Interpretable Multi-scale Statistical Methods for Genetic Association Studies

by

Wei Cheng

A dissertation submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy at the Center of Computational Molecular Biology at Brown University

Providence, Rhode Island
May 2022
This dissertation by Wei Cheng is accepted in its present form by
the Center of Computational Molecular Biology as satisfying the dissertation requirement
for the degree of Doctor of Philosophy.

Date ________________  Dr. Sohini Ramachandran, co-Advisor

Date ________________  Dr. Lorin Crawford, co-Advisor

Recommended to the Graduate Council

Date ________________  Dr. Eliezer Upfal, Reader

Date ________________  Dr. Ritambhara Singh, Reader

Date ________________  Dr. Barbara Engelhardt, Reader

Approved by the Graduate Council

iii

Date ________________  Dr. Andrew Campbell, Dean
  of the Graduate School
Wei Cheng
164 Angell Street, Brown University, Providence, RI, 02912, USA
weicheng1226@gmail.com • +1(401)330-9294 • weichengv.github.io

EDUCATION

Brown University
Ph.D. in Computer Science and Computational Biology Aug 2017 - Present
M.S. in Computer Science

Cornell University

China Agricultural University
B.S. in Computational Biology Aug 2012 - May 2014

PROFESSIONAL EXPERIENCE

Facebook - Signal Intelligence team Jun 2021 - Sep 2021
Summer Intern, Software Engineer (Machine Learning).
• Built end-to-end encoding and decoding pipelines to recover ads signals under privacy friendly attribution API. Developed tools to compute statistical metrics for evaluating encoding and decoding pipeline performance.
• Implemented scalable conversion calibration algorithms using feature compression method with low bandwidth training to mitigate signal loss under extreme information bottleneck. Developed new non-linear feature compression method that consistently outperforms baseline measured by multiple metrics.

PUBLICATIONS

(* CO-FIRST AUTHORS)


SOFTWARE

ESNN: Ensemble of Single-effect Neural Network
BANNs: Biologically Annotated Neural Networks
HEBAE: Hierarchical Empirical Bayes Auto-Encoder
gene-ε: A Recalibrated Hypothesis Test for Sets of SNP-Level Summary Statistics

CONFERENCES


SKILLS

R, Python, TensorFlow, PyTorch, Matlab, Linux, JAVA.

COURSES


TEACHING

Teaching assistant – Statistical Analysis of Biological Data

SERVICE

Reviewer - AISTATS 2022
Acknowledgements

First, I would like to thank my advisor Sohini Ramachandran. Not only does she teach me how to be a scientist, but also she shows me how to be a leader, a teacher, a speaker, and a friend. Sohini is one of the most patient and open-minded people I know and I am deeply grateful to become her advisee. I would also like to thank my co-advisor Lorin Crawford. I want to thank him for teaching me how to be a statistician and the time we spent together on reading, coding, derivation, and discussion. Besides, I would also like to thank my committee members Eli Upfal, Ritambhara Singh, and Barbara Engelhardt for all the valuable advice and support.

I also want to thank the members of the Ramachandran lab: Sam, Sahar, Kaleigh, Priya, Lauren, Elsie, Cole, Selena, Kate, Carlos, Michael, and Greg for being collaborators and for being my friends. In addition, I’d like to thank the program manager, Nathaniel Gill, and all other staff in CCMB, EEB, CS, and Brown for their support.

Lastly, I would like to thank my family and all my other friends, especially my parents Shengyi Cheng and Huaidi Wu for their love, support, kindness, and encouragement. I would also like to thank my girlfriend Yue You for her companion along the road. Many thanks to all the people who have kindly helped me as I am pursuing my dream and becoming who I really am today.
Abstract

In genetics, one fundamental goal is to understand which mutations in the human genome have significant effects on complex traits and diseases. Achieving this goal requires performing statistical inference such as hypothesis testing to generate interpretable results. Genome-wide association studies (GWAS) often use marginal linear regression to test the association between each mutation with the trait independently. Even though GWAS enjoys simplicity and scalability, it still suffers from several problems due to the complex nature of the genetic data. For example, the strong correlation structure of genotype states along the genome can induce a large number of false positives under the GWAS framework. Further, linear models can only explain additive variation in traits and cannot account for nonlinear effects such as dominance effects and epistatic interactions. This dissertation addresses these problems via multiple projects and two conceptual themes. In theme 1, we developed a gene-level association method called gene-$\epsilon$ that uses a reformulated null model for association testing with shrinkage on GWAS summary statistics. We show that gene-$\epsilon$ reduces false positives compared to GWAS and competing gene-level association approaches using extensive simulations, and we apply gene-$\epsilon$ to quantitative traits in UKBiobank data to identify novel associated genes. In addition, we explored the potential of gene-$\epsilon$ to replicate findings in multiple ancestries and in case-control studies. In theme 2, we develop two nonlinear methods: Biologically Annotated Neural Networks (BANN) and the Ensemble of Single-effect Neural Network (ESNN). BANN uses coordinate ascent variational inference to simultaneously perform association tests on SNPs and genes together. ESNN, on the other hand, uses black-box variational inference and can be used to quantify uncertainty for associations when genetic data are highly collinear by providing
credible sets. ESNN can be applied to both continuous traits and binary traits with any neural network architecture. For BANN and ESNN, we demonstrate their power and interpretability for association studies using extensive simulations. We also apply them to different real-world datasets including both quantitative and binary traits for biological discovery.
Contents

List of Tables xiv

List of Figures xx

1 Estimation of Non-null SNP Effect Size Distributions Enables the Detection of Enriched Genes Underlying Complex Traits 4
1.1 Abstract 4
1.2 Introduction 5
1.3 Results 8

1.3.1 Overview of gene-ε 8
1.3.2 Performance Comparisons in Simulation Studies 10
1.3.3 Characterizing Genetic Architecture of Quantitative Traits in the UK Biobank 13
1.3.4 gene-ε Identifies Refined List of Genetic Enrichments 14
1.4 Discussion 20
1.5 Materials and Methods 23

1.5.1 Traditional Association Tests using Summary Statistics 23
1.5.2 Regularized Regression for GWA Summary Statistics 24
1.5.3 Estimating the SNP-Level Null Threshold 26
1.5.4 Regularized GWA Summary Statistics under the Null Hypothesis . . . . . . . 29
1.5.5 Using the SNP-Level Null Threshold to Detect Enriched Genes . . . . . . . 31
1.5.6 Simulation Studies . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 33
1.6 Software Details . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 34
1.7 URLs . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 34

2 Multi-scale Inference of Genetic Trait Architecture using Biologically Anno-
tated Neural Networks 36
2.1 Abstract . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 36
2.2 Author Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 37
2.3 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 38
2.4 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 40
  2.4.1 BANNs Framework Overview . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 40
  2.4.2 Power to Detect SNPs and SNP-Sets in Simulation Studies . . . . . . . . . . 44
  2.4.3 Estimating Total Phenotypic Variance Explained in Simulation Studies . . . . 52
  2.4.4 Mapping Genomic Enrichment in Heterogenous Stock of Mice . . . . . . . . . 53
  2.4.5 Analyzing Lipoproteins in the Framingham Heart Study . . . . . . . . . . . . 57
  2.4.6 Independent Lipoprotein Study using the UK Biobank . . . . . . . . . . . . . 61
2.5 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 62
2.6 URLs . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 65
2.7 Acknowledgements . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 66
2.8 Funding . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 67
2.9 Author Contributions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 67
2.10 Material and Methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 68
# List of Tables

1.1 Computational time for running gene-ε and other gene-level association approaches, as a function of the total number genes analyzed and the number of SNPs within each gene. ................................................................. 12

1.2 Top three newly identified candidate genes reported by gene-ε for the six quantitative traits studied in the UK Biobank (using imputed genotypes with gene boundaries defined by the NCBI’s RefSeq database in the UCSC Genome Browser (195)). . . . . . . 19

2.1 Notable enriched SNP-sets after applying the BANNs framework to six quantitative traits in heterogenous stock of mice from the Wellcome Trust Centre for Human Genetics. ................................................................. 56

2.2 Top three enriched SNP-sets after applying the BANNs framework to high-density and low-density lipoprotein (HDL and LDL, respectively) traits in the Framingham Heart Study. ................................................................. 60

A.1 Empirical power and false discovery rates (FDR) for detecting enriched genes (genes containing at least one causal SNP) after correcting for multiple hypothesis testing in simulations ($N = 5,000; h^2 = 0.2$). ................................................................. 133

A.2 Empirical power and false discovery rates (FDR) for detecting enriched genes (genes containing at least one causal SNP) after correcting for multiple hypothesis testing in simulations ($N = 10,000; h^2 = 0.2$). ................................................................. 134
A.3 Empirical power and false discovery rates (FDR) for detecting enriched genes (genes containing at least one causal SNP) after correcting for multiple hypothesis testing in simulations ($N = 5,000; h^2 = 0.6$).

A.4 Empirical power and false discovery rates (FDR) for detecting enriched genes (genes containing at least one causal SNP) after correcting for multiple hypothesis testing in simulations ($N = 10,000; h^2 = 0.6$).

A.5 Empirical power and false discovery rates (FDR) for detecting enriched genes (genes containing at least one causal SNP) after correcting for multiple hypothesis testing in simulations with population stratification ($N = 5,000; h^2 = 0.2$).

A.6 Empirical power and false discovery rates (FDR) for detecting enriched genes (genes containing at least one causal SNP) after correcting for multiple hypothesis testing in simulations with population stratification ($N = 10,000; h^2 = 0.2$).

A.7 Empirical power and false discovery rates (FDR) for detecting enriched genes (genes containing at least one causal SNP) after correcting for multiple hypothesis testing in simulations with population stratification ($N = 5,000; h^2 = 0.6$).

A.8 Empirical power and false discovery rates (FDR) for detecting enriched genes (genes containing at least one causal SNP) after correcting for multiple hypothesis testing in simulations with population stratification ($N = 10,000; h^2 = 0.6$).

A.9 Empirical power and false discovery rates (FDR) for detecting enriched genes (genes containing at least one causal SNP) after correcting for multiple hypothesis testing in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer ($N = 5,000; h^2 = 0.2$).

A.10 Empirical power and false discovery rates (FDR) for detecting enriched genes (genes containing at least one causal SNP) after correcting for multiple hypothesis testing in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer ($N = 10,000; h^2 = 0.2$).
A.11 Empirical power and false discovery rates (FDR) for detecting enriched genes (genes containing at least one causal SNP) after correcting for multiple hypothesis testing in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer (\(N = 5,000; h^2 = 0.6\)). .......................................................... 143

A.12 Empirical power and false discovery rates (FDR) for detecting enriched genes (genes containing at least one causal SNP) after correcting for multiple hypothesis testing in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer (\(N = 10,000; h^2 = 0.6\)). .......................................................... 144

A.13 Empirical power and false discovery rates (FDR) for detecting enriched genes (genes containing at least one causal SNP) after correcting for multiple hypothesis testing in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer and with population stratification (\(N = 5,000; h^2 = 0.2\)). .......................................................... 145

A.14 Empirical power and false discovery rates (FDR) for detecting enriched genes (genes containing at least one causal SNP) after correcting for multiple hypothesis testing in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer and with population stratification (\(N = 10,000; h^2 = 0.2\)). .......................................................... 146

A.15 Empirical power and false discovery rates (FDR) for detecting enriched genes (genes containing at least one causal SNP) after correcting for multiple hypothesis testing in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer and with population stratification (\(N = 5,000; h^2 = 0.6\)). .......................................................... 147

A.16 Empirical power and false discovery rates (FDR) for detecting enriched genes (genes containing at least one causal SNP) after correcting for multiple hypothesis testing in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer and with population stratification (\(N = 10,000; h^2 = 0.6\)). .......................................................... 148

A.17 Empirical type I error estimates using different gene-\(c\) approaches. .......................... 149

A.18 Characterization of the genetic architectures of six traits assayed in European-ancestry individuals in the UK Biobank. .......................................................... 150

A.19 Significant genes for body height in the UK Biobank analysis using gene-\(c\)-EN. .......................... 151

xvi
A.20 Significant genes for body mass index (BMI) in the UK Biobank analysis using gene-\(\varepsilon\)-EN. .......................................................... 151
A.21 Significant genes for mean corpuscular volume (MCV) in the UK Biobank analysis using gene-\(\varepsilon\)-EN. .......................................................... 152
A.22 Significant genes for mean platelet volume (MPV) in the UK Biobank analysis using gene-\(\varepsilon\)-EN. .......................................................... 152
A.23 Significant genes for platelet count (PLC) in the UK Biobank analysis using gene-\(\varepsilon\)-EN. .......................................................... 153
A.24 Significant genes for waist-hip ratio (WHR) in the UK Biobank analysis using gene-\(\varepsilon\)-EN. .......................................................... 153
A.25 Characterization of the genetic architectures of six traits assayed in European-ancestry individuals in the UK Biobank (using un-imputed genotypes). ....................... 154
A.26 Comparison of the different gene-\(\varepsilon\) approaches on the six quantitative traits assayed in European-ancestry individuals from the UK Biobank un-imputed genotyped data. 155
A.27 Comparison of the different gene-\(\varepsilon\) approaches on the six quantitative traits assayed in European-ancestry individuals from the UK Biobank un-imputed genotyped data with gene boundaries augmented by a 50 kilobase (kb) buffer. ....................... 156
B.1 Comparing the empirical power and false discovery rates (FDR) of the BANNs framework against competing SNP and SNP-set mapping approaches in simulations. .... 219
B.2 Comparing the empirical power and false discovery rates (FDR) of the BANNs framework against competing SNP and SNP-set mapping approaches in simulations. .... 220
B.3 Comparing the empirical power and false discovery rates (FDR) of the BANNs framework against competing SNP and SNP-set mapping approaches in simulations with population structure (European cohort). ....................... 221
B.4 Comparing the empirical power and false discovery rates (FDR) of the BANNs framework against competing SNP and SNP-set mapping approaches in simulations with population structure (European cohort). ....................... 222

xvii
B.5 Comparing the empirical power and false discovery rates (FDR) of the BANNs framework against competing SNP and SNP-set mapping approaches in simulations. . . . 223
B.6 Comparing the empirical power and false discovery rates (FDR) of the BANNs framework against competing SNP and SNP-set mapping approaches in simulations. . . . 224
B.7 Comparing the empirical power and false discovery rates (FDR) of the BANNs framework against competing SNP and SNP-set mapping approaches in simulations with
population structure (European cohort). . . . . . . . . . . . . . . . . . . . . . . . . . 225
B.8 Comparing the empirical power and false discovery rates (FDR) of the BANNs framework against competing SNP and SNP-set mapping approaches in simulations with
population structure (European cohort). . . . . . . . . . . . . . . . . . . . . . . . . . 226
B.9 Computational time for running Bayesian annotated neural networks (BANNs) and
other SNP-level association mapping approaches, as a function of the total number
SNPs analyzed and the number of samples in the data. . . . . . . . . . . . . . . . . . 227
B.10 Computational time for running Bayesian annotated neural networks (BANNs) and
other SNP-set level enrichment approaches, as a function of the total number SNPsets analyzed and the number of SNPs within each SNP-set. . . . . . . . . . . . . . . 228
B.11 SNP and SNP-set results for body mass index (BMI) in the heterogenous stock of
mice from the Wellcome Trust Centre for Human Genetics. . . . . . . . . . . . . . . 229
B.12 SNP and SNP-set results for body weight in the heterogenous stock of mice from the
Wellcome Trust Centre for Human Genetics. . . . . . . . . . . . . . . . . . . . . . . . 229
B.13 SNP and SNP-set results for percentage of CD8+ cells in the heterogenous stock of
mice from the Wellcome Trust Centre for Human Genetics. . . . . . . . . . . . . . . 230
B.14 SNP and SNP-set results for high-density lipoprotein (HDL) cholesterol in the heterogenous stock of mice from the Wellcome Trust Centre for Human Genetics. . . . . 230
B.15 SNP and SNP-set results for low-density lipoprotein (LDL) cholesterol in the heterogenous stock of mice from the Wellcome Trust Centre for Human Genetics. . . . . 231
B.16 SNP and SNP-set results for mean corpuscular hemoglobin (MCH) in the heterogenous stock of mice from the Wellcome Trust Centre for Human Genetics. . . . . . . 231
xviii


B.17 Notable enriched SNP-sets after applying the BANNs framework to high-density and low-density lipoprotein (HDL and LDL, respectively) traits in the Framingham Heart Study where each SNP-set annotation has been augmented with a ±500 kilobase (kb) buffer to account for possible regulatory elements. ........................................ 232

B.18 SNP and SNP-set results for high-density lipoprotein (HDL) cholesterol in individuals assayed within the Framingham Heart Study. ........................................ 233

B.19 SNP and SNP-set results for low-density lipoprotein (LDL) cholesterol in individuals assayed within the Framingham Heart Study. ........................................ 233

B.20 Complete summary of the results after applying BANNs, SuSiE, and RSS to high-density lipoprotein (HDL) and low-density lipoprotein (LDL) cholesterol in both individuals assayed within the Framingham Heart Study and ten thousand randomly sampled individuals of European ancestry from the UK Biobank. 234

B.21 SNP and SNP-set results for high-density lipoprotein (HDL) cholesterol in ten thousand randomly sampled individuals of European ancestry from the UK Biobank. 234

B.22 SNP and SNP-set results for low-density lipoprotein (LDL) cholesterol in ten thousand randomly sampled individuals of European ancestry from the UK Biobank. 235
List of Figures

1.1 Illustration of null hypothesis assumptions for the distribution of GWA SNP-level effect sizes according to different views on underlying genetic architectures. .... 6
1.2 Schematic overview of gene-ε: our new gene-level association approach accounting for spurious nonzero SNP-level effects ...................................................................... 16
1.3 Receiver operating characteristic (ROC) and precision-recall curves comparing the performance of gene-ε and competing approaches in simulations ($N = 10,000; h^2 = 0.6$) ................................................................. 17
1.4 Gene-level association results from applying gene-ε to body height (panels A and C) and mean platelet volume (MPV; panels B and D), assayed in European-ancestry individuals in the UK Biobank ......................................................... 18

2.1 Biologically annotated neural networks (BANNs) allow for efficient multi-scale genotype-phenotype analyses in a unified probabilistic framework by leveraging the hierarchical nature of enrichment studies to define network architecture. ................. 42
2.2 Receiver operating characteristic (ROC) curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations (British cohort). ................. 49
2.3 Scatter plots comparing how the integrative neural network training procedure enables the ability to identify associated SNPs and enriched SNP-sets in simulations (British cohort). ......................................................... 50
2.4 Manhattan plot of variant-level association mapping results for high-density and low-density lipoprotein (HDL and LDL, respectively) traits in the Framingham Heart Study 59
3.1 Credible sets comparisons of ESNN with SuSiE. ........................................ 85
3.2 Receiver Operating Characteristic (ROC) curves for simulation studies of different scenarios. Results are based on 200 data replicates. ................................. 87
3.3 Posterior inclusion probabilities (PIP) of ESNN and SuSiE in the WTCCC analysis. 89

A.1 Simulation study results showing the Pearson correlation between various degrees of gene-ε regularized SNP-level effect size estimates and the true effect sizes that generated the complex traits. ................................................................. 104

A.2 (A, C) Receiver operating characteristic (ROC) and (B, D) precision-recall curves comparing the performance of gene-ε and competing approaches in simulations (N = 5,000; h² = 0.2). ................................................................. 105

A.3 (A, C) Receiver operating characteristic (ROC) and (B, D) precision-recall curves comparing the performance of gene-ε and competing approaches in simulations (N = 10,000; h² = 0.2). ................................................................. 106

A.4 (A, C) Receiver operating characteristic (ROC) and (B, D) precision-recall curves comparing the performance of gene-ε and competing approaches in simulations (N = 5,000; h² = 0.6). ................................................................. 107

A.5 (A, C) Receiver operating characteristic (ROC) and (B, D) precision-recall curves comparing the performance of gene-ε and competing approaches in simulations with population stratification (N = 5,000; h² = 0.2). ................................................................. 108

A.6 (A, C) Receiver operating characteristic (ROC) and (B, D) precision-recall curves comparing the performance of gene-ε and competing approaches in simulations with population stratification (N = 10,000; h² = 0.2). ................................................................. 109

A.7 (A, C) Receiver operating characteristic (ROC) and (B, D) precision-recall curves comparing the performance of gene-ε and competing approaches in simulations with population stratification (N = 5,000; h² = 0.6). ................................................................. 110

xxi
A.8 (A, C) Receiver operating characteristic (ROC) and (B, D) precision-recall curves comparing the performance of gene-ε and competing approaches in simulations with population stratification ($N = 10,000; h^2 = 0.6$). 111

A.9 (A, C) Receiver operating characteristic (ROC) and (B, D) precision-recall curves comparing the performance of gene-ε and competing approaches in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer ($N = 5,000; h^2 = 0.2$). 112

A.10 (A, C) Receiver operating characteristic (ROC) and (B, D) precision-recall curves comparing the performance of gene-ε and competing approaches in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer ($N = 10,000; h^2 = 0.2$). 113

A.11 (A, C) Receiver operating characteristic (ROC) and (B, D) precision-recall curves comparing the performance of gene-ε and competing approaches in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer ($N = 5,000; h^2 = 0.6$). 114

A.12 (A, C) Receiver operating characteristic (ROC) and (B, D) precision-recall curves comparing the performance of gene-ε and competing approaches in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer ($N = 10,000; h^2 = 0.6$). 115

A.13 (A, C) Receiver operating characteristic (ROC) and (B, D) precision-recall curves comparing the performance of gene-ε and competing approaches in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer and with population stratification ($N = 5,000; h^2 = 0.2$). 116

A.14 (A, C) Receiver operating characteristic (ROC) and (B, D) precision-recall curves comparing the performance of gene-ε and competing approaches in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer and with population stratification ($N = 10,000; h^2 = 0.2$). 117

A.15 (A, C) Receiver operating characteristic (ROC) and (B, D) precision-recall curves comparing the performance of gene-ε and competing approaches in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer and with population stratification ($N = 5,000; h^2 = 0.6$). 118
A.16 (A, C) Receiver operating characteristic (ROC) and (B, D) precision-recall curves comparing the performance of gene-ε and competing approaches in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer and with population stratification \((N = 10,000; h^2 = 0.6)\). .................................................... 119

A.17 Scatter plots assessing how regularization on SNP-level summary statistics affects the ability to identify enriched genes in simulations \((h^2 = 0.2)\). .............................................. 120

A.18 Scatter plots assessing how regularization on SNP-level summary statistics affects the ability to identify enriched genes in simulations \((h^2 = 0.6)\). .............................................. 121

A.19 Scatter plots assessing how regularization on SNP-level summary statistics affects the ability to identify enriched genes in simulations with population stratification \((h^2 = 0.2)\). .................................................... 122

A.20 Scatter plots assessing how regularization on SNP-level summary statistics affects the ability to identify enriched genes in simulations with population stratification \((h^2 = 0.6)\). .................................................... 123

A.21 Scatter plots assessing how regularization on SNP-level summary statistics affects the ability to identify enriched genes in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer \((h^2 = 0.2)\). .................................................... 124

A.22 Scatter plots assessing how regularization on SNP-level summary statistics affects the ability to identify enriched genes in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer \((h^2 = 0.6)\). .................................................... 125

A.23 Scatter plots assessing how regularization on SNP-level summary statistics affects the ability to identify enriched genes in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer and with population stratification \((h^2 = 0.2)\). .............. 126

A.24 Scatter plots assessing how regularization on SNP-level summary statistics affects the ability to identify enriched genes in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer and with population stratification \((h^2 = 0.6)\). .............. 127
A.25 Gene-level association results from applying gene-ε to body height (panels A and C) and mean platelet volume (MPV; panels B and D), assayed in European-ancestry individuals in the UK Biobank with UCSC RefSeq gene boundaries augmented by a 50 kilobase (kb) buffer. 128

A.26 Gene-level association results from applying gene-ε to body mass index (BMI), assayed in European-ancestry individuals in the UK Biobank. 129

A.27 Gene-level association results from applying gene-ε to mean corpuscular volume (MCV), assayed in European-ancestry individuals in the UK Biobank. 130

A.28 Gene-level association results from applying gene-ε to platelet count (PLC), assayed in European-ancestry individuals in the UK Biobank. 131

A.29 Gene-level association results from applying gene-ε to waist-hip ratio (WHR), assayed in European-ancestry individuals in the UK Biobank. 132

A.30 Receiver Operating Characteristic (ROC) curves and Precision and Recall curves of GWAS for sparse case-control simulations. 158

A.31 Receiver Operating Characteristic (ROC) curves and Precision and Recall curves of GWAS for polygenic case-control simulations. 159

A.32 Receiver Operating Characteristic (ROC) curves and Precision and Recall curves of gene level association methods for sparse case-control simulations. 160

A.33 Receiver Operating Characteristic (ROC) curves and Precision and Recall curves of gene level association methods for polygenic case-control simulations. 161

B.1 Biologically annotated neural networks also take in GWA summary statistics (BANN-SS) for multi-scale genotype-phenotype by specifying a partially connected architecture based on the hierarchical nature of enrichment studies. 184

B.2 Receiver operating characteristic (ROC) curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations cohort. 186
B.3 Receiver operating characteristic (ROC) curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations with population structure (European cohort).

B.4 Receiver operating characteristic (ROC) curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations with population structure (European cohort).

B.5 Receiver operating characteristic (ROC) curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations (British cohort).

B.6 Receiver operating characteristic (ROC) curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations (British cohort).

B.7 Receiver operating characteristic (ROC) curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations with population structure (European cohort).

B.8 Receiver operating characteristic (ROC) curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations with population structure (European cohort).

B.9 Precision-recall curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations (British cohort).

B.10 Precision-recall curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations (British cohort).

B.11 Precision-recall curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations with population structure (European cohort).
B.12 Precision-recall curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations with population structure (European cohort). ........................................ 196
B.13 Precision-recall curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations (British cohort). ......................................................... 197
B.14 Precision-recall curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations (British cohort). ......................................................... 198
B.15 Precision-recall curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations with population structure (European cohort). ......................... 199
B.16 Precision-recall curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations with population structure (European cohort). ......................... 200
B.17 Scatter plots comparing how the integrative neural network training procedure enables the ability to identify associated SNPs and enriched SNP-sets in simulations (British cohort). ......................................................... 201
B.18 Scatter plots comparing how the integrative neural network training procedure enables the ability to identify associated SNPs and enriched SNP-sets in simulations with population structure (European cohort). ......................... 202
B.19 Scatter plots comparing how the integrative neural network training procedure enables the ability to identify associated SNPs and enriched SNP-sets in simulations with population structure (European cohort). ......................... 203
B.20 Scatter plots comparing how the integrative neural network training procedure enables the ability to identify associated SNPs and enriched SNP-sets in simulations (British cohort). ......................................................... 204
B.21 Scatter plots comparing how the integrative neural network training procedure enables the ability to identify associated SNPs and enriched SNP-sets in simulations (British cohort). ................................................................. 205
B.22 Scatter plots comparing how the integrative neural network training procedure enables the ability to identify associated SNPs and enriched SNP-sets in simulations with population structure (European cohort). ......................... 206
B.23 Scatter plots comparing how the integrative neural network training procedure enables the ability to identify associated SNPs and enriched SNP-sets in simulations with population structure (European cohort). ......................... 207
B.24 Assessments of posterior inclusion probability (PIP) calibration for both SNP-level associations and enrichment of SNP-sets. ................................................................. 208
B.25 Receiver operating characteristic (ROC) curves comparing the performance of the BANNs models with different modifications via an ablation test. ......................... 209
B.26 Boxplots depicting the ability of the BANNs and BANN-SS models to estimate the phenotypic variation explained (PVE) by SNPs (pink) and SNP-sets (blue) for complex traits in simulations. ................................................................. 210
B.27 Boxplots depicting the ability of the BANNs and BANN-SS models to estimate the phenotypic variation explained (PVE) by SNPs (pink) and SNP-sets (blue) for complex traits in simulations. ................................................................. 211
B.28 Manhattan plots of variant-level fine mapping results for six traits in heterogenous stock of mice from the Wellcome Trust Centre for Human Genetics. ......................... 212
B.29 Gene set enrichment analyses using the significant SNP-sets identified by BANNs for high-density and low-density lipoprotein (HDL and LDL, respectively) traits in the Framingham Heart Study. ................................................................. 213
B.30 Gene set enrichment analyses using the significant SNP-sets identified by BANNs for high-density and low-density lipoprotein (HDL and LDL, respectively) traits in the Framingham Heart Study. ................................................................. 214
B.31 Manhattan plot of variant-level association mapping results for high-density and low-density lipoprotein (HDL and LDL, respectively) traits in ten thousand randomly sampled individuals of European ancestry from the UK Biobank. 215

B.32 Gene set enrichment analyses using the significant SNP-sets identified by BANNs for high-density and low-density lipoprotein (HDL and LDL, respectively) traits in ten thousand randomly sampled individuals of European ancestry from the UK Biobank. 216

B.33 Gene set enrichment analyses using the significant SNP-sets identified by BANNs for high-density and low-density lipoprotein (HDL and LDL, respectively) traits in ten thousand randomly sampled individuals of European ancestry from the UK Biobank. 217

C.1 An example of a single-effect neural network (SNN) with only the first input variable having an effect on the outcome. 239

C.2 Toy example demonstrating importance of accounting for nonlinearity when performing variable selection. 240

C.3 An example of absolute correlation matrix for genotype data. 241

C.4 Assessments of posterior inclusion probability (PIP) calibration for ESNN and SuSiE. 242

C.5 Precision and recall curves for simulation studies of different scenarios. 243

C.6 Posterior inclusion probabilities (PIP) of ESNN and SuSiE for the heterogenous stock of mice dataset from the Wellcome Trust Centre for Human Genetics. 244
Introduction

Learning the genetic basis of traits helps us better understand the underlying etiology and provide guidance for the treatment of disease. Genome-wide association studies (GWAS) aim to solve this problem by performing hypothesis testing using marginal linear regressions. Though GWAS have identified thousands of genetic loci of complex traits and common disease, there remain a lot of limitations in the current GWAS framework due to the complex nature of the genetic data (254; 166; 255; 18; 280; 166; 288; 255). For example, traditional GWAS tests the association between the trait with each Single Nucleotide Polymorphism (SNP) independently. However, genetic data often has strong correlation structure among SNPs which is also known as linkage disequilibrium (LD). Therefore, when non-causal SNPs are correlated with true causal SNPs, the strong LD can cause a large amount of false positive signals under the GWAS framework (305; 289; 21; 280). Besides, this univariate SNP-level test approach can be underpowered for “polygenic” traits which are generated by many mutations of small effect (166; 255; 305; 289; 21; 280). Many recent work has extended variable selection techniques to identify enriched gene or pathway-level associations, where one would combine groups of SNPs within a particular genomic region (commonly known as a SNP-set) to detect aggregated signals instead (154; 281; 28; 56; 132; 183; 307; 237). We will also show that these methods can also generate a lot of false positives as most of them do not take LD into account.

