Multi-scale Genomic Inference using Biologically Annotated Neural Networks

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

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

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

Abstract

With the emergence of large-scale genomic datasets, there is a unique opportunity to integrate machine learning approaches as standard tools within genome-wide association (GWA) studies. Unfortunately, while machine learning methods have been shown to account for nonlinear data structures and exhibit greater predictive power over classic linear models, these same algorithms have also become criticized as "black box" techniques. Here, we present biologically annotated neural networks (BANNs), a novel probabilistic framework that makes machine learning fully amenable for GWA applications. 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. Part of our key innovation is to treat the weights and connections of the network as random variables with prior distributions that reflect how genetic effects are manifested at different genomic scales. The BANN software uses scalable variational inference to provide fully interpretable posterior summaries which allow researchers to simultaneously perform (*i*) fine-mapping with SNPs and (*ii*) enrichment analyses with SNP-sets on complex traits. Through simulations and real GWA data applications, we show that our method improves upon state-of-the-art approaches in both settings across a wide range of genetic architectures.

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 mixed models (LMMs) [1–13]. The flexibility and interpretability of LMMs make them a widely used tool in genome-wide association (GWA) studies where the goal is often to test for associations between individual single nucleotide polymorphisms (SNPs) and a phenotype of interest. In these cases, traditional LMMs provide a set of P-values or posterior inclusion probabilities (PIPs) which lend statistical evidence on how important each variant is to explaining the overall genetic architecture of a trait. However, this univariate SNP-level 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 the LMM framework to identify enriched gene or pathway-level associations, where 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 LMM 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 LMM framework and integrate machine learning techniques as standard statistical tools within GWA analyses. In some consortia efforts hundreds of thousands of individuals have been genotyped at millions of markers [29]. Indeed, machine learning methods such as neural networks are well known to be most powered in settings when large training data is available [30]. It is also well known that these nonlinear statistical approaches often exhibit greater predictive accuracy than LMMs, particularly for complex traits with broad-sense heritability that is driven by non-additive genetic variation (e.g., gene-by-gene interactions) [31]. One of the key characteristics that leads to better predictive performance from machine learning approaches is the automatic inclusion of higher order interactions between variables being put into the model [32, 33]. For example, neural networks leverage nonlinear activation functions between layers that implicitly enumerate all possible (polynomial) interaction effects [34]. 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 machine learning methods since they do not naturally produce interpretable significance measures (e.g., P-values or PIPs) like traditional LMMs [33, 35].

In this work, we develop biologically annotated neural networks (BANNs), a novel probabilistic framework that makes machine learning amenable for fine mapping and discovery in high-dimensional genomic association studies (Fig. 1). BANNs are feedforward Bayesian models with partially connected architectures that are guided by predefined SNP-set annotations (Fig. 1a). There are three key scientific contributions to our approach. First, the partially connected network architecture yields a fully interpretable model where the input layer encodes SNP-level effects, and the hidden layer models accounts for the effects among SNP-sets (Fig. 1b). Second, 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 make for a powerful machine learning strategy for conducting fine 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, 36–39]. We also apply the BANN framework to six quantitative traits assayed in a heterogenous stock of mice from Wellcome Trust Centre for Human Genetics [40], and two quantitative traits in individuals from the Framingham Heart Study [41].

Results

BANN 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 BANN framework simply requires individual-level genotype/phenotype data and a predefined list of SNP-set annotations (Fig. 1a). The method can also take in summary statistics where SNP-level effect size estimates are treated as the phenotype and an estimate of the linkage disequilibrium

(LD) matrix is used as input data (Supplementary Fig. 1). Structurally, sequential layers of the BANN 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 work, we define SNP-sets as collections of variants that fall within a chromosomal window or neighborhood. For example, when studying human GWA data, we use gene annotations as defined by the NCBI's Reference Sequence (RefSeq) database in the UCSC Genome Browser [42] (Methods). The BANN 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 desert, 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 BANN 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.

We frame the BANN methodology as a Bayesian nonlinear mixed model with which we can perform classic variable selection (Fig. 1c; see Methods). Here, we leverage the fact that using nonlinear activation functions for the neurons in the hidden layer implicitly account for both additive and non-additive effects between SNPs within a given SNP-set (Supplementary Notes). Part of our key innovation is to 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 ($\boldsymbol{\theta}$) to flexibly estimate a wide range of SNP-level effect size distributions [10, 44–46]. Similarly, we follow previous works 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 on the weights in the second layer ($\mathbf{w}$). With these point mass mixture distributions, 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 widely used variational inference algorithm for neural networks [47], we jointly infer posterior inclusion probabilities (PIPs) for SNPs ($\boldsymbol{\gamma}_\theta$) and SNP-sets ($\boldsymbol{\gamma}_w$). The PIPs are defined as the posterior probability that the weight of a given connection is nonzero. We use this information to select 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. Details and derivations of the BANN framework can be found in Methods and Supplementary Notes.

Power to Detect SNPs and SNP-Sets in Simulation Studies

In order to assess the performance of models under the BANN framework, we followed previous work [9,28] to simulate 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 [29] (Methods). After quality control procedures, our simulations included 36,518 SNPs (Supplementary Notes). Next, we used the NCBI's Reference Sequence (RefSeq) database in the UCSC Genome Browser [42] to annotate SNPs with the appropriate genes. Unannotated SNPs located within the same genomic desert 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 ($H^2 = 0.2$ 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 help 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 scenario, we consider traits being generated with and without additional population structure (Methods). In the former setting, traits are simulated while also using the top ten principal components of the genotype matrix as covariates to create stratification. The genetic contributions of the principal components are fixed to be 10% of the total phenotypic variance. Throughout this section, we assess the performance for two versions of the BANN 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) without any control for population structure. All results are based on 100 different simulated phenotypes for each parameter combination (Supplementary Notes).

The main utility of the BANN 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, 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 (Supplementary Tables 1-8). 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) [48]. 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 Supplementary Figs. 2-16). Specific takeaways about these analyses are given below.

Fine Mapped SNP-Level Results. For SNP-level comparisons, we used three fine-mapping methods as benchmarks: CAVIAR [38], SuSiE [39], and FINEMAP [37]. 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. Notedly, CAVIAR exhaustively and FINEMAP (approximately) search over different models to find the best combination of associated SNPs with nonzero effects. 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 an empirical estimate of the LD matrix.

