergence between the exact and approximate posterior distributions, respectively. To begin, we assign exchangeable uniform hyper-priors over a grid of values on the log-scale for $\boldsymbol{\pi}_\theta$ and π_w [15]. We then run the EM algorithm while iterating through each combination of these values. In the E-step, we use co-ordinate ascent to update the free parameters of the approximate variational posterior. In the M-step, we derive updates for the model hyper-parameters by solving for the roots of their gradients. Finally, in the last step, we empirically compute (approximate) posterior values for the network connection weights $(\boldsymbol{\theta}, \mathbf{w})$ and their corresponding inclusion probabilities $(\boldsymbol{\gamma}_\theta, \boldsymbol{\gamma}_w)$ by marginalizing over the different model combinations for $\boldsymbol{\pi}_\theta$ and π_w with normalized importance weights [19, 20]. A complete overview of the algorithm is given below.

Given the formulation of the BANNs model and the partially connected neural network architecture, the weights in the second layer are conditionally independent of the weights in the input layer given the activations (or outputs) from the first layer. This means that we can break up the model fitting procedure into two integrative parts and assess two different lower bounds for the input and hidden layer weights, respectively, to ensure convergence. Specifically estimates on the SNP-level are first maximized with respect to the trait of interest; while, parameters corresponding to the SNP-set level are maximized with respect to the observed trait. The software code iterates between the "inner" lower bound and the "outer" lower bound each step of the algorithm until convergence. Iterations in the algorithm are terminated when either one of two stopping criteria are met: (i) the difference between the lower bound of two consecutive updates are within some small range (specified by tolerance argument ϵ), or (ii) a maximum number of iterations is reached. For the simulations and real data analyses ran in this paper, we set $\epsilon = 1 \times 10^{-4}$ for the first criterion and used a maximum of 10,000 iterations for the second.

2.1 Input Layer (SNP-Level) Updates

For the SNP-level effects in the input layer, we aim to find a distribution $q(\boldsymbol{\theta}, \boldsymbol{\gamma}_\theta)$ that approximates the true posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{\gamma}_\theta \,|\, \mathcal{D})$, where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G)$ and $\mathcal{D}$ is used to denote the individual-level data and all relevant hyper-parameters. The similarity between these two distributions is maximized by minimizing the Kullback-Leibler (KL) divergence between them. This is formulated by

$$\mathrm{KL}(q(\boldsymbol{\theta}, \boldsymbol{\gamma}_\theta) \,\|\, p(\boldsymbol{\theta}, \boldsymbol{\gamma}_\theta \,|\, \mathcal{D})) = \int \log \left[\frac{q(\boldsymbol{\theta}, \boldsymbol{\gamma}_\theta)}{p(\boldsymbol{\theta}, \boldsymbol{\gamma}_\theta \,|\, \mathcal{D})} \right] q(\boldsymbol{\theta}, \boldsymbol{\gamma}_\theta) \, \mathrm{d}\boldsymbol{\theta} \, \mathrm{d}\boldsymbol{\gamma}_\theta. \tag{7}$$

In this work, we specify the following variational mixture distribution for each of the individual weights

$$q(\theta_j, \gamma_{\theta j}; \phi_j) = \begin{cases} \sum_{k=1}^{K-1} \alpha_{jk} \, \mathcal{N}(m_{jk}, s_{jk}^2) & \text{if} \quad \gamma_{\theta j} = 1 \\ (1 - \sum_{k=1}^{K-1} \alpha_{jk}) \, \delta_0 & \text{if} \quad \gamma_{\theta j} = 0 \end{cases} \tag{8}$$

where, in addition to previous notation, $\phi_j = \{\alpha_{jk}, m_{jk}, s_{jk}^2\}_{k=1}^{K-1}$ is a collection of free parameters. To compute the approximations, we make the mean-field assumption that the variational posterior can be factorized over $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_J)$ [26, 27]. The basic idea behind the variational approximation is to formulate a lower bound to the marginal likelihood, then to iteratively adjust the free parameters in $\boldsymbol{\phi}$ so that this bound becomes as tight as possible [19, 20, 25]. Finding the "best" factorized variational distribution amounts to finding the free parameters $\boldsymbol{\phi}$ that make the Kullback-Leibler divergence in Eq. (7) as small as possible. The specific class of variational distributions in Eq. (8) yields the following analytical expression for the lower bound on the inner layer (or SNP-level)

$$\begin{aligned} \mathrm{LB}(\boldsymbol{\pi}_\theta, \boldsymbol{\sigma}_\theta^2, \tau_\theta^2) = & -\frac{N}{2} \log(2\pi\tau_\theta^2) - \frac{1}{2\tau_\theta^2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_\theta\|_2^2 - \frac{1}{2} \sum_{j=1}^{J} (\mathbf{X}^\intercal \mathbf{X})_{jj} \mathbb{V}[\theta_j] \\ & - \sum_{j=1}^{J} \sum_{k=1}^{K} \alpha_{jk} \log \left(\frac{\alpha_{jk}}{\pi_{\theta k}} \right) + \frac{1}{2} \sum_{j=1}^{J} \sum_{k=1}^{K-1} \alpha_{jk} \left[1 + \log \left(\frac{s_{jk}^2}{\tau_\theta^2 \sigma_{\theta k}^2} \right) - \frac{s_{jk}^2 + m_{jk}^2}{\tau_\theta^2 \sigma_{\theta k}^2} \right] \end{aligned} \tag{9}$$

where $\tau_\theta^2 \approx \mathbb{V}[\mathbf{h}(\boldsymbol{\theta}) - \mathbf{X}\boldsymbol{\beta}_\theta]$ estimates the variance of residual training error in the input layer; $\| \bullet \|_2$ is the Euclidean norm; $\boldsymbol{\beta}_\theta$ is a J-dimensional estimate of the posterior mean for $\boldsymbol{\theta}$ with individual elements $\beta_{\theta j} = \sum_{k=1}^{K-1} \alpha_{jk} m_{jk}$; the term $(\mathbf{X}^\intercal \mathbf{X})_{jj}$ is the j-th diagonal component of the matrix $(\mathbf{X}^\intercal \mathbf{X})$; and $\mathbb{V}[\theta_j] = \sum_{k=1}^{K-1} \alpha_{jk}(m_{jk}^2 + s_{jk}^2) - (\sum_{k=1}^{K-1} \alpha_{jk} m_{jk})^2$ is the variance of the j-th weight under the approximating distribution in Eq. (8). As part of our contribution, we point out that the lower bound in Eq. (9) is implicitly maximized with respect to the hidden neurons during the backpropagation step in the algorithm. We now describe the expectation and maximization steps of the approximate EM algorithm below (see Software Details in Supplementary Note, Section 9).

1. **E-Step: Update the Variational Free Parameters.** In the E-step of the algorithm, we take the partial derivatives of the lower bound with respect to the free parameters in $\boldsymbol{\phi}$ and set them equal to zero. Solving for m_{jk}, s^2_{jk}, and α_{jk} yields the following co-ordinate ascent updates

$$m_{jk} = \frac{s^2_{jk}}{\tau^2_\theta}\left[(\mathbf{X}^\intercal\mathbf{y})_j - \sum_{l\neq j}(\mathbf{X}^\intercal\mathbf{X})_{jl}\beta_{\theta l}\right] \tag{10}$$

$$s^2_{jk} = \tau^2_\theta\left[(\mathbf{X}^\intercal\mathbf{X})_{jj} + \frac{1}{\sigma^2_{\theta k}}\right]^{-1} \tag{11}$$

$$\alpha_{jk} = \text{Sigmoid}\left(\log\left(\frac{\pi_{\theta k}}{1-\pi_{\theta k}}\right) + \log\left(\frac{s_{jk}}{\sigma_{\theta k}\tau_\theta}\right) + \frac{m^2_{jk}}{2s^2_{jk}}\right) \tag{12}$$

   where $\sum_k \alpha_{jk} \approx \Pr[\gamma_{\theta j} = 1 \,|\, \mathbf{y}, \mathbf{X}, \boldsymbol{\pi}_\theta, \sigma^2_\theta, \tau^2_\theta]$ and the sigmoid function is set to be the standard logistic function. Intuitively, the E-step of the algorithm for the input layer produces a collection of $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jK-1})$ values to determine whether each SNP has a nonzero effect on the phenotypic variance.

2. **M-Step: Update the Variance Hyper-Parameters.** In the M-step of the algorithm, we fix values of the variational free parameters and derive the following updates for each $\sigma^2_{\theta k}$ and τ^2_θ,

$$\sigma^2_{\theta k} = \left[\sum_{j=1}^{J}\sum_{k=1}^{K-1}\alpha_{jk}(m^2_{jk} + s^2_{jk})\right] \Big/ \left(\tau^2_\theta \sum_{j=1}^{J}\sum_{k=1}^{K-1}\alpha_{jk}\right) \tag{13}$$

$$\tau^2_\theta = \left(N + \sum_{j=1}^{J}\sum_{k=1}^{K-1}\alpha_{jk}\right)^{-1}\left[\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_\theta\|^2_2 + \sum_{j=1}^{J}(\mathbf{X}^\intercal\mathbf{X})_{jj}\mathbb{V}[\theta_j] + \sum_{j=1}^{J}\sum_{k=1}^{K-1}\frac{\alpha_{jk}}{\sigma^2_{\theta k}}(m^2_{jk} + s^2_{jk})\right] \tag{14}$$

   where N is equal to the dimensionality of the trait vector (i.e., the sample size when modeling individual-level data).

Following previous work [7,14–16,19,20], we account for our lack of *a priori* knowledge about the "correct" proportion of associated SNPs with nonzero effects by placing an exchangeable uniform hyper-prior distribution over an L-valued grid of possible values where $\{\boldsymbol{\pi}^{(1)}_\theta, \ldots, \boldsymbol{\pi}^{(L)}_\theta\} \in [1/J, 1]$. We then use the lower bound to the likelihood in Eq. (9) to approximate the posterior distribution of $\boldsymbol{\pi}_\theta$. Formally, we approximate $\Pr[\boldsymbol{\pi}_\theta = \boldsymbol{\pi}^{(l)}_\theta \,|\, \mathbf{y}, \mathbf{X}]$ with the normalized importance weights

$$\lambda^{(l)}_\theta = \frac{\text{LB}(\boldsymbol{\pi}^{(l)}_\theta, \boldsymbol{\sigma}^2_\theta, \tau^2_\theta)}{\sum_{l'=1}^{L}\text{LB}(\boldsymbol{\pi}^{(l')}_\theta, \boldsymbol{\sigma}^2_\theta, \tau^2_\theta)}. \tag{15}$$

As a final step in the model fitting procedure, we empirically compute (approximate) SNP-level posterior inclusion probabilities $\boldsymbol{\gamma}_\theta$ by marginalizing over the different grid combinations for $\boldsymbol{\pi}_\theta$. Namely,

$$\Pr[\gamma_{\theta j} = 1 \,|\, \mathbf{y}, \mathbf{X}] \approx \sum_{l=1}^{L}\lambda^{(l)}_\theta \Pr[\gamma_{\theta j} = 1 \,|\, \mathbf{y}, \mathbf{X}, \boldsymbol{\pi}^{(l)}_\theta, \boldsymbol{\sigma}^2_\theta, \tau^2_\theta]. \tag{16}$$

This final step can be viewed as an analogy to Bayesian model averaging where marginal distributions are estimated via a weighted average of conditional distributions multiplied by importance weights [19,20,28].

2.2 Outer Layer (SNP-Set Level) Updates

In this section, we detail the posterior computation for parameters in the outer layer of the partially connected neural network. We are now interested in finding a distribution $q(\mathbf{w}, \boldsymbol{\gamma}_w)$ that approximates the true posterior $p(\mathbf{w}, \boldsymbol{\gamma}_w \,|\, \mathcal{D}, \boldsymbol{\theta})$. Here, it is important to note that, due to the integrative setup of the joint likelihood used in the BANNs framework, the true posterior for the weights in the outer layer is conditionally dependent upon the posterior estimates for the weights in the input layer. Since we assume that enriched SNP-sets contain at least one SNP with a nonzero effect, we consider a simpler family of variational distributions

$$q(w_g, \gamma_{wg}; \psi_g) = \begin{cases} \alpha_g \, \mathcal{N}(m_g, s_g^2) & \text{if} \quad \gamma_{wg} = 1 \\ (1 - \alpha_g) \, \delta_0 & \text{if} \quad \gamma_{wg} = 0 \end{cases} \tag{17}$$

where $\psi_g = (\alpha_g, m_g, s_g^2)$ is used to describe a new set free parameters for the g-th SNP-set. Once again, our goal is to find the "best" factorized variational distribution amounts with free parameters $\boldsymbol{\psi}$ that minimize the Kullback-Leibler divergence between the exact and approximate posteriors. The specific class of variational distributions in Eq. (17) yields the following analytical expression for the lower bound on the outer layer (or SNP-set level)

$$\begin{aligned} \text{LB}(\pi_w, \sigma_w^2, \tau_w^2 \,|\, \boldsymbol{\theta}) = &- \frac{N}{2}\log(2\pi\tau_w^2) - \frac{1}{2\tau_w^2}\|\mathbf{y} - \mathbf{H}(\boldsymbol{\theta})\boldsymbol{\beta}_w\|_2^2 - \frac{1}{2}\sum_{g=1}^{G}\left\{\mathbf{H}(\boldsymbol{\theta})^\intercal\mathbf{H}(\boldsymbol{\theta})\right\}_{gg}\mathbb{V}[w_g] \\ &- \sum_{g=1}^{G}\alpha_g \log\left(\frac{\alpha_g}{\pi_w}\right) - \sum_{g=1}^{G}(1-\alpha_g)\log\left(\frac{1-\alpha_g}{1-\pi_w}\right) \\ &+ \frac{1}{2}\sum_{g=1}^{G}\alpha_g\left[1 + \log\left(\frac{s_g^2}{\sigma_w^2\tau_w^2}\right) - \frac{m_g^2 + s_g^2}{\sigma_w^2\tau_w^2}\right] \end{aligned} \tag{18}$$

where, similar to the input layer updates, $\tau_w^2 \approx \mathbb{V}[\mathbf{y} - \mathbf{H}(\boldsymbol{\theta})\boldsymbol{\beta}_w]$ estimates the variance of residual training error in the outer layer; the term $\boldsymbol{\beta}_w$ is a G-dimensional estimate of the posterior mean for $\mathbf{w}$ with elements $\beta_{wg} = \alpha_g m_g$ for the g-th SNP-set; the matrix $\mathbf{H}(\boldsymbol{\theta})^\intercal\mathbf{H}(\boldsymbol{\theta})$ is deterministically computed given posterior estimates of the weights $\boldsymbol{\theta}$ from the input layer, and $\mathbb{V}[w_g] = \alpha_g(m_g^2 + s_g^2) + \alpha_g^2 m_g^2$ is the variance of the g-th weight under the approximating distributional family in Eq. (17). We describe the explicit expectation and maximization steps of the approximate EM algorithm for the outer layer below.

1. **E-Step: Update the Variational Free Parameters.** In the E-step of the algorithm, we this time take the partial derivatives of the lower bound in Eq. (18) with respect to the free parameters in $\boldsymbol{\psi}$ and set them equal to zero. Solving for m_g, s_g^2, and α_g yields the following co-ordinate updates

$$m_g = \frac{s_g^2}{\tau_w^2}\left[\left(\mathbf{H}(\boldsymbol{\theta})^\intercal\mathbf{y}\right)_g - \sum_{l \neq g}\left\{\mathbf{H}(\boldsymbol{\theta})^\intercal\mathbf{H}(\boldsymbol{\theta})\right\}_{lg}\beta_{wl}\right] \tag{19}$$

$$s_g^2 = \tau_w^2\left[\left\{\mathbf{H}(\boldsymbol{\theta})^\intercal\mathbf{H}(\boldsymbol{\theta})\right\}_{gg} + \frac{1}{\sigma_w^2}\right]^{-1} \tag{20}$$

$$\alpha_g = \text{Sigmoid}\left(\log\left(\frac{\pi_w}{1-\pi_w}\right) + \log\left(\frac{s_g}{\sigma_w\tau_w}\right) + \frac{m_g^2}{2s_g^2}\right) \tag{21}$$

where $\alpha_g \approx \Pr[\gamma_{wg} = 1 \,|\, \mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \pi_w, \sigma_w^2, \tau_w^2]$ and, again, we set the sigmoid function to be the standard logistic function.

2. **M-Step: Update the Variance Hyper-Parameters.** In the M-step of the algorithm, we fix values of the variational free parameters and derive the following updates for σ_w^2 and τ_w^2 as

$$\sigma_w^2 = \left[\sum_{g=1}^{G} \alpha_g (m_g^2 + s_g^2)\right] \Big/ \left(\tau_w^2 \sum_{g=1}^{G} \alpha_g\right) \tag{22}$$

$$\tau_w^2 = \left(N + \sum_{g=1}^{G} \alpha_g\right)^{-1} \left[\|\mathbf{y} - \mathbf{H}(\boldsymbol{\theta})\boldsymbol{\beta}_w\|_2^2 + \sum_{g=1}^{G} \{\mathbf{H}(\boldsymbol{\theta})^\intercal \mathbf{H}(\boldsymbol{\theta})\}_{gg} \mathbb{V}[w_g] + \frac{1}{\sigma_w^2} \sum_{g=1}^{G} \alpha_g (m_g^2 + s_g^2)\right] \tag{23}$$

where, again, N is equal to the dimensionality of the phenotypic response vector $\mathbf{y}$.

Similar to the algorithmic updates in the input layer, we account for our lack of *a priori* knowledge about the "correct" proportion of enriched SNP-sets by placing another exchangeable uniform hyper-prior distribution over an L-valued grid of possible values where $\{\pi_w^{(1)}, \ldots, \pi_w^{(L)}\} \in [1/G, 1]$. Here, we now use the variational lower bound in Eq. (18) to approximate the posterior distribution of π_w. As a final step in the model fitting procedure, we again conduct a Bayesian model averaging-like procedure by integrating over the different grid combinations for π_w and computing marginal posterior inclusion probabilities for each of the G-annotated SNP-sets as the following

$$\Pr[\gamma_{wg} = 1 \,|\, \mathbf{y}, \mathbf{X}, \boldsymbol{\theta}] \approx \sum_{l=1}^{L} \lambda_w^{(l)} \Pr[\gamma_{wg} = 1 \,|\, \mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \pi_w^{(l)}, \sigma_w^2, \tau_w^2]. \tag{24}$$

where each importance weight $\lambda_w^{(l)}$ takes on a form similar to the normalized ratio described in Eq. (14).

Algorithm 1 BANNs Model with Individual Level Data

Input genotype data $\mathbf{X}$, continuous trait $\mathbf{y}$, and annotations $\{\mathcal{S}_1, \ldots, \mathcal{S}_G\}$.
Choose the number of models L, number of maximum iterations T, and tolerance parameter ϵ. Randomly initialize variational parameters $\phi_j = \{\alpha_{jk}, m_{jk}, s_{jk}^2\}_{k=1}^{K-1}$ and τ_θ^2, and $\psi_g = (\alpha_g, m_g, s_g^2)$ and τ_w^2 for the inner and outer layer, respectively, across the L models.
for $l = 1 \rightarrow L$ **do**
    Compute hidden layer neurons $\mathbf{H}(\boldsymbol{\theta})$.
    Compute inner and outer lower bounds `LB_inner_new` and `LB_outer_new`.
  for $t = 1 \rightarrow T$ **do**
      Set `LB_inner` = `LB_inner_new` and `LB_outer` = `LB_outer_new`.
      Update inner layer parameters $\{\phi_j, \theta_j\}$ for $j = 1, \ldots, J$ SNPs.
      Update hidden layer neurons $\mathbf{H}(\boldsymbol{\theta})$.
      Update outer layer parameters $\{\psi_g, w_g\}$ for $g = 1, \ldots, G$ SNP-sets.
      Update lower bounds `LB_inner_new` and `LB_outer_new`.
    if `LB_inner_new` $-$ `LB_inner` $\leq \epsilon$ and `LB_outer_new` $-$ `LB_outer` $\leq \epsilon$ **then**
        Break
    end if
  end for
  Save maximized lower bounds.
end for
Compute normalized importance weights $\lambda_\theta^{(l)}$ and $\lambda_g^{(l)}$ for $l = 1, \ldots, L$ models.
Compute (marginal) posterior means for network weights $\boldsymbol{\theta}$ and $\mathbf{w}$
Compute (marginal) posterior inclusion probabilities (PIPs) $\boldsymbol{\gamma}_\theta$ and $\boldsymbol{\gamma}_w$.
Compute the phenotypic variance explained by the input and hidden layers PVE($\boldsymbol{\theta}$) and PVE($\mathbf{w}$).
Return $\{\boldsymbol{\theta}, \mathbf{w}, \boldsymbol{\gamma}_\theta, \boldsymbol{\gamma}_w, \text{PVE}(\boldsymbol{\theta}), \text{PVE}(\mathbf{w})\}$.

Algorithm 2 BANN-SS Model with GWA Summary Statistics

Input LD matrix $\mathbf{R}$, OLS effect size estimates $\widehat{\boldsymbol{\theta}}$, and annotations $\{\mathcal{S}_1, \ldots, \mathcal{S}_G\}$.
Choose the number of models L, number of maximum iterations T, and tolerance parameter ϵ.
Randomly initialize variational parameters $\phi_j = \{\alpha_{jk}, m_{jk}, s^2_{jk}\}_{k=1}^{K-1}$ and τ^2_θ, and $\psi_g = (\alpha_g, m_g, s^2_g)$ and τ^2_w for the inner and outer layer, respectively, across the L models.
for $l = 1 \rightarrow L$ **do**
    Compute hidden layer neurons $\mathbf{H}(\boldsymbol{\theta})$.
    Compute inner and outer lower bounds `LB_inner_new` and `LB_outer_new`.
  for $t = 1 \rightarrow T$ **do**
      Set `LB_inner` = `LB_inner_new` and `LB_outer` = `LB_outer_new`.
      Update inner layer parameters $\{\phi_j, \theta_j\}$ for $j = 1, \ldots, J$ SNPs.
      Update hidden layer neurons $\mathbf{H}(\boldsymbol{\theta})$.
      Update outer layer parameters $\{\psi_g, w_g\}$ for $g = 1, \ldots, G$ SNP-sets.
      Update lower bounds `LB_inner_new` and `LB_outer_new`.
    if `LB_inner_new` $-$ `LB_inner` $\leq \epsilon$ and `LB_outer_new` $-$ `LB_outer` $\leq \epsilon$ **then**
        Break
    end if
  end for
  Save maximized lower bounds.
end for
Compute normalized importance weights $\lambda^{(l)}_\theta$ and $\lambda^{(l)}_g$ for $l = 1, \ldots, L$ models.
Compute (marginal) posterior means for network weights $\boldsymbol{\theta}$ and $\mathbf{w}$
Compute (marginal) posterior inclusion probabilities (PIPs) $\boldsymbol{\gamma}_\theta$ and $\boldsymbol{\gamma}_w$.
Compute the phenotypic variance explained by the input and hidden layers PVE($\boldsymbol{\theta}$) and PVE($\mathbf{w}$).
Return $\{\boldsymbol{\theta}, \mathbf{w}, \boldsymbol{\gamma}_\theta, \boldsymbol{\gamma}_w, \text{PVE}(\boldsymbol{\theta}), \text{PVE}(\mathbf{w})\}$.

3 Accounting for Non-Additive Genetic Effects

As mentioned in the main text, the BANNs framework jointly models the proportion of phenotypic variance that is explained by sparse genetic effects (both additive and non-additive) and random effects collectively [15]. This primarily done through the inclusion of the nonlinear Leaky ReLU activation function $h(\bullet)$ in the hidden layer [1]. In other areas of statistical genetics, similar nonlinear functions have been used to model non-additive random effects that contribute to phenotypic variation [29–35]. For example, it has been shown that the Taylor series expansion of the Gaussian kernel function enumerates all higher-order interaction terms between SNPs [36–39], thus alleviating potential combinatorial concerns with exhaustive searches [40]. The ReLU function family, generally defined as $h(\mathbf{X}_g\boldsymbol{\theta}_g) = \max(0, \mathbf{X}_g\boldsymbol{\theta}_g)$ for SNPs in the g-th SNP-set, shares this same property. To see this, we take the infinitely differentiable smooth ReLU (softplus) approximation $h(\mathbf{X}_g\boldsymbol{\theta}_g) \approx \log(1 + \exp\{\mathbf{X}_g\boldsymbol{\theta}_g\})$ such that function can be rewritten with infinite terms like the Gaussian kernel. The Taylor series expansion of the inside term is

$$1 + \exp\{\mathbf{X}_g\boldsymbol{\theta}_g\} = 1 + \sum_{m=1}^{\infty} \frac{1}{m!} \left(\sum_{j=1}^{|\mathcal{S}_g|} \mathbf{x}_j \theta_j \right)^m, \tag{25}$$

where the term on the right hand side includes the summation of first order effects in the form $\mathbf{x}_j^\intercal \mathbf{x}_{j'}$ for the j-th and j'-th SNP, and also includes the $m \rightarrow \infty$ higher-order (polynomial) interactions between SNPs. Notice that the first order effect terms are elements of the conventional relatedness matrix in linear mixed models, which has been well known to effectively control for population structure in genetic association studies [41–45]. Through our simulation studies, we demonstrate the ability to accurately

prioritize/rank associated SNPs and enriched SNP-sets in the BANNs framework, both in the presence of pairwise SNP-by-SNP interactions, as well as random effects driven by population structure.

4 Estimating Phenotypic Variance Explained (PVE)

As described in the main text, we are able to provide an estimate of phenotypic variance explained (PVE) within the BANNs framework as the total proportion of phenotypic variance that is explained by sparse genetic effects (both additive and non-additive) and random effects collectively [15]. Given the true values of the neural network parameters, we define this proportion on the SNP-level in the inner layer and SNP-set level in the outer layer as the following

$$\text{PVE}(\boldsymbol{\theta}) \approx \frac{\mathbb{V}[\mathbf{X}\boldsymbol{\theta}]}{\mathbb{V}[\mathbf{y}]}, \qquad \text{PVE}(\mathbf{w}) \approx \frac{\mathbb{V}[\mathbf{H}(\boldsymbol{\theta})\mathbf{w}]}{\mathbb{V}[\mathbf{y}]}, \tag{26}$$

where, as a reminder, $\mathbb{V}[\bullet]$ is the variance function and $\mathbf{H}(\boldsymbol{\theta}) = [h(\mathbf{X}_1\boldsymbol{\theta}_1 + \mathbf{1}b_1^{(1)}), \ldots, h(\mathbf{X}_G\boldsymbol{\theta}_G + \mathbf{1}b_G^{(1)})]$ denotes the matrix of deterministic nonlinear neurons in the hidden layer given estimates of the input layer weights. In practice, we estimate PVE using posterior values of the network parameters derived from the variational EM algorithm described in the previous section. Specifically, after averaging over the grid of different models, we use the (approximate) marginal posterior means $\boldsymbol{\beta}_\theta$ and $\boldsymbol{\beta}_w$ for the input and outer layer weights from Eqs. (9) and (17), respectively. We also approximate the variance of residual error that is observed during the training phase of both layers with estimates of τ_θ^2 and τ_w^2 from Eqs. (14) and (23). This yields the following empirical estimate for the PVE of complex traits

$$\text{PVE}(\boldsymbol{\theta}) \approx \frac{\mathbb{V}[\mathbf{X}\boldsymbol{\beta}_\theta]}{\mathbb{V}[\mathbf{X}\boldsymbol{\beta}_\theta] + \tau_\theta^2}, \qquad \text{PVE}(\mathbf{w}) \approx \frac{\mathbb{V}[\mathbf{H}(\boldsymbol{\beta}_\theta)\boldsymbol{\beta}_w]}{\mathbb{V}[\mathbf{H}(\boldsymbol{\beta}_\theta)\boldsymbol{\beta}_w] + \tau_w^2}, \tag{27}$$

where the matrix hidden neurons is empirically estimated as $\mathbf{H}(\boldsymbol{\beta}_\theta) = [h(\mathbf{X}_1\boldsymbol{\beta}_{\theta 1} + b_1^{(1)}), \ldots, h(\mathbf{X}_G\boldsymbol{\beta}_{\theta G} + b_G^{(1)})]$. Note that this formula is similar to the traditional form used for estimating PVE, except here we also consider the contribution of both non-additive and random genetic effects through the nonlinear Leaky ReLU activation function $h(\bullet)$ [1]. Through various simulations, we demonstrate the ability to accurately estimate PVE in the BANNs framework under additive sparse architectures (see Supplementary Figs. 24 and 25). We underestimate PVE in both polygenic traits and traits with pairwise SNP-by-SNP interactions, which we believe is caused by a misestimation of the approximate posterior mean for network weights during the variational EM algorithm. Similar observations have been noted when using variational inference [9, 19, 25, 46]. Results from other work also suggest that the sparsity assumption on the SNP-level effects can lead to the underestimation of the PVE [14, 15].

5 Data Quality Control Procedures for Stock of Mice

Some of the real data analysis results in this work made use of GWA data from the Wellcome Trust Centre for Human Genetics (`http://mtweb.cs.ucl.ac.uk/mus/www/mouse/index.shtml`). This study contains N = 1,814 heterogenous stock of mice from 85 families (all descending from eight inbred progenitor strains) [47], and 131 quantitative traits that are classified into 6 broad categories including behavior, diabetes, asthma, immunology, haematology, and biochemistry (`http://mtweb.cs.ucl.ac.uk/mus/www/GSCAN/index.shtml/index.old.shtml`). In the main text, we focused on six specific phenotypes from these categories including: body mass index (BMI) (`Obesity.BMI`), body weight (`Glucose.BodyWeight`), percentage of CD8+ cells (`Imm.PctCD8`), mean corpuscular hemoglobin (MCH) (`Haem.MCH`), high-density lipoprotein content (`Biochem.HDL`), and low-density lipoprotein content (`Biochem.LDL`). All phenotypes were previously corrected for sex, age, body weight, season, year, and cage effects [47]. For individuals

with missing genotypes, we imputed values by the mean genotype of that SNP in their corresponding family. Only polymorphic SNPs with minor allele frequency above 5% were kept for the analyses. This left a total of $J = 10{,}227$ autosomal SNPs that were available for all mice. For annotations, we used the Mouse Genome Informatics database (`http://www.informatics.jax.org`) to map SNPs to the closest neighboring gene(s). Here, pseudogenes, quantitative trait loci (QTL), and genes with only 1 annotated SNP within their boundary were excluded from the analyses. Unannotated SNPs located within the same genomic region were labeled as being within the "intergenic region" between two genes. Altogether, a total of $G = 1{,}925$ SNP-sets were analyzed.

6 Data Quality Control Procedures for Framingham Heart Study

The other real data analysis results made use of human GWA data from the Framingham Heart Study (`https://www.ncbi.nlm.nih.gov/gap`) [48]. This study originally contains $N = 6{,}950$ individuals and $J = 394{,}174$ SNPs. For quality control on these data, we removed (i) SNPs with minor allele frequency less than 2.5%, (ii) SNPs not in Hardy-Weinberg Equilibrium (Fisher's exact test $P > 1 \times 10^{-4}$), and (iii) SNPs in high linkage disequilibrium (using the flag `--indep-pairwise 50 5 0.8` with PLINK 1.9 [49]). This resulted in a final dataset containing $J = 372{,}131$ SNPs, where any missing values for a given SNP were imputed by using the estimated mean genotype of that SNP. Next, we used the NCBI's Reference Sequence (RefSeq) database in the UCSC Genome Browser [50] to annotate SNPs with appropriate genes. Gene annotations with only 1 SNP within their boundary were excluded from all analyses. Unannotated SNPs located within the same genomic region were labeled as being within the "intergenic region" between two genes. Altogether, a total of $G = 18{,}364$ SNP-sets were analyzed—which included 8,658 intergenic SNP-sets and 9,706 annotated genes.

