PLOS GENETICS

RESEARCH ARTICLE

Multi-scale inference of genetic trait architecture using biologically annotated neural networks

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

1 Department of Computer Science, Brown University, Providence, Rhode Island, United States of America, **2** Center for Computational Molecular Biology, Brown University, Providence, Rhode Island, United States of America, **3** Department of Ecology and Evolutionary Biology, Brown University, Providence, Rhode Island, United States of America, **4** Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, United States of America, **5** Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan, United States of America, **6** Microsoft Research New England, Cambridge, Massachusetts, United States of America, **7** Department of Biostatistics, Brown University, Providence, Rhode Island, United States of America

☯ These authors contributed equally to this work.
* lcrawford@microsoft.com

Citation: Demetci P, Cheng W, Darnell G, Zhou X, Ramachandran S, Crawford L (2021) Multi-scale inference of genetic trait architecture using biologically annotated neural networks. PLoS Genet 17(8): e1009754. https://doi.org/10.1371/journal.pgen.1009754

Editor: Vincent Plagnol, University College London, UNITED KINGDOM

Received: November 23, 2020

Accepted: July 31, 2021

Published: August 19, 2021

Data Availability Statement: The mice dataset from the Wellcome Trust Centre for Human Genetics can be found at http://mtweb.cs.ucl.ac.uk/mus/www/mouse/index.shtml. Data from the UK Biobank Resource (https://www.ukbiobank.ac.uk) was made available under Application Number 22419. The Framingham Heart Study genotype and phenotype data is available in dbGaP (https://www.ncbi.nlm.nih.gov/gap) with accession number phs000007.

Abstract

In this article, we present Biologically Annotated Neural Networks (BANNs), a nonlinear probabilistic framework for association mapping in genome-wide association (GWA) studies. BANNs are feedforward models with partially connected architectures that are based on biological annotations. This setup yields a fully interpretable neural network where the input layer encodes SNP-level effects, and the hidden layer models the aggregated effects among SNP-sets. We treat the weights and connections of the network as random variables with prior distributions that reflect how genetic effects manifest at different genomic scales. The BANNs software uses variational inference to provide posterior summaries which allow researchers to simultaneously perform (*i*) mapping with SNPs and (*ii*) enrichment analyses with SNP-sets on complex traits. Through simulations, we show that our method improves upon state-of-the-art association mapping and enrichment approaches across a wide range of genetic architectures. We then further illustrate the benefits of BANNs by analyzing real GWA data assayed in approximately 2,000 heterogenous stock of mice from the Wellcome Trust Centre for Human Genetics and approximately 7,000 individuals from the Framingham Heart Study. Lastly, using a random subset of individuals of European ancestry from the UK Biobank, we show that BANNs is able to replicate known associations in high and low-density lipoprotein cholesterol content.

Author summary

A common goal in genome-wide association (GWA) studies is to characterize the relationship between genotypic and phenotypic variation. Linear models are widely used tools

Funding: This research was supported in part by grants P20GM109035 (COBRE Center for Computational Biology of Human Disease; PI Rand) and P20GM103645 (COBRE Center for Central Nervous; PI Sanes) from the NIH NIGMS, 2U10CA180794-06 from the NIH NCI and the Dana Farber Cancer Institute (PIs Gray and Gatsonis), an Alfred P. Sloan Research Fellowship, and a David & Lucile Packard Fellowship for Science and Engineering awarded to L. Crawford. This research was also partly supported by US National Institutes of Health (NIH) grant R01 GM118652, and National Science Foundation (NSF) CAREER award DBI-1452622 to S. Ramachandran. G. Darnell was supported by NSF Grant No. DMS-1439786 while in residence at the Institute for Computational and Experimental Research in Mathematics (ICERM) in Providence, RI. X. Zhou was supported by the NIH grant R01 HG009124 and the NSF Grant DMS-1712933. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

in GWA analyses, in part, because they provide significance measures which detail how individual single nucleotide polymorphisms (SNPs) are statistically associated with a trait or disease of interest. However, traditional linear regression largely ignores non-additive genetic variation, and the univariate SNP-level mapping approach has been shown to be underpowered and challenging to interpret for certain trait architectures. While nonlinear methods such as neural networks are well known to account for complex data structures, these same algorithms have also been criticized as "black box" since they do not naturally carry out statistical hypothesis testing like classic linear models. This limitation has prevented nonlinear regression approaches from being used for association mapping tasks in GWA applications. Here, we present Biologically Annotated Neural Networks (BANNs): a flexible class of feedforward models with partially connected architectures that are based on biological annotations. The BANN framework uses approximate Bayesian inference to provide interpretable probabilistic summaries which can be used for simultaneous (*i*) mapping with SNPs and (*ii*) enrichment analyses with SNP-sets (e.g., genes or signaling pathways). We illustrate the benefits of our method over state-of-the-art approaches using extensive simulations. We also demonstrate the ability of BANNs to recover novel and previously discovered genomic associations using quantitative traits from the Wellcome Trust Centre for Human Genetics, the Framingham Heart Study, and the UK Biobank.

Introduction

Over the two last decades, a considerable amount of methodological research in statistical genetics has focused on developing and improving the utility of linear models [1–13]. The flexibility and interpretability of linear models make them a widely used tool in genome-wide association (GWA) studies, where the goal is to test for statistical associations between individual single nucleotide polymorphisms (SNPs) and a phenotype of interest. In these cases, traditional variable selection approaches provide a set of *P*-values or posterior inclusion probabilities (PIPs) which lend statistical evidence on how important each variant is for explaining the overall genetic architecture of a trait. However, this univariate SNP-level mapping approach can be underpowered for "polygenic" traits which are generated by many mutations of small effect [14–19]. To mitigate this issue, more recent work has extended variable selection techniques to identify enriched gene or pathway-level associations, where groups of SNPs within a particular genomic region are combined (commonly known as a SNP-set) to detect biologically relevant disease mechanisms underlying the trait [20–27]. Still, the performance of standard SNP-set methods can be hampered by strict additive modeling assumptions; and the most powerful of these statistical approaches rely on algorithms that are computationally inefficient and unreliable for large-scale sets of data [28].

The explosion of large-scale genomic datasets has provided the unique opportunity to move beyond the traditional linear regression framework and integrate nonlinear modeling techniques as standard statistical tools within GWA analyses. Indeed, nonlinear methods such as neural networks are well known to be most powered in settings when large training data is available [29]. This includes GWA applications where consortiums have data sets that include hundreds of thousands of individuals genotyped at millions of markers and phenotyped for thousands of traits [30, 31]. It is also well known that these nonlinear statistical approaches often exhibit greater predictive accuracy than linear models, particularly for complex traits with broad-sense heritability that is driven by non-additive genetic variation (e.g., gene-by-gene interactions) [32, 33]. One of the key characteristics that leads to better predictive

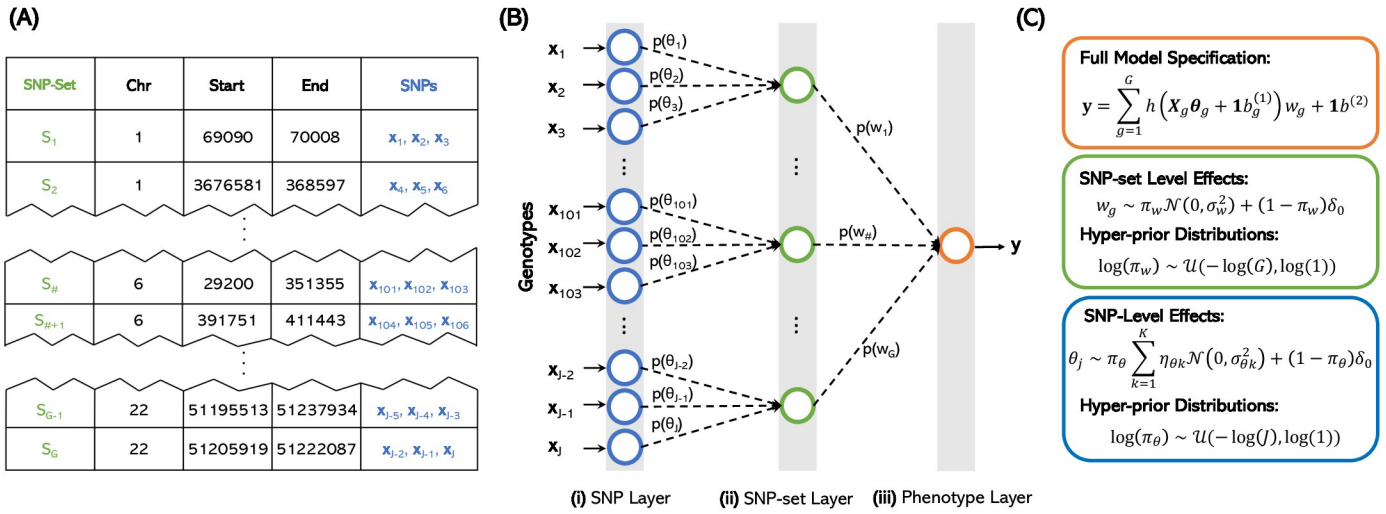

Fig 1. Biologically annotated neural networks (BANNs) allow for efficient multi-scale genotype-phenotype analyses in a unified probabilistic framework by leveraging the hierarchical nature of enrichment studies to define network architecture. (A) The BANNs framework requires an $N \times J$ matrix of individual-level genotypes $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_J]$, an N-dimensional phenotypic vector $\mathbf{y}$, 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 regression 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 $\text{PIP}(j) \equiv \Pr[\theta_j \neq 0 \mid \mathbf{y}, \mathbf{X}]$ and $\text{PIP}(g) \equiv \Pr[w_g \neq 0 \mid \mathbf{y}, \mathbf{X}, \boldsymbol{\theta}_g]$ summarize associations at the SNP and SNP-set level, respectively. The BANNs framework uses variational inference for efficient network training and incorporates nonlinear processing between network layers for accurate estimation of phenotypic variance explained (PVE).

https://doi.org/10.1371/journal.pgen.1009754.g001

performance from nonlinear approaches is the automatic inclusion of higher order interactions between variables being put into the model [34, 35]. For example, neural networks leverage activation functions between layers that implicitly enumerate all possible (polynomial) interaction effects [36]. While this is a partial mathematical explanation for model improvement, in many biological applications, we often wish to know precisely which subsets of variants are most important in defining the architecture of a trait. Unfortunately, the classic statistical idea of variable selection and hypothesis testing is lost within nonlinear methods since they do not naturally produce interpretable significance measures (e.g., P-values or PIPs) like traditional linear regression [35, 37].

In this work, we develop biologically annotated neural networks (BANNs), a nonlinear probabilistic framework for mapping and variable selection in high-dimensional genomic association studies (Fig 1). BANNs are a class of feedforward Bayesian models with partially connected architectures that are guided by predefined SNP-set annotations (Fig 1A). The interpretability of our approach stems from a combination of three key properties. First, the partially connected network architecture yields a hierarchical model where the input layer encodes SNP-level effects, and the single hidden layer models the effects among SNP-sets (Fig 1B). Second, inspired by previous work in the Bayesian neural network literature [38–42], we treat the weights and connections of the network as random variables with sparse prior distributions, which flexibly allows us to model a wide range of sparse and polygenic genetic architectures (Fig 1C). Third, we perform an integrative model fitting procedure where the enrichment of SNP-sets in the hidden layer are directly influenced by the distribution of associated SNPs with nonzero effects on the input layer. These three components collectively make for an effective nonlinear variable selection strategy for conducting association mapping and

enrichment analyses simultaneously on complex traits. With detailed simulations, we assess the power of BANNs to identify significant SNPs and SNP-sets under a variety of genetic architectures, and compare its performance against multiple competing approaches [21, 23, 25–27, 43–46]. We also apply the BANNs framework to six quantitative traits assayed in a heterogenous stock of mice from Wellcome Trust Centre for Human Genetics [47], and two quantitative traits in individuals from the Framingham Heart Study [48]. For the latter, we include an additional study where we independently analyze the same traits in a subset of individuals of European ancestry from the UK Biobank [31].

Results

BANNs framework overview

Biologically annotated neural networks (BANNs) are feedforward models with partially connected architectures that are inspired by the hierarchical nature of biological enrichment analyses in GWA studies (Fig 1). The BANNs software takes in one of two data types: (*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 (Fig 1A); or (*ii*) GWA summary statistics $\mathcal{D} = \{\mathbf{R}, \hat{\boldsymbol{\theta}}\}$ where $\mathbf{R}$ is a $J \times J$ empirical linkage disequilibrium (LD) matrix of pairwise correlations between SNPs and $\hat{\boldsymbol{\theta}}$ are marginal effect size estimates for each SNP computed using ordinary least squares (OLS) (S1 Fig). In both settings, 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. Structurally, sequential layers of the BANNs model represent different scales of genomic units. The first layer of the network takes SNPs as inputs, with each unit corresponding to information about a single SNP. The second layer of the network represents SNP-sets. All SNPs that have been annotated for the same SNP-set are then connected to the same neuron in the second layer (Fig 1B).

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 $\hat{\boldsymbol{\theta}}$ (see Materials and methods). 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 regression 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. Here, we define $h(\bullet)$ to be a Leaky rectified linear unit (Leaky ReLU) activation function [49], where $h(\boldsymbol{x}) = \boldsymbol{x}$ if $\boldsymbol{x} > \mathbf{0}$ and $0.01\boldsymbol{x}$ otherwise. Lastly, for convenience, we assume that the genotype matrix (column-wise) and trait of interest have been mean-centered and standardized.

In this work, we define SNP-sets as collections of contiguous regions of the genome that contain variants within some chromosomal window or neighborhood. More specifically, when

studying real mice and human GWA data, we use gene annotations as defined by the Mouse Genome Informatics database [51] and the NCBI's Reference Sequence (RefSeq) database in the UCSC Genome Browser [50], respectively (Materials and methods). The BANNs framework flexibly allows for overlapping annotations. In this way, SNPs may be connected to multiple hidden layer units if they are located within the intersection of multiple gene boundaries. SNPs that are unannotated, but located within the same genomic region, are connected to their own units in the second layer and represent the intergenic region between two annotated genes. Given the natural biological interpretation of both layers, the partially connected architecture of the BANNs model creates a unified framework for comprehensibly understanding SNP and SNP-set level contributions to the broad-sense heritability of complex traits and phenotypes. Notably, this framework may be easily extended to other biological annotations and applications.

The framing of the BANNs methodology as a Bayesian nonlinear model helps facilitate our ability to perform classic variable selection (Fig 1C; see Materials and methods). Here, we leverage the fact that using nonlinear activation functions for the neurons in the hidden layer implicitly accounts for both additive and non-additive effects between SNPs within a given SNP-set (S1 Text). Following previous work in the literature [38–42], we treat the weights and connections of the neural network as random variables with prior distributions that reflect how genetic effects are manifested at different genomic scales. For the input layer, we assume that the effect size distribution of non-null SNPs can take vastly different forms depending on both the degree and nature of trait polygenicity [28]. For example, polygenic traits are generated by many mutations of small effect, while other phenotypes can be driven by just a few clusters of SNPs with effect sizes much larger in magnitude [19]. To this end, we place a normal mixture prior on the input layer weights to flexibly estimate a wide range of SNP-level effect size distributions [10, 52–54]

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

where δ_0 is a point mass at zero; $\boldsymbol{\sigma}^2_\theta = (\sigma^2_{\theta 1}, \ldots, \sigma^2_{\theta K})$ are variance of the K-nonzero mixture components; $\boldsymbol{\eta}_\theta = (\eta_{\theta 1}, \ldots, \eta_{\theta K})$ represents the marginal (unconditional) probability that a randomly selected SNP belongs to the k-th mixture component such that $\Sigma_k \eta_{\theta k} = 1$; and π_θ denotes the total proportion of SNPs that have a nonzero effect on the trait of interest. Here, we fix $K = 3$ which emulates the hypothesis that SNPs can have large, moderate, and small nonzero effects on phenotypic variation [28]. Similarly, we follow other previous work and assume that enriched SNP-sets contain at least one SNP with a nonzero effect on the trait of interest [26]. This is formulated by placing a spike and slab prior distribution on the weights in the second layer

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

where, in addition to previous notation, π_w denotes the total proportion of SNP-sets that have a nonzero effect on the trait of interest.

By using these point mass mixture distributions in Eqs (2) and (3), we assume that each connection in the neural network has a nonzero weight with: (*i*) probability π_θ for SNP-to-SNP-set connections, and (*ii*) probability π_w for SNP-set-to-phenotype connections. By modifying a variational inference algorithm assuming point-normal priors in multiple linear regression [55, 56] to the neural network setting, we jointly infer posterior inclusion probabilities (PIPs) for SNPs and SNP-sets. These quantities are defined as the posterior probability that the weight of a given connection in the neural network is nonzero, $\text{PIP}(j) \equiv \Pr[\theta_j \neq 0 \mid \mathbf{y}, \mathbf{X}]$ and

$\text{PIP}(g) \equiv \Pr[w_g \neq 0 \mid \mathbf{y}, \mathbf{X}, \boldsymbol{\theta}_g]$. We use this information to prioritize statistically associated SNPs and SNP-sets that significantly contribute to the broad-sense heritability of the trait of interest. With biologically annotated units and the ability to perform statistical inference on explicitly defined parameters, our model presents a fully interpretable extension of neural networks to GWA applications. Additional details and derivations of the BANNs framework can be found in Materials and methods and Supporting information.