Meanwhile, most of the existing GWAS methods and SNP-set methods are built based on linear models and ignore the non-linearity such as dominance effects and interactions among SNPs which will cause the method to be underpowered (269; 9; 113; 51; 174; 50; 200; 146). On the other hand, nonlinear methods such as neural networks are well-known to have great power to model complex
nonlinearity (135). However, the lack of interpretation prevents them from being used as formal hypothesis testing methods in genetics (51; 52).

The main contribution of this thesis is to overcome these limitations by developing several multiscale statistical inference methods including SNP level association, SNP-set association, linear and nonlinear methods. This thesis highlighted three methods: (i) a SNP-set association method — gene-ε — that can reduce false positive signals by using a reformulated null model and incorporating the LD into the model along with the shrinkage with summary statistics. (ii). Biologically Annotated Neural Networks (BANNs), a nonlinear probabilistic framework that can perform association tests on SNP level and SNP-set simultaneously with a shrinkage prior. (iii). the Ensemble of Single-effect Neural Network (ESNN), a Bayesian neural network that can be used to quantify uncertainty for variable selection when the variables are highly correlated which is one of the key features in the genetic fine-mapping. These three methods are presented separately in three chapters.

In Chapter I, we introduce — gene-ε — a new approach for identifying statistical associations between sets of variants and quantitative traits. The key insight is using a reformulate the genome-wide SNP-level null hypothesis to identify spurious small-to-intermediate SNP effects and classify them as non-causal. gene-ε first uses regularized regression (e.g. LASSO, Elastic Net, or Ridge Regression) with summary statistics (GWA effect size estimators) and LD matrix to derive shrunk effect sizes. Next, it takes shrunk effect sizes as input and uses an EM algorithm to estimate a null threshold for SNPs. This null threshold is used to model the variance of the effect sizes for the SNPs that have strong correlations with the truly associated SNPs. Finally, gene-ε computes the gene level test statistics in a quadratic form using shrunk effect sizes and derives p-value by constructing the null distribution with the null threshold. We show that gene-ε can outperform competing methods in various simulation scenarios. We apply gene-ε to UKBiobank data and identify novel enriched genes. This project is joint work with Sohini Ramachandran and Lorin Crawford and is published in Plos Genetics (40).

As an extension project, we also explore the utility of gene-varepsilon on binary traits (case-control studies). In case-control GWAS, the actual prevalence of the disease is often small which means the number of cases is far less than the controls. But cases are normally oversampled for
study purposes. Such ascertainment bias has been shown to confound heritability estimation and GWAS (138; 86; 69). In our preliminary results, we show that gene-$\varepsilon$ can provide robust associations mapping under different disease prevalence especially when the traits are sparse and heritable. This project is joint work with Lorin Crawford and Sohini Ramachandran.

In Chapter II, we present Biologically Annotated Neural Networks (BANNs), a nonlinear probabilistic framework for association mapping on SNP and SNP-set levels. BANNs are feedforward models with partially connected architectures that are built using biological annotations. BANNs use coordinate ascent variational inference to derive posterior summary such as posterior inclusion probabilities that can be used for association mapping. Simulation studies show that BANNs can achieve state-of-art performance for association mapping on both SNP and SNP-set levels. We apply BANNs to three datasets: stock mice data from Wellcome Trust Centre, Framingham Heart Study, and UKBiobank data. We show that BANNs can identify novel associations and also replicate known associations. This chapter corresponds to joint work with Pinar Demetci, Gregory Darnell, Xiang Zhou, Sohini Ramachandran and Lorin Crawford. And the work is published in Plos Genetics (40).

In Chapter III, we propose a novel nonlinear association method called the Ensemble of Single-effect Neural Network (ESNN). ESNN is a Bayesian neural network that and it provides posterior summary such as credible sets which can be used to quantify uncertainty for variable selection (association) when the input variables (e.g. SNPs) are highly correlated. We implement ESNN using black-box variational inference and several approximation techniques. ESNN can be applied to both continuous and binary traits and used with any neural network architecture. We use extensive simulation to demonstrate that the credible sets and posterior inclusion probabilities generated by ESNN have higher coverage and better performance for association compared with competing methods. We further apply ESNN to sock mice data and Wellcome Trust Case Control Consortium data to show that ESNN can identify meaningful new association signals. This chapter is based on joint work with Sohini Ramachandran and Lorin Crawford.
Chapter 1

Estimation of Non-null SNP Effect Size Distributions Enables the Detection of Enriched Genes Underlying Complex Traits

1.1 Abstract

Traditional univariate genome-wide association studies generate false positives and negatives due to difficulties distinguishing associated variants from variants with spurious nonzero effects that do not directly influence the trait. Recent efforts have been directed at identifying genes or signaling pathways enriched for mutations in quantitative traits or case-control studies, but these can be computationally costly and hampered by strict model assumptions. Here, we present gene-ε, a new approach for identifying statistical associations between sets of variants and quantitative traits. Our key insight is that enrichment studies on the gene-level are improved when we reformulate the
genome-wide SNP-level null hypothesis to identify spurious small-to-intermediate SNP effects and classify them as non-causal. gene-ε efficiently identifies enriched genes under a variety of simulated genetic architectures, achieving greater than a 90% true positive rate at 1% false positive rate for polygenic traits. Lastly, we apply gene-ε to summary statistics derived from six quantitative traits using European-ancestry individuals in the UK Biobank, and identify enriched genes that are in biologically relevant pathways.

1.2 Introduction

Over the last decade, there has been an evolving debate about the types of insight genome-wide single-nucleotide polymorphism (SNP) genotype data offer into the genetic architecture of complex traits (254; 166; 255; 18; 280). In the traditional genome-wide association (GWA) framework, individual SNPs are tested independently for association with a trait of interest. While this approach can have drawbacks (166; 288; 255), more recent approaches that combine SNPs within a region have gained power to detect biologically relevant genes and pathways enriched for correlations with complex traits (154; 28; 109; 56; 132; 183; 266; 307). Reconciling these two observations is crucial for biomedical genomics.

In the traditional GWA model, each SNP is assumed to either (i) directly influence (or perfectly tag a variant that directly influences) the trait of interest; or (ii) have no affect on the trait at all (see Fig. 1.2A). Throughout this manuscript, for simplicity, we refer to SNPs under the former as “associated” and those under latter as “non-associated”. These classifications are based on ordinary least squares (OLS) effect size estimates for each SNP in a regression framework, where the null hypothesis assumes that the true effects of non-associated SNPs are zero ($H_0: \beta_j = 0$). The traditional GWA model is agnostic to trait architecture, and is underpowered with a high false-positive rate for “polygenic” traits or traits which are generated by many mutations of small effect (305; 289; 21; 280).
Figure 1.1. Illustration of null hypothesis assumptions for the distribution of GWA SNP-level effect sizes according to different views on underlying genetic architectures. The effect sizes of “non-associated” (pink), “spurious non-associated” (red), and “associated” (blue) SNPs were drawn from normal distributions with successively larger variances. (A) The traditional GWA model of complex traits simply assumes SNPs are associated or non-associated. Under the corresponding null hypothesis, associated SNPs are likely to emit nonzero effect sizes while non-associated SNPs will have effect sizes of zero. When there are many causal variants, we refer to the traits as polygenic. (B) Under our reformulated GWA model, there are three categories: associated SNPs, non-associated SNPs that emit spurious nonzero effect sizes, and non-associated SNPs with effect sizes of zero. We propose a multi-component framework (see also (299)), in which null SNPs can emit different levels of statistical signals based on (i) different degrees of connectedness (e.g., through linkage disequilibrium), or (ii) its regulated gene interacts with an enriched gene. While truly associated SNPs are still more likely to emit large effect sizes than SNPs in the other categories, null SNPs can have intermediate effect sizes. Here, our goal is to treat spurious SNPs with small-to-intermediate nonzero effects as being non-associated with the trait of interest.

Suppose that in truth each SNP in a GWA dataset instead belongs to one of three categories depending on the underlying distribution of their effects on the trait of interest: (i) associated SNPs; (ii) non-associated SNPs that emit spurious nonzero statistical signals; and (iii) non-associated SNPs with zero-effects (Fig. 1.2B) (299). Associated SNPs may lie in enriched genes that directly influence the trait of interest. The phenomenon of a non-associated SNP emitting nonzero statistical signal can occur due to multiple reasons. For example, spurious nonzero SNP effects can be due to some varying degree of linkage disequilibrium (LD) with associated SNPs (102); or alternatively, non-associated SNPs can have a trans-interaction effect with SNPs located within an enriched gene. In either setting, spurious SNPs can emit small-to-intermediate statistical noise (in some cases, even appearing indistinguishable from truly associated SNPs), thereby confounding traditional GWA tests (Fig. 1.2B). Hereafter, we refer to this noise as “epsilon-genic effects” (denoted in shorthand
as “ε-genic effects”). There is a need for a computational framework that has the ability to identify mutations associated with a wide range of traits, regardless of whether narrow-sense heritability is sparsely or uniformly distributed across the genome.

Here, we develop a new and scalable quantitative approach for testing aggregated sets of SNP-level GWA summary statistics for enrichment of associated mutations in a given quantitative trait. In practice, our approach can be applied to any user-specified set of genomic regions, such as regulatory elements, intergenic regions, or gene sets. In this study, for simplicity, we refer to our method as a gene-level test (i.e., an annotated collection of SNPs within the boundary of a gene). The key contribution of our approach is that gene-level association tests should treat spurious SNPs with ε-genic effects as non-associated variants. Conceptually, this requires assessing whether SNPs explain more than some “epsilon” proportion of the phenotypic variance. In this generalized model, we reformulate the GWA null hypothesis to assume approximately no association for spurious non-associated SNPs where

$$H_0: \beta_j \approx 0, \quad \beta_j \sim \mathcal{N}(0, \sigma^2_{\varepsilon}), \quad j = 1, \ldots, J \text{ SNPs.}$$

Here, $\sigma^2_{\varepsilon}$ denotes a “SNP-level null threshold” and represents the maximum proportion of phenotypic variance explained (PVE) that is contributed by spurious non-associated SNPs. This null hypothesis can be equivalently restated as $H_0: \mathbb{E}[\beta_j^2] \leq \sigma^2_{\varepsilon}$ (Fig. 1.2B). Non-enriched genes are then defined as genes that only contain SNPs with ε-genic effects (i.e., $0 \leq \mathbb{E}[\beta_j^2] \leq \sigma^2_{\varepsilon}$ for every $j$-th SNP within that region). Enriched genes, on the other hand, are genes that contain at least one associated SNP (i.e., $\mathbb{E}[\beta_j^2] > \sigma^2_{\varepsilon}$ for at least one SNP $j$ within that region). By accounting for the presence of spurious ε-genic effects (i.e., through different values of $\sigma^2_{\varepsilon}$ which the user can subjectively control), our approach flexibly constructs an appropriate GWA SNP-level null hypothesis for a wide range of traits with genetic architectures that land anywhere on the polygenic spectrum (see Materials and Methods).

We refer to our gene-level association framework as “gene-ε” (pronounced “genie”). gene-ε leverages our modified SNP-level null hypothesis to lower false positive rates and increases power for
identifying gene-level enrichment within GWA studies. This happens via two key conceptual insights. First, gene-$\varepsilon$ regularizes observed (and inflated) GWA summary statistics so that SNP-level effect size estimates are positively correlated with the assumed generative model of complex traits. Second, it examines the distribution of regularized effect sizes to offer the user choices for an appropriate SNP-level null threshold $\sigma_{\varepsilon}^2$ to distinguish associated SNPs from spurious non-associated SNPs. This makes for an improved and refined hypothesis testing strategy for identifying enriched genes underlying complex traits. With detailed simulations, we assess the power of gene-$\varepsilon$ to identify significant genes under a variety of genetic architectures, and compare its performance against multiple competing approaches (154; 281; 56; 183; 307). We also apply gene-$\varepsilon$ to the SNP-level summary statistics of six quantitative traits assayed in individuals of European ancestry from the UK Biobank (24).

1.3 Results

1.3.1 Overview of gene-$\varepsilon$

The gene-$\varepsilon$ framework requires two inputs: GWA SNP-level effect size estimates, and an empirical linkage disequilibrium (LD, or variance-covariance) matrix. The LD matrix can be estimated directly from genotype data, or from an ancestry-matched set of samples if genotype data are not available to the user. We use these inputs to both estimate gene-level contributions to narrow-sense heritability $h^2$, and perform gene-level enrichment tests. After preparing the input data, there are three steps implemented in gene-$\varepsilon$, which are detailed below (Fig. 1.2).

First, we shrink the observed GWA effect size estimates via regularized regression (Figs. 1.2A and B; Eq. (1.4) in Materials and Methods). This shrinkage step reduces the inflation of OLS effect sizes for spurious SNPs (233), and increases their correlation with the assumed generative model for the trait of interest (particularly for traits with high heritability; Fig. S1). When assessing the performance of gene-$\varepsilon$ in simulations, we considered different types of regularization for the effect size estimates: the Least Absolute Shrinkage And Selection Operator (gene-$\varepsilon$-LASSO) (243), the Elastic Net solution (gene-$\varepsilon$-EN) (308), and Ridge Regression (gene-$\varepsilon$-RR) (99). We also assessed
our framework using the observed ordinary least squares (OLS) estimates without any shrinkage (gene-ε-OLS) to serve as motivation for having regularization as a step in the framework.

Second, we fit a $K$-mixture Gaussian model to all regularized effect sizes genome-wide with the goal of classifying SNPs as associated, non-associated with spurious statistical signal, or non-associated with zero-effects (Figs. 1.2B and 1.2C; see also (299)). Each successive Gaussian mixture component has distinctly smaller variances ($\sigma_1^2 > \cdots > \sigma_K^2$) with the $K$-th component fixed at $\sigma_K^2 = 0$. Estimating these variance components helps determine an appropriate $k$-th category to serve as the cutoff for SNPs with null effects (i.e., choosing some variance component $\sigma_k^2$ to be the null threshold $\sigma^2_2$). The gene-ε software allows users to determine this cutoff subjectively. Intuitively, enriched genes are likely to contain important variants with relatively larger effects that are categorized in the early-to-middle mixture components. Since the biological interpretation of the middle components may not be consistent across trait architectures, we take a conservative approach in our selection of a cutoff when determining associated SNPs. Without loss of generality, we assume non-null SNPs appear in the first mixture component with the largest variance, while null SNPs appear in the latter components. By this definition, non-associated SNPs with spurious $\varepsilon$-genic or zero-effects then have PVEs that fall at or below the variance of the second component (i.e., $\sigma_2^2 = \sigma_2^2$ and $H_0: E[|\beta^2_j|] \leq \sigma_2^2$ for the $j$-th SNP). gene-ε allows for flexibility in the number of Gaussians that specify the range of null and non-null SNP effects. To achieve genome-wide scalability, we estimate parameters of the $K$-mixture model using an expectation-maximization (EM) algorithm.

Third, we group the regularized GWA summary statistics according to gene boundaries (or user-specified SNP-sets) and compute a gene-level enrichment statistic based on a commonly used quadratic form (Fig. 1.2D) (154; 281; 183). In expectation, these test statistics can be naturally interpreted as the contribution of each gene to the narrow-sense heritability. We use Imhof’s method (108) to derive a $P$-value for assessing evidence in support of an association between a given gene and the trait of interest. Details for each of these steps can be found in Materials and Methods, as well as in Supporting Information.
1.3.2 Performance Comparisons in Simulation Studies

To assess the performance of gene-ε, we simulated complex traits under multiple genetic architectures using real genotype data on chromosome 1 from individuals of European ancestry in the UK Biobank (Materials and Methods). Following quality control procedures, our simulations included 36,518 SNPs (Supporting Information). Next, we used the NCBI's Reference Sequence (RefSeq) database in the UCSC Genome Browser (195) to annotate SNPs with the appropriate genes. Simulations were conducted using two different SNP-to-gene assignments. In the first, we directly used the UCSC annotations which resulted in 1,408 genes to be used in the simulation study. In the second, we augmented the UCSC gene boundaries to include SNPs within ±50kb, which resulted in 1,916 genes in the simulation study. For both cases, we assumed a linear additive model for quantitative traits, while varying the following parameters: sample size (N = 5,000 or 10,000); narrow-sense heritability (h² = 0.2 or 0.6); and the percentage of enriched genes (set to 1% or 10%). In each scenario, we considered traits being generated with and without additional population structure. In the latter setting, traits are simulated while also using the top ten principal components of the genotype matrix as covariates to create stratification. Regardless of the setting, GWA summary statistics were computed by fitting a single-SNP univariate linear model (via OLS) without any control for population structure. Comparisons were based on 100 different simulated runs for each parameter combination.

We compared the performance of gene-ε against that of five competing gene-level association or enrichment methods: SKAT (281), VEGAS (154), MAGMA (56), PEGASUS (183), and RSS (307) (Supporting Information). As previously noted, we also explored the performance of gene-ε while using various degrees of regularization on effect size estimates, with gene-ε-OLS being treated as a baseline. SKAT, VEGAS, and PEGASUS are frequentist approaches, in which SNP-level GWA P-values are drawn from a correlated chi-squared distribution with covariance estimated using an empirical LD matrix (139). 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 (56). RSS is a Bayesian model-based 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 (306). Conceptually, SKAT, MAGMA, VEGAS, and PEGASUS assume null models under the traditional GWA framework, while RSS and gene-ε allow for traits to have architectures with more complex SNP effect size distributions.

For all methods, we assess the power and false discovery rates (FDR) for identifying correct genes at a Bonferroni-corrected threshold \( P = 0.05/1408 \) genes = 3.55×10⁻⁵ and \( P = 0.05/1916 \) genes = 2.61×10⁻⁵, depending on if the ±50kb buffer was used) or median probability model (posterior enrichment probability > 0.5; see (6)) (Tables S1-S16). We also compare their ability to rank true positives over false positives via receiver operating characteristic (ROC) and precision-recall curves (Figs. 1.3 and A.2 - A.16). While we find gene-ε and RSS have the best tradeoff between true and false positive rates, RSS does not scale well for genome-wide analyses (Table 1.1). In many settings, gene-ε has similar power to RSS (while maintaining a considerably lower FDR), and generally outperforms RSS in precision-versus-recall. gene-ε also stands out as the best approach in scenarios where the observed OLS summary statistics were produced without first controlling for confounding stratification effects in more heritable traits (i.e., \( h^2 = 0.6 \)). Computationally, gene-ε gains speed by directly assessing evidence for rejecting the gene-level null hypothesis, whereas RSS must compute the posterior probability of being an enriched gene (which can suffer from convergence issues; Supporting Information). For context, an analysis of just 1,000 genes takes gene-ε an average of 140 seconds to run on a personal laptop, while RSS takes around 9,400 seconds to complete.

When using GWA summary statistics to identify genotype-phenotype associations, modeling the appropriate trait architecture is crucial. As expected, all methods we compared in this study have relatively more power for traits with high \( h^2 \). However, our simulation studies confirm the expectation that the max utility for methods assuming the traditional GWA framework (i.e., SKAT, MAGMA, VEGAS, and PEGASUS) is limited to scenarios where heritability is low, phenotypic variance is dominated by just a few enriched genes with large effects, and summary statistics are not confounded by population structure (Figs. A.2, A.3, A.9 and A.19). RSS, gene-ε-EN, and gene-ε-LASSO robustly outperform these methods for the other trait architectures (Figs. 1.3, A.4 - A.8, and A.11 - A.16). One major reason for this result is that shrinkage and penalized regression
Table 1.1. Computational time for running gene-ε and other gene-level association approaches, as a function of the total number genes analyzed and the number of SNPs within each gene. Methods compared include: gene-ε, PEGASUS (183), VEGAS (154), RSS (307), MAGMA (56), and SKAT(281). Here, we simulated 10 datasets for each pair of parameter values (number of genes analyzed, and number of SNPs within each gene). Each table entry represents the average computation time (in seconds) it takes each approach to analyze a dataset of the size indicated. Run times were measured on a MacBook Pro (Processor: 3.1-gigahertz (GHz) Intel Core i5, Memory: 8GB 2133-megahertz (MHz) LPDDR3). Only a single core on the machine was used. PEGASUS, SKAT, and MAGMA are score-based methods and, thus, are expected to take the least amount of time to run. Both gene-ε and RSS are regression-based methods, but gene-ε is scalable in both the number of genes and the number of SNPs per gene. The increased computational burden of RSS results from its need to do Bayesian posterior inference; however, gene-ε is able to scale because it leverages regularization and point estimation for hypothesis testing. Regularization also allows gene-ε to preserve type 1 error when traits are generated under the null hypothesis of no gene enrichment. Importantly, our method is relatively conservative when GWA summary statistics are less precise and derived from studies with smaller sample sizes (e.g., $N = 5,000$; Table A.17).
1.3.3 Characterizing Genetic Architecture of Quantitative Traits in the UK Biobank

We applied gene-ε to 1,070,306 genome-wide SNPs and six quantitative traits — height, body mass index (BMI), mean red blood cell volume (MCV), mean platelet volume (MPV), platelet count (PLC), waist-hip ratio (WHR) — assayed in 349,414 European-ancestry individuals in the UK Biobank (Supporting Information) (24). After quality control, we regressed the top ten principal components of the genotype data onto each trait to control for population structure, and then we derived OLS SNP-level effect sizes using the traditional GWA framework. For completeness, we then analyzed these GWA effect size estimates with the four different implementations of gene-ε. In the main text, we highlight results under the Elastic Net solution; detailed findings with the other gene-ε approaches can be found in Supporting Information.

While estimating ε-genic effects, gene-ε provides insight into the genetic architecture of a trait (Table A.18). For example, past studies have shown human height to have a higher narrow-sense heritability (estimates ranging from 45-80%; (288; 294; 279; 96; 218; 284; 78; 227; 167; 258)). Using Elastic Net regularized effect sizes, gene-ε estimated approximately 11% of SNPs in the UK Biobank to be statistically associated with height. This meant approximately 110,000 SNPs had marginal PVEs $\mathbb{E}[\beta_j^2] > 0$ (Materials and Methods). This number is similar to the 93,000 and 100,000 height associated variants previously estimated by Goldstein (87) and Boyle et al. (18), respectively. Additionally, gene-ε identified approximately 2% of SNPs to be “causal” (meaning they had PVEs greater than the SNP-level null threshold, $\mathbb{E}[\beta_j^2] > \sigma^2$); again similar to the Boyle et al. (18) estimate of 3.8% causal SNPs for height using data from the GIANT Consortium (279), and the Lello et al. (141) estimate of 3.1% causal SNPs for height using European-ancestry individuals in the UK Biobank.

Compared to body height, narrow-sense heritability estimates for BMI have been considered both high and low (estimates ranging from 25-60%; (253; 294; 290; 96; 218; 78; 204; 227; 205)). Such inconsistency is likely due to difference in study design (e.g., twin, family, population-based studies), many of which have been known to produce different levels of bias (204). Here, our
results suggest BMI to have a lower narrow-sense heritability than height, with a slightly different distribution of null and non-null SNP effects. Specifically, we found BMI to have 13% associated SNPs and 6% causal SNPs.

In general, we found our genetic architecture characterizations in the UK Biobank to reflect the same general themes we saw in the simulation study. Less aggressive shrinkage approaches (e.g., OLS and Ridge) are subject to misclassifications of associated, spurious, and non-associated SNPs. As a result, these methods struggle to reproduce well-known narrow-sense heritability estimates from the literature, across all six traits. This once again highlights the need for computational frameworks that are able to appropriately correct for inflation in summary statistics.

### 1.3.4 gene-ε Identifies Refined List of Genetic Enrichments

Next, we applied gene-ε to the summary statistics from the UK Biobank and generated genome-wide gene-level association \( P \)-values (panels A and B of Figs. 1.4 and A.25 - A.29). As in the simulation study, we conducted two separate analyses using two different SNP-to-gene annotations: (i) we used the RefSeq database gene boundary definitions directly, or (b) we augmented the gene boundaries by adding SNPs within a ±50 kilobase (kb) buffer to account for possible regulatory elements. A total of 14,322 genes were analyzed when using the UCSC boundaries as defined, and a total of 17,680 genes were analyzed when including the 50kb buffer. The ultimate objective of gene-ε is to identify enriched genes, which we define as containing at least one associated SNP and achieving a gene-level association \( P \)-value below a Bonferroni-corrected significance threshold (in our two analyses, \( P = 0.05/14322 \) genes = \( 3.49 \times 10^{-6} \) and \( P = 0.05/17680 \) genes = \( 2.83 \times 10^{-6} \), respectively; Tables S19-S24). As a validation step, we compared gene-ε \( P \)-values to RSS posterior enrichment probabilities for each gene. We also used the gene set enrichment analysis tool Enrichr (34) to identify dbGaP categories with an overrepresentation of significant genes reported by gene-ε (panels C and D of Figs. 1.4 and A.25 - A.29). A comparison of gene-level associations and gene set enrichments between the different gene-ε approaches are also listed (Tables S25-S27).

Many of the candidate enriched genes we identified by applying gene-ε were not previously annotated as having trait-specific associations in either dbGaP or the GWAS catalog (Fig. 1.4):
however, many of these same candidate genes have been identified by past publications as related to the phenotype of interest (Table 1.2). It is worth noting that multiple genes would not have been identified by standard GWA approaches since the top SNP in the annotated region had a marginal association below a genome-wide threshold (see Table 1.2 and highlighted rows in Tables A.19 - A.24). Additionally, 45% of the genes selected by gene-ε were also selected by RSS. For example, gene-ε reports C1orf150 as having a significant gene-level association with MPV \( (P = 1 \times 10^{-20}) \) and RSS posterior enrichment probability of 1), which is known to be associated with germinal center signaling and the differentiation of mature B cells that mutually activate platelets (203; 65; 110).