Overall, the two BANN models and SuSiE (with high $\ell = 3000$) consistently achieve the greatest empirical power and lowest FDR across all genetic architectures we considered. These three approaches also stand out in terms of true-versus-false positive rates and precision-versus-recall. Notedly, the choice of the ℓ parameter proved to be a large factor for the performance of SuSiE, as it was consistently the worst model 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 (Supplementary Table 9). For more context, an analysis on just 4,000 individuals and 10,000 SNPs takes the BANN 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 the BANN model on individual-level data clearly 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 Supplementary Figs. 5-8 and 13-16)—particularly when traits are lowly heritable with polygenic architectures (e.g., $H^2 = 0.2$). A direct comparison of the PIPs derived by BANNs and SuSiE shows that the integrative and nonlinear neural network training procedure of BANNs enables its ability to identify associated SNPs even in these more complex phenotypic architectures (Fig. 3 and Supplementary Figs. 17-23). Ultimately, this result is unsurprising since the ReLU activation functions in the hidden layers of the BANN framework implicitly enumerates the interactions between SNPs within the a given SNP-set (Supplementary Notes). 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) partition or marginalize out proportions of the phenotypic variance that are contributed by nonlinearities [9, 13].

Enriched SNP-Set Level Results. For SNP-set level comparisons, we consider six gene or SNP-set enrichment approaches including: RSS [26], PEGASUS [25], GBJ [27], SKAT [21], GSEA [36], 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 [49]. 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 [50]. Lastly, RSS is a Bayesian linear mixed model enrichment 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 [51]. It is worth noting that, while RSS and the BANN framework are conceptually different, the two methods utilize very similar variational approximation algorithms for posterior inference [47] (Methods and Supplementary Notes).

Similar to the conclusions drawn during the SNP-level assessments, both the BANN 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. 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 BANN and RSS can be found in Fig. 3 and Supplementary Figs. 17-23. While RSS also performs generally well for the additive trait architectures, the algorithm for the model often takes longer than either of the BANN implementations to converge (Supplementary Table 10). 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; however, importantly, the sparse and partially connected architecture of the BANN framework allows it to scale more favorably for larger dimensional datasets. 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 high, sparse phenotypic architectures that are dominated by additive genetic effects, and total phenotypic variance that is not confounded by additional population stratification (Fig. 2 and Supplementary Figs. 2-16). 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. While our main focus is on conducting multi-scale genomic inference, because the BANN 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 fixed genetic effects (both additive and non-additive) and random effects (e.g., population stratification), collectively [16]. Within the BANN framework, this estimation can be done on both the SNP and SNP-set level while using either genotype-phenotype data or summary statistics (Supplementary Notes). For our simulation studies, the true PVE $= H^2 + 10\%$ and H^2 for traits generated with and without including the top ten genotypic principal components as covariates, respectively. We assess the ability of BANN to recover these true estimates using root mean square error (RMSE) (Supplementary Figs. 24 and 25). 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. BANN 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 BANN 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 (Supplementary Notes). Similar observations have been noted when using variational inference [52, 53].

Fine Mapping and Genomic Enrichment in Heterogenous Stock of Mice

We apply the BANN framework to individual-level genotypes and six quantitative traits in a heterogeneous stock of mice dataset from the Wellcome Trust Centre for Human Genetics [40]. 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 (Supplementary Notes). For SNP-set annotations, we used the Mouse Genome Informatics database (`http://www.informatics.jax.org`) [54] to map SNPs to the closest neighboring gene(s). Unannotated SNPs located within the same genomic desert 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. Specifically, the mice in this study are known to be related 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 [33].

For each trait, we provide a summary table which lists the PIPs for SNPs and SNP-sets after fitting the BANN model to the individual-level genotypes and phenotype data (Supplementary Tables 11-16). We use manhattan plots to visually display the variant-level fine 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 (Supplementary Fig. 26). Importantly, many of the candidate genes and intergenic regions selected by the BANN model have been previously validated by past publications as having some functional relationship with the traits of interest (Table 1). For example, BANN reports the genes *Btbd9* and *hlb156* as being enriched for the percentage of CD8+ cells in mice (PIP = 0.87 and 0.72, respectively). This same chromosomal region on chromosome 17 was also reported in the original study as having highly significant quantitative trait loci for CD8+ cells (bootstrap posterior probability equal to 1.00) [40]. Similarly, the X chromosome is well known

to majorly influence adiposity and metabolism in mice [55]. As expected, in body weight and BMI, our approach identified significant enrichment in this region—headlined by the dystrophin gene *Dmd* in both cases [56]. Finally, we want to point out 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 BANN reported the region between *Gm22219* and *Mc4r* on chromosome 18 as having a relatively high PIP of 0.74. Recently, a large-scale GWA study on individuals from the UK Biobank showed that variants around *MC4R* protect against obesity in humans [57].

Overall, the results from this smaller GWA study highlight three key characteristics resulting from the sparse probabilistic assumptions underlying the BANN 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 [47]. 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 our findings with BANNs are not biased by the sheer size of SNP-sets. The enrichment of a SNP-set is instead strictly determined by the relative posterior distribution of zero and nonzero SNP-level effect sizes within its annotated genomic window (Supplementary Tables 11-16). In other words, a SNP-set is not guaranteed to have a high inclusion probability just because it contains a 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 BANN 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 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 Supplementary Fig. 26).

Analyzing Cholesterol Traits in the Framingham Heart Study

Next, we apply the BANN 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 [41] genotyped at 394,174 SNPs. Following quality control procedures, we regressed the top ten principal components of the genotype data onto each trait to control for population structure (Supplementary Notes). Next, we used the gene boundaries listed in the NCBI's RefSeq database from the UCSC Genome Browser [42] to define SNP-sets. Similar to the previous sections, unannotated SNPs located within the same genomic desert were labeled as being within the "intergenic region" between two genes. This resulted in a total of 18,364 SNP-sets to be analyzed.

For each trait, we again fit the BANN 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 (Supplementary Tables 17-18). In Fig. 4, we show a manhattan plot of the variant-level fine mapping results, 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 [58, 59] to identify the categories that they overrepresent in the database of Genotypes and Phenotypes (dbGaP) and the NHGRI-EBI GWAS Catalog (Supplementary Figure 27). Similar to the previous section, the BANN framework identified many SNPs and SNP-sets that have been shown to be associated with cholesterol-related processes in past publications (Table 2). For example, in HDL, BANN identified an enriched intergenic region between the *HERPUD1* and *CETP* (PIP = 0.99) which has been also replicated in multiple GWA studies with multiethnic cohorts [60–63]. The Enrichr analyses were also consistent with intuition. 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 BANN model are either directly related to lipoproteins and

cholesterol (e.g., "Alpolipoprotein 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 some key consequences from the variational inference used in the BANN software. SNP-level 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 approximately 1% of the phenotypic variation in circulating LDL levels [64] Since these two SNPs are physically closed to each other and sit in a high LD block ($R^2 \approx 0.63$ with $P < 1 \times 10^{-4}$ [65]), the spike and slab prior in the BANN framework will maintain the nonzero weight of one and penalize the estimated effect of the other. Indeed, in our analysis, rs4970834 was reported to be associated with LDL (PIP = 0.947), while the effect size of rs599839 was shrunk towards 0 (PIP = 1×10^{-4}). Again, this is most likely due to the variational approximations utilized by BANNs (Methods and Supplementary Notes) where, if two SNPs are in strong LD, the model will tend to select just one of them [26,47]. Lastly, we obtain PVE estimates on the SNP and SNP-set level for each trait. As previously demonstrated in our simulation study, the variational Bayesian inference algorithm used in the BANN software yields misestimated approximate posterior means for the network weights and overestimated variance for the residual training error for complex polygenic traits (Supplementary Notes). As a result, BANNs estimates the PVE to be 0.11 and 0.04 for HDL and LDL, respectively. In general, these values are smaller than what is typically reported in the literature for these complex phenotypes (PVE $\geq$ 27% for HDL and PVE $\geq$ 21% for LDL, respectively) [66].