7 Data Quality Control Procedures for UK Biobank

The simulation results and lipoprotein replication study presented in the main text made use of imputed data released from the UK Biobank [51]. Quality control procedures for these data are as follows. First, we only studied individuals who self-identified as "white British" people. From this cohort, we further excluded individuals identified by the UK Biobank to have high heterozygosity, excessive relatedness, or aneuploidy (1,550 individuals removed). We also removed individuals whose kinship coefficient was greater than 0.0442 (i.e., close relatives). Next, we removed (i) monomorphic SNPs, (ii) SNPs with minor allele frequency less than 2.5%, (iii) SNPs not in Hardy-Weinberg Equilibrium (Fisher's exact test $P > 1 \times 10^{-6}$), (iv) SNPs with missingness greater than 1%, and (v) SNPs in high linkage disequilibrium (using the flag `--indep-pairwise 50 5 0.9` with PLINK 1.9 [49]). After all QC steps, we had a final dataset of $N = 349{,}414$ individuals and $J = 1{,}070{,}306$ SNPs. Next, we used the NCBI's Reference Sequence (RefSeq) database in the UCSC Genome Browser [50] to annotate SNPs with appropriate genes. Gene annotations with only 1 SNP within their boundary were excluded from all analyses. Unannotated SNPs located within the same genomic region were labeled as being within the "intergenic region" between two genes. Altogether, a total of $G = 28{,}644$ SNP-sets were analyzed.

8 Simulation Setup and Scenarios

In our simulation studies, we used the following general simulation scheme to generate quantitative traits using real genotype data on chromosome 1 from ten thousand randomly sampled individuals of European ancestry in the UK Biobank [51]. This setup follows mostly from previous studies [13, 38–40]. We will denote this genotype matrix as $\mathbf{X}$, with $\mathbf{x}_j$ denoting the genotypic vector for the j-th SNP. Following quality control procedures detailed in the previous section, our simulations included $J = 36{,}518$ SNPs

distributed across genome. Again, we used the NCBI's RefSeq database in the UCSC Genome Browser to assign SNPs to genes which resulted in 1,408 genes to be used in the simulation study. We also consider the unnanotated SNPs between two genes to be located within intergenic regions. Altogether, a total of $G = 2{,}816$ SNP-sets were analyzed.

After the annotation step, we assume that all simulated traits have been standardized such that $\mathbb{V}[\mathbf{y}] = 1$ and that all observed genetic effects explain a fixed proportion of this value (i.e., broad-sense heritability, H^2). Next, we use the $N \times J$ matrix of genotypes $\mathbf{X}$ to generate real-valued phenotypes that mirror genetic architectures affected by a combination of linear (additive) and interaction (epistatic) effects. We randomly select a certain percentage of truly associated SNP-sets and denote the SNPs that they contain as $\mathcal{C}$. Within $\mathcal{C}$, we select causal SNPs in a way such that each associated SNP-set contains at least two SNPs with non-zero effects. The additive effect size for all causal SNPs are assumed to come from a standard normal distribution, $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Next, we create a separate matrix $\mathbf{W}$ which holds the pairwise interactions between the causal SNPs in enriched SNP-sets. This is done by taking the Hadamard (element-wise) product between genotypic vectors of SNPs within $\mathcal{C}$. The corresponding interaction effect sizes are drawn as $\boldsymbol{\varphi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. We scale both the additive and pairwise genetic effects so that collectively they explain a fixed proportion of genetic variance. Namely, the additive effects make up $\rho\%$ while the pairwise interactions make up the remaining $(1-\rho)\%$. Alternatively, the proportion of the heritability explained by additivity is said to be $\mathbb{V}[\sum \mathbf{x}_c\theta_c] = \rho H^2$, while the proportion detailed by genetic interactions is given as $\mathbb{V}[\mathbf{W}\boldsymbol{\varphi}] = (1-\rho)H^2$. We consider two choices for the parameter $\rho = \{0.5, 1\}$. Intuitively, $\rho = 1$ represents the limiting case where the variation of a trait is driven by solely additive effects. For $\rho = 0.5$, the additive and pairwise interaction effects are assumed to equally contribute to the phenotypic variance. Once we obtain the final effect sizes for all causal variants, we draw normally distributed random errors as $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to make up the remaining percentage of the total variance. Quantitative continuous traits are then generated under the following two general linear models:

(i) Standard Model: $\mathbf{y} = \sum_{c\in\mathcal{C}} \mathbf{x}_c\theta_c + \mathbf{W}\boldsymbol{\varphi} + \boldsymbol{\varepsilon}$;

(ii) Population Stratification Model: $\mathbf{y} = \mathbf{Z}\boldsymbol{\mu} + \sum_{c\in\mathcal{C}} \mathbf{x}_c\theta_c + \mathbf{W}\boldsymbol{\varphi} + \boldsymbol{\varepsilon}$;

where $\mathbf{Z}$ contains the top 10 genotype principal components (PCs) representing additional population structure, and $\boldsymbol{\mu}$ are the corresponding fixed effects which are also assumed to follow a standard multivariate normal distribution. Alternatively, one can think of the term $\mathbf{Z}\boldsymbol{\mu}$ as an additional genetic random effect [15]. To this end, simulations under model (ii) assume that the genotypic PCs explain an additional 10% of the overall phenotypic variation explained (PVE) within the trait [15]. Therefore, under model (i) the total PVE $= H^2$; while under model (ii) the total PVE $= H^2 + 10\%$. It is important to note that genotype PCs are not included in any of the model fitting procedures, and no other preprocessing normalizations were carried out to account for the added population structure.

Given the simulation procedure above, we randomly sample $N = 10{,}000$ individuals and simulate a wide range of scenarios for comparing the performance of both SNP and SNP-set level association methods. Here, we vary the following simulation parameters:

- Broad-sense heritability: $H^2 = 0.2$ and 0.6;
- Contribution of interaction effects: $(1-\rho) = 0$ and 0.5;
- Percentage of associated SNP-sets: 1% (sparse architecture) and 10% (polygenic architecture);

Lastly, we set the number of causal SNPs with non-zero effects to be some fixed percentage of all SNPs located within the selected associated SNP-sets. We set this percentage to be 0.125% in the 1% associated SNP-set case, and 3% in the 10% associated SNP-set case. All performance comparisons are based on 100 different simulated runs for each parameter combination. For evaluating the performance of each method, we assessed the following:

- The power and false discovery rates when identifying causal SNPs or associated SNP-sets at a Bonferroni-corrected threshold for frequentist approaches ($P = 0.05/36518 = 1.37 \times 10^{-6}$ at the SNP-level and $P = 0.05/2816 = 1.78 \times 10^{-5}$ at the SNP-set level) or according to the median probability model for Bayesian methods (posterior enrichment probability > 0.5) [52];
- The ability to rank true positive (TP) genes over false positives (FP) via receiver operating characteristic (ROC) and precision-recall curves.

All figures and tables show the mean performances (and standard errors) across all simulated replicates.

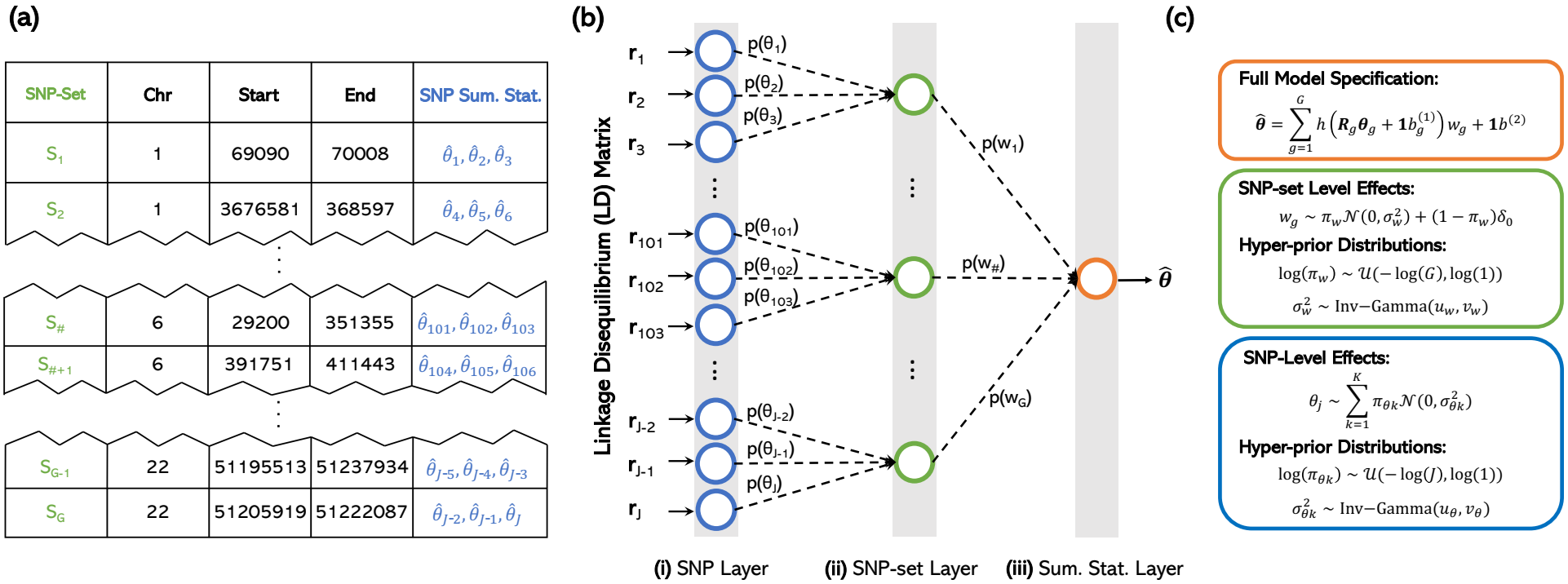
9 Software Details

Source code for the BANNs framework is freely available at `https://github.com/lcrawlab/BANNs` and is licensed under the GNU General Public License (version 3.0). We have released two versions of the BANNs software: one implemented within `Python 3` (release version 3.7.7) and other within `R` (compatible with versions 3.3.2 through 3.6.3). The BANNs GitHub repository includes example data, documentation, and instructions for how to execute the code within both coding languages. Results in the main text and Supplementary Notes are based on the `Python 3` implementation which depends on the `pandas` library (version 1.0.1) [53] for automatically creating partial neural network architectures based on the biological annotations provided by the user; the `NumPy` (version 1.18-19) [54] and `Numba` (version 0.48.0) [55] packages for efficient matrix operations; and the `multiprocessing` library (version 2.6) [56] for parallelizing posterior computation over multiple threads and providing faster execution. Training, estimation of the network parameters, and optimization was done by using an Adam optimizer [57] in `TensorFlow` (version 1.5). While the software can be run directly using the source code, it can also be installed as a package through `pip` with the command: `pip3 install BANNs`. All dependencies are also automatically installed with the package.

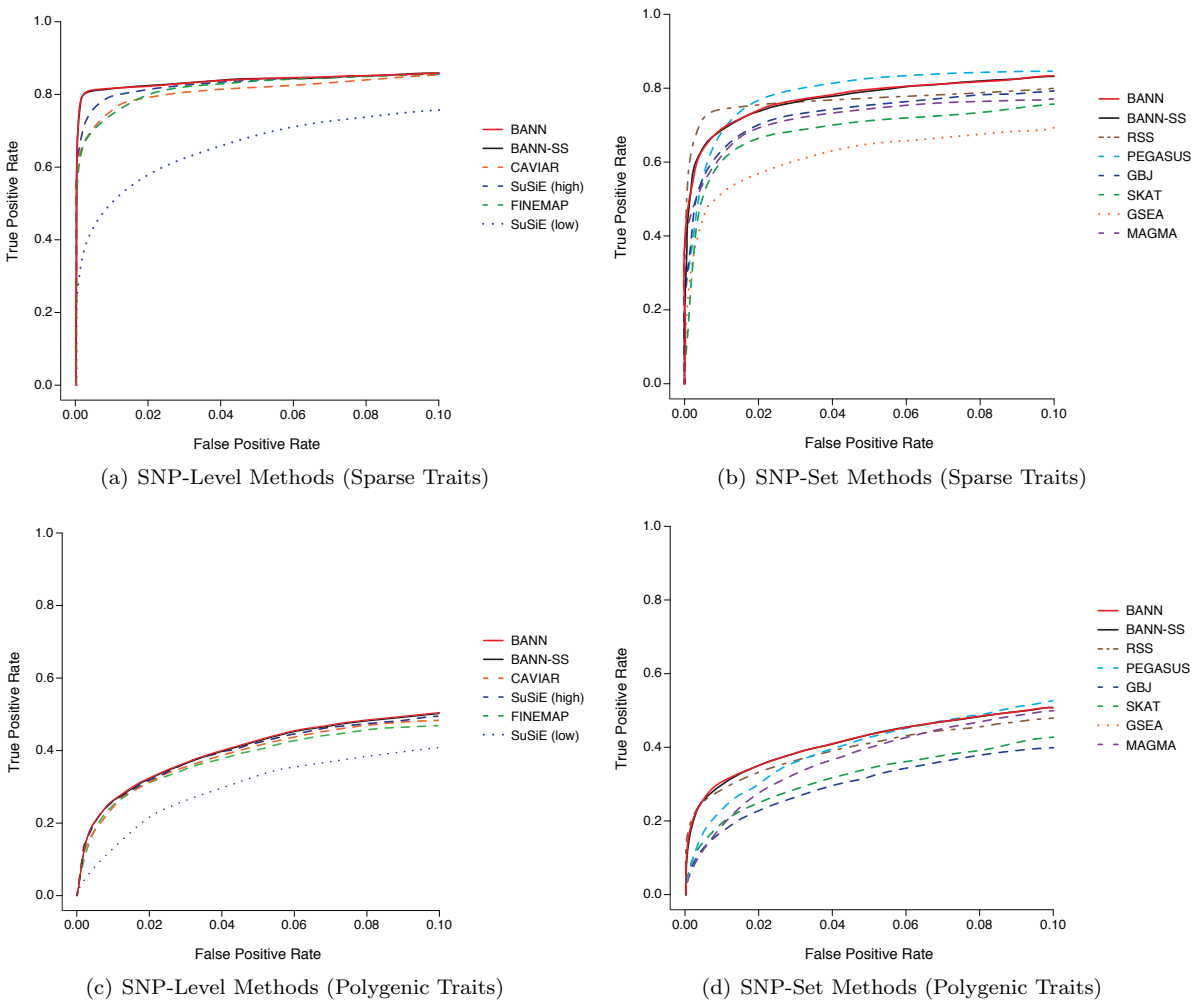
The `R` implementation uses the `dplyr` package (version 0.8.5) [58] for automatically creating partial neural network architectures based on the biological annotations provided by the user; the `Matrix` package (version 1.2-18) [59] for efficient matrix operations; and the `doParallel` (version 1.0.15) [60], `forEach` (version 1.4.8) [61], `iterators` (version 1.0.12) [62], and standard `parallel` packages for parallelized execution of the variational expectation-maximization algorithm. Similarly, the `R` implementation of the software can be run by directly downloading the source code or it can be installed using `devtools` [63] with the commands: `devtools::install("lcrawlab/BANNs")` and `library(BANNs)`.

Software Details for Competing Approaches. In this work, comparisons to SNP-level association mapping methods were made using software for CAVIAR (version 2.0.0; `http://genetics.cs.ucla.edu/caviar/`), FINEMAP (version 1.4; `http://www.christianbenner.com`), and SuSiE (version 0.9.0; `https://github.com/stephenslab/susieR`). Comparisons to SNP-set mapping methods were made using software for GBJ (version 0.5.3; `https://cran.r-project.org/web/packages/GBJ/`), GSEA (`https://www.nr.no/en/projects/software-genomics`), MAGMA (version 1.07b; `https://ctg.cncr.nl/software/magma`), PEGASUS (version 1.3.0; `https://github.com/ramachandran-lab/PEGASUS`), RSS (version 1.0.0; `https://github.com/stephenslab/rss`), and SKAT (version 1.3.2.1; `https://www.hsph.harvard.edu/skat`), which are also publicly available. All software for competing methods were fit using the default settings, unless otherwise stated in the main text and Supplementary Notes.

10 Supplementary Figures



Supplementary Figure 1. Biologically annotated neural networks also take in GWA summary statistics (BANN-SS) for multi-scale genotype-phenotype by specifying a partially connected architecture based on the hierarchical nature of enrichment studies. **(a)** The BANN-SS framework requires a J-dimensional vector of SNP-level GWA marginal effect size (OLS) estimates $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_J)$; an empirical $J \times J$ linkage disequilibrium (LD) matrix $\mathbf{R} = [\mathbf{r}_1, \ldots, \mathbf{r}_J]$, where $\mathbf{r}_j = [r(\mathbf{x}_j, \mathbf{x}_1), \ldots, r(\mathbf{x}_j, \mathbf{x}_J)]$ is a vector of correlation coefficients between the j-th SNP and all other SNPs in the study; and a list of G-predefined SNP-sets $\{\mathcal{S}_1, \ldots, \mathcal{S}_G\}$. In this work, SNP-sets are defined as genes and intergenic regions (between genes) given by the NCBI's Reference Sequence (RefSeq) database in the UCSC Genome Browser [50]. **(b)** A partially connected Bayesian neural network is constructed based on the annotated SNP groups. In the first hidden layer, only SNPs within the boundary of a gene are connected to the same node. Similarly, SNPs within the same intergenic region between genes are connected to the same node. Completing this specification for all SNPs gives the hidden layer the natural interpretation of being the "SNP-set" layer. **(c)** The hierarchical nature of the network is represented as nonlinear mixed model. The corresponding weights in both the SNP ($\boldsymbol{\theta}$) and SNP-set ($\mathbf{w}$) layers are treated as random variables with biologically motivated sparse prior distributions. Posterior inclusion probabilities (PIPs) $\boldsymbol{\gamma}_\theta$ and $\boldsymbol{\gamma}_w$ summarize associations at the SNP and SNP-set level, respectively. The BANN-SS framework uses the same variational inference procedure that is used when we have access to individual-level data.



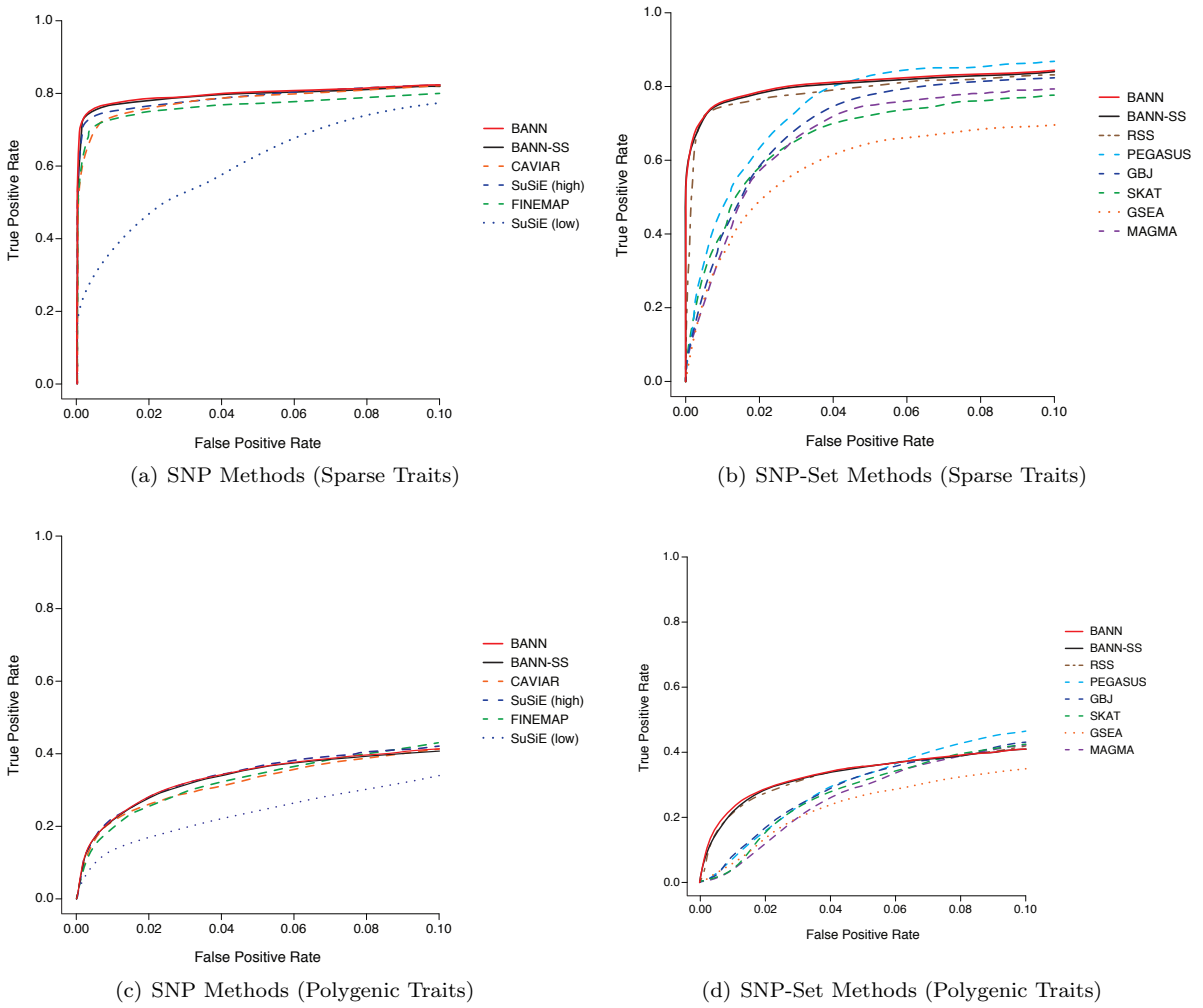
(a) SNP-Level Methods (Sparse Traits)

(b) SNP-Set Methods (Sparse Traits)

(c) SNP-Level Methods (Polygenic Traits)

(d) SNP-Set Methods (Polygenic Traits)

Supplementary Figure 2. Receiver operating characteristic (ROC) curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations. Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with only contributions from additive effects (i.e., $\rho = 1$). We show power versus false positive rate for two different trait architectures: **(a, b)** sparse where only 1% of SNP-sets are enriched for the trait; and **(c, d)** polygenic where 10% of SNP-sets are enriched. We then set the number of causal SNPs with non-zero effects to be 0.125% and 3% of all SNPs located within the selected enriched SNP-sets, respectively. To derive results, the full genotype matrix and phenotypic vector are given to the BANNs model and all competing methods that require individual-level data. For the BANN-SS model and other competing methods that take GWA summary statistics, we compute standard GWA SNP-level effect sizes and P-values (estimated using ordinary least squares). **(a, c)** Competing SNP-level mapping approaches include: CAVIAR [64], SuSiE [65], and FINEMAP [66]. The software for SuSiE requires an input ℓ which fixes the maximum number of causal SNPs in the model. We display results when this input number is high ($\ell = 3000$) and when this input number is low ($\ell = 10$). **(b, d)** Competing SNP-set mapping approaches include: RSS [7], PEGASUS [67], GBJ [68], SKAT [69], GSEA [70], and MAGMA [71]. Note that the upper limit of the x-axis has been truncated at 0.1. All results are based on 100 replicates (see Supplementary Note, Section 8).



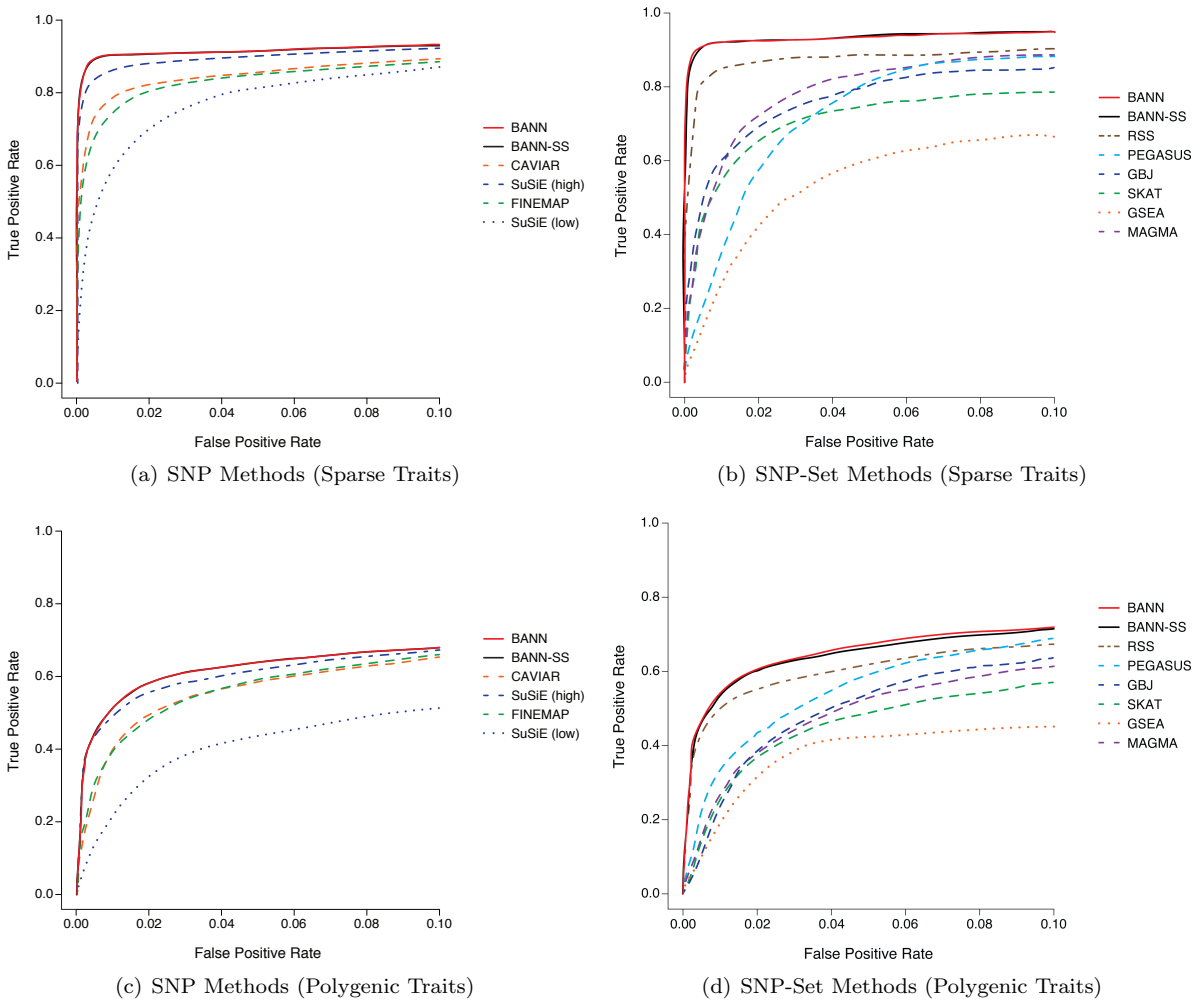
(a) SNP Methods (Sparse Traits)

(b) SNP-Set Methods (Sparse Traits)

(c) SNP Methods (Polygenic Traits)

(d) SNP-Set Methods (Polygenic Traits)

Supplementary Figure 3. Receiver operating characteristic (ROC) curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations with population stratification. Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with only contributions from additive effects (i.e., $\rho = 1$). In these simulations, traits were generated while using the top ten principal components (PCs) of the genotype matrix as covariates. We show power versus false positive rate for two different trait architectures: **(a, b)** sparse where only 1% of SNP-sets are enriched for the trait; and **(c, d)** polygenic where 10% of SNP-sets are enriched. We then set the number of causal SNPs with non-zero effects to be 0.125% and 3% of all SNPs located within the selected enriched SNP-sets, respectively. To derive results, the full genotype matrix and phenotypic vector are given to the BANNs model and all competing methods that require individual-level data. For the BANN-SS model and other competing methods that take GWA summary statistics, we compute standard GWA SNP-level effect sizes and P-values (estimated using ordinary least squares). **(a, c)** Competing SNP-level mapping approaches include: CAVIAR [64], SuSiE [65], and FINEMAP [66]. The software for SuSiE requires an input ℓ which fixes the maximum number of causal SNPs in the model. We display results when this input number is high ($\ell = 3000$) and when this input number is low ($\ell = 10$). **(b, d)** Competing SNP-set mapping approaches include: RSS [7], PEGASUS [67], GBJ [68], SKAT [69], GSEA [70], and MAGMA [71]. Note that the upper limit of the x-axis has been truncated at 0.1. All results are based on 100 replicates (see Supplementary Note, Section 8).



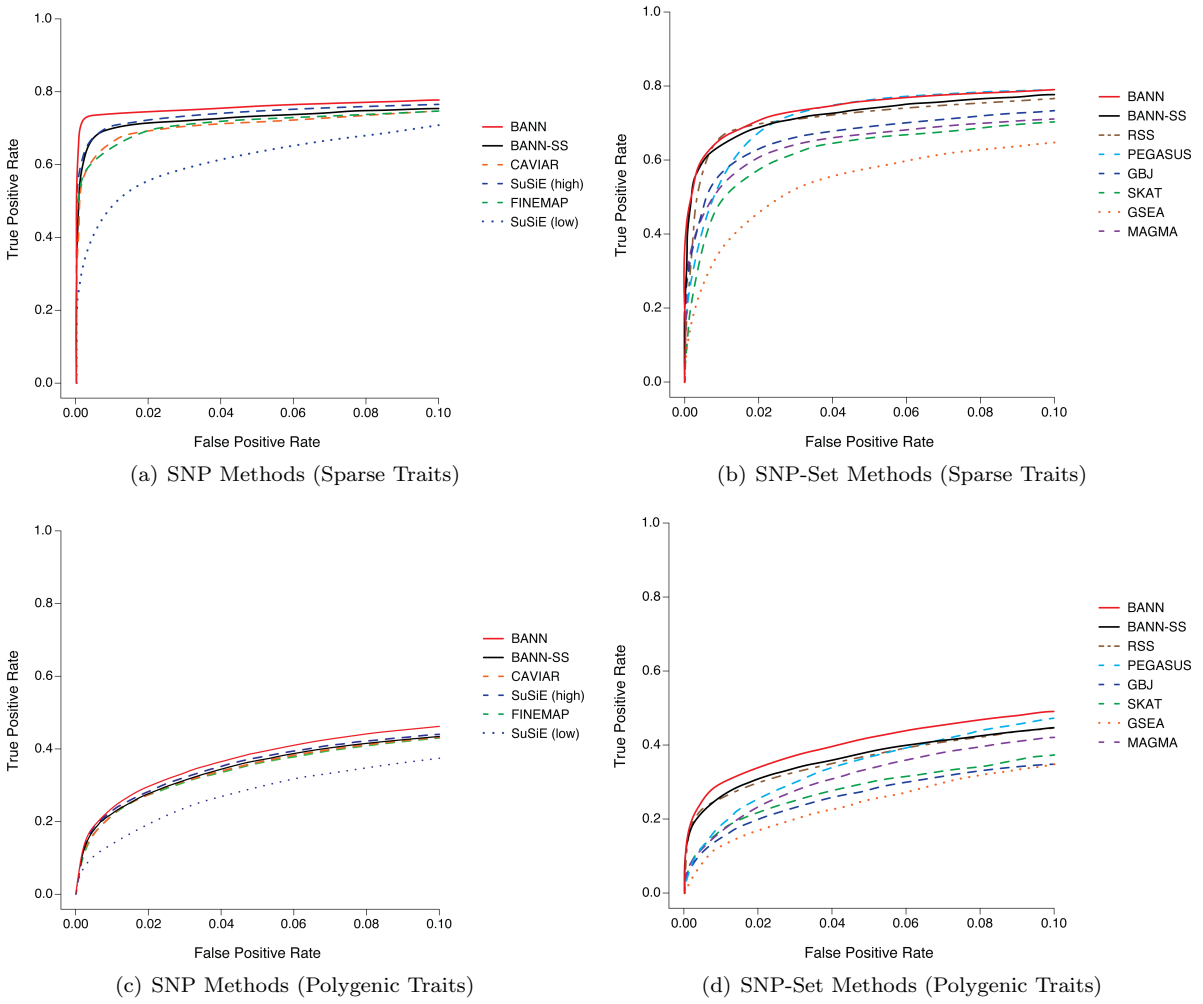
(a) SNP Methods (Sparse Traits)

(b) SNP-Set Methods (Sparse Traits)

(c) SNP Methods (Polygenic Traits)

(d) SNP-Set Methods (Polygenic Traits)

Supplementary Figure 4. Receiver operating characteristic (ROC) curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations with population stratification. Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.6$ with only contributions from additive effects (i.e., $\rho = 1$). In these simulations, traits were generated while using the top ten principal components (PCs) of the genotype matrix as covariates. We show power versus false positive rate for two different trait architectures: **(a, b)** sparse where only 1% of SNP-sets are enriched for the trait; and **(c, d)** polygenic where 10% of SNP-sets are enriched. We then set the number of causal SNPs with non-zero effects to be 0.125% and 3% of all SNPs located within the selected enriched SNP-sets, respectively. To derive results, the full genotype matrix and phenotypic vector are given to the BANNs model and all competing methods that require individual-level data. For the BANN-SS model and other competing methods that take GWA summary statistics, we compute standard GWA SNP-level effect sizes and P-values (estimated using ordinary least squares). **(a, c)** Competing SNP-level mapping approaches include: CAVIAR [64], SuSiE [65], and FINEMAP [66]. The software for SuSiE requires an input ℓ which fixes the maximum number of causal SNPs in the model. We display results when this input number is high ($\ell = 3000$) and when this input number is low ($\ell = 10$). **(b, d)** Competing SNP-set mapping approaches include: RSS [7], PEGASUS [67], GBJ [68], SKAT [69], GSEA [70], and MAGMA [71]. Note that the upper limit of the x-axis has been truncated at 0.1. All results are based on 100 replicates (see Supplementary Note, Section 8).