Importantly, nearly all of the genes reported by gene-ε had evidence of overrepresentation in gene set categories that were at least related to the trait of interest. As expected, the top categories with Enrichr \( Q \)-values smaller than 0.05 for height and MPV were "Body Height" and "Platelet Count", respectively. Even for the less heritable MCV, the top significant gene sets included hematological categories such as "Transferrin", "Erythrocyte Indices", "Hematocrit", "Narcolepsy", and "Iron" — all of which have verified and clinically relevant connections to trait (73; 3; 12; 33; 225; 75; 145; 4).
Figure 1.2. Schematic overview of gene-$\varepsilon$: our new gene-level association approach accounting for spurious nonzero SNP-level effects. **(A)** gene-$\varepsilon$ takes SNP-level GWA marginal effect sizes (OLS estimates $\hat{\beta}$) and a linkage disequilibrium (LD) matrix ($\Sigma$) as input. It is well-known that OLS effect size estimates are inflated due to LD (i.e., correlation structures) among genome-wide genotypes. **(B)** gene-$\varepsilon$ first uses its inputs to derive regularized effect size estimates ($\tilde{\beta}$) through shrinkage methods (LASSO, Elastic Net and Ridge Regression; we explore performance of each solution under a variety of simulated trait architectures in Supporting Information). **(C)** A unique feature of gene-$\varepsilon$ is that it treats SNPs with spurious nonzero effects as non-associated. gene-$\varepsilon$ assumes a reformulated null distribution of SNP-level effects $\tilde{\beta}_j \sim N(0, \sigma_j^2)$, where $\sigma_j^2$ is the SNP-level null threshold and represents the maximum proportion of phenotypic variance explained (PVE) by a spurious or non-associated SNP. This leads to the reformulated SNP-level null hypothesis $H_0$: $E[\tilde{\beta}_j^2] \leq \sigma_j^2$. To infer an appropriate $\sigma_j^2$, gene-$\varepsilon$ fits a $K$-mixture of normal distributions over the regularized effect sizes with successively smaller variances ($\sigma_1^2 > \cdots > \sigma_K^2$; with $\sigma_K^2 = 0$). In this study (without loss of generality), we assume that associated SNPs will appear in the first set, while spurious and non-associated SNPs appear in the latter sets. By definition, the SNP-level null threshold is then $\sigma_1^2 = \sigma^2_1$. **(D)** Lastly, gene-$\varepsilon$ computes gene-level association test statistics $\tilde{Q}_g$ using quadratic forms and corresponding $P$-values using Imhoh’s method. This assumes the common gene-level null $H_0$: $Q_g = 0$, where the null distribution of $Q_g$ is dependent upon the SNP-level null threshold $\sigma^2_1$. For more details, see Materials and Methods.
Figure 1.3. Receiver operating characteristic (ROC) and precision-recall curves comparing the performance of gene-$\varepsilon$ and competing approaches in simulations ($N = 10,000; h^2 = 0.6$). We simulate complex traits under different genetic architectures and GWA study scenarios, varying the following parameters: narrow sense heritability, proportion of associated genes, and sample size (Supporting Information). Here, the sample size $N = 10,000$ and the narrow-sense heritability $h^2 = 0.6$. We compute standard GWA SNP-level effect sizes (estimated using ordinary least squares). Results for gene-$\varepsilon$ are shown with LASSO (blue), Elastic Net (EN; red), and Ridge Regression (RR; purple) regularizations. We also show the results of gene-$\varepsilon$ without regularization to illustrate the importance of this step (labeled OLS; orange). We further compare gene-$\varepsilon$ with five existing methods: PEGASUS (brown) (183), VEGAS (teal) (154), the Bayesian approach RSS (black) (307), SKAT (green) (281), and MAGMA (peach) (56). (A, C) ROC curves show power versus false positive rate for each approach of sparse (1% associated genes) and polygenic (10% associated genes) architectures, respectively. Note that the upper limit of the x-axis has been truncated at 0.1. (B, D) Precision-Recall curves for each method applied to the simulations. Note that, in the sparse case (1% associated genes), the top ranked genes are always true positives, and therefore the minimal recall is not 0. All results are based on 100 replicates.
Figure 1.4. Gene-level association results from applying gene-ε to body height (panels A and C) and mean platelet volume (MPV; panels B and D), assayed in European-ancestry individuals in the UK Biobank. Body height has been estimated to have a narrow-sense heritability $h^2$ in the range of 0.45 to 0.80 (288; 294; 279; 96; 218; 284; 78; 227; 167; 258); while, MPV has been estimated to have $h^2$ between 0.50 and 0.70 (198; 96; 218). Manhattan plots of gene-ε gene-level association $P$-values using Elastic Net regularized effect sizes for (A) body height and (B) MPV. The purple dashed line indicates a log-transformed Bonferroni-corrected significance threshold ($P = 3.49 \times 10^{-6}$ correcting for 14,322 autosomal genes analyzed). We color code all significant genes identified by gene-ε in orange, and annotate genes overlapping with the database of Genotypes and Phenotypes (dbGaP). In (C) and (D), we conduct gene set enrichment analysis using Enrichr (34;129) to identify dbGaP categories enriched for significant gene-level associations reported by gene-ε. We highlight categories with $Q$-values (i.e., false discovery rates) less than 0.05 and annotate corresponding genes in the Manhattan plots in (A) and (B), respectively. For height, the only significant dbGaP category is “Body Height”, with nine of the genes identified by gene-ε appearing in this category. For MPV, the two significant dbGaP categories are “Platelet Count” and “Face” — the first of which is directly connected to trait (4; 142; 180).
Weaver Syndrome (characterized by rapid growth). BMI is reported to be associated with obesity. MCV which is known to be involved with germinal center signaling and differentiation of mature B cells that mutually activate platelets. THF4 which is involved with canonical Wnt signaling and can affect platelet formation. WDR6 which is also known as DCAF7 which has been shown to bind Huntingtin-associated protein 1 (HAP1) and affect weight. MLL which is also known as the 26S proteasome which is found to be important for platelet production.

### Table 1.2

<table>
<thead>
<tr>
<th>Trait</th>
<th>Gene</th>
<th>Chr</th>
<th>gene(\epsilon) Value</th>
<th>Rank</th>
<th>Post. Prob.</th>
<th>Biological Relevance to Trait</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>EZH2</td>
<td>7</td>
<td>(9.34 \times 10^{-8})</td>
<td>61</td>
<td>1.000</td>
<td>Associated with diseases Adamantinoma of Long Bone and Weaver Syndrome (characterized by rapid growth).</td>
</tr>
<tr>
<td>Height</td>
<td>C17orf42</td>
<td>17</td>
<td>(5.38 \times 10^{-9})</td>
<td>52</td>
<td>1.000</td>
<td>Known as the transcription elongation factor of mitochondria (TEFM) which regulates transcription and can affect body height.</td>
</tr>
<tr>
<td>Height</td>
<td>KISS1R</td>
<td>19</td>
<td>(1 \times 10^{-20})</td>
<td>1*</td>
<td>0.970</td>
<td>Associated with disorders of puberty and final height.</td>
</tr>
<tr>
<td>BMI</td>
<td>ZC3H4</td>
<td>19</td>
<td>(1.62 \times 10^{-14})</td>
<td>20</td>
<td>1.000</td>
<td>BMI-inducer known to be associated with adiposity and obesity.</td>
</tr>
<tr>
<td>BMI</td>
<td>PTOV1</td>
<td>19</td>
<td>(1 \times 10^{-20})</td>
<td>1*</td>
<td>0.990</td>
<td>Found to be overexpressed in prostate adenocarcinomas which can be induced by obesity.</td>
</tr>
<tr>
<td>BMI</td>
<td>FBXO45</td>
<td>3</td>
<td>(6.52 \times 10^{-7})</td>
<td>23</td>
<td>0.029</td>
<td>Reported to be involved in children syndromic obesity.</td>
</tr>
<tr>
<td>MCV</td>
<td>SLC24A1</td>
<td>15</td>
<td>(1.74 \times 10^{-4})</td>
<td>50</td>
<td>0.140</td>
<td>Encoded protein is involved in glucose transportation pathway and MCV is reported to be associated with glucose level.</td>
</tr>
<tr>
<td>MCV</td>
<td>PDX1</td>
<td>13</td>
<td>(1 \times 10^{-20})</td>
<td>1*</td>
<td>0.019</td>
<td>Associated with Glycated hemoglobin which is affected by MCV</td>
</tr>
<tr>
<td>MCV</td>
<td>RHOD</td>
<td>11</td>
<td>(1 \times 10^{-20})</td>
<td>1*</td>
<td>0.002</td>
<td>Associated with Wiskott-Aldrich Syndrome which is characterized by abnormal immune system function (immune deficiency) and a reduced ability to form blood clots.</td>
</tr>
<tr>
<td>MPV</td>
<td>C1orf150</td>
<td>1</td>
<td>(1 \times 10^{-20})</td>
<td>1*</td>
<td>1.000</td>
<td>Known as GCSAM which is involved with germinal center signaling and differentiation of mature B cells that mutually activate platelets.</td>
</tr>
<tr>
<td>MPV</td>
<td>KIAA0922</td>
<td>4</td>
<td>(3.20 \times 10^{-6})</td>
<td>64</td>
<td>1.000</td>
<td>Known as TEME131L which is associated with canonical Wnt signaling and can effect platelet formation.</td>
</tr>
<tr>
<td>MPV</td>
<td>TPT1</td>
<td>13</td>
<td>(1 \times 10^{-20})</td>
<td>1*</td>
<td>0.051</td>
<td>mRNA expression is identified in platelets.</td>
</tr>
<tr>
<td>PLC</td>
<td>C1orf150</td>
<td>1</td>
<td>(1 \times 10^{-20})</td>
<td>1*</td>
<td>1.000</td>
<td>Known as GCSAM which is involved with germinal center signaling and differentiation of mature B cells that mutually activate platelets.</td>
</tr>
<tr>
<td>PLC</td>
<td>PSMD2</td>
<td>3</td>
<td>(1.42 \times 10^{-3})</td>
<td>29</td>
<td>1.000</td>
<td>Also known as the 26S proteasome which is found to be important for platelet production.</td>
</tr>
<tr>
<td>PLC</td>
<td>APOB48R</td>
<td>16</td>
<td>(1 \times 10^{-20})</td>
<td>1*</td>
<td>0.003</td>
<td>Involved in Lipoprotein metabolism pathway which can affect platelet.</td>
</tr>
<tr>
<td>WHR</td>
<td>TFAP2B</td>
<td>6</td>
<td>(3.92 \times 10^{-7})</td>
<td>21</td>
<td>1.000</td>
<td>Dietary protein associated with weight maintenance.</td>
</tr>
<tr>
<td>WHR</td>
<td>WDR68</td>
<td>17</td>
<td>(1.05 \times 10^{-8})</td>
<td>20</td>
<td>0.990</td>
<td>Also known as APOB48R which has been shown to bind Huntingtin-associated protein 1 (HAP1) and affect weight.</td>
</tr>
<tr>
<td>WHR</td>
<td>MLL</td>
<td>11</td>
<td>(8.14 \times 10^{-8})</td>
<td>19</td>
<td>0.940</td>
<td>Orthologous gene in mice that affects skeleton, body size, and growth.</td>
</tr>
</tbody>
</table>

Table 1.2. Top three newly identified candidate genes reported by gene-\(\epsilon\) for the six quantitative traits studied in the UK Biobank (using imputed genotypes with gene boundaries defined by the NCBI’s RefSeq database in the UCSC Genome Browser (195)). We call these novel candidate genes because they are not listed as being associated with the trait of interest in either the GWAS catalog or dbGaP, and they have top posterior enrichment probabilities with the trait using RSS analysis. Each gene is annotated with past functional studies that link them to the trait of interest. We also report each gene’s overall trait-specific significance rank (out of 14,322 autosomal genes analyzed for each trait), as well as their heritability estimates from gene-\(\epsilon\) using Elastic Net to regularize GWA SNP-level effect size estimates. The traits are: height; BMI; mean corpuscular volume (MCV); mean platelet volume (MPV); platelet count (PLC); and waist-hip ratio (WHR). *: Enriched genes whose top SNP is not marginally significant according to a genome-wide Bonferroni-corrected threshold \(P = 4.67 \times 10^{-8}\) correcting for 1,070,306 SNPs analyzed; see highlighted rows in Supplementary Tables S19-S24 for complete list). *: Multiple genes were tied for this ranking. 

(195; 120)
Lastly, gene-ε also identified genes with rare causal variants. For example, ZNF628 (which is not mapped to height in the GWAS catalog) was detected by gene-ε with a significant $P$-value of $1 \times 10^{-20}$ (and $P = 4.58 \times 10^{-8}$ when the gene annotation included a 50kb buffer). Previous studies have shown a rare variant $\text{rs147110934}$ within this gene to significantly affect adult height (167). Rare and low-frequency variants are generally harder to detect under the traditional GWA framework. However, rare variants have been shown to be important for explaining the variation of complex traits (282; 139; 140; 309; 77). With regularization and testing for spurious ε-genic effects, gene-ε is able to distinguish between rare variants that are causal and SNPs with larger effect sizes due to various types of correlations. This only enhances the power of gene-ε to identify potential novel enriched genes.

1.4 Discussion

During the past decade, it has been repeatedly observed that the traditional GWA framework can struggle to accurately differentiate between associated and spurious SNPs (which we define as SNPs that covary with associated SNPs but do not directly influence the trait of interest). As a result, the traditional GWA approach is prone to generating false positives, and detects variant-level associations spread widely across the genome rather than aggregated sets in disease-relevant pathways (18). While this observation has spurred to many interesting lines of inquiry — such as investigating the role of rare variants in generating complex traits (282; 139; 109; 140), comparing the efficacy of tagging causal variants in different ancestries (277; 168), and integrating GWA data with functional -omics data (90; 283; 286) — the focus of GWA studies and studies integrating GWA data with other -omics data is still largely based on the role of individual variants, acting independently.

Here, our objective is to identify biologically significant underpinnings of the genetic architecture of complex traits by modifying the traditional GWA null hypothesis from $H_0: \beta_j = 0$ (i.e., the $j$-th SNP has zero statistical association with the trait of interest) to $H_0: \beta_j \approx 0$. We accomplish
this by testing for ε-genic effects: spurious small-to-intermediate effect sizes emitted by truly non-associated SNPs. We use an empirical Bayesian approach to learn the effect size distributions of null and non-null SNP effects, and then we aggregate (regularized) SNP-level association signals into a gene-level test statistic that represents the gene’s contribution to the narrow-sense heritability of the trait of interest. Together, these two steps reduce false positives and increase power to identify the mutations, genes, and pathways that directly influence a trait’s genetic architecture. By considering different thresholds for what constitutes a null SNP effect (i.e., different values of $\sigma^2_\varepsilon$ for spurious non-associated SNPs; Figs. 1.2 and 1.2), gene-ε offers the flexibility to construct an appropriate null hypothesis for a wide range of traits with genetic architectures that land anywhere on the polygenic spectrum. It is important to stress that while we repeatedly point to our improved ability distinguish “causal” variants in enriched genes, gene-ε is by no means a causal inference procedure. Instead, it is an association test which highlights genes in enriched pathways that are most likely to be associated with the trait of interest.

Through simulations, we showed the gene-ε framework outperforms other widely used gene-level association methods (particularly for highly heritable traits), while also maintaining scalability for genome-wide analyses (Figs. 1.3 and A.2 - A.24, and Tables 1 and A.1 - A.17). Indeed, all the approaches we compared in this study showed improved performance when they used summary statistics derived from studies with larger sample sizes (i.e., simulations with $N = 10,000$). This is because the quality of summary statistics also improves in these settings (via the asymptotic properties of OLS estimates). Nonetheless, our results suggest that applying gene-ε to summary statistics from previously published studies will increase the return made on investments in GWA studies over the last decade.

Like any aggregated SNP-set association method, gene-ε has its limitations. Perhaps the most obvious limitation is that annotations can bias the interpretation of results and lead to erroneous scientific conclusions (i.e., might cause us to highlight the “wrong” gene (222; 42; 307)). We observed some instances of this during the UK Biobank analyses. For example, when studying MPV, $CAPN10$ only appeared to be a significant gene after its UCSC annotated boundary was augmented by a ±50kb buffer window ($P = 1.85 \times 10^{-1}$ and $P = 1.17 \times 10^{-7}$ before and after
the buffer was added, respectively; Table A.22). After further investigation, this result occurred because the augmented definition of \emph{CAPN10} included nearly all causal SNPs from the significant neighboring gene \emph{RNPEPL1} ($P = 1 \times 10^{-20}$ and $P = 2.07 \times 10^{-9}$ before and after the buffer window was added, respectively). While this shows the need for careful biological interpretation of the results, it also highlights the power of gene-$\varepsilon$ to prioritize true genetic signal effectively.

Another limitation of gene-$\varepsilon$ is that it relies on the user to determine an appropriate SNP-level null threshold $\sigma^2_\varepsilon$ to serve as a cutoff between null and non-null SNP effects. In the current study, we use a $K$-mixture Gaussian model to classify SNPs into different categories and then (without loss of generality) we subjectively assume that associated SNPs only appear in the component with the largest variance (i.e., we choose $\sigma^2_\varepsilon = \sigma^2_{2}$). Indeed, there can be many scenarios where this particular threshold choice is not optimal. For example, if there is one very strongly associated locus, the current implementation of the algorithm will assign it to its own mixture component and all other SNPs will be assumed to be not associated with the trait, regardless of the size of their corresponding variances. As previously mentioned, one practical guideline would be to select $\sigma^2_\varepsilon$ based on some \emph{a priori} knowledge about a trait’s architecture. However, a more robust approach would be to select the SNP-null hypothesis threshold based on the data at hand. One way to do this would be to take a fully Bayesian approach and allow posterior inference on $\sigma^2_\varepsilon$ to be dependent upon how much heritability is explained by SNPs placed in the top few largest components of the normal mixture. Recently, sparse Bayesian parametric (155) and nonparametric (295) Gaussian mixture models have been proposed for improved polygenic prediction with summary statistics. Combining these modeling strategies with our modified SNP-level null hypothesis could make for a more unified and data-driven implementation of the gene-$\varepsilon$ framework.

There are several other potential extensions for the gene-$\varepsilon$ framework. First, in the current study, we only focused on applying gene-$\varepsilon$ to quantitative traits (Figs. 1.4 and A.25 - A.29, and Tables 2 and A.18 - A.27). Future studies extending this approach to binary traits (e.g., case-control studies) should explore controlling for additional confounders that can occur within these phenotypes, such as ascertainment (137; 85; 268). Second, we only focus on data consisting of common variants; however, it would be interesting to extend gene-$\varepsilon$ for (i) rare variant association
testing and (ii) studies that consider the combined effect between rare and common variants. A significant challenge, in either case, would be to adaptively adjust the strength of the regularization penalty on the observed OLS summary statistics for causal rare variants, so as to not misclassify them as spurious non-associated SNPs. Previous approaches with specific re-weighting functions for rare variants may help here (282; 139; 109) (Materials and Methods). A final related extension of gene-$\varepsilon$ is to include information about standard errors when estimating $\varepsilon$-genic effects. In our analyses using the UK Biobank, some of the newly identified candidate genes contained SNPs that had large effect sizes but insignificant $P$-values in the original GWA analysis (after Bonferroni-correction; Tables 1.2 and A.19 - A.24). While this could be attributed to the modified SNP-level null distribution assumed by gene-$\varepsilon$, it also motivates a regularization model that accounts for the standard error of effect size estimates from GWA studies (233; 306; 307).

1.5 Materials and Methods

1.5.1 Traditional Association Tests using Summary Statistics

gene-$\varepsilon$ requires two inputs: genome-wide association (GWA) marginal effect size estimates $\tilde{\beta}$, and an empirical linkage disequilibrium (LD) matrix $\Sigma$. We assumed the following generative linear model for complex traits

$$y = X\beta + e, \quad e \sim \mathcal{N}(0, \tau^2 I),$$

where $y$ denotes an $N$-dimensional vector of phenotypic states for a quantitative trait of interest measured in $N$ individuals; $X$ is an $N \times J$ matrix of genotypes, with $J$ denoting the number of single nucleotide polymorphisms (SNPs) encoded as $\{0, 1, 2\}$ copies of a reference allele at each locus; $\beta$ is a $J$-dimensional vector containing the additive effect sizes for an additional copy of the reference allele at each locus on $y$; $e$ is a normally distributed error term with mean zero and scaled variance $\tau^2$; and $I$ is an $N \times N$ identity matrix. For convenience, we assumed that the genotype matrix (column-wise) and trait of interest have been mean-centered and standardized. We also treat $\beta$ as a fixed effect. A central step in GWA studies is to infer $\beta$ for each SNP, given both genotypic
and phenotypic measurements for each individual sample. For every SNP $j$, gene-$\varepsilon$ takes in the ordinary least squares (OLS) estimates based on Eq. (1.1)

$$\hat{\beta}_j = (x_j^T x_j)^{-1} x_j^T y,$$

(1.2)

where $x_j$ is the $j$-th column of the genotype matrix $X$, and $\hat{\beta}_j$ is the $j$-th entry of the vector $\hat{\beta}$. In traditional GWA studies, the null hypothesis for statistical association tests assumes $H_0 : \beta_j = 0$ for all $j = 1, \ldots, J$ SNPs. It can be shown that two genotypic variants $x_j$ and $x_{j'}$ in linkage disequilibrium (LD) will produce effect size estimates $\hat{\beta}_j$ and $\hat{\beta}_{j'}$ ($j \neq j'$) that are correlated (306). This can lead to confounded statistical tests. For the applications considered here, the LD matrix 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).

### 1.5.2 Regularized Regression for GWA Summary Statistics

gene-$\varepsilon$ uses regularization on the observed GWA summary statistics to reduce inflation of SNP-level effect size estimates and increase their correlation with the assumed generative model of complex traits. For large sample size $N$, note that the asymptotic relationship between the observed GWA effect size estimates $\hat{\beta}$ and the true coefficient values $\beta$ is (103; 104; 299)

$$\mathbb{E}[\hat{\beta}_j] = \sum_{j'=1}^{J} \rho(x_j, x_{j'}) \beta_{j'} \iff \mathbb{E}[\hat{\beta}] = \Sigma \beta,$$

(1.3)

where $\Sigma_{j,j'} = \rho(x_j, x_{j'})$ denotes the correlation coefficient between SNPs $x_j$ and $x_{j'}$. The above mirrors a high-dimensional regression model with the misestimated OLS summary statistics as the response variables and the LD matrix as the design matrix. Theoretically, the resulting output coefficients from this model are the desired true effect size estimates. Due to the multi-collinear structure of GWA data, we cannot reuse the ordinary least squares solution reliably (278). Thus,
we derive the general regularization

\[ \tilde{\beta} = \arg \min_{\beta} \| \tilde{\beta} - \Sigma \beta \|_2^2, \quad \text{subject to } (1 - \alpha) \| \beta \|_1 + \alpha \| \beta \|_2^2 \leq t \text{ for some } t, \]  

(1.4)

where, in addition to previous notation, the solution \( \tilde{\beta} \) is used to denote the regularized solution of the observed GWA effect sizes \( \hat{\beta} \); and \( \| \cdot \|_1 \) and \( \| \cdot \|_2^2 \) denote \( L_1 \) and \( L_2 \) penalties, respectively. The free regularization parameter \( t \) is chosen based off a grid \([\log t_{\text{min}}, \log t_{\text{max}}]\) with 100 sequential steps of size 0.01. Here, \( t_{\text{max}} \) is the minimum value such that all summary statistics are shrunk to zero. We then select the \( t \) that results in a model with an \( R^2 \) within one standard error of the best fitted model. In other words, we choose the \( t \) that (i) results in a more sparse solution than the best fitted model, but (ii) cannot be distinguished from the best fitted model in terms of overall variance explained.

The term \( \alpha \) in Eq. (1.4) distinguishes the type of regularization used, and can be chosen to induce various degrees of shrinkage on the effect size estimates. Specifically, \( \alpha = 0 \) corresponds to the “Least Absolute Shrinkage and Selection Operator” or LASSO solution (243), \( \alpha = 1 \) equates to Ridge Regression (99), while \( 0 < \alpha < 1 \) results in the Elastic Net (308). The LASSO solution forces some inflated coefficients to be zero; while the Ridge shrinks the magnitudes of all coefficients but does not set any of them to be exactly zero. Intuitively, the LASSO will create a regularized set of effect sizes where associated SNPs have larger effects, non-associated SNPs with spurious small-to-intermediate (or \( \varepsilon \)-genic) effects, and non-associated SNPs with zero-effects. It has been suggested that the \( L_1 \)-penalty can suffer from a lack of stability (32). Therefore, in the main text, we also highlighted gene-\( \varepsilon \) using the Elastic Net (with \( \alpha = 0.5 \)). The Elastic Net is a convex combination of the LASSO and Ridge penalties, but still produces distinguishable sets of associated, spurious, and non-associated SNPs. Note that for large GWA studies (e.g., the UK Biobank analysis in the main text), it can be impractical to construct a genome-wide LD matrix; therefore, we regularize OLS effect size estimates based on partitioned chromosome specific LD matrices. Results comparing each of the gene-\( \varepsilon \) regularization implementations are given in the main text (Fig. 3) and Supporting Information (Figs. S2-S24 and Tables S1-18 and 25-27). We will describe how we approximate the
null distribution for these regularized GWA summary statistics over the next two sections.

### 1.5.3 Estimating the SNP-Level Null Threshold

The main innovation of gene-$\varepsilon$ is to treat spurious SNPs with $\varepsilon$-genic effects as non-associated. This leads to reformulating the GWA SNP-level null hypothesis to assume non-associated SNPs can make small-to-intermediate contributions to the phenotypic variance. Formally, we write this as

$$H_0: \beta_j \approx 0, \quad \beta_j \sim \mathcal{N}(0, \sigma^2_{\varepsilon}), \quad j = 1, \ldots, J$$

(1.5)

where $\sigma^2_{\varepsilon}$ denotes the “SNP-level null threshold” and represents the maximum proportion of phenotypic variance explained (PVE) that is contributed by spurious SNPs. Based on Eq. (1.5), we equivalently say

$$H_0: \mathbb{E}[eta_j^2] \leq \sigma^2_{\varepsilon}.$$  

(1.6)

To estimate the threshold $\sigma^2_{\varepsilon}$ for null SNP-level effects, we use an empirical Bayesian approach and fit a $K$-mixture of normal distributions over the (regularized) effect size estimates (299),

$$\tilde{\beta}_j \mid z_j = k \sim \mathcal{N}(0, \sigma^2_k), \quad \Pr[z_j = k] = \pi_k,$$  

(1.7)

where $z_j \in \{1, \ldots, K\}$ is a latent variable representing the categorical membership for the $j$-th SNP. When summing over all components, Eq. (1.7) corresponds to the following marginal distribution

$$\tilde{\beta}_j \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(0, \sigma^2_k),$$  

(1.8)

where $\pi_k$ is a mixture weight representing the marginal (unconditional) probability that a randomly selected SNP belongs to the $k$-th component, with $\sum_k \pi_k = 1$. The above mixture allows for distinct clusters of nonzero effects through $K$ different variance components ($\sigma^2_k, k = 1, \ldots, K$) (299). Here, we consider sequential fractions ($\pi_1, \ldots, \pi_K$) of SNPs to correspond to distinctly smaller effects ($\sigma^2_1 > \cdots > \sigma^2_K = 0$) (299). The goal of the mixture model is to “bin” each of the (regularized)
SNP-level effects and determine an appropriate category \( k \) to serve as the cutoff for SNPs with null effects (i.e., choosing the threshold \( \sigma^2_\varepsilon \) based on some \( \sigma^2_k \)). Such a threshold can be chosen based on \textit{a priori} knowledge about the phenotype of interest. It is intuitive to assume that enriched genes will contain non-null SNPs that classify within the early-to-middle mixture components; unfortunately, the biological interpretations of the middle components may not be consistent across trait architectures. Therefore, without loss of generality in this paper, we take a conservative approach in our definition of associated SNPs within enriched genes. Here, we subjectively set the SNP-level null threshold as \( \sigma^2_\varepsilon = \sigma^2_2 \). Thus, non-null SNPs are assumed to appear in the largest fraction (i.e., the alternative \( H_A : \mathbb{E}[\beta^2_j] > \sigma^2_2 \)), while null SNPs with belong to the latter groups (i.e., the null \( H_0 : \mathbb{E}[\beta^2_j] \leq \sigma^2_2 \)). Given Eqs. (1.7) and (1.8), we write the joint log-likelihood for all \( J \) SNPs as the following

\[
\log p(\vec{\beta} | \Theta) = \sum_{j=1}^{J} \log p(\tilde{\beta}_j | \Theta) = \sum_{j=1}^{J} \log \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(0, \sigma^2_k) \right\},
\]

where \( \Theta = (\pi_1, \ldots, \pi_K, \sigma^2_1, \ldots, \sigma^2_K) \) is the complete set of parameters for the mixture model. Since there is not a closed-form solution for the maximum likelihood estimate (MLE), so we use an expectation-maximization (EM) algorithm to estimate the parameters in \( \Theta \) (59; 10; 171).

**Derivation of the EM Algorithm.** To derive an EM solution, we use Eqs. (1.7) and (1.8) to write the joint distribution of the \( J \)-regularized SNP-level effect sizes and the \( J \)-latent random variables \( z = (z_1, \ldots, z_J) \), conditioned on the mixture parameters \( \Theta \),

\[
p(\vec{\beta}, z | \Theta) = p(\vec{\beta} | z, \Theta)p(z) = \prod_{j=1}^{J} \prod_{k=1}^{K} [\pi_k \mathcal{N}(0, \sigma^2_k)]^{I(z_j = k)},
\]

where \( I(z_j = k) \) is an indicator function and equates to one if \( z_j = k \) and zero otherwise. Taking the log of this distribution yields the following

\[
\log p(\vec{\beta}, z | \Theta) = \sum_{j=1}^{J} \log p(\tilde{\beta}_j, z_j | \Theta) = \sum_{j=1}^{J} \sum_{k=1}^{K} I(z_j = k) \left[ \log \pi_k + \log \mathcal{N}(0, \sigma^2_k) \right].
\]
As opposed to Eq. (1.9), the augmented log-likelihood in Eq. (1.11) is a much simpler function for which to find a solution. The formal steps of the EM algorithm are now detailed below:

1. **E-Step: Update the Probability of Fraction Assignment.** In the E-step of the EM algorithm, we estimate the probability that the $j$-th SNP belongs to one of the $K$ fraction groups. To begin, we use Bayes theorem to find

$$p(z | \tilde{\beta}, \Theta) \propto p(\tilde{\beta} | z, \Theta)p(z) = \prod_{j=1}^{J} \prod_{k=1}^{K} [\pi_k \mathcal{N}(0, \sigma_k^2)]^{\delta_{z_j=k}}. \quad (1.12)$$

Next, we take the expectation of the complete log-likelihood log $p(\tilde{\beta}, z | \Theta)$, with respect to the conditional distribution $p(z | \tilde{\beta}, \Theta)$, under current value of the mixture parameters $\tilde{\Theta}$. This yields

$$\mathbb{E}_{z | \tilde{\beta}, \tilde{\Theta}}[\log p(\tilde{\beta}, z | \Theta)] = \sum_{j=1}^{J} \sum_{k=1}^{K} \hat{\gamma}_k^{(j)} \left[ \log \pi_k + \log \mathcal{N}(0, \sigma_k^2) \right], \quad (1.13)$$

where $\hat{\gamma}_k^{(j)}$ is referred to as the “responsibility of the $k$-th mixture component”, and is given as

$$\hat{\gamma}_k^{(j)} = \Pr[z_j = k | \tilde{\beta}_j, \tilde{\Theta}] = \frac{\hat{\pi}_k \mathcal{N}(0, \hat{\sigma}_k^2)}{\sum_{k'=1}^{K} \hat{\pi}_{k'} \mathcal{N}(0, \hat{\sigma}_{k'}^2)}. \quad (1.14)$$

Intuitively, the EM algorithm uses the collection of these responsibility values to assign SNPs to one of the $K$ fraction groups. This key step may be interpreted as determining the category of SNP effects (which is determined by identifying the $k$-th component with the largest $\hat{\gamma}_k^{(j)}$ for each $j$-th SNP).