Power to detect SNPs and SNP-sets in simulation studies

In order to assess the performance of models under the BANNs framework, we simulated complex traits under multiple genetic architectures using real genotype data on chromosome 1 from ten thousand randomly sampled individuals of European ancestry in the UK Biobank [31] (see Materials and methods and previous work [9, 28]). After quality control procedures, our simulations included 36,518 SNPs (S1 Text). Next, we used the NCBI's Reference Sequence (RefSeq) database in the UCSC Genome Browser [50] to annotate SNPs with the appropriate genes. Unannotated SNPs located within the same genomic region were labeled as being within the "intergenic region" between two genes. Altogether, this left a total of $G = 2{,}816$ SNP-sets to be included in the simulation study.

After the annotation step, we assume a linear model to generate quantitative traits while varying the following parameters: broad-sense heritability (modestly set to $H^2 = 0.2$ and 0.6); the proportion of broad-sense heritability that is being contributed by additive effects versus pairwise *cis*-interaction effects ($\rho = 1$ and 0.5); and the percentage of enriched SNP-sets that influence the trait (set to 1% for sparse and 10% for polygenic architectures, respectively). We use the parameter ρ to assess the neural network's robustness in the presence of non-additive genetic effects between causal SNPs. To this end, $\rho = 1$ 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.

In each simulation scenario, we consider traits being generated with and without additional population structure (Materials and methods, and Supporting information). To do so, we consider two different data compositions with individuals from the UK Biobank. In the first, we simulate synthetic traits only using individuals who self-identify as being of "white British" ancestry. In the second, we simulate traits by randomly subsampling 3,000 individuals who self-identify as being of "white British" ancestry, 3,000 individuals who self-identify as being of "white Irish" ancestry, and 4,000 individuals who identify as being of "any other white background". Note that the latter composition introduces additional population structure into the problem. In the main text and Supporting information, we refer to these datasets as the "British" and "European" cohorts, respectively.

Throughout this section, we assess the performance for two versions of the BANNs framework. The first takes in individual-level genotype and phenotype data; while, the second models GWA summary statistics (hereafter referred to as BANN-SS). For the latter, GWA summary statistics are computed by fitting a single-SNP univariate linear model (via ordinary least squares) after quality control to obtain: effect size estimates, standard errors, and *P*-values for all SNPs in the data. We also use the in-sample genotypes to compute the LD matrix between SNPs. All results are based on 100 different simulated phenotypes for each parameter combination (S1 Text).

The main utility of the BANNs framework is having the ability to detect associated SNPs and enriched SNP-sets simultaneously. Therefore, we compare the performance of BANNs to state-of-the-art SNP and SNP-set level approaches [21, 23, 25–27, 43–46], with the primary idea that our method should be competitive in both settings. For each method, we assess the

empirical power and false discovery rates (FDR) for identifying either the correct causal SNPs or the correct SNP-sets containing causal SNPs (S1–S8 Tables). Frequentist approaches are evaluated at a Bonferroni-corrected threshold for multiple hypothesis testing (e.g., $P = 0.05/36518 = 1.37 \times 10^{-6}$ at the SNP-level and $P = 0.05/2816 = 1.78 \times 10^{-5}$ at the SNP-set level, respectively); while, Bayesian methods are evaluated according to the median probability model (PIPs and posterior enrichment probability ≥ 0.5) [57]. We also compare each method's ability to rank true positives over false positives via receiver operating characteristic (ROC) and precision-recall curves (Fig 2 and S2–S16 Figs). Specific results about these analyses are given below.

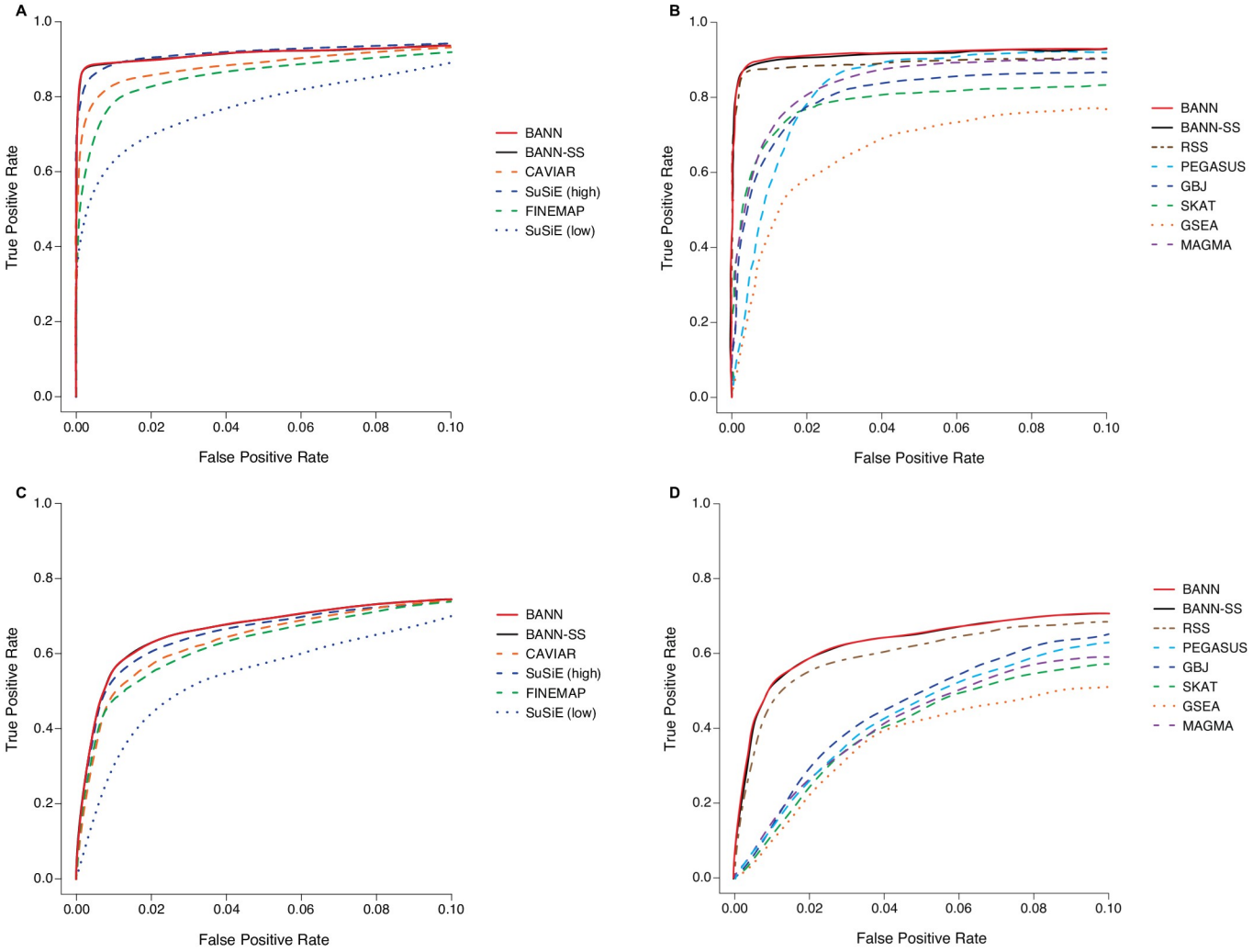

Fig 2. Receiver operating characteristic (ROC) curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations (British cohort). Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.6$ with only contributions from additive effects set (i.e., $\rho = 1$). We 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 set the number of causal SNPs with nonzero effects to be 1% and 10% 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 [45], SuSiE [46], and FINEMAP [44]. 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 [26], PEGASUS [25], GBJ [27], SKAT [21], GSEA [43], and MAGMA [23]. Note that the upper limit of the x-axis has been truncated at 0.1. All results are based on 100 replicates (see S1 Text).

https://doi.org/10.1371/journal.pgen.1009754.g002

Mapped SNP-level results. For SNP-level comparisons, we used three fine-mapping methods as benchmarks: CAVIAR [45], SuSiE [46], and FINEMAP [44]. Each of these methods implement Bayesian variable selection strategies, in which different sparse prior distributions are placed on the "true" effect sizes of each SNP and posterior inclusion probabilities (PIPs) are used to summarize their statistical relevance to the trait of interest. Notably, both CAVIAR (exhaustively) and FINEMAP (approximately) search over different models to find the best combination of associated SNPs with nonzero effects on a given phenotype. On the other hand, the software for SuSiE requires an input ℓ which fixes the maximum number of causal SNPs to include in the model. In this section, we consider results when this input number is high ($\ell = 3000$) and when this input number is low ($\ell = 10$). While SuSiE is applied to individual-level data, both CAVIAR and FINEMAP require summary statistics where marginal z-scores are treated as a phenotype and modeled with in-sample estimate of the LD matrix.

Overall, BANNs, BANN-SS, and SuSiE (with high $\ell = 3000$) generally achieve the greatest empirical power and lowest FDR across all genetic architectures we considered (S1–S8 Tables). These three approaches also stand out in terms of true-versus-false positive rates and precision-versus-recall (Fig 2 and S2–S16 Figs). Notably, the choice of the ℓ parameter largely influenced the performance of SuSiE, as it was consistently the worst performing method when we underestimated the number of causal SNPs with nonzero effects *a priori* (i.e., $\ell = 10$). Importantly, these performance gains come with a cost: the computational run time of SuSiE becomes much slower as ℓ increases (S9 Table). For more context, an analysis on just 4,000 individuals and 10,000 SNPs takes the BANNs methods an average of 319 seconds to run on a CPU; while, SuSiE can take up to nearly twice as long to complete as ℓ increases (e.g., average runtimes of 23 and 750 seconds for $\ell = 10$ and 3000, respectively).

Training BANNs on individual-level data relatively becomes the best approach when the broad-sense heritability of complex traits is partly made up of pairwise genetic interaction effects between causal SNPs (e.g., $\rho = 0.5$; see S5–S8 Figs and S13–S16 Figs)—particularly when traits have low heritability with polygenic architectures (e.g., $H^2 = 0.2$). A direct comparison of the PIPs derived by BANNs and SuSiE shows that the proposed neural network training procedure enables the ability to identify associated SNPs even in these more complex phenotypic architectures (Fig 3 and S17–S23 Figs). It is important to note that the inclusion probabilities were not perfectly calibrated for either BANNs or SuSiE in our simulations (S24 Fig), despite FDR still being reasonably well controlled for both methods (S1–S8 Tables). We hypothesize that the quality of PIP calibration for BANNs is a direct consequence of its variational inference algorithm which tends to favor sparse solutions and can lead to greater type II versus type I error rates [46, 55]. To investigate how choices in the BANNs model setup contributed to improved variable selection over SuSiE, we also performed an "ablation analysis" [58, 59] where we modified parts of the algorithm independently and observed their direct effect on method performance (S25 Fig). Ultimately, these results for BANNs were enabled by a combination of (*i*) using ReLU activation functions in the hidden layers of the BANNs framework, which implicitly enumerates the interactions between SNPs within a given SNP-set, and (*ii*) using model averaging to estimate the inclusion probabilities for the network weights (S1 Text). Note the absence of the nonlinear activation function only affected the power of BANNs in simulations where there were non-additive genetic effects (e.g., S25(C) and S25(D) Fig).

As a final comparison, the BANN-SS, CAVIAR, and FINEMAP methods see a decline in performance for these same scenarios with genetic interactions. Assuming that the additive and non-additive genetic effects are uncorrelated, this result is also expected since summary statistics are often derived from simple linear additive regression models that (in theory)

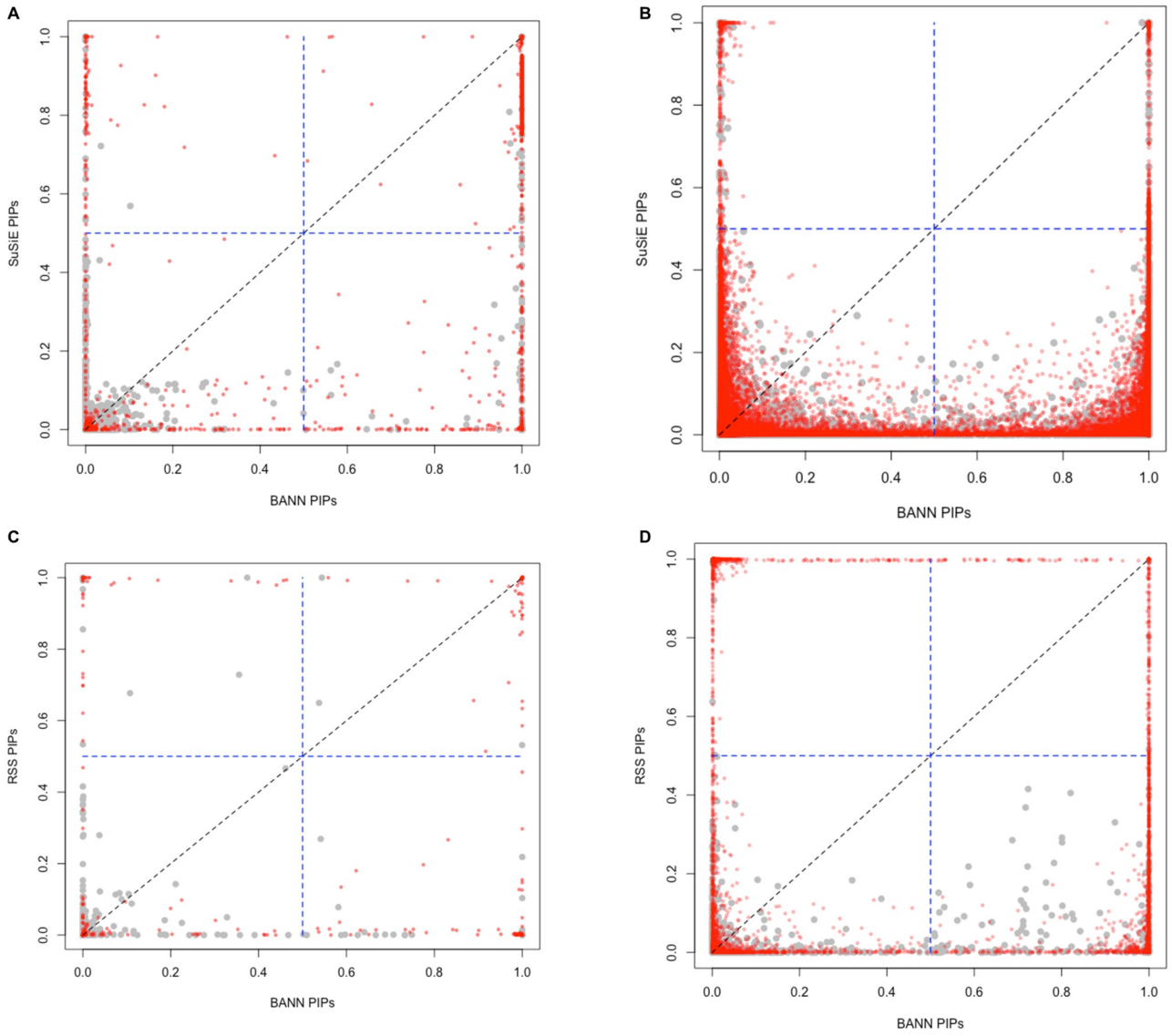

Fig 3. Scatter plots comparing how the integrative neural network training procedure enables the ability to identify associated SNPs and enriched SNP-sets in simulations (British cohort). Quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.6$ with only contributions from additive 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 nonzero effects to be 1% and 10% of all SNPs located within the enriched SNP-sets, respectively. Results are shown comparing the posterior inclusion probabilities (PIPs) derived by the BANNs model on the x-axis and **(A, C)** SuSiE [46] and **(B, D)** RSS [26] 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) [57]. 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 S1 Text).

https://doi.org/10.1371/journal.pgen.1009754.g003

partition or marginalize out proportions of the phenotypic variance that are contributed by nonlinearities [9, 13].

Enriched SNP-set level results. For comparisons between SNP-set level methods, we consider six gene or SNP-set enrichment approaches including: RSS [26], PEGASUS [25], GBJ [27], SKAT [21], GSEA [43], and MAGMA [23]. SKAT, VEGAS, and PEGASUS fall within the same class of frequentist approaches, in which SNP-set GWA *P*-values are assumed to be

drawn from a correlated chi-squared distribution with covariance estimated using an empirical LD matrix [60]. MAGMA is also a frequentist approach in which gene-level *P*-values are derived from distributions of SNP-level effect sizes using an *F*-test [23]. GBJ attempts to improve upon the previously mentioned methods by generalizing the Berk-Jones statistic to account for complex correlation structures and adaptively adjust the size of annotated SNP-sets to only SNPs that maximize power [61]. Lastly, RSS is a Bayesian linear regression method which places a likelihood on the observed SNP-level GWA effect sizes (using their standard errors and LD estimates), and assumes a spike-and-slab shrinkage prior on the true SNP effects to derive a probability of enrichment for genes or other annotated units [62]. It is worth noting that, while RSS and the BANNs framework are conceptually different, the two methods utilize very similar variational approximation algorithms for posterior inference [55] (Materials and methods, and Supporting information).