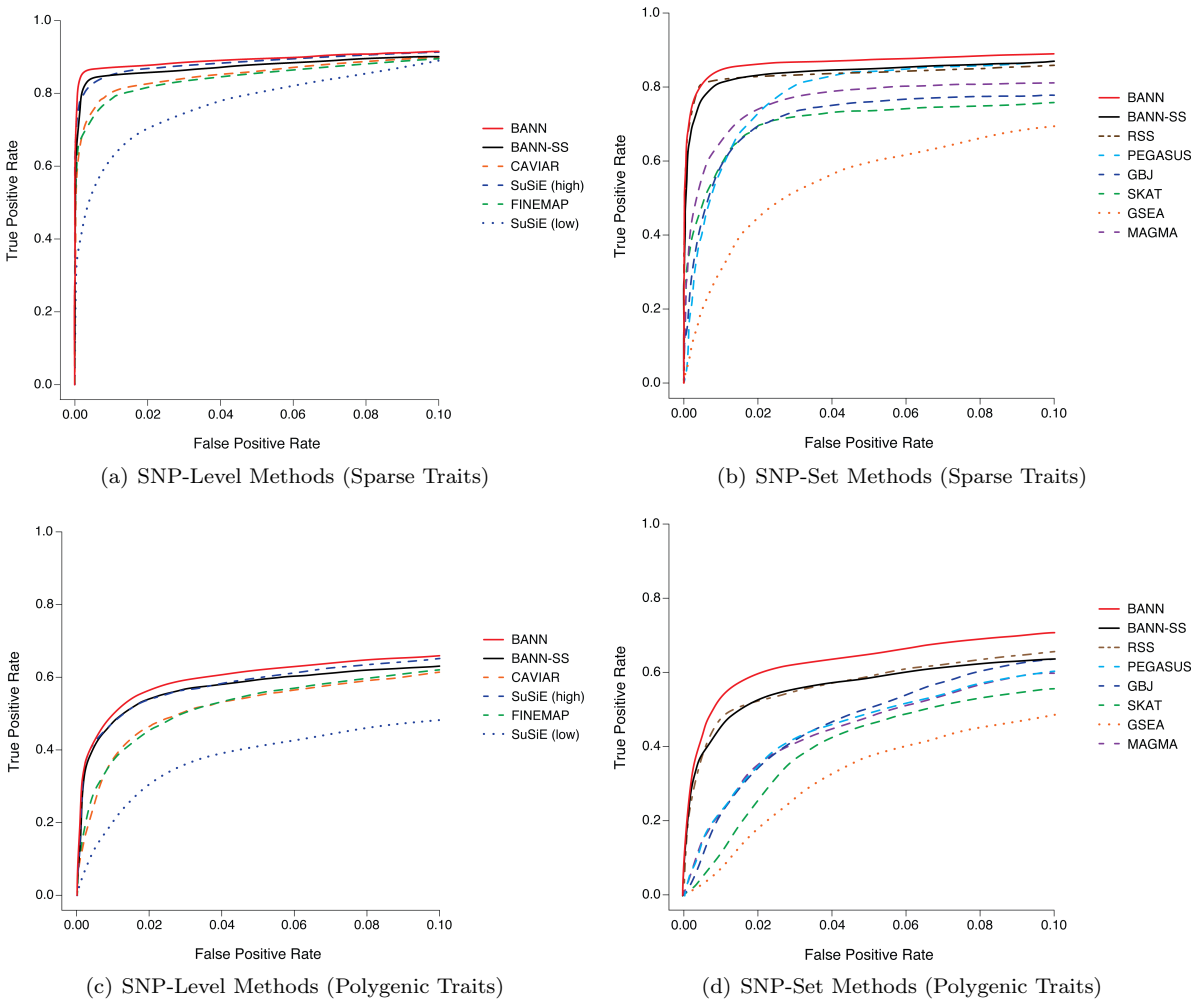
(a) SNP Methods (Sparse Traits)

(b) SNP-Set Methods (Sparse Traits)

(c) SNP Methods (Polygenic Traits)

(d) SNP-Set Methods (Polygenic Traits)

Supplementary Figure 5. Receiver operating characteristic (ROC) curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations. Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with equal contributions from additive effects and epistatic interactions (i.e., $\rho = 0.5$). We show power versus false positive rate for two different trait architectures: **(a, b)** sparse where only 1% of SNP-sets are enriched for the trait; and **(c, d)** polygenic where 10% of SNP-sets are enriched. We then set the number of causal SNPs with non-zero effects to be 0.125% and 3% of all SNPs located within the enriched SNP-sets, respectively. To derive results, the full genotype matrix and phenotypic vector are given to the BANNs model and all competing methods that require individual-level data. For the BANN-SS model and other competing methods that take GWA summary statistics, we compute standard GWA SNP-level effect sizes and P-values (estimated using ordinary least squares). **(a, c)** Competing SNP-level mapping approaches include: CAVIAR [64], SuSiE [65], and FINEMAP [66]. The software for SuSiE requires an input ℓ which fixes the maximum number of causal SNPs in the model. We display results when this input number is high ($\ell = 3000$) and when this input number is low ($\ell = 10$). **(b, d)** Competing SNP-set mapping approaches include: RSS [7], PEGASUS [67], GBJ [68], SKAT [69], GSEA [70], and MAGMA [71]. Note that the upper limit of the x-axis has been truncated at 0.1. All results are based on 100 replicates (see Supplementary Note, Section 8).



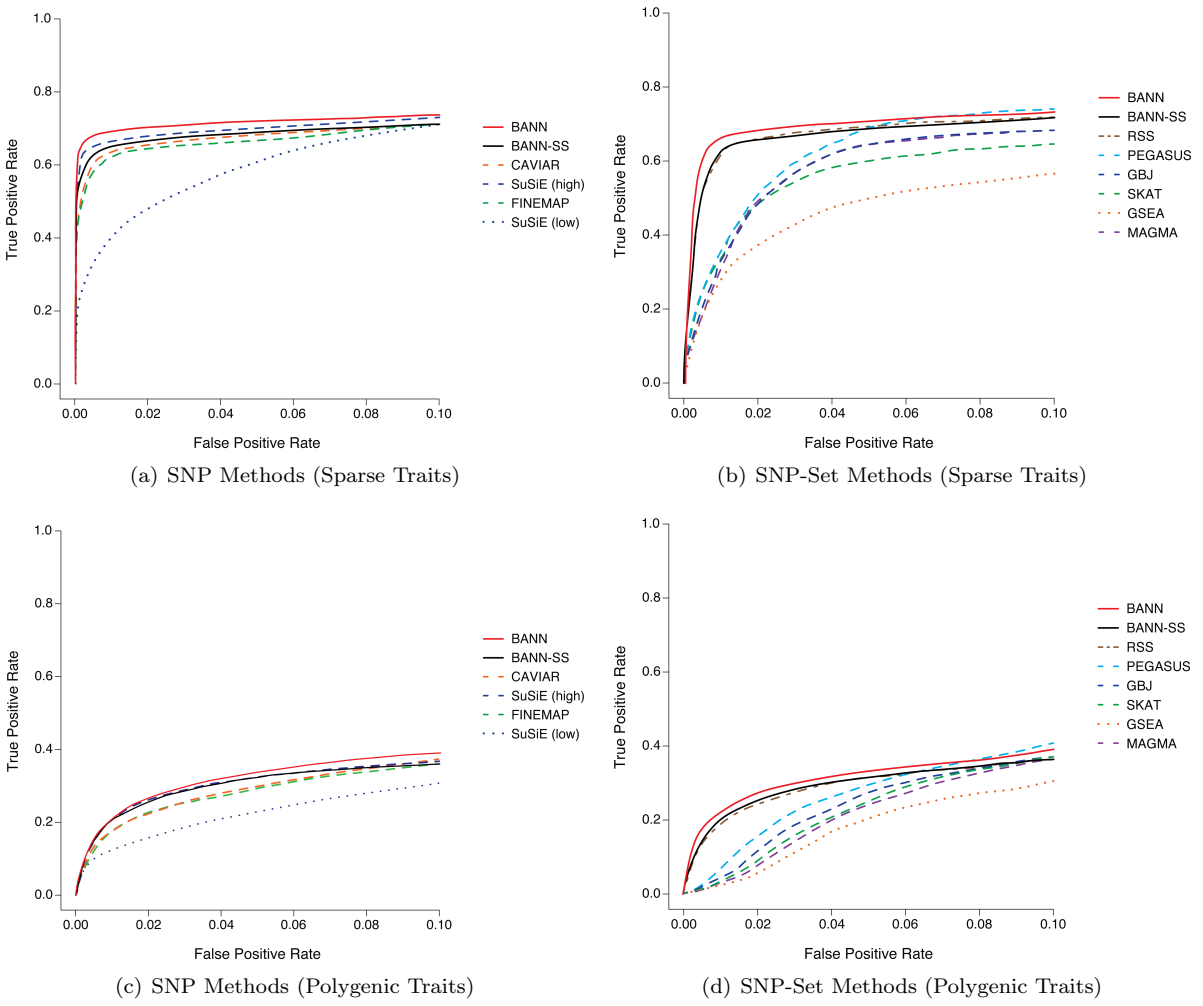
(a) SNP-Level Methods (Sparse Traits)

(b) SNP-Set Methods (Sparse Traits)

(c) SNP-Level Methods (Polygenic Traits)

(d) SNP-Set Methods (Polygenic Traits)

Supplementary Figure 6. Receiver operating characteristic (ROC) curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations. Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.6$ with equal contributions from additive effects and epistatic interactions (i.e., $\rho = 0.5$). We show power versus false positive rate for two different trait architectures: **(a, b)** sparse where only 1% of SNP-sets are enriched for the trait; and **(c, d)** polygenic where 10% of SNP-sets are enriched. We then set the number of causal SNPs with non-zero effects to be 0.125% and 3% of all SNPs located within the enriched SNP-sets, respectively. To derive results, the full genotype matrix and phenotypic vector are given to the BANNs model and all competing methods that require individual-level data. For the BANN-SS model and other competing methods that take GWA summary statistics, we compute standard GWA SNP-level effect sizes and P-values (estimated using ordinary least squares). **(a, c)** Competing SNP-level mapping approaches include: CAVIAR [64], SuSiE [65], and FINEMAP [66]. The software for SuSiE requires an input ℓ which fixes the maximum number of causal SNPs in the model. We display results when this input number is high ($\ell = 3000$) and when this input number is low ($\ell = 10$). **(b, d)** Competing SNP-set mapping approaches include: RSS [7], PEGASUS [67], GBJ [68], SKAT [69], GSEA [70], and MAGMA [71]. Note that the upper limit of the x-axis has been truncated at 0.1. All results are based on 100 replicates (see Supplementary Note, Section 8).



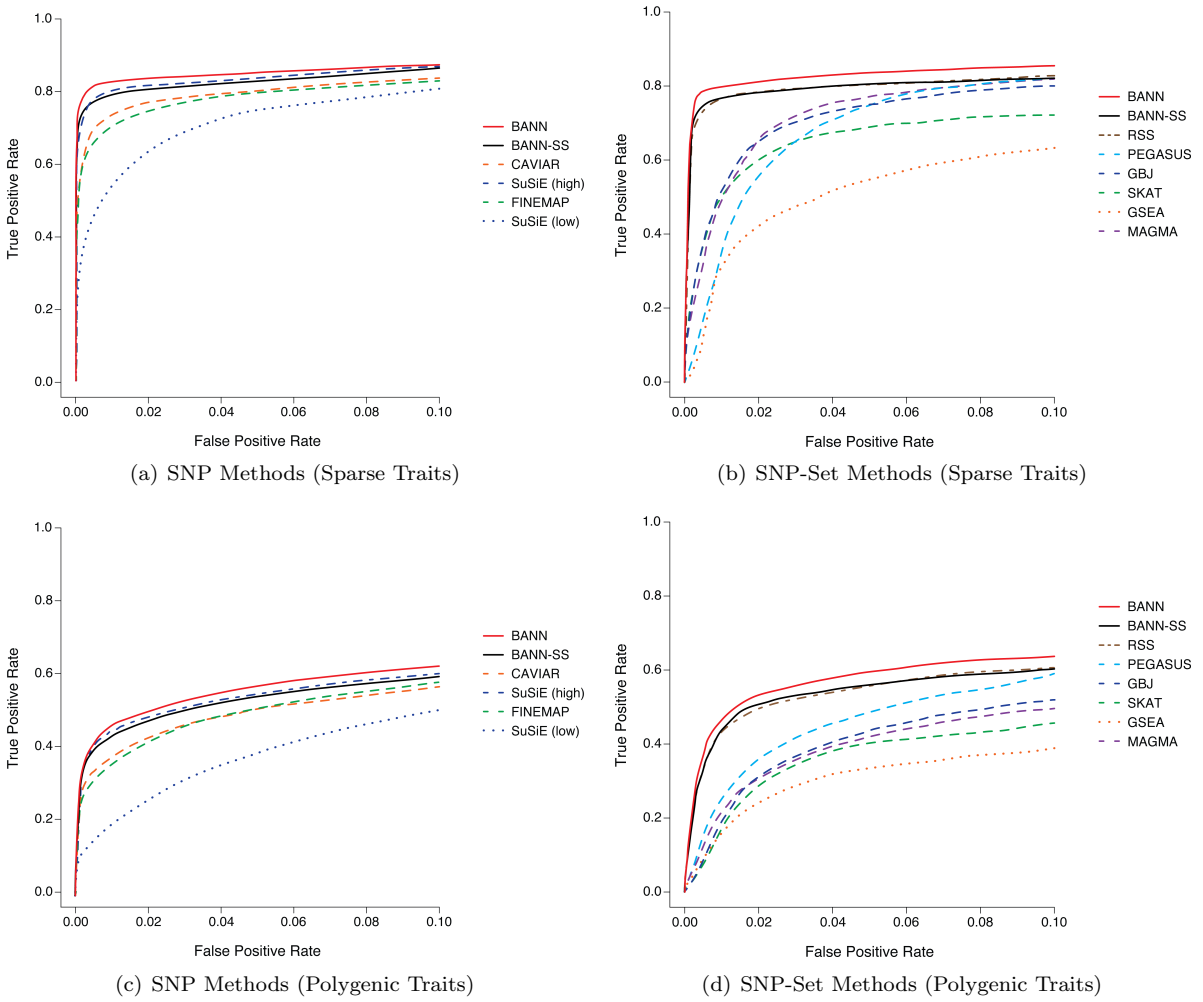
(a) SNP Methods (Sparse Traits)

(b) SNP-Set Methods (Sparse Traits)

(c) SNP Methods (Polygenic Traits)

(d) SNP-Set Methods (Polygenic Traits)

Supplementary Figure 7. Receiver operating characteristic (ROC) curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations with population stratification. Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with equal contributions from additive effects and epistatic interactions (i.e., $\rho = 0.5$). In these simulations, traits were generated while using the top ten principal components (PCs) of the genotype matrix as covariates. We show power versus false positive rate for two different trait architectures: **(a, b)** sparse where only 1% of SNP-sets are enriched for the trait; and **(c, d)** polygenic where 10% of SNP-sets are enriched. We then set the number of causal SNPs with non-zero effects to be 0.125% and 3% of all SNPs located within the enriched SNP-sets, respectively. To derive results, the full genotype matrix and phenotypic vector are given to the BANNs model and all competing methods that require individual-level data. For the BANN-SS model and other competing methods that take GWA summary statistics, we compute standard GWA SNP-level effect sizes and P-values (estimated using ordinary least squares). **(a, c)** Competing SNP-level mapping approaches include: CAVIAR [64], SuSiE [65], and FINEMAP [66]. The software for SuSiE requires an input ℓ which fixes the maximum number of causal SNPs in the model. We display results when this input number is high ($\ell = 3000$) and when this input number is low ($\ell = 10$). **(b, d)** Competing SNP-set mapping approaches include: RSS [7], PEGASUS [67], GBJ [68], SKAT [69], GSEA [70], and MAGMA [71]. Note that the upper limit of the x-axis has been truncated at 0.1. All results are based on 100 replicates (see Supplementary Note, Section 8).



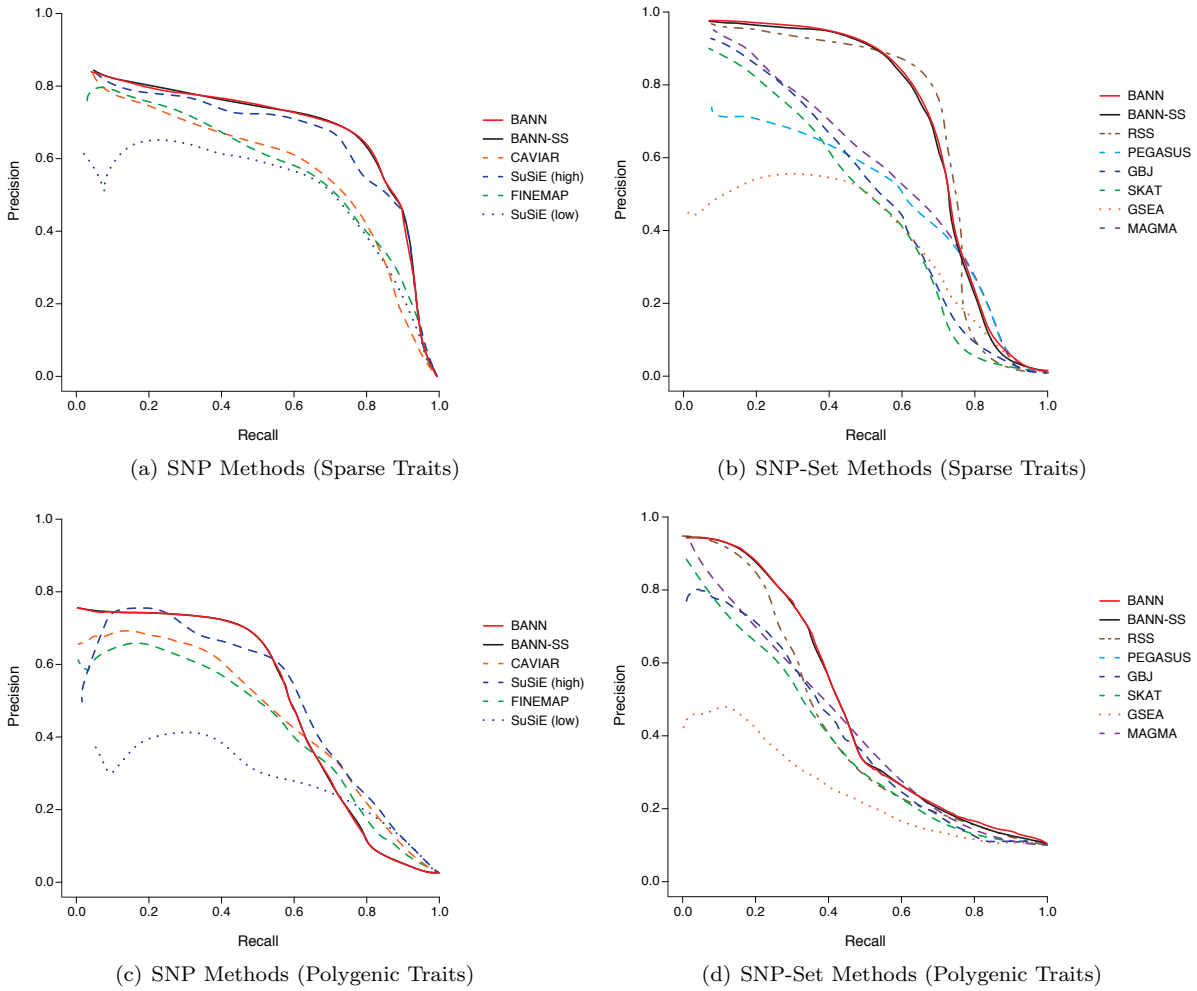
(a) SNP Methods (Sparse Traits)

(b) SNP-Set Methods (Sparse Traits)

(c) SNP Methods (Polygenic Traits)

(d) SNP-Set Methods (Polygenic Traits)

Supplementary Figure 8. Receiver operating characteristic (ROC) curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations with population stratification. Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.6$ with equal contributions from additive effects and epistatic interactions (i.e., $\rho = 0.5$). In these simulations, traits were generated while using the top ten principal components (PCs) of the genotype matrix as covariates. We show power versus false positive rate for two different trait architectures: **(a, b)** sparse where only 1% of SNP-sets are enriched for the trait; and **(c, d)** polygenic where 10% of SNP-sets are enriched. We then set the number of causal SNPs with non-zero effects to be 0.125% and 3% of all SNPs located within the enriched SNP-sets, respectively. To derive results, the full genotype matrix and phenotypic vector are given to the BANNs model and all competing methods that require individual-level data. For the BANN-SS model and other competing methods that take GWA summary statistics, we compute standard GWA SNP-level effect sizes and P-values (estimated using ordinary least squares). **(a, c)** Competing SNP-level mapping approaches include: CAVIAR [64], SuSiE [65], and FINEMAP [66]. The software for SuSiE requires an input ℓ which fixes the maximum number of causal SNPs in the model. We display results when this input number is high ($\ell = 3000$) and when this input number is low ($\ell = 10$). **(b, d)** Competing SNP-set mapping approaches include: RSS [7], PEGASUS [67], GBJ [68], SKAT [69], GSEA [70], and MAGMA [71]. Note that the upper limit of the x-axis has been truncated at 0.1. All results are based on 100 replicates (see Supplementary Note, Section 8).



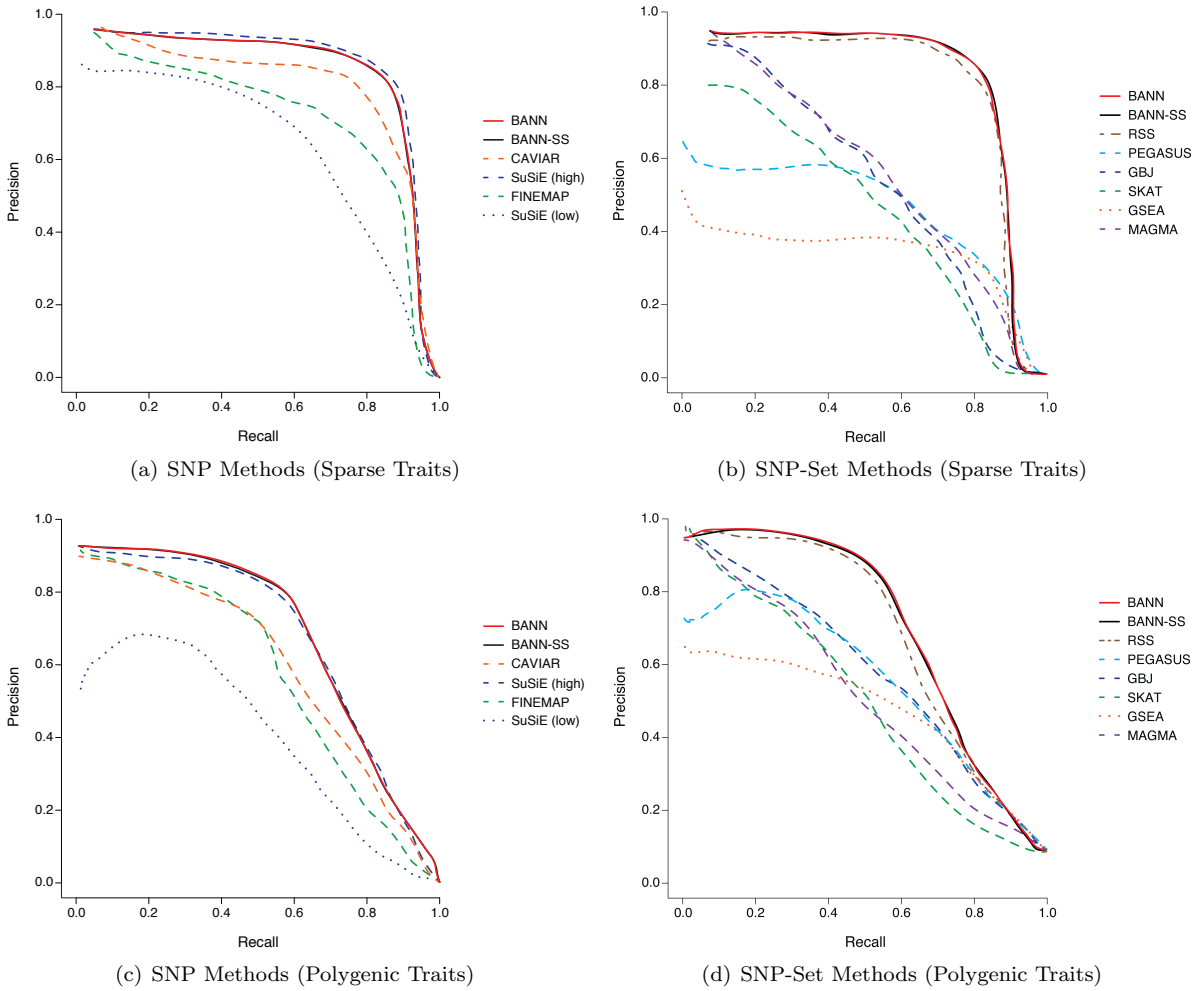
(a) SNP Methods (Sparse Traits)

(b) SNP-Set Methods (Sparse Traits)

(c) SNP Methods (Polygenic Traits)

(d) SNP-Set Methods (Polygenic Traits)

Supplementary Figure 9. Precision-recall curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations. Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with only contributions from additive effects (i.e., $\rho = 1$). We show precision versus recall for two different trait architectures: **(a, b)** sparse where only 1% of SNP-sets are enriched for the trait; and **(c, d)** polygenic where 10% of SNP-sets are enriched. We then set the number of causal SNPs with non-zero effects to be 0.125% and 3% of all SNPs located within the selected enriched SNP-sets, respectively. To derive results, the full genotype matrix and phenotypic vector are given to the BANNs model and all competing methods that require individual-level data. For the BANN-SS model and other competing methods that take GWA summary statistics, we compute standard GWA SNP-level effect sizes and P-values (estimated using ordinary least squares). **(a, c)** Competing SNP-level mapping approaches include: CAVIAR [64], SuSiE [65], and FINEMAP [66]. The software for SuSiE requires an input ℓ which fixes the maximum number of causal SNPs in the model. We display results when this input number is high ($\ell = 3000$) and when this input number is low ($\ell = 10$). **(b, d)** Competing SNP-set mapping approaches include: RSS [7], PEGASUS [67], GBJ [68], SKAT [69], GSEA [70], and MAGMA [71]. Note that, for traits with sparse architectures, the top ranked SNPs and SNP-sets are always true positives, and therefore the minimal recall is not 0. All results are based on 100 replicates (see Supplementary Note, Section 8).



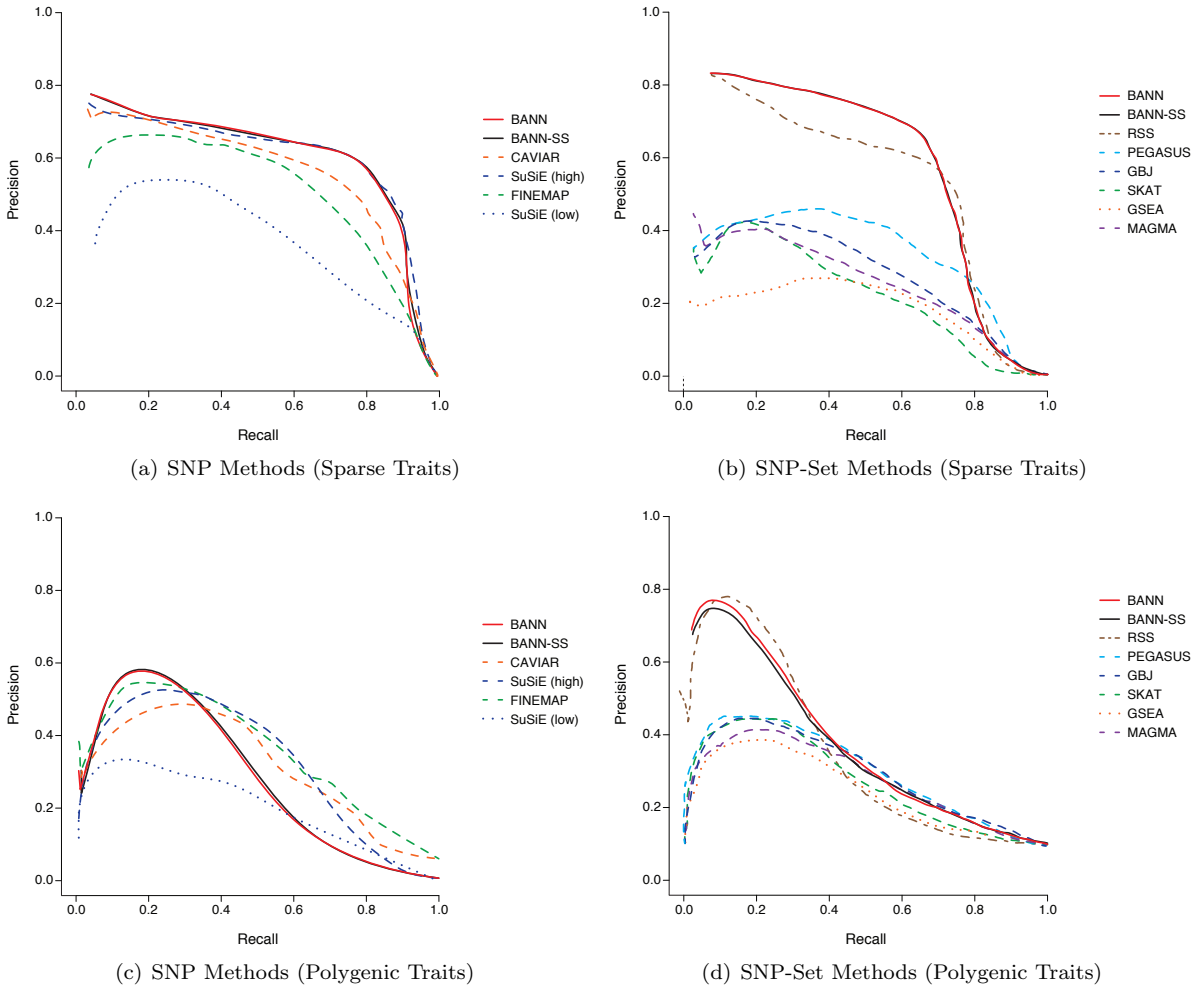
(a) SNP Methods (Sparse Traits)

(b) SNP-Set Methods (Sparse Traits)

(c) SNP Methods (Polygenic Traits)

(d) SNP-Set Methods (Polygenic Traits)

Supplementary Figure 10. Precision-recall curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations. Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.6$ with only contributions from additive effects (i.e., $\rho = 1$). We show precision versus recall for two different trait architectures: **(a, b)** sparse where only 1% of SNP-sets are enriched for the trait; and **(c, d)** polygenic where 10% of SNP-sets are enriched. We then set the number of causal SNPs with non-zero effects to be 0.125% and 3% of all SNPs located within the selected enriched SNP-sets, respectively. To derive results, the full genotype matrix and phenotypic vector are given to the BANNs model and all competing methods that require individual-level data. For the BANN-SS model and other competing methods that take GWA summary statistics, we compute standard GWA SNP-level effect sizes and P-values (estimated using ordinary least squares). **(a, c)** Competing SNP-level mapping approaches include: CAVIAR [64], SuSiE [65], and FINEMAP [66]. The software for SuSiE requires an input ℓ which fixes the maximum number of causal SNPs in the model. We display results when this input number is high ($\ell = 3000$) and when this input number is low ($\ell = 10$). **(b, d)** Competing SNP-set mapping approaches include: RSS [7], PEGASUS [67], GBJ [68], SKAT [69], GSEA [70], and MAGMA [71]. Note that, for traits with sparse architectures, the top ranked SNPs and SNP-sets are always true positives, and therefore the minimal recall is not 0. All results are based on 100 replicates (see Supplementary Note, Section 8).