2. **M-Step: Update the Component Variances and Mixture Weights.** In the M-step of the EM algorithm, we now fix the responsibility values and maximize the expectation in Eq. (1.13), with respect to the parameters in $\tilde{\Theta}$. Namely, we compute the following closed-form solutions:

$$\hat{\sigma}_k^2 = \frac{1}{J_k} \sum_{j=1}^{J} \hat{\gamma}_k^{(j)} \tilde{\beta}_j^2, \quad \hat{\pi}_k = \frac{J_k}{J} \quad (1.15)$$

where $J_k = \sum_j \hat{\gamma}_k^{(j)}$ is the sum of the membership weights for the $k$-th mixture component.
and represents the number of SNPs assigned to that component. The $\hat{\sigma}_k^2$ estimates are used to set the SNP-level null threshold $\hat{\sigma}_k^2$.

The gene-$\varepsilon$ software implements the above EM algorithm using the mclust (215) package in R. Results in the main text and Supporting Information are based on 100 iterations from 10 different parallel chains to ensure convergence. To implement the above algorithm, we use the mclust software package which can fit a Gaussian mixture with up to $K = 10$ distinct components (see Software Details). Here, the function will compare the Bayesian Information Criterion (BIC) approximation to the Bayes factor for each possible $K$ (214), and produces a resulting output for the $K$ value that has the largest BIC value. Note that since the EM updates do not involve any large LD matrices, the algorithm scales to be fit efficiently over all SNPs genome-wide.

1.5.4 Regularized GWA Summary Statistics under the Null Hypothesis

With an estimate of the SNP-level null threshold $\sigma_\varepsilon^2$, we now describe the probabilistic distribution of the regularized GWA summary statistics under the null hypothesis. Without loss of generality, we demonstrate this property using the general regularization approach where we fix $\alpha \in [0, 1]$ and have the following (approximate) closed form solution for the regularized effect size estimates (99; 243; 308)

$$
\tilde{\beta} \simeq \mathbf{H}\hat{\beta}, \quad \mathbf{H} = (\Sigma + \vartheta\mathbf{D}^{-1})^{-1}
$$

with $\vartheta \geq 0$ being a penalization parameter that has one-to-one correspondence with $t$ in Eq. (1.4). Here, $\mathbf{H}$ is commonly referred to as the “linear shrinkage estimator” [citation], where $\mathbf{D}$ is a diagonal weight matrix with nonzero elements dictated by the type of regularization that is being used. For example, $\mathbf{D} = \mathbf{I}$ while performing ridge regression (99), and $\mathbf{D} = \text{diag}(|\tilde{\beta}_1|, \ldots, |\tilde{\beta}_p|)$ while using ridge-based approximations for the elastic net and lasso solutions (243; 308). From Eq. (1.16), it is clear that $\tilde{\beta}$ may be interpreted as a marginal estimator of SNP-level effects after accounting for LD structure. Using Eqs. (1.2)-(1.3), it is straightforward to show the (approximate)
relationship between the regularized effect size estimates and the true coefficient values

\[ \mathbb{E}[\tilde{\beta}] \simeq \mathbf{H}\Sigma\beta. \]  

(1.17)

As described in the main text, the accuracy of this relationship is dependent upon both the sample size and narrow-sense heritability of the trait of interest (Fig. A.1). Indeed, if \( \Sigma \) is full rank and regularization is no longer implemented (i.e., \( \vartheta = 0 \)), \( \tilde{\beta} \) is simply the ordinary least squares solution for marginal GWA summary statistics with asymptotic variance-covariance \( \mathbb{V}[\tilde{\beta}] \simeq \Sigma \) under the null model (103; 104; 299). In the limiting case where the number of observations in a GWA study is large (i.e., \( N \to \infty \)) and the trait of interest is highly heritable, \( \beta \) converges onto \( \tilde{\beta} \) in expectation; and thus is assumed to be independently and normally distributed under the null hypothesis with asymptotic variance \( \sigma^2_\varepsilon \mathbf{I} \) (previously discussed in Eq. (1.5)). As empirically demonstrated for synthetic traits in the current study, we are rarely in situations where we expect the regularized effect size estimates to have completely converged onto the true generative SNP-level coefficients (again see Fig. A.1). This effectively means that we cannot expect each \( \tilde{\beta}_j \) to be completely independent under the null hypothesis in practice. We accommodate this realization by assuming that under the null model

\[ \mathbb{V}[\tilde{\beta}] = \sigma^2_\varepsilon \Sigma, \quad \lim_{\sigma^2_\varepsilon \to 0} \sigma^2_\varepsilon \Sigma = \sigma^2_\varepsilon \mathbf{I} \]  

(1.18)

Our reasoning for the formulation above is that, for most quality controlled studies, SNPs in perfect LD will have been pruned such that \( \rho(x_j, x_{j'}) < \rho(x_j, x_j) \) for all \( j \neq j' \) variants in the data. Therefore, when traits are generated under the idealized null scenario with large sample sizes and no genetic effects, the estimate of \( \sigma^2_\varepsilon \to 0 \) and the off-diagonals of \( \sigma^2_\varepsilon \Sigma \) will approach zero quicker than the diagonal elements; thus, allowing the regularized \( \tilde{\beta} \) to asymptotically converge onto the true coefficients \( \beta \). When this scenario does not occur, we are able to appropriately deal with the remaining correlation structure (e.g., all the simulation scenarios explored in this work; see Figs. 1.3 and A.2 - A.24, and Tables 1 and A.1 - A.17).
1.5.5 Using the SNP-Level Null Threshold to Detect Enriched Genes

We now formalize the hypothesis test for identifying significantly enriched genes conditioned on the SNP-level null threshold $\sigma^2$, which we compute using the variance component estimates from the EM algorithm detailed in the previous section. The gene-$\varepsilon$ gene-level test statistic is based on a quadratic form using GWA summary statistics, which is a common approach for generating gene-level test statistics for complex traits. Let gene (or genomic region) $g$ represent a known set of SNPs $j \in J_g$; for example, $J_g$ may include SNPs within the boundaries of $g$ and/or within its corresponding regulatory region. Here, we conformably partition the regularized GWA effect size estimates $\tilde{\beta}$ and define the gene-level test statistic

$$Q_g = \tilde{\beta}_g^T A \tilde{\beta}_g,$$  \hspace{1cm} (1.19)

where $A$ is an arbitrary symmetric and positive semi-definite weight matrix. We set to $A = I$ to be the identity matrix for all analyses in the current study; hence, $Q_g$ simplifies to a sum of squared SNP effects in the $g$-th gene. Indeed, similar quadratic forms have been implemented to assess the enrichment of mutations at the gene level (154; 183) and across general SNP-sets (281; 282; 139; 109). A key feature of the gene-$\varepsilon$ framework is to assess the statistics in Eq. (1.19) against a gene-level enrichment null hypothesis $H_0: Q_g = 0$ that is dependent on the SNP-level null threshold $\sigma^2$. Due to the normality assumption for each SNP effect in Eq. (1.5), $Q_g$ is theoretically assumed to follow a mixture of chi-square distributions,

$$Q_g \sim \sum_{j=1}^{|J_g|} \lambda_j \chi^2_{1,j},$$  \hspace{1cm} (1.20)

where $|J_g|$ denotes the cardinality of the set of SNPs $J_g$; $\chi^2_{1,j}$ are standard chi-square random variables with one degree of freedom; and $(\lambda_1, \ldots, \lambda_{|J_g|})$ are the eigenvalues of the matrix (302; 49)

$$\forall [\tilde{\beta}_g]^{1/2} A [\tilde{\beta}_g]^{1/2} = \sigma^2 \Sigma_g^{1/2} A \Sigma_g^{1/2}.$$
Again, in the current study, $\sigma^2 = \hat{\sigma}^2$ from the estimates in Eq. (1.15), and $\Sigma_g$ denotes a subset of the LD matrix only containing SNPs annotated in the $g$-th SNP-set. Again, when $A = I$, the eigenvalues are based on a scaled version of the local gene-specific LD matrix. Several approximate and exact methods have been suggested to obtain $P$-values under a mixture of chi-square distributions. In this study, we use Imhof’s method (108) where we empirically compute an estimate of the weighted sum in Eq. (1.20) and compare this distribution to the observed test statistic in Eq. (1.19) (see Software Details). It is important to note here that the gene-level null hypothesis is the same for gene-\(\varepsilon\) and other similar competing enrichment methods (281; 282; 139; 109; 183); the defining characteristic that sets gene-\(\varepsilon\) apart is that it assumes a different null distribution for effects on the SNP-level.

**Estimating Gene Specific Contributions to the PVE.** In the main text, we highlight some of the additional features of the gene-\(\varepsilon\) gene-level association test statistic. First, the expected enrichment for trait-associated mutations in a given gene is equal to the heritability explained by the SNPs contained in said gene. Formally, consider the expansion of Eq. (1.19) derived from the expectation of quadratic forms,

$$
\mathbb{E}[\tilde{Q}_g] = \sum_{j=1}^{n_g} \sum_{j'=1}^{n_g} a_{jj'} \mathbb{E}[\beta_j \beta_{j'}] = h^2_{g}, \quad (1.21)
$$

where denotes the heritability contributed by gene $g$. When $A = I$ (as in the current study), the gene-\(\varepsilon\) hypothesis test for identifying enriched genes is based on the individual SNP contributions to the narrow-sense heritability (i.e., the sum of the expectation of squared SNP effects; see also (218))

$$
\mathbb{E}[\tilde{Q}_g] = \sum_{j=1}^{n_g} \mathbb{E}[\tilde{\beta}_j^2] = h^2_{g}, \quad (1.22)
$$

Alternatively, one could choose to re-weight these contributions by specifying $A$ otherwise (281; 183; 37; 300; 302). For example, if SNP $j$ has a small effect size but is known to be functionally associated with the trait of interest, then increasing $A_{jj}$ will reflect this knowledge. Specific weight functions have also been suggested for dealing with rarer variants (282; 139; 109).
1.5.6 Simulation Studies

We used a simulation scheme to generate SNP-level summary statistics for GWA studies. First, we randomly select a set of enriched genes and assume that complex traits (under various genetic architectures) are generated via a linear model

\[
y = Wb + \sum_{c \in \mathcal{C}} x_c \beta_c + e, \quad e \sim \mathcal{N}(0, \tau^2 I),
\]

where \( y \) is an \( N \)-dimensional vector containing all the phenotypes; \( \mathcal{C} \) represents the set of causal SNPs contained within the associated genes; \( x_c \) is the genotype for the \( c \)-th causal SNP encoded as 0, 1, or 2 copies of a reference allele; \( \beta_c \) is the additive effect size for the \( c \)-th SNP; \( W \) is an \( N \times M \) matrix of covariates representing additional population structure (e.g., the top ten principal components from the genotype matrix) with corresponding fixed effects \( b \); and \( e \) is an \( N \)-dimensional vector of environmental noise. The phenotypic variance is assumed \( \mathbb{V}[y] = 1 \). The effect sizes of SNPs in enriched genes are randomly drawn from standard normal distributions and then rescaled so they explain a fixed proportion of the narrow-sense heritability \( \mathbb{V}[\sum x_c \beta_c] = h^2 \).

The covariate coefficients are also drawn from standard normal distributions and then rescaled such that \( \mathbb{V}[Wb] + \mathbb{V}[e] = (1 - h^2) \). GWA summary statistics are then computed by fitting a single-SNP univariate linear model via ordinary least squares (OLS): \( \hat{\beta}_j = (x_j^T x_j)^{-1} x_j^T y \) for every SNP in the data \( j = 1, \ldots J \). These effect size estimates, along with an LD matrix \( \Sigma \) computed directly from the full \( N \times J \) genotype matrix \( X \), are given to gene-\( \varepsilon \). We also retain standard errors and \( P \)-values for implementation of the competing methods (VEGAS, PEGASUS, RSS, SKAT, and MAGMA).

Given different model parameters, we simulate data mirroring a wide range of genetic architectures (Supporting Information).
1.6 Software Details

Source code implementing gene-$\varepsilon$ and tutorials are freely available at https://github.com/ramachandran-lab/genee and was written in R (version 3.3.3). Within this software, regularization of the OLS SNP-level effect sizes is done using the package glmnet (version 2.0-16) (74). For large datasets, such as the UK Biobank, the software also offers regularization using the biglasso (version 1.3-6) (297) to help with memory and scalability requirements. Note that selection of the free parameter $t$ is done the same way using both the glmnet and biglasso packages. Both packages also take in an $\alpha \in [0,1]$ to specify fitting the Ridge, Elastic Net or Lasso regularization to the OLS SNP-level effect sizes. The fitting of a $K$-mixture of Gaussian distributions for the estimation of the SNP-level null threshold $\sigma^2_\varepsilon$ is done using the package mclust (version 5.4.3) (215). Lastly, the package CompQuadForm (version 1.4.3) was used to compute gene-$\varepsilon$ gene-level $P$-values with Imhof’s method (108; 64). Comparisons in this work were made using software for MAGMA (version 1.07b; https://ctg.cnclr.nl/software/magma), PEGASUS (version 1.3.0; https://github.com/ramachandran-lab/PEGASUS), RSS (version 1.0.0; https://github.com/stephenslab/rss), SKAT (version 1.3.2.1; https://www.hsph.harvard.edu/skat), VEGAS (version 2.0.0; https://vegas2.qimrberghofer.edu.au) which are also publicly available. See all other relevant URLs below.

1.7 URLs

Chapter 2

Multi-scale Inference of Genetic Trait Architecture using Biologically Annotated Neural Networks

2.1 Abstract

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

2.2 Author Summary

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

2.3 Introduction

Over the two last decades, a considerable amount of methodological research in statistical genetics has focused on developing and improving the utility of linear models (118; 119; 194; 152; 128; 303; 94; 96; 49; 297; 157; 112; 207). The flexibility and interpretability of linear models make them a widely used tool in genome-wide association (GWA) studies, where the goal is to test for statistical associations between individual single nucleotide polymorphisms (SNPs) and a phenotype of interest. In these cases, traditional variable selection approaches provide a set of $P$-values or posterior inclusion probabilities (PIPs) which lend statistical evidence on how important each variant is for explaining the overall genetic architecture of a trait. However, this univariate SNP-level mapping approach can be underpowered for “polygenic” traits which are generated by many mutations of small effect (166; 255; 305; 289; 21; 280). To mitigate this issue, more recent work has extended variable selection techniques to identify enriched gene or pathway-level associations, where groups of SNPs within a particular genomic region are combined (commonly known as a SNP-set) to detect biologically relevant disease mechanisms underlying the trait (154; 281; 28; 56; 132; 183; 307; 237). Still, the performance of standard SNP-set methods can be hampered by strict additive modeling assumptions; and the most powerful of these statistical approaches rely on algorithms that are computationally inefficient and unreliable for large-scale sets of data (39).

The explosion of large-scale genomic datasets has provided the unique opportunity to move beyond the traditional linear regression framework and integrate nonlinear modeling techniques as standard statistical tools within GWA analyses. Indeed, nonlinear methods such as neural networks are well known to be most powered in settings when large training data is available (134). This includes GWA applications where consortiums have data sets that include hundreds of thousands of individuals genotyped at millions of markers and phenotyped for thousands of traits (182; 24). It is also well known that these nonlinear statistical approaches often exhibit
greater predictive accuracy than linear models, particularly for complex traits with broad-sense heritability that is driven by non-additive genetic variation (e.g., gene-by-gene interactions) (269; 9). One of the key characteristics that leads to better predictive performance from nonlinear approaches is the automatic inclusion of higher order interactions between variables being put into the model (113; 51). For example, neural networks leverage activation functions between layers that implicitly enumerate all possible (polynomial) interaction effects (257). While this is a partial mathematical explanation for model improvement, in many biological applications, we often wish to know precisely which subsets of variants are most important in defining the architecture of a trait. Unfortunately, the classic statistical idea of variable selection and hypothesis testing is lost within nonlinear methods since they do not naturally produce interpretable significance measures (e.g., P-values or PIPs) like traditional linear regression (51; 52).

In this work, we develop biologically annotated neural networks (BANNs), a nonlinear probabilistic framework for mapping and variable selection in high-dimensional genomic association studies (Fig. 2.1). BANNs are a class of feedforward Bayesian models with partially connected architectures that are guided by predefined SNP-set annotations (Fig. 2.1a). The interpretability of our approach stems from a combination of three key properties. First, the partially connected network architecture yields a hierarchical model where the input layer encodes SNP-level effects, and the single hidden layer models the effects among SNP-sets (Fig. 2.1b). Second, inspired by previous work in the Bayesian neural network literature (47; 60; 229; 147; 81), 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. 2.1c). Third, we perform an integrative model fitting procedure where the enrichment of SNP-sets in the hidden layer are directly influenced by the distribution of associated SNPs with nonzero effects on the input layer. These three components collectively make for an effective nonlinear variable selection strategy for conducting association mapping and enrichment analyses simultaneously on complex traits. With detailed simulations, we assess the power of BANNs to identify significant SNPs and SNP-sets under a variety of genetic architectures, and compare its performance against multiple
competing approaches (101; 281; 56; 11; 104; 183; 307; 237; 262). We also apply the BANNs framework to six quantitative traits assayed in a heterogeneous stock of mice from Wellcome Trust Centre for Human Genetics (249), and two quantitative traits in individuals from the Framingham Heart Study (228). For the latter, we include an additional study where we independently analyze the same traits in a subset of individuals of European ancestry from the UK Biobank (24).

2.4 Results

2.4.1 BANNs Framework Overview

Biologically annotated neural networks (BANNs) are feedforward models with partially connected architectures that are inspired by the hierarchical nature of biological enrichment analyses in GWA studies (Fig. 2.1). The BANNs software takes in one of two data types: (i) individual-level data $\mathcal{D} = \{ \mathbf{X}, \mathbf{y} \}$ where $\mathbf{X}$ is an $N \times J$ matrix of genotypes with $J$ denoting the number of single nucleotide polymorphisms (SNPs) encoded as $\{0, 1, 2\}$ copies of a reference allele at each locus and $\mathbf{y}$ is an $N$-dimensional vector of quantitative traits (Fig. 2.1a); or (ii) GWA summary statistics $\mathcal{D} = \{ \mathbf{R}, \mathbf{\theta} \}$ where $\mathbf{R}$ is a $J \times J$ empirical linkage disequilibrium (LD) matrix of pairwise correlations between SNPs and $\mathbf{\theta}$ are marginal effect size estimates for each SNP computed using ordinary least squares (OLS) (Fig. B.1). In both settings, the BANNs software also requires a predefined list of SNP-set annotations $\{S_1, \ldots, S_G\}$ to construct partially connected network layers that represent different scales of genomic units. Structurally, sequential layers of the BANNs model represent different scales of genomic units. The first layer of the network takes SNPs as inputs, with each unit corresponding to information about a single SNP. The second layer of the network represents SNP-sets. All SNPs that have been annotated for the same SNP-set are then connected to the same neuron in the second layer (Fig. 2.1b).

In this section, we review the hierarchical probabilistic specification of the BANNs framework for individual data; however, note that extensions to summary statistics is straightforward and only requires substituting the genotypes $\mathbf{X}$ for the LD matrix $\mathbf{R}$ and substituting the phenotypes $\mathbf{y}$ for the OLS effect sizes $\mathbf{\theta}$ (see Material and Methods). Without loss of generality, let SNP-set $g$
represent an annotated collection of SNPs $j \in S_g$ with cardinality $|S_g|$. The BANNs framework is probabilistically represented as a nonlinear regression model

$$y = \sum_{g=1}^{G} h(X_g \theta_g + 1b^{(1)}_g)w_g + 1b^{(2)},$$  \hspace{1cm} (2.1)$$

where $X_g = [x_1, \ldots, x_{|S_g|}]$ is the subset of SNPs annotated for SNP-set $g$; $\theta_g = (\theta_1, \ldots, \theta_{|S_g|})$ are the corresponding inner layer weights; $h(\bullet)$ denotes the nonlinear activations defined for neurons in the hidden layer; $w = (w_1, \ldots, w_G)$ are the weights for the $G$-predefined SNP-sets in the hidden layer; $b^{(1)} = (b^{(1)}_1, \ldots, b^{(1)}_G)$ and $b^{(2)}$ are deterministic biases that are produced during the network training phase in the input and hidden layers, respectively; and $1$ is an $N$-dimensional vector of ones. Here, we define $h(\bullet)$ to be a Leaky rectified linear unit (Leaky ReLU) activation function, where $h(x) = x$ if $x > 0$ and $0.01x$ otherwise. Lastly, for convenience, we assume that the genotype matrix (column-wise) and trait of interest have been mean-centered and standardized.
Figure 2.1. Biologically annotated neural networks (BANNs) allow for efficient multi-scale genotype-phenotype analyses in a unified probabilistic framework by leveraging the hierarchical nature of enrichment studies to define network architecture. (a) The BANNs framework requires an $N \times J$ matrix of individual-level genotypes $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_J]$, an $N$-dimensional phenotypic vector $\mathbf{y}$, and a list of $G$-predefined SNP-sets $\{S_1, \ldots, S_G\}$. 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 (195). (b) A partially connected Bayesian neural network is constructed based on the annotated SNP groups. In the first hidden layer, only SNPs within the boundary of a gene are connected to the same node. Similarly, SNPs within the same intergenic region between genes are connected to the same node. Completing this specification for all SNPs gives the hidden layer the natural interpretation of being the “SNP-set” layer. (c) The hierarchical nature of the network is represented as nonlinear regression model. The corresponding weights in both the SNP ($\theta$) and SNP-set ($w$) layers are treated as random variables with biologically motivated sparse prior distributions. Posterior inclusion probabilities $\text{PIP}(j) \equiv \Pr(\theta_j \neq 0 \mid \mathbf{y}, \mathbf{X})$ and $\text{PIP}(g) \equiv \Pr(w_g \neq 0 \mid \mathbf{y}, \mathbf{X}, \theta_g)$ summarize associations at the SNP and SNP-set level, respectively. The BANNs framework uses variational inference for efficient network training and incorporates nonlinear processing between network layers for accurate estimation of phenotypic variance explained (PVE).

In this work, we define SNP-sets as collections of contiguous regions of the genome that contain variants within some chromosomal window or neighborhood. More specifically, when studying real mice and human GWA data, we use gene annotations as defined by the Mouse Genome Informatics database (23) and the NCBI’s Reference Sequence (RefSeq) database in the UCSC Genome Browser (195), respectively (Materials and Methods). The BANNs framework flexibly allows for overlapping annotations. In this way, SNPs may be connected to multiple hidden layer units if they are located within the intersection of multiple gene boundaries. SNPs that are unannotated, but located within the same genomic region, are connected to their own units in the second layer and represent the intergenic region between two annotated genes. Given the natural biological
interpretation of both layers, the partially connected architecture of the BANNs model creates a unified framework for comprehensibly understanding SNP and SNP-set level contributions to the broad-sense heritability of complex traits and phenotypes. Notably, this framework may be easily extended to other biological annotations and applications.

The framing of the BANNs methodology as a Bayesian nonlinear model helps facilitate our ability to perform classic variable selection (Fig. 2.1c; see Materials and Methods). Here, we leverage the fact that using nonlinear activation functions for the neurons in the hidden layer implicitly accounts for both additive and non-additive effects between SNPs within a given SNP-set (Supporting Information). Following previous work in the literature (47; 60; 229; 147; 81), we treat the weights and connections of the neural network as random variables with prior distributions that reflect how genetic effects are manifested at different genomic scales. For the input layer, we assume that the effect size distribution of non-null SNPs can take vastly different forms depending on both the degree and nature of trait polygenicity (39). 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 (280). To this end, we place a normal mixture prior on the input layer weights to flexibly estimate a wide range of SNP-level effect size distributions (179; 297; 155)

\[
\theta_j \sim \pi_\theta \sum_{k=1}^{K} \eta_{\theta k} N(0, \sigma_{\theta k}^2) + (1 - \pi_\theta) \delta_0
\]

where \( \delta_0 \) is a point mass at zero; \( \sigma_{\theta}^2 = (\sigma_{\theta 1}^2, \ldots, \sigma_{\theta K}^2) \) are variance of the \( K \)-nonzero mixture components; \( \eta_{\theta} = (\eta_{\theta 1}, \ldots, \eta_{\theta K}) \) represents the marginal (unconditional) probability that a randomly selected SNP belongs to the \( k \)-th mixture component such that \( \sum_k \eta_{\theta k} = 1 \); and \( \pi_\theta \) denotes the total proportion of SNPs that have a nonzero effect on the trait of interest. Here, we fix \( K = 3 \) which emulates the hypothesis that SNPs can have large, moderate, and small nonzero effects on phenotypic variation (39). Similarly, we follow other previous work and assume that enriched SNP-sets contain at least one SNP with a nonzero effect on the trait of interest (307). This is formulated
by placing a spike and slab prior distribution on the weights in the second layer

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

(2.3)

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

By using these point mass mixture distributions in Eqs. (2.2)-(2.3), we assume that each connection in the neural network has a nonzero weight with: (i) probability \( \pi_{\theta} \) for SNP-to-SNP-set connections, and (ii) probability \( \pi_w \) for SNP-set-to-phenotype connections. By modifying a variational inference algorithm assuming point-normal priors in multiple linear regression (27; 30) to the neural network setting, we jointly infer posterior inclusion probabilities (PIPs) for SNPs and SNP-sets. These quantities are defined as the posterior probability that the weight of a given connection in the neural network is nonzero, \( \text{PIP}(j) \equiv \Pr[\theta_j \neq 0 | y, X] \) and \( \text{PIP}(g) \equiv \Pr[w_g \neq 0 | y, X, \theta_g] \).

We use this information to prioritize statistically associated SNPs and SNP-sets that significantly contribute to the broad-sense heritability of the trait of interest. With biologically annotated units and the ability to perform statistical inference on explicitly defined parameters, our model presents a fully interpretable extension of neural networks to GWA applications. Additional details and derivations of the BANNs framework can be found in Materials and Methods and Supporting Information.

2.4.2 Power to Detect SNPs and SNP-Sets in Simulation Studies

In order to assess the performance of models under the BANNs framework, we simulated complex traits under multiple genetic architectures using real genotype data on chromosome 1 from ten thousand randomly sampled individuals of European ancestry in the UK Biobank (24) (see Materials and Methods and previous work (49; 39)). After quality control procedures, our simulations included 36,518 SNPs (Supporting Information). Next, we used the NCBI’s Reference Sequence (RefSeq) database in the UCSC Genome Browser (195) to annotate SNPs with the appropriate genes. Unannotated SNPs located within the same genomic region were labeled as being within
the “intergenic region” between two genes. Altogether, this left a total of $G = 2,816$ SNP-sets to be included in the simulation study.

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

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

Throughout this section, we assess the performance for two versions of the BANNs framework. The first takes in individual-level genotype and phenotype data; while, the second models GWA summary statistics (hereafter referred to as BANN-SS). For the latter, GWA summary statistics are computed by fitting a single-SNP univariate linear model (via ordinary least squares) after quality control to obtain: effect size estimates, standard errors, and $P$-values for all SNPs in the data. We also use the in-sample genotypes to compute the LD matrix between SNPs. All results are based on 100 different simulated phenotypes for each parameter combination (Supporting Information).
The main utility of the BANNs framework is having the ability to detect associated SNPs and enriched SNP-sets simultaneously. Therefore, we compare the performance of BANNs to state-of-the-art SNP and SNP-set level approaches (101; 281; 56; 11; 104; 183; 307; 237; 262), 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 (Tables B.1-B.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 \( \geq 0.5 \)) (6). We also compare each method’s ability to rank true positives over false positives via receiver operating characteristic (ROC) and precision-recall curves (Fig. 2.2 and Figs. B.2-B.16). Specific results about these analyses are given below.