Similar to the conclusions drawn during the SNP-level assessments, both the BANNs and BANN-SS implementations had among the best tradeoffs between true and false positive rates for detecting enriched SNP-sets across all simulations—once again, including those scenarios which also considered pairwise interactions between causal SNPs (Fig 2 and S2–S16 Figs, and S1–S8 Tables). Since RSS is an additive model, it sees a decline in performance for the more complex genetic architectures that we simulated. A direct comparison between the PIPs from BANNs and RSS can be found in Fig 3 and S17–S24 Figs. Once again, training BANNs on individual-level data becomes the best approach when the broad-sense heritability of complex traits is partly made up of non-additive genetic variation. Our ablation analysis results suggest that the nonlinear activation function plays an important role here (S25 Fig). While RSS also performs generally well for the additive trait architectures, the algorithm for the model often takes twice as long than either of the BANNs implementations to converge (S10 Table). PEGASUS, GBJ, SKAT, and MAGMA are score-based methods and, thus, are expected to take the least amount of time to run. BANNs and RSS are hierarchical regression-based methods and the increased computational burden of these approaches results from their need to do (approximate) Bayesian posterior inference. Previous work has suggested that, when using GWA summary statistics to identify genotype-phenotype associations at the SNP-set level, having the ability to adaptively account for possibly inflated SNP-level effect sizes and/or *P*-values is crucial [28]. Therefore, it is understandable why the score-based methods consistently struggle relative to the regression-based approaches even in the simplest simulation cases where traits are generated to have high broad-sense heritability, sparse phenotypic architectures that are dominated by additive genetic effects, and total phenotypic variance that is not confounded by additional population structure (Fig 2 and S2–S16 Figs). Both the BANN-SS and RSS methods use shrinkage priors to correct for potential inflation in GWA summary statistics and recover estimates that are better correlated with the true generative model for the trait of interest.

Estimating total phenotypic variance explained in simulation studies

While our main focus is on conducting multi-scale inference of genetic trait architecture, because the BANNs framework provides posterior estimates for all weights in the neural network, we are able to also provide an estimate of phenotypic variance explained (PVE). Here, we define PVE as the total proportion of phenotypic variance that is explained by genetic effects, both additive and non-additive, collectively [16]. Within the BANNs framework, this estimation can be done on both the SNP and SNP-set level while using either genotype-phenotype data or summary statistics (S1 Text). As a reminder, for our simulation studies, the true PVE is set to $H^2 = 0.2$ and 0.6, respectively. We assess the ability of BANNs to recover these true estimates using root mean square error (RMSE) (S26 and S27 Figs). In order to be

successful at this task, the neural network needs to accurately estimate both the individual effects of causal SNPs in the input layer, as well as their cumulative effects for SNP-sets in the outer layer. BANNs and BANN-SS exhibit the most success with traits have additive sparse architectures (with and without additional population structure)—achieving PVE estimates with RMSEs as low as 4.54×10^{-3} and 4.78×10^{-3} on the SNP and SNP-set levels for highly heritable phenotypes, respectively. However, both models underestimate the total PVE in polygenic traits and traits with pairwise SNP-by-SNP interactions. Therefore, even though the BANNs framework is still able to correctly prioritize the appropriate SNPs and SNP-sets, in these more complicated settings, we misestimate the approximate posterior means for the network weights and overestimate the variance of the residual training error (S1 Text). Similar observations have been noted when using variational inference [63, 64]. Results from other work also suggest that the sparsity assumption on the SNP-level effects can lead to the underestimation of the PVE [16, 65].

Mapping genomic enrichment in heterogenous stock of mice

We apply the BANNs framework to individual-level genotypes and six quantitative traits in a heterogeneous stock of mice dataset from the Wellcome Trust Centre for Human Genetics [47]. This data contains approximately 2,000 individuals genotyped at approximately 10,000 SNPs—with specific numbers varying slightly depending on the quality control procedure for each phenotype (S1 Text). For SNP-set annotations, we used the Mouse Genome Informatics database (http://www.informatics.jax.org) [51] to map SNPs to the closest neighboring gene (s). Unannotated SNPs located within the same genomic region were labeled as being within the "intergenic region" between two genes. Altogether, a total of 2,616 SNP-sets were analyzed. The six traits that we consider are grouped based on their category and include: body mass index (BMI) and body weight; percentage of CD8+ cells and mean corpuscular hemoglobin (MCH); and high-density and low-density lipoprotein (HDL and LDL, respectively). We choose to analyze these particular traits because their architectures represent a realistic mixture of the simulation scenarios we detailed in the previous section (i.e., varying different values of ρ). Specifically, the mice in this study are known to be genetically related with population structure and these particular traits have been shown to have various levels of broad-sense heritability with different contributions from both additive and non-additive genetic effects [35, 37, 47, 66–68].

For each trait, we provide a summary table which lists the PIPs for SNPs and SNP-sets after fitting the BANNs model to the individual-level genotypes and phenotype data (S11–S16 Tables). We use Manhattan plots to visually display the variant-level mapping results across each of the six traits, where chromosomes are shown in alternating colors for clarity and associated SNPs with PIPs above the median probability model threshold are highlighted (S28 Fig). As a comparison, we also report the corresponding SNP and SNP-set level PIPs after running SuSiE [46] and RSS [26] on these same data, respectively. Across all traits, BANNs identified 71 associated SNPs and 57 enriched SNP-sets (according to the median probability model threshold). In comparison, SuSiE identified 22 associated SNPs (11 of which were also identified by BANNs) and RSS identified 14 enriched SNP-sets (6 of which were also identified by BANNs). Importantly, many of the candidate genes and intergenic regions selected by the BANNs model have been previously discovered by past publications as having some functional relationship with the traits of interest (Table 1). For example, BANNs reports the genes *Btbd9* and *hlb156* as being enriched for the percentage of CD8+ cells in mice (PIP = 0.87 and 0.72 versus RSS PIP = 0.02 and 0.68, respectively). This same chromosomal region on chromosome 17 was also reported in the original study as having highly significant quantitative trait loci and

Table 1. Notable enriched SNP-sets after applying the BANNs framework to six quantitative traits in heterogenous stock of mice from the Wellcome Trust Centre for Human Genetics. [47]. The traits include: body mass index (BMI), percentage of CD8+ cells, high-density lipoprotein (HDL), low-density lipoprotein (LDL), mean corpuscular hemoglobin (MCH), and body weight. Here, SNP-set annotations are based on gene boundaries defined by the Mouse Genome Informatics database (see URLs). 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 table. 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 enriched if it has a PIP(g) $\geq$ 0.5 (i.e., the "median probability model" threshold [57]). We report the "top" associated SNP within each region and its corresponding PIP(j). We also report the corresponding SNP and SNP-set level results after running SuSiE [46] and RSS [26] on these same traits, respectively. The last column details references and literature sources that have previously suggested some level of association or enrichment between the each genomic region and the traits of interest. See S11–S16 Tables for the complete list of SNP and SNP-set level results.

Trait	SNP-Set	Chr	PIP(g)	Rank	RSS PIP	RSS Rank	Top SNP	PIP(j)	SuSiE PIP	SuSiE Rank	Ref(s)
BMI	*Dmd*	X	0.900	1	0.862	2	rs3090667	0.600	0.140	10	[69]
	Mir466q-Slc2a2	3	0.816	3	0.371	3	rs6269713	0.477	0.009	124	[71]
	Gm22219-Mc4r	18	0.740	5	0.001	81	rs3696955	0.039	0.264	7	[70]
CD8+	*Gm46177-Gm30088*	1	0.968	1	0.307	4	mhcCD8a3	1.000	0.998	3	[72–74]
	Btbd9	17	0.866	7	0.019	7	CEL-17_31069801	1.000	0.080	38	[75, 76]
	hlb156	17	0.720	8	0.683	3	CEL-17_31069801	1.000	0.080	38	[51]
HDL	*Pphc2*	4	0.976	3	0.395	6	rs3724711	1.000	0.136	16	[77]
	Ctnna2	6	0.886	8	0.908	2	rs3710419	1.000	0.712	4	[78]
	hlb156	17	0.589	9	0.481	4	CEL-17_31069801	1.000	0.142	15	[51]
LDL	*Btbd9*	17	0.983	1	0.275	6	CEL-17_31069801	1.000	0.163	6	[79, 80]
	Pphc2	4	0.941	3	0.428	4	rs3724711	1.000	0.452	2	[77]
	Syt14	1	0.852	7	0.070	47	rs3654706	0.001	0.002	626	[81–84]
MCH	*Btbd9*	17	0.905	2	0.387	5	CEL-17_31069801	1.000	0.421	9	[75, 76]
	Picalm	7	0.648	8	0.723	3	rs3704554	0.070	0.263	13	[85]
	Ebf1	11	0.500	10	0.108	11	rs3693846	0.009	2.82×10^{-4}	424	[86]
Weight	*Wdpcp*	11	0.969	1	0.412	6	rs13481023	1.000	0.391	12	[79, 80]
	Chrm2	6	0.882	3	0.195	13	rs3676478	0.012	0.102	26	[87, 88]
	Csmd1	8	0.759	5	0.408	7	rs3709567	0.001	1.32×10^{-9}	2166	[89]

https://doi.org/10.1371/journal.pgen.1009754.t001

contributing non-additive variation for CD8+ cells (bootstrap posterior probability equal to 1.00) [47]. Similarly, the X chromosome is well known to strongly influence adiposity and metabolism in mice [66]. As expected, in body weight and BMI, our approach identified significant enrichment in this region—headlined by the dystrophin gene *Dmd* in both cases [69]. Finally, we note that including intergenic regions in our analyses allows us to discover trait relevant genomic associations outside the immediate gene annotations provided by the Mouse Genome Informatics database. This proved important for BMI where BANNs reported the region between *Gm22219* and *Mc4r* on chromosome 18 as having a relatively high PIP of 0.74 (versus an RSS PIP = 1×10^{-3} for reference). Recently, a large-scale GWA study on individuals from the UK Biobank showed that variants around *MC4R* protect against obesity in humans [70].

Overall, the results from this smaller GWA study highlight three key characteristics resulting from the sparse probabilistic assumptions underlying the BANNs framework. First, the variational spike and slab prior placed on the weights of the neural network will select no more than a few variants in a given LD block [55]. This is important since traditional naïve SNP-set methods will often exhibit high false positive rates due to many of these correlated regions along the genome [28]. Second, we see that the enrichment of a SNP-set is influenced by the relative posterior distribution of zero and nonzero SNP-level effect sizes within its annotated genomic window (S11–S16 Tables). In other words, a SNP-set is not guaranteed to have a high inclusion probability just because it contains one SNP with a large nonzero effect; however,

BANNs will report a SNP-set as insignificant if the total ratio of non-causal SNPs within the set heavily outweighs the number of causal SNPs that have been annotated for the same region. To this end, in the presence of large SNP-sets, the BANNs framework will favor preserving false discovery rates at the expense of having slightly more false negatives. Lastly, the careful modeling of the SNP-level effect size distributions and considering genetic interactions enhances our ability to conduct multi-scale genomic inference. In this particular study, we show the power to still find trait relevant SNP-sets with variants that are not marginally strong enough to be detected individually, but have notable genetic signal when their weights are aggregated together (again see Table 1 and S28 Fig).

Analyzing lipoproteins in the Framingham Heart Study

Next, we apply the BANNs framework to two continuous plasma trait measurements—high-density lipoprotein (HDL) and low-density lipoprotein (LDL) cholesterol—assayed in 6,950 individuals from the Framingham Heart Study [48] genotyped at 394,174 SNPs genome-wide. Following quality control procedures, we regressed out the top ten principal components of the genotype data from each trait to control for population structure (S1 Text). Next, we used the gene boundaries listed in the NCBI's RefSeq database from the UCSC Genome Browser [50] to define SNP-sets. In this analysis, we define genes with boundaries in two ways: (*a*) we use the UCSC gene boundary definitions directly, or (*b*) we augment the gene boundaries by adding SNPs within a ±500 kilobase (kb) buffer to account for possible regulatory elements. Genes with only 1 SNP within their boundary were excluded from either analysis. 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—using the UCSC boundaries. When including the 500kb buffer, a total of $G = 35{,}871$ SNP-sets were analyzed.

For each trait, we again fit the BANNs model to the individual-level genotype-phenotype data and used the median probability model threshold as evidence of statistical significance for all weights in the neural network (S17–S19 Tables). We also again report the corresponding SNP and SNP-set level PIPs after running SuSiE and RSS on these same data. Note that while BANNs is run on the genome-wide data jointly, for computational considerations, SuSiE and RSS are run on a chromosome-by-chromosome basis. A complete breakdown of the overlap of findings between BANNs, SuSiE, and RSS can be found on the first page of S20 Table. In Fig 4, we show Manhattan plots of the variant-level association mapping results for BANNs, where each significant SNP is color coded according to its SNP-set annotation. As an additional validation step, we took the enriched SNP-sets identified by BANNs in each trait and used the gene set enrichment analysis tool Enrichr [90, 91] to identify the categories that they overrepresent in the database of Genotypes and Phenotypes (dbGaP) and the NHGRI-EBI GWAS Catalog (S29 and S30 Figs). Similar to our results in the previous section, the BANNs framework identified many SNPs and SNP-sets that have been shown to be associated with cholesterol-related processes in past publications (Table 2 with UCSC gene boundary definitions and S17 Table with augmented buffer). For example, in HDL, BANNs identified an enriched intergenic region between the genes *HERPUD1* and *CETP* (PIP = 1.00 versus RSS PIP = 0.78) which has been also replicated in multiple GWA studies with diverse cohorts [92–95]. The Enrichr analyses were also consistent with published results (S29 and S30 Figs). For example, the top ten significant enriched categories in the GWAS Catalog (i.e., Bonferroni-correct threshold P-value $< 1 \times 10^{-5}$ or Q-value < 0.05) for HDL-associated SNP-sets selected by the BANNs model are either directly related to lipoproteins and cholesterol (e.g., "Alpolipoprotein

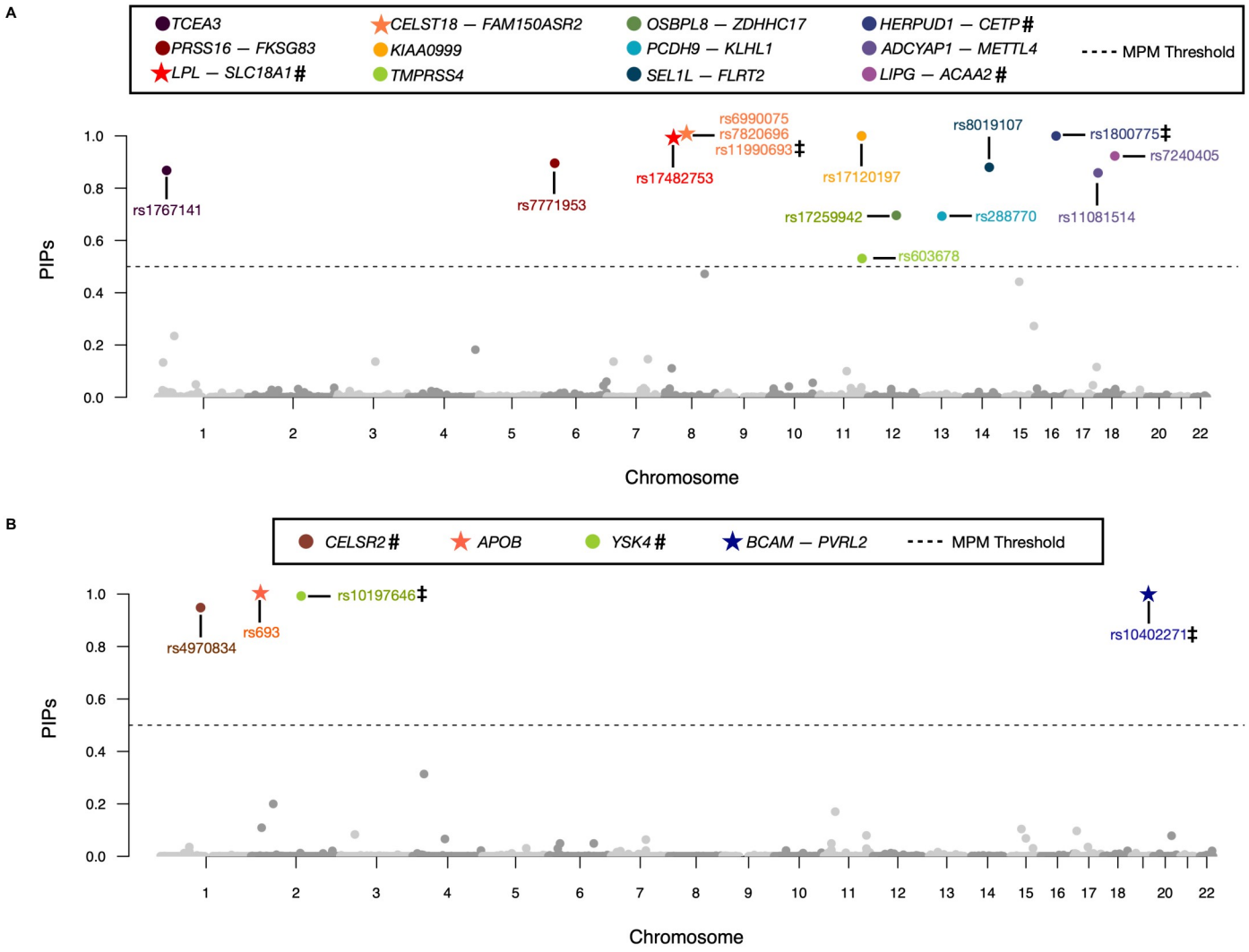

Fig 4. Manhattan plot of variant-level association mapping results for high-density and low-density lipoprotein (HDL and LDL, respectively) traits in the Framingham Heart Study [48]. Posterior inclusion probabilities (PIP) for the neural network weights are derived from the BANNs model fit on individual-level 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 [57]. 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. Double daggers (‡) denote SNPs that are also identified when using SuSiE [46] to analyze the same traits, and hashtag symbols (#) denote SNP-sets that are identified by RSS [26]. Stars (★) denote SNPs and SNP-sets identified by BANNs that replicate in our analyses of HDL and LDL using ten thousand randomly sampled individuals of European ancestry from the UK Biobank [31]. Gene set enrichment analyses for these SNP-sets identified by BANNs can be found in S29 and S30 Figs. A complete list of PIPs for all SNPs and SNP-sets computed in these two traits can be found in S18 and S19 Tables. Results for the additional study with the independent UK Biobank dataset [31] are illustrated in S31–S33 Figs and full results are listed in S21 and S22 Tables.

https://doi.org/10.1371/journal.pgen.1009754.g004

A1 levels", "HDL cholesterol levels") or related to metabolic functions (e.g., "Lipid metabolism phenotypes", "Metabolic syndrome").