Discussion

Historically, machine learning approaches have largely been limited to prediction-based tasks in GWA applications [67–70]. Biologically annotated neural networks (BANNs) are a class of feedforward probabilistic models that overcome this central limitation by incorporating partially connected architectures that are guided by predefined SNP-set annotations. This creates a fully interpretable framework where the first layer of the neural network encodes SNP-level effects and the neurons within hidden layer represent the different SNP-set groupings. We frame the BANN methodology as a Bayesian nonlinear mixed model and use sparse prior distributions to perform variable selection on the network weights. By implementing a novel and integrative variational inference algorithm, we are able to derive posterior inclusion probabilities (PIPs) which allows researchers to carry out SNP-level fine-mapping and SNP-set enrichment analyses, simultaneously. Most importantly, we want to stress that while we focus on biological motivations, the concept of partially connected neural networks may extend to any scientific application where annotations can help guide the groupings of variables.

Through an extensive simulation study, we demonstrated the utility of the BANN framework on individual-level data (Fig. 1) and GWA summary statistics (Supplementary Fig. 1). Here, we showed that both implementations consistently outperform widely used SNP-level fine-mapping methods and state-of-the-art SNP-set enrichment methods in a wide range of genetic architectures (Figs. 2-3, Supplementary Figs. 2-23, and Supplementary Tables 1-8). This advantage was most clear when the broad-sense heritability of the complex traits included pairwise genetic interactions. Lastly, 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 works (Fig. 4, Supplementary Figs. 26-27, Tables 1-2, and Supplementary Tables 11-18).

As with any methodology, the current implementation of the BANN framework is not without its limitations. 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 [71, 72]). 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 BANN model comes from the fact that it uses variational inference to estimate its parameters. While the current implementation is scalable for large datasets (Supplementary Tables 9 and 10), we showed that the variational algorithm can lead to underestimated approximations of the PVE (Supplementary Figs. 24 and 25) and will occasionally miss causal SNPs if they are in high LD with other non-causal SNPs in the dataset. Exploring alternative ways to carry out approximate Bayesian inference is something to consider for future work [73].

There are several other potential extensions for the BANN 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. This would involve integrating information from curated databases such as MSigDB [74, 75]. Second, the current BANN model only takes in genetic information and ignores other sources of variation (e.g., population structure). In the future, we would like to expand the framework to also take in covariates as fixed effects in the model. Third, we have only focused on analyzing one phenotype at a time in this study. However, many previous works have extensively shown that modeling multiple phenotypes can often dramatically increase power [76]. Therefore, it would be interesting to extend the BANN framework to take advantage of phenotype correlations to identify pleiotropic epistatic effects. Modeling strategies based on the multivariate linear mixed model (mvLMM) [77] and matrix variate Gaussian process (mvGP) [78] 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 BANN framework to be expressed as a (nonlinear) logistic mixed model which is an approach that has been well-established in the statistics literature [79–81]. Unfortunately, it is not straightforward to define broad-sense heritability under the logistic mixed model and controlling for additional confounders that can occur within case-control studies (e.g., ascertainment) can be difficult. As an alternative, the liability threshold mixed model avoids these issues by assuming that binary traits can be modeled via continuous latent liability scores [82–84]. 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, we would 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.

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.

Acknowledgements

This research research was conducted in part using computational resources and services at the Center for Computation and Visualization at Brown University. 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.

Funding

This research was supported in part by US National Institutes of Health (NIH) grant R01 GM118652, and National Science Foundation (NSF) CAREER award DBI-1452622 to S. Ramachandran. This research was also partly supported 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), as well as by an Alfred P. Sloan Research Fellowship (No. FG-2019-11622) awarded to L. Crawford. 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 R01HG009124 and the NSF Grant DMS1712933. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of any of the funders.

Author Contributions

LC conceived the methods. PD developed the software. All authors carried out the analyses. All authors wrote and reviewed the manuscript.

Competing Interests

The authors declare no competing interests.

Methods

Annotations

We used the NCBI's Reference Sequence (RefSeq) database in the UCSC Genome Browser [42] to annotate SNPs with appropriate SNP-sets. In the main text, we consider a SNP being "inside" a gene using the UCSC gene boundary definitions directly. Genes with only one SNP within their boundary were excluded from either analysis. Unannotated SNPs located within the same genomic desert are labeled as being within the "intergenic region" between two genes. Altogether, with annotated genes and labeled intergenic regions, a total of 28,644 SNP-sets were analyzed.

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 $\{\mathcal{S}_1, \ldots, \mathcal{S}_G\}$ (Figure 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 BANN 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 (Figure 1b). Formally, we specify this network as a nonlinear regression model (Figure 1c)

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

where $\mathbf{X}_g = [\mathbf{x}_1, \ldots, \mathbf{x}_{|\mathcal{S}_g|}]$ is the subset of SNPs annotated for SNP-set g; $\boldsymbol{\theta}_g = (\theta_1, \ldots, \theta_{|\mathcal{S}_g|})$ are the corresponding inner layer weights; $h(\bullet)$ denotes the nonlinear activations defined for neurons in the hidden layer; $\mathbf{w} = (w_1, \ldots, w_G)$ are the weights for the G-predefined SNP-sets in the hidden layer; $\mathbf{b}^{(1)} = (b_1^{(1)}, \ldots, b_G^{(1)})$ and $b^{(2)}$ are deterministic biases that are produced during the network training phase in the input and hidden layers, respectively; and $\mathbf{1}$ is an N-dimensional vector of ones. For convenience, we assume that the genotype matrix (column-wise) and trait of interest have been mean-centered and standardized. In the main text, $h(\bullet)$ is defined as a Leaky rectified linear unit (Leaky ReLU) activation function [43], where $h(\boldsymbol{x}) = \boldsymbol{x}$ if $\boldsymbol{x} > \mathbf{0}$ and $0.01\boldsymbol{x}$ otherwise. Note that Eq. (1) can be seen as a nonlinear take on classic integrative and structural regression models [22, 26, 85–88] frequently used in GWA analyses.