**Mapped SNP-Level Results.** For SNP-level comparisons, we used three fine-mapping methods as benchmarks: CAVIAR (104), SuSiE (262), and FINEMAP (11). Each of these methods implement Bayesian variable selection strategies, in which different sparse prior distributions are placed on the “true” effect sizes of each SNP and posterior inclusion probabilities (PIPs) are used to summarize their statistical relevance to the trait of interest. Notably, both CAVIAR (exhaustively) and FINEMAP (approximately) search over different models to find the best combination of associated SNPs with nonzero effects on a given phenotype. On the other hand, the software for SuSiE requires an input \( \ell \) which fixes the maximum number of causal SNPs to include in the model. In this section, we consider results when this input number is high (\( \ell = 3000 \)) and when this input number is low (\( \ell = 10 \)). While SuSiE is applied to individual-level data, both CAVIAR and FINEMAP require summary statistics where marginal z-scores are treated as a phenotype and modeled with in-sample estimate of the LD matrix.

Overall, BANNs, BANN-SS, and SuSiE (with high \( \ell = 3000 \)) generally achieve the greatest empirical power and lowest FDR across all genetic architectures we considered (Tables B.1-B.8).
These three approaches also stand out in terms of true-versus-false positive rates and precision-versus-recall (Fig. 2.2 and Figs. B.2-B.16). Notably, the choice of the $\ell$ parameter largely influenced the performance of SuSiE, as it was consistently the worst performing method when we underestimated the number of causal SNPs with nonzero effects \textit{a priori} (i.e., $\ell = 10$). Importantly, these performance gains come with a cost: the computational run time of SuSiE becomes much slower as $\ell$ increases (Table B.9). For more context, an analysis on just 4,000 individuals and 10,000 SNPs takes the BANNs methods an average of 319 seconds to run on a CPU; while, SuSiE can take up to nearly twice as long to complete as $\ell$ increases (e.g., average runtimes of 23 and 750 seconds for $\ell = 10$ and 3000, respectively).

Training BANNs on individual-level data relatively becomes the best approach when the broad-sense heritability of complex traits is partly made up of pairwise genetic interaction effects between causal SNPs (e.g., $\rho = 0.5$; see Figs. B.5-B.8 and B.13-B.16)—particularly when traits have low heritability with polygenic architectures (e.g., $H^2 = 0.2$). A direct comparison of the PIPs derived by BANNs and SuSiE shows that the proposed neural network training procedure enables the ability to identify associated SNPs even in these more complex phenotypic architectures (Fig. 2.3 and Figs. B.17-B.23). It is important to note that the inclusion probabilities were not perfectly calibrated for either BANNs or SuSiE in our simulations (Fig. B.24), despite FDR still being reasonably well controlled for both methods (Tables B.1-B.8). We hypothesize that the quality of PIP calibration for BANNs is a direct consequence of its variational inference algorithm which tends to favor sparse solutions and can lead to greater type II versus type I error rates (27; 262). To investigate how choices in the BANNs model setup contributed to improved variable selection over SuSiE, we also performed an “ablation analysis” (77; 78) where we modified parts of the algorithm independently and observed their direct effect on method performance (Fig. B.25). Ultimately, these results for BANNs were enabled by a combination of (i) using ReLU activation functions in the hidden layers of the BANNs framework, which implicitly enumerates the interactions between SNPs within a given SNP-set, and (ii) using model averaging to estimate the inclusion probabilities for the network weights (Supporting Information). Note the absence of the nonlinear activation function only affected the power of BANNs in simulations where there were non-additive genetic
effects (e.g., Figs. B.25c and B.25d).

As a final comparison, the BANN-SS, CAVIAR, and FINEMAP methods see a decline in performance for these same scenarios with genetic interactions. Assuming that the additive and non-additive genetic effects are uncorrelated, this result is also expected since summary statistics are often derived from simple linear additive regression models that (in theory) partition or marginalize out proportions of the phenotypic variance that are contributed by nonlinearities (49; 207).
Figure 2.2. Receiver operating characteristic (ROC) curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations (British cohort). Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.6$ with only contributions from additive effects set (i.e., $\rho = 1$). We show power versus false positive rate for two different trait architectures: (a, b) sparse where only 1% of SNP-sets are enriched for the trait; and (c, d) polygenic where 10% of SNP-sets are enriched. We set the number of causal SNPs with nonzero effects to be 1% and 10% of all SNPs located within the enriched SNP-sets, respectively. To derive results, the full genotype matrix and phenotypic vector are given to the BANNs model and all competing methods that require individual-level data. For the BANN-SS model and other competing methods that take GWA summary statistics, we compute standard GWA SNP-level effect sizes and $P$-values (estimated using ordinary least squares). (a, c) Competing SNP-level mapping approaches include: CAVIAR (104), SuSiE (262), and FINEMAP (11). The software for SuSiE requires an input $\ell$ 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 (307), PEGASUS (183), GBJ (237), SKAT (281), GSEA (101), and MAGMA (56). Note that the upper limit of the x-axis has been truncated at 0.1. All results are based on 100 replicates (see Supporting Information).
Figure 2.3. Scatter plots comparing how the integrative neural network training procedure enables the ability to identify associated SNPs and enriched SNP-sets in simulations (British cohort). Quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.6$ with only contributions from additive effects set (i.e., $\rho = 1$). We consider two different trait architectures: (a, b) sparse where only 1% of SNP-sets are enriched for the trait; and (c, d) polygenic where 10% of SNP-sets are enriched. We set the number of causal SNPs with nonzero effects to be 1% and 10% of all SNPs located within the enriched SNP-sets, respectively. Results are shown comparing the posterior inclusion probabilities (PIPs) derived by the BANNs model on the x-axis and (a, c) SuSiE (262) and (b, d) RSS (307) 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) (6). True positive causal variants used to generate the synthetic phenotypes are colored in red, while non-causal variants are given in grey. SNPs and SNP-sets in the top right quadrant are selected by both approaches; while, elements in the bottom right and top left quadrants are uniquely identified by BANNs and SuSiE/RSS, respectively. Each plot combines results from 100 simulated replicates (see Supporting Information).
Enriched SNP-Set Level Results. For comparisons between SNP-set level methods, we consider six gene or SNP-set enrichment approaches including: RSS (307), PEGASUS (183), GBJ (237), SKAT (281), GSEA (101), and MAGMA (56). 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 (139). 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 (56). 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 (13). Lastly, RSS is a Bayesian linear regression method which places a likelihood on the observed SNP-level GWA effect sizes (using their standard errors and LD estimates), and assumes a spike-and-slab shrinkage prior on the true SNP effects to derive a probability of enrichment for genes or other annotated units (306). It is worth noting that, while RSS and the BANNs framework are conceptually different, the two methods utilize very similar variational approximation algorithms for posterior inference (27) (Materials and Methods, and Supporting Information).

Similar to the conclusions drawn during the SNP-level assessments, both the BANNs and BANN-SS implementations had among the best tradeoffs between true and false positive rates for detecting enriched SNP-sets across all simulations—once again, including those scenarios which also considered pairwise interactions between causal SNPs (Fig. 2.2, Figs. B.2-B.16, and Tables B.1-B.8). Since RSS is an additive model, it sees a decline in performance for the more complex genetic architectures that we simulated. A direct comparison between the PIPs from BANNs and RSS can be found in Fig. 2.3 and Figs. B.17-B.24. Once again, training BANNs on individual-level data becomes the best approach when the broad-sense heritability of complex traits is partly made up of non-additive genetic variation. Our ablation analysis results suggest that the nonlinear activation function plays an important role here (Fig. B.25). While RSS also performs generally well for the additive trait architectures, the algorithm for the model often takes twice as long than either of the BANNs implementations to converge (Table B.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. 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 (39). Therefore, it is understandable why the score-based methods consistently struggle relative to the regression-based approaches even in the simplest simulation cases where traits are generated to have high broad-sense heritability, sparse phenotypic architectures that are dominated by additive genetic effects, and total phenotypic variance that is not confounded by additional population structure (Fig. 2.2 and Figs. B.2-B.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.

### 2.4.3 Estimating Total Phenotypic Variance Explained in Simulation Studies.

While our main focus is on conducting multi-scale inference of genetic trait architecture, because the BANNs framework provides posterior estimates for all weights in the neural network, we are able to also provide an estimate of phenotypic variance explained (PVE). Here, we define PVE as the total proportion of phenotypic variance that is explained by genetic effects, both additive and non-additive, collectively (305). Within the BANNs framework, this estimation can be done on both the SNP and SNP-set level while using either genotype-phenotype data or summary statistics (Supporting Information). As a reminder, for our simulation studies, the true PVE is set to \( H^2 = 0.2 \) and 0.6, respectively. We assess the ability of BANNs to recover these true estimates using root mean square error (RMSE) (Figs. B.26 and B.27). In order to be successful at this task, the neural network needs to accurately estimate both the individual effects of causal SNPs in the input layer, as well as their cumulative effects for SNP-sets in the outer layer. BANNs and BANN-SS exhibit the most success with traits have additive sparse architectures (with and without additional population structure)—achieving PVE estimates with RMSEs as low as \( 4.54 \times 10^{-3} \)
and $4.78 \times 10^{-3}$ on the SNP and SNP-set levels for highly heritable phenotypes, respectively. However, both models underestimate the total PVE in polygenic traits and traits with pairwise SNP-by-SNP interactions. Therefore, even though the BANNs framework is still able to correctly prioritize the appropriate SNPs and SNP-sets, in these more complicated settings, we misestimate the approximate posterior means for the network weights and overestimate the variance of the residual training error (Supporting Information). Similar observations have been noted when using variational inference (17; 84). Results from other work also suggest that the sparsity assumption on the SNP-level effects can lead to the underestimation of the PVE (91; 305).

2.4.4 Mapping Genomic Enrichment in Heterogenous Stock of Mice

We apply the BANNs framework to individual-level genotypes and six quantitative traits in a heterogeneous stock of mice dataset from the Wellcome Trust Centre for Human Genetics (249). 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 (Supporting Information). For SNP-set annotations, we used the Mouse Genome Informatics database (http://www.informatics.jax.org) (23) to map SNPs to the closest neighboring gene(s). Unannotated SNPs located within the same genomic region were labeled as being within the “intergenic region” between two genes. Altogether, a total of 2,616 SNP-sets were analyzed. The six traits that we consider are grouped based on their category and include: body mass index (BMI) and body weight; percentage of CD8+ cells and mean corpuscular hemoglobin (MCH); and high-density and low-density lipoprotein (HDL and LDL, respectively). We choose to analyze these particular traits because their architectures represent a realistic mixture of the simulation scenarios we detailed in the previous section (i.e., varying different values of $\rho$). Specifically, the mice in this study are known to be genetically related with population structure and these particular traits have been shown to have various levels of broad-sense heritability with different contributions from both additive and non-additive genetic effects (249; 36? ? ; 51; 52).

For each trait, we provide a summary table which lists the PIPs for SNPs and SNP-sets after fitting the BANNs model to the individual-level genotypes and phenotype data (Tables B.11-B.16).
We use Manhattan plots to visually display the variant-level mapping results across each of the six traits, where chromosomes are shown in alternating colors for clarity and associated SNPs with PIPs above the median probability model threshold are highlighted (Fig. B.28). As a comparison, we also report the corresponding SNP and SNP-set level PIPs after running SuSiE (262) and RSS (307) on these same data, respectively. Across all traits, BANNs identified 71 associated SNPs and 57 enriched SNP-sets (according to the median probability model threshold). In comparison, SuSiE identified 22 associated SNPs (11 of which were also identified by BANNs) and RSS identified 14 enriched SNP-sets (6 of which were also identified by BANNs). Importantly, many of the candidate genes and intergenic regions selected by the BANNs model have been previously discovered by past publications as having some functional relationship with the traits of interest (Table 2.1). For example, BANNs reports the genes Btbd9 and hlb156 as being enriched for the percentage of CD8+ cells in mice (PIP = 0.87 and 0.72 versus RSS PIP = 0.02 and 0.68, respectively). This same chromosomal region on chromosome 17 was also reported in the original study as having highly significant quantitative trait loci and contributing non-additive variation for CD8+ cells (bootstrap posterior probability equal to 1.00) (249). Similarly, the X chromosome is well known to strongly influence adiposity and metabolism in mice (36). As expected, in body weight and BMI, our approach identified significant enrichment in this region—headlined by the dystrophin gene Dmd in both cases (235). Finally, we note that including intergenic regions in our analyses allows us to discover trait relevant genomic associations outside the immediate gene annotations provided by the Mouse Genome Informatics database. This proved important for BMI where BANNs reported the region between Gm22219 and Mc4r on chromosome 18 as having a relatively high PIP of 0.74 (versus an RSS PIP = $1 \times 10^{-3}$ for reference). Recently, a large-scale GWA study on individuals from the UK Biobank showed that variants around $MC4R$ protect against obesity in humans (158).

Overall, the results from this smaller GWA study highlight three key characteristics resulting from the sparse probabilistic assumptions underlying the BANNs framework. First, the variational spike and slab prior placed on the weights of the neural network will select no more than a few variants in a given LD block (27). 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 (39).
Second, we see that the enrichment of a SNP-set is influenced by the relative posterior distribution of zero and nonzero SNP-level effect sizes within its annotated genomic window (Tables B.11-B.16). In other words, a SNP-set is not guaranteed to have a high inclusion probability just because it contains one SNP with a large nonzero effect; however, BANNs will report a SNP-set as insignificant if the total ratio of non-causal SNPs within the set heavily outweighs the number of causal SNPs that have been annotated for the same region. To this end, in the presence of large SNP-sets, the BANNs framework will favor preserving false discovery rates at the expense of having slightly more false negatives. Lastly, the careful modeling of the SNP-level effect size distributions and considering genetic interactions enhances our ability to conduct multi-scale genomic inference. In this particular study, we show the power to still find trait relevant SNP-sets with variants that are not marginally strong enough to be detected individually, but have notable genetic signal when their weights are aggregated together (again see Table 2.1 and Fig. B.28).
Table 2.1. Notable enriched SNP-sets after applying the BANNs framework to six quantitative traits in heterogeneous stock of mice from the Wellcome Trust Centre for Human Genetics. (249). The traits include: body mass index (BMI), percentage of CD8+ cells, high-density lipoprotein (HDL), low-density lipoprotein (LDL), mean corpuscular hemoglobin (MCH), and body weight. Here, SNP-set annotations are based on gene boundaries defined by the Mouse Genome Informatics database (see URLs). Unannotated SNPs located within the same genomic region were labeled as being within the “intergenic region” between two genes. These regions are labeled as Gene1-Gene2 in the table. Posterior inclusion probabilities (PIP) for the input and hidden layer weights are derived by fitting the BANNs model on individual-level data. A SNP-set is considered enriched if it has a PIP($g$) ≥ 0.5 (i.e., the “median probability model” threshold (6)). We report the “top” associated SNP within each region and its corresponding PIP($j$). We also report the corresponding SNP and SNP-set level results after running SuSiE (262) and RSS (307) on these same traits, respectively. The last column details references and literature sources that have previously suggested some level of association or enrichment between the each genomic region and the traits of interest. See Tables B.11-B.16 for the complete list of SNP and SNP-set level results.
2.4.5 Analyzing Lipoproteins in the Framingham Heart Study

Next, we apply the BANNs framework to two continuous plasma trait measurements — high-density lipoprotein (HDL) and low-density lipoprotein (LDL) cholesterol — assayed in 6,950 individuals from the Framingham Heart Study (228) genotyped at 394,174 SNPs genome-wide. Following quality control procedures, we regressed out the top ten principal components of the genotype data from each trait to control for population structure (Supporting Information). Next, we used the gene boundaries listed in the NCBI’s RefSeq database from the UCSC Genome Browser (195) to define SNP-sets. In this analysis, we define genes with boundaries in two ways: (a) we use the UCSC gene boundary definitions directly, or (b) we augment the gene boundaries by adding SNPs within a ±500 kilobase (kb) buffer to account for possible regulatory elements. Genes with only 1 SNP within their boundary were excluded from either analysis. Unannotated SNPs located within the same genomic region were labeled as being within the “intergenic region” between two genes. Altogether, a total of $G = 18,364$ SNP-sets were analyzed—which included 8,658 intergenic SNP-sets and 9,706 annotated genes—using the UCSC boundaries. When including the 500kb buffer, a total of $G = 35,871$ SNP-sets were analyzed.

For each trait, we again fit the BANNs model to the individual-level genotype-phenotype data and used the median probability model threshold as evidence of statistical significance for all weights in the neural network (Tables B.17-B.19). We also again report the corresponding SNP and SNP-set level PIPs after running SuSiE and RSS on these same data. Note that while BANNs is run on the genome-wide data jointly, for computational considerations, SuSiE and RSS are run on a chromosome-by-chromosome basis. A complete breakdown of the overlap of findings between BANNs, SuSiE, and RSS can be found on the first page of Table B.20. In Fig. 2.4, we show Manhattan plots of the variant-level association mapping results for BANNs, where each significant SNP is color coded according to its SNP-set annotation. As an additional validation step, we took the enriched SNP-sets identified by BANNs in each trait and used the gene set enrichment analysis tool Enrichr (34; 129) to identify the categories that they overrepresent in the database of Genotypes and Phenotypes (dbGaP) and the NHGRI-EBI GWAS Catalog (Figs. B.29 and B.30).
Similar to our results in the previous section, the BANNs framework identified many SNPs and SNP-sets that have been shown to be associated with cholesterol-related processes in past publications (Table 2.2 with UCSC gene boundary definitions and Table B.17 with augmented buffer). For example, in HDL, BANNs identified an enriched intergenic region between the genes HERPUD1 and CETP (PIP = 1.00 versus RSS PIP = 0.78) which has been also replicated in multiple GWA studies with diverse cohorts (212; 208; 127; 95). The Enrichr analyses were also consistent with published results (Figs. B.29 and B.30). For example, the top ten significant enriched categories in the GWAS Catalog (i.e., Bonferroni-correct threshold $P$-value $< 1 \times 10^{-5}$ or $Q$-value $< 0.05$) for HDL-associated SNP-sets selected by the BANNs model are either directly related to lipoproteins and cholesterol (e.g., “Alipoprotein A1 levels”, “HDL cholesterol levels”) or related to metabolic functions (e.g., “Lipid metabolism phenotypes”, “Metabolic syndrome”).

As in the previous analysis, the results from this analysis also highlight insight into complex trait architecture enabled by the variational inference used in the BANNs software. SNP-level and SNP-set results remain consistent with the qualitative assumptions underlying our probabilistic hierarchical model. For instance, previous studies have estimated that rs599839 (chromosome 1, bp: 109822166) and rs4970834 (chromosome 1, bp: 109814880) explain approximately 1% of the phenotypic variation in circulating LDL levels (210). Since these two SNPs are physically closed to each other and sit in a high LD block ($r^2 \approx 0.63$ with $P < 1 \times 10^{-4}$ (161)), the spike and slab prior in the BANNs framework will maintain the nonzero weight for one and penalize the estimated effect of the other. Indeed, in our analysis, rs4970834 was reported to be associated with LDL (PIP = 0.95 versus SuSiE PIP = 1.12 $\times 10^{-4}$), while the effect size of rs599839 was shrunk towards 0 (PIP = $1 \times 10^{-4}$ versus SuSiE PIP = 0.99). A similar issue can occur in correctly identifying enriched SNP-sets when nearby sets contain SNPs in tight LD. For example, when augmenting the boundary of SNP-set annotations by a ±500 kilobase buffer, BANNs tends to shrink the PIP of at least one member of overlapping or correlated sets. Due to the variational approximations utilized by BANNs (Materials and Methods, and Supporting Information), if two SNPs or SNP-sets are in strong LD, the model will tend to select just one of them (27; 307).
Figure 2.4. Manhattan plot of variant-level association mapping results for high-density and low-density lipoprotein (HDL and LDL, respectively) traits in the Framingham Heart Study (228). Posterior inclusion probabilities (PIP) for the neural network weights are derived from the BANNs model fit on individual-level data and are plotted for each SNP against their genomic positions. Chromosomes are shown in alternating colors for clarity. The black dashed line is marked at 0.5 and represents the “median probability model” threshold (6). 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 (195). Unannotated SNPs located within the same genomic region were labeled as being within the “intergenic region” between two genes. These regions are labeled as Gene1-Gene2 in the legend. Double daggers (‡) denote SNPs that are also identified when using SuSiE (262) to analyze the same traits, and hashtag symbols (#) denote SNP-sets that are identified by RSS (307). Stars (★) denote SNPs and SNP-sets identified by BANNs that replicate in our analyses of HDL and LDL using ten thousand randomly sampled individuals of European ancestry from the UK Biobank (24). Gene set enrichment analyses for these SNP-sets identified by BANNs can be found in Figs. B.29 and B.30. A complete list of PIPs for all SNPs and SNP-sets computed in these two traits can be found in Tables B.18 and B.19. Results for the additional study with the independent UK Biobank dataset (24) are illustrated in Figures B.31-B.33 and full results are listed in Tables B.21 and B.22.
## Table 2.2: Top three enriched SNP-sets after applying the BANNs framework to high-density and low-density lipoprotein (HDL and LDL, respectively) traits in the Framingham Heart Study (228).

<table>
<thead>
<tr>
<th>Trait</th>
<th>SNP-Set</th>
<th>Chr</th>
<th>PIP(g)</th>
<th>Rank</th>
<th>RSS PIP</th>
<th>RSS Rank</th>
<th>Top SNP</th>
<th>PIP(j)</th>
<th>SuSiE PIP</th>
<th>SuSiE Rank</th>
<th>Ref(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDL</td>
<td>HERPUD1-CETP♣</td>
<td>16</td>
<td>0.999</td>
<td>1*</td>
<td>0.781</td>
<td>5</td>
<td>rs1800775♣</td>
<td>1.000</td>
<td>1.000</td>
<td>1</td>
<td>(212; 208; 127; 95)</td>
</tr>
<tr>
<td></td>
<td>ST18-FAM150A</td>
<td>8</td>
<td>0.999</td>
<td>1*</td>
<td>0.869</td>
<td>3</td>
<td>rs6990075</td>
<td>1.000</td>
<td>0.006</td>
<td>107</td>
<td>(242)</td>
</tr>
<tr>
<td></td>
<td>TCEA3</td>
<td>1</td>
<td>0.989</td>
<td>2</td>
<td>1.22 × 10^{-4}</td>
<td>15056</td>
<td>rs1767141</td>
<td>0.868</td>
<td>0.039</td>
<td>21</td>
<td>(125)</td>
</tr>
<tr>
<td>LDL</td>
<td>CELSR2</td>
<td>1</td>
<td>0.989</td>
<td>1</td>
<td>0.972</td>
<td>2</td>
<td>rs4970834</td>
<td>0.948</td>
<td>1.12 × 10^{-4}</td>
<td>4559</td>
<td>(213; 275; 187)</td>
</tr>
<tr>
<td></td>
<td>BCAM-PVRL2♣</td>
<td>19</td>
<td>0.987</td>
<td>2</td>
<td>0.237</td>
<td>12</td>
<td>rs10402271♣</td>
<td>0.998</td>
<td>0.966</td>
<td>2</td>
<td>(240; 193; 176)</td>
</tr>
<tr>
<td></td>
<td>APOB♣</td>
<td>2</td>
<td>0.976</td>
<td>3</td>
<td>0.167</td>
<td>18</td>
<td>rs693♣</td>
<td>0.999</td>
<td>0.278</td>
<td>6</td>
<td>(240; 153; 202)</td>
</tr>
</tbody>
</table>

Here, SNP-set annotations are based on gene boundaries defined by the NCBI's RefSeq database in the UCSC Genome Browser (195). Unannotated SNPs located within the same genomic region were labeled as being within the “intergenic region” between two genes. These regions are labeled as **Gene1-Gene2** in the table. Posterior inclusion probabilities (PIP) for the input and hidden layer weights are derived by fitting the BANNs model on individual-level data. A SNP-set is considered enriched if it has a PIP(g) ≥ 0.5 (i.e., the “median probability model” threshold (6)). We report the “top” associated SNP within each region and its corresponding PIP(j). We also report the corresponding SNP and SNP-set level results after running SuSiE (262) and RSS (307) on these same traits, respectively. The last column details references and literature sources that have previously suggested some level of association or enrichment between the each genomic region and the traits of interest. See Tables B.18 and B.19 for the complete list of SNP and SNP-set level results. ♣: Multiple SNP-sets were tied for this ranking. ♣: SNPs and SNP-sets replicated in an independent analysis of ten thousand randomly sampled individuals of European ancestry from the UK Biobank (24).


2.4.6 Independent Lipoprotein Study using the UK Biobank

To further validate our results from the Framingham Heart Study, we also independently apply BANNs to analyze HDL and LDL cholesterol traits in ten thousand randomly sampled individuals of European ancestry from the UK Biobank (24). Here, we filter the imputed genotypes from the UK Biobank to keep only the same 394,174 SNPs that were used in the Framingham Heart Study analyses from the previous section. We then apply BANNs, SuSiE, and RSS to the individual-level data and in-sample derived summary statistics using the same (a) 18,364 SNP-set annotations based on the NCBI’s RefSeq database from the UCSC Genome Browser (195) and (b) 35,849 SNP-sets when applying the augmented ±500 kilobase buffer. It is important to note that we restrict this analysis to just ten thousand individuals due to computational considerations for BANNs and SuSiE since they take in individual level data. In Fig. B.31, we show the BANNs variant-level Manhattan plots for the independent UK Biobank cohort with significant SNPs color coded according to their SNP-set annotation. Once again, we use the median probability model threshold to determine statistical significance for all weights in the neural network, and a complete breakdown of the overlap of findings between BANNs, SuSiE, and RSS between the traits can be found in Table B.20. Lastly, Tables B.21 and B.22 give the complete list of all SNP and SNP-set level results in this additional UK Biobank study.

Despite the UK Biobank being a completely independent dataset, we found that BANNs was able to replicate two SNPs and two SNP-sets in HDL and two SNPs and one SNP-set that we observed in the Framingham Heart Study analysis (see specially marked rows in Table 2.2 and Table B.17, as well as the overlap summary given in Table B.20). For example, in HDL, both the variants rs1800775 (PIP = 1.00 versus SuSiE PIP = 1.00) and rs17482753 (PIP = 1.00 versus SuSiE PIP = 0.73) were replicated. BANNs also identified the corresponding intergenic region between the genes HERPUD1 and CETP as being enriched (PIP = 1.00 versus RSS PIP = 1.00). In our analysis of LDL, BANNs replicated two out of the four associated SNPs: rs693 within the APOB gene, and rs10402271 which falls within the intergenic region between genes BCAM and PVRL2.

There were a few scenarios where a given SNP-set was replicated but the leading SNP in that
region differed between the two studies. For instance, while the intergenic region between \textit{LIPG} and \textit{ACAA2} was enriched in both cohorts, the variant rs7240405 was found to be most associated with HDL in the Framingham Heart Study; a different SNP, rs7244811, was identified in the UK Biobank (Fig. 2.4 and Fig. B.31) Similarly, in the analysis with the \(\pm 500\) kilobase buffer for SNP-set annotations, rs4939883 in the intergenic region between \textit{LIPG} and \textit{ACAA2} was found to be significant for HDL in the UK Biobank instead of rs7244811 which was selected in the Framingham Heart Study. These discrepancies at the variant level are likely due to: (i) the sparsity assumption imposed by BANNs, which lead the model to select one of two variants in high LD; and (ii) ancestry differences among individuals from the two studies likely also generate different LD structures in the same genomic region.

As a final step, we took the enriched SNP-sets identified by BANNs in the UK Biobank and used Enrichr (34; 129) to ensure that we were still obtaining trait relevant results (Figs. B.32 and B.33). Indeed, for both HDL and LDL, the most overrepresented categories in dbGaP and the GWAS Catalog (i.e., Bonferroni-correct threshold \(P\)-value < \(1 \times 10^{-5}\) or \(Q\)-value < 0.05) was consistently the trait of interest—followed by other functionally related gene sets such as “Metabolic syndrome” and “Cholesterol levels”. This story remained largely consistent even when augmenting SNP-set annotations with a \(\pm 500\) kilobase buffer (Figs. B.32 and B.33). Overall, the sensible results from performing mapping on the variant-level and enrichment analyses on the SNP-set level in two different independent datasets, only further enhances our confidence about the potential impact of the BANNs framework in GWA studies.

### 2.5 Discussion

Recently, nonlinear approaches have been applied in biomedical genomics for prediction-based tasks, particularly using GWA datasets with the objective of predicting phenotypes (189; 122; 97; 114). However, since the classical idea of variable selection and hypothesis testing is lost within these statistical algorithms, they have not been widely used for association mapping where the goal is to identify significant SNPs or genes underlying complex traits. Here, we present Biologically
Annotated Neural Networks (BANNs): a class of feedforward probabilistic models which overcome this limitation by incorporating partially connected architectures that are guided by predefined SNP-set annotations. This creates an interpretable and integrative framework where the first layer of the neural network encodes SNP-level effects and the neurons within the hidden layer represent the different SNP-set groupings. We frame the BANNs methodology as a Bayesian nonlinear regression model and use sparse prior distributions to perform variable selection on the network weights. By modifying a well established variational inference algorithm, we are able to derive posterior inclusion probabilities (PIPs) which allows researchers to carry out SNP-level mapping and SNP-set enrichment analyses, simultaneously. While we focus on genomic motivations in this study, the concept of partially connected neural networks may extend to any scientific application where annotations can help guide the groupings of variables.