As in the previous analysis, the results from this analysis also highlight insight into complex trait architecture enabled by the variational inference used in the BANNs software. SNP-level and SNP-set results remain consistent with the qualitative assumptions underlying our probabilistic hierarchical model. For instance, previous studies have estimated that rs599839 (chromosome 1, bp: 109822166) and rs4970834 (chromosome 1, bp: 109814880) explain

Table 2. Top three enriched SNP-sets after applying the BANNs framework to 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. These regions are labeled as *Gene1-Gene2* in the table. 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 enriched if it has a PIP(g) $\geq$ 0.5 (i.e., the "median probability model" threshold [57]). We report the "top" associated SNP within each region and its corresponding PIP(j). We also report the corresponding SNP and SNP-set level results after running SuSiE [46] and RSS [26] on these same traits, respectively. The last column details references and literature sources that have previously suggested some level of association or enrichment between the each genomic region and the traits of interest. See S18 and S19 Tables for the complete list of SNP and SNP-set level results. *: Multiple SNP-sets were tied for this ranking. ♣: SNPs and SNP-sets replicated in an independent analysis of ten thousand randomly sampled individuals of European ancestry from the UK Biobank [31].

Trait	SNP-Set	Chr	PIP(g)	Rank	RSS PIP	RSS Rank	Top SNP	PIP(j)	SuSiE PIP	SuSiE Rank	Ref(s)
HDL	*HERPUD1-CETP*♣	16	0.999	1*	0.781	5	rs1800775♣	1.000	1.000	1	[92–95]
	ST18-FAM150A	8	0.999	1*	0.869	3	rs6990075	1.000	0.006	107	[98]
	TCEA3	1	0.989	2	1.22×10^{-4}	15056	rs1767141	0.868	0.039	21	[99]
LDL	*CELSR2*	1	0.989	1	0.972	2	rs4970834	0.948	1.12×10^{-4}	4559	[100–102]
	BCAM-PVRL2♣	19	0.987	2	0.237	12	rs10402271♣	0.998	0.966	2	[103–105]
	APOB♣	2	0.976	3	0.167	18	rs693♣	0.999	0.278	6	[103, 106, 107]

https://doi.org/10.1371/journal.pgen.1009754.t002

approximately 1% of the phenotypic variation in circulating LDL levels [96]. Since these two SNPs are physically closed to each other and sit in a high LD block ($r^2 \approx 0.63$ with $P < 1 \times 10^{-4}$ [97]), the spike and slab prior in the BANNs framework will maintain the non-zero weight for one and penalize the estimated effect of the other. Indeed, in our analysis, rs4970834 was reported to be associated with LDL (PIP = 0.95 versus SuSiE PIP = 1.12×10^{-4}), while the effect size of rs599839 was shrunk towards 0 (PIP = 1×10^{-4} versus SuSiE PIP = 0.99). A similar issue can occur in correctly identifying enriched SNP-sets when nearby sets contain SNPs in tight LD. For example, when augmenting the boundary of SNP-set annotations by a ±500 kilobase buffer, BANNs tends to shrink the PIP of at least one member of overlapping or correlated sets. Due to the variational approximations utilized by BANNs (Materials and methods, and Supporting information), if two SNPs or SNP-sets are in strong LD, the model will tend to select just one of them [26, 55].

Independent lipoprotein study using the UK Biobank

To further validate our results from the Framingham Heart Study, we also independently apply BANNs to analyze HDL and LDL cholesterol traits in ten thousand randomly sampled individuals of European ancestry from the UK Biobank [31]. Here, we filter the imputed genotypes from the UK Biobank to keep only the same 394,174 SNPs that were used in the Framingham Heart Study analyses from the previous section. We then apply BANNs, SuSiE, and RSS to the individual-level data and in-sample derived summary statistics using the same (a) 18,364 SNP-set annotations based on the NCBI's RefSeq database from the UCSC Genome Browser [50] and (b) 35,849 SNP-sets when applying the augmented ±500 kilobase buffer. It is important to note that we restrict this analysis to just ten thousand individuals due to computational considerations for BANNs and SuSiE since they take in individual level data. In S31 Fig, we show the BANNs variant-level Manhattan plots for the independent UK Biobank cohort with significant SNPs color coded according to their SNP-set annotation. Once again, we use the median probability model threshold to determine statistical significance for all weights in the neural network, and a complete breakdown of the overlap of findings between BANNs, SuSiE, and RSS between the traits can be found in S20 Table. Lastly, S21 and S22 Tables give the complete list of all SNP and SNP-set level results in this additional UK Biobank study.

Despite the UK Biobank being a completely independent dataset, we found that BANNs was able to replicate two SNPs and two SNP-sets in HDL and two SNPs and one SNP-set that we observed in the Framingham Heart Study analysis (see specially marked rows in Table 2 and S17 Table, as well as the overlap summary given in S20 Table). For example, in HDL, both the variants rs1800775 (PIP = 1.00 versus SuSiE PIP = 1.00) and rs17482753 (PIP = 1.00 versus SuSiE PIP = 0.73) were replicated. BANNs also identified the corresponding intergenic region between the genes *HERPUD1* and *CETP* as being enriched (PIP = 1.00 versus RSS PIP = 1.00). In our analysis of LDL, BANNs replicated two out of the four associated SNPs: rs693 within the *APOB* gene, and rs10402271 which falls within the intergenic region between genes *BCAM* and *PVRL2*.

There were a few scenarios where a given SNP-set was replicated but the leading SNP in that region differed between the two studies. For instance, while the intergenic region between *LIPG* and *ACAA2* was enriched in both cohorts, the variant rs7240405 was found to be most associated with HDL in the Framingham Heart Study; a different SNP, rs7244811, was identified in the UK Biobank (Fig 4 and S31 Fig) Similarly, in the analysis with the ±500 kilobase buffer for SNP-set annotations, rs4939883 in the intergenic region between *LIPG* and *ACAA2* was found to be significant for HDL in the UK Biobank instead of rs7244811 which was selected in the Framingham Heart Study. These discrepancies at the variant level are likely due to: (*i*) the sparsity assumption imposed by BANNs, which lead the model to select one of two variants in high LD; and (*ii*) ancestry differences among individuals from the two studies likely also generate different LD structures in the same genomic region.

As a final step, we took the enriched SNP-sets identified by BANNs in the UK Biobank and used Enrichr [90, 91] to ensure that we were still obtaining trait relevant results (S32 and S33 Figs). Indeed, for both HDL and LDL, the most overrepresented categories in dbGaP and the GWAS Catalog (i.e., Bonferroni-correct threshold P-value $< 1 \times 10^{-5}$ or Q-value < 0.05) was consistently the trait of interest—followed by other functionally related gene sets such as "Metabolic syndrome" and "Cholesterol levels". This story remained largely consistent even when augmenting SNP-set annotations with a ±500 kilobase buffer (S32 and S33 Figs). Overall, the sensible results from performing mapping on the variant-level and enrichment analyses on the SNP-set level in two different independent datasets, only further enhances our confidence about the potential impact of the BANNs framework in GWA studies.

Discussion

Recently, nonlinear approaches have been applied in biomedical genomics for prediction-based tasks, particularly using GWA datasets with the objective of predicting phenotypes [108–112]. However, since the classical idea of variable selection and hypothesis testing is lost within these statistical algorithms, they have not been widely used for association mapping where the goal is to identify significant SNPs or genes underlying complex traits. Here, we present Biologically Annotated Neural Networks (BANNs): a class of feedforward probabilistic models which overcome this limitation by incorporating partially connected architectures that are guided by predefined SNP-set annotations. This creates an interpretable and integrative framework where the first layer of the neural network encodes SNP-level effects and the neurons within the hidden layer represent the different SNP-set groupings. We frame the BANNs methodology as a Bayesian nonlinear regression model and use sparse prior distributions to perform variable selection on the network weights. By modifying a well established variational inference algorithm, we are able to derive posterior inclusion probabilities (PIPs) which allows researchers to carry out SNP-level mapping and SNP-set enrichment analyses, simultaneously. While we focus on genomic motivations in this study, the concept of partially connected

neural networks may extend to any scientific application where annotations can help guide the groupings of variables.

Through extensive simulation studies, we demonstrate the utility of the BANNs framework on individual-level data (Fig 1) and GWA summary statistics (S1 Fig). Here, we showed that both implementations are consistently competitive with commonly used SNP-level association mapping methods and state-of-the-art SNP-set enrichment methods in a wide range of genetic architectures (Figs 2 and 3, and S2–S23 Figs, and S1–S8 Tables). The advantage of our approach was most clear when the broad-sense heritability of the complex traits included pairwise genetic interactions. In two real GWA datasets, we demonstrated the ability of BANNs to prioritize trait relevant SNPs and SNP-sets that have been identified by previous publications and functional validation studies (Fig 4 and S28–S30 Figs, and Tables 1 and 2 and S11–S19 Tables). Lastly, using a third real dataset, we assess the ability of BANNs to statistically replicate a subset of these findings in an independent cohort (S31–S33 Figs and S21 and S22 Tables).

The current implementation of the BANNs framework offers many directions for future development and applications. Perhaps the most obvious limitation is that ill-annotated SNP-sets can bias the interpretation of results and lead to misplaced scientific conclusions (i.e., might cause us to highlight the "wrong" gene [113, 114]). This is a common issue in most enrichment methods [28]; however, similar to other hierarchical methods like RSS [26], BANNs is likely to rank SNP-set enrichments that are driven by just a single SNP as less reliable than enrichments driven by multiple SNPs with nonzero effects. Another current limitation for the BANNs model comes from the fact that it uses a variational inference to estimate its parameters. While the current implementation works reasonably well for large datasets (S9 and S10 Tables), we showed that our sparse prior assumption combined with the variational expectation-maximization algorithm can lead to slightly miscalibrated PIPs (S24 Fig), underestimated approximations of the PVE (S27 and S28 Figs), and will occasionally miss causal SNPs if they are in high LD with other non-causal SNPs in the dataset. For example, in the application to the Framingham Heart Study, BANNs estimates the PVE for HDL and LDL to be 0.11 and 0.04, respectively. Similarly, in the UK Biobank study, BANNs estimates the PVE for HDL and LDL to be 0.12 and 0.06, respectively. In general, these values are lower than what is typically reported in the literature for these complex phenotypes (PVE $\geq$ 27% for HDL and PVE $\geq$ 21% for LDL, respectively) [115]. Exploring different prior assumptions and considering other (more scalable) ways to carry out approximate Bayesian inference is something to consider for future work [116]. For example, the Bayesian sparse linear mixed modeling (BSLMM) framework [16, 117, 118] extends the traditional spike-and-slab prior and could provide a useful, yet alternative, hierarchical specification for BANNs.

There are several other potential extensions for the BANNs framework. First, in the current study, we only consider a single hidden layer based on the annotations of gene boundaries and intergenic region between genes. One natural direction for future work would be to a take more of a deep learning approach by including additional hidden layers to the neural network where genes are grouped based on signaling pathways or other functional ontologies (e.g., transcription factor binding). This would involve integrating information from curated databases such as MSigDB [119, 120] or CADD [121]. Second, while BANNs is able to account for nonlinear genetic effects, it cannot be used to directly identify the component (i.e., linear vs. nonlinear) that is driving individual SNP or SNP-set associations. A key part of our future work is learning how to disentangle this information and provide detailed summaries of variant-level and gene-by-gene interaction effects [122]. Third, the current BANNs model only takes in genetic information and, in its current form, ignores unobserved environmental covariates (and potential gene-by-environment or G×E interactions) that explain

variation in complex traits. In the future, we would like to expand the framework to also take in covariates as fixed effects in the model. Fourth, we have only focused on analyzing one phenotype at a time in this study. However, many previous studies have extensively shown that modeling multiple phenotypes can often dramatically increase power [123]. Therefore, it would be interesting to extend the BANNs framework to take advantage of phenotype correlations to identify pleiotropic epistatic effects. Modeling strategies based on the multivariate linear mixed model (mvLMM) [124] and matrix variate Gaussian process (mvGP) [125] could be helpful here.

As a final avenue for future work, we only focused on applying BANNs to quantitative traits. For studies interested in extending this approach to binary traits (i.e., case-control studies), one might be tempted to simply place a sigmoid or logistic link function on the penultimate layer of the neural network. Indeed, this would allow the BANNs framework to be expressed as a (nonlinear) logistic classification model which is an approach that has been well-established in the statistics literature [126–128]. Unfortunately, it is not straightforward to define broad-sense heritability under the traditional logistic regression framework. As one alternative, we could implement a penalized quasi-likelihood approach [129] which has been shown to enable effective heritability estimation and differential analyses using the generalized linear mixed model framework. As a second alternative, the liability threshold model avoids issues by assuming that binary traits can be modeled via continuous latent liability scores [130–132]. Therefore, a potentially effective way to extend BANNs to case-control studies would be to develop a two-step algorithmic procedure where: in the first step, we find the posterior mean of the liability scores be using existing software packages and then, in the second step, treat those empirical liability estimates as observed traits in the neural network. Regardless of the modeling strategy, new algorithms are likely needed to maximize the appropriateness of BANNs for non-continuous phenotypes.

Materials and methods

Annotations

We used the NCBI's Reference Sequence (RefSeq) database in the UCSC Genome Browser [50] to annotate SNPs with appropriate SNP-sets. In the main text, we define genes with boundaries in two ways: (*a*) we use the UCSC gene boundary definitions directly, or (*b*) we augment the gene boundaries by adding SNPs within a ±500 kilobase (kb) buffer to account for possible regulatory elements. Genes with only 1 SNP within their boundary were excluded from either analysis. 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 kept for analysis using the UCSC boundaries and a total of $G = 35{,}849$ SNP-sets were kept for analysis when including the 500kb buffer.

Biologically annotated neural networks

Consider a genome-wide association (GWA) study with N individuals. We have an N-dimensional vector of quantitative traits $\mathbf{y}$, an $N \times J$ matrix of genotypes $\mathbf{X}$, with J denoting the number of single nucleotide polymorphisms (SNPs) encoded as {0, 1, 2} copies of a reference allele at each locus, and a list of G-predefined SNP-sets $\{\mathcal{S}_1, \ldots, \mathcal{S}_G\}$ (Fig 1A). Let each SNP-set g represent a known collection of annotated SNPs $j \in \mathcal{S}_g$ with cardinality $|\mathcal{S}_g|$. For example, $\mathcal{S}_g$ may include SNPs within the regulatory region of a gene. The BANNs framework assumes a partially connected Bayesian neural network architecture based on SNP-set annotations to learn the phenotype of interest for each observation in the data (Fig 1B). Formally, we specify

this network as a nonlinear regression model (Fig 1C)

$$\mathbf{y} = \sum_{g=1}^{G} h(\mathbf{X}_g \boldsymbol{\theta}_g + \mathbf{1} b_g^{(1)}) w_g + \mathbf{1} b^{(2)}, \tag{4}$$

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 [49], where $h(\boldsymbol{x}) = \boldsymbol{x}$ if $\boldsymbol{x} > \mathbf{0}$ and $0.01\boldsymbol{x}$ otherwise. Note that Eq (4) can be seen as a nonlinear take on classic integrative and structural regression models [22, 26, 133–136] frequently used in GWA analyses.