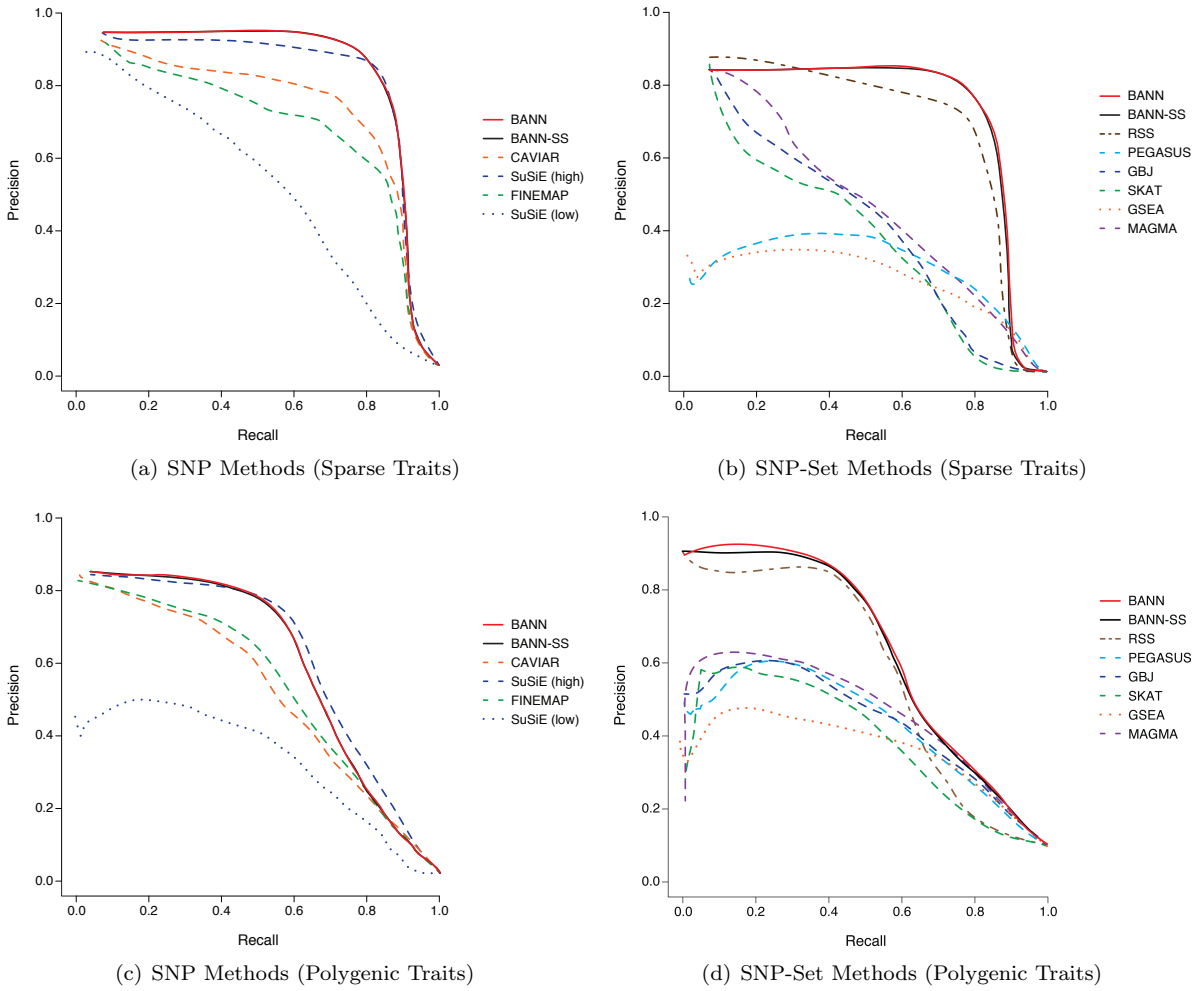
(a) SNP Methods (Sparse Traits)

(b) SNP-Set Methods (Sparse Traits)

(c) SNP Methods (Polygenic Traits)

(d) SNP-Set Methods (Polygenic Traits)

Supplementary Figure 11. Precision-recall curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations with population stratification. Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with only contributions from additive effects (i.e., $\rho = 1$). In these simulations, traits were generated while using the top ten principal components (PCs) of the genotype matrix as covariates. We show precision versus recall for two different trait architectures: **(a, b)** sparse where only 1% of SNP-sets are enriched for the trait; and **(c, d)** polygenic where 10% of SNP-sets are enriched. We then set the number of causal SNPs with non-zero effects to be 0.125% and 3% of all SNPs located within the selected enriched SNP-sets, respectively. To derive results, the full genotype matrix and phenotypic vector are given to the BANNs model and all competing methods that require individual-level data. For the BANN-SS model and other competing methods that take GWA summary statistics, we compute standard GWA SNP-level effect sizes and P-values (estimated using ordinary least squares). **(a, c)** Competing SNP-level mapping approaches include: CAVIAR [64], SuSiE [65], and FINEMAP [66]. The software for SuSiE requires an input ℓ which fixes the maximum number of causal SNPs in the model. We display results when this input number is high ($\ell = 3000$) and when this input number is low ($\ell = 10$). **(b, d)** Competing SNP-set mapping approaches include: RSS [7], PEGASUS [67], GBJ [68], SKAT [69], GSEA [70], and MAGMA [71]. Note that, for traits with sparse architectures, the top ranked SNPs and SNP-sets are always true positives, and therefore the minimal recall is not 0. All results are based on 100 replicates (see Supplementary Note, Section 8).



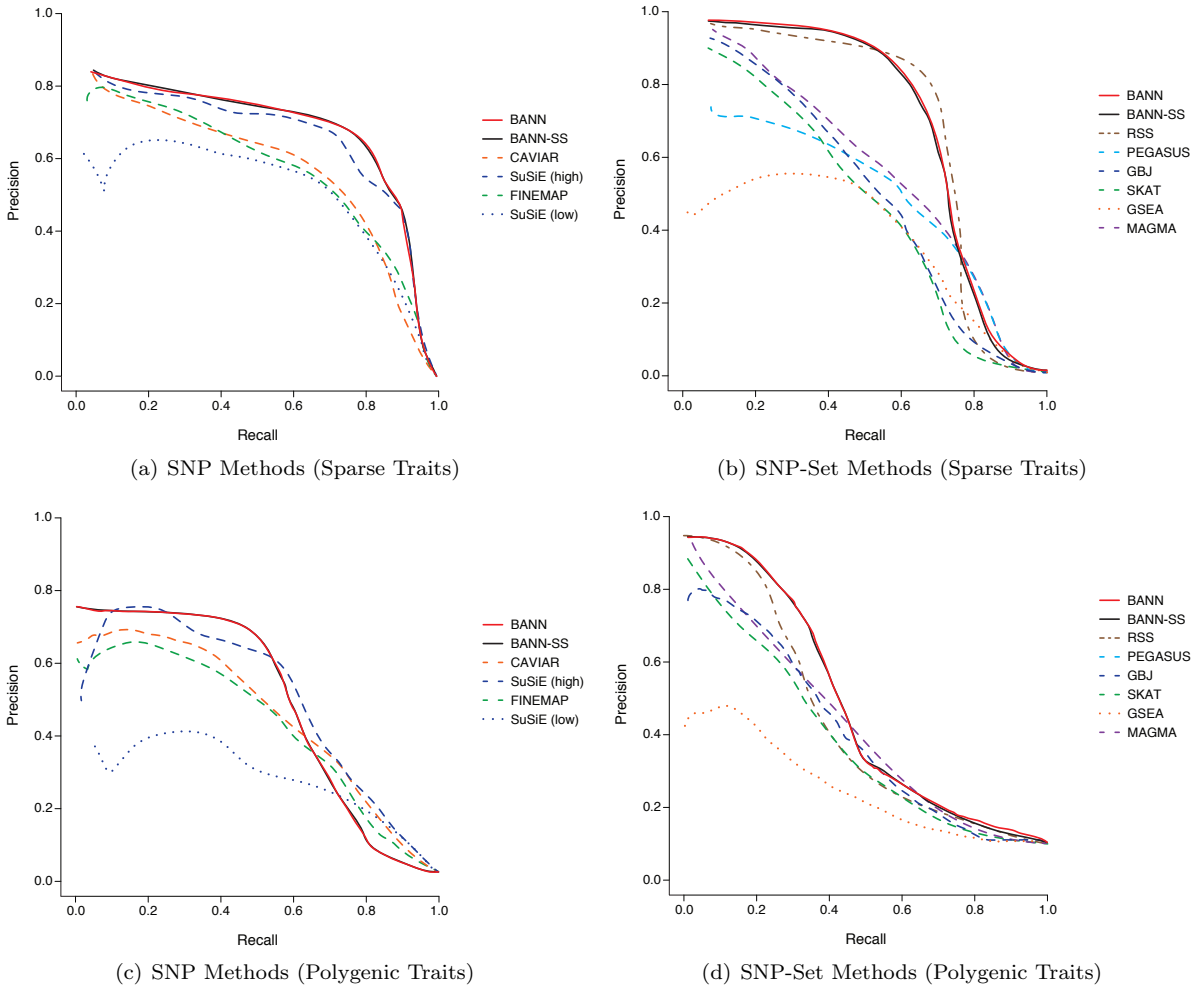
(a) SNP Methods (Sparse Traits)

(b) SNP-Set Methods (Sparse Traits)

(c) SNP Methods (Polygenic Traits)

(d) SNP-Set Methods (Polygenic Traits)

Supplementary Figure 12. Precision-recall curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations with population stratification. Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.6$ with only contributions from additive effects (i.e., $\rho = 1$). In these simulations, traits were generated while using the top ten principal components (PCs) of the genotype matrix as covariates. We show precision versus recall for two different trait architectures: **(a, b)** sparse where only 1% of SNP-sets are enriched for the trait; and **(c, d)** polygenic where 10% of SNP-sets are enriched. We then set the number of causal SNPs with non-zero effects to be 0.125% and 3% of all SNPs located within the selected enriched SNP-sets, respectively. To derive results, the full genotype matrix and phenotypic vector are given to the BANNs model and all competing methods that require individual-level data. For the BANN-SS model and other competing methods that take GWA summary statistics, we compute standard GWA SNP-level effect sizes and P-values (estimated using ordinary least squares). **(a, c)** Competing SNP-level mapping approaches include: CAVIAR [64], SuSiE [65], and FINEMAP [66]. The software for SuSiE requires an input ℓ which fixes the maximum number of causal SNPs in the model. We display results when this input number is high ($\ell = 3000$) and when this input number is low ($\ell = 10$). **(b, d)** Competing SNP-set mapping approaches include: RSS [7], PEGASUS [67], GBJ [68], SKAT [69], GSEA [70], and MAGMA [71]. Note that, for traits with sparse architectures, the top ranked SNPs and SNP-sets are always true positives, and therefore the minimal recall is not 0. All results are based on 100 replicates (see Supplementary Note, Section 8).



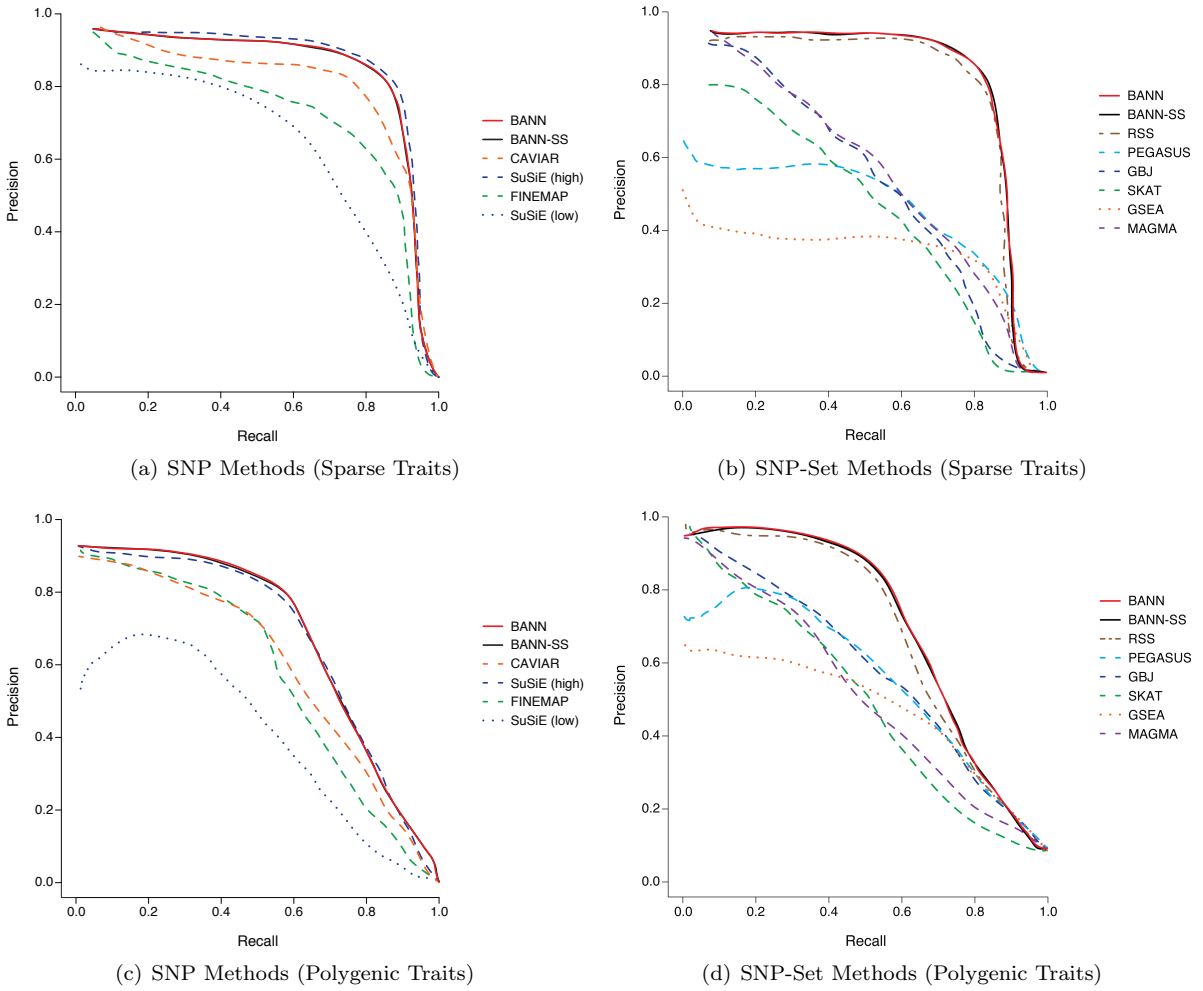
(a) SNP Methods (Sparse Traits)

(b) SNP-Set Methods (Sparse Traits)

(c) SNP Methods (Polygenic Traits)

(d) SNP-Set Methods (Polygenic Traits)

Supplementary Figure 13. Precision-recall curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations. Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with equal contributions from additive effects and epistatic interactions (i.e., $\rho = 0.5$). We show precision versus recall for two different trait architectures: **(a, b)** sparse where only 1% of SNP-sets are enriched for the trait; and **(c, d)** polygenic where 10% of SNP-sets are enriched. We then set the number of causal SNPs with non-zero effects to be 0.125% and 3% of all SNPs located within the selected enriched SNP-sets, respectively. To derive results, the full genotype matrix and phenotypic vector are given to the BANNs model and all competing methods that require individual-level data. For the BANN-SS model and other competing methods that take GWA summary statistics, we compute standard GWA SNP-level effect sizes and P-values (estimated using ordinary least squares). **(a, c)** Competing SNP-level mapping approaches include: CAVIAR [64], SuSiE [65], and FINEMAP [66]. The software for SuSiE requires an input ℓ which fixes the maximum number of causal SNPs in the model. We display results when this input number is high ($\ell = 3000$) and when this input number is low ($\ell = 10$). **(b, d)** Competing SNP-set mapping approaches include: RSS [7], PEGASUS [67], GBJ [68], SKAT [69], GSEA [70], and MAGMA [71]. Note that, for traits with sparse architectures, the top ranked SNPs and SNP-sets are always true positives, and therefore the minimal recall is not 0. All results are based on 100 replicates (see Supplementary Note, Section 8).



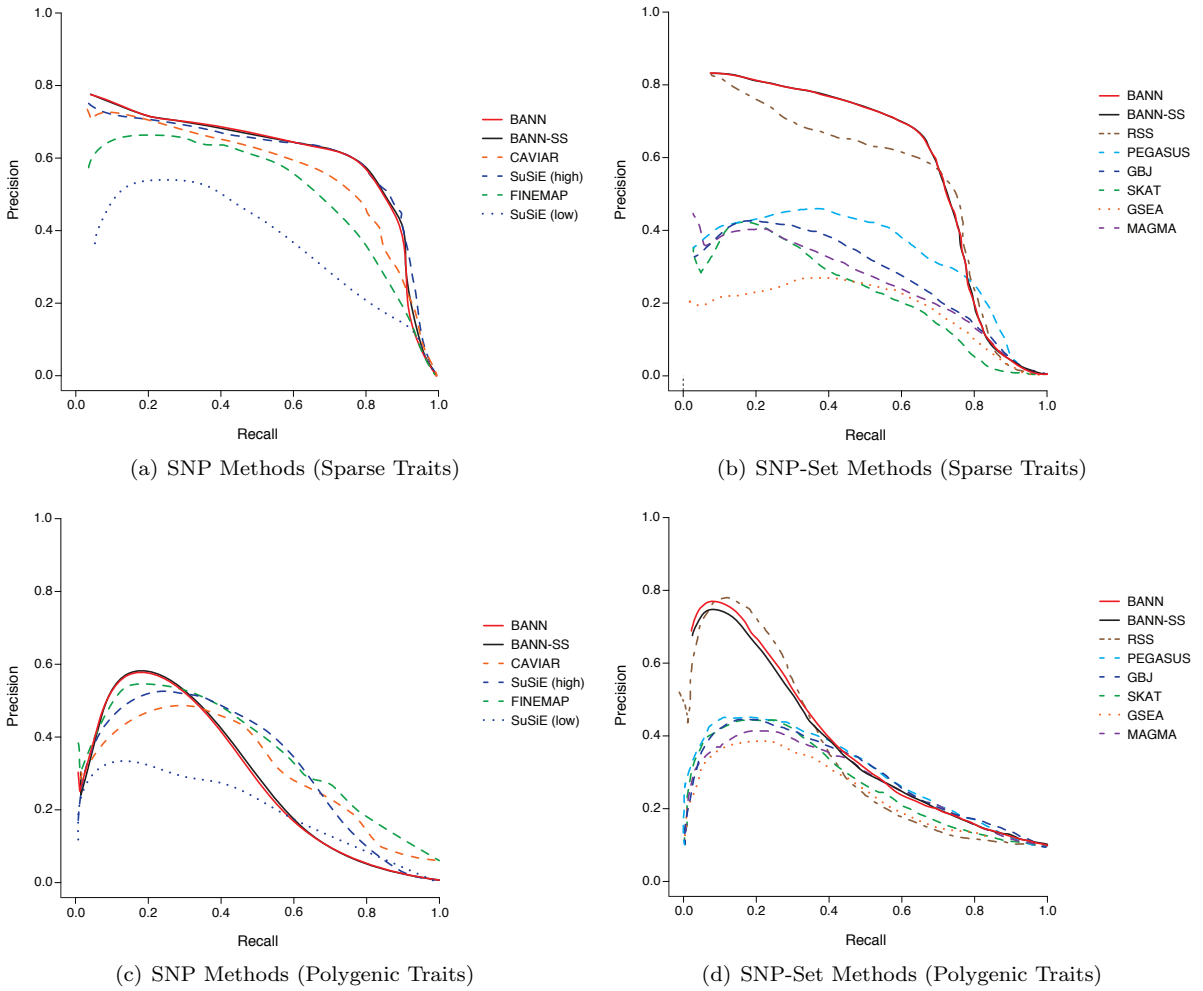
(a) SNP Methods (Sparse Traits)

(b) SNP-Set Methods (Sparse Traits)

(c) SNP Methods (Polygenic Traits)

(d) SNP-Set Methods (Polygenic Traits)

Supplementary Figure 14. Precision-recall curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations. Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.6$ with equal contributions from additive effects and epistatic interactions (i.e., $\rho = 0.5$). We show precision versus recall for two different trait architectures: **(a, b)** sparse where only 1% of SNP-sets are enriched for the trait; and **(c, d)** polygenic where 10% of SNP-sets are enriched. We then set the number of causal SNPs with non-zero effects to be 0.125% and 3% of all SNPs located within the selected enriched SNP-sets, respectively. To derive results, the full genotype matrix and phenotypic vector are given to the BANNs model and all competing methods that require individual-level data. For the BANN-SS model and other competing methods that take GWA summary statistics, we compute standard GWA SNP-level effect sizes and P-values (estimated using ordinary least squares). **(a, c)** Competing SNP-level mapping approaches include: CAVIAR [64], SuSiE [65], and FINEMAP [66]. The software for SuSiE requires an input ℓ which fixes the maximum number of causal SNPs in the model. We display results when this input number is high ($\ell = 3000$) and when this input number is low ($\ell = 10$). **(b, d)** Competing SNP-set mapping approaches include: RSS [7], PEGASUS [67], GBJ [68], SKAT [69], GSEA [70], and MAGMA [71]. Note that, for traits with sparse architectures, the top ranked SNPs and SNP-sets are always true positives, and therefore the minimal recall is not 0. All results are based on 100 replicates (see Supplementary Note, Section 8).



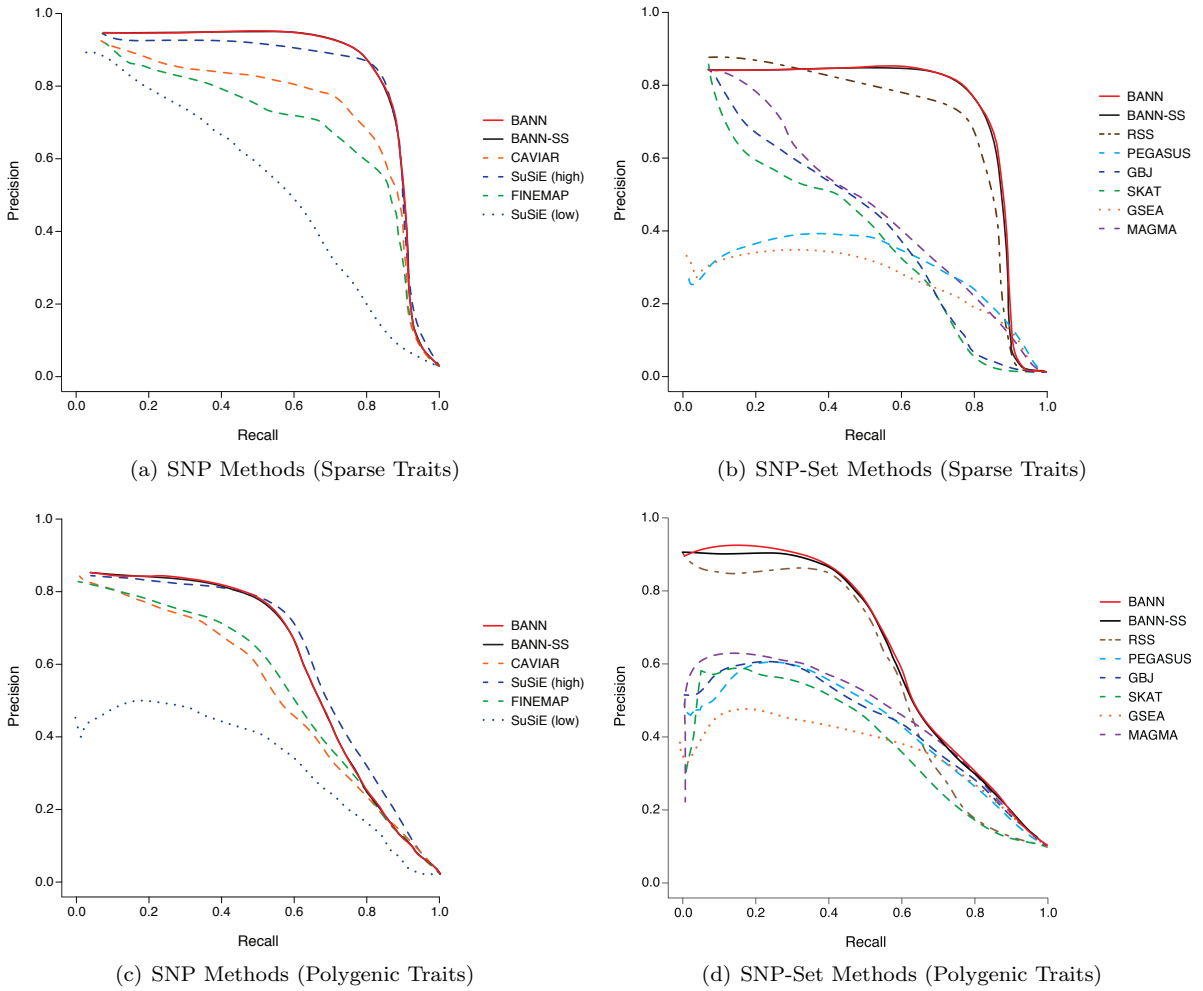
(a) SNP Methods (Sparse Traits)

(b) SNP-Set Methods (Sparse Traits)

(c) SNP Methods (Polygenic Traits)

(d) SNP-Set Methods (Polygenic Traits)

Supplementary Figure 15. Precision-recall curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations with population stratification. Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with equal contributions from additive effects and epistatic interactions (i.e., $\rho = 0.5$). In these simulations, traits were generated while using the top ten principal components (PCs) of the genotype matrix as covariates. We show precision versus recall for two different trait architectures: **(a, b)** sparse where only 1% of SNP-sets are enriched for the trait; and **(c, d)** polygenic where 10% of SNP-sets are enriched. We then set the number of causal SNPs with non-zero effects to be 0.125% and 3% of all SNPs located within the selected enriched SNP-sets, respectively. To derive results, the full genotype matrix and phenotypic vector are given to the BANNs model and all competing methods that require individual-level data. For the BANN-SS model and other competing methods that take GWA summary statistics, we compute standard GWA SNP-level effect sizes and P-values (estimated using ordinary least squares). **(a, c)** Competing SNP-level mapping approaches include: CAVIAR [64], SuSiE [65], and FINEMAP [66]. The software for SuSiE requires an input ℓ which fixes the maximum number of causal SNPs in the model. We display results when this input number is high ($\ell = 3000$) and when this input number is low ($\ell = 10$). **(b, d)** Competing SNP-set mapping approaches include: RSS [7], PEGASUS [67], GBJ [68], SKAT [69], GSEA [70], and MAGMA [71]. Note that, for traits with sparse architectures, the top ranked SNPs and SNP-sets are always true positives, and therefore the minimal recall is not 0. All results are based on 100 replicates (see Supplementary Note, Section 8).