Through extensive simulation studies, we demonstrate the utility of the BANNs framework on individual-level data (Fig. 2.1) and GWA summary statistics (Fig. B.1). Here, we showed that both implementations are consistently competitive with commonly used SNP-level association mapping methods and state-of-the-art SNP-set enrichment methods in a wide range of genetic architectures (Figs. 2.2-2.3, Figs. B.2-B.23, and Tables B.1-B.8). The advantage of our approach was most clear when the broad-sense heritability of the complex traits included pairwise genetic interactions. In two real GWA datasets, we demonstrated the ability of BANNs to prioritize trait relevant SNPs and SNP-sets that have been identified by previous publications and functional validation studies (Fig. 2.4, Figs. B.28-B.30, Tables 2.1-2.2, and Tables B.11-B.19). Lastly, using a third real dataset, we assess the ability of BANNs to statistically replicate a subset of these findings in an independent cohort (Figs. B.31-B.33 and Tables B.21-B.22).

The current implementation of the BANNs framework offers many directions for future development and applications. Perhaps the most obvious limitation is that ill-annotated SNP-sets can bias the interpretation of results and lead to misplaced scientific conclusions (i.e., might cause us to highlight the “wrong” gene (222; 42)). This is a common issue in most enrichment methods (39); however, similar to other hierarchical methods like RSS (307), BANNs is likely to rank SNP-set enrichments that are driven by just a single SNP as less reliable than enrichments driven by multiple
SNPs with nonzero effects. Another current limitation for the BANNs model comes from the fact that it uses a variational inference to estimate its parameters. While the current implementation works reasonably well for large datasets (Tables B.9 and B.10), we showed that our sparse prior assumption combined with the variational expectation-maximization algorithm can lead to slightly miscalibrated PIPs (Figs. B.24), underestimated approximations of the PVE (Figs. B.27 and B.28), and will occasionally miss causal SNPs if they are in high LD with other non-causal SNPs in the dataset. For example, in the application to the Framingham Heart Study, BANNs estimates the PVE for HDL and LDL to be 0.11 and 0.04, respectively. Similarly, in the UK Biobank study, BANNs estimates the PVE for HDL and LDL to be 0.12 and 0.06, respectively. In general, these values are lower than what is typically reported in the literature for these complex phenotypes (PVE ≥ 27% for HDL and PVE ≥ 21% for LDL, respectively) (117). Exploring different prior assumptions and considering other (more scalable) ways to carry out approximate Bayesian inference is something to consider for future work (298). For example, the Bayesian sparse linear mixed modeling (BSLMM) framework (305; 296; 292) extends the traditional spike-and-slab prior and could provide a useful, yet alternative, hierarchical specification for BANNs.

There are several other potential extensions for the BANNs framework. First, in the current study, we only consider a single hidden layer based on the annotations of gene boundaries and intergenic region between genes. One natural direction for future work would be to take more of a deep learning approach by including additional hidden layers to the neural network where genes are grouped based on signaling pathways or other functional ontologies (e.g., transcription factor binding). This would involve integrating information from curated databases such as MSigDB (177; 236) or CADD (201). Second, while BANNs is able to account for nonlinear genetic effects, it cannot be used to directly identify the component (i.e., linear vs. nonlinear) that is driving individual SNP or SNP-set associations. A key part of our future work is learning how to disentangle this information and provide detailed summaries of variant-level and gene-by-gene interaction effects (245). Third, the current BANNs model only takes in genetic information and, in its current form, ignores unobserved environmental covariates (and potential gene-by-environment or G×E interactions) that explain variation in complex traits. In the future, we would like to expand the
framework to also take in covariates as fixed effects in the model. Fourth, we have only focused on analyzing one phenotype at a time in this study. However, many previous studies have extensively shown that modeling multiple phenotypes can often dramatically increase power (206). Therefore, it would be interesting to extend the BANNs framework to take advantage of phenotype correlations to identify pleiotropic epistatic effects. Modeling strategies based on the multivariate linear mixed model (mvLMM) (304) and matrix variate Gaussian process (mvGP) (159) could be helpful here. As a final avenue for future work, we only focused on applying BANNs to quantitative traits. For studies interested in extending this approach to binary traits (i.e., case-control studies), one might be tempted to simply place a sigmoid or logistic link function on the penultimate layer of the neural network. Indeed, this would allow the BANNs framework to be expressed as a (nonlinear) logistic classification model which is an approach that has been well-established in the statistics literature (19; 20; 150). Unfortunately, it is not straightforward to define broad-sense heritability under the traditional logistic regression framework. As one alternative, we could implement a penalized quasi-likelihood approach (238) which has been shown to enable effective heritability estimation and differential analyses using the generalized linear mixed model framework. As a second alternative, the liability threshold model avoids issues by assuming that binary traits can be modeled via continuous latent liability scores (137; 85; 268). Therefore, a potentially effective way to extend BANNs to case-control studies would be to develop a two-step algorithmic procedure where: in the first step, we find the posterior mean of the liability scores be using existing software packages and then, in the second step, treat those empirical liability estimates as observed traits in the neural network. Regardless of the modeling strategy, new algorithms are likely needed to maximize the appropriateness of BANNs for non-continuous phenotypes.

### 2.6 URLs

Biologically annotated neural networks (BANNs) software, [https://github.com/lcrawlab/BANNs](https://github.com/lcrawlab/BANNs);

UK Biobank, [https://www.ukbiobank.ac.uk](https://www.ukbiobank.ac.uk);


2.7 Acknowledgements

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

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

2.9 Author Contributions

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

2.10.1 Annotations

We used the NCBI’s Reference Sequence (RefSeq) database in the UCSC Genome Browser (195) to annotate SNPs with appropriate SNP-sets. In the main text, we define genes with boundaries in two ways: (a) we use the UCSC gene boundary definitions directly, or (b) we augment the gene boundaries by adding SNPs within a ±500 kilobase (kb) buffer to account for possible regulatory elements. Genes with only 1 SNP within their boundary were excluded from either analysis. Unannotated SNPs located within the same genomic region were labeled as being within the “intergenic region” between two genes. Altogether, a total of \( G = 28,644 \) SNP-sets were kept for analysis using the UCSC boundaries and a total of \( G = 35,849 \) SNP-sets were kept for analysis when including the 500kb buffer.

2.10.2 Biologically Annotated Neural Networks

Consider a genome-wide association (GWA) study with \( N \) individuals. We have an \( N \)-dimensional vector of quantitative traits \( y \), an \( N \times J \) matrix of genotypes \( X \), 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 \( \{S_1, \ldots, S_G\} \) (Fig. 2.1a). Let each SNP-set \( g \) represent a known collection of annotated SNPs \( j \in S_g \) with cardinality \( |S_g| \). For example, \( S_g \) may include SNPs within the regulatory region of a gene. The BANNs framework assumes a partially connected Bayesian neural network architecture based on SNP-set annotations to learn the phenotype of interest for each observation in the data (Fig. 2.1b). Formally, we specify this network as a nonlinear regression model (Fig. 2.1c)

\[
y = \sum_{g=1}^{G} h(X_g \theta_g + 1b^{(1)}_g)w_g + 1b^{(2)},
\]

where \( X_g = [x_1, \ldots, x_{|S_g|}] \) is the subset of SNPs annotated for SNP-set \( g \); \( \theta_g = (\theta_1, \ldots, \theta_{|S_g|}) \) are the corresponding inner layer weights; \( h(\bullet) \) denotes the nonlinear activations defined for neurons in the hidden layer; \( w = (w_1, \ldots, w_G) \) are the weights for the \( G \)-predefined SNP-sets in the hidden
layer; \( b^{(1)} = (b^{(1)}_1, \ldots, b^{(1)}_G) \) 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 \((? )\), where \( h(x) = x \) if \( x > 0 \) and \( 0.01x \) otherwise. Note that Eq. (2.4) can be seen as a nonlinear take on classic integrative and structural regression models (265; 25; 28; 291; 307; 121) frequently used in GWA analyses.

A key methodological aspect in the BANNs framework is to treat the weights of the input \((\theta_j)\) and hidden layers \((w_g)\) as random variables. This, in part, enables us to perform interpretable association mapping on both SNPs and SNP-sets, simultaneously. For the weights on the input layer, our goal is to approximate a wide range of possible SNP-level effect size distributions underlying complex traits. To this end, we assume that SNP-level effects follow a \( K \)-mixture of normal distributions \((179; 297 ?; 155)\)

\[
\theta_j \sim \pi_\theta \sum_{k=1}^{K} \eta_{\theta k} N(0, \sigma^2_{\theta k}) + (1 - \pi_\theta)\delta_0, \quad \log(\pi_\theta) \sim \mathcal{U}(-\log(J), \log(1)) \tag{2.5}
\]

where \( \delta_0 \) is a point mass at zero; \( \mathbf{\sigma}^2_{\theta} = (\sigma^2_{\theta 1}, \ldots, \sigma^2_{\theta K}) \) are variance of the \( K \)-nonzero mixture components; \( \eta_\theta = (\eta_{\theta 1}, \ldots, \eta_{\theta K}) \) represents the marginal (unconditional) probability that a randomly selected SNP belongs to the \( k \)-th mixture component such that \( \sum_k \eta_{\theta k} = 1 \); and \( \pi_\theta \) denotes the total proportion of SNPs that have a nonzero effect on the trait of interest. We allow sequential fractions of SNPs \((\eta_{\theta 1}, \ldots, \eta_{\theta K})\) to correspond to distinctly smaller effects \((\sigma^2_{\theta 1} > \cdots > \sigma^2_{\theta K})\) \((? )\). Intuitively, specifying a larger \( K \) allows the neural network to learn general SNP effect size distributions spanning over a diverse class of trait architectures. For results in the main text, we fix \( K = 3 \) for computational reasons. This corresponds to the hypothesis that SNPs can have large, moderate, and small nonzero effects on phenotypic variation \((39)\). We assume a uniform prior on \( \log \pi_\theta \) to coincide with the observation that the number of SNPs in each of these categories can vary greatly depending on how heritability is distributed across the genome \((91; 305)\) (see Supporting Information).
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

$$w_g \sim \pi_w \mathcal{N}(0, \sigma^2_w) + (1 - \pi_w)\delta_0, \quad \log(\pi_w) \sim \mathcal{U}(-\log(G), \log(1))$$ (2.6)

where, in addition to previous notation, the parameter $\pi_w$ 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. (2.4), the magnitude of association for a SNP-set will be directly influenced by the effect size distribution of the SNPs it contains.

We use a variational Bayesian algorithm to estimate all model parameters (Supporting Information). As the BANNs model is trained, the posterior mean for the weights of non-associated SNP and SNP-sets will trend towards zero as the neural network attempts to identify a subset of neurons that are associated with the phenotype. We use posterior inclusion probabilities (PIPs) as a general summaries of evidence for SNPs and SNP-sets being associated with phenotypic variation. Here, we respectively define

$$\text{PIP}(j) \equiv \Pr[\theta_j \neq 0 | y, X], \quad \text{PIP}(g) \equiv \Pr[w_g \neq 0 | y, X, \theta_g]$$ (2.7)

where, again for the latter, the enrichment of SNP-sets is conditioned on the association of individual SNPs. Overall, the Bayesian formulation in the BANNs framework enables network sparsity to be targeted for GWA applications through contextually motivated sparse shrinkage prior distributions in Eqs. (2.5)-(2.6). Moreover, posterior inference on PIP$(j)$ and PIP$(g)$ detail the degree to which nonzero weights occur.

### 2.10.3 Posterior Computation with Variational Inference

We combine the likelihood in Eq. (2.4) and the prior distributions in Eqs. (2.5)-(2.7) to perform Bayesian inference. With the size of high-throughput GWA datasets, it is less feasible to implement
traditional Markov Chain Monte Carlo (MCMC) algorithms due to the large dimensionality of the parameter space. For model fitting, we modify a previously established variational expectation-maximization (EM) algorithm (27; 30) for integrative neural network parameter estimation. The overall goal of variational inference is to approximate the true posterior distribution for network parameters with a "best match" distribution from an approximating family (17). 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" (260). The algorithm then follows three general steps. First, we assign exchangeable uniform hyper-priors over a grid of values on the log-scale for $\pi_\theta$ and $\pi_w$ (27). 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 ($\theta, w$) and their corresponding inclusion probabilities by marginalizing over the different hyper-parameter combinations. This final step can be viewed as an analogy to Bayesian model averaging where marginal distributions are estimated via a weighted average of conditional distributions multiplied by importance sampling weights (100).

Throughout the model fitting procedure, we assess two different lower bounds for the input and hidden layers to check convergence of the algorithm. The first lower bound is maximized with respect to the SNP-level effects on the observed trait of interest; while, the second lower bound focuses on the SNP-set level enrichments. The software code first iterates over the "inner" lower bound until convergence and then uses those weights to compute the hidden neurons and maximize the "outer" lower bound. Detailed steps in the variational EM algorithm, explicit co-ordinate ascent updates for network parameters, and pseudocode are given in Supporting Information.

Parameters in the variational EM algorithm are initialized by taking a random draws from their assumed prior distributions. Iterations in the algorithm are terminated when either one of two stopping criteria are met: (i) the difference between the lower bound of two consecutive updates are within some small range (specified by argument $\epsilon$), or (ii) a maximum number of iterations is reached. For the simulations and real data analyses ran in this paper, we set $\epsilon = 1 \times 10^{-4}$ for the
first criterion and used a maximum of 10,000 iterations for the second.

### 2.10.4 Extensions to Summary Statistics

The BANNs framework also models summary statistics in the event that individual-level genotype and phenotype data are not accessible. Here, the software takes alternative inputs: GWA marginal effect size estimates $\hat{\theta}$ as the response variable, and an empirical linkage disequilibrium (LD) matrix $R$ as the design matrix. In the main text, we refer to this version of the method as the BANN-SS model. We assume that GWA summary statistics are derived from the following generative linear model for complex traits

$$ y = X\theta + e, \quad e \sim N(0, \tau^2 I) \quad (2.8) $$

where $e$ is a normally distributed error term with mean zero and scaled variance $\tau^2$, and $I$ is an $N \times N$ identity matrix. For every $j$-th SNP, the ordinary least squares (OLS) estimates are based on the generative model $\hat{\theta}_j = (x_j^T x_j)^{-1} x_j^T y$, where $x_j$ is the $j$-th column of the individual-level genotype matrix $X$ and $\hat{\theta}_j$ is the $j$-th entry of the vector $\hat{\theta}$. In practice, the LD matrix $R$ can be empirically estimated directly from the in-sample GWA study data or from external data (e.g., using an LD map from a population with genomic ancestry similar to individuals in the original study). Note that all results presented in the main text are based on estimating $R$ with the in-sample genotype data. The BANN-SS model treats the observed OLS estimates and LD matrix as "proxies" for the unobserved phenotype and genotypes, respectively. Specifically, for large sample size $N$, we consider the asymptotic relationship between the expectation of the observed GWA effect size estimates $\hat{\theta}$ and the true coefficient values $\theta$ is (103; 104; 39)

$$ \mathbb{E}[\hat{\theta}_j] = \sum_{j'=1}^J r(x_j, x_{j'}) \theta_{j'} \quad (2.9) $$

where $r(x_j, x_{j'})$ denotes the correlation coefficient between SNPs $x_j$ and $x_{j'}$. The above resembles a high-dimensional regression model with the OLS effect sizes $\hat{\theta}$ as the response variables, the LD matrix $R$ as the design matrix, and the true coefficients $\theta$ being the SNP-level effects that generated
the phenotype. Note that this observation is also utilized by other GWA summary-level statistical methods (e.g., CAVIAR (104) and RSS (306; 307)). With this relationship in mind, the BANN-SS framework implements the following sparse nonlinear regression for inferring multi-scale genomic effects from summary statistics (Fig. B.1)

$$\hat{\theta} = \sum_{g=1}^{G} h(R_g \theta_g + 1b^{(1)})w_g + 1b^{(2)},$$

(2.10)

where, in addition to previous notation, $R_g$ is the subset of the LD matrix involving all SNPs annotated for the $g$-th SNP-set. Using the rewritten joint likelihood in Eq. (2.10), posterior Bayesian inference for the parameters in the BANN-SS model directly mirrors the procedure used when we have access to individual-level data (i.e., as described previously in Eqs. (2.5)-(2.7) and given in detail in the Supporting Information). Again, we use measurements $\text{PIP}(j)$ and $\text{PIP}(g)$ to summarize whether the true SNP-level effects and aggregated effects on the SNP-set level are statistically associated with the trait of interest.

2.10.5 Simulation Studies

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

$$y = \sum_{c \in C} x_c \theta_c + W \varphi + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I),$$

(2.11)

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

### 2.10.6 Data and Software Availability

Source code (with versions in both \texttt{R} and \texttt{Python 3}) and tutorials for implementing biologically annotated neural networks (BANNs) is publicly available online at \url{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 (24) (\url{https://www.ukbiobank.ac.uk}) was made available under Application Number 22419. The FHS genotype and phenotype data is available in dbGaP (228) (\url{https://www.ncbi.nlm.nih.gov/gap}) with accession number phs000007.
Chapter 3

Uncertainty Quantification in Variable Selection for Genetic Fine-Mapping using Bayesian Neural Networks

In this paper, we propose a new approach for variable selection using a collection of Bayesian neural networks with a focus on quantifying uncertainty over which variables are selected. Motivated by fine-mapping applications in statistical genetics, we refer to our framework as an “ensemble of single-effect neural networks” (ESNN) which generalizes the “sum of single-effects” regression framework by both accounting for nonlinear structure in genotypic data (e.g., dominance effects) and having the capability to model discrete phenotypes (e.g., case-control studies). Through extensive simulations, we demonstrate our method’s ability to produce calibrated posterior summaries such as credible sets and posterior inclusion probabilities, particularly for traits with genetic architectures that have significant proportions of non-additive variation driven by correlated variants. Lastly, we use real
data to demonstrate that the ESNN framework improves upon the state-of-the-art for identifying true effect variables underlying various complex traits.

3.1 Introduction

Variable selection is a fundamental problem in high-dimensional statistical learning that arises in a wide range of application domains (80; 70; 29; 287). An important benefit of incorporating sparsity when building a predictive model is that it provides interpretations on which input variables are most important in explaining variation across the output variables. Such a property is particularly desirable when the end goal of an application also includes scientific discovery. For example, the goal of many genome-wide association (GWA) studies is not just to predict the disease status or phenotypic risk of a patient but also to identify the (subsets of) single nucleotide polymorphisms (SNPs) that are statistically associated with the genetic architecture of the disease (165; 163). This can further help with downstream clinical applications such as drug development.

While many methods for variable selection have been developed in the literature (80; 70; 29; 287; 308; 243), some significant challenges still remain. One important challenge is assessing the uncertainty in which variables should be selected when they are highly correlated (263; 29). As an extreme case, imagine there are two variables that are completely collinear. In this context, it becomes statistically impossible to distinguish them, and many traditional regularization and shrinkage methods will arbitrarily select one SNP as being associated with the trait of interest and disregard the other (263). While such a strategy suffices if the goal is to build a predictive model, it becomes limiting for scientific discovery because the conclusions rely on selecting the correct subset of genetic variants for downstream investigation. Recently, Wang et al. (263) introduced the “sum of single-effects” model called SuSiE to address these issues. More specifically, SuSiE assesses the uncertainty of variables by providing “credible sets” which, in the case of our extreme example, effectively summarize that “either SNP 1 or 2 are relevant but we are unsure as to which one”. SuSiE uses an iterative Bayesian stepwise selection (IBSS) procedure where it will iteratively regresses out effect variables and feeds the corresponding residuals to the next iteration for training.
The main limitation of SuSiE is that it is a linear model and therefore does not capture non-linear effects in data. In GWA studies, it is well known that the genetic architecture of complex traits can be driven by phenomena such as dominance and epistasis (174; 50; 200; 146). Indeed, machine learning models are most powered in settings when large sets of training data are available and often exhibit greater predictive accuracy than linear models in applications driven by non-additive variation. In this paper, we introduce the “ensemble of single-effect neural networks” (ESNN) framework which overcomes the limitations of SuSiE while preserving the ability to assess uncertainty for variable selection. We demonstrate our approach in a simulation study and on two real GWA datasets.

3.2 The Sum of Single-Effects Regression Model

In this section, we provide background on single-effects regression (SER) and state a rigorous definition of credible sets for variable selection. The original SER model (216; 191) assumes that exactly one of \( J \) input variables has a non-zero coefficient. More specifically,

\[
\begin{align*}
\mathbf{y} &= \mathbf{Xb} + \mathbf{e}, \quad \mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2_{\mathbf{e}} \mathbf{I}) \\
\mathbf{b} &= b\mathbf{\gamma}, \quad b \sim \mathcal{N}(0, \sigma^2_b), \quad \mathbf{\gamma} \sim \text{Mult}(1, \pi)
\end{align*}
\]

where \( \mathbf{y} \) is an \( N \)-dimensional response vector (e.g., continuous phenotypes); \( \mathbf{X} \) is an \( N \times J \) design matrix (e.g., genotypes); \( \mathbf{e} \) is an \( N \)-dimensional error term; \( \mathbf{b} \) is a \( J \)-dimensional vector of regression coefficients; \( \mathbf{\gamma} \) is a binary indicator that determines which regression coefficient is to be non-zero; and \( \text{Mult}(m, \pi) \) denotes the multinomial distribution with \( m \) samples drawn with class probability distribution \( \pi \). For simplicity, we will consider a uniform prior such that \( \pi = (1/J, \ldots, 1/J) \). Note that \( m \) is set to equal to one so that the coefficient vector \( \mathbf{b} \) has exactly one non-zero entry for modeling the single-effect. To estimate the statistical association of each variable, one would fit \( J \)-univariate models corresponding to regressing each \( j \)-th column \( \mathbf{x}_j \) of \( \mathbf{X} \) onto the response \( \mathbf{y} \) and computing posterior inclusion probabilities defined as \( \text{PIP}_j \equiv \Pr[\hat{b}_j \neq 0 | \mathbf{y}, \mathbf{X}] \).

In the context of statistical genetics, the original SER model only assumes one causal SNP.
However, we know that many real-world applications, it is desired to have a method that flexibly allows for many variants to have an effect on trait architecture (29). The SuSiE framework is based on an extension of summing over L-multiple SER models (263). Here, the main idea is to construct an overall effect vector $b$ from multiple single-effect coefficients $b_1, \ldots, b_L$ via the following

$$y = Xb + e, \quad e \sim \mathcal{N}(0, \sigma^2_y I),$$

$$b = \sum_{l=1}^L b^{(l)}, \quad b^{(l)} = b^{(l)} \gamma^{(l)}, \quad b^{(l)} \sim \mathcal{N}(0, \sigma^2_l), \quad \gamma^{(l)} \sim \text{Mult}(1, \pi).$$

In practice, SuSiE uses an iterative Bayesian stepwise selection (IBSS) algorithm (i.e., coordinate ascent variational inference) to estimate the model parameters. More specifically, at each iteration, it fits the SER model for $b^{(l)}$ using the residuals from the model $y - \sum_{l' \neq l} Xb^{(l')}$.

At the end of training, the SuSiE model provides $L$ estimated coefficient vectors $\hat{b}$ and $L$ corresponding PIP vectors $\alpha^{(l)} = \{\Pr[b^{(l)}_1 \neq 0 | y, X], \ldots, \Pr[b^{(l)}_j \neq 0 | y, X]\}$. Computation of a final inclusion probability assumes that effects are independent across the $L$ different models and is computed as

$$\text{PIP}_j \equiv \Pr[\hat{b}_j \neq 0 | y, X] \approx 1 - \prod_{l=1}^L \left(1 - \alpha^{(l)}_j\right).$$

A key component of SuSiE is that it uses these PIPs to naturally construct credible sets. Effectively, a level $\rho$ credible set $S(\alpha, \rho)$ can be estimated by simply sorting variables in descending order and then including variables into the set until their cumulative probability exceeds $\rho$ (263). Below we give the rigorous definition for credible sets.

**Definition 1 (Wang et al. (2020) (263))** In the context of a multiple-regression model, a level $\rho$ credible set is defined to be a subset of variables that has probability $\rho$ or greater of containing at least one effect variable (i.e., a variable with non-zero regression coefficient). Equivalently, the probability that all variables in the credible set have zero regression coefficients is $1 - \rho$ or less.

The definition above yields a metric for assessing the uncertainty when conducting variable selection. A credible set will determine if a subset of collinear variables have effects on the response.
even when we are unclear as to which specific ones. This differs from the results produced by
the conventional regularization and shrinkage methods (29; 243; 308) where the effect sizes for an
arbitrarily selected subset of correlated variables will be penalized while the others are retained.

3.3 The Ensemble of Single-Effect Neural Networks

In this section, we detail the full specification of our proposed nonlinear framework for variable
selection. While there exist many nonlinear models, neural networks are well known to have the
ability to approximate complex systems (143; 7). For simplicity, we will focus on multi-layer per-
ceptrons throughout this paper; however, we also want to emphasize that the theoretical concepts
we describe can also be applied broadly to other architectures (e.g., convolutional neural networks).
Formally, we specify a $K$-layer probabilistic neural network as a generalized nonlinear model

$$
g(\mu) = f = Z_K \Theta_K + \epsilon_K, \ldots, \quad z_k = h(Z_{k-1} \Theta_{k-1} + \epsilon_{k-1}), \ldots, \quad z_1 = h(XW + \epsilon_0)
$$

(3.4)

where, in expectation, the response variable is related to the input data by $E[y | X] = \mu$; $f$ is an
$N$-dimensional latent vector to be learned; $g(\bullet)$ denotes a general cumulative link function which,
for example, is set to be the identity if $y$ is continuous or the logit if $y$ is binary; $Z_k$ denotes the
matrix of nonlinear neurons from the $k$-th hidden layer with corresponding weight matrix $\Theta_k$; $\epsilon_k$
are deterministic biases that are produced during the network training phase for the $k$-th hidden
layer; $h(\bullet)$ is a nonlinear activation function (e.g., ReLU or tanh); and $W$ is a matrix of weights
for the input layer.

Similar to the SER model, the key design that leads to our ability to model single-effect is
through the prior we place on the input layer weights in $W$. Let $H_k$ represent the number of
neurons in the $k$-th hidden layer such that $W$ is $J \times H_1$ dimensions (i.e., the number of input
variables by the number of neurons in the first hidden layer). Next, let $w_{j\bullet}$ denote the $j$-th row of
the weight matrix $\mathbf{W}$. We place a grouped “single-effect” shrinkage prior on the input weights

$$
\mathbf{W} = \mathbf{A} \circ \mathbf{\Gamma}, \quad \mathbf{a} \sim \mathcal{N}(\mathbf{0}, \sigma^2_1 \mathbf{I}), \quad \gamma \sim \text{Mult}(1, \pi)
$$

(3.5)

where $\mathbf{\Gamma}$ is a matrix that is $H_1$ copies of the binary vector $\gamma$, $\mathbf{a}$ is an $H_1$-dimensional row-vector of continuous weights in $\mathbf{A} = [\mathbf{a}_1 \ldots, \mathbf{a}_j \ldots]$, and $\circ$ denotes the Hadamard product between two parameters. Note that this shrinkage prior mimics the sparse assumption of previous neural network architectures in the literature (72; 82), except that the binary indicator variable $\gamma$ is assumed to be multinomial with one trial. Hence, since the $j$-th row of $\mathbf{W}$ contains the weights connected to the $j$-th column in $\mathbf{X}$, when only $\gamma_j = 1$, the rest of the input variables are excluded from the model (see proof-of-concept example in Fig. S1 in the Appendix). Together, we refer to the model above as a “single-effect neural network” (SNN). The SNN resembles the SER model in that it assumes that only one input variable has an effect on the response and, thus, posterior summaries of $\gamma$ can be similarly used to compute credible sets.