Part of the key methodological innovation in the BANN framework is to treat the weights of the input (θ_j) and hidden layers (w_g) as random variables. This 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, 44–46]

$$\theta_j \sim \sum_{k=1}^{K} \pi_{\theta k} \mathcal{N}(0, \sigma_{\theta k}^2), \qquad \log(\pi_{\theta k}) \sim \mathcal{U}(-\log(J), \log(1)), \qquad \sigma_{\theta k}^2 \sim \text{Inv-Gamma}(u_\theta, v_\theta) \tag{2}$$

where $\boldsymbol{\pi}_\theta = (\pi_{\theta 1}, \ldots, \pi_{\theta K})$ represents the marginal (unconditional) probability that a randomly selected SNP belongs to the k-th mixture component (with $\sum_k \pi_{\theta k} = 1$). The prior in Eq. (2) models distinct types of nonzero SNP-level effects through the K different variance components $\boldsymbol{\sigma}_\theta^2 = (\sigma_{\theta 1}^2, \ldots, \sigma_{\theta K}^2)$. We allow sequential fractions of SNPs $(\pi_{\theta 1}, \ldots, \pi_{\theta K})$ to correspond to distinctly smaller effects $(\sigma_{\theta 1}^2 > \cdots > \sigma_{\theta K}^2 = 0)$ [45]. 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 effects on phenotypic variation [28]. We place a uniform prior on $\log \pi_{\theta k}$ 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, 89]. Similarly, because we do not know the magnitude for SNP effects in each category, we place relatively diffuse inverse-gamma priors on each of the variance components to allow the posterior of $\boldsymbol{\theta}$ to be primarily driven by information contained within the genotype data at hand (see Supplementary Notes).

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

$$w_g \sim \pi_w \mathcal{N}(0, \sigma_w^2) + (1 - \pi_w)\delta_0, \qquad \log(\pi_w) \sim \mathcal{U}(-\log(G), \log(1)), \qquad \sigma_w^2 \sim \text{Inv-Gamma}(u_w, v_w) \quad (3)$$

where, in addition to previous notation, δ_0 is a point mass at zero, and 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. (1), 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 scalable variational Bayesian algorithm to estimate all model parameters (Supplemental Note). As the BANN network is trained, the posterior mean for the weights of non-associated SNP and SNP-sets are set to zero, leaving only a sparse subset of trait relevant neurons to predict 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

$$\gamma_{\theta j} = \Pr[\theta_j \neq 0 \,|\, \mathbf{y}, \mathbf{X}], \qquad \gamma_{wg} = \Pr[w_g \neq 0 \,|\, \mathbf{y}, \mathbf{X}, \boldsymbol{\theta}_g] \quad (4)$$

where, again for the latter, the enrichment of SNP-sets is conditioned on the association of individual SNPs. The goal of the sparse shrinkage priors in Eqs. (2)-(3) is similar to that of regularization via "dropout" in the machine and deep learning literature where the connections between units in a neural network are dropped according to a penalized loss function [90]. The Bayesian formulation in the BANN framework makes network sparsity more targeted for GWA applications through contextually motivated prior distributions. Moreover, posterior inference on $\boldsymbol{\gamma}_\theta = (\gamma_{\theta 1}, \ldots, \gamma_{\theta J})$ and $\boldsymbol{\gamma}_w = (\gamma_{w1}, \ldots, \gamma_{wG})$ detail the degree to which nonzero weights occur.

Posterior Computation with Variational Inference

We combine the likelihood in Eq. (1) and the prior distributions in Eqs. (2)-(4) 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 scalable model fitting we modify a previously established variational expectation-maximization (EM) algorithm for integrative network parameter estimation [47]. 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 [52]. 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" [91]. The algorithm then follows three general steps. First, we assign exchangeable uniform hyper-priors over a grid of values on the log-scale for $\boldsymbol{\pi}_\theta$ and π_w [47]. 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 $(\boldsymbol{\gamma}_\theta, \boldsymbol{\gamma}_w)$ 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 [92]. 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 on the SNP-set level enrichments. The software code iterates between the "inner" lower bound and the "outer" lower bound each step of the algorithm until convergence. Detailed steps in the variational EM algorithm and explicit co-ordinate ascent updates for network parameters are given in Supplementary Notes.

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 BANN 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 $\widehat{\boldsymbol{\theta}}$, and an empirical linkage disequilibrium (LD) matrix $\mathbf{R}$. 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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{e}, \qquad \mathbf{e} \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbf{I}) \tag{5}$$

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 $\widehat{\theta}_j = (\mathbf{x}_j^\intercal \mathbf{x}_j)^{-1}\mathbf{x}_j^\intercal \mathbf{y}$, where $\mathbf{x}_j$ is the j-th column of the genotype matrix $\mathbf{X}$ and $\widehat{\theta}_j$ is the j-th entry of the vector $\widehat{\boldsymbol{\theta}}$. We assume the LD matrix $\mathbf{R}$ is empirically estimated from external data (e.g., directly from GWA study data, or using an LD map from a population with similar genomic ancestry to that of the samples analyzed in the GWA study). 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 $\widehat{\boldsymbol{\theta}}$ and the true coefficient values $\boldsymbol{\theta}$ is [28, 38, 45, 93]

$$\mathbb{E}[\widehat{\theta}_j] = \sum_{j'=1}^{J} r(\mathbf{x}_j, \mathbf{x}_{j'})\theta_{j'} \tag{6}$$

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 $\widehat{\boldsymbol{\theta}}$ 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. With this relationship in mind, the BANN-SS framework implements the following sparse nonlinear regression for inferring multi-scale genomic effects from summary statistics (Supplementary Figure 1)

$$\widehat{\boldsymbol{\theta}} = \sum_{g=1}^{G} h(\mathbf{R}_g\boldsymbol{\theta}_g + \mathbf{1}b_g^{(1)})w_g + \mathbf{1}b^{(2)}, \tag{7}$$

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. (7), 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. (2)-(4); Supplementary Note). Again, we use PIPs $\boldsymbol{\gamma}_\theta$ and $\boldsymbol{\gamma}_w$ 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 used a simulation scheme to generate quantitative traits under multiple genetic architectures 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 mixed model

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\mu} + \sum_{c \in \mathcal{C}} \mathbf{x}_c \theta_c + \mathbf{W}\boldsymbol{\varphi} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \tau^2 \mathbf{I}), \tag{8}$$