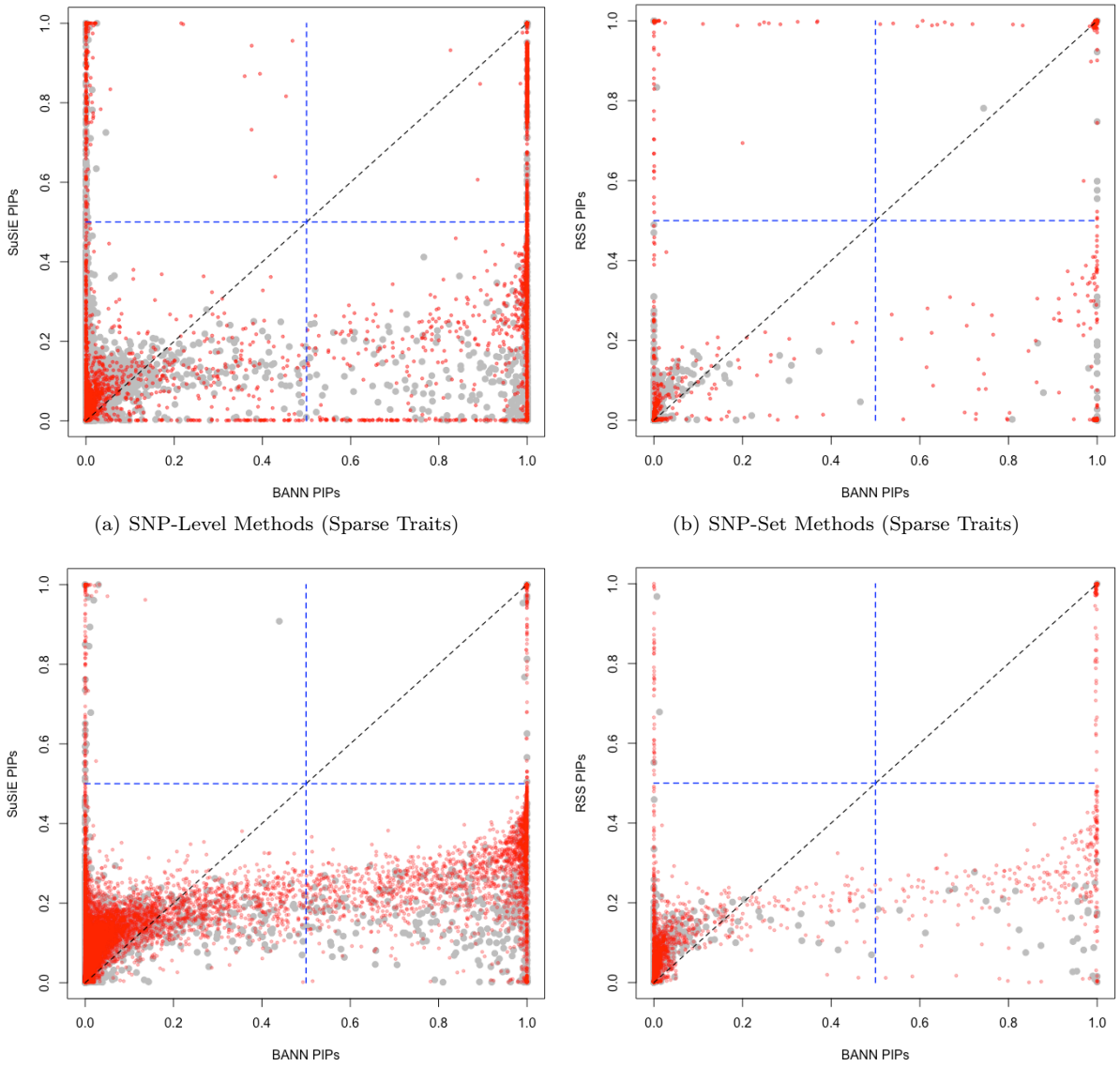
(a) SNP Methods (Sparse Traits)

(b) SNP-Set Methods (Sparse Traits)

(c) SNP Methods (Polygenic Traits)

(d) SNP-Set Methods (Polygenic Traits)

Supplementary Figure 16. Precision-recall curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations with population stratification. Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with equal contributions from additive effects and epistatic interactions (i.e., $\rho = 0.5$). In these simulations, traits were generated while using the top ten principal components (PCs) of the genotype matrix as covariates. We show precision versus recall for two different trait architectures: **(a, b)** sparse where only 1% of SNP-sets are enriched for the trait; and **(c, d)** polygenic where 10% of SNP-sets are enriched. We then set the number of causal SNPs with non-zero effects to be 0.125% and 3% of all SNPs located within the selected enriched SNP-sets, respectively. To derive results, the full genotype matrix and phenotypic vector are given to the BANNs model and all competing methods that require individual-level data. For the BANN-SS model and other competing methods that take GWA summary statistics, we compute standard GWA SNP-level effect sizes and P-values (estimated using ordinary least squares). **(a, c)** Competing SNP-level mapping approaches include: CAVIAR [64], SuSiE [65], and FINEMAP [66]. The software for SuSiE requires an input ℓ which fixes the maximum number of causal SNPs in the model. We display results when this input number is high ($\ell = 3000$) and when this input number is low ($\ell = 10$). **(b, d)** Competing SNP-set mapping approaches include: RSS [7], PEGASUS [67], GBJ [68], SKAT [69], GSEA [70], and MAGMA [71]. Note that, for traits with sparse architectures, the top ranked SNPs and SNP-sets are always true positives, and therefore the minimal recall is not 0. All results are based on 100 replicates (see Supplementary Note, Section 8).



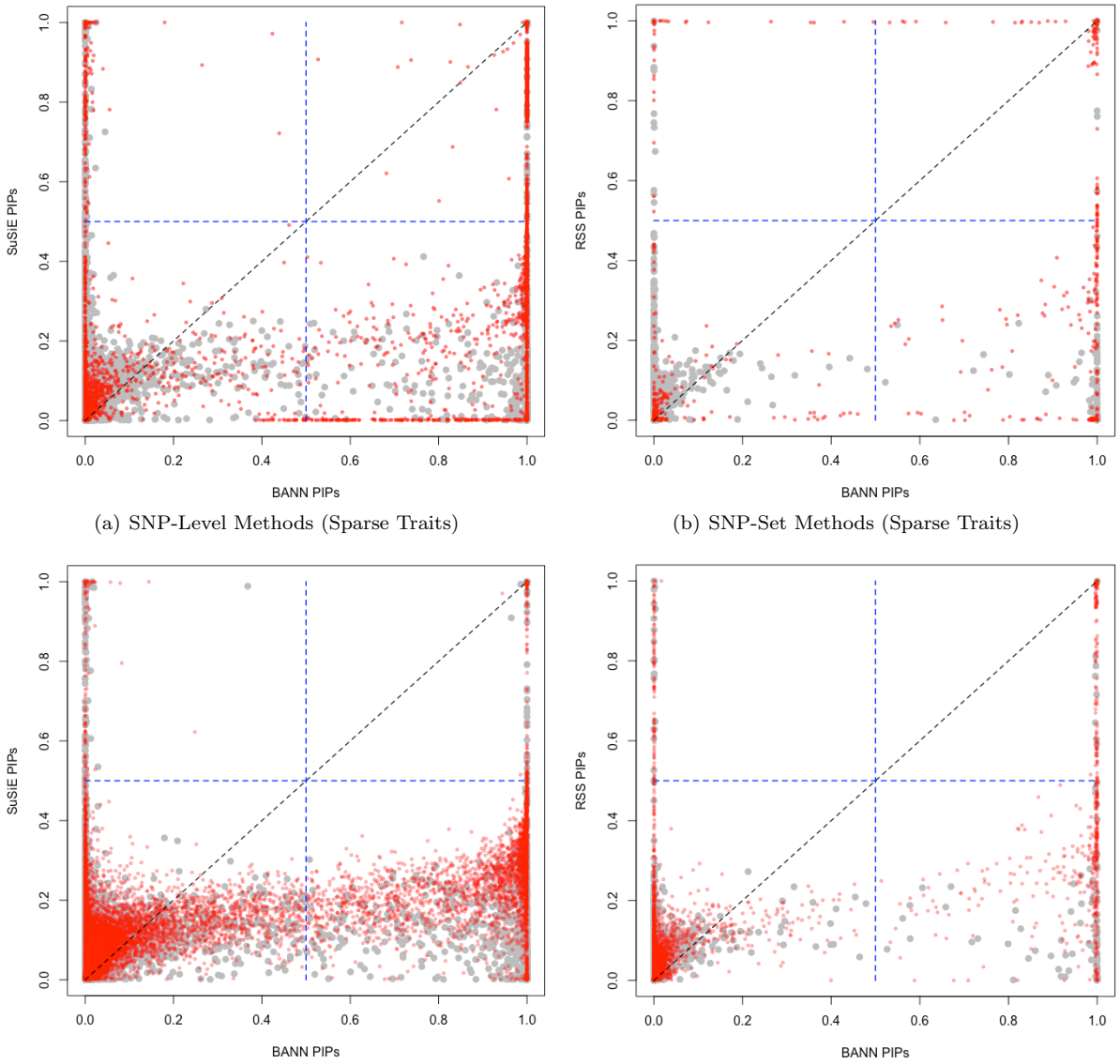
(a) SNP-Level Methods (Sparse Traits)

(b) SNP-Set Methods (Sparse Traits)

(c) SNP-Level Methods (Polygenic Traits)

(d) SNP-Set Methods (Polygenic Traits)

Supplementary Figure 17. Scatter plots comparing how the integrative neural network training procedure enables the ability to identify associated SNPs and enriched SNP-sets in simulations. Quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with only contributions from additive effects set (i.e., $\rho = 1$). We consider two different trait architectures: **(a, b)** sparse where only 1% of SNP-sets are enriched for the trait; and **(c, d)** polygenic where 10% of SNP-sets are enriched. We set the number of causal SNPs with non-zero effects to be 0.125% and 3% of all SNPs located within the enriched SNP-sets, respectively. Results are shown comparing the posterior inclusion probabilities (PIPs) derived by the BANNs model fit with individual-level data on the x-axis and **(a, c)** SuSiE [65] and **(b, d)** RSS [7] on the y-axis, respectively. Here, SuSie is fit while assuming a high maximum number of causal SNPs ($\ell = 3000$). The blue horizontal and vertical dashed lines are marked at the "median probability criterion" (i.e., PIPs for SNPs and SNP-sets greater than 0.5) [52]. True positive causal variants used to generate the synthetic phenotypes are colored in red, while non-causal variants are given in grey. SNPs and SNP-sets in the top right quadrant are selected by both approaches; while, elements in the bottom right and top left quadrants are uniquely identified by BANNs and SuSie/RSS, respectively. Each plot combines results from 100 simulated replicates (see Section 8).



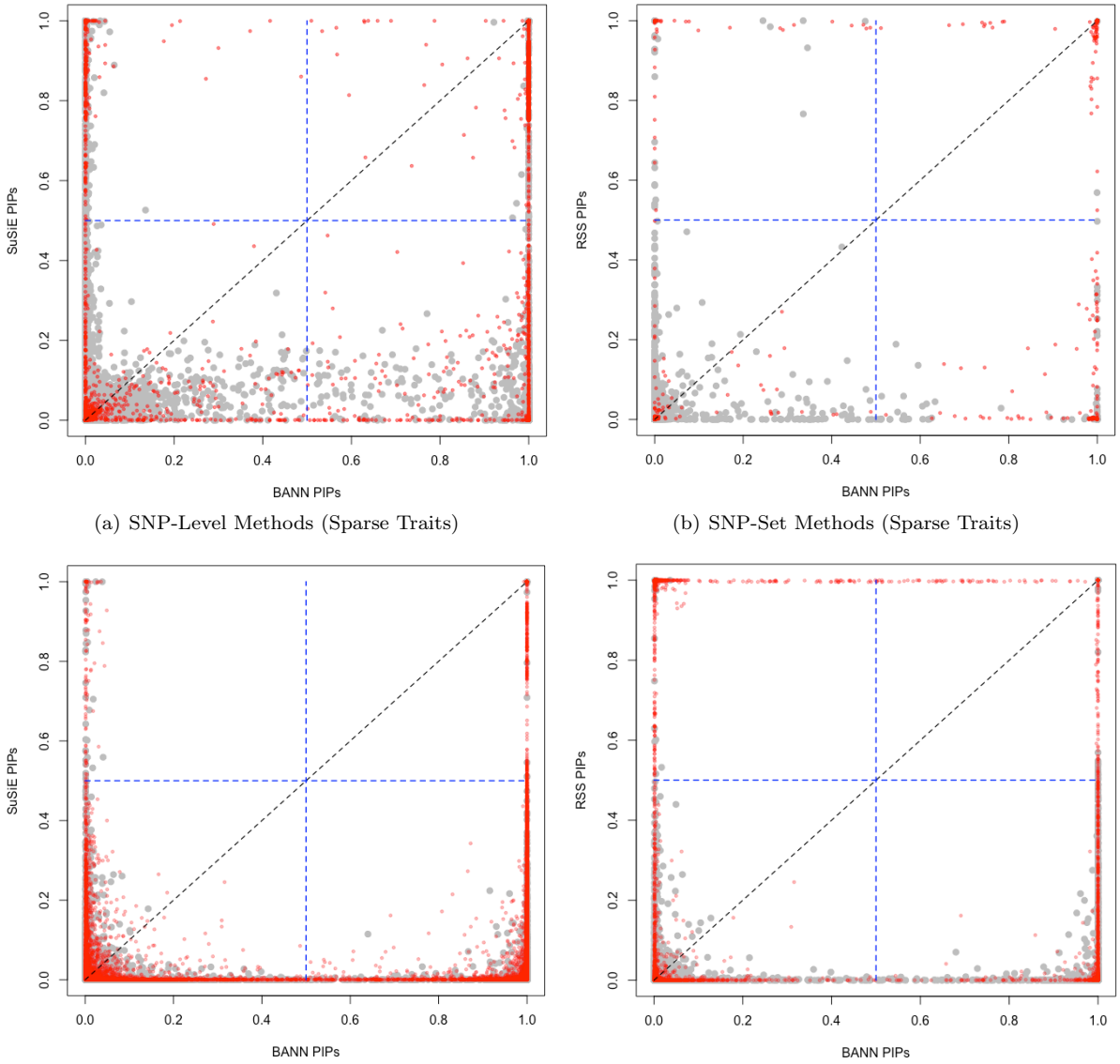
(a) SNP-Level Methods (Sparse Traits)

(b) SNP-Set Methods (Sparse Traits)

(c) SNP-Level Methods (Polygenic Traits)

(d) SNP-Set Methods (Polygenic Traits)

Supplementary Figure 18. Scatter plots comparing how the integrative neural network training procedure enables the ability to identify associated SNPs and enriched SNP-sets in simulations with population stratification. Quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with only contributions from additive effects set (i.e., $\rho = 1$). We consider two different trait architectures: **(a, b)** sparse where only 1% of SNP-sets are enriched for the trait; and **(c, d)** polygenic where 10% of SNP-sets are enriched. We set the number of causal SNPs with non-zero effects to be 0.125% and 3% of all SNPs located within the enriched SNP-sets, respectively. In these simulations, traits were generated while also using the top ten principal components (PCs) of the genotype matrix as covariates. Results are shown comparing the posterior inclusion probabilities (PIPs) derived by the BANNs model fit with individual-level data on the x-axis and **(a, c)** SuSiE [65] and **(b, d)** RSS [7] on the y-axis, respectively. Here, SuSie is fit while assuming a high maximum number of causal SNPs ($\ell = 3000$). The blue horizontal and vertical dashed lines are marked at the "median probability criterion" (i.e., PIPs for SNPs and SNP-sets greater than 0.5) [52]. True positive causal variants used to generate the synthetic phenotypes are colored in red, while non-causal variants are given in grey. SNPs and SNP-sets in the top right quadrant are selected by both approaches; while, elements in the bottom right and top left quadrants are uniquely identified by BANNs and SuSie/RSS, respectively. Each plot combines results from 100 simulated replicates (see Section 8).



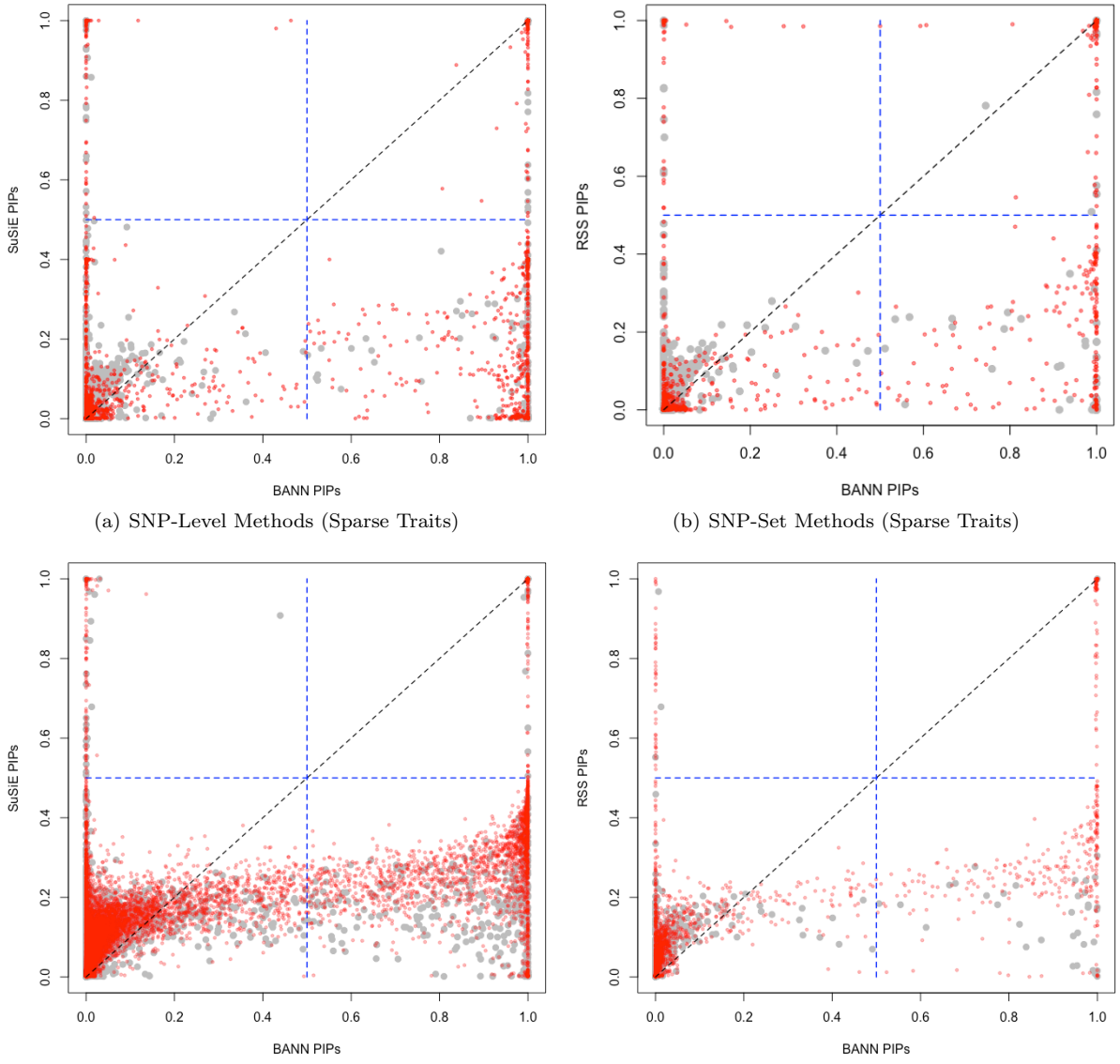
(a) SNP-Level Methods (Sparse Traits)

(b) SNP-Set Methods (Sparse Traits)

(c) SNP-Level Methods (Polygenic Traits)

(d) SNP-Set Methods (Polygenic Traits)

Supplementary Figure 19. Scatter plots comparing how the integrative neural network training procedure enables the ability to identify associated SNPs and enriched SNP-sets in simulations with population stratification. Quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.6$ with only contributions from additive effects set (i.e., $\rho = 1$). We consider two different trait architectures: **(a, b)** sparse where only 1% of SNP-sets are enriched for the trait; and **(c, d)** polygenic where 10% of SNP-sets are enriched. We set the number of causal SNPs with non-zero effects to be 0.125% and 3% of all SNPs located within the enriched SNP-sets, respectively. In these simulations, traits were generated while also using the top ten principal components (PCs) of the genotype matrix as covariates. Results are shown comparing the posterior inclusion probabilities (PIPs) derived by the BANNs model fit with individual-level data on the x-axis and **(a, c)** SuSiE [65] and **(b, d)** RSS [7] on the y-axis, respectively. Here, SuSie is fit while assuming a high maximum number of causal SNPs ($\ell = 3000$). The blue horizontal and vertical dashed lines are marked at the "median probability criterion" (i.e., PIPs for SNPs and SNP-sets greater than 0.5) [52]. True positive causal variants used to generate the synthetic phenotypes are colored in red, while non-causal variants are given in grey. SNPs and SNP-sets in the top right quadrant are selected by both approaches; while, elements in the bottom right and top left quadrants are uniquely identified by BANNs and SuSie/RSS, respectively. Each plot combines results from 100 simulated replicates (see Section 8).



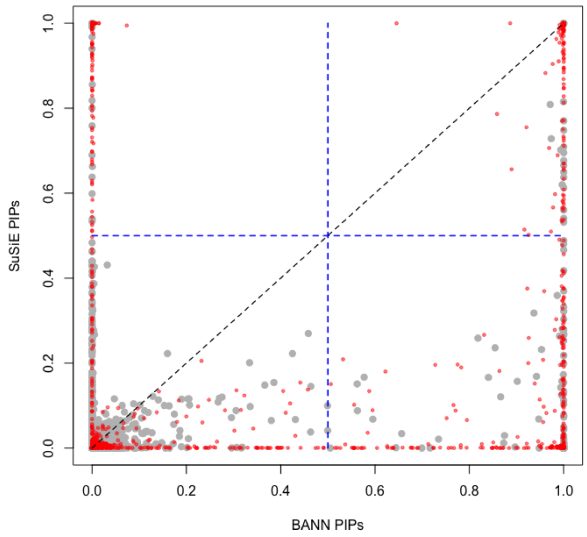
(a) SNP-Level Methods (Sparse Traits)

(b) SNP-Set Methods (Sparse Traits)

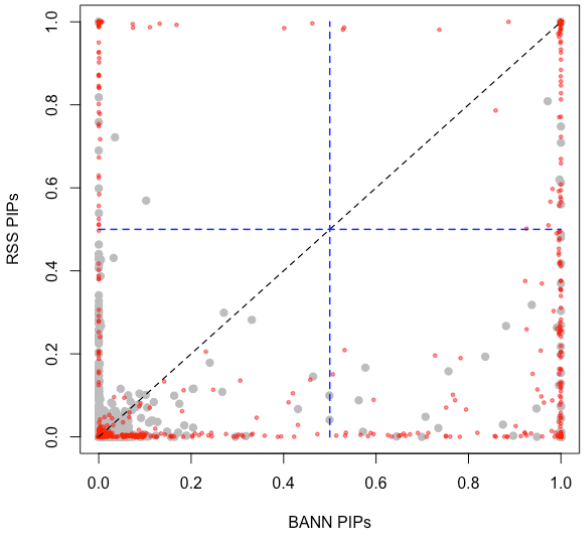
(c) SNP-Level Methods (Polygenic Traits)

(d) SNP-Set Methods (Polygenic Traits)

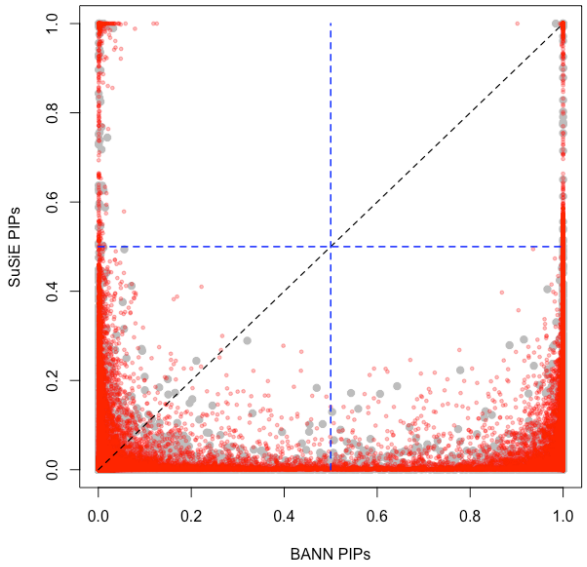
Supplementary Figure 20. Scatter plots comparing how the integrative neural network training procedure enables the ability to identify associated SNPs and enriched SNP-sets in simulations. Quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with equal contributions from additive effects and epistatic interactions (i.e., $\rho = 0.5$). We consider two different trait architectures: **(a, b)** sparse where only 1% of SNP-sets are enriched for the trait; and **(c, d)** polygenic where 10% of SNP-sets are enriched. We set the number of causal SNPs with non-zero effects to be 0.125% and 3% of all SNPs located within the enriched SNP-sets, respectively. Results are shown comparing the posterior inclusion probabilities (PIPs) derived by the BANNs model fit with individual-level data on the x-axis and **(a, c)** SuSiE [65] and **(b, d)** RSS [7] on the y-axis, respectively. Here, SuSie is fit while assuming a high maximum number of causal SNPs ($\ell = 3000$). The blue horizontal and vertical dashed lines are marked at the "median probability criterion" (i.e., PIPs for SNPs and SNP-sets greater than 0.5) [52]. True positive causal variants used to generate the synthetic phenotypes are colored in red, while non-causal variants are given in grey. SNPs and SNP-sets in the top right quadrant are selected by both approaches; while, elements in the bottom right and top left quadrants are uniquely identified by BANNs and SuSie/RSS, respectively. Each plot combines results from 100 simulated replicates (see Section 8).



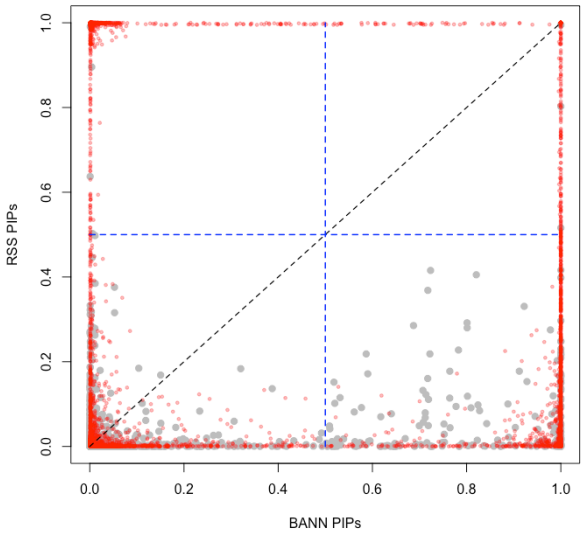
(a) SNP-Level Methods (Sparse Traits)



(b) SNP-Set Methods (Sparse Traits)

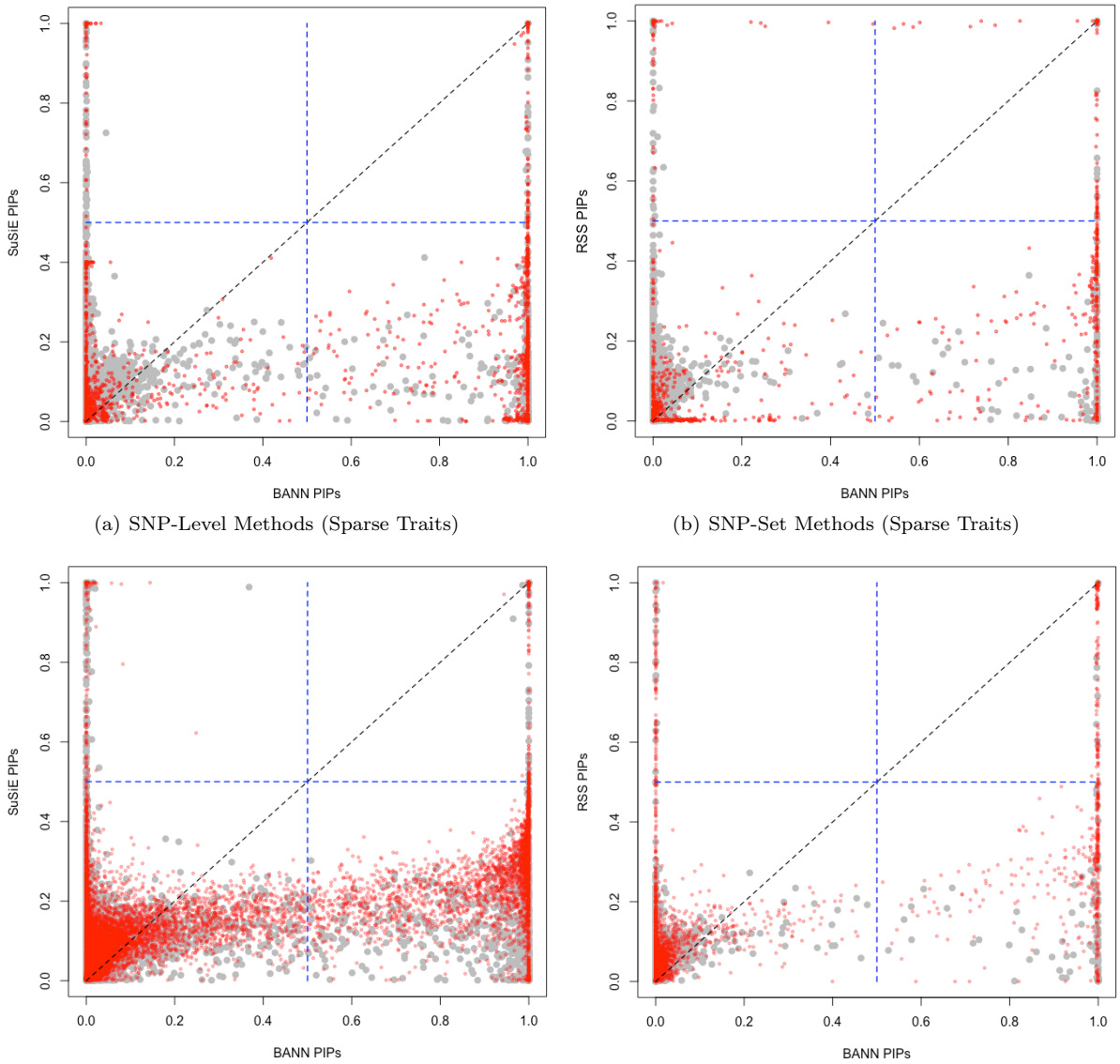


(c) SNP-Level Methods (Polygenic Traits)



(d) SNP-Set Methods (Polygenic Traits)

Supplementary Figure 21. Scatter plots comparing how the integrative neural network training procedure enables the ability to identify associated SNPs and enriched SNP-sets in simulations. Quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.6$ with equal contributions from additive effects and epistatic interactions (i.e., $\rho = 0.5$). We consider two different trait architectures: **(a, b)** sparse where only 1% of SNP-sets are enriched for the trait; and **(c, d)** polygenic where 10% of SNP-sets are enriched. We set the number of causal SNPs with non-zero effects to be 0.125% and 3% of all SNPs located within the enriched SNP-sets, respectively. Results are shown comparing the posterior inclusion probabilities (PIPs) derived by the BANNs model fit with individual-level data on the x-axis and **(a, c)** SuSiE [65] and **(b, d)** RSS [7] on the y-axis, respectively. Here, SuSie is fit while assuming a high maximum number of causal SNPs ($\ell = 3000$). The blue horizontal and vertical dashed lines are marked at the "median probability criterion" (i.e., PIPs for SNPs and SNP-sets greater than 0.5) [52]. True positive causal variants used to generate the synthetic phenotypes are colored in red, while non-causal variants are given in grey. SNPs and SNP-sets in the top right quadrant are selected by both approaches; while, elements in the bottom right and top left quadrants are uniquely identified by BANNs and SuSie/RSS, respectively. Each plot combines results from 100 simulated replicates (see Section 8).