We now extend the SNN to incorporate multiple effect variables. Analogous to the SuSiE framework, we now consider training on the response variable to be based on an ensemble of single-effect neural networks (ESNN). Probabilistically, the ESNN maybe specified as a summation of $L$-latent nonlinear models of the form

$$
\mathbf{f}^{(l)} = \mathbf{Z}^{(l)} \mathbf{\Theta}^{(l)} + \epsilon^{(l)}_K, \quad \ldots, \quad \mathbf{z}^{(l)}_k = h\left(\mathbf{Z}^{(l)}_k \mathbf{\Theta}^{(l)}_{k-1} + \epsilon^{(l)}_{k-1}\right), \quad \ldots, \quad \mathbf{z}^{(l)}_1 = h\left(\mathbf{X} \mathbf{W}^{(l)} + \epsilon^{(l)}_0\right)
$$

(3.6)

where, in expectation, the response variable is now related to the input data as $\mathbb{E}[y | \mathbf{X}] = g(\sum_l \mathbf{f}^{(l)})$ and the sparse prior for the weights of the network are now specified as the following

$$
\mathbf{W}^{(l)} = \mathbf{A}^{(l)} \circ \mathbf{\Gamma}^{(l)}, \quad \mathbf{a}^{(l)} \sim \mathcal{N}(\mathbf{0}, \sigma^2_1 \mathbf{I}), \quad \gamma^{(l)} \sim \text{Mult}(1, \pi).
$$

(3.7)

Notice that at the end of training, each $l$-th neural network will also yield an estimated set of input layer weights $\hat{\mathbf{W}}$ and a corresponding set of inclusion probabilities $\mathbf{\alpha}^{(l)} = \{\Pr|\hat{\mathbf{w}}^{(l)}_{ij} \neq 0\}$. 

\[0 \mid y, X, \ldots, \Pr[w_{j*}^{(l)} \neq 0 \mid y, X]\] which each assess whether all weights connected to the \(j\)-th input node are equal to zero. Then, given these \(L\) posterior summaries, we can compute credible sets \(S(\alpha, \rho)\) in the same way as SuSiE by defining the overall posterior inclusion probabilities as

\[
\text{PIP}_j \equiv \Pr[\hat{w}_{j*} \neq 0 \mid y, X] \approx 1 - \prod_{l=1}^{L} \left(1 - \alpha_j^{(l)}\right)
\]

which we use to determine variable significance.

### 3.4 Posterior Inference via Variational Bayes

As the size of many high-throughput genome-wide sequencing studies continue to grow, both in the number of individuals and the number of genetic variants, it has become less feasible to implement traditional Markov Chain Monte Carlo (MCMC) algorithms for inference. To this end, we use variational inference to approximate the posterior distribution of the weights and hyper-parameters within the ESNN framework. We take the hierarchical model specified in Eqs. (3.6)-(3.7) and replace the intractable true posterior distribution over the parameters \(p(W_{1:L}, \Gamma_{1:L} \mid D)\) with an approximating family of distributions \(q(W_{1:L}, \Gamma_{1:L}; \phi_{1:L})\)—where we use shorthand \(1 : L = 1, \ldots, L\) to represent the \(L\) models in the ensemble, \(\phi_{1:L}\) represent the collection of free parameters in the approximations, and \(D\) is used to denote the observed data and all relevant hyper-parameters. The basic idea behind the variational inference is to iteratively adjust the free parameters such that they minimize the the difference between the two distributions, which amounts to maximizing the so-called evidence lower bound (ELBO)

\[
\mathcal{L}(\phi_{1:L}) = \mathbb{E}_q \left[\log p(y \mid W_{1:L}, \Gamma_{1:L}, D)\right] + \text{KL}(q(W_{1:L}, \Gamma_{1:L}; \phi_{1:L}) \parallel p(W_{1:L}, \Gamma_{1:L}))
\]

Here, the first term is the expectation of the log-likelihood taken with respect to the variational distribution, and the second term is the Kullback-Leibler divergence which measures the similarity between two distributions. We then use a stochastic gradient descent based method to train models under the ESNN framework. In this work, we choose the variational distributions to factorize across
$L$ models and for each model we have the following proposals

$$q(W^{(l)}, \Gamma^{(l)}) = q(A^{(l)}q(\Gamma^{(l)})), \quad q(a^{(l)}) = \mathcal{N}(m, \tau^2_1 I), \quad q(\gamma^{(l)}) = \text{Mult}(1, \kappa),$$  \hspace{1cm} (3.10)

Based on these choices, the gradients of the KL term are available in closed form, while the expectation of the log-likelihood is evaluated using Monte Carlo samples and the local re-parameterization trick (see Appendix for theoretical details and corresponding pseudocode). In a regression task with continuous responses, the log-likelihood term is chosen to be Gaussian and maximizing the lower bound corresponds to minimizing mean square error. In classification tasks for case-control studies, the log-likelihood term is taken to be a binomial distribution which corresponds to minimizing the cross-entropy loss. Since we use gradient descent based method for optimization, the ESNN can be applied for both types of data analyses.

### 3.4.1 Iterative Bayesian Stepwise Selection

Similar to the SuSiE framework, the ESNN model also uses an iterative Bayesian stepwise selection (IBSS) procedure where it trains $L$ models by first fitting one model with a coordinate ascent algorithm and then regressing out that model to compute residuals for training next model. By doing so, we can generate credible sets (263). It is worth noting that, when the model is uncertain about which variables to choose (e.g., when there are no significant effect variables), $\alpha$ will become diffuse such that $\mathcal{S}(\alpha, \rho)$ will contain many variables that are not correlated. Under these scenarios, it makes sense to ignore those sets. Previous work have outlined the concept of “purity” as the smallest absolute correlation between all pairs of variables within a credible set which can be used as a criteria for filtering out nonsensical results (263). This same strategy is not particularly useful on its own for the ESNN framework. An intuitive explanation for this is because since the optimizing objective for neural networks is non-convex, training algorithms can get stuck in local optima where the estimated variational parameters $\phi$ are not optimal. In the scenario where the model is unable to find correct effect variable, regressing out $\phi$ will only introduce noise during training. Therefore, we take an extra approach where we also check to ensure that a trained model is informative before
computing the residuals. One simple way to do this is by monitoring whether the likelihood is larger with the \( l \)-th model trained versus it be excluded from consideration. More specifically, the criteria to include the \( l \)-th model can be expressed via the (approximate) likelihood ratio

\[
\lambda^{(l)} = \frac{L(\phi_1, \ldots, \phi_{l-1}, \phi_l, \phi_{l+1}, \ldots, \phi_L)}{L(\phi_1, \ldots, \phi_{l-1}, \phi_{l+1}, \ldots, \phi_L)}
\]

(3.11)

where we keep models that satisfy \( \lambda^{(l)} > 1 \). Note that we only regress out variables on continuous data as this is the scenario where it is meaningful to compute the residuals. For the binary classification case, we simply fix the trained models and add up the logits if the criteria is satisfied.

### 3.5 Results

In this section, we first examine the utility of the ESNN model in simulations motivated by fine-mapping applications for continuous and binary traits in genome-wide association studies. We also apply our method to real world GWA datasets from the Wellcome Trust Case Control Consortium (WTCCC) and the Wellcome Trust Centre of Human Genetics.

#### 3.5.1 Simulations with Continuous Phenotypes

In order to evaluate the performance of our model on continuous traits, we simulate data using real genotypes from chromosome 1 of \( N = 5000 \) randomly sampled individuals of self-identified European ancestry in the UK Biobank. After quality control (\( ? \)), this dataset had 36,518 SNPs. To simulate fine-mapping applications, we used the NCBI’s Reference Sequence (RefSeq) database in the UCSC Genome Browser (196) to annotate SNPs to genes. Here, we randomly sampled 200 genes on this chromosome where the annotations included both SNPs located within the gene boundary and SNPs that fall within a \( \pm 500 \) kilobase (kb) window of the boundary to also include regulatory elements.

In this study, each gene is considered to be its own dataset with its own complex correlation structure (see Fig. S3) and unique number of SNPs (ranging from \( J = 50 \) to 417 variants) encoded
as \( \{0, 1, 2\} \) copies of a reference allele where 0 and 2 represent “homozygotes” and 1 represents “heterozygotes”. For each dataset, we assign 5 effect SNPs and use the following generative model

\[
y = \sum_{j \in C} x_j \beta_j \mathbb{I}(x_j = 0 \text{ or } 2) + \sum_{j \in C} x_j \omega_j \mathbb{I}(x_j = 1) + e, \quad e \sim \mathcal{N}(0, \sigma_y^2 \mathbf{I})
\] (3.12)

where \( C \) represents the set of causal SNPs and \( \mathbb{I}(\cdot) \) is an indicator function. Here, \( \beta \) and \( \omega \) are different effect sizes for heterozygotes and homozygotes, respectively. Both variables are randomly sampled from standard normal distribution and rescaled according to their frequencies. The error term \( e \) is also assumed to be normally distributed and is rescaled during the simulation such that the causal SNPs explain a certain proportion of the variance in the synthetic trait (i.e., the narrow-sense heritability, \( h^2 \)). We consider different scenarios where \( h^2 = \{0.05, 0.1, 0.4\} \).

We compare our method with SuSiE (fit under its default parameter settings). We set \( L = 10 \) for both approaches. For the ESNN we used a simple sparse architecture with 5 hidden neurons and tanh activation functions. Here, we set the maximum number of epochs to be 30: the hyper-parameter \( \pi \) for the indicator \( \gamma \) is chosen from a uniform distribution; we fix \( \sigma^2_y = 1 \) for all \( L \) models; and, during training, we take 100 Monte Carlo samples to evaluate the log-likelihood. Finally, we used an Adam optimizer with a learning rate of 0.005 and a decay rate of 0.995 after every epoch, and we used an early stopping rule if the likelihood on validation data stopped increasing (based on 85/15 training/validation splits).

To assess the performance, we consider three different metrics. First, we begin by assessing the probability that each method create a credible set containing at least 1 effect SNP (first row Fig. 3.1(a)). Ideally, a 95% level credible set should have at least 95% coverage. When heritability is high (e.g., \( h^2 = 0.4 \)), signals are easier to detect, and both ESNN and SuSiE achieve the appropriate coverage. However, for lowly heritable traits (e.g., \( h^2 = 0.1 \) and 0.05), the coverage of SuSiE drops while the coverage of ESNN remains. The second metric we check is the average number of effect variables included in all credible sets (first row Fig. 3.1(b)). In practice, each method can report multiple credible sets. Therefore, this metric essentially helps evaluate the empirical power of ESNN and SuSiE. In these simulations, our method is consistently better regardless of trait heritability.
For the final metric, we assess the ability of ESNN and SuSiE to accurately prioritize causal variants according to the PIPs that each method provides. Here, we use receiver operating characteristic (ROC) and precision-recall curves to compare their ability to rank true positives over false positives (first row of Figs. 3.2 and S5). As $h^2$ decreases, accuracy of the PIPs for both method decrease but our method is relatively better powered for all scenarios. Importantly, the PIPs from ESNN and SuSiE are calibrated similarly (Fig. S4) (263).

Figure 3.1. Panel (a) shows comparisons of coverage for ESNN and SuSiE in simulation studies under different levels of heritability. Panel (b) shows the average number of effect variables included in all credible sets for each simulation replicate. Results are based on 200 data replicates.

### 3.5.2 Simulations with Binary Phenotypes

We now assess the performance of ESNN on binary traits (e.g., case-control studies). We consider two generative models for the class labels: (1) logistic regression and (2) a liability threshold (LT) model (138; 86; 69). In the former, we simply use the genotypes from chromosome 1 of the $N =$
5000 randomly sampled individuals from the UK Biobank to assume that

$$\mathbf{y} \sim \text{Bern}(p), \quad \log \left( \frac{p}{1-p} \right) = \sum_{j \in C} \mathbf{x}_j \beta_j \mathbb{I}(\mathbf{x}_j = 0 \text{ or } 2) + \sum_{j \in C} \mathbf{x}_j \omega_j \mathbb{I}(\mathbf{x}_j = 1)$$

(3.13)

where, in addition to previous notation, the binary traits follow a Bernoulli distribution with probability $p$. In the latter simulation model, we take into account disease prevalence and ascertainment bias which can occur in case-control studies. Here, we adopt the LT model which assumes an latent liability $l_i \sim \mathcal{N}(0, 1)$ for each observation. With some known prevalence $k$, one can determine a threshold $t = \Phi^{-1}(k)$ using the quantile function of normal distribution such that an individual is a case $y_i = 1$ if $l_i > t$. To simulate data under the LT model, we first generate one million individuals each with $J = 200$ SNPs (with minor allele frequency uniformly sampled between 0.05 and 0.5). Next, we select 5 causal SNPs and generate continuous liabilities with a controlled heritability $h^2 = \{0.05, 0.1, 0.4\}$ using a model similar to Eq. (3.12). Then we consider a prevalence $k \in \{50\%, 10\%, 1\%\}$ and define cases-controls labels for each of the million individuals. Finally, we subsample 2500 cases and 2500 controls for the analysis.

The SuSiE framework was originally designed for continuous traits, so we consider two adaptations of the model for the binary data. In the first, we simply treat the class labels as continuous and run the model as is. In the second, which we refer to as LT-SuSiE, we use an MCMC to estimate continuous liability scores as phenotypes (71; 69; 54). Here, we use all the same parameter setting as in the regression simulation study, except that we set the learning rate for ESNN to be 0.01.

Similarly, we compared powers of two methods using coverage (Fig. 3.1(a)), the number of effect variables included in all credible sets per dataset (Fig. 3.1(b)), ROC curves (Fig. 3.2), and precision-recall curves (Fig. S5). Overall, performances follow a similar trend to the regression simulations such that ESNN consistently outperforms SuSiE. When disease prevalence is very low (e.g., $k = 1\%$), cases are assumed to come from “tail” of the distribution. In this scenario, statistical models are generally better powered (48). As the prevalence $k$ becomes greater, such that the liability threshold moves from the tails to the center of the distribution, it will become harder for a classifier
to distinguish cases from controls. This also results in lower power for variable selection. Notably, even in these cases, our method remains robust.

**Figure 3.2.** Receiver Operating Characteristic (ROC) curves for simulation studies of different scenarios. Results are based on 200 data replicates.

### 3.5.3 Fine-Mapping in Heterogenous Stock of Mice

We apply ESNN and SuSiE to two continuous traits: high-density and low-density lipoprotein (HDL and LDL, respectively) in a heterogeneous stock of mice dataset from the Wellcome Trust Centre for Human Genetics (250). This dataset contains $J = 10,346$ SNPs with $N = 1594$ samples for HDL and $N = 1637$ samples for LDL. To run both methods, we simply partition the whole genome into 21 windows where each window contains 500 SNPs. By doing so, we fine-map SNPs in annotated genes as well as SNPs in intergenic regions. We used the same hyper-parameter settings as in the regression simulations for both methods.
For HDL and LDL, ESNN finds 41 and 19 credible sets while SuSiE finds 62 and 26 credible sets, respectively. Our method finding less credible sets is potentially could due to the criteria that we only include an SNN model if it increases the likelihood. This criteria demonstrated to ensure that a credible set generated by ESNN would have high coverage in simulations (Fig. 3.1). There were 12 SNPs that were included in the credible sets of both methods for HDL and 5 for LDL. This potentially means that these SNPs contributed additive effects to the phenotypic variation. SNPs that are only identified by ESNN probably contribute nonlinear effects (e.g., dominance). We highlighted one region for each trait in Fig. S6. One SNP found by both methods, *rs3090325* in LDL (Fig. S6(a)), can be mapped to the *Smarca2* gene which has been found to be associated with cholesterol regulation (172). In HDL, SNP *gnf04.147.942* can be mapped to the *Panc1* gene which regulates pancreatic activity and has been shown to be linked with HDL (164). Furthermore, SNPs such as *rs13483562* (which is only found by ESNN in the LDL analysis), can be mapped to the *Aldh1a7* gene which also has been demonstrated to affect related traits such as lipid, cholesterol level, and obesity in mice (293; 136).

### 3.5.4 Fine-Mapping in the WTCCC 1 Study

We next apply ESNN and SuSiE to two binary traits: type 1 diabetes (T1D) and type 2 diabetes (T2D) from the Wellcome Trust Case Control Consortium (WTCCC) 1 study (44). This dataset has $N = 1963$ cases and $N = 2938$ controls for T1D, $N = 1924$ cases and $N = 2938$ controls for T2D, along with $J = 458,868$ genotyped SNPs for each individual. Similarly, we run ESNN and SuSiE with a window size of 500 SNPs and used the same model settings as in the binary simulations.

ESNN identifies 32 and 19 credible sets for T1D and T2D, respectively, whereas SuSiE finds and 67 and 30 sets for each trait. There are 5 SNPs that are found by both methods for T1D, but none for T2D. This is likely due to the fact that SuSiE was not originally developed for binary traits and also due to the potential role of nonlinear genetic architecture. We highlight two interesting results in Fig. 3.3 where we plot the PIPs of SNPs computed by ESNN and SuSiE. In panel (a), we show a window near the *HLA* region on chromosome 6 which has been well studied in the literature.
and found to be associated with the T1D (106; 184; 67; 185). One of the two SNPs found only by ESNN, rs3129051, is located upstream (within 50kb) of the HLA-G gene which is a well-known gene that is related to T1D. The other SNP, rs16894900, is located between MAS1L (within 50kb downstream) and UBD (within 50kb upstream), both of which have been shown to be related to T1D (185). In panel (b), we highlight the region around NOS1AP on chromosome 1. This gene has been found to be linked with T2D in several studies (105; 41; 199). Our method identified 2 SNPs in this region, but SuSiE reports none. It has been suggested that this region may not play a dominant role in susceptibility to T2D, but a minor effect may exist (105). Similar to SuSiE, these conclusion were previously made use linear models. We hypothesis this region may contribute to T2D nonlinearly and thus the traditional hypothesis testing methods will have missed this signal.
3.6 Discussion

In this paper, we present the ensemble of single-effect neural network (ESNN) which generalizes the sum of single-effects regression framework by accounting for nonlinear genetic architecture and extending to noncontinuous phenotypes. The ESNN approach provides posterior inclusion probabilities and credible sets that can guide variable selection. While we focus on genetic fine-mapping, this method is also applicable to other fields especially when data are correlated and sparse. We provide a variational algorithm with several relaxation techniques that enables scalable inference. We show that ESNN can effectively increase power for variable selection using simulations. We applied ESNN to two real world genetic datasets and demonstrated its ability to make discoveries that are biologically meaningful. There are a few limitations to the current ESNN framework. Similar to most deep learning models, our method requires large sample sizes for training and requires hyper-parameter fine-tuning. For high-dimensional settings, we currently run the method by splitting the whole dataset into small windows so that the training algorithm can quickly converge. However, this may ignore some long-range interactions. Therefore, a focus for future work will be extending the current model with more complex network architectures.

3.7 Software Availability

Source code and tutorials for implementing the “ensemble of single-effect neural networks” (ESNN) framework are publicly available online at https://github.com/ramachandran-lab/ESNN.

3.8 Data Availability

The heterogenous stock of mice dataset from the Wellcome Trust Centre for Human Genetics can be found at http://mtweb.cs.ucl.ac.uk/mus/www/mouse/index.shtml. Data from the UK Biobank Resource (https://www.ukbiobank.ac.uk) was made available under Application Number 22419. This study also makes use of data generated by the Wellcome Trust Case Control Consortium (WTCCC). A full list of the investigators who contributed to the generation of the data is available
from www.wtccc.org.uk. Funding for the WTCCC project was provided by the Wellcome Trust under award 076113, 085475, and 090355.
Appendix A

Supporting Information to
“Estimation of Non-null SNP Effect Size Distributions Enables the Detection of Enriched Genes Underlying Complex Traits”

A.1 Supporting Information

A.1.1 Data Quality Control Procedures

The results presented in the main text made use of imputed data released from the UK Biobank (24). Quality control procedures for these data are as follows. First, we only studied individuals who self-identified as “white British” people. From this cohort, we further excluded individuals
identified by the UK Biobank to have high heterozygosity, excessive relatedness, or aneuploidy (1,550 individuals removed). We also removed individuals whose kinship coefficient was greater than 0.0442 (i.e., close relatives). Next, we removed (i) monomorphic SNPs, (ii) ambiguous A/T or C/G SNPs, (iii) SNPs with minor allele frequency (MAF) less than 2.5%, (iv) SNPs not in Hardy-Weinberg Equilibrium (Fisher’s exact test $P > 10^{-6}$), (v) SNPs with missingness greater than 1%, and (vi) SNPs in high linkage disequilibrium (using the flag `--indep-pairwise 50 5 0.9` with PLINK 1.9 (197)). After all QC steps, we had a final dataset of 349,414 individuals and 1,070,306 SNPs. Next, we used the NCBI’s Reference Sequence (RefSeq) database in the UCSC Genome Browser (195) to annotate SNPs with the appropriate genes. Recall that in both the simulation studies and real data analysis, we define genes with boundaries in two ways: (a) we use the UCSC gene boundary definitions directly, or (b) we augment the gene boundaries by adding SNPs within a ±50 kilobase (kb) buffer to account for possible regulatory elements. Genes with only 1 SNP within their boundary were excluded from either analysis. A total of 14,322 autosomal genes were analyzed when using the UCSC boundaries, and a total of 17,680 autosomal genes were analyzed when including the 50kb buffer.

### A.1.2 Simulation Setup and Scenarios

In our simulation studies, we used the following general simulation scheme to generate SNP-level summary statistics for GWA studies using real genotype data on chromosome 1 from individuals of European ancestry in the UK Biobank (24). We will denote this genotype matrix as $X$, with $x_j$ denoting the genotypic vector for the $j$-th SNP. Following quality control procedures detailed in the previous section, our simulations included $J = 36,518$ SNPs distributed across genome. Again, we used the NCBI’s RefSeq database in the UCSC Genome Browser to assign SNPs to genes. Simulations were conducted using two different SNP-to-gene assignments. In the first, we directly used the UCSC annotations which resulted in 1,408 genes to be used in the simulation study. In the second, we augmented the UCSC gene boundaries to include SNPs within ±50kb resulting in 1,916 genes for analysis. Regardless of annotation type, we simulated phenotypes by first assuming that the total phenotypic variance $\mathbb{V}[\mathbf{y}] = 1$ and that all observed genetic effects explained a fixed
proportion of this value (i.e., narrow-sense heritability, $h^2$). Next, we randomly selected a certain percentage of enriched genes and denoted the sets of SNPs that they contained as $C$. Within $C$, we select causal SNPs in a way such that each associated gene at least contains one SNP with non-zero effect size. Quantitative continuous traits were then generated under the following two general linear models:

1. Standard Model: $y = \sum_{c \in C} x_c \beta_c + e$

2. Population Stratification Model: $y = Wb + \sum_{c \in C} x_c \beta_c + e$

where $y$ is an $N$-dimensional vector containing all the phenotypes; $x_c$ is the genotype for the $c$-th causal SNP encoded as 0, 1, or 2 copies of a reference allele; $\beta_c$ is the additive effect size for the $c$-th SNP; and $e \sim \mathcal{N}(0, \tau^2 I)$ is an $N$-dimensional vector of normally distributed environmental noise. Additionally, in model (ii), $W$ is an $N \times M$ matrix of the top five principal components (PCs) from the genotype matrix and represents additional population structure with corresponding fixed effects $b$. The effect sizes of SNPs in enriched genes are randomly drawn from standard normal distributions and then rescaled so they explain a fixed proportion of the narrow-sense heritability $\mathbb{V}[\sum x_c \beta_c] = h^2$. The coefficients for the genotype PCs are also drawn from standard normal distributions and rescaled such that $\mathbb{V}[Wb] = 10\%$ of the total phenotypic variance, with the variance of all non-genetic effects contributing $\mathbb{V}[Wb] + \mathbb{V}[e] = (1 - h^2)$. For any simulations conducted under model (ii), genotype PCs are not included in any of the model fitting procedures, and no other preprocessing normalizations were carried out to account for the additional population structure. More specifically, GWA summary statistics are then computed by fitting a single-SNP univariate linear model via ordinary least squares (OLS):

$$\hat{\beta}_j = (x_j^T x_j)^{-1} x_j^T y; \quad (A.1)$$

for every SNP in the data $j = 1, \ldots, J$. These OLS effect size estimates, along with an empirically LD matrix $\Sigma$ computed directly from the full $N \times J$ genotype matrix $X$, are given to gene-$\varepsilon$. We also retain standard errors and $P$-values for the implementation of competing methods (i.e., VEGAS,
PEGASUS, RSS, SKAT, and MAGMA). Given the simulation procedure above, we simulate a wide range of scenarios for comparing the performance of gene-level association approaches by varying the following parameters:

- Number of individuals: \( N = 5,000 \) and \( 10,000 \);
- Narrow-sense heritability: \( h^2 = 0.2 \) and \( 0.6 \);
- Percentage of enriched genes: 1% and 10%;

Furthermore, we set the number of causal SNPs with non-zero effects to be some fixed percentage of all SNPs located within the designated enriched genes. In the setting where we have 1,408 genes with boundaries defined strictly by RefSeq in UCSC Genome Browser, we set this percentage to be 0.125% in the 1% associated gene case, and 3% in the 10% associated gene case. In the setting where we have 1,916 genes with boundaries augmented by the ±50kb buffer, we set this percentage to be 0.125% in the 1% associated gene case, and 8% in the 10% associated gene case. Lastly, for each simulated dataset, we also selected some number of intergenic SNPs (i.e., SNPs not mapped to any gene) to have non-zero effect sizes. This was done to mimic genetic associations in unannotated regulatory elements. Specifically, 5 randomly selected intergenic SNPs were given non-zero contributions to the trait heritability in the 1% enriched genes case, and 30 intergenic SNPs were selected in the 10% enriched genes case.

All performance comparisons are based on 100 different simulated runs for each parameter combination. We computed gene-level \( P \)-values for the gene-\( \varepsilon \) approaches, PEGASUS, VEGAS, SKAT, and MAGMA. For evaluating the performance of RSS, we compute posterior enrichment probabilities. For all approaches, we assessed:

- The power and false discovery rates when identifying enriched genes at a Bonferroni-corrected threshold (\( P = 0.05/1,408 \) genes = \( 3.55 \times 10^{-5} \); \( P = 0.05/1,916 \) genes = \( 2.61 \times 10^{-5} \) if the ±50kb buffer was used) or median probability model (posterior enrichment probability > 0.5) (6);
• The ability to rank true positive (TP) genes over false positives (FP) via receiver operating characteristic (ROC) and precision-recall curves.

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

A.1.3 Review of Other Gene-Level Association Methods

In this section, we give a comprehensive review of the three gene-level association tests that we compare with the gene-ε approach. To facilitate the understanding of these summaries, we adapt notation from the original references that first introduced these methods to mirror the notation we use in this study.

Precise, Efficient Gene Association Score Using SNPs (PEGASUS). Consider a gene \( g \) with \( |\mathcal{J}_g| \) SNPs, where \( |\mathcal{J}_g| \) represents the cardinality of the set of SNPs \( \mathcal{J}_g \). Also assume that we have access to corresponding \( |\mathcal{J}_g| \) GWA SNP-level \( P \)-values. We denote the \( P \)-values for SNPs within a given gene boundary as \( \tilde{p}_g = \{\tilde{p}_1, \ldots, \tilde{p}_{|\mathcal{J}_g|}\} \). PEGASUS computes a gene-level test statistic \( \hat{Q}_g \) via the following quadratic form

\[
\hat{Q}_g = \hat{\beta}_g^T A \hat{\beta}_g
\]

where \( \hat{\beta}_g = F^{-1}(\tilde{p}_g) \), and \( F^{-1}(\bullet) \) is the quantile function of the standard chi-square distribution with one degree of freedom, and \( A \) is a predefined symmetric and positive semi-definite weight matrix. Probabilistically, under the null hypothesis, \( \hat{\beta}_g \) is assumed to jointly follow a multivariate normal distribution with mean \( 0 \) and covariance matrix \( \Sigma_g \), where each matrix element \( \rho(x_j, x_l) \) is the LD between the \( j \)-th and \( l \)-th SNPs contained within gene \( g \). Therefore, also under the null hypothesis, \( Q_g \) is assumed to follow a mixture of chi-square distributions,

\[
Q_g \sim \sum_{j=1}^{|\mathcal{J}_g|} \lambda_j U_j^2
\]
where each \( U_j \) is a mutually independent standard normal variables, and \((\lambda_1, \ldots, \lambda_{|J_g|})\) are the eigenvalues of the matrix product \( \Sigma_g A \). \( P \)-values are computed numerically using Davies’ exact method (55). See (183) for more details. Note that in our implementation of PEGASUS, \( A = I \) is set to be the identity matrix.

**Versatile Gene-based Association Study (VEGAS).** Again consider a gene \( g \) with \(|J_g|\) SNPs. Under the null hypothesis, a non-associated gene will contain only non-causal SNPs and is assumed to be represented by a \(|J_g|\)-dimensional multivariate normal vector \( \beta_g^* = (\beta_1^*, \ldots, \beta_{|J_g|}^*) \) for which

\[
\beta_g^* \sim \mathcal{N}(0, \Sigma_g),
\]  

(A.4)

where \( \Sigma_g \) is the LD matrix for all SNPs within gene \( g \). VEGAS generates gene scores by: (i) simulating the random vector \( \beta_g^* \) upwards of one million times, (ii) transforming the elements of each vector into correlated chi-square variables with one degree of freedom where \( q_j = \beta_j^{*2} \) and \( Q_g^* = (q_1, \ldots, q_{|J_g|}) \), (iii) acquiring realizations from the null distribution by summing over all the components in each \( Q_g^* \), and (iv) computing an empirical gene-level \( P \)-value based on the proportion of times an observed test statistic is smaller than the simulated null statistics \( \Pr[\sum Q_g < \sum Q_g^*] \) across all simulations. See (154) for more details.