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 $\boldsymbol{\varphi}$; $\mathbf{Z}$ is an $N \times M$ matrix of covariates representing additional population structure (e.g., the top ten genotype principal components from the genotype matrix) with corresponding fixed effects $\boldsymbol{\mu}$; and $\boldsymbol{\varepsilon}$ is an N-dimensional vector of environmental noise. The 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 $\mathbb{V}[\sum \mathbf{x}_c \theta_c] + \mathbb{V}[\mathbf{W}\boldsymbol{\varphi}] = H^2$. Together with the centered and scaled genetic random effects, we get a total phenotypic variance explained for each trait $\text{PVE} = H^2 + \mathbb{V}[\mathbf{Z}\boldsymbol{\mu}]$. Lastly the environment noise is rescaled such that $\mathbb{V}[\boldsymbol{\varepsilon}] = 1 - \text{PVE}$. The full genotype matrix and phenotypic vector are given to the BANN 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 $\widehat{\theta}_j = (\mathbf{x}_j^\intercal \mathbf{x}_j)^{-1} \mathbf{x}_j^\intercal \mathbf{y}$, standard errors $\widehat{s}_j^2 = J^{-1}(\mathbf{y} - \mathbf{x}_j \widehat{\theta}_j)^\intercal (\mathbf{y} - \mathbf{x}_j \widehat{\theta}_j) / \mathbf{x}_j^\intercal \mathbf{x}_j$, 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 genotype matrix. Given different model parameters, we simulate data mirroring a wide range of genetic architectures (Supplementary Notes).

Data and Software Availability

Source code (with versions in both `R` and `Python 3`) and tutorials for implementing biologically annotated neural networks (BANNs) is publicly available online at `https://github.com/lcrawlab/BANNs`. All software for competing methods were fit using the default settings, unless otherwise stated in the main text. Links to competing methods, WTCHG mice data, and other relevant sources are also provided (See URLs). Data from the UK Biobank Resource [29] (`https://www.ukbiobank.ac.uk`) was made available under Application Number 22419. The FHS genotype and phenotype data is available in dbGaP [41] (`https://www.ncbi.nlm.nih.gov/gap`) with accession number phs000007.

Figures and Tables

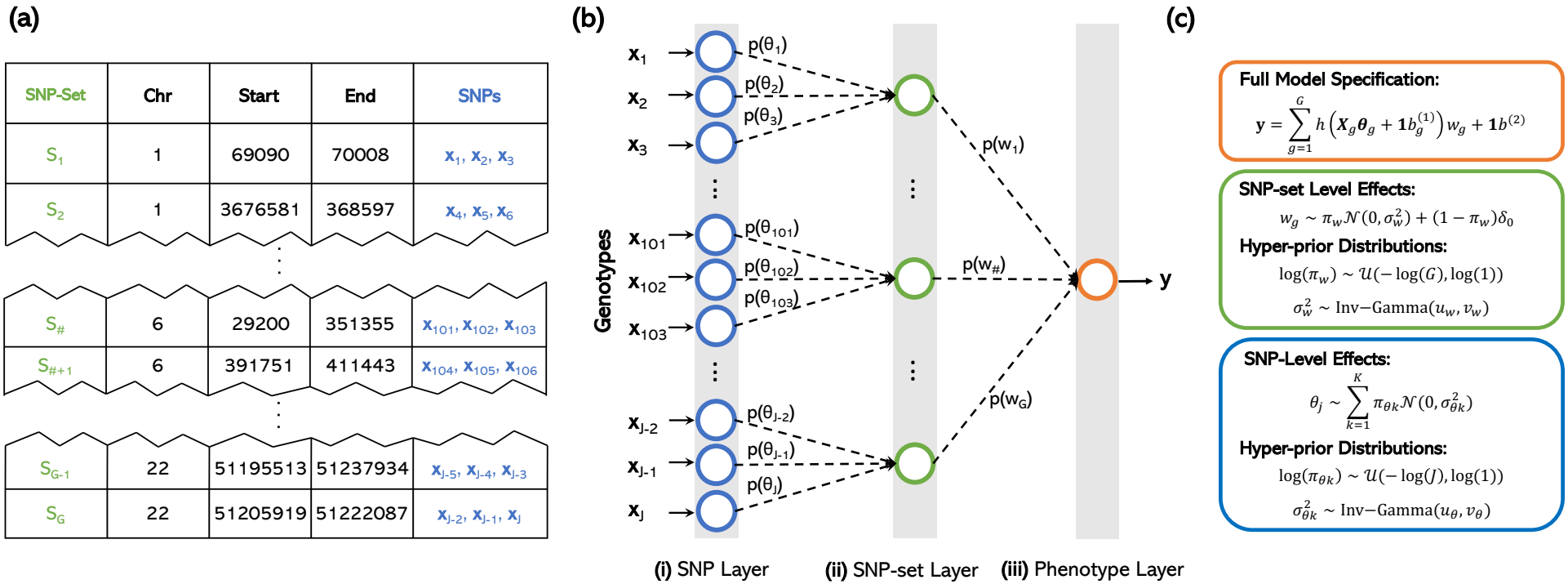
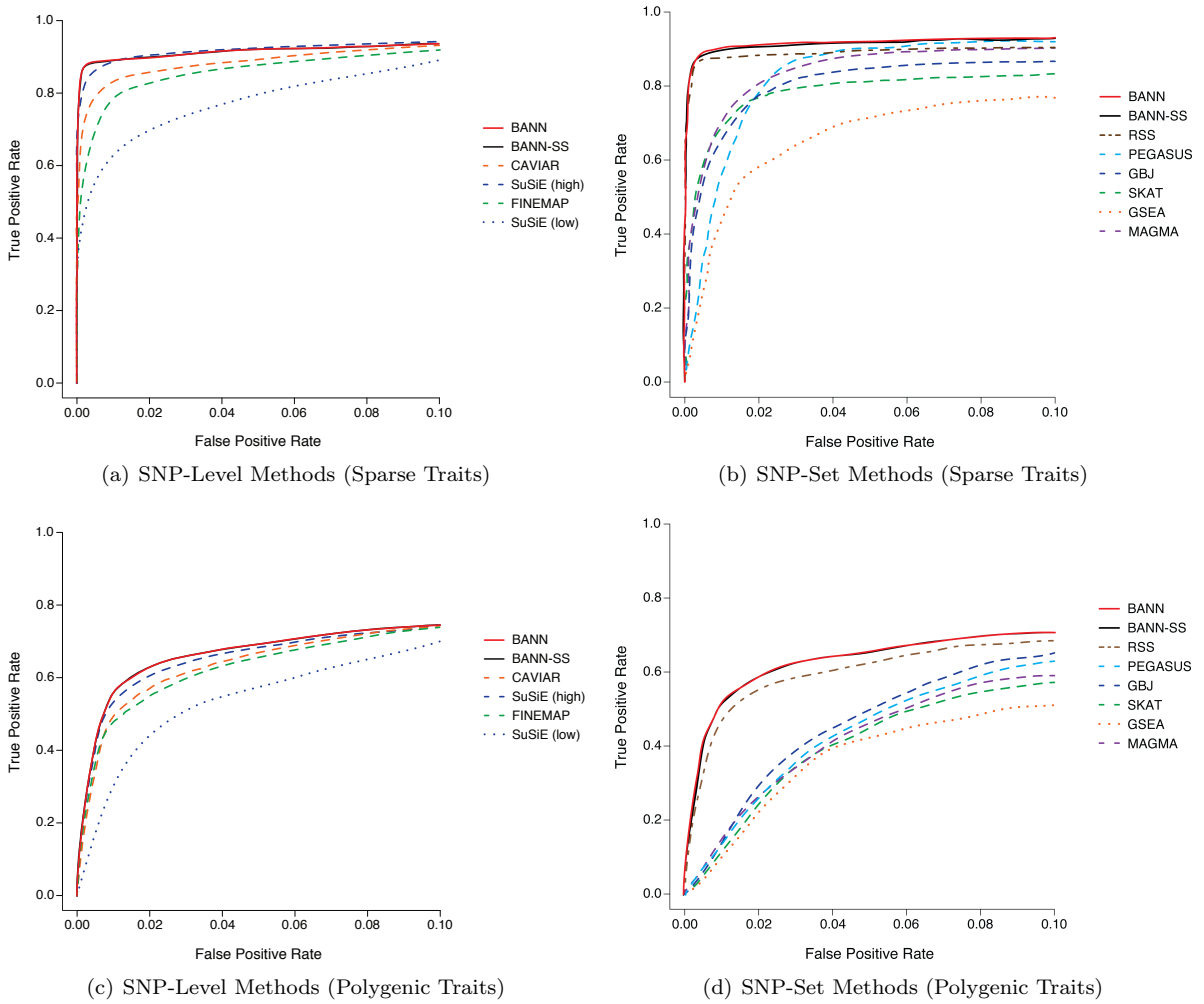