A key methodological aspect in the BANNs framework is to treat the weights of the input (θ_j) and hidden layers (w_g) as random variables. This, in part, enables us to perform interpretable association mapping on both SNPs and SNP-sets, simultaneously. For the weights on the input layer, our goal is to approximate a wide range of possible SNP-level effect size distributions underlying complex traits. To this end, we assume that SNP-level effects follow a K-mixture of normal distributions [10, 52–54]

$$\theta_j \sim \pi_\theta \sum_{k=1}^{K} \eta_{\theta k} \mathcal{N}(0, \sigma_{\theta k}^2) + (1 - \pi_\theta)\delta_0, \qquad \log(\pi_\theta) \sim \mathcal{U}(-\log(J), \log(1)) \tag{5}$$

where δ_0 is a point mass at zero; $\boldsymbol{\sigma}_\theta^2 = (\sigma_{\theta 1}^2, \ldots, \sigma_{\theta K}^2)$ are variance of the K-nonzero mixture components; $\boldsymbol{\eta}_\theta = (\eta_{\theta 1}, \ldots, \eta_{\theta K})$ represents the marginal (unconditional) probability that a randomly selected SNP belongs to the k-th mixture component such that $\sum_k \eta_{\theta k} = 1$; and π_θ denotes the total proportion of SNPs that have a nonzero effect on the trait of interest. We allow sequential fractions of SNPs ($\eta_{\theta 1}, \ldots, \eta_{\theta K}$) to correspond to distinctly smaller effects ($\sigma_{\theta 1}^2 > \cdots > \sigma_{\theta K}^2$) [53]. 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 results in the main text, we fix $K = 3$ for computational reasons. This corresponds to the hypothesis that SNPs can have large, moderate, and small nonzero effects on phenotypic variation [28]. We assume a uniform prior on $\log \pi_\theta$ to coincide with the observation that the number of SNPs in each of these categories can vary greatly depending on how heritability is distributed across the genome [16, 65] (see S1 Text).

For inference on the hidden layer, we assume that enriched SNP-sets contain at least one SNP with a nonzero effect. This criterion is formulated by placing a spike and slab prior on the hidden layer weights

$$w_g \sim \pi_w \mathcal{N}(0, \sigma_w^2) + (1 - \pi_w)\delta_0, \qquad \log(\pi_w) \sim \mathcal{U}(-\log(G), \log(1)) \tag{6}$$

where, in addition to previous notation, the parameter π_w denotes the total proportion of annotated SNP-sets that are enriched for the trait of interest. Given the structural form of the joint likelihood in Eq (4), the magnitude of association for a SNP-set will be directly influenced by the effect size distribution of the SNPs it contains.

We use a variational Bayesian algorithm to estimate all model parameters (S1 Text). As the BANNs model is trained, the posterior mean for the weights of non-associated SNP and SNP-sets will trend towards zero as the neural network attempts to identify a subset of neurons that

are associated with the phenotype. We use posterior inclusion probabilities (PIPs) as a general summaries of evidence for SNPs and SNP-sets being associated with phenotypic variation. Here, we respectively define

$$\text{PIP}(j) \equiv \Pr[\theta_j \neq 0 \,|\, \mathbf{y}, \mathbf{X}], \qquad \text{PIP}(g) \equiv \Pr[w_g \neq 0 \,|\, \mathbf{y}, \mathbf{X}, \boldsymbol{\theta}_g] \tag{7}$$

where, again for the latter, the enrichment of SNP-sets is conditioned on the association of individual SNPs. Overall, the Bayesian formulation in the BANNs framework enables network sparsity to be targeted for GWA applications through contextually motivated sparse shrinkage prior distributions in Eqs (5) and (6). Moreover, posterior inference on PIP(j) and PIP(g) detail the degree to which nonzero weights occur.

Posterior computation with variational inference

We combine the likelihood in Eq (4) and the prior distributions in Eqs (5) and (7) to perform Bayesian inference. With the size of high-throughput GWA datasets, it is less feasible to implement traditional Markov Chain Monte Carlo (MCMC) algorithms due to the large dimensionality of the parameter space. For model fitting, we modify a previously established variational expectation-maximization (EM) algorithm [55, 56] for integrative neural network parameter estimation. The overall goal of variational inference is to approximate the true posterior distribution for network parameters with a "best match" distribution from an approximating family [63]. The EM algorithm we use aims to minimize the Kullback-Leibler divergence between the exact and approximate posterior distributions.

To compute the variational approximations, we make the mean-field assumption that the true posterior can be "fully-factorized" [137]. The algorithm then follows three general steps. First, we assign exchangeable uniform hyper-priors over a grid of values on the log-scale for π_θ and π_w [55]. Next, we iterate through each combination of hyper-parameter values and compute variational updates for the other parameters using co-ordinate ascent. Lastly, we empirically compute (approximate) posterior values for the network connections ($\boldsymbol{\theta}$, $\mathbf{w}$) and their corresponding inclusion probabilities by marginalizing over the different hyper-parameter combinations. 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 sampling weights [138]. Throughout the model fitting procedure, we assess two different lower bounds for the input and hidden layers to check convergence of the algorithm. The first lower bound is maximized with respect to the SNP-level effects on the observed trait of interest; while, the second lower bound focuses on the SNP-set level enrichments. The software code first iterates over the "inner" lower bound until convergence and then uses those weights to compute the hidden neurons and maximize the "outer" lower bound. Detailed steps in the variational EM algorithm, explicit co-ordinate ascent updates for network parameters, and pseudocode are given in Supporting information.

Parameters in the variational EM algorithm are initialized by taking a random draws from their assumed prior distributions. 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 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.

Extensions to summary statistics

The BANNs framework also models summary statistics in the event that individual-level genotype and phenotype data are not accessible. Here, the software takes alternative inputs: GWA

marginal effect size estimates $\hat{\boldsymbol{\theta}}$ as the response variable, and an empirical linkage disequilibrium (LD) matrix $\mathbf{R}$ as the design matrix. In the main text, we refer to this version of the method as the BANN-SS model. We assume that GWA summary statistics are derived from the following generative linear model for complex traits

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{e}, \qquad \mathbf{e} \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbf{I}) \tag{8}$$

where $\mathbf{e}$ is a normally distributed error term with mean zero and scaled variance τ^2, and $\mathbf{I}$ is an $N \times N$ identity matrix. For every j-th SNP, the ordinary least squares (OLS) estimates are based on the generative model $\hat{\theta}_j = (\mathbf{x}_j^\top \mathbf{x}_j)^{-1}\mathbf{x}_j^\top \mathbf{y}$, where $\mathbf{x}_j$ is the j-th column of the individual-level genotype matrix $\mathbf{X}$ and $\hat{\theta}_j$ is the j-th entry of the vector $\hat{\boldsymbol{\theta}}$. In practice, the LD matrix $\mathbf{R}$ can be empirically estimated directly from the in-sample GWA study data or from external data (e.g., using an LD map from a population with genomic ancestry similar to individuals in the orginial study). Note that all results presented in the main text are based on estimating $\mathbf{R}$ with the in-sample genotype data. The BANN-SS model treats the observed OLS estimates and LD matrix as "proxies" for the unobserved phenotype and genotypes, respectively. Specifically, for large sample size N, we consider the asymptotic relationship between the expectation of the observed GWA effect size estimates $\hat{\boldsymbol{\theta}}$ and the true coefficient values $\boldsymbol{\theta}$ is [28, 45, 53, 139]

$$\mathbb{E}[\hat{\theta}_j] = \sum_{j'=1}^{J} r(\mathbf{x}_j, \mathbf{x}_{j'})\theta_{j'} \tag{9}$$

where $r(\mathbf{x}_j, \mathbf{x}_{j'})$ denotes the correlation coefficient between SNPs $\mathbf{x}_j$ and $\mathbf{x}_{j'}$. The above resembles a high-dimensional regression model with the OLS effect sizes $\hat{\boldsymbol{\theta}}$ as the response variables, the LD matrix $\mathbf{R}$ as the design matrix, and the true coefficients $\boldsymbol{\theta}$ being the SNP-level effects that generated the phenotype. Note that this observation is also utilized by other GWA summary-level statistical methods (e.g., CAVIAR [45] and RSS [26, 62]). With this relationship in mind, the BANN-SS framework implements the following sparse nonlinear regression for inferring multi-scale genomic effects from summary statistics (S1 Fig)

$$\hat{\boldsymbol{\theta}} = \sum_{g=1}^{G} h(\mathbf{R}_g\boldsymbol{\theta}_g + \mathbf{1}b_g^{(1)})w_g + \mathbf{1}b^{(2)}, \tag{10}$$

where, in addition to previous notation, $\mathbf{R}_g$ is the subset of the LD matrix involving all SNPs annotated for the g-th SNP-set. Using the rewritten joint likelihood in Eq (10), posterior Bayesian inference for the parameters in the BANN-SS model directly mirrors the procedure used when we have access to individual-level data (i.e., as described previously in Eqs (5)–(7) and given in detail in the Supporting information). Again, we use measurements PIP(j) and PIP(g) to summarize whether the true SNP-level effects and aggregated effects on the SNP-set level are statistically associated with the trait of interest.

Simulation studies

We implement a simulation scheme to generate quantitative traits under multiple genetic architectures by using real genotype data on chromosome 1 from individuals of European ancestry in the UK Biobank. First, we randomly select a subset of associated SNP-sets (i.e., collections of genomic regions) and assume that complex traits are generated via the linear

regression model

$$\mathbf{y} = \sum_{c \in \mathcal{C}} \mathbf{x}_c \theta_c + \mathbf{W}\boldsymbol{\varphi} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \tau^2 \mathbf{I}), \tag{11}$$

where $\mathbf{y}$ is an N-dimensional vector containing all the phenotypes; $\mathcal{C}$ represents the set of causal SNPs contained within the associated SNP-sets; $\mathbf{x}_c$ is the genotype for the c-th causal SNP encoded as 0, 1, or 2 copies of a reference allele; θ_c is the additive effect size for the c-th SNP; $\mathbf{W}$ is an $N \times E$ matrix which holds all pairwise interactions between the causal SNPs with corresponding effects $\boldsymbol{\varphi}$; and $\boldsymbol{\varepsilon}$ is an N-dimensional vector of environmental noise. The total phenotypic variance is assumed $\mathbb{V}[\mathbf{y}] = 1$. The additive and interaction effect sizes of SNPs in associated SNP-sets are randomly drawn from standard normal distributions and then rescaled so they explain a fixed proportion of the broad-sense heritability $\mathbb{V}[\sum \mathbf{x}_c \theta_c] + \mathbb{V}[\mathbf{W}\boldsymbol{\varphi}] = H^2$. Lastly the environment noise is rescaled such that $\mathbb{V}[\boldsymbol{\varepsilon}] = 1 - H^2$. The full genotype matrix and phenotypic vector are given to the BANNs model and all other competing methods that require individual-level data. For the BANN-SS model and other competing methods that take GWA summary statistics, we fit a single-SNP univariate linear model via ordinary least squares (OLS) to obtain: coefficient estimates $\hat{\theta}_j = (\mathbf{x}_j^\top \mathbf{x}_j)^{-1} \mathbf{x}_j^\top \mathbf{y}$, standard errors $\hat{s}_j^2 = J^{-1}(\mathbf{y} - \mathbf{x}_j \hat{\theta}_j)^\top (\mathbf{y} - \mathbf{x}_j \hat{\theta}_j) / \mathbf{x}_j^\top \mathbf{x}_j$, and P-values for all SNPs in the data. We also obtain an empirical estimate of the linkage disequilibrium (LD) matrix for these methods $\mathbf{R}$, which we compute directly from the full in-sample genotype matrix. Given different model parameters, we simulate data mirroring a wide range of genetic architectures (S1 Text).

URLs

Biologically annotated neural networks (BANNs) software, https://github.com/lcrawlab/BANNs; UK Biobank, https://www.ukbiobank.ac.uk; Database of Genotypes and Phenotypes (dbGaP), https://www.ncbi.nlm.nih.gov/gap; Framingham Heart Study (FHS), https://www.ncbi.nlm.nih.gov/gap; NHGRI-EBI GWAS Catalog, https://www.ebi.ac.uk/gwas/; UCSC Genome Browser, https://genome.ucsc.edu/index.html; Enrichr software, http://amp.pharm.mssm.edu/Enrichr/; Wellcome Trust Centre for Human Genetics, http://mtweb.cs.ucl.ac.uk/mus/www/mouse/index.shtml; Mouse Genome Informatics database, http://www.informatics.jax.org; CAusal Variants Identification in Associated Regions (CAVIAR) software, http://genetics.cs.ucla.edu/caviar/; Efficient variable selection using summary data from GWA studies (FINEMAP) software, http://www.christianbenner.com; Generalized Berk-Jones (GBJ) test for set-based inference software, https://cran.r-project.org/web/packages/GBJ/; Gene Set Enrichment Analysis (GSEA) software, https://www.nr.no/en/projects/software-genomics; SNP-set (Sequence) Kernel Association Test (SKAT) software, https://www.hsph.harvard.edu/skat; Sum of Single Effects (SuSiE) variable selection software, https://github.com/stephenslab/susieR; Multi-marker Analysis of GenoMic Annotation (MAGMA) software, https://ctg.cncr.nl/software/magma; Precise, Efficient Gene Association Score Using SNPs (PEGASUS) software, https://github.com/ramachandran-lab/PEGASUS; and Regression with Summary Statistics (RSS) enrichment software, https://github.com/stephenslab/rss.

Supporting information

S1 Fig. 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

$\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_J)$; 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 regression 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 $\text{PIP}(j) \equiv \Pr[\theta_j \neq 0 \mid \mathbf{y}, \mathbf{X}]$ and $\text{PIP}(g) \equiv \Pr[w_g \neq 0 \mid \mathbf{y}, \mathbf{X}, \boldsymbol{\theta}_g]$ 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.
(PDF)

S2 Fig. 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 (British cohort). 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 1% and 10% 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 [45], SuSiE [46], and FINEMAP [44]. 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 [26], PEGASUS [25], GBJ [27], SKAT [21], GSEA [43], and MAGMA [23]. Note that the upper limit of the x-axis has been truncated at 0.1. All results are based on 100 replicates (see S1 Text).
(PDF)

S3 Fig. 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 structure (European cohort). 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 1% and 10% 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 [45], SuSiE [46], and FINEMAP [44]. 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 [26], PEGASUS [25], GBJ [27], SKAT [21], GSEA [43], and MAGMA [23]. Note that the upper limit of the x-axis has been truncated at 0.1. All results are based on 100 replicates (see S1 Text).
(PDF)

S4 Fig. 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 structure (European cohort). 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 1% and 10% 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 [45], SuSiE [46], and FINEMAP [44]. 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 [26], PEGASUS [25], GBJ [27], SKAT [21], GSEA [43], and MAGMA [23]. Note that the upper limit of the x-axis has been truncated at 0.1. All results are based on 100 replicates (see S1 Text).
(PDF)

S5 Fig. 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 (British cohort). 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 1% and 10% 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 [45], SuSiE [46], and FINEMAP [44]. 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 [26], PEGASUS [25], GBJ [27], SKAT [21], GSEA [43], and MAGMA [23]. Note that the upper limit of the x-axis has been truncated at 0.1. All results

are based on 100 replicates (see S1 Text).
(PDF)

S6 Fig. 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 (British cohort). 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 1% and 10% 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 [45], SuSiE [46], and FINEMAP [44]. 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 [26], PEGASUS [25], GBJ [27], SKAT [21], GSEA [43], and MAGMA [23]. Note that the upper limit of the x-axis has been truncated at 0.1. All results are based on 100 replicates (see S1 Text).
(PDF)

S7 Fig. 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 structure (European cohort). 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 1% and 10% 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 [45], SuSiE [46], and FINEMAP [44]. 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 [26], PEGASUS [25], GBJ [27], SKAT [21], GSEA [43], and MAGMA [23]. Note that the upper limit of the x-axis has been truncated at 0.1. All results are based on 100 replicates (see S1 Text).
(PDF)

S8 Fig. 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 structure (European cohort). 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 1% and 10% 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 [45], SuSiE [46], and FINEMAP [44]. 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 [26], PEGASUS [25], GBJ [27], SKAT [21], GSEA [43], and MAGMA [23]. Note that the upper limit of the x-axis has been truncated at 0.1. All results are based on 100 replicates (see S1 Text).
(PDF)

S9 Fig. 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 (British cohort). 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 1% and 10% 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 [45], SuSiE [46], and FINEMAP [44]. 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 [26], PEGASUS [25], GBJ [27], SKAT [21], GSEA [43], and MAGMA [23]. 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 S1 Text).
(PDF)

S10 Fig. 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 (British cohort). 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 1% and 10% 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 [45], SuSiE [46], and FINEMAP [44]. 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 [26], PEGASUS [25], GBJ [27], SKAT [21], GSEA [43], and MAGMA [23]. 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 S1 Text).
(PDF)

S11 Fig. 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 structure (European cohort). 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 1% and 10% 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 [45], SuSiE [46], and FINEMAP [44]. 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 [26], PEGASUS [25], GBJ [27], SKAT [21], GSEA [43], and MAGMA [23]. 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 S1 Text).
(PDF)