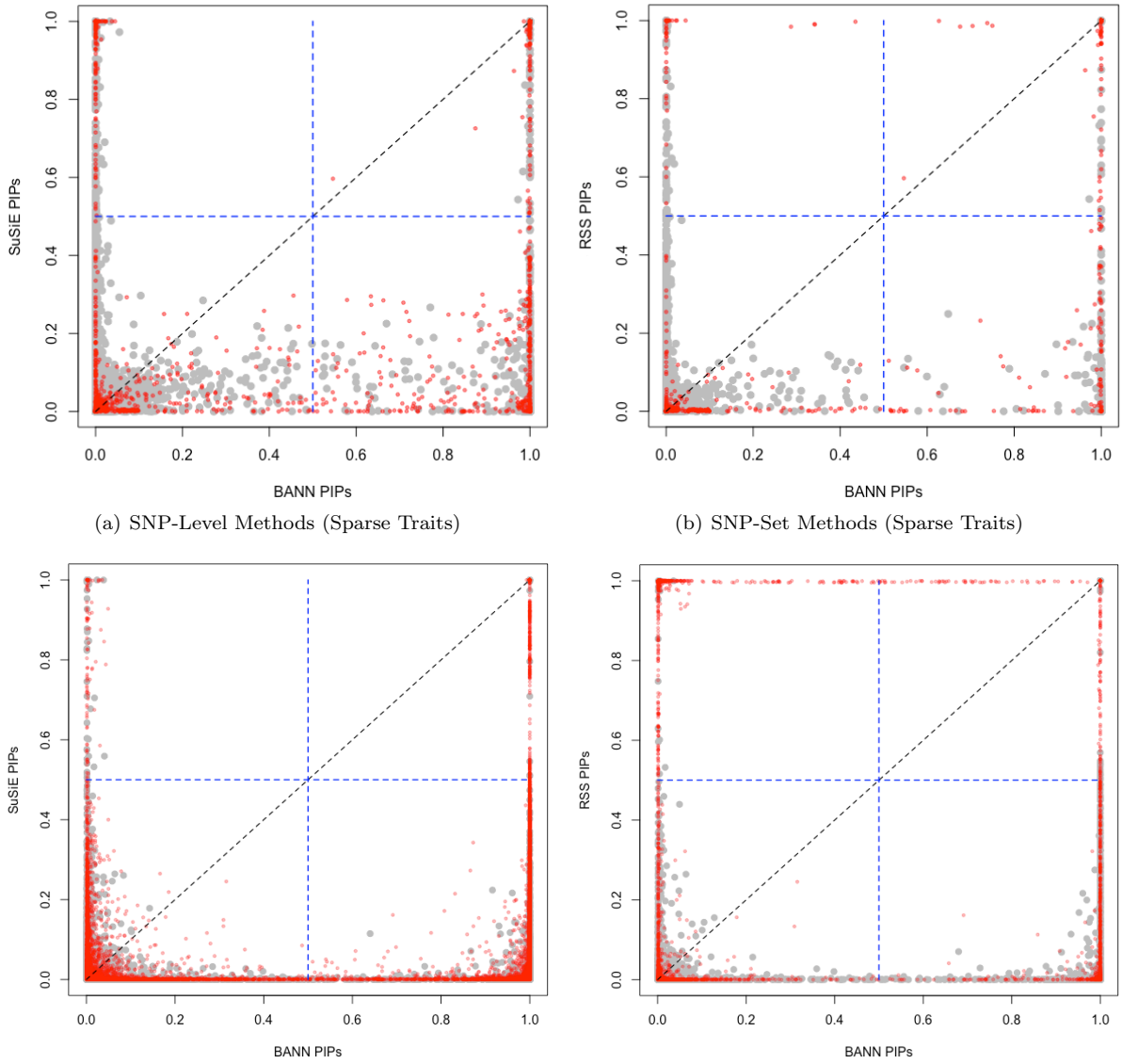
(a) SNP-Level Methods (Sparse Traits)

(b) SNP-Set Methods (Sparse Traits)

(c) SNP-Level Methods (Polygenic Traits)

(d) SNP-Set Methods (Polygenic Traits)

Supplementary Figure 22. Scatter plots comparing how the integrative neural network training procedure enables the ability to identify associated SNPs and enriched SNP-sets in simulations with population stratification. Quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with equal contributions from additive effects and epistatic interactions (i.e., $\rho = 0.5$). We consider two different trait architectures: **(a, b)** sparse where only 1% of SNP-sets are enriched for the trait; and **(c, d)** polygenic where 10% of SNP-sets are enriched. We set the number of causal SNPs with non-zero effects to be 0.125% and 3% of all SNPs located within the enriched SNP-sets, respectively. In these simulations, traits were generated while also using the top ten principal components (PCs) of the genotype matrix as covariates. Results are shown comparing the posterior inclusion probabilities (PIPs) derived by the BANNs model fit with individual-level data on the x-axis and **(a, c)** SuSiE [65] and **(b, d)** RSS [7] on the y-axis, respectively. Here, SuSie is fit while assuming a high maximum number of causal SNPs ($\ell = 3000$). The blue horizontal and vertical dashed lines are marked at the "median probability criterion" (i.e., PIPs for SNPs and SNP-sets greater than 0.5) [52]. True positive causal variants used to generate the synthetic phenotypes are colored in red, while non-causal variants are given in grey. SNPs and SNP-sets in the top right quadrant are selected by both approaches; while, elements in the bottom right and top left quadrants are uniquely identified by BANNs and SuSie/RSS, respectively. Each plot combines results from 100 simulated replicates (see Section 8).



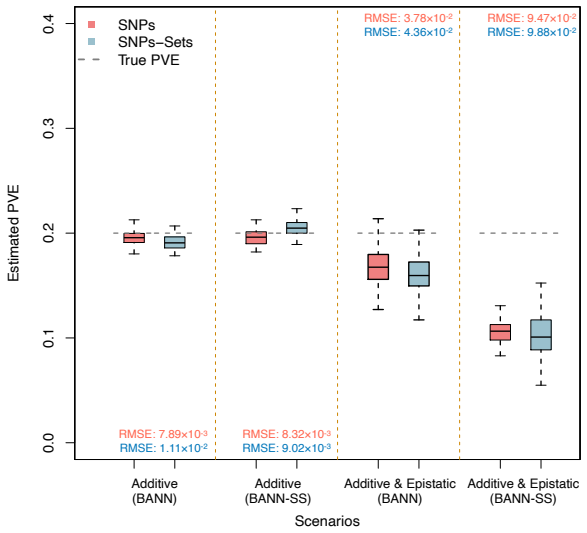
(a) SNP-Level Methods (Sparse Traits)

(b) SNP-Set Methods (Sparse Traits)

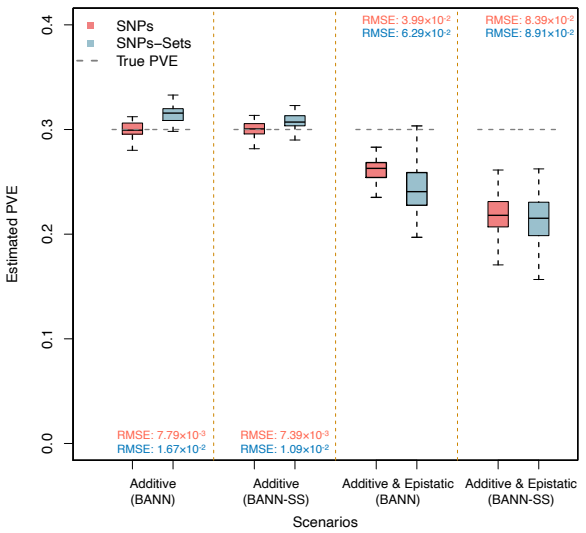
(c) SNP-Level Methods (Polygenic Traits)

(d) SNP-Set Methods (Polygenic Traits)

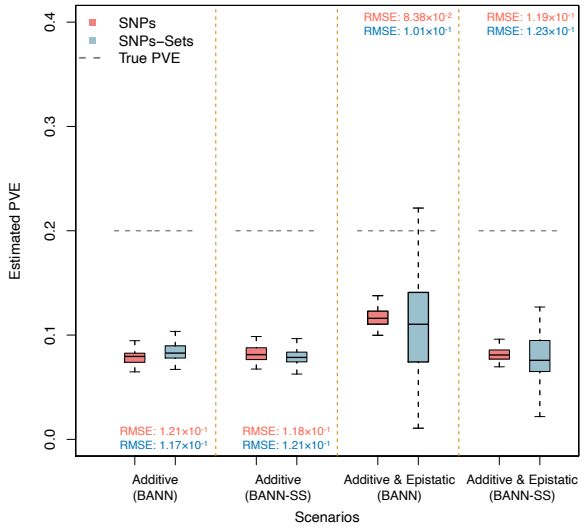
Supplementary Figure 23. Scatter plots comparing how the integrative neural network training procedure enables the ability to identify associated SNPs and enriched SNP-sets in simulations with population stratification. Quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.6$ with equal contributions from additive effects and epistatic interactions (i.e., $\rho = 0.5$). We consider two different trait architectures: **(a, b)** sparse where only 1% of SNP-sets are enriched for the trait; and **(c, d)** polygenic where 10% of SNP-sets are enriched. We set the number of causal SNPs with non-zero effects to be 0.125% and 3% of all SNPs located within the enriched SNP-sets, respectively. In these simulations, traits were generated while also using the top ten principal components (PCs) of the genotype matrix as covariates. Results are shown comparing the posterior inclusion probabilities (PIPs) derived by the BANNs model fit with individual-level data on the x-axis and **(a, c)** SuSiE [65] and **(b, d)** RSS [7] on the y-axis, respectively. Here, SuSie is fit while assuming a high maximum number of causal SNPs ($\ell = 3000$). The blue horizontal and vertical dashed lines are marked at the "median probability criterion" (i.e., PIPs for SNPs and SNP-sets greater than 0.5) [52]. True positive causal variants used to generate the synthetic phenotypes are colored in red, while non-causal variants are given in grey. SNPs and SNP-sets in the top right quadrant are selected by both approaches; while, elements in the bottom right and top left quadrants are uniquely identified by BANNs and SuSie/RSS, respectively. Each plot combines results from 100 simulated replicates (see Section 8).



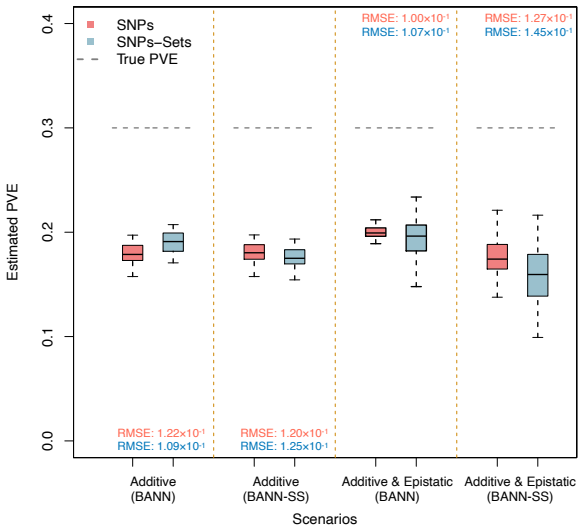
(a) Sparse Traits



(b) Sparse Traits with Population Structure

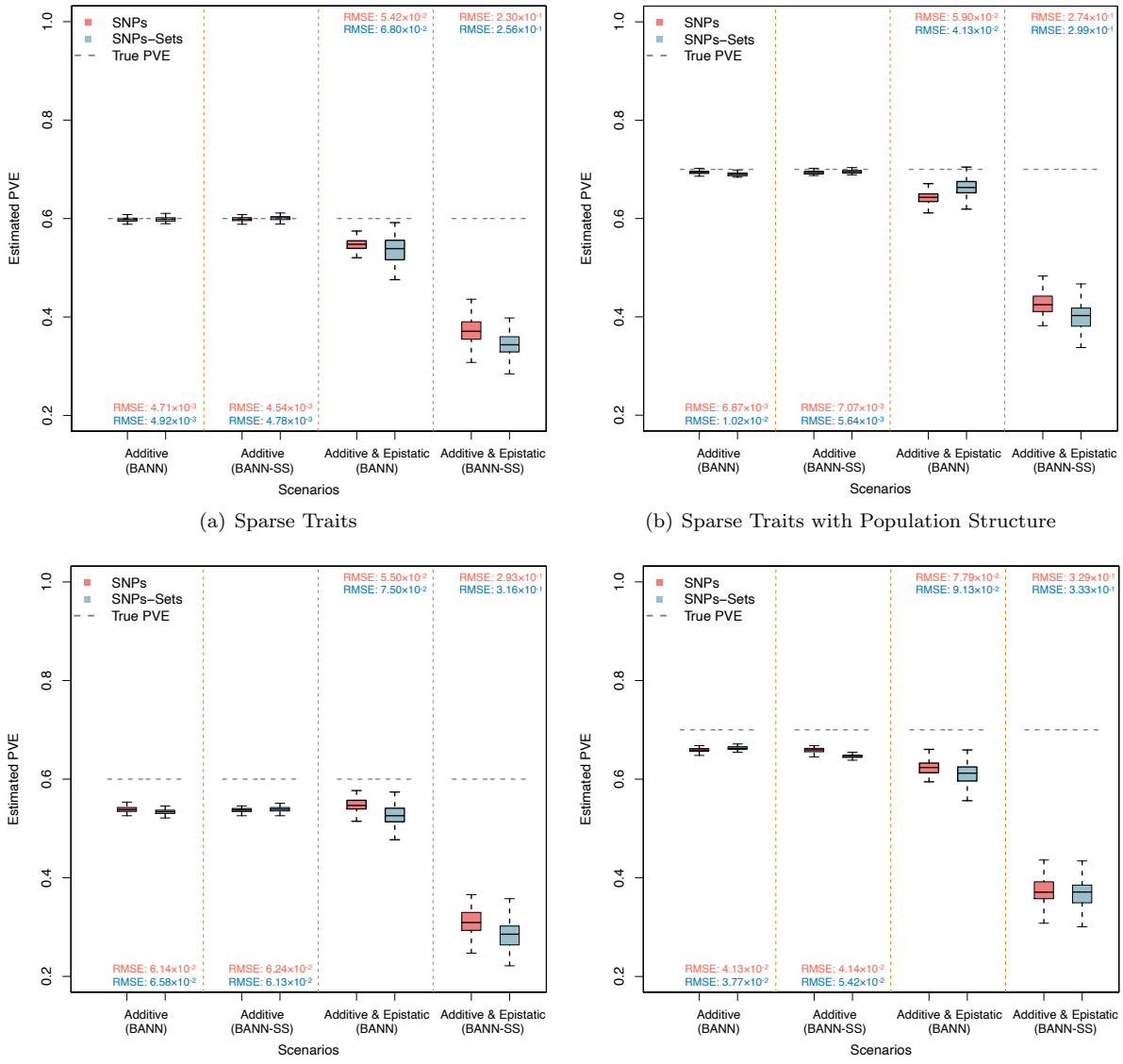


(c) Polygenic Traits



(d) Polygenic Traits with Population Structure

Supplementary Figure 24. Boxplots depicting the ability of the BANNs and BANN-SS models to estimate the phenotypic variation explained (PVE) by SNPs (pink) and SNP-sets (blue) for complex traits in simulations. In this work, we define PVE as the total proportion of phenotypic variance that is explained by fixed and random genetic effects, collectively [15]. Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with different levels of contributions from additive effects and epistatic interactions. We consider two different trait architectures: **(a, b)** sparse where only 1% of SNP-sets are enriched for the trait; and **(c, d)** polygenic where 10% of SNP-sets are enriched. The number of causal SNPs with non-zero effects is set to be 0.125% and 3% of all SNPs located within the selected enriched SNP-sets, respectively. For panels **(b, d)**, traits were generated while also assuming that the top ten principal components (PCs) of the genotype matrix contribute 10% to the phenotypic variance. Therefore, in panels **(a, c)**, the true PVE $= H^2 = 20\%$; while, in panels **(b, d)**, the true total PVE $= H^2 + 10\% = 30\%$. These true values are shown as the dashed grey horizontal lines. The root mean square error (RMSE) between the BANNs model estimates of the PVE and the true values are also provided.



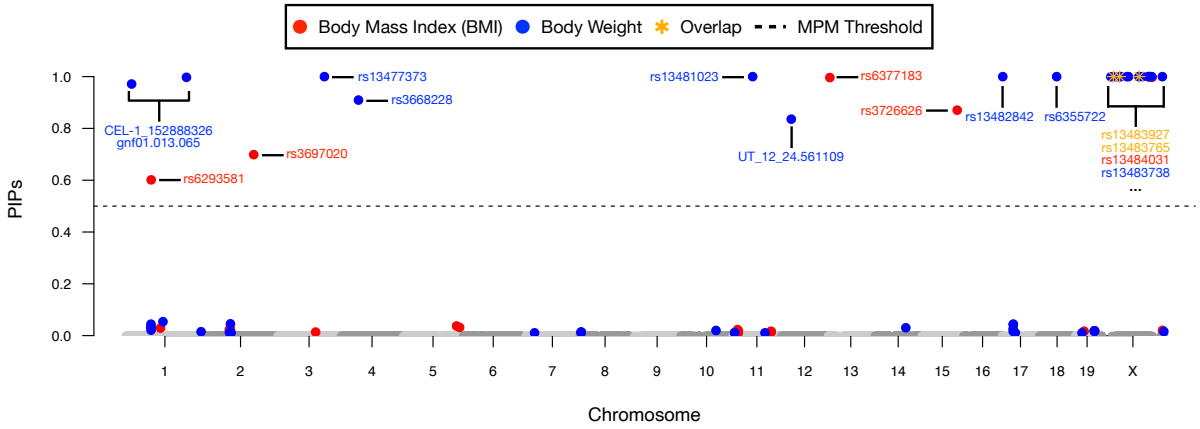
(a) Sparse Traits

(b) Sparse Traits with Population Structure

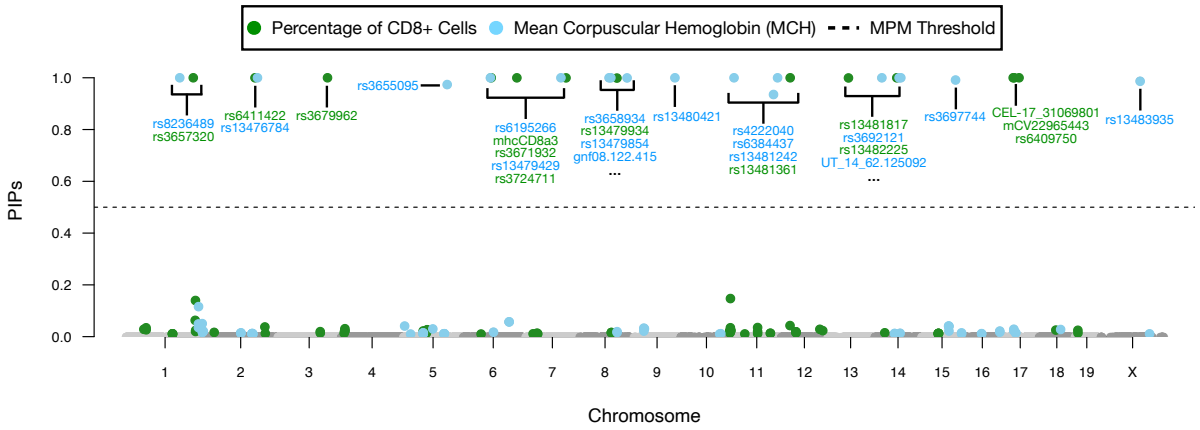
(c) Polygenic Traits

(d) Polygenic Traits with Population Structure

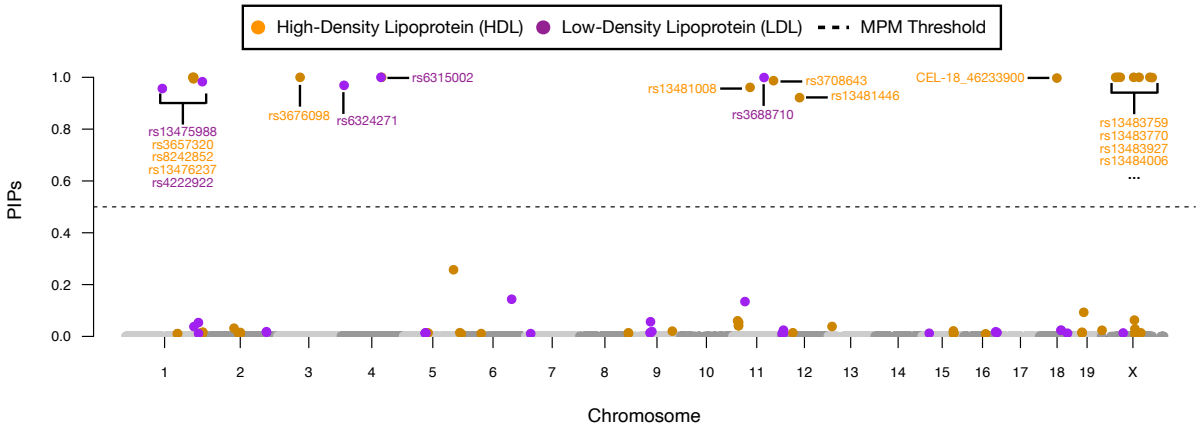
Supplementary Figure 25. Boxplots depicting the ability of the BANNs and BANN-SS models to estimate the phenotypic variation explained (PVE) by SNPs (pink) and SNP-sets (blue) for complex traits in simulations. In this work, we define PVE as the total proportion of phenotypic variance that is explained by fixed and random genetic effects, collectively [15]. Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.6$ with different levels of contributions from additive effects and epistatic interactions. We consider two different trait architectures: **(a, b)** sparse where only 1% of SNP-sets are enriched for the trait; and **(c, d)** polygenic where 10% of SNP-sets are enriched. The number of causal SNPs with non-zero effects is set to be 0.125% and 3% of all SNPs located within the selected enriched SNP-sets, respectively. For panels **(b, d)**, traits were generated while also assuming that the top ten principal components (PCs) of the genotype matrix contribute 10% to the phenotypic variance. Therefore, in panels **(a, c)**, the true PVE = H^2 = 60%; while, in panels **(b, d)**, the true total PVE = $H^2 + 10\% = 70\%$. These true values are shown as the dashed grey horizontal lines. The root mean square error (RMSE) between the BANNs model estimates of the PVE and the true values are also provided.



(a) Body Mass Index and Body Weight



(b) % CD8+ cells and Mean Corpuscular Hemoglobin



(c) High-Density and Low-Density Lipoprotein

Supplementary Figure 26. Manhattan plots of variant-level fine mapping results for six traits in heterogenous stock of mice from the Wellcome Trust Centre for Human Genetics. Traits are grouped based on their category and include: **(a)** body mass index (BMI) and body weight, **(b)** percentage of CD8+ cells and mean corpuscular hemoglobin (MCH), and **(c)** high-density and low-density lipoprotein (HDL and LDL, respectively) cholesterol. Posterior inclusion probabilities (PIP) for the input layer weights are derived from the BANNs model fit on individual-level data and are plotted for each SNP against their genomic positions. Chromosomes are shown in alternating colors for clarity. The black dashed line is marked at 0.5 and represents the "median probability model (MPM)" threshold [52]. Here, we only color code SNPs that had a PIP greater than 1% in either trait. SNPs with PIPs exceeding 1% in both traits are marked by a star and denoted as falling in the "overlap" category. BANNs estimated the following PVEs on the SNP and SNP-set levels for these traits, respectively: (i) 0.09 and 0.08 for BMI, (ii) 0.39 and 0.40 for body weight, (iii) 0.51 and 0.48 for percentage of CD8+ cells, (iv) 0.34 and 0.32 for MCH, (v) 0.34 and 0.28 for HDL, and (vi) 0.15 and 0.15 for LDL.

(a)

	p value	q value	Odds.ratio	Combined score	# of sig. genes in dbGaP
Metabolic Syndrome X	1.316e-04	4.541e-02	115.94	1036.01	2
Lipoproteins, HDL	9.365e-04	1.616e-01	43.72	304.85	2
Cholesterol, HDL	2.188e-03	2.516e-01	11.20	68.63	3
Apolipoprotein A-I	8.221e-03	7.091e-01	121.21	581.95	1
Natriuretic Peptide, Brain	1.342e-02	9.26e-01	74.07	319.33	1
Triglycerides	1.402e-02	8.06e-01	10.93	46.64	2
Lipids	6.332e-02	1.000	15.33	42.29	1
Arteries	6.473e-02	1.000	14.98	41.01	1
Alcoholism	6.755e-02	1.000	14.34	38.64	1
Iron	1.095e-01	1.000	8.66	19.15	1

(b)

	p value	q value	Odds.ratio	Combined score	# of sig. genes in dbGaP
1-Alkyl-2-acetylglycerophosphocholine Esterase	2.098e-03	7.239e-01	476.19	2936.47	1
Lipoproteins, LDL	1.075e-02	1.000	92.59	419.68	1
Alzheimer Disease	2.376e-02	1.000	41.67	155.82	1
Coronary Disease	2.435e-02	1.000	40.65	151.02	1
Coronary Artery Disease	5.995e-02	1.000	16.26	45.76	1
Myocardial Infarction	6.677e-02	1.000	14.56	39.40	1
Triglycerides	7.101e-02	1.000	13.66	36.13	1
Cholesterol	7.776e-02	1.000	12.44	31.77	1
Cholesterol, LDL	8.781e-02	1.000	10.96	26.67	1
Cholesterol, HDL	1.024e-01	1.000	9.34	21.27	1

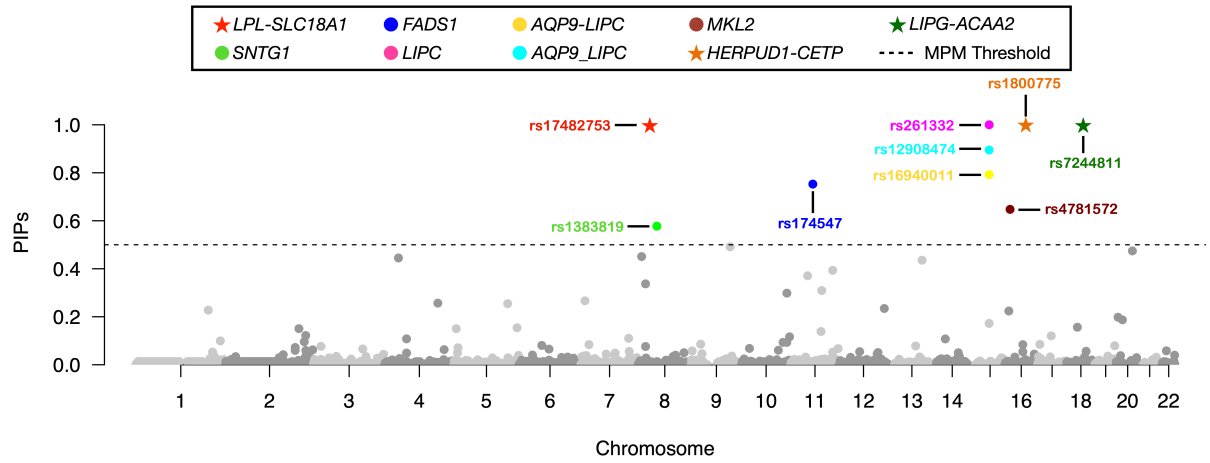
(c)

	p value	q value	Odds.ratio	Combined score	# of sig. genes in GWAS Catalog
Apolipoprotein A1 levels	3.409e-09	5.922e-06	800	15597.40	3
High density lipoprotein cholesterol levels	1.932e-07	1.678e-04	75.12	1161.28	4
HDL cholesterol levels	5.749e-07	3.329e-04	57.35	824.03	4
HDL cholesterol	6.234e-07	2.707e-04	28.01	400.23	5
Triglyceride levels	9.011e-07	3.130e-04	51.28	713.83	4
Metabolite levels (lipoprotein measures)	9.875e-07	2.859e-04	148.15	2048.61	3
Lipid metabolism phenotypes	2.611e-06	6.480e-04	108.11	1389.80	3
Metabolic syndrome	4.752e-06	1.032e-03	88.89	1089.51	3
Mean diameter of HDL particles	7.861e-06	1.517e-03	444.44	5223.81	2
Total cholesterol levels	9.924e-06	1.724e-03	28.07	323.38	4

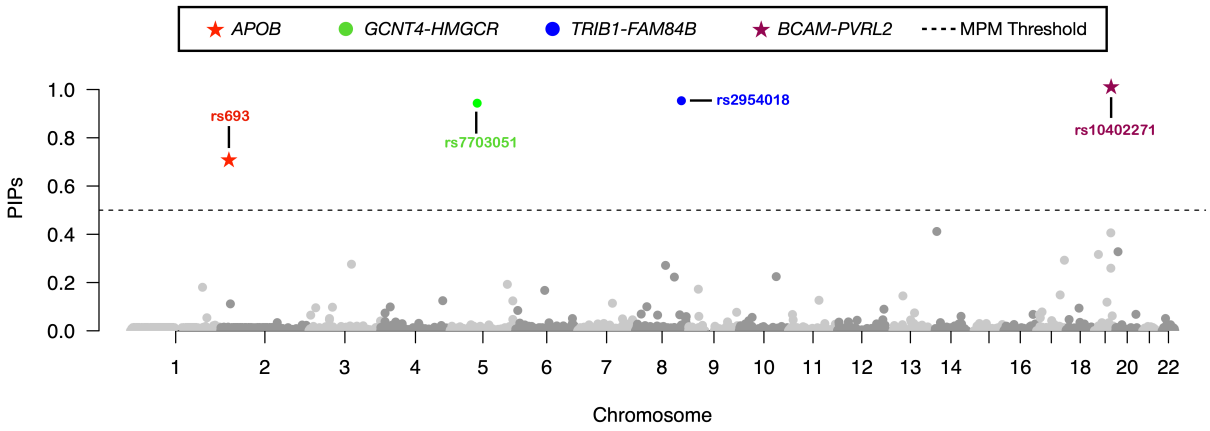
(d)

	p value	q value	Odds.ratio	Combined score	# of sig. genes in GWAS Catalog
Waist-to-hip circumference ratio (smoking years interaction)	7.497e-07	1.302e-03	1333.33	18804.85	2
Cerebrospinal fluid t-tau levels in mild cognitive impairment	1.018e-05	8.841e-03	392.16	4507.90	2
Cerebrospinal AB1-42 levels in mild cognitive impairment	1.280e-05	7.409e-03	350.88	3953.12	2
Cerebrospinal AB1-42 levels in Alzheimer's disease dementia	1.571e-05	6.822e-03	317.46	3511.50	2
Logical memory (immediate recall)	1.892e-05	6.573e-03	289.86	3152.23	2
Logical memory (delayed recall)	2.064e-05	5.121e-03	277.78	2996.75	2
Cerebrospinal fluid p-tau levels	2.064e-05	5.975e-03	277.78	2996.75	2
Cerebrospinal fluid p-tau levels in mild cognitive impairment	2.43e-05	5.275e-03	256.41	2724.40	2
Cerebrospinal fluid t-tau levels	2.624e-05	5.064e-03	246.91	2604.53	2
Bladder cancer	2.825e-05	4.907e-03	238.10	2493.89	2

Supplementary Figure 27. Gene set enrichment analyses using the significant SNP-sets identified by BANNs for high-density and low-density lipoprotein (HDL and LDL, respectively) traits in the Framingham Heart Study [48]. Here, SNP-set annotations are based on gene boundaries defined by the NCBI's RefSeq database in the UCSC Genome Browser [50]. Unannotated SNPs located within the same genomic region were labeled as being within the "intergenic region" between two genes. Posterior inclusion probabilities (PIP) for the input and hidden layer weights are derived by fitting the BANNs model on individual-level data. A SNP-set is considered significant if it has a PIP $\boldsymbol{\gamma}_w \geq 0.5$ (i.e., the "median probability model" threshold [52]). We take these significant SNP-sets and conduct "gene set enrichment analysis" using Enrichr [72, 73] to identify the categories they overrepresent in **(a, b)** the database of Genotypes and Phenotypes (dbGaP) and **(c, d)** the GWAS Catalog (2019). We highlight categories with Q-values (i.e., false discovery rates) below 0.05. Nearly all enriched categories are related with **(a, c)** HDL and **(b, d)** LDL, respectively. Note that in LDL, the BANNs framework identified the gene *APOB* as having a high PIP ($\boldsymbol{\gamma}_w = 0.976$). There have been hypotheses connecting LDL to cognitive traits [74, 75], and *APOB* has been shown to be related to cerebrospinal fluid and memory [76–78]. Therefore, we argue that results in panel **(d)** are also relevant.