Figure 1. Biologically annotated neural networks (BANNs) allow for efficient multi-scale genomic analyses in a unified probabilistic framework by leveraging the hierarchical nature of enrichment studies to define network architecture. (a) The BANN framework requires an $N \times J$ matrix of individual-level 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 [42]. **(b)** A partially connected Bayesian neural network is constructed based on the annotated SNP groups. In the first hidden layer, only SNPs within the boundary of a gene are connected to the same node. Similarly, SNPs within the same intergenic region between genes are connected to the same node. Completing this specification for all SNPs gives the hidden layer the natural interpretation of being the "SNP-set" layer. **(c)** The hierarchical nature of the network is represented as nonlinear mixed model. The corresponding weights in both the SNP ($\boldsymbol{\theta}$) and SNP-set ($\mathbf{w}$) layers are treated as random variables with biologically motivated sparse prior distributions. Posterior inclusion probabilities (PIPs) $\boldsymbol{\gamma}_\theta$ and $\boldsymbol{\gamma}_w$ summarize associations at the SNP and SNP-set level, respectively. The BANN framework uses variational inference for efficient network training and incorporates nonlinear processing between network layers for accurate estimation of phenotypic variance explained (PVE).



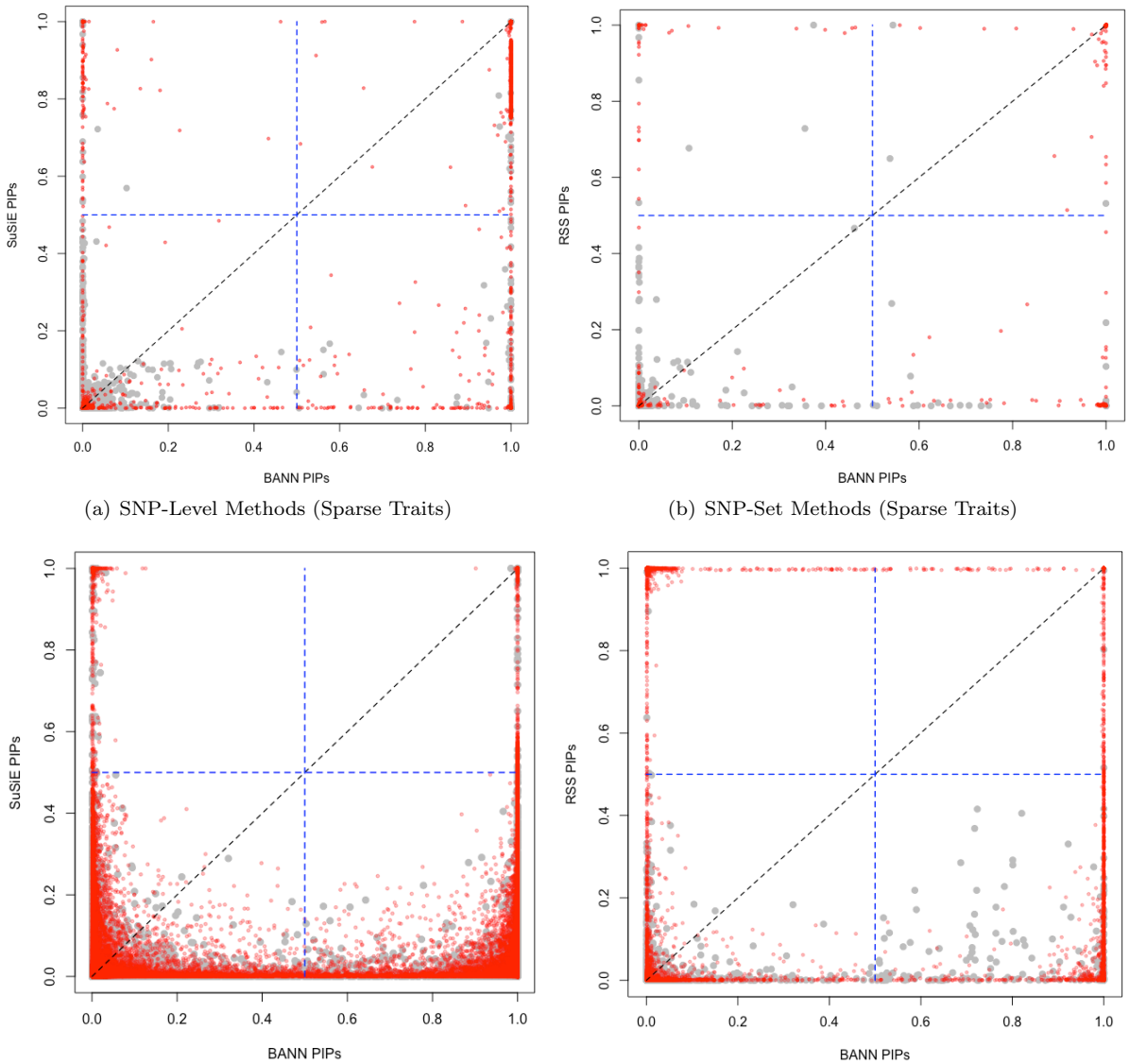
(a) SNP-Level Methods (Sparse Traits)

(b) SNP-Set Methods (Sparse Traits)

(c) SNP-Level Methods (Polygenic Traits)

(d) SNP-Set Methods (Polygenic Traits)

Figure 2. Receiver operating characteristic (ROC) curves comparing the performance of the BANN (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations. Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.6$ with only contributions from additive effects 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 0.125% and 3% of all SNPs located within the enriched SNP-sets, respectively. To derive results, the full genotype matrix and phenotypic vector are given to the BANN 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 [38], SuSiE [39], and FINEMAP [37]. 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 [36], 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 Supplementary Note, Section 8).