S12 Fig. 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 structure (European cohort). 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 1% and 10% 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 [45], SuSiE [46], and FINEMAP [44]. 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 [26], PEGASUS [25], GBJ [27], SKAT [21], GSEA [43], and MAGMA [23]. 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 S1 Text).
(PDF)

S13 Fig. 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 (British cohort). 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 1% and 10% 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 [45], SuSiE [46], and FINEMAP [44]. 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 [26], PEGASUS [25], GBJ [27], SKAT [21], GSEA [43], and MAGMA [23]. 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 S1 Text).
(PDF)

S14 Fig. 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 (British cohort). 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 1% and 10% 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 [45], SuSiE [46], and FINEMAP [44]. 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 [26], PEGASUS [25], GBJ [27], SKAT [21], GSEA [43], and MAGMA [23]. 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 S1 Text).
(PDF)

S15 Fig. 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 structure (European cohort). 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 1% and 10% 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 [45], SuSiE [46], and FINEMAP [44]. 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 [26], PEGASUS [25], GBJ [27], SKAT [21], GSEA [43], and MAGMA [23]. 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 S1 Text).
(PDF)

S16 Fig. 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 structure (European cohort). 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 1% and 10% 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 [45], SuSiE [46], and FINEMAP [44]. 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 [26], PEGASUS [25], GBJ [27], SKAT [21], GSEA [43], and MAGMA [23]. 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 S1 Text).
(PDF)

S17 Fig. Scatter plots comparing how the integrative neural network training procedure enables the ability to identify associated SNPs and enriched SNP-sets in simulations (British cohort). 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 ∼ 1% and ∼ 10% 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 [46] and **(B, D)** RSS [26] 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) [57]. 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 S1 Text).
(PDF)

S18 Fig. 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 structure (European cohort). 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 1% and 10% 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 [46] and **(B, D)** RSS [26] 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) [57]. 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 S1 Text).
(PDF)

S19 Fig. 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 structure (European cohort). 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 1% and 10% 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 [46] and **(B, D)** RSS [26] 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) [57]. 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 S1 Text).
(PDF)

S20 Fig. Scatter plots comparing how the integrative neural network training procedure enables the ability to identify associated SNPs and enriched SNP-sets in simulations (British cohort). 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 1% and 10% 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 [46] and **(B, D)** RSS [26] 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) [57]. 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 S1 Text).
(PDF)

S21 Fig. Scatter plots comparing how the integrative neural network training procedure enables the ability to identify associated SNPs and enriched SNP-sets in simulations (British cohort). 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 1% and 10% 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 [46] and **(B, D)** RSS [26] 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) [57]. 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 S1 Text).
(PDF)

S22 Fig. 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 structure (European cohort). 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 1% and 10% 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 [46] and **(B, D)** RSS [26] 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) [57]. 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 S1 Text).
(PDF)

S23 Fig. 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 structure (European cohort). 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 1% and 10% 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 [46] and **(B, D)** RSS [26] 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) [57]. 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 S1 Text).
(PDF)

S24 Fig. Assessments of posterior inclusion probability (PIP) calibration for both SNP-level associations and enrichment of SNP-sets. This experiment follows largely from previous work [46, 65]. Here, SNPs and SNP-sets across simulations are grouped into bins according to their reported PIPs (using 20 equally spaced bins, from 0 to 1). The plots show the average PIP for each bin against the proportion of causal SNPs or SNP-sets in that bin. A well calibrated method should produce points near the x-axis = y-axis line (i.e., the diagonal red lines). Gray error bars show ±2 standard errors. Panel **(A, B)** shows the comparison of BANNs SNP layer with SuSiE [46], and **(C, D)** shows the comparison of BANNs SNP-set layer with RSS [26]. While the inclusion probabilities are not perfectly calibrated for any of the methods, the empirical power and false discovery rate (FDR) above the "median probability criterion" (i.e., PIPs for SNPs and SNP-sets greater than 0.5) [57] are still reasonably well controlled (see S1–S8 Tables). We hypothesize that these calibration results are due both to consequences of both variational inference and the level of polygenicity with which we simulated synthetic phenotypes.
(PDF)

S25 Fig. Receiver operating characteristic (ROC) curves comparing the performance of the BANNs models with different modifications via an ablation test. To investigate how choices in the model setup contribute to variable selection, we performed an "ablation analysis" where we modified parts of the BANNs framework independently and observed their direct effect on model performance (see S1 Text). We considered two different modifications to our model: (1) removing the activation function and training a fully linear hierarchical model, and (2) removing the approximate Bayesian model averaging approach and updating the probabilities π_θ and π_w as additional parameters in the variational EM algorithm. In the normal BANNs setup, we initialize L different models with varying priors for inclusion probabilities specified over a grid $\{\pi_\theta^{(1)}, \ldots, \pi_\theta^{(L)}\} \in [1/J, 1]$ and $\{\pi_w^{(1)}, \ldots, \pi_w^{(L)}\} \in [1/G, 1]$, respectively. However, in the case of the latter ablation modification, we initialize $\pi_\theta = 1/J$ and $\pi_w = 1/G$ as an analogy to the "single causal variant" assumption frequently used in fine mapping [46]. Next, we update their values in the M-step of the algorithm according to the following analytic expressions: **(A, C)** $\pi_\theta/1 - \pi_\theta = \sum_j \sum_k \alpha_{jk}/\sum_j \sum_k (1 - \alpha_{jk})$, and **(B, D)** $\pi_w/1 - \pi_w = \sum_g \alpha_g/\sum_g (1 - \alpha_g)$. Results here are shown using simulations with the self-identified "white British" ancestry cohort from the UK Biobank on synthetic traits that have broad-sense heritability $H^2 = 0.6$ with sparse genetic architecture. Each plot combines results from 100 simulated replicates (see S1 Text).
(PDF)

S26 Fig. 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 sparse genetic effects (both additive and non-additive) [16]. 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. Panels **(A, C)** show heritability estimates on simulations with genetic data from individuals who self-identify as being of "white British" ancestry in the UK Biobank; while, panels **(B, D)** show heritability estimates on simulations with genetic data from individuals who more broadly identify as being of European ancestry. True heritability 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.
(PDF)

S27 Fig. 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 sparse genetic effects (both additive and non-additive) [16]. 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. Panels **(A, C)** show heritability estimates on simulations with genetic data from individuals who self-identify as being of "white British" ancestry in the UK Biobank; while, panels **(B, D)** show heritability estimates on simulations with genetic data from individuals who more broadly identify as being of European ancestry. True heritability 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.
(PDF)

S28 Fig. 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 [57]. 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.
(PDF)

S29 Fig. 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(g) $\geq$ 0.5 (i.e., the "median probability model" threshold [57]). We take these significant SNP-sets and conduct "gene set enrichment analysis" using Enrichr [90, 91] to identify the categories they overrepresent in **(A, B)** the database of Genotypes and Phenotypes (dbGaP) and **(C, D)** the GWAS Catalog (2019). 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 = 0.976. There have been hypotheses connecting LDL to cognitive traits [140, 141], and *APOB* has been shown to be related to cerebrospinal fluid and memory [142–144]. Therefore, we argue that results in panel **(D)** are also relevant.
(PDF)

S30 Fig. 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. In this analysis, each gene boundary annotation is modfied by adding SNPs within a ±500 kilobase (kb) buffer to account for possible regulatory elements. 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 (g) $\geq$ 0.5 (i.e., the "median probability model" threshold [57]). We take these significant SNP-sets and conduct "gene set enrichment analysis" using Enrichr [90, 91] to identify the categories they overrepresent in **(A, B)** the database of Genotypes and Phenotypes (dbGaP) and **(C, D)** the GWAS Catalog (2019). Nearly all enriched categories are related with **(A, C)** HDL and **(B, D)** LDL, respectively.
(PDF)

S31 Fig. Manhattan plot of variant-level association 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 [31]. 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 [57]. 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 S31 Fig. Stars (★) denote SNPs and SNP-sets that replicate findings from our analyses of HDL and LDL in the Framingham Heart Study (see Fig 4 in the main text).
(PDF)

S32 Fig. 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 [31]. 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(g) $\geq$ 0.5 (i.e., the "median probability model" threshold [57]). We take these significant SNP-sets and conduct "gene set enrichment analysis" using Enrichr [90, 91] to identify the categories they overrepresent in **(A, B)** the database of Genotypes and Phenotypes (dbGaP) and **(C, D)** the GWAS Catalog (2019). 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 [140, 141], and *APOB* has been shown to be related to cerebrospinal fluid and memory [142–144]. Therefore, we argue that results in panel **(B)** are also relevant (a similar argument can be made for S33 Fig).
(PDF)

S33 Fig. 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 [31]. 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. In this analysis, each gene boundary annotation is modfied by adding SNPs within a ±500 kilobase (kb) buffer to account for possible regulatory elements. 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(g) $\geq$ 0.5 (i.e., the "median probability model" threshold [57]). We take these significant SNP-sets and conduct "gene set enrichment analysis" using Enrichr [90, 91] to identify the categories they overrepresent in **(A, B)** the database of Genotypes and Phenotypes (dbGaP) and **(C, D)** the GWAS Catalog (2019). Note that for panel **(A)**, BANNs did not find many enriched SNP-sets with PIPs meeting the "median probability model" threshold and so we used a lower SNP-set threshold (PIP $\geq$ 0.1) to enable

Enrichr to find associated dbGaP categories.
(PDF)

S1 Table. 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 1% and 10% of all SNPs located within the enriched SNP-sets, respectively. **(Top)** Competing SNP-level mapping approaches include: CAVIAR [45], SuSiE [46], and FINEMAP [44]. 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 [26], PEGASUS [25], GBJ [27], SKAT [21], GSEA [43], and MAGMA [23]. 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) [57]. 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.
(PDF)

S2 Table. 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 1% and 10% of all SNPs located within the enriched SNP-sets, respectively. **(Top)** Competing SNP-level mapping approaches include: CAVIAR [45], SuSiE [46], and FINEMAP [44]. 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 [26], PEGASUS [25], GBJ [27], SKAT [21], GSEA [43], and MAGMA [23]. 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) [57]. 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.
(PDF)

S3 Table. 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 structure (European cohort). 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 1% and 10% of all SNPs located within the enriched SNP-sets, respectively. **(Top)** Competing SNP-level mapping approaches include: CAVIAR [45], SuSiE [46], and FINEMAP [44]. 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 [26], PEGASUS [25], GBJ [27], SKAT [21], GSEA [43], and MAGMA [23]. 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) [57]. 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.
(PDF)

S4 Table. 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 structure (European cohort). 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 1% and 10% of all SNPs located within the enriched SNP-sets, respectively. **(Top)** Competing SNP-level mapping approaches include: CAVIAR [45], SuSiE [46], and FINEMAP [44]. 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 [26], PEGASUS [25], GBJ [27], SKAT [21], GSEA [43], and MAGMA [23]. 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) [57]. 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.
(PDF)

S5 Table. 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 1% and 10% of all SNPs located within the enriched SNP-sets, respectively. **(Top)** Competing SNP-level mapping approaches include: CAVIAR [45], SuSiE [46], and FINEMAP [44]. 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

[26], PEGASUS [25], GBJ [27], SKAT [21], GSEA [43], and MAGMA [23]. 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) [57]. 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.
(PDF)

S6 Table. 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 1% and 10% of all SNPs located within the enriched SNP-sets, respectively. **(Top)** Competing SNP-level mapping approaches include: CAVIAR [45], SuSiE [46], and FINEMAP [44]. 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 [26], PEGASUS [25], GBJ [27], SKAT [21], GSEA [43], and MAGMA [23]. 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) [57]. 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.
(PDF)

S7 Table. 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 structure (European cohort). 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 1% and 10% of all SNPs located within the enriched SNP-sets, respectively. **(Top)** Competing SNP-level mapping approaches include: CAVIAR [45], SuSiE [46], and FINEMAP [44]. 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 [26], PEGASUS [25], GBJ [27], SKAT [21], GSEA [43], and MAGMA [23]. 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) [57]. 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.
(PDF)

S8 Table. 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 structure (European cohort). 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 1% and 10% of all SNPs located within the enriched SNP-sets, respectively. **(Top)** Competing SNP-level mapping approaches include: CAVIAR [45], SuSiE [46], and FINEMAP [44]. 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 [26], PEGASUS [25], GBJ [27], SKAT [21], GSEA [43], and MAGMA [23]. 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) [57]. 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.
(PDF)

S9 Table. 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 [45], SuSiE [46], and FINEMAP [44]. 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.
(PDF)

S10 Table. 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 [26], PEGASUS [25], GBJ [27], SKAT [21], GSEA [43], and MAGMA [23]. 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.
(PDF)

S11 Table. 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 (Supporting information). 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 [57] (i.e., PIP ≥ 0.5). Page #1 provides the variant-level association mapping results with columns corresponding to: (1) chromosome; (2) SNP ID; (3) chromosomal position in base-pair (bp) coordinates; (4) SNP PIP; and (5) SuSiE PIP, which corresponds to SNP-level posterior inclusion probabilities computed by SuSiE [46]. 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) RSS PIP, which corresponds to the posterior inclusion probabilities computed by RSS [26]; (7) the number of SNPs that have been annotated within each SNP-set; (8) the "top" associated SNP within each SNP-set; (9) the PIP of each top SNP.
(XLSX)

S12 Table. 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 (Supporting information). 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 [57] (i.e., PIP ≥ 0.5). Page #1 provides the variant-level association mapping results with columns corresponding to: (1) chromosome; (2) SNP ID; (3) chromosomal position in base-pair (bp) coordinates; (4) SNP PIP; and (5) SuSiE PIP, which corresponds to SNP-level posterior inclusion probabilities computed by SuSiE [46]. 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) RSS PIP, which corresponds to the posterior inclusion probabilities computed by RSS [26]; (7) the number of SNPs that have been annotated within each SNP-set; (8) the "top" associated SNP within each SNP-set; (9) the PIP of each top SNP.
(XLSX)

S13 Table. 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 (Supporting information). 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 [57] (i.e., PIP ≥ 0.5). Page #1 provides the variant-level association mapping results with columns corresponding to: (1) chromosome; (2) SNP ID; (3) chromosomal position in base-pair (bp) coordinates; (4) SNP PIP; and (5) SuSiE PIP, which corresponds to SNP-level posterior inclusion probabilities computed by SuSiE [46]. 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) RSS PIP, which corresponds to the posterior inclusion probabilities computed by RSS [26]; (7) the number of SNPs that have been annotated within each SNP-set; (8) the "top" associated SNP within each SNP-set; (9) the PIP of each top SNP.
(XLSX)

S14 Table. 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 (Supporting information). 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 [57] (i.e., PIP ≥ 0.5). Page #1 provides the variant-level association mapping results with columns corresponding to: (1) chromosome; (2) SNP ID; (3) chromosomal position in base-pair (bp) coordinates; (4) SNP PIP; and (5) SuSiE PIP, which corresponds to SNP-level posterior inclusion probabilities computed by SuSiE [46]. 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) RSS PIP, which corresponds to the posterior inclusion probabilities computed by RSS [26]; (7) the number of SNPs that have been annotated within each SNP-set; (8) the "top" associated SNP within each SNP-set; (9) the PIP of each top SNP.
(XLSX)

S15 Table. 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 (Supporting information). 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 [57] (i.e., PIP $\geq$ 0.5). Page #1 provides the variant-level association mapping results with columns corresponding to: (1) chromosome; (2) SNP ID; (3) chromosomal position in base-pair (bp) coordinates; (4) SNP PIP; and (5) SuSiE PIP, which corresponds to SNP-level posterior inclusion probabilities computed by SuSiE [46]. 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) RSS PIP, which corresponds to the posterior inclusion probabilities computed by RSS [26]; (7) the number of SNPs that have been annotated within each SNP-set; (8) the "top" associated SNP within each SNP-set; (9) the PIP of each top SNP.
(XLSX)

S16 Table. 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 (Supporting information). 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 [57] (i.e., PIP $\geq$ 0.5). Page #1 provides the variant-level association mapping results with columns corresponding to: (1) chromosome; (2) SNP ID; (3) chromosomal position in base-pair (bp) coordinates; (4) SNP PIP; and (5) SuSiE PIP, which corresponds to SNP-level posterior inclusion probabilities computed by SuSiE [46]. 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) RSS PIP, which corresponds to the posterior inclusion probabilities computed by RSS [26]; (7) the number of SNPs that have been annotated within each SNP-set; (8) the "top" associated SNP within each SNP-set; (9) the PIP of each top SNP.
(XLSX)

S17 Table. Notable enriched SNP-sets after applying the BANNs framework to high-density and low-density lipoprotein (HDL and LDL, respectively) traits in the Framingham Heart Study [48] where each SNP-set annotation has been augmented with a ±500 kilobase (kb) buffer to account for possible regulatory elements. 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 table. 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 enriched if it has a PIP(g) $\geq$ 0.5 (i.e., the "median probability model" threshold [57]). We report the "top" associated SNP within each region and its corresponding PIP(j). We also report the corresponding SNP and SNP-set level results after running SuSiE [46] and RSS [26] on these same traits, respectively. The last column details references and literature sources that have previously suggested some level of association or enrichment between the each genomic region and the traits of interest. See S18 and S19 Tables for the complete list of SNP and SNP-set level results. ♣: SNPs and SNP-sets replicated in an independent analysis of ten thousand

randomly sampled individuals of European ancestry from the UK Biobank [31].
(PDF)