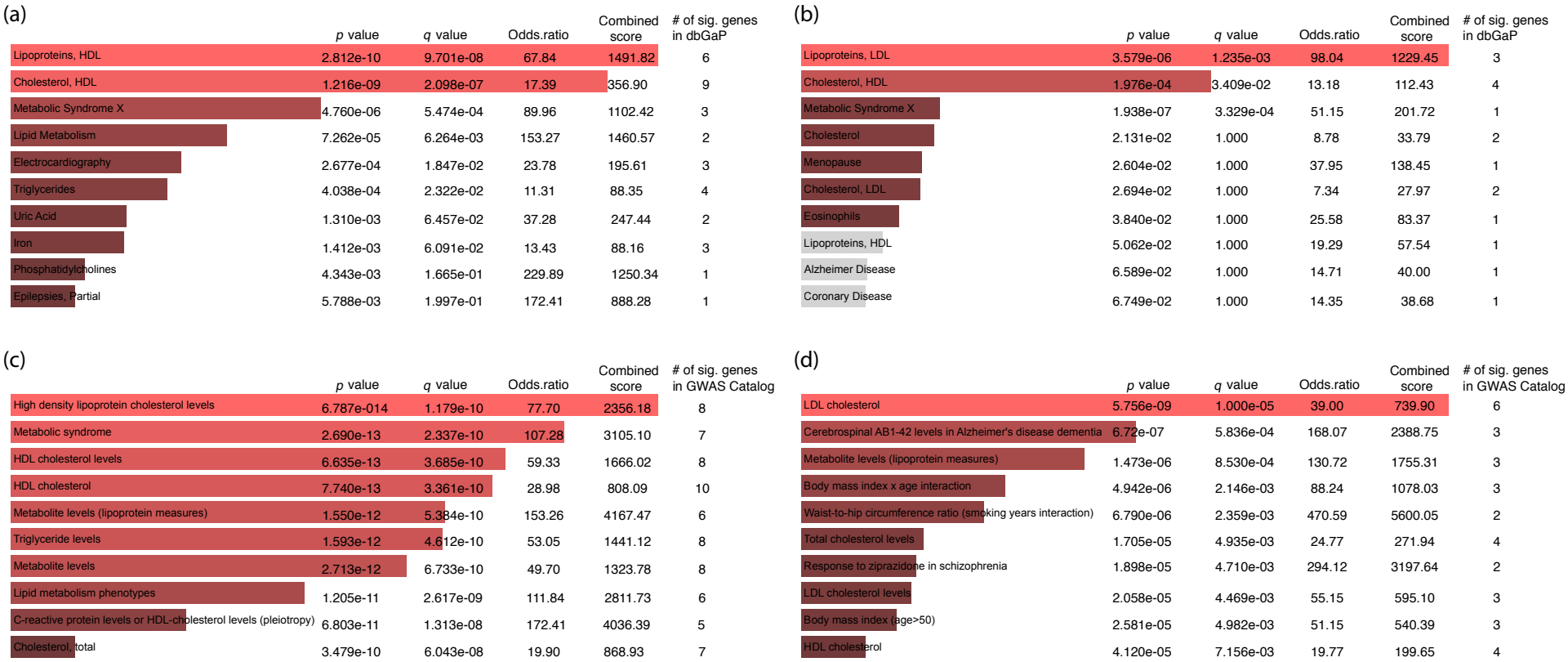


(a) High-Density Lipoprotein (HDL)



(b) Low-Density Lipoprotein (LDL)

Supplementary Figure 28. Manhattan plot of variant-level fine mapping results for high-density and low-density lipoprotein (HDL and LDL, respectively) traits in ten thousand randomly sampled individuals of European ancestry from the UK Biobank [51]. Posterior inclusion probabilities (PIP) for the neural network weights are derived from the BANNs model fit on individual-level data and are plotted for each SNP against their genomic positions. Chromosomes are shown in alternating colors for clarity. The black dashed line is marked at 0.5 and represents the "median probability model" threshold [52]. SNPs with PIPs above that threshold are color coded based on their SNP-set annotation. Here, SNP-set annotations are based on gene boundaries defined by the NCBI's RefSeq database in the UCSC Genome Browser [50]. Unannotated SNPs located within the same genomic region were labeled as being within the "intergenic region" between two genes. These regions are labeled as *Gene1-Gene2* in the legend. Gene set enrichment analyses for these SNP-sets can be found in Supplementary Figure 29. Stars denote SNPs and SNP-sets that replicate findings from our analyses of HDL and LDL in the Framingham Heart Study (See Figure 4 in the main text).



Supplementary Figure 29. Gene set enrichment analyses using the significant SNP-sets identified by BANNs for high-density and low-density lipoprotein (HDL and LDL, respectively) traits in ten thousand randomly sampled individuals of European ancestry from the UK Biobank [51]. Here, SNP-set annotations are based on gene boundaries defined by the NCBI's RefSeq database in the UCSC Genome Browser [50]. Unannotated SNPs located within the same genomic region were labeled as being within the "intergenic region" between two genes. Posterior inclusion probabilities (PIP) for the input and hidden layer weights are derived by fitting the BANNs model on individual-level data. A SNP-set is considered significant if it has a PIP $\boldsymbol{\gamma}_w \geq 0.5$ (i.e., the "median probability model" threshold [52]). We take these significant SNP-sets and conduct "gene set enrichment analysis" using Enrichr [72, 73] to identify the categories they overrepresent in **(a, b)** the database of Genotypes and Phenotypes (dbGaP) and **(c, d)** the GWAS Catalog (2019). We highlight categories with Q-values (i.e., false discovery rates) below 0.05. Nearly all enriched categories are related with **(a, c)** HDL and **(b, d)** LDL, respectively. Note that in LDL, the BANNs framework again identifies the gene *APOB* as having a high PIP (replicating the finding in the Framingham Heart Study). There have been hypotheses connecting LDL to cognitive traits [74, 75], and *APOB* has been shown to be related to cerebrospinal fluid and memory [76–78]. Therefore, we argue that results in panel **(b)** are also relevant (a similar argument can be made for Supplementary Figure 27).

11 Supplementary Tables

		SNP-Level Approaches					
Trait Type	**Metric**	**BANN**	**BANN-SS**	**SuSiE (High)**	**SuSiE (Low)**	**CAVIAR**	**FINEMAP**
Sparse	**Power**	**0.614 (0.106)**	0.609 (0.106)	0.608 (0.087)	0.424 (0.103)	0.597 (0.167)	0.529 (0.156)
	FDR	0.211(0.092)	0.210 (0.104)	**0.207 (0.052)**	0.512 (0.088)	0.331 (0.155)	0.346 (0.048)
Polygenic	**Power**	0.119 (0.039)	**0.120 (0.053)**	0.103 (0.021)	0.072 (0.037)	0.098 (0.032)	0.101 (0.044)
	FDR	0.224 (0.098)	0.227 (0.107)	**0.217 (0.034)**	0.491 (0.051)	0.244 (0.086)	0.257 (0.103)

		SNP-Set Level Approaches						
Trait Type	**Metric**	**BANN**	**BANN-SS**	**RSS**	**PEGASUS**	**SKAT**	**MAGMA**	**GSEA**
Sparse	**Power**	0.608 (0.132)	0.602 (0.161)	0.609 (0.127)	0.672 (0.118)	0.543 (0.142)	**0.737 (0.123)**	0.431 (0.138)
	FDR	**0.052 (0.107)**	0.067 (0.098)	0.088 (0.103)	0.514 (0.137)	0.466 (0.156)	0.571 (0.126)	0.579 (0.149)
Polygenic	**Power**	0.078 (0.027)	0.081 (0.041)	0.073 (0.024)	0.148 (0.032)	0.112 (0.042)	**0.152 (0.039)**	0.081 (0.031)
	FDR	0.166 (0.151)	0.163 (0.126)	**0.065 (0.112)**	0.179 (0.098)	0.181 (0.101)	0.191 (0.018)	0.221 (0.141)

Supplementary Table 1. Comparing the empirical power and false discovery rates (FDR) of the BANNs framework against competing SNP and SNP-set mapping approaches in simulations. Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with only contributions from additive effects set (i.e., $\rho = 1$). We consider two different trait architectures: sparse where only 1% of SNP-sets are enriched for the trait; and polygenic where 10% of SNP-sets are enriched. We set the number of causal SNPs with non-zero effects to be 0.125% and 3% of all SNPs located within the enriched SNP-sets, respectively. **(Top)** Competing SNP-level mapping approaches include: CAVIAR [64], SuSiE [65], and FINEMAP [66]. The software for SuSiE requires an input ℓ which fixes the maximum number of causal SNPs in the model. We display results when this input number is high ($\ell = 3000$) and when this input number is low ($\ell = 10$). **(Bottom)** Competing SNP-set mapping approaches include: RSS [7], PEGASUS [67], GBJ [68], SKAT [69], GSEA [70], and MAGMA [71]. Results for the BANN, BANN-SS, and other Bayesian methods are evaluated based on the "median probability criterion" (i.e., PIPs for SNPs and SNP-sets greater than 0.5) [52]. Results for the frequentist approaches are based on Bonferroni-corrected thresholds for multiple hypothesis testing ($P = 0.05/36518 = 1.37 \times 10^{-6}$ at the SNP-level and $P = 0.05/2816 = 1.78 \times 10^{-5}$ at the SNP-set level, respectively). All results are based on 100 replicates and standard deviations of the estimates across runs are given in the parentheses. Approaches with the greatest power are bolded in purple, while methods with the lowest FDR is bolded in blue.

		SNP-Level Approaches					
Trait Type	**Metric**	**BANN**	**BANN-SS**	**SuSiE (High)**	**SuSiE (Low)**	**CAVIAR**	**FINEMAP**
Sparse	**Power**	**0.851 (0.096)**	0.812 (0.081)	0.803 (0.034)	0.631 (0.098)	0.774 (0.159)	0.722 (0.132)
	FDR	0.201 (0.027)	0.196 (0.029)	**0.185 (0.063)**	0.522 (0.106)	0.196 (0.093)	0.248 (0.083)
Polygenic	**Power**	**0.374 (0.071)**	0.369 (0.067)	0.296 (0.074)	0.198 (0.061)	0.319 (0.106)	0.332 (0.044)
	FDR	0.212 (0.018)	**0.198 (0.026)**	0.208 (0.022)	0.414 (0.042)	0.205 (0.031)	0.307 (0.109)

		SNP-Set Level Approaches						
Trait Type	**Metric**	**BANN**	**BANN-SS**	**RSS**	**PEGASUS**	**SKAT**	**MAGMA**	**GSEA**
Sparse	**Power**	**0.823 (0.108)**	0.821 (0.112)	0.783 (0.105)	0.815 (0.102)	0.757 (0.114)	0.821 (0.097)	0.627 (0.123)
	FDR	0.121 (0.087)	0.127 (0.091	**0.099 (0.096)**	0.713 (0.081)	0.692 (0.089)	0.742 (0.061)	0.581 (0.051)
Polygenic	**Power**	0.276 (0.083)	0.279 (0.097)	0.272 (0.029)	0.416 (0.076)	0.331 (0.038)	**0.451 (0.049)**	0.241 (0.022)
	FDR	0.171 (0.034)	0.166 (0.041)	**0.069 (0.040)**	0.309 (0.087)	0.282 (0.079)	0.383 (0.083)	0.322 (0.018)

Supplementary Table 2. Comparing the empirical power and false discovery rates (FDR) of the BANNs framework against competing SNP and SNP-set mapping approaches in simulations. Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.6$ with only contributions from additive effects set (i.e., $\rho = 1$). We consider two different trait architectures: sparse where only 1% of SNP-sets are enriched for the trait; and polygenic where 10% of SNP-sets are enriched. We set the number of causal SNPs with non-zero effects to be 0.125% and 3% of all SNPs located within the enriched SNP-sets, respectively. **(Top)** Competing SNP-level mapping approaches include: CAVIAR [64], SuSiE [65], and FINEMAP [66]. The software for SuSiE requires an input ℓ which fixes the maximum number of causal SNPs in the model. We display results when this input number is high ($\ell = 3000$) and when this input number is low ($\ell = 10$). **(Bottom)** Competing SNP-set mapping approaches include: RSS [7], PEGASUS [67], GBJ [68], SKAT [69], GSEA [70], and MAGMA [71]. Results for the BANN, BANN-SS, and other Bayesian methods are evaluated based on the "median probability criterion" (i.e., PIPs for SNPs and SNP-sets greater than 0.5) [52]. Results for the frequentist approaches are based on Bonferroni-corrected thresholds for multiple hypothesis testing ($P = 0.05/36518 = 1.37 \times 10^{-6}$ at the SNP-level and $P = 0.05/2816 = 1.78 \times 10^{-5}$ at the SNP-set level, respectively). All results are based on 100 replicates and standard deviations of the estimates across runs are given in the parentheses. Approaches with the greatest power are bolded in purple, while methods with the lowest FDR is bolded in blue.

		SNP-Level Approaches					
Trait Type	**Metric**	**BANN**	**BANN-SS**	**SuSiE (High)**	**SuSiE (Low)**	**CAVIAR**	**FINEMAP**
Sparse	**Power**	**0.604 (0.117)**	0.600 (0.098)	0.598 (0.133)	0.424 (0.097)	0.537 (0.152)	0.529 (0.146)
	FDR	0.211 (0.087)	0.223 (0.103)	**0.207 (0.067)**	0.512 (0.062)	0.331 (0.128)	0.346 (0.099)
Polygenic	**Power**	**0.119 (0.091)**	0.108 (0.088)	0.103 (0.046)	0.070 (0.017)	0.096 (0.026)	0.101 (0.056)
	FDR	0.224 (0.074)	**0.219 (0.091)**	0.217 (0.025)	0.491 (0.083)	0.244 (0.073)	0.257 (0.091)

		SNP-Set Level Approaches						
Trait Type	**Metric**	**BANN**	**BANN-SS**	**RSS**	**PEGASUS**	**SKAT**	**MAGMA**	**GSEA**
Sparse	**Power**	0.624 (0.126)	0.641 (0.131)	0.611 (0.114)	0.694 (0.122)	0.583 (0.123)	**0.760 (0.119)**	0.491 (0.123)
	FDR	0.312 (0.073)	**0.308 (0.081)**	0.325 (0.109)	0.793 (0.074)	0.755 (0.089)	0.807 (0.073)	0.812 (0.121)
Polygenic	**Power**	0.131 (0.062)	0.129 (0.053)	0.081 (0.027)	0.152 (0.023)	0.121 (0.021)	**0.164 (0.028)**	0.151 (0.034)
	FDR	**0.217 (0.114)**	0.231 (0.096)	0.259 (0.123)	0.541 (0.124)	0.542 (0.184)	0.551 (0.109)	0.623 (0.115)

Supplementary Table 3. Comparing the empirical power and false discovery rates (FDR) of the BANNs framework against competing SNP and SNP-set mapping approaches in simulations with population stratification. Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with only contributions from additive effects set (i.e., $\rho = 1$). We consider two different trait architectures: sparse where only 1% of SNP-sets are enriched for the trait; and polygenic where 10% of SNP-sets are enriched. We set the number of causal SNPs with non-zero effects to be 0.125% and 3% of all SNPs located within the enriched SNP-sets, respectively. In these simulations, traits were generated while also using the top ten principal components (PCs) of the genotype matrix as covariates. **(Top)** Competing SNP-level mapping approaches include: CAVIAR [64], SuSiE [65], and FINEMAP [66]. The software for SuSiE requires an input ℓ which fixes the maximum number of causal SNPs in the model. We display results when this input number is high ($\ell = 3000$) and when this input number is low ($\ell = 10$). **(Bottom)** Competing SNP-set mapping approaches include: RSS [7], PEGASUS [67], GBJ [68], SKAT [69], GSEA [70], and MAGMA [71]. Results for the BANN, BANN-SS, and other Bayesian methods are evaluated based on the "median probability criterion" (i.e., PIPs for SNPs and SNP-sets greater than 0.5) [52]. Results for the frequentist approaches are based on Bonferroni-corrected thresholds for multiple hypothesis testing ($P = 0.05/36518 = 1.37 \times 10^{-6}$ at the SNP-level and $P = 0.05/2816 = 1.78 \times 10^{-5}$ at the SNP-set level, respectively). All results are based on 100 replicates and standard deviations of the estimates across runs are given in the parentheses. Approaches with the greatest power are bolded in purple, while methods with the lowest FDR is bolded in blue.

		SNP-Level Approaches					
Trait Type	**Metric**	**BANN**	**BANN-SS**	**SuSiE (High)**	**SuSiE (Low)**	**CAVIAR**	**FINEMAP**
Sparse	**Power**	**0.817 (0.126)**	0.798 (0.117)	0.792 (0.092)	0.563 (0.104)	0.752 (0.134)	0.726 (0.128)
	FDR	**0.182 (0.038)**	0.191 (0.045)	0.346 (0.057)	0.467 (0.075)	0.337 (0.076)	0.382 (0.084)
Polygenic	**Power**	**0.348 (0.109)**	0.319 (0.094)	0.305 (0.081)	0.211 (0.039)	0.327 (0.093)	0.338 (0.053)
	FDR	0.239 (0.047)	**0.221 (0.038)**	0.224 (0.042)	0.385 (0.035)	0.309 (0.041)	0.325 (0.091)

		SNP-Set Level Approaches						
Trait Type	**Metric**	**BANN**	**BANN-SS**	**RSS**	**PEGASUS**	**SKAT**	**MAGMA**	**GSEA**
Sparse	**Power**	0.781 (0.109)	0.749 (0.117)	0.743 (0.105)	**0.814 (0.112)**	0.773 (0.127)	0.802 (0.091)	0.699 (0.118)
	FDR	**0.121 (0.104)**	0.124 (0.098)	0.312 (0.099)	0.827 (0.056)	0.805 (0.065)	0.833 (0.051)	0.841 (0.077)
Polygenic	**Power**	0.294 (0.042)	0.281 (0.053)	0.301 (0.034)	0.419 (0.047)	0.341 (0.038)	**0.465 (0.038)**	0.318 (0.078)
	FDR	0.166 (0.054)	**0.159 (0.071)**	0.178 (0.062)	0.452 (0.089)	0.418 (0.095)	0.471 (0.079)	0.516 (0.214)

Supplementary Table 4. Comparing the empirical power and false discovery rates (FDR) of the BANNs framework against competing SNP and SNP-set mapping approaches in simulations with population stratification. Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.6$ with only contributions from additive effects set (i.e., $\rho = 1$). We consider two different trait architectures: sparse where only 1% of SNP-sets are enriched for the trait; and polygenic where 10% of SNP-sets are enriched. We set the number of causal SNPs with non-zero effects to be 0.125% and 3% of all SNPs located within the enriched SNP-sets, respectively. In these simulations, traits were generated while also using the top ten principal components (PCs) of the genotype matrix as covariates. **(Top)** Competing SNP-level mapping approaches include: CAVIAR [64], SuSiE [65], and FINEMAP [66]. The software for SuSiE requires an input ℓ which fixes the maximum number of causal SNPs in the model. We display results when this input number is high ($\ell = 3000$) and when this input number is low ($\ell = 10$). **(Bottom)** Competing SNP-set mapping approaches include: RSS [7], PEGASUS [67], GBJ [68], SKAT [69], GSEA [70], and MAGMA [71]. Results for the BANN, BANN-SS, and other Bayesian methods are evaluated based on the "median probability criterion" (i.e., PIPs for SNPs and SNP-sets greater than 0.5) [52]. Results for the frequentist approaches are based on Bonferroni-corrected thresholds for multiple hypothesis testing ($P = 0.05/36518 = 1.37 \times 10^{-6}$ at the SNP-level and $P = 0.05/2816 = 1.78 \times 10^{-5}$ at the SNP-set level, respectively). All results are based on 100 replicates and standard deviations of the estimates across runs are given in the parentheses. Approaches with the greatest power are bolded in purple, while methods with the lowest FDR is bolded in blue.

		SNP-Level Approaches					
Trait Type	**Metric**	**BANN**	**BANN-SS**	**SuSiE (High)**	**SuSiE (Low)**	**CAVIAR**	**FINEMAP**
Sparse	**Power**	**0.522 (0.122)**	0.406 (0.094)	0.402 (0.082)	0.312 (0.077)	0.309 (0.093)	0.307 (0.087)
	FDR	0.296 (0.113)	0.311 (0.102)	**0.187 (0.061)**	0.421 (0.089)	0.207 (0.099)	0.214 (0.068)
Polygenic	**Power**	**0.104 (0.058)**	0.088 (0.033)	0.094 (0.042)	0.053 (0.032)	0.081 (0.027)	0.092 (0.031)
	FDR	0.217 (0.061)	**0.186 (0.042)**	0.193 (0.063)	0.398 (0.092)	0.194 (0.054)	0.203 (0.059)

		SNP-Set Level Approaches						
Trait Type	**Metric**	**BANN**	**BANN-SS**	**RSS**	**PEGASUS**	**SKAT**	**MAGMA**	**GSEA**
Sparse	**Power**	**0.528 (0.113)**	0.421 (0.098)	0.476 (0.092)	0.504 (0.109)	0.426 (0.113)	0.443 (0.128)	0.378 (0.099)
	FDR	**0.073 (0.024)**	0.095 (0.032)	0.079 (0.023)	0.201 (0.098)	0.231 (0.106)	0.312 (0.129)	0.347 (0.127)
Polygenic	**Power**	0.069 (0.029)	0.051 (0.041)	0.057 (0.024)	0.056 (0.032)	0.112 (0.042)	**0.152 (0.039)**	0.081 (0.031)
	FDR	0.166 (0.151)	0.163 (0.126)	**0.065 (0.112)**	0.179 (0.098)	0.181 (0.101)	0.191 (0.018)	0.221 (0.141)

Supplementary Table 5. Comparing the empirical power and false discovery rates (FDR) of the BANNs framework against competing SNP and SNP-set mapping approaches in simulations. Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with contributions from both additive and epistatic effects set (i.e., $\rho = 0.5$). We consider two different trait architectures: sparse where only 1% of SNP-sets are enriched for the trait; and polygenic where 10% of SNP-sets are enriched. We set the number of causal SNPs with non-zero effects to be 0.125% and 3% of all SNPs located within the enriched SNP-sets, respectively. **(Top)** Competing SNP-level mapping approaches include: CAVIAR [64], SuSiE [65], and FINEMAP [66]. The software for SuSiE requires an input ℓ which fixes the maximum number of causal SNPs in the model. We display results when this input number is high ($\ell = 3000$) and when this input number is low ($\ell = 10$). **(Bottom)** Competing SNP-set mapping approaches include: RSS [7], PEGASUS [67], GBJ [68], SKAT [69], GSEA [70], and MAGMA [71]. Results for the BANN, BANN-SS, and other Bayesian methods are evaluated based on the "median probability criterion" (i.e., PIPs for SNPs and SNP-sets greater than 0.5) [52]. Results for the frequentist approaches are based on Bonferroni-corrected thresholds for multiple hypothesis testing ($P = 0.05/36518 = 1.37 \times 10^{-6}$ at the SNP-level and $P = 0.05/2816 = 1.78 \times 10^{-5}$ at the SNP-set level, respectively). All results are based on 100 replicates and standard deviations of the estimates across runs are given in the parentheses. Approaches with the greatest power are bolded in purple, while methods with the lowest FDR is bolded in blue.

		SNP-Level Approaches					
Trait Type	**Metric**	**BANN**	**BANN-SS**	**SuSiE (High)**	**SuSiE (Low)**	**CAVIAR**	**FINEMAP**
Sparse	**Power**	**0.851 (0.096)**	0.812 (0.081)	0.803 (0.034)	0.631 (0.098)	0.774 (0.159)	0.722 (0.132)
	FDR	0.201 (0.027)	0.196 (0.029)	**0.185 (0.063)**	0.522 (0.106)	0.196 (0.093)	0.248 (0.083)
Polygenic	**Power**	**0.374 (0.071)**	0.369 (0.067)	0.296 (0.074)	0.198 (0.061)	0.319 (0.106)	0.332 (0.044)
	FDR	0.212 (0.018)	**0.198 (0.026)**	0.208 (0.022)	0.414 (0.042)	0.205 (0.031)	0.307 (0.109)

		SNP-Set Level Approaches						
Trait Type	**Metric**	**BANN**	**BANN-SS**	**RSS**	**PEGASUS**	**SKAT**	**MAGMA**	**GSEA**
Sparse	**Power**	**0.823 (0.108)**	0.821 (0.112)	0.783 (0.105)	0.815 (0.102)	0.757 (0.114)	0.821 (0.097)	0.627 (0.123)
	FDR	0.121 (0.087)	0.127 (0.091)	**0.099 (0.096)**	0.713 (0.081)	0.692 (0.089)	0.742 (0.061)	0.581 (0.051)
Polygenic	**Power**	0.276 (0.083)	0.279 (0.097)	0.272 (0.029)	0.416 (0.076)	0.331 (0.038)	**0.451 (0.049)**	0.241 (0.022)
	FDR	0.171 (0.034)	0.166 (0.041)	**0.069 (0.040)**	0.309 (0.087)	0.282 (0.079)	0.383 (0.083)	0.322 (0.018)

Supplementary Table 6. Comparing the empirical power and false discovery rates (FDR) of the BANNs framework against competing SNP and SNP-set mapping approaches in simulations. Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.6$ with contributions from both additive and epistatic effects set (i.e., $\rho = 0.5$). We consider two different trait architectures: sparse where only 1% of SNP-sets are enriched for the trait; and polygenic where 10% of SNP-sets are enriched. We set the number of causal SNPs with non-zero effects to be 0.125% and 3% of all SNPs located within the enriched SNP-sets, respectively. **(Top)** Competing SNP-level mapping approaches include: CAVIAR [64], SuSiE [65], and FINEMAP [66]. The software for SuSiE requires an input ℓ which fixes the maximum number of causal SNPs in the model. We display results when this input number is high ($\ell = 3000$) and when this input number is low ($\ell = 10$). **(Bottom)** Competing SNP-set mapping approaches include: RSS [7], PEGASUS [67], GBJ [68], SKAT [69], GSEA [70], and MAGMA [71]. Results for the BANN, BANN-SS, and other Bayesian methods are evaluated based on the "median probability criterion" (i.e., PIPs for SNPs and SNP-sets greater than 0.5) [52]. Results for the frequentist approaches are based on Bonferroni-corrected thresholds for multiple hypothesis testing ($P = 0.05/36518 = 1.37 \times 10^{-6}$ at the SNP-level and $P = 0.05/2816 = 1.78 \times 10^{-5}$ at the SNP-set level, respectively). All results are based on 100 replicates and standard deviations of the estimates across runs are given in the parentheses. Approaches with the greatest power are bolded in purple, while methods with the lowest FDR is bolded in blue.

		SNP-Level Approaches					
Trait Type	**Metric**	**BANN**	**BANN-SS**	**SuSiE (High)**	**SuSiE (Low)**	**CAVIAR**	**FINEMAP**
Sparse	**Power**	**0.499 (0.094)**	0.372 (0.061)	0.371 (0.082)	0.263 (0.060)	0.332 (0.094)	0.327 (0.091)
	FDR	0.243 (0.100)	0.245 (0.113)	**0.228 (0.074)**	0.436 (0.126)	0.364 (0.142)	0.381 (0.109)
Polygenic	**Power**	**0.105 (0.042)**	0.087 (0.034)	0.094 (0.032)	0.063 (0.016)	0.095 (0.015)	0.083 (0.035)
	FDR	**0.203 (0.064)**	0.241 (0.101)	0.238 (0.037)	0.449 (0.091)	0.268 (0.081)	0.283 (0.101)

		SNP-Set Level Approaches						
Trait Type	**Metric**	**BANN**	**BANN-SS**	**RSS**	**PEGASUS**	**SKAT**	**MAGMA**	**GSEA**
Sparse	**Power**	**0.483 (0.102)**	0.378 (0.082)	0.397 (0.105)	0.403 (0.076)	0.321 (0.077)	0.472 (0.074)	0.304 (0.076)
	FDR	0.104 (0.04)	0.138 (0.087)	**0.088 (0.099)**	0.372 (0.082)	0.430 (0.098)	0.672 (0.087)	0.593 (0.0133)
Polygenic	**Power**	**0.103 (0.049)**	0.079 (0.032)	0.082 (0.017)	0.094 (0.015)	0.071 (0.013)	0.102 (0.018)	0.094 (0.021)
	FDR	**0.249 (0.103)**	0.254 (0.106)	0.281 (0.135)	0.472 (0.134)	0.491 (0.102)	0.606 (0.129)	0.653 (0.137)

Supplementary Table 7. Comparing the empirical power and false discovery rates (FDR) of the BANNs framework against competing SNP and SNP-set mapping approaches in simulations with population stratification. Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with contributions from both additive and epistatic effects set (i.e., $\rho = 0.5$). We consider two different trait architectures: sparse where only 1% of SNP-sets are enriched for the trait; and polygenic where 10% of SNP-sets are enriched. We set the number of causal SNPs with non-zero effects to be 0.125% and 3% of all SNPs located within the enriched SNP-sets, respectively. In these simulations, traits were generated while also using the top ten principal components (PCs) of the genotype matrix as covariates. **(Top)** Competing SNP-level mapping approaches include: CAVIAR [64], SuSiE [65], and FINEMAP [66]. The software for SuSiE requires an input ℓ which fixes the maximum number of causal SNPs in the model. We display results when this input number is high ($\ell = 3000$) and when this input number is low ($\ell = 10$). **(Bottom)** Competing SNP-set mapping approaches include: RSS [7], PEGASUS [67], GBJ [68], SKAT [69], GSEA [70], and MAGMA [71]. Results for the BANN, BANN-SS, and other Bayesian methods are evaluated based on the "median probability criterion" (i.e., PIPs for SNPs and SNP-sets greater than 0.5) [52]. Results for the frequentist approaches are based on Bonferroni-corrected thresholds for multiple hypothesis testing ($P = 0.05/36518 = 1.37 \times 10^{-6}$ at the SNP-level and $P = 0.05/2816 = 1.78 \times 10^{-5}$ at the SNP-set level, respectively). All results are based on 100 replicates and standard deviations of the estimates across runs are given in the parentheses. Approaches with the greatest power are bolded in purple, while methods with the lowest FDR is bolded in blue.

		SNP-Level Approaches					
Trait Type	**Metric**	**BANN**	**BANN-SS**	**SuSiE (High)**	**SuSiE (Low)**	**CAVIAR**	**FINEMAP**
Sparse	**Power**	**0.817 (0.126)**	0.798 (0.117)	0.792 (0.092)	0.563 (0.104)	0.752 (0.134)	0.726 (0.128)
	FDR	**0.182 (0.038)**	0.191 (0.045)	0.346 (0.057)	0.467 (0.075)	0.337 (0.076)	0.382 (0.084)
Polygenic	**Power**	**0.348 (0.109)**	0.319 (0.094)	0.305 (0.081)	0.211 (0.039)	0.327 (0.093)	0.338 (0.053)
	FDR	0.239 (0.047)	**0.221 (0.038)**	0.224 (0.042)	0.385 (0.035)	0.309 (0.041)	0.325 (0.091)

		SNP-Set Level Approaches						
Trait Type	**Metric**	**BANN**	**BANN-SS**	**RSS**	**PEGASUS**	**SKAT**	**MAGMA**	**GSEA**
Sparse	**Power**	0.781 (0.109)	0.749 (0.117)	0.743 (0.105)	**0.814 (0.112)**	0.773 (0.127)	0.802 (0.091)	0.699 (0.118)
	FDR	**0.121 (0.104)**	0.124 (0.098)	0.312 (0.099)	0.827 (0.056)	0.805 (0.065)	0.833 (0.051)	0.841 (0.077)
Polygenic	**Power**	0.294 (0.042)	0.281 (0.053)	0.301 (0.034)	0.419 (0.047)	0.341 (0.038)	**0.465 (0.038)**	0.318 (0.078)
	FDR	0.166 (0.054)	**0.159 (0.071)**	0.178 (0.062)	0.452 (0.089)	0.418 (0.095)	0.471 (0.079)	0.516 (0.214)

Supplementary Table 8. Comparing the empirical power and false discovery rates (FDR) of the BANNs framework against competing SNP and SNP-set mapping approaches in simulations with population stratification. Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.6$ with contributions from both additive and epistatic effects set (i.e., $\rho = 0.5$). We consider two different trait architectures: sparse where only 1% of SNP-sets are enriched for the trait; and polygenic where 10% of SNP-sets are enriched. We set the number of causal SNPs with non-zero effects to be 0.125% and 3% of all SNPs located within the enriched SNP-sets, respectively. In these simulations, traits were generated while also using the top ten principal components (PCs) of the genotype matrix as covariates. **(Top)** Competing SNP-level mapping approaches include: CAVIAR [64], SuSiE [65], and FINEMAP [66]. The software for SuSiE requires an input ℓ which fixes the maximum number of causal SNPs in the model. We display results when this input number is high ($\ell = 3000$) and when this input number is low ($\ell = 10$). **(Bottom)** Competing SNP-set mapping approaches include: RSS [7], PEGASUS [67], GBJ [68], SKAT [69], GSEA [70], and MAGMA [71]. Results for the BANN, BANN-SS, and other Bayesian methods are evaluated based on the "median probability criterion" (i.e., PIPs for SNPs and SNP-sets greater than 0.5) [52]. Results for the frequentist approaches are based on Bonferroni-corrected thresholds for multiple hypothesis testing ($P = 0.05/36518 = 1.37 \times 10^{-6}$ at the SNP-level and $P = 0.05/2816 = 1.78 \times 10^{-5}$ at the SNP-set level, respectively). All results are based on 100 replicates and standard deviations of the estimates across runs are given in the parentheses. Approaches with the greatest power are bolded in purple, while methods with the lowest FDR is bolded in blue.