(a) SNP-Level Methods (Sparse Traits)

(b) SNP-Set Methods (Sparse Traits)

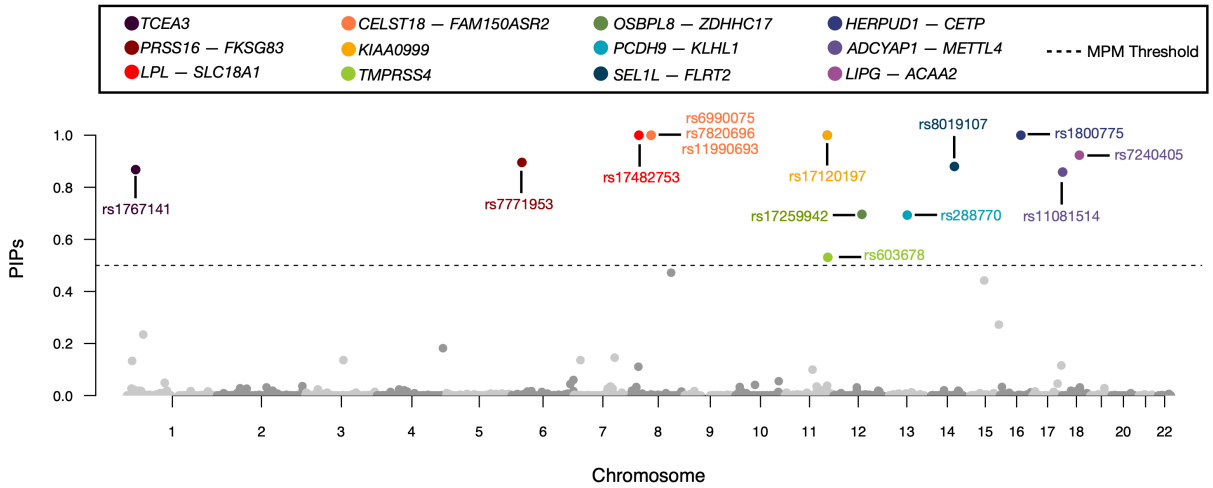
(c) SNP-Level Methods (Polygenic Traits)

(d) SNP-Set Methods (Polygenic Traits)

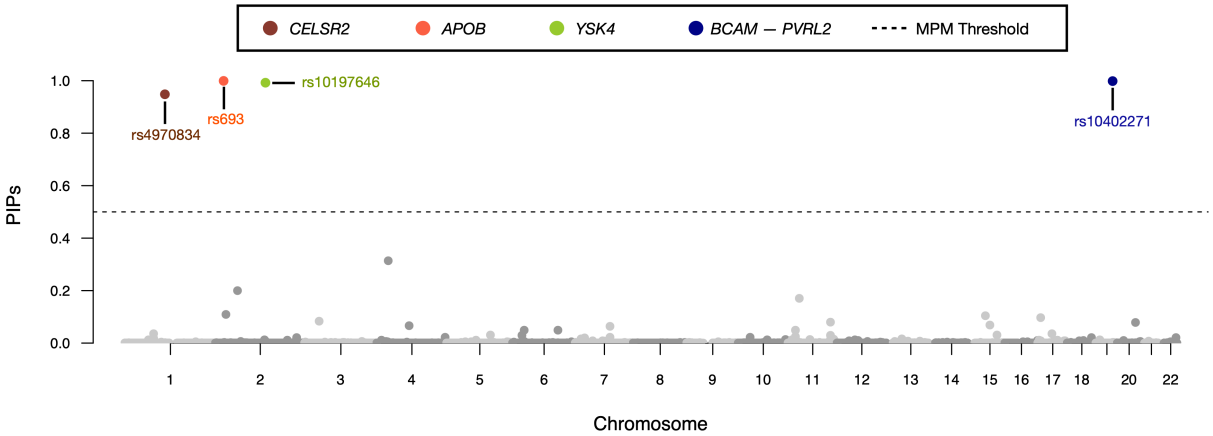
Figure 3. Scatter plots comparing how the integrative neural network training procedure enables the ability to identify associated SNPs and enriched SNP-sets in simulations. Quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.6$ with only contributions from additive effects 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 0.125% and 3% of all SNPs located within the enriched SNP-sets, respectively. Results are shown comparing the posterior inclusion probabilities (PIPs) derived by the BANN model on the x-axis and **(a, c)** SuSiE [39] 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) [48]. 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 BANN and SuSie/RSS, respectively. Each plot combines results from 100 simulated replicates (see Supplementary Notes).

Trait	SNP-Set	Chr	PIP (γ_θ)	Rank	Top SNP	PIP (γ_w)	Biological Relevance to Trait	Ref(s)
BMI	*Dmd*	X	0.900	1	rs3090667	0.600	Dystrophin loss has integrative effects on metabolic function	[56]
	Mir466q-Slc2a2	3	0.816	3	rs6269713	0.477	Encodes *GLUT2* and shown to vary with BMI in humans	[94]
	Gm22219-Mc4r	18	0.740	5	rs3696955	0.039	*MC4R* variants protect against obesity in humans	[57]
CD8+	*Gm46177-Gm30088*	1	0.968	1	mhcCD8a3	1.000	Intergenic region containing lupus related QTL that are linked to CD8+ T cell differentiation	[95–97]
	Btbd9	17	0.866	7	CEL-17_31069801	1.000	Contains SNPs associated with restless leg syndrome and is positioned within a QTL associated with iron concentration	[98, 99]
	hlb156	17	0.720	8	CEL-17_31069801	1.000	Heart, lung, and blood functionally related gene	[54]
HDL	*Pphc2*	4	0.976	3	rs3724711	1.000	Involved in cholesterol metabolic processes	[100]
	Ctnna2	6	0.886	8	rs3710419	1.000	Shown to be associated with the abnormality of cholesterol metabolism in different GWA studies	[101]
	hlb156	17	0.589	8	CEL-17_31069801	1.000	Heart, lung, and blood functionally related gene	[54]
LDL	*Btbd9*	17	0.983	1	CEL-17_31069801	1.000	Mutations in this gene have been linked to Bardet-Biedl syndrome, for which truncal obesity is a cardinal symptom	[102, 103]
	Pphc2	4	0.941	3	rs3724711	1.000	Involved in cholesterol metabolic processes	[100]
	Syt14	1	0.852	7	rs3654706	0.001	Also known as the RIKEN gene and involved in processes dealing with lipid binding	[104–107]
MCH	*Btbd9*	17	0.905	2	CEL-17_31069801	1.000	Contains SNPs associated with restless leg syndrome and is positioned within a QTL associated with iron concentration	[98, 99]
	Picalm	7	0.648	8	rs3704554	0.070	Mutations in this gene are responsible for the hematopoietic and iron metabolism abnormalities in mice	[108]
	Ebf1	11	0.500	10	rs3693846	0.009	Knockout experiments with this gene have been linked to B-cell deficiency and other hematopoietic system changes in mice	[109]
Weight	*Wdpcp*	11	0.969	1	rs13481023	1.000	Mutations in this gene have been linked to Bardet-Biedl syndrome, for which truncal obesity is a cardinal symptom	[102, 103]
	Chrm2	6	0.882	3	rs3676478	0.012	The genotypic variance of this gene has been shown to be predictive of longitudinal BMI and obesity status	[110, 111]
	Csmd1	8	0.759	5	rs3709567	0.001	Knockout experiments with this gene have been linked to weight gain in mice	[112]