S18 Table. 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 [57] (i.e., PIP ≥ 0.5). Page #1 provides the variant-level association mapping results with columns corresponding to: (1) chromosome; (2) SNP ID; (3) chromosomal position in base-pair (bp) coordinates; (4) SNP PIP; and (5) SuSiE PIP, which corresponds to SNP-level posterior inclusion probabilities computed by SuSiE [46]. 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) RSS PIP, which corresponds to the posterior inclusion probabilities computed by RSS [26]; (7) the number of SNPs that have been annotated within each SNP-set; (8) the "top" associated SNP within each SNP-set; (9) the PIP of each top SNP. Pages #3 and #4 provide similar results based on analyses where each SNP-set annotation has been augmented with a ±500 kilobase (kb) buffer to account for possible regulatory elements.
(ZIP)

S19 Table. 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 [57] (i.e., PIP ≥ 0.5). Page #1 provides the variant-level association mapping results with columns corresponding to: (1) chromosome; (2) SNP ID; (3) chromosomal position in base-pair (bp) coordinates; (4) SNP PIP; and (5) SuSiE PIP, which corresponds to SNP-level posterior inclusion probabilities computed by SuSiE [46]. 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) RSS PIP, which corresponds to the posterior inclusion probabilities computed by RSS [26]; (7) the number of SNPs that have been annotated within each SNP-set; (8) the "top" associated SNP within each SNP-set; (9) the PIP of each top SNP. Pages #3 and #4 provide similar results based on analyses where each SNP-set annotation has been augmented with a ±500 kilobase (kb) buffer to account for possible regulatory elements.
(ZIP)

S20 Table. Complete summary of the results after applying BANNs, SuSiE [46], and RSS [26] to high-density lipoprotein (HDL) and low-density lipoprotein (LDL) cholesterol in both individuals assayed within the Framingham Heart Study and ten thousand randomly sampled individuals of European ancestry from the UK Biobank. The first page compares the overlap of significant SNPs and SNP-sets found by each method according to the "median

probability model" threshold [57] (i.e., PIP $\geq$ 0.5) in the Framingham Heart Study. The second page lists how many SNPs and SNP-sets were replicated for each method when analyzing the independent dataset from the UK Biobank. Results are based on defining gene boundaries in two ways: (*a*) we use the UCSC gene boundary definitions directly, and (*b*) we augment the gene boundaries by adding SNPs within a ±500 kilobase (kb) buffer to account for possible regulatory elements.
(XLSX)

S21 Table. 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 [57] (i.e., PIP $\geq$ 0.5). Page #1 provides the variant-level association mapping results with columns corresponding to: (1) chromosome; (2) SNP ID; (3) chromosomal position in base-pair (bp) coordinates; (4) SNP PIP; and (5) SuSiE PIP, which corresponds to SNP-level posterior inclusion probabilities computed by SuSiE [46]. 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) RSS PIP, which corresponds to the posterior inclusion probabilities computed by RSS [26]; (7) the number of SNPs that have been annotated within each SNP-set; (8) the "top" associated SNP within each SNP-set; (9) the PIP of each top SNP. Pages #3 and #4 provide similar results based on analyses where each SNP-set annotation has been augmented with a ±500 kilobase (kb) buffer to account for possible regulatory elements.
(ZIP)

S22 Table. 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 [57] (i.e., PIP $\geq$ 0.5). Page #1 provides the variant-level association mapping results with columns corresponding to: (1) chromosome; (2) SNP ID; (3) chromosomal position in base-pair (bp) coordinates; (4) SNP PIP; and (5) SuSiE PIP, which corresponds to SNP-level posterior inclusion probabilities computed by SuSiE [46]. 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) RSS PIP, which corresponds to the posterior inclusion probabilities computed by RSS [26]; (7) the number of SNPs that have been annotated within each SNP-set; (8) the "top" associated SNP within each SNP-set; (9) the PIP of each top SNP. Pages #3 and #4 provide similar results based on analyses where each SNP-set annotation has been augmented with a ±500 kilobase

(kb) buffer to account for possible regulatory elements.
(ZIP)

S1 Text. Supplementary and background information for results mentioned in the main text. Specifically, we give description of the variational inference algorithm for the BANN framework, the data quality control procedures, simulation setup and scenarios, and additional results for the traits analyzed from the Framingham Heart Study and the UK Biobank.
(PDF)

Acknowledgments

This research was conducted in part using computational resources and services at the Center for Computation and Visualization at Brown University. This research was conducted using the UK Biobank Resource under Application Number 22419. This research was also conducted in part using data and resources from the Framingham Heart Study of the NHLBI and Boston University School of Medicine, which was partially supported by the NHLBI Framingham Heart Study (Contract No. N01-HC-25195) and its contract with Affymetrix, Inc for genotyping services (Contract No. N02-HL-6-4278). We thank all participants and staff from the Framingham Heart Study.

Author Contributions

Conceptualization: Lorin Crawford.

Data curation: Xiang Zhou, Sohini Ramachandran, Lorin Crawford.

Formal analysis: Pinar Demetci, Wei Cheng.

Funding acquisition: Xiang Zhou, Sohini Ramachandran, Lorin Crawford.

Investigation: Pinar Demetci, Wei Cheng, Gregory Darnell, Lorin Crawford.

Methodology: Pinar Demetci, Wei Cheng, Lorin Crawford.

Project administration: Lorin Crawford.

Resources: Xiang Zhou, Sohini Ramachandran, Lorin Crawford.

Software: Pinar Demetci, Wei Cheng.

Supervision: Xiang Zhou, Sohini Ramachandran, Lorin Crawford.

Validation: Pinar Demetci, Wei Cheng, Gregory Darnell.

Visualization: Pinar Demetci, Wei Cheng.

Writing – original draft: Pinar Demetci, Wei Cheng, Gregory Darnell, Xiang Zhou, Sohini Ramachandran, Lorin Crawford.

Writing – review & editing: Pinar Demetci, Wei Cheng, Gregory Darnell, Xiang Zhou, Sohini Ramachandran, Lorin Crawford.

References

1. 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. PMID: 18385116
2. 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. PMID: 20208533

3. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. Nat Rev Genet. 2010; 11(7):459–463. https://doi.org/10.1038/nrg2813 PMID: 20548291

4. 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. PMID: 21892150

5. 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.

6. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012; 44(7):821–825. https://doi.org/10.1038/ng.2310 PMID: 22706312

7. Hayeck TJ, Zaitlen NA, Loh PR, Vilhjalmsson B, Pollack S, Gusev A, et al. Mixed model with correction for case-control ascertainment increases association power. Am J Hum Genet. 2015; 96(5):720–730. Available from: https://pubmed.ncbi.nlm.nih.gov/25892111. https://doi.org/10.1016/j.ajhg.2015.03.004

8. Heckerman D, Gurdasani D, Kadie C, Pomilla C, Carstensen T, Martin H, et al. Linear mixed model for heritability estimation that explicitly addresses environmental variation. Proc Natl Acad Sci USA. 2016; 113(27):7377. Available from: http://www.pnas.org/content/113/27/7377.abstract. PMID: 27382152

9. 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. PMID: 28746338

10. 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. PMID: 28878256

11. Loh PR, Kichaev G, Gazal S, Schoech AP, Price AL. Mixed-model association for biobank-scale datasets. Nat Genet. 2018; 50(7):906–908. Available from: https://pubmed.ncbi.nlm.nih.gov/29892013. https://doi.org/10.1038/s41588-018-0144-6

12. Jiang L, Zheng Z, Qi T, Kemper KE, Wray NR, Visscher PM, et al. A resource-efficient tool for mixed model association analysis of large-scale data. Nat Genet. 2019; 51(12):1749–1755. Available from: https://doi.org/10.1038/s41588-019-0530-8. PMID: 31768069

13. Runcie DE, Crawford L. Fast and flexible linear mixed models for genome-wide genetics. PLoS Genet. 2019; 15(2):e1007978. Available from: https://doi.org/10.1371/journal.pgen.1007978. PMID: 30735486

14. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009; 461(7265):747–753. Available from: https://www.ncbi.nlm.nih.gov/pubmed/19812666. https://doi.org/10.1038/nature08494

15. Visscher PM, Brown MA, McCarthy MI, Yang J. Five Years of GWAS Discovery. Am J Hum Genet. 2012; 90(1):7–24. Available from: http://www.sciencedirect.com/science/article/pii/S0002929711005337. PMID: 22243964

16. Zhou X, Carbonetto P, Stephens M. Polygenic modeling with Bayesian sparse linear mixed models. PLoS Genet. 2013; 9(2):e1003264. https://doi.org/10.1371/journal.pgen.1003264 PMID: 23408905

17. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. Nat Genet. 2014; 46(2):100–106. https://doi.org/10.1038/ng.2876 PMID: 24473328

18. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, of the Psychiatric Genomics Consortium SWG, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015; 47:291–295. Available from: http://dx.doi.org/10.1038/ng.3211. PMID: 25642630

19. Wray NR, Wijmenga C, Sullivan PF, Yang J, Visscher PM. Common disease is more complex than implied by the core gene omnigenic model. Cell. 2018; 173(7):1573–1580. Available from: https://doi.org/10.1016/j.cell.2018.05.051. PMID: 29906445

20. Liu JZ, Mcrae AF, Nyholt DR, Medland SE, Wray NR, Brown KM, et al. A versatile gene-based test for genome-wide association studies. Am J Hum Genet. 2010; 87(1):139–145. https://doi.org/10.1016/j.ajhg.2010.06.009 PMID: 20598278

21. 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. https://doi.org/10.1016/j.ajhg.2010.05.002 PMID: 20560208

22. 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. PMID: 24098138

23. 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. PMID: 25885710

24. Lamparter D, Marbach D, Rueedi R, Kutalik Z, Bergmann S. Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. PLoS Comput Biol. 2016; 12(1):e1004714. Available from: https://doi.org/10.1371/journal.pcbi.1004714. PMID: 26808494

25. 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. PMID: 27489002

26. 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. https://doi.org/10.1038/s41467-018-06805-x PMID: 30341297

27. 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. PMID: 30875371

28. 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. PMID: 32542026

29. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015; 521(7553):436–444. Available from: https://doi.org/10.1038/nature14539. PMID: 26017442

30. Nagai A, Hirata M, Kamatani Y, Muto K, Matsuda K, Kiyohara Y, et al. Overview of the BioBank Japan Project: study design and profile. J Epidemiol. 2017; 27(3S):S2–S8. Available from: https://pubmed.ncbi.nlm.nih.gov/28189464. https://doi.org/10.1016/j.je.2016.12.005

31. 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. PMID: 30305743

32. 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. PMID: 27302636

33. Bellot P, de los Campos G, Pérez-Enciso M. Can deep learning improve genomic prediction of complex human traits? Genetics. 2018 11; 210(3):809–819. Available from: http://www.genetics.org/content/210/3/809.abstract. PMID: 30171033

34. Jiang Y, Reif JC. Modeling epistasis in genomic selection. Genetics. 2015; 201:759–768. https://doi.org/10.1534/genetics.115.177907 PMID: 26219298

35. Crawford L, Wood KC, Zhou X, Mukherjee S. Bayesian approximate kernel regression with variable selection. J Am Stat Assoc. 2018; 113(524):1710–1721. https://doi.org/10.1080/01621459.2017.1361830 PMID: 30799887

36. Wahba G. Splines models for observational data. vol. 59 of Series in Applied Mathematics. Philadelphia, PA: SIAM; 1990.

37. 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. https://doi.org/10.1214/18-aoas1222 PMID: 32542104

38. Courville A, Bergstra J, Bengio Y. Unsupervised models of images by spike-and-slab RBMs. In: Proceedings of the 28th International Conference on International Conference on Machine Learning. ICML'11. Madison, WI, USA: Omnipress; 2011. p. 1145–1152.

39. Deng W, Zhang X, Liang F, Lin G. An adaptive empirical Bayesian method for sparse deep learning. Advances in Neural Information Processing Systems. 2019 12; 2019:5563–5573. Available from: https://pubmed.ncbi.nlm.nih.gov/33244209.

40. Srinivas S, Subramanya A, Venkatesh Babu R. Training sparse neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops; 2017. p. 455–462.

41. Liang F, Li Q, Zhou L. Bayesian neural networks for selection of drug sensitive genes. J Am Stat Assoc. 2018; 113(523):955–972. Available from: https://pubmed.ncbi.nlm.nih.gov/31354179. https://doi.org/10.1080/01621459.2017.1409122

42. Ghosh S, Yao J, Doshi-Velez F. Model selection in Bayesian neural networks via horseshoe priors. J Mach Learn Res. 2019; 20(182):1–46. Available from: http://jmlr.org/papers/v20/19-236.html.

43. 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. https://doi.org/10.1093/bioinformatics/btn516 PMID: 18854360

44. 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. https://doi.org/10.1093/bioinformatics/btw018

45. 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. PMID: 27866706

46. Wang G, Sarkar A, Carbonetto P, Stephens M. A simple new approach to variable selection in regression, with application to genetic fine-mapping. J R Stat Soc B. 2020; 82:1273–1300. BioRxiv. Available from: http://biorxiv.org/content/early/2019/07/29/501114.abstract.

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. PMID: 16832355

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. https://doi.org/10.1093/aje/kwm021 PMID: 17372189

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

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. https://doi.org/10.1093/nar/gki025 PMID: 15608248

51. Bult CJ, Blake JA, Smith CL, Kadin JA, Richardson JE. Mouse Genome Database (MGD). Nucleic Acids Res. 2019; 47(D1):D801–D806. https://doi.org/10.1093/nar/gky1056 PMID: 30407599

52. 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. https://doi.org/10.1371/journal.pgen.1004969

53. 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. https://doi.org/10.1038/s41588-018-0193-x PMID: 30104760

54. 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. PMID: 31704910

55. 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. https://doi.org/10.1214/12-BA703

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

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

58. Uijlings JRR, van de Sande KEA, Gevers T, Smeulders AWM. Selective search for object recognition. Int J Comput Vis. 2013; 104(2):154–171. Available from: https://doi.org/10.1007/s11263-013-0620-5.

59. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition; 2014. p. 580–587.

60. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet. 2012; 91(2):224–237. Available from: http://www.sciencedirect.com/science/article/pii/S0002929712003163. PMID: 22863193

61. Berk RH, Jones DH. Goodness-of-fit test statistics that dominate the Kolmogorov statistics. Z Wahrsch Verw Gebiete. 1979; 47(1):47–59. Available from: https://doi.org/10.1007/BF00533250.

62. 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. PMID: 29399241

63. Blei DM, Kucukelbir A, McAuliffe JD. Variational inference: A review for statisticians. J Am Stat Assoc. 2017; 112(518):859–877. https://doi.org/10.1080/01621459.2017.1285773

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

65. 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.