Simulation Parameters		Average Run Time (seconds)				
SNPs	**Samples Sizes**	**BANN**	**SuSiE (low)**	**SuSiE (high)**	**CAVIAR**	**FINEMAP**
2500	1000	3.34	1.89	4.22	8.21	56.99
	2000	6.71	2.87	8.72	8.21	56.99
	4000	10.82	8.42	13.63	8.21	56.99
5000	1000	7.42	2.49	7.12	31.48	102.58
	2000	13.21	5.04	21.84	31.48	102.58
	4000	21.34	9.45	32.81	31.48	102.58
10000	1000	31.39	3.52	52.24	118.98	145.51
	2000	127.18	10.22	159.97	118.98	145.51
	4000	318.81	22.62	754.63	118.98	145.51

Supplementary Table 9. Computational time for running Bayesian annotated neural networks (BANNs) and other SNP-level association mapping approaches, as a function of the total number SNPs analyzed and the number of samples in the data. Methods compared include: BANNs, CAVIAR [64], SuSiE [65], and FINEMAP [66]. Each table entry represents the average computation time (in seconds) it takes each approach to analyze a dataset of the size indicated. Run times were measured on an Intel i5-8259U CPU with base frequency of 2.30GHz, turbo frequency of 3.80GHz, and memory 16GB 2133 MHz LPDDR3. Here, we used 4 cores for parallelization when applicable. The software for SuSiE requires an input ℓ which fixes the maximum number of causal SNPs in the model. We display results when this input parameter is high ($\ell = 3000$) and when this input parameter is low ($\ell = 10$). Note that we implemented BANNs using the `Python 3` version of the software, and the timing for its variational algorithm includes inference on both SNPs and SNP-sets. CAVIAR and FINEMAP are set up to work with GWA summary statistics, so their inputs (and timing) are the same irrespective of the sample size.

Simulation Parameters		Average Run Time (seconds)						
SNP-Sets	**SNPs per SNP-set**	**BANN**	**RSS**	**PEGASUS**	**GBJ**	**SKAT**	**MAGMA**	**GSEA**
250	10	12.58	13.12	2.41	2.68	2.13	0.03	2.48
	20	44.32	58.21	2.13	5.18	3.82	0.08	4.68
	40	189.44	224.62	2.22	9.64	6.47	0.18	8.51
500	10	48.92	59.31	5.11	5.37	5.23	0.09	5.31
	20	223.14	244.07	5.02	11.26	9.22	0.21	11.12
	40	965.48	1026.12	5.72	27.91	14.84	0.24	20.36
1000	10	194.62	249.57	8.67	12.27	11.31	0.72	11.41
	20	1213.19	2176.33	8.93	27.62	18.16	1.48	24.93
	40	6823.31	14495.72	10.21	61.37	30.83	4.26	60.82

Supplementary Table 10. Computational time for running Bayesian annotated neural networks (BANNs) and other SNP-set level enrichment approaches, as a function of the total number SNP-sets analyzed and the number of SNPs within each SNP-set. Methods compared include: BANNs, RSS [7], PEGASUS [67], GBJ [68], SKAT [69], GSEA [70], and MAGMA [71]. Here, we simulated 10 datasets for each pair of parameter values (number of SNP-sets analyzed and number of SNPs within each SNP-set). Sample size was held constant at $n = 10{,}000$ individuals. Each table entry represents the average computation time (in seconds) it takes each approach to analyze a dataset of the size indicated. Run times were measured on an Intel i5-8259U CPU with base frequency of 2.30GHz, turbo frequency of 3.80GHz, and memory 16GB 2133 MHz LPDDR3. Here, we used 4 cores for parallelization when applicable. Note that PEGASUS, GBJ, SKAT, and MAGMA are score-based methods and, thus, are expected to take the least amount of time to run. Both the BANNs framework and RSS are regression-based methods. The increased computational burden of these approaches results from its need to do (approximate) Bayesian posterior inference; however, the sparse and partially connected architecture of the BANNs model allows it to scale more favorably for larger dimensional datasets. Note that we implemented BANNs using the `Python 3` version of the software, and the timing for its variational algorithm includes inference on both SNPs and SNP-sets.

Supplementary Table 11. SNP and SNP-set results for body mass index (BMI) in the heterogenous stock of mice from the Wellcome Trust Centre for Human Genetics. We analyze $J \approx 10{,}000$ SNPs and $G = 1,925$ SNP-sets from $N = 1{,}814$ mice—with specific numbers varying slightly depending on the quality control procedure for each phenotype (Supplementary Note, Section 5). Here, SNP-set annotations are based on gene boundaries defined by the Mouse Genome Informatics database (see URLs listed in the main text). Unannotated SNPs located within the same genomic region were labeled as being within the "intergenic region" between two genes. This file gives the posterior inclusion probabilities (PIPs) for the input and hidden layer neural network weights after fitting the BANNs model on the individual-level data. We assess significance for both SNPs and SNP-sets according to the "median probability model" threshold [52] (i.e., PIP ≥ 0.5). Page #1 provides the variant-level fine mapping results with columns corresponding to: (1) chromosome; (2) SNP ID; (3) chromosomal position in base-pair (bp) coordinates; and (4) SNP PIP. Page #2 provides the SNP-set level enrichment results with columns corresponding to: (1) chromosome; (2) SNP-set ID; (3-4) the starting and ending position of the SNP-set chromosomal boundaries; (5) SNP-set PIP; (6) the number of SNPs that have been annotated within each SNP-set; (7) the "top" associated SNP within each SNP-set; and (8) the PIP of each top SNP. (XLSX)

Supplementary Table 12. SNP and SNP-set results for body weight in the heterogenous stock of mice from the Wellcome Trust Centre for Human Genetics. We analyze $J \approx 10{,}000$ SNPs and $G = 1,925$ SNP-sets from $N = 1{,}814$ mice—with specific numbers varying slightly depending on the quality control procedure for each phenotype (Supplementary Note, Section 5). Here, SNP-set annotations are based on gene boundaries defined by the Mouse Genome Informatics database (see URLs in the main text). Unannotated SNPs located within the same genomic region were labeled as being within the "intergenic region" between two genes. This file gives the posterior inclusion probabilities (PIPs) for the input and hidden layer neural network weights after fitting the BANNs model on the individual-level data. We assess significance for both SNPs and SNP-sets according to the "median probability model" threshold [52] (i.e., PIP ≥ 0.5). Page #1 provides the variant-level fine mapping results with columns corresponding to: (1) chromosome; (2) SNP ID; (3) chromosomal position in base-pair (bp) coordinates; and (4) SNP PIP. Page #2 provides the SNP-set level enrichment results with columns corresponding to: (1) chromosome; (2) SNP-set ID; (3-4) the starting and ending position of the SNP-set chromosomal boundaries; (5) SNP-set PIP; (6) the number of SNPs that have been annotated within each SNP-set; (7) the "top" associated SNP within each SNP-set; and (8) the PIP of each top SNP. (XLSX)

Supplementary Table 13. SNP and SNP-set results for percentage of CD8+ cells in the heterogenous stock of mice from the Wellcome Trust Centre for Human Genetics. We analyze $J \approx 10{,}000$ SNPs and $G = 1,925$ SNP-sets from $N = 1{,}814$ mice—with specific numbers varying slightly depending on the quality control procedure for each phenotype (Supplementary Note, Section 5). Here, SNP-set annotations are based on gene boundaries defined by the Mouse Genome Informatics database (see URLs listed in the main text). Unannotated SNPs located within the same genomic region were labeled as being within the "intergenic region" between two genes. This file gives the posterior inclusion probabilities (PIPs) for the input and hidden layer neural network weights after fitting the BANNs model on the individual-level data. We assess significance for both SNPs and SNP-sets according to the "median probability model" threshold [52] (i.e., PIP ≥ 0.5). Page #1 provides the variant-level fine mapping results with columns corresponding to: (1) chromosome; (2) SNP ID; (3) chromosomal position in base-pair (bp) coordinates; and (4) SNP PIP. Page #2 provides the SNP-set level enrichment results with columns corresponding to: (1) chromosome; (2) SNP-set ID; (3-4) the starting and ending position of the SNP-set chromosomal boundaries; (5) SNP-set PIP; (6) the number of SNPs that have been annotated within each SNP-set; (7) the "top" associated SNP within each SNP-set; and (8) the PIP of each top SNP. (XLSX)

Supplementary Table 14. SNP and SNP-set results for high-density lipoprotein (HDL) cholesterol in the heterogenous stock of mice from the Wellcome Trust Centre for Human Genetics. We analyze $J \approx 10{,}000$ SNPs and $G = 1,925$ SNP-sets from $N = 1{,}814$ mice—with specific numbers varying slightly depending on the quality control procedure for each phenotype (Supplementary Note, Section 5). Here, SNP-set annotations are based on gene boundaries defined by the Mouse Genome Informatics database (see URLs listed in the main text). Unannotated SNPs located within the same genomic region were labeled as being within the "intergenic region" between two genes. This file gives the posterior inclusion probabilities (PIPs) for the input and hidden layer neural network weights after fitting the BANNs model on the individual-level data. We assess significance for both SNPs and SNP-sets according to the "median probability model" threshold [52] (i.e., PIP ≥ 0.5). Page #1 provides the variant-level fine mapping results with columns corresponding to: (1) chromosome; (2) SNP ID; (3) chromosomal position in base-pair (bp) coordinates; and (4) SNP PIP. Page #2 provides the SNP-set level enrichment results with columns corresponding to: (1) chromosome; (2) SNP-set ID; (3-4) the starting and ending position of the SNP-set chromosomal boundaries; (5) SNP-set PIP; (6) the number of SNPs that have been annotated within each SNP-set; (7) the "top" associated SNP within each SNP-set; and (8) the PIP of each top SNP. (XLSX)

Supplementary Table 15. SNP and SNP-set results for low-density lipoprotein (LDL) cholesterol in the heterogenous stock of mice from the Wellcome Trust Centre for Human Genetics. We analyze $J \approx 10{,}000$ SNPs and $G = 1,925$ SNP-sets from $N = 1{,}814$ mice—with specific numbers varying slightly depending on the quality control procedure for each phenotype (Supplementary Note, Section 5). Here, SNP-set annotations are based on gene boundaries defined by the Mouse Genome Informatics database (see URLs listed in the main text). Unannotated SNPs located within the same genomic region were labeled as being within the "intergenic region" between two genes. This file gives the posterior inclusion probabilities (PIPs) for the input and hidden layer neural network weights after fitting the BANNs model on the individual-level data. We assess significance for both SNPs and SNP-sets according to the "median probability model" threshold [52] (i.e., PIP ≥ 0.5). Page #1 provides the variant-level fine mapping results with columns corresponding to: (1) chromosome; (2) SNP ID; (3) chromosomal position in base-pair (bp) coordinates; and (4) SNP PIP. Page #2 provides the SNP-set level enrichment results with columns corresponding to: (1) chromosome; (2) SNP-set ID; (3-4) the starting and ending position of the SNP-set chromosomal boundaries; (5) SNP-set PIP; (6) the number of SNPs that have been annotated within each SNP-set; (7) the "top" associated SNP within each SNP-set; and (8) the PIP of each top SNP. (XLSX)

Supplementary Table 16. SNP and SNP-set results for mean corpuscular hemoglobin (MCH) in the heterogenous stock of mice from the Wellcome Trust Centre for Human Genetics. We analyze $J \approx 10{,}000$ SNPs and $G = 1,925$ SNP-sets from $N = 1{,}814$ mice—with specific numbers varying slightly depending on the quality control procedure for each phenotype (Supplementary Note, Section 5). Here, SNP-set annotations are based on gene boundaries defined by the Mouse Genome Informatics database (see URLs listed in the main text). Unannotated SNPs located within the same genomic region were labeled as being within the "intergenic region" between two genes. This file gives the posterior inclusion probabilities (PIPs) for the input and hidden layer neural network weights after fitting the BANNs model on the individual-level data. We assess significance for both SNPs and SNP-sets according to the "median probability model" threshold [52] (i.e., PIP ≥ 0.5). Page #1 provides the variant-level fine mapping results with columns corresponding to: (1) chromosome; (2) SNP ID; (3) chromosomal position in base-pair (bp) coordinates; and (4) SNP PIP. Page #2 provides the SNP-set level enrichment results with columns corresponding to: (1) chromosome; (2) SNP-set ID; (3-4) the starting and ending position of the SNP-set chromosomal boundaries; (5) SNP-set PIP; (6) the number of SNPs that have been annotated within each SNP-set; (7) the "top" associated SNP within each SNP-set; and (8) the PIP of each top SNP. (XLSX)

Supplementary Table 17. SNP and SNP-set results for high-density lipoprotein (HDL) cholesterol in individuals assayed within the Framingham Heart Study. We analyze $J =$ 394,174 SNPs and $G = 18,364$ SNP-sets from $N =$ 6,950 people. Here, SNP-set annotations are based on gene boundaries defined by the NCBI's RefSeq database in the UCSC Genome Browser [50]. Unannotated SNPs located within the same genomic region were labeled as being within the "intergenic region" between two genes. This file gives the posterior inclusion probabilities (PIPs) for the input and hidden layer neural network weights after fitting the BANNs model on the individual-level data. We assess significance for both SNPs and SNP-sets according to the "median probability model" threshold [52] (i.e., PIP ≥ 0.5). Page #1 provides the variant-level fine mapping results with columns corresponding to: (1) chromosome; (2) SNP ID; (3) chromosomal position in base-pair (bp) coordinates; and (4) SNP PIP. Page #2 provides the SNP-set level enrichment results with columns corresponding to: (1) chromosome; (2) SNP-set ID; (3-4) the starting and ending position of the SNP-set chromosomal boundaries; (5) SNP-set PIP; (6) the number of SNPs that have been annotated within each SNP-set; (7) the "top" associated SNP within each SNP-set; and (8) the PIP of each top SNP. (XLSX)

Supplementary Table 18. SNP and SNP-set results for low-density lipoprotein (LDL) cholesterol in individuals assayed within the Framingham Heart Study. We analyze $J =$ 394,174 SNPs and $G = 18,364$ SNP-sets from $N =$ 6,950 people. Here, SNP-set annotations are based on gene boundaries defined by the NCBI's RefSeq database in the UCSC Genome Browser [50]. Unannotated SNPs located within the same genomic region were labeled as being within the "intergenic region" between two genes. This file gives the posterior inclusion probabilities (PIPs) for the input and hidden layer neural network weights after fitting the BANNs model on the individual-level data. We assess significance for both SNPs and SNP-sets according to the "median probability model" threshold [52] (i.e., PIP ≥ 0.5). Page #1 provides the variant-level fine mapping results with columns corresponding to: (1) chromosome; (2) SNP ID; (3) chromosomal position in base-pair (bp) coordinates; and (4) SNP PIP. Page #2 provides the SNP-set level enrichment results with columns corresponding to: (1) chromosome; (2) SNP-set ID; (3-4) the starting and ending position of the SNP-set chromosomal boundaries; (5) SNP-set PIP; (6) the number of SNPs that have been annotated within each SNP-set; (7) the "top" associated SNP within each SNP-set; and (8) the PIP of each top SNP. (XLSX)

Supplementary Table 19. SNP and SNP-set results for high-density lipoprotein (HDL) cholesterol in ten thousand randomly sampled individuals of European ancestry from the UK Biobank. We analyze the same $J =$ 394,174 SNPs and $G = 18,364$ SNP-sets used in the Framingham Heart Study analyses. Here, SNP-set annotations are based on gene boundaries defined by the NCBI's RefSeq database in the UCSC Genome Browser [50]. Unannotated SNPs located within the same genomic region were labeled as being within the "intergenic region" between two genes. This file gives the posterior inclusion probabilities (PIPs) for the input and hidden layer neural network weights after fitting the BANNs model on the individual-level data. We assess significance for both SNPs and SNP-sets according to the "median probability model" threshold [52] (i.e., PIP ≥ 0.5). Page #1 provides the variant-level fine mapping results with columns corresponding to: (1) chromosome; (2) SNP ID; (3) chromosomal position in base-pair (bp) coordinates; and (4) SNP PIP. Page #2 provides the SNP-set level enrichment results with columns corresponding to: (1) chromosome; (2) SNP-set ID; (3-4) the starting and ending position of the SNP-set chromosomal boundaries; (5) SNP-set PIP; (6) the number of SNPs that have been annotated within each SNP-set; (7) the "top" associated SNP within each SNP-set; and (8) the PIP of each top SNP. (XLSX)

Supplementary Table 20. SNP and SNP-set results for low-density lipoprotein (LDL) cholesterol in ten thousand randomly sampled individuals of European ancestry from the UK Biobank. We analyze the same $J = 394{,}174$ SNPs and $G = 18,364$ SNP-sets used in the Framingham Heart Study analyses. Here, SNP-set annotations are based on gene boundaries defined by the NCBI's RefSeq database in the UCSC Genome Browser [50]. Unannotated SNPs located within the same genomic region were labeled as being within the "intergenic region" between two genes. This file gives the posterior inclusion probabilities (PIPs) for the input and hidden layer neural network weights after fitting the BANNs model on the individual-level data. We assess significance for both SNPs and SNP-sets according to the "median probability model" threshold [52] (i.e., PIP ≥ 0.5). Page #1 provides the variant-level fine mapping results with columns corresponding to: (1) chromosome; (2) SNP ID; (3) chromosomal position in base-pair (bp) coordinates; and (4) SNP PIP. Page #2 provides the SNP-set level enrichment results with columns corresponding to: (1) chromosome; (2) SNP-set ID; (3-4) the starting and ending position of the SNP-set chromosomal boundaries; (5) SNP-set PIP; (6) the number of SNPs that have been annotated within each SNP-set; (7) the "top" associated SNP within each SNP-set; and (8) the PIP of each top SNP. (XLSX)

References

1. Xu B, Wang N, Chen T, Li M. Empirical evaluation of rectified activations in convolutional network; 2015. ArXiv.

2. Wang L, Zhang B, Wolfinger RD, Chen X. An integrated approach for the analysis of biological pathways using mixed models. PLoS Genet. 2008;4(7):e1000115. Available from: https://doi.org/10.1371/journal.pgen.1000115.

3. Califano A, Butte AJ, Friend S, Ideker T, Schadt E. Leveraging models of cell regulation and GWAS data in integrative network-based association studies. Nat Genet. 2012;44(8):841–847. Available from: https://doi.org/10.1038/ng.2355.

4. Carbonetto P, Stephens M. Integrated enrichment analysis of variants and pathways in genome-wide association studies indicates central role for IL-2 signaling genes in type 1 diabetes, and cytokine signaling genes in Crohn's disease. PLoS Genet. 2013;9(10):e1003770. Available from: https://doi.org/10.1371/journal.pgen.1003770.

5. Yang J, Fritsche LG, Zhou X, Abecasis G, Consortium IARMDG. A scalable Bayesian method for integrating functional information in genome-wide association studies. Am J Hum Genet. 2017;101(3):404–416.

6. van der Wijst MGP, de Vries DH, Brugge H, Westra HJ, Franke L. An integrative approach for building personalized gene regulatory networks for precision medicine. Genome Med. 2018;10(1):96. Available from: https://doi.org/10.1186/s13073-018-0608-4.

7. Zhu X, Stephens M. Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. Nat Comm. 2018;9(1):4361.

8. Kichaev G, Bhatia G, Loh PR, Gazal S, Burch K, Freund MK, et al. Leveraging Polygenic Functional Enrichment to Improve GWAS Power. Am J Hum Genet. 2019;104(1):65–75. Available from: https://pubmed.ncbi.nlm.nih.gov/30595370.

9. Zeng P, Zhou X. Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. Nat Comm. 2017;8:456. Available from: https://doi.org/10.1038/s41467-017-00470-2.

10. Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM. Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. PLoS Genet. 2015;11(4):e1004969. Available from: https://pubmed.ncbi.nlm.nih.gov/25849665.

11. Zhang Y, Qi G, Park JH, Chatterjee N. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. Nat Genet. 2018;50(9):1318–1326.

12. Lloyd-Jones LR, Zeng J, Sidorenko J, Yengo L, Moser G, Kemper KE, et al. Improved polygenic prediction by Bayesian multiple regression on summary statistics. Nat Comm. 2019;10(1):5086. Available from: https://doi.org/10.1038/s41467-019-12653-0.

13. Cheng W, Ramachandran S, Crawford L. Estimation of non-null SNP effect size distributions enables the detection of enriched genes underlying complex traits. PLoS Genet. 2020;16(6):e1008855. Available from: https://doi.org/10.1371/journal.pgen.1008855.

14. Guan Y, Stephens M. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. Ann Appl Stat. 2011;5(3):1780–1815. Available from: https://projecteuclid.org:443/euclid.aoas/1318514285.

15. Zhou X, Carbonetto P, Stephens M. Polygenic modeling with Bayesian sparse linear mixed models. PLoS Genet. 2013;9(2):e1003264.

16. Zhu X, Stephens M. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. Ann Appl Stat. 2017;11(3):1561–1592. Available from: https://projecteuclid.org:443/euclid.aoas/1507168840.

17. George EI, McCulloch RE. Variable selection via Gibbs sampling. J Am Stat Assoc. 1993;88(423):881–889.

18. Blei DM, Jordan MI. Variational inference for Dirichlet process mixtures. Bayesian Anal. 2006;1(1):121–143.

19. Carbonetto P, Stephens M. Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. Bayesian Anal. 2012;7(1):73–108.

20. Carbonetto P, Zhou X, Stephens M. varbvs: Fast variable selection for large-scale regression; 2017. ArXiv.

21. Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK. An introduction to variational methods for graphical models. Mach Learn. 1999;37(2):183–233.

22. Bishop CM. Pattern recognition and machine learning. Springer; 2006.

23. Ormerod JT, Wand MP. Explaining variational approximations. Am Stat. 2010;64(2):140–153.

24. Grimmer J. An introduction to Bayesian inference via variational approximations. Political Anal. 2011;19(1):32–47.

25. Blei DM, Kucukelbir A, McAuliffe JD. Variational inference: A review for statisticians. J Am Stat Assoc. 2017;112(518):859–877.

26. Wand MP, Ormerod JT, Padoan SA, Frühwirth R. Mean field variational Bayes for elaborate distributions. Bayesian Anal. 2011;6(4):847–900.

27. Pham TH, Ormerod JT, Wand MP. Mean field variational Bayesian inference for nonparametric regression with measurement error. Comput Stat Data Anal. 2013;68:375–387.

28. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial (with comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors. Statist Sci. 1999;14(4):382–417. Available from: https://projecteuclid.org:443/euclid.ss/1009212519.

29. de los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, et al. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics. 2009;182(1):375–385. Available from: http://www.genetics.org/content/182/1/375.abstract.

30. de los Campos G, Gianola D, Rosa GJM, Weigel KA, Crossa J. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet Res (Camb). 2010;92(4):295–308.

31. Morota G, Gianola D. Kernel-based whole-genome prediction of complex traits: a review. Front Genet. 2014;5:363.

32. Swain PS, Stevenson K, Leary A, Montano-Gutierrez LF, Clark IBN, Vogel J, et al. Inferring time derivatives including cell growth rates using Gaussian processes. Nat Comm. 2016;7(1):13766. Available from: https://doi.org/10.1038/ncomms13766.

33. Weissbrod O, Geiger D, Rosset S. Multikernel linear mixed models for complex phenotype prediction. Genome Res. 2016;26(7):969–979. Available from: http://genome.cshlp.org/content/26/7/969.abstract.

34. Cheng L, Ramchandran S, Vatanen T, Lietzén N, Lahesmaa R, Vehtari A, et al. An additive Gaussian process regression model for interpretable non-parametric analysis of longitudinal data. Nat Comm. 2019;10(1):1–11.

35. Weissbrod O, Kaufman S, Golan D, Rosset S. Maximum likelihood for Gaussian process classification and generalized linear mixed models under case-control sampling. J Mach Learn Res. 2019;20(108):1–30.

36. Cotter A, Keshet J, Srebro N. Explicit approximations of the Gaussian kernel; 2011. ArXiv.

37. Jiang Y, Reif JC. Modeling epistasis in genomic selection. Genetics. 2015;201:759–768.

38. Crawford L, Wood KC, Zhou X, Mukherjee S. Bayesian approximate kernel regression with variable selection. J Am Stat Assoc. 2018;113(524):1710–1721.

39. Crawford L, Flaxman SR, Runcie DE, West M. Variable prioritization in nonlinear black box methods: A genetic association case study. Ann Appl Stat. 2019;13(2):958–989.

40. Crawford L, Zeng P, Mukherjee S, Zhou X. Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. PLoS Genet. 2017;13(7):e1006869. Available from: https://doi.org/10.1371/journal.pgen.1006869.

41. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, et al. Efficient control of population structure in model organism association mapping. Genetics. 2008;178(3):1709–1723. Available from: http://www.genetics.org/content/178/3/1709.abstract.

42. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong Sy, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010;42(4):348–354. Available from: http://dx.doi.org/10.1038/ng.548.

43. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. Nat Meth. 2011;8(10):833–835. Available from: http://dx.doi.org/10.1038/nmeth.1681.

44. Korte A, Vilhjálmsson BJ, Segura V, Platt A, Long Q, Nordborg M. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat Genet. 2012;44(9):1066–1071. Available from: https://pubmed.ncbi.nlm.nih.gov/22902788.

45. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012;44(7):821–825.

46. Giordano R, Broderick T, Jordan MI. Covariances, robustness and variational bayes. J Mach Learn Res. 2018;19(1):1981–2029.

47. Valdar W, Solberg LC, Gauguier D, Burnett S, Klenerman P, Cookson WO, et al. Genome-wide genetic association of complex traits in heterogeneous stock mice. Nat Genet. 2006;38(8):879–887. Available from: http://dx.doi.org/10.1038/ng1840.

48. Splansky GL, Corey D, Yang Q, Atwood LD, Cupples LA, Benjamin EJ, et al. The Third Generation Cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. Am J Epidemiol. 2007;165(11):1328–1335.

49. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–575.

50. Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2005;33(Database issue):D501–4.

51. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562(7726):203–209. Available from: https://doi.org/10.1038/s41586-018-0579-z.

52. Barbieri MM, Berger JO. Optimal predictive model selection. Ann Statist. 2004;32(3):870–897. Available from: http://projecteuclid.org/euclid.aos/1085408489.

53. McKinney W. Data Structures for Statistical Computing in Python. In: van der Walt S, Millman J, editors. Proceedings of the 9th Python in Science Conference; 2010. p. 51–56.

54. Walt Svd, Colbert SC, Varoquaux G. The NumPy array: a structure for efficient numerical computation. Comput Sci Eng. 2011;13(2):22–30. Available from: https://aip.scitation.org/doi/abs/10.1109/MCSE.2011.37.

55. Lam SK, Pitrou A, Seibert S. Numba: A LLVM-Based Python JIT Compiler. In: Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC. LLVM '15. New York, NY, USA: Association for Computing Machinery; 2015. p. 1–6. Available from: https://doi.org/10.1145/2833157.2833162.

56. Mckerns MM, Sullivan T, Fang A, Aivazis MA. Aivazis, Building a framework for predictive science. In: Proceedings of the 10th Python in Science Conference. Citeseer; 2011. .

57. Kingma DP, Ba J. Adam: a method for stochastic optimization; 2014.

58. Wickham H, François R, Henry L, Müller K. dplyr: A Grammar of Data Manipulation; 2018. R package version 0.8.5. Available from: https://CRAN.R-project.org/package=dplyr.

59. Bates D, Maechler M. Matrix: Sparse and Dense Matrix Classes and Methods; 2019. R package version 1.2-18. Available from: https://CRAN.R-project.org/package=Matrix.

60. Corporation M, Weston S. doParallel: Foreach Parallel Adaptor for the 'parallel' Package; 2019. R package version 1.0.15. Available from: https://CRAN.R-project.org/package=doParallel.

61. Microsoft, Weston S. foreach: Provides Foreach Looping Construct; 2020. R package version 1.4.8. Available from: https://CRAN.R-project.org/package=foreach.

62. Analytics R, Weston S. iterators: Provides Iterator Construct; 2019. R package version 1.0.12. Available from: https://CRAN.R-project.org/package=iterators.

63. Wickham H, Hester J, Chang W. devtools: Tools to Make Developing R Packages Easier; 2019. R package version 2.2.1. Available from: https://CRAN.R-project.org/package=devtools.

64. Hormozdiari F, van de Bunt M, Segrè AV, Li X, Joo JWJ, Bilow M, et al. Colocalization of GWAS and eQTL signals detects target genes. Am J Hum Genet. 2016;99(6):1245–1260. Available from: https://doi.org/10.1016/j.ajhg.2016.10.003.

65. Wang G, Sarkar A, Carbonetto P, Stephens M. A simple new approach to variable selection in regression, with application to genetic fine-mapping; 2019. BioRxiv. Available from: http://biorxiv.org/content/early/2019/07/29/501114.abstract.

66. Benner C, Spencer CCA, Havulinna AS, Salomaa V, Ripatti S, Pirinen M. FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics. 2016;32(10):1493–1501. Available from: https://pubmed.ncbi.nlm.nih.gov/26773131.

67. Nakka P, Raphael BJ, Ramachandran S. Gene and network analysis of common variants reveals novel associations in multiple complex diseases. Genetics. 2016;204(2):783–798. Available from: http://www.genetics.org/content/204/2/783.abstract.

68. Sun R, Hui S, Bader GD, Lin X, Kraft P. Powerful gene set analysis in GWAS with the Generalized Berk-Jones statistic. PLOS Genetics. 2019;15(3):e1007530. Available from: https://doi.org/10.1371/journal.pgen.1007530.

69. Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, et al. Powerful SNP-set analysis for case-control genome-wide association studies. Am J Hum Genet. 2010;86(6):929–942.

70. Holden M, Deng S, Wojnowski L, Kulle B. GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. Bioinformatics. 2008;24(23):2784–2785.

71. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol. 2015;11(4):e1004219–. Available from: https://doi.org/10.1371/journal.pcbi.1004219.

72. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinform. 2013;14(1):128. Available from: https://doi.org/10.1186/1471-2105-14-128.

73. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44(W1):W90–W97. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27141961.

74. Chen X, Hui L, Geiger JD. Role of LDL cholesterol and endolysosomes in amyloidogenesis and Alzheimer's disease. J Neurol Neurophysiol. 2014;5(5):236. Available from: https://pubmed.ncbi.nlm.nih.gov/26413387.

75. Wang H, Eckel RH. What are lipoproteins doing in the brain? Trends Endocrinol Metab. 2014;25(1):8–14. Available from: https://pubmed.ncbi.nlm.nih.gov/24189266.

76. Pitas RE, Boyles JK, Lee SH, Hui D, Weisgraber KH. Lipoproteins and their receptors in the central nervous system. Characterization of the lipoproteins in cerebrospinal fluid and identification of apolipoprotein B,E(LDL) receptors in the brain. J Biol Chem. 1987;262(29):14352–14360.

77. Kay AD, Day SP, Nicoll JAR, Packard CJ, Caslake MJ. Remodelling of cerebrospinal fluid lipoproteins after subarachnoid hemorrhage. Atherosclerosis. 2003;170(1):141–146.

78. Hui L, Han M, Du XD, Zhang BH, He SC, Shao TN, et al. Serum ApoB levels in depressive patients: associated with cognitive deficits. Scientific Rep. 2017;7(1):39992. Available from: https://doi.org/10.1038/srep39992.