Table 1. Notable enriched SNP-sets after applying the BANN framework to six quantitative traits in heterogenous stock of mice from the Wellcome Trust Centre for Human Genetics. [40]. 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 desert 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 BANN model on individual-level data. A SNP-set is considered enriched if it has a PIP $\boldsymbol{\gamma}_w \geq 0.5$ (i.e., the "median probability model" threshold [48]). We also report the "top" associated SNP within each region and its corresponding PIP $\boldsymbol{\gamma}_\theta$. The reference column details literature sources that have previously suggested some level of association between the each genomic region and the traits of interest. See Supplementary Tables 11-16 for the complete list of SNP and SNP-set level results. *: Multiple SNP-sets were tied for this ranking.



(a) High-Density Lipoprotein (HDL)



(b) Low-Density Lipoprotein (LDL)

Figure 4. Manhattan plot of variant-level fine mapping results for high-density and low-density lipoprotein (HDL and LDL, respectively) traits in the Framingham Heart Study [41]. Posterior inclusion probabilities (PIP) for the neural network weights are derived from the BANN 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 [48]. 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 [42]. Unannotated SNPs located within the same genomic desert were labeled as being within the "intergenic region" between two genes. These regions are labeled as *Gene1-Gene2* in the legend. Gene set enrichment analyses for these SNP-sets can be found in Supplementary Figure 27.

Trait	SNP-Set	Chr	PIP (γ_w)	Rank	Top SNP	PIP (γ_θ)	Biological Relevance to Trait	Ref(s)
HDL	*HERPUD1-CETP*	16	0.999	1*	rs7240405	0.923	Previously found to be associated with HDL in multiple multiethnic GWA studies	[60–63]
	ST18-FAM150A	8	0.999	1*	rs6990075	1.000	Suppression of mouse ortholog has been shown facilitate high glucose-induced cell death	[113]
	TCEA3	1	0.989	2	rs1767141	0.868	Found to be commonly associated with total cholesterol measurement across multiple cohorts	[114]
LDL	*CELSR2*	1	0.989	1	rs4970834	0.948	Member of the cadherin superfamily and commonly found to be associated with LDL across multiple multiethnic cohorts	[115–117]
	BCAM-PVRL2	19	0.987	2	rs10402271	0.998	*BCAM* encodes a Lutheran blood group glycoprotein, while *PVRL2* is a cholesterol-responsive gene. Both have been linked to LDL response	[118–120]
	APOB	2	0.976	3	rs693	0.999	This gene produces the main apolipoprotein of chylomicrons and low density lipoproteins (LDL), and is the ligand for the LDL receptor	[118, 121, 122]

Table 2. Top three enriched SNP-sets after applying the BANN framework to high-density and low-density lipoprotein (HDL and LDL, respectively) traits in the Framingham Heart Study [41]. Here, SNP-set annotations are based on gene boundaries defined by the NCBI's RefSeq database in the UCSC Genome Browser [42]. Unannotated SNPs located within the same genomic desert 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 BANN model on individual-level data. A SNP-set is considered enriched if it has a PIP $\boldsymbol{\gamma}_w \geq 0.5$ (i.e., the "median probability model" threshold [48]). We also report the "top" associated SNP within each region and its corresponding PIP ($\boldsymbol{\gamma}_\theta$). The reference column details literature sources that have previously suggested some level of association between the each genomic region and the traits of interest. See Supplementary Tables 17 and 18 for the complete list of SNP and SNP-set level results. *: Multiple SNP-sets were tied for this ranking.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

29. 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.

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

31. 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.

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

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

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

35. 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.

36. 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.

37. 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.

38. 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.

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

40. 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.

41. 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.

42. 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.

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

44. 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.

45. 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.

46. 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.

47. 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.

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

49. 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.

50. 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.

51. 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.

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

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

54. Bult CJ, Blake JA, Smith CL, Kadin JA, Richardson JE. Mouse Genome Database (MGD). Nucleic Acids Res. 2019;47(D1):D801–D806.

55. 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.

56. 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.

57. 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.

58. 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.

59. 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.

60. 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.

61. 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.

62. 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.

63. 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.

64. 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.

65. 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.

66. 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.

67. 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.

68. 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.

69. 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.

70. 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.

71. 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.

72. 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.

73. 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.

74. 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.

75. 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.

76. 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.

77. 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.

78. 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.

79. 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.

80. Breslow NE, Lin X. Bias correction in generalised linear mixed models with a single component of dispersion. Biometrika. 1995 2020/06/24/;82(1):81–91. Available from: www.jstor.org/stable/2337629.

81. 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.

82. 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.

83. 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.

84. 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.

85. 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.

86. 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.

87. 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.

88. 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.

89. 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.

90. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15(1):1929–1958.

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

92. 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.

93. 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.

94. 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.

95. 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.

96. 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.

97. 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.

98. 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.

99. 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.

100. 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.

101. 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.

102. 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.

103. 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.

104. 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.

105. 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.

106. Diez-Roux G, Banfi S, Sultan M, Geffers L, Anand S, Rozado D, et al. A High-Resolution Anatomical Atlas of the Transcriptome in the Mouse Embryo. PLoS Biol. 2011;9(1):e1000582. Available from: https://doi.org/10.1371/journal.pbio.1000582.

107. 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.

108. 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.

109. Lin H, Grosschedl R. Failure of B-cell differentiation in mice lacking the transcription factor EBF. Nature. 1995;376(6537):263–267.

110. 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.

111. 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.

112. 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.

113. 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.

114. 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.

115. 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.

116. 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.

117. 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.

118. 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.

119. 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.

120. 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.

121. 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.

122. 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.