66. Chen X, McClusky R, Chen J, Beaven SW, Tontonoz P, Arnold AP, et al. The number of X chromosomes causes sex differences in adiposity in mice. PLoS Genet. 2012; 8(5):e1002709. Available from: https://doi.org/10.1371/journal.pgen.1002709. PMID: 22589744

67. Mackay TFC. Epistasis and quantitative traits: using model organisms to study gene–gene interactions. Nat Rev Genet. 2014; 15(1):22–33. Available from: https://doi.org/10.1038/nrg3627. PMID: 24296533

68. Tyler AL, Donahue LR, Churchill GA, Carter GW. Weak epistasis generally stabilizes phenotypes in a mouse intercross. PLoS Genet. 2016; 12(2):e1005805. Available from: https://doi.org/10.1371/journal.pgen.1005805. PMID: 26828925

69. Strakova J, Kamdar F, Kulhanek D, Razzoli M, Garry DJ, Ervasti JM, et al. Integrative effects of dystrophin loss on metabolic function of the mdx mouse. Scientific Rep. 2018; 8(1):13624. Available from: https://pubmed.ncbi.nlm.nih.gov/30206270. https://doi.org/10.1038/s41598-018-31753-3

70. Lotta LA, Mokrosiński J, Mendes de Oliveira E, Li C, Sharp SJ, Luan J, et al. Human gain-of-function MC4R variants show signaling bias and protect against obesity. Cell. 2019; 177(3):597–607. https://doi.org/10.1016/j.cell.2019.03.044 PMID: 31002796

71. Zhou K, Yee SW, Seiser EL, van Leeuwen N, Tavendale R, Bennett AJ, et al. Variation in the glucose transporter gene SLC2A2 is associated with glycemic response to metformin. Nat Genet. 2016; 48 (9):1055–1059. Available from: https://pubmed.ncbi.nlm.nih.gov/27500523. https://doi.org/10.1038/ng.3632

72. Blanco P, Pitard V, Viallard JF, Taupin JL, Pellegrin JL, Moreau JF. Increase in activated CD8+ T lymphocytes expressing perforin and granzyme B correlates with disease activity in patients with systemic lupus erythematosus. Arthritis Rheum. 2005; 52(1):201–211. https://doi.org/10.1002/art.20745 PMID: 15641052

73. Li H, Adamopoulos IE, Moulton VR, Stillman IE, Herbert Z, Moon JJ, et al. Systemic lupus erythematosus favors the generation of IL-17 producing double negative T cells. Nat Comm. 2020; 11(1):2859. Available from: https://doi.org/10.1038/s41467-020-16636-4. PMID: 32503973

74. Sharabi A, Tsokos GC. T cell metabolism: new insights in systemic lupus erythematosus pathogenesis and therapy. Nat Rev Rheumatol. 2020; 16(2):100–112. Available from: https://doi.org/10.1038/s41584-019-0356-x. PMID: 31949287

75. Stefansson H, Rye DB, Hicks A, Petursson H, Ingason A, Thorgeirsson TE, et al. A genetic risk factor for periodic limb movements in sleep. N Engl J Med. 2007; 357(7):639–647. https://doi.org/10.1056/NEJMoa072743 PMID: 17634447

76. Winkelmann J, Schormair B, Lichtner P, Ripke S, Xiong L, Jalilzadeh S, et al. Genome-wide association study of restless legs syndrome identifies common variants in three genomic regions. Nat Genet. 2007; 39(8):1000–1006. https://doi.org/10.1038/ng2099 PMID: 17637780

77. Vaithilingam DS, Antao V, Kakis G. Regulation of polyunsaturated fat induced postprandial hypercholesterolemia by a novel gene Phc-2. Mol Cell Biochem. 1994; 130(1):67–74. https://doi.org/10.1007/BF01084269 PMID: 8190122

78. Silver M, Chen P, Li R, Cheng CY, Wong TY, Tai ES, et al. Pathways-Driven Sparse Regression Identifies Pathways and Genes Associated with High-Density Lipoprotein Cholesterol in Two Asian Cohorts. PLoS Genet. 2013; 9(11):e1003939. Available from: https://doi.org/10.1371/journal.pgen.1003939. PMID: 24278029

79. Cui C, Chatterjee B, Lozito TP, Zhang Z, Francis RJ, Yagi H, et al. Wdpcp, a PCP Protein Required for Ciliogenesis, Regulates Directional Cell Migration and Cell Polarity by Direct Modulation of the Actin Cytoskeleton. PLoS Biol. 2013; 11(11):e1001720. Available from: https://doi.org/10.1371/journal.pbio.1001720. PMID: 24302887

80. Wang DX, Kaur Y, Alyass A, Meyre D. A candidate-gene approach identifies novel associations between common variants in/near syndromic obesity genes and BMI in pediatric and adult European populations. Diabetes. 2019; 68(4):724–732. https://doi.org/10.2337/db18-0986 PMID: 30692245

81. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature. 2002; 420(6915):563–573. Available from: https://doi.org/10.1038/nature01266. PMID: 12466851

82. Hansen GM, Markesich DC, Burnett MB, Zhu Q, Dionne KM, Richter LJ, et al. Large-scale gene trapping in C57BL/6N mouse embryonic stem cells. Genome Res. 2008; 18(10):1670–1679. https://doi.org/10.1101/gr.078352.108 PMID: 18799693

83. Diez-Roux G, Banfi S, Sultan M, Geffers L, Anand S, Rozado D, et al. A high-resolution anatomical ttlas of the transcriptome in the mouse embryo. PLoS Biol. 2011; 9(1):e1000582. Available from: https://doi.org/10.1371/journal.pbio.1000582. PMID: 21267068

84. Skarnes WC, Rosen B, West AP, Koutsourakis M, Bushell W, Iyer V, et al. A conditional knockout resource for the genome-wide study of mouse gene function. Nature. 2011; 474(7351):337–342. Available from: https://doi.org/10.1038/nature10163. PMID: 21677750

85. Klebig ML, Wall MD, Potter MD, Rowe EL, Carpenter DA, Rinchik EM. Mutations in the clathrin-assembly gene *Picalm* are responsible for the hematopoietic and iron metabolism abnormalities in *fit1* mice. Proc Natl Acad Sci USA. 2003; 100(14):8360. Available from: http://www.pnas.org/content/100/14/8360.abstract. PMID: 12832620

86. Lin H, Grosschedl R. Failure of B-cell differentiation in mice lacking the transcription factor EBF. Nature. 1995; 376(6537):263–267. https://doi.org/10.1038/376263a0 PMID: 7542362

87. Laramie JM, Wilk JB, Williamson SL, Nagle MW, Latourelle JC, Tobin JE, et al. Multiple genes influence BMI on chromosome 7q31-34: the NHLBI Family Heart Study. Obesity (Silver Spring). 2009; 17 (12):2182–2189. https://doi.org/10.1038/oby.2009.141 PMID: 19461589

88. Lichenstein SD, Jones BL, O'Brien JW, Zezza N, Stiffler S, Holmes B, et al. Familial risk for alcohol dependence and developmental changes in BMI: the moderating influence of addiction and obesity genes. Pharmacogenomics. 2014; 15(10):1311–1321. Available from: https://pubmed.ncbi.nlm.nih.gov/25155933. https://doi.org/10.2217/pgs.14.86

89. Steen VM, Nepal C, Ersland KM, Holdhus R, Nævdal M, Ratvik SM, et al. Neuropsychological deficits in mice depleted of the schizophrenia susceptibility gene CSMD1. PLoS One. 2013; 8(11):e79501. https://doi.org/10.1371/journal.pone.0079501 PMID: 24244513

90. 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. PMID: 23586463

91. 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. https://doi.org/10.1093/nar/gkw377

92. Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PIW, Chen H, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science. 2007; 316(5829):1331–1336. https://doi.org/10.1126/science.1142358 PMID: 17463246

93. Sabatti C, Service SK, Hartikainen AL, Pouta A, Ripatti S, Brodsky J, et al. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. Nat Genet. 2009; 41(1):35–46. Available from: https://doi.org/10.1038/ng.271. PMID: 19060910

94. Ko A, Cantor RM, Weissglas-Volkov D, Nikkola E, Reddy PMVL, Sinsheimer JS, et al. Amerindian-specific regions under positive selection harbour new lipid variants in Latinos. Nat Comm. 2014; 5 (1):3983. Available from: https://doi.org/10.1038/ncomms4983. PMID: 24886709

95. Hebbar P, Nizam R, Melhem M, Alkayal F, Elkum N, John SE, et al. Genome-wide association study identifies novel recessive genetic variants for high TGs in an Arab population. J Lipid Res. 2018; 59 (10):1951–1966. https://doi.org/10.1194/jlr.P080218 PMID: 30108155

96. Sandhu MS, Waterworth DM, Debenham SL, Wheeler E, Papadakis K, Zhao JH, et al. LDL-cholesterol concentrations: a genome-wide association study. Lancet. 2008; 371(9611):483–491. Available from: https://pubmed.ncbi.nlm.nih.gov/18262040. https://doi.org/10.1016/S0140-6736(08)60208-1

97. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics. 2015; 31 (21):3555–3557. Available from: https://pubmed.ncbi.nlm.nih.gov/26139635. https://doi.org/10.1093/bioinformatics/btv402

98. Tennant BR, Vanderkruk B, Dhillon J, Dai D, Verchere CB, Hoffman BG. Myt3 suppression sensitizes islet cells to high glucose-induced cell death via Bim induction. Cell Death Dis. 2016; 7(5):e2233–e2233. Available from: https://doi.org/10.1038/cddis.2016.141. PMID: 27195679

99. Klarin D, Damrauer SM, Cho K, Sun YV, Teslovich TM, Honerlaw J, et al. Genetics of blood lipids among 300,000 multi-ethnic participants of the Million Veteran Program. Nat Genet. 2018; 50 (11):1514–1523. Available from: https://doi.org/10.1038/s41588-018-0222-9. PMID: 30275531

100. Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, et al. Mapping the Genetic Architecture of Gene Expression in Human Liver. PLoS Biol. 2008; 6(5):e107. Available from: https://doi.org/10.1371/journal.pbio.0060107. PMID: 18462017

101. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet. 2008; 40(2):161–169. Available from: https://doi.org/10.1038/ng.76. PMID: 18193043

102. Oni-Orisan A, Haldar T, Ranatunga DK, Medina MW, Schaefer C, Krauss RM, et al. The impact of adjusting for baseline in pharmacogenomic genome-wide association studies of quantitative change. npj Genom Med. 2020; 5(1):1. Available from: https://doi.org/10.1038/s41525-019-0109-4. https://doi.org/10.1038/s41525-019-0109-4 PMID: 31969989

103. Talmud PJ, Drenos F, Shah S, Shah T, Palmen J, Verzilli C, et al. Gene-centric association signals for lipids and apolipoproteins identified via the HumanCVD BeadChip. Am J Hum Genet. 2009; 85 (5):628–642. Available from: http://www.sciencedirect.com/science/article/pii/S0002929709004698. PMID: 19913121

104. Postmus I, Trompet S, Deshmukh HA, Barnes MR, Li X, Warren HR, et al. Pharmacogenetic meta-analysis of genome-wide association studies of LDL cholesterol response to statins. Nat Comm. 2014; 5(1):5068. Available from: https://doi.org/10.1038/ncomms6068. PMID: 25350695

105. Mo X, Lei S, Zhang Y, Zhang H. Genome-wide enrichment of m6A-associated single-nucleotide polymorphisms in the lipid loci. Pharmacogenomics J. 2019; 19(4):347–357. Available from: https://doi.org/10.1038/s41397-018-0055-z. PMID: 30262821

106. Liu DJ, Peloso GM, Yu H, Butterworth AS, Wang X, Mahajan A, et al. Exome-wide association study of plasma lipids in >300,000 individuals. Nat Genet. 2017; 49(12):1758–1766. Available from: https://pubmed.ncbi.nlm.nih.gov/29083408. https://doi.org/10.1038/ng.3977

107. Richardson TG, Sanderson E, Palmer TM, Ala-Korpela M, Ference BA, Davey Smith G, et al. Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable Mendelian randomisation analysis. PLoS Med. 2020; 17(3):e1003062. Available from: https://doi.org/10.1371/journal.pmed.1003062. PMID: 32203549

108. Paré G, Mao S, Deng WQ. A machine-learning heuristic to improve gene score prediction of polygenic traits. Scientific Rep. 2017; 7(1):12665. Available from: https://doi.org/10.1038/s41598-017-13056-1. PMID: 28979001

109. Ma J, Yu MK, Fong S, Ono K, Sage E, Demchak B, et al. Using deep learning to model the hierarchical structure and function of a cell. Nat Meth. 2018; 15(4):290–298. Available from: https://doi.org/10.1038/nmeth.4627. PMID: 29505029

110. Kim BJ, Kim SH. Prediction of inherited genomic susceptibility to 20 common cancer types by a supervised machine-learning method. Proc Natl Acad Sci USA. 2018; 115(6):1322–1327. Available from: http://www.pnas.org/content/115/6/1322.abstract. PMID: 29358382

111. Ho DSW, Schierding W, Wake M, Saffery R, O'Sullivan J. Machine learning SNP based prediction for precision medicine. Front Genet. 2019; 10:267. Available from: https://www.frontiersin.org/article/10.3389/fgene.2019.00267. PMID: 30972108

112. Jonsson BA, Bjornsdottir G, Thorgeirsson TE, Ellingsen LM, Walters GB, Gudbjartsson DF, et al. Brain age prediction using deep learning uncovers associated sequence variants. Nat Comm. 2019; 10(1):5409. Available from: https://doi.org/10.1038/s41467-019-13163-9. PMID: 31776335

113. Smemo S, Tena JJ, Kim KH, Gamazon ER, Sakabe NJ, Gomez-Marin C, et al. Obesity-associated variants within FTO form long-range functional connections with IRX3. Nature. 2014; 507(7492):371–375. https://doi.org/10.1038/nature13138 PMID: 24646999

114. Claussnitzer M, Dankel SN, Kim KH, Quon G, Meuleman W, Haugen C, et al. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. N Engl J Med. 2015; 373(10):895–907. Available from: https://doi.org/10.1056/NEJMoa1502214. PMID: 26287746

115. Kaess B, Fischer M, Baessler A, Stark K, Huber F, Kremer W, et al. The lipoprotein subfraction profile: heritability and identification of quantitative trait loci. J Lipid Res. 2008; 49(4):715–723. https://doi.org/10.1194/jlr.M700338-JLR200 PMID: 18165655

116. Zhang C, Shahbaba B, Zhao H. Variational Hamiltonian monte carlo via score matching. Bayesian Anal. 2018; 13(2):485–506. Available from: https://projecteuclid.org:443/euclid.ba/1500948232.

117. Zeng P, Zhou X, Huang S. Prediction of gene expression with cis-SNPs using mixed models and regularization methods. BMC Genomics. 2017; 18(1):368. Available from: https://doi.org/10.1186/s12864-017-3759-6. PMID: 28490319

118. Yang S, Zhou X. Accurate and scalable construction of polygenic scores in large biobank data sets. Am J Hum Genet. 2020; 106(5):679–693. Available from: http://www.sciencedirect.com/science/article/pii/S0002929720301099. PMID: 32330416

119. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet. 2003; 34(3):267–273. Available from: https://doi.org/10.1038/ng1180. PMID: 12808457

120. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005; 102(43):15545–15550. Available from: http://www.pnas.org/content/102/43/15545.abstract. PMID: 16199517

121. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019; 47(D1):D886–D894. Available from: https://doi.org/10.1093/nar/gky1016. PMID: 30371827

122. Tsang M, Cheng D, Liu Y. Detecting statistical interactions from neural network weights. In: International Conference on Learning Representations; 2018. p. 1–21.

123. Runcie D, Cheng H, Crawford L. Mega-scale linear mixed models for genomic predictions with thousands of traits. bioRxiv. 2020;p. 2020.05.26.116814. Available from: http://biorxiv.org/content/early/2020/05/29/2020.05.26.116814.abstract.

124. Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat Meth. 2014; 11(4):407–409. Available from: https://pubmed.ncbi.nlm.nih.gov/24531419. https://doi.org/10.1038/nmeth.2848

125. Louizos C, Welling M. Structured and Efficient Variational Deep Learning with Matrix Gaussian Posteriors. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning—Volume 48. ICML'16. JMLR.org; 2016. p. 1708–1716.

126. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. J Am Stat Assoc. 1993; 88(421):9–25. Available from: www.jstor.org/stable/2290687.

127. Breslow NE, Lin X. Bias correction in generalised linear mixed models with a single component of dispersion. Biometrika. 1995; 82(1):81–91. Available from: www.jstor.org/stable/2337629.

128. Lin X, Breslow NE. Bias correction in generalized linear mixed models with multiple components of dispersion. J Am Stat Assoc. 1996; 91(435):1007–1016. Available from: www.jstor.org/stable/2291720.

129. Sun S, Zhu J, Mozaffari S, Ober C, Chen M, Zhou X. Heritability estimation and differential analysis of count data with generalized linear mixed models in genomic sequencing studies. Bioinformatics. 2019; 35(3):487–496. Available from: https://pubmed.ncbi.nlm.nih.gov/30020412. https://doi.org/10.1093/bioinformatics/bty644

130. Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet. 2011; 88(3):294–305. Available from: https://pubmed.ncbi.nlm.nih.gov/21376301. https://doi.org/10.1016/j.ajhg.2011.02.002

131. Golan D, Lander ES, Rosset S. Measuring missing heritability: Inferring the contribution of common variants. Proc Natl Acad Sci USA. 2014; 111(49):5272–5281. Available from: http://www.pnas.org/content/111/49/E5272.abstract. PMID: 25422463

132. Weissbrod O, Lippert C, Geiger D, Heckerman D. Accurate liability estimation improves power in ascertained case-control studies. Nat Meth. 2015; 12(4):332–334. Available from: https://doi.org/10.1038/nmeth.3285. PMID: 25664543

133. 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. PMID: 18852846

134. 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. PMID: 22836096

135. 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. https://doi.org/10.1016/j.ajhg.2017.08.002 PMID: 28844487

136. 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. https://doi.org/10.1016/j.ajhg.2018.11.008

137. Wand MP, Ormerod JT, Padoan SA, Frühwirth R. Mean field variational Bayes for elaborate distributions. Bayesian Anal. 2011; 6(4):847–900. https://doi.org/10.1214/11-BA631

138. 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.

139. Hormozdiari F, Kostem E, Kang EY, Pasaniuc B, Eskin E. Identifying causal variants at loci with multiple signals of association. Genetics. 2014; 198(2):497–508. Available from: https://pubmed.ncbi.nlm.nih.gov/25104515. https://doi.org/10.1534/genetics.114.167908

140. 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. https://doi.org/10.4172/2155-9562.1000236

141. 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. https://doi.org/10.1016/j.tem.2013.10.003

142. 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. https://doi.org/10.1016/S0021-9258(18)47945-8 PMID: 3115992

143. Kay AD, Day SP, Nicoll JAR, Packard CJ, Caslake MJ. Remodelling of cerebrospinal fluid lipoproteins after subarachnoid hemorrhage. Atherosclerosis. 2003; 170(1):141–146. https://doi.org/10.1016/S0021-9150(03)00249-1 PMID: 12957692

144. 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. PMID: 28054633