**Regression with Summary Statistics (RSS) Enrichment.** Consider a GWA study with \( N \) individuals typed on \( P \) SNPs. For the \( j \)-th SNP, assume that we are given corresponding effect sizes \( \hat{\beta}_j \) and standard error \( \hat{s}_j \) via a single-SNP linear model fit using OLS. RSS then implements the following likelihood to model the GWA summary statistics (306)

\[
\hat{\beta} \sim \mathcal{N}(\hat{S}\Sigma\hat{S}^{-1}\beta, \hat{S}\Sigma\hat{S})
\]  

(A.5)

where \( \hat{S} = \text{diag}(\hat{s}) \) is a \( J \times J \) diagonal matrix of standard errors, \( \Sigma \) is again used to represent some empirical estimate of the LD matrix (i.e., using some external reference panel with ancestry
matching the cohort of interest), and $\beta$ are the true (unobserved) SNP-level effect sizes. To model
gene-level enrichment, RSS assumes the following hierarchical prior structure on the true effect sizes

$$\beta_j \sim \pi_j \mathcal{N}(0, \sigma^2_\beta) + (1 - \pi_j) \delta_0,$$  \hspace{1cm} (A.6)

$$\sigma^2_\beta = h^2 \left( \sum_{j=1}^{J} \pi_j N^{-1} \hat{s}_j^{-2} \right)^{-1},$$  \hspace{1cm} (A.7)

$$\pi_j = \left( 1 + 10^{-\left( \theta_0 + a_j \theta \right)} \right)^{-1},$$  \hspace{1cm} (A.8)

where $\delta_0$ is point mass centered at zero, $h^2$ denotes the narrow-sense heritability of the trait, $a_j$
is an indicator detailing whether the $j$-th SNP is inside a particular gene, $\theta_0$ is the background
proportion of trait-associated SNPs, and $\theta$ reflects the increase in probability (on the log$_{10}$-odds
scale) when a SNP within a gene has non-zero effect. Here, the authors follow earlier works (28)
and place independent uniform grid priors on the hyper-parameters $\{h^2, \theta_0, \theta\}$. Note that, unlike
other methods, RSS does not calculate a $P$-value for assessing gene-level association. Instead, RSS
produces a posterior enrichment probability that at least one SNP in a given gene boundary is
associated with the trait

$$P_g := 1 - \Pr [\beta_j = 0, \forall j \in J_g | D]$$  \hspace{1cm} (A.9)

where $D$ represents all of the input data including the GWA summary statistics $\{\hat{\beta}, \hat{s}\}$, the estimated
LD matrix $\Sigma$, and any applicable SNP annotations or weights $a = (a_1, \ldots, a_J)$. See (306; 307) for
more details on preferred hyper-parameter settings. As noted in the main text, RSS is relies on a
Markov chain Monte Carlo (MCMC) scheme for sampling posterior distributions and estimating
model parameters. As a result, its algorithm can be subject to convergence issues if these (or the
random seed) are not chosen properly.

**SNP-set (Sequence) Kernel Association Test (SKAT).** The implementation of SKAT re-
quired access to raw phenotype $y$ and genotype $X$ information for $N$ individuals typed on $J$ SNPs.
To assess enrichment of the $|J_g|$ variants within gene $g$, consider the linear model with sub-matrix
where $\beta_0$ is an intercept term, $\beta_g = (\beta_1, \ldots, \beta_{|J_g|})$ is a vector of regression coefficients for the SNPs within the gene of interest, and $e$ is a normally distributed error term with mean zero and scaled variance $\tau^2$. For model flexibility, gene-specific SNP effects $\beta_j$ are assumed to follow an arbitrary distribution with mean zero and marginal variances $a_j \sigma^2_\beta$, where $\sigma^2_\beta$ is a variance component and $a_j$ is a pre-specified weight for the $j$-th SNP. To this end, SKAT uses a variance component scoring approach and tests the null hypothesis $H_0: \beta = 0$, or equivalently $H_0: \sigma^2_\beta = 0$. The corresponding gene-level test statistic $\hat{Q}_g$ then takes on the familiar quadratic form

$$
\hat{Q}_g = (y - \hat{\beta}_0)^T K_g (y - \hat{\beta}_0)
$$

where $\hat{\beta}_0$ is the predicted mean of trait under the null hypothesis, and is computed by projecting $y$ onto the column space of the intercept (i.e., a vector of ones). The term $K_g = X_g A_g A_g X_g^T$ is commonly referred to as an $N \times N$ kernel matrix, where $A_g = \text{diag}(a_1, \ldots, a_{|J_g|})$ is used to denote a diagonal weight matrix that changes for each gene $g$. Each element of $K_g$ is computed via the linear kernel function

$$
k(x_i, x_i') = \sum_{j=1}^{|J_g|} a_j x_{ij} x_{i'j}.
$$

While implementing SKAT in this work, we follow previous works and set each weight to be $\sqrt{\sigma_j} = \text{Beta}(\text{MAF}_j, 1, 25)$ — the beta distribution density function with pre-specified parameters evaluated at the sample minor allele frequency (MAF) for the $j$-th SNP in the gene region. For more details, see (281; 282; 139; 109).

**Multi-marker Analysis of GenoMic Annotation (MAGMA).** In the current study, gene analyses with MAGMA also required access to raw phenotype $y$ and genotype $X$ information for $N$
individuals typed on $J$ SNPs. This approach is based on a multiple principal components regression model. In the first step, MAGMA projects the sub-genotype matrix for a gene $X_g$ onto its principal components. Next, it prunes away PCs with very small eigenvalues, and then uses those reduced vectors as predictors for the phenotype in the linear regression model. Consider the following linear regression and singular-value decomposition (SVD) of the genotype matrix

$$y = \beta_0 + X\beta + e, \quad X = U\Lambda V^\top, \quad e \sim \mathcal{N}(0, \tau^2 I)$$

(A.13)

where, in addition to the aforementioned notation, $\Lambda$ is an $N \times J$ rectangular diagonal matrix of singular values, and $U$ and $V$ are $N \times N$ and $J \times J$ matrices of orthogonal unit vectors, respectively. For numerical stability and reduction of computational complexity, vectors corresponding to small eigenvalues can be truncated. Therefore, without loss of generality, MAGMA considers $V$ and $\Lambda$ to be of dimensions $J^* \times J^*$ and $N \times J^*$, respectively. Here, $J^*$ denotes the top eigenvalues explaining $99.9\%$ of the cumulative variance in $X_g$. By defining $G = U\Lambda$, the model above simplifies to

$$y = \beta_0 + G\vartheta + e, \quad e \sim \mathcal{N}(0, \tau^2 I)$$

(A.14)

where $\vartheta = V^\top\beta$ represents the lower-dimensional genetic effect. To derive a $P$-value for a single gene’s association with the phenotype, MAGMA uses an F-test under the null hypothesis $H_0: \vartheta = 0$ or, equivalently, $H_0: V^\top\beta = 0$. See (56) for more details.

A.1.4 Additional Detailed Results for Traits in the UK Biobank

In this section, we present additional detailed findings and results from applying gene-$\varepsilon$ to the six quantitative traits — height, body mass index (BMI), mean red blood cell volume (MCV), mean platelet volume (MPV), platelet count (PLC), waist-hip ratio (WHR) — assayed in self-identified European-ancestry individuals in the UK Biobank (24). For these extra set of analyses, we obtained the genotype data release (without imputed genotypes) and implemented the same quality control procedure that was used in the main text (Section A.1.1). This resulted in a final dataset of
$N = 349,468$ individuals and $J = 410,172$ genome-wide SNPs. Once again, we used the NCBI’s Reference Sequence (RefSeq) database in the UCSC Genome Browser (195) to annotate SNPs with the appropriate genes in one of two ways. In the first setting, we use the UCSC gene boundary definitions directly; while in the second setting, we augment the gene boundaries by adding SNPs within a ±50 kilobase (kb) buffer to account for possible regulatory elements. Genes with only 1 SNP in their boundary were excluded from the respective analysis. For these data, a total of 13,029 autosomal genes were analyzed when using the UCSC boundaries as defined; while, a total of 17,680 autosomal genes were analyzed when including the 50kb buffer. Lastly, we regressed the top ten principal components of the genotype data onto each trait to control for population structure, and then we derived OLS SNP-level effect sizes using the traditional GWA framework. Here, our goal is to compare how the four different implementations of gene-$\varepsilon$ (i.e., OLS with no regularization, Ridge Regression, Elastic Net, and LASSO) analyze these summary statistics.

As shown in the main text, we begin with assessing how the various regularization solutions result in different characterizations of genetic architectures (S25 Table). In general, we find the same general themes we saw in our simulation study. Less aggressive shrinkage approaches (e.g., OLS and Ridge) are subject to misclassifications of associated, spurious, and non-associated SNPs. As result, these methods struggle to avoid identifying false positive SNP-level associations, across all six traits. For example, gene-$\varepsilon$-OLS assumes that approximately 54% and 50% of the SNPs analyzed are associated with BMI and WHR, respectively. This once again highlights the need for computational frameworks that are able to appropriately correct for inflation in summary statistics.

Lastly, we applied each version of gene-$\varepsilon$ to the (regularized) GWA summary statistics and generated genome-wide gene-level association $P$-values. Recall that we are motivated to identify enriched genes, which we define as a gene containing at least one associated SNP and achieving a gene-level association $P$-value below a Bonferroni-corrected significance threshold. In our analyses, this significance threshold is $P = 0.05/13029$ autosomal genes $= 3.84 \times 10^{-6}$ when the UCSC gene boundaries are used directly, and $P = 0.05/17680$ autosomal genes $= 2.83 \times 10^{-6}$ when the ±50kb buffer is applied, respectively. As a validation step, we used the gene set enrichment analysis tool Enrichr (34) to identify dbGaP categories with an overrepresentation of significant genes reported
by the four different implementations of gene-$\varepsilon$. A comparison of gene-level associations and gene set enrichments between the each gene-$\varepsilon$ approaches are also listed (S26 and S27 Tables). Note that, similar to the main text, we use the findings of gene-$\varepsilon$-EN as the reference.
A.2 Supplementary Figures

Figure A.1. Simulation study results showing the Pearson correlation between various degrees of gene-ε regularized SNP-level effect size estimates and the true effect sizes that generated the complex traits. Assessed regularization techniques are the (A) LASSO (243), (B) Elastic Net (308), (C) Ridge Regression (99), and (D) no regularization of ordinary least squares (OLS) effect sizes which serves as a baseline. Here, we take real genotype data on chromosome 19 from N = 5,000 randomly chosen individuals of European ancestry in the UK Biobank (see Section A.1.1). We then assumed a simple linear additive model for quantitative traits while varying the narrow-sense heritability (h^2 = \{0.01, 0.05, 0.10, 0.15, 0.20, 0.25\}). We considered two scenarios where traits are generated with and without additional population structure (colored as pink and blue lines, respectively). In the former setting, phenotypes are simulated while also using the top five principal components (PCs) of the genotype matrix as covariates to create stratification. These PCs contributed to 10% of the phenotypic variance. In both settings, GWA SNP-level effect sizes were derived via OLS without accounting for any additional structure. The y-axis shows Pearson correlation between gene-ε regularized effect sizes and the truth. On the x-axis of each plot, we vary the number of causal SNPs for each trait (i.e., 1, 5, 10, 15, 20, 25)%. Results are based on ten replicates (see Section A.1.2), with the error bars representing standard errors across runs.
Figure A.2. (A, C) Receiver operating characteristic (ROC) and (B, D) precision-recall curves comparing the performance of gene-$\varepsilon$ and competing approaches in simulations ($N = 5,000$; $h^2 = 0.2$). Here, the sample size $N = 5,000$ and the narrow-sense heritability of the simulated quantitative trait is $h^2 = 0.2$. We compute standard GWA SNP-level effect sizes (estimated using ordinary least squares). Results for gene-$\varepsilon$ are shown with LASSO (blue), Elastic Net (EN; red), and Ridge Regression (RR; purple) regularizations. We also show the results of gene-$\varepsilon$ without regularization to illustrate the importance of the regularization step (labeled OLS; orange). We compare gene-$\varepsilon$ with five existing methods: PEGASUS (brown) (183), VEGAS (teal) (154), the Bayesian approach RSS (black) (307), SKAT (green) (281), and MAGMA (peach) (56). (A, C) ROC curves show power versus false positive rate for each approach of sparse (1% enriched genes) and polygenic (10% enriched genes) architectures, respectively. Note that the upper limit of the x-axis has been truncated at 0.1. (B, D) Precision-Recall curves for each method applied to the simulations. Note that, in the sparse case (1% enriched genes), the top ranked genes are always true positives, and therefore the minimal recall is not 0. All results are based on 100 replicates (see Section A.1.2).
Figure A.3. (A, C) Receiver operating characteristic (ROC) and (B, D) precision-recall curves comparing the performance of gene-ε and competing approaches in simulations ($N = 10,000; h^2 = 0.2$). Here, the sample size $N = 10,000$ and the narrow-sense heritability of the simulated quantitative trait is $h^2 = 0.2$. We compute standard GWA SNP-level effect sizes (estimated using ordinary least squares). Results for gene-ε are shown with LASSO (blue), Elastic Net (EN; red), and Ridge Regression (RR; purple) regularizations. We also show the results of gene-ε without regularization to illustrate the importance of the regularization step (labeled OLS; orange). We compare gene-ε with five existing methods: PEGASUS (brown) (183), VEGAS (teal) (154), the Bayesian approach RSS (black) (307), SKAT (green) (281), and MAGMA (peach) (56). (A, C) ROC curves show power versus false positive rate for each approach of sparse (1% enriched genes) and polygenic (10% enriched genes) architectures, respectively. Note that the upper limit of the x-axis has been truncated at 0.1. (B, D) Precision-Recall curves for each method applied to the simulations. Note that, in the sparse case (1% enriched genes), the top ranked genes are always true positives, and therefore the minimal recall is not 0. All results are based on 100 replicates (see Section A.1.2).
Figure A.4. (A, C) Receiver operating characteristic (ROC) and (B, D) precision-recall curves comparing the performance of gene-ε and competing approaches in simulations ($N = 5,000$; $h^2 = 0.6$). Here, the sample size $N = 5,000$ and the narrow-sense heritability of the simulated quantitative trait is $h^2 = 0.6$. We compute standard GWA SNP-level effect sizes (estimated using ordinary least squares). Results for gene-ε are shown with LASSO (blue), Elastic Net (EN; red), and Ridge Regression (RR; purple) regularizations. We also show the results of gene-ε without regularization to illustrate the importance of the regularization step (labeled OLS; orange). We compare gene-ε with five existing methods: PEGASUS (brown) (183), VEGAS (teal) (154), the Bayesian approach RSS (black) (307), SKAT (green) (281), and MAGMA (peach) (56). (A, C) ROC curves show power versus false positive rate for each approach of sparse (1% enriched genes) and polygenic (10% enriched genes) architectures, respectively. Note that the upper limit of the x-axis has been truncated at 0.1. (B, D) Precision-Recall curves for each method applied to the simulations. Note that, in the sparse case (1% enriched genes), the top ranked genes are always true positives, and therefore the minimal recall is not 0. All results are based on 100 replicates (see Section A.1.2).
Figure A.5. (A, C) Receiver operating characteristic (ROC) and (B, D) precision-recall curves comparing the performance of gene-$\varepsilon$ and competing approaches in simulations with population stratification ($N = 5,000; h^2 = 0.2$). Here, the sample size $N = 5,000$ and the narrow-sense heritability of the simulated quantitative trait is $h^2 = 0.2$. In this simulation, traits were generated while using the top five principal components (PCs) of the genotype matrix as covariates. GWA summary statistics were computed by fitting a single-SNP univariate linear model (via ordinary least squares) without any control for the additional structure. Results for gene-$\varepsilon$ are shown with LASSO (blue), Elastic Net (EN; red), and Ridge Regression (RR; purple) regularizations. We also show the results of gene-$\varepsilon$ without regularization to illustrate the importance of the regularization step (labeled OLS; orange). We compare gene-$\varepsilon$ with five existing methods: PEGASUS (brown) (183), VEGAS (teal) (154), the Bayesian approach RSS (black) (307), SKAT (green) (281), and MAGMA (peach) (56). Note that each was method implemented without using any covariates. (A, C) ROC curves show power versus false positive rate for each approach of sparse (1% enriched genes) and polygenic (10% enriched genes) architectures, respectively. Note that the upper limit of the x-axis has been truncated at 0.1. (B, D) Precision-Recall curves for each method applied to the simulations. Note that, in the sparse case (1% enriched genes), the top ranked genes are always true positives, and therefore the minimal recall is not 0. All results are based on 100 replicates (see Section A.1.2).
Figure A.6. (A, C) Receiver operating characteristic (ROC) and (B, D) precision-recall curves comparing the performance of gene-ε and competing approaches in simulations with population stratification ($N = 10,000$; $h^2 = 0.2$). Here, the sample size $N = 10,000$ and the narrow-sense heritability of the simulated quantitative trait is $h^2 = 0.2$. In this simulation, traits were generated while using the top five principal components (PCs) of the genotype matrix as covariates. GWA summary statistics were computed by fitting a single-SNP univariate linear model (via ordinary least squares) without any control for the additional structure. Results for gene-ε are shown with LASSO (blue), Elastic Net (EN; red), and Ridge Regression (RR; purple) regularizations. We also show the results of gene-ε without regularization to illustrate the importance of the regularization step (labeled OLS; orange). We compare gene-ε with five existing methods: PEGASUS (brown) (183), VEGAS (teal) (154), the Bayesian approach RSS (black) (307), SKAT (green) (281), and MAGMA (peach) (56). Note that each was method implemented without using any covariates. (A, C) ROC curves show power versus false positive rate for each approach of sparse (1% enriched genes) and polygenic (10% enriched genes) architectures, respectively. Note that the upper limit of the x-axis has been truncated at 0.1. (B, D) Precision-Recall curves for each method applied to the simulations. Note that, in the sparse case (1% enriched genes), the top ranked genes are always true positives and therefore the minimal recall is not 0. All results are based on 100 replicates (see Section A.1.2).
Figure A.7. (A, C) Receiver operating characteristic (ROC) and (B, D) precision-recall curves comparing the performance of gene-ε and competing approaches in simulations with population stratification (N = 5,000; h^2 = 0.6). Here, the sample size N = 5,000 and the narrow-sense heritability of the simulated quantitative trait is h^2 = 0.6. In this simulation, traits were generated while using the top five principal components (PCs) of the genotype matrix as covariates. GWA summary statistics were computed by fitting a single-SNP univariate linear model (via ordinary least squares) without any control for the additional structure. Results for gene-ε are shown with LASSO (blue), Elastic Net (EN; red), and Ridge Regression (RR; purple) regularizations. We also show the results of gene-ε without regularization to illustrate the importance of the regularization step (labeled OLS; orange). We compare gene-ε with five existing methods: PEGASUS (brown) (183), VEGAS (teal) (154), the Bayesian approach RSS (black) (307), SKAT (green) (281), and MAGMA (peach) (56). Note that each was method implemented without using any covariates. (A, C) ROC curves show power versus false positive rate for each approach of sparse (1% enriched genes) and polygenic (10% enriched genes) architectures, respectively. Note that the upper limit of the x-axis has been truncated at 0.1. (B, D) Precision-Recall curves for each method applied to the simulations. Note that, in the sparse case (1% enriched genes), the top ranked genes are always true positives, and therefore the minimal recall is not 0. All results are based on 100 replicates (see Section A.1.2).
Figure A.8. (A, C) Receiver operating characteristic (ROC) and (B, D) precision-recall curves comparing the performance of gene-ε and competing approaches in simulations with population stratification (N = 10,000; h^2 = 0.6). Here, the sample size N = 10,000 and the narrow-sense heritability of the simulated quantitative trait is h^2 = 0.6. In this simulation, traits were generated while using the top five principal components (PCs) of the genotype matrix as covariates. GWA summary statistics were computed by fitting a single-SNP univariate linear model (via ordinary least squares) without any control for the additional structure. Results for gene-ε are shown with LASSO (blue), Elastic Net (EN; red), and Ridge Regression (RR; purple) regularizations. We also show the results of gene-ε without regularization to illustrate the importance of the regularization step (labeled OLS; orange). We compare gene-ε with five existing methods: PEGASUS (brown) (183), VEGAS (teal) (154), the Bayesian approach RSS (black) (307), SKAT (green) (281), and MAGMA (peach) (56). Note that each was method implemented without using any covariates. (A, C) ROC curves show power versus false positive rate for each approach of sparse (1% enriched genes) and polygenic (10% enriched genes) architectures, respectively. Note that the upper limit of the x-axis has been truncated at 0.1. (B, D) Precision-Recall curves for each method applied to the simulations. Note that, in the sparse case (1% enriched genes), the top ranked genes are always true positives, and therefore the minimal recall is not 0. All results are based on 100 replicates (see Section A.1.2).
Figure A.9. (A, C) Receiver operating characteristic (ROC) and (B, D) precision-recall curves comparing the performance of gene-$\varepsilon$ and competing approaches in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer ($N = 5,000$; $h^2 = 0.2$). Here, the sample size $N = 5,000$ and the narrow-sense heritability of the simulated quantitative trait is $h^2 = 0.2$. We compute standard GWA SNP-level effect sizes (estimated using ordinary least squares). Results for gene-$\varepsilon$ are shown with LASSO (blue), Elastic Net (EN; red), and Ridge Regression (RR; purple) regularizations. We also show the results of gene-$\varepsilon$ without regularization to illustrate the importance of the regularization step (labeled OLS; orange). We compare gene-$\varepsilon$ with five existing methods: PEGASUS (brown) (183), VEGAS (teal) (154), the Bayesian approach RSS (black) (307), SKAT (green) (281), and MAGMA (peach) (56). (A, C) ROC curves show power versus false positive rate for each approach of sparse (1% enriched genes) and polygenic (10% enriched genes) architectures, respectively. Note that the upper limit of the x-axis has been truncated at 0.1. (B, D) Precision-Recall curves for each method applied to the simulations. Note that, in the sparse case (1% enriched genes), the top ranked genes are always true positives, and therefore the minimal recall is not 0. All results are based on 100 replicates (see Section A.1.2).
Figure A.10. (A, C) Receiver operating characteristic (ROC) and (B, D) precision-recall curves comparing the performance of gene-ε and competing approaches in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer ($N = 10,000$; $h^2 = 0.2$). Here, the sample size $N = 10,000$ and the narrow-sense heritability of the simulated quantitative trait is $h^2 = 0.2$. We compute standard GWA SNP-level effect sizes (estimated using ordinary least squares). Results for gene-ε are shown with LASSO (blue), Elastic Net (EN; red), and Ridge Regression (RR; purple) regularizations. We also show the results of gene-ε without regularization to illustrate the importance of the regularization step (labeled OLS; orange). We compare gene-ε with five existing methods: PEGASUS (brown) (183), VEGAS (teal) (154), the Bayesian approach RSS (black) (307), SKAT (green) (281), and MAGMA (peach) (56). (A, C) ROC curves show power versus false positive rate for each approach of sparse (1% enriched genes) and polygenic (10% enriched genes) architectures, respectively. Note that the upper limit of the x-axis has been truncated at 0.1. (B, D) Precision-Recall curves for each method applied to the simulations. Note that, in the sparse case (1% enriched genes), the top ranked genes are always true positives, and therefore the minimal recall is not 0. All results are based on 100 replicates (see Section A.1.2).
Figure A.11. (A, C) Receiver operating characteristic (ROC) and (B, D) precision-recall curves comparing the performance of gene-$\epsilon$ and competing approaches in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer ($N = 5,000$; $h^2 = 0.6$). Here, the sample size $N = 5,000$ and the narrow-sense heritability of the simulated quantitative trait is $h^2 = 0.6$. We compute standard GWA SNP-level effect sizes (estimated using ordinary least squares). Results for gene-$\epsilon$ are shown with LASSO (blue), Elastic Net (EN; red), and Ridge Regression (RR; purple) regularizations. We also show the results of gene-$\epsilon$ without regularization to illustrate the importance of the regularization step (labeled OLS; orange). We compare gene-$\epsilon$ with five existing methods: PEGASUS (brown) (183), VEGAS (teal) (154), the Bayesian approach RSS (black) (307), SKAT (green) (281), and MAGMA (peach) (56). (A, C) ROC curves show power versus false positive rate for each approach of sparse (1% enriched genes) and polygenic (10% enriched genes) architectures, respectively. Note that the upper limit of the x-axis has been truncated at 0.1. (B, D) Precision-Recall curves for each method applied to the simulations. Note that, in the sparse case (1% enriched genes), the top ranked genes are always true positives, and therefore the minimal recall is not 0. All results are based on 100 replicates (see Section A.1.2).
Figure A.12. (A, C) Receiver operating characteristic (ROC) and (B, D) precision-recall curves comparing the performance of gene-ε and competing approaches in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer (N = 10,000; h² = 0.6). Here, the sample size N = 10,000 and the narrow-sense heritability of the simulated quantitative trait is h² = 0.6. We compute standard GWA SNP-level effect sizes (estimated using ordinary least squares). Results for gene-ε are shown with LASSO (blue), Elastic Net (EN; red), and Ridge Regression (RR; purple) regularizations. We also show the results of gene-ε without regularization to illustrate the importance of the regularization step (labeled OLS; orange). We compare gene-ε with five existing methods: PEGASUS (brown) (183), VEGAS (teal) (154), the Bayesian approach RSS (black) (307), SKAT (green) (281), and MAGMA (peach) (56). (A, C) ROC curves show power versus false positive rate for each approach of sparse (1% enriched genes) and polygenic (10% enriched genes) architectures, respectively. Note that the upper limit of the x-axis has been truncated at 0.1. (B, D) Precision-Recall curves for each method applied to the simulations. Note that, in the sparse case (1% enriched genes), the top ranked genes are always true positives, and therefore the minimal recall is not 0. All results are based on 100 replicates (see Section A.1.2).
Figure A.13. (A, C) Receiver operating characteristic (ROC) and (B, D) precision-recall curves comparing the performance of gene-ε and competing approaches in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer and with population stratification ($N = 5,000; h^2 = 0.2$). Here, the sample size $N = 5,000$ and the narrow-sense heritability of the simulated quantitative trait is $h^2 = 0.2$. In this simulation, traits were generated while using the top five principal components (PCs) of the genotype matrix as covariates. GWA summary statistics were computed by fitting a single-SNP univariate linear model (via ordinary least squares) without any control for the additional structure. Results for gene-ε are shown with LASSO (blue), Elastic Net (EN; red), and Ridge Regression (RR; purple) regularizations. We also show the results of gene-ε without regularization to illustrate the importance of the regularization step (labeled OLS; orange). We compare gene-ε with five existing methods: PEGASUS (brown) (183), VEGAS (teal) (154), the Bayesian approach RSS (black) (307), SKAT (green) (281), and MAGMA (peach) (56). Note that each was method implemented without using any covariates. (A, C) ROC curves show power versus false positive rate for each approach of sparse (1% enriched genes) and polygenic (10% enriched genes) architectures, respectively. Note that the upper limit of the x-axis has been truncated at 0.1. (B, D) Precision-Recall curves for each method applied to the simulations. Note that, in the sparse case (1% enriched genes), the top ranked genes are always true positives, and therefore the minimal recall is not 0. All results are based on 100 replicates (see Section A.1.2).
Figure A.14. (A, C) Receiver operating characteristic (ROC) and (B, D) precision-recall curves comparing the performance of gene-$\varepsilon$ and competing approaches in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer and with population stratification ($N = 10,000$; $h^2 = 0.2$). Here, the sample size $N = 10,000$ and the narrow-sense heritability of the simulated quantitative trait is $h^2 = 0.2$. In this simulation, traits were generated while using the top five principal components (PCs) of the genotype matrix as covariates. GWA summary statistics were computed by fitting a single-SNP univariate linear model (via ordinary least squares) without any control for the additional structure. Results for gene-$\varepsilon$ are shown with LASSO (blue), Elastic Net (EN; red), and Ridge Regression (RR; purple) regularizations. We also show the results of gene-$\varepsilon$ without regularization to illustrate the importance of the regularization step (labeled OLS; orange). We compare gene-$\varepsilon$ with five existing methods: PEGASUS (brown) (183), VEGAS (teal) (154), the Bayesian approach RSS (black) (307), SKAT (green) (281), and MAGMA (peach) (56). Note that each was method implemented without using any covariates. (A, C) ROC curves show power versus false positive rate for each approach of sparse (1% enriched genes) and polygenic (10% enriched genes) architectures, respectively. Note that the upper limit of the x-axis has been truncated at 0.1. (B, D) Precision-Recall curves for each method applied to the simulations. Note that, in the sparse case (1% enriched genes), the top ranked genes are always true positives, and therefore the minimal recall is not 0. All results are based on 100 replicates (see Section A.1.2).
Figure A.15. (A, C) Receiver operating characteristic (ROC) and (B, D) precision-recall curves comparing the performance of gene-$\varepsilon$ and competing approaches in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer and with population stratification ($N = 5,000$; $h^2 = 0.6$). Here, the sample size $N = 5,000$ and the narrow-sense heritability of the simulated quantitative trait is $h^2 = 0.6$. In this simulation, traits were generated while using the top five principal components (PCs) of the genotype matrix as covariates. GWA summary statistics were computed by fitting a single-SNP univariate linear model (via ordinary least squares) without any control for the additional structure. Results for gene-$\varepsilon$ are shown with LASSO (blue), Elastic Net (EN; red), and Ridge Regression (RR; purple) regularizations. We also show the results of gene-$\varepsilon$ without regularization to illustrate the importance of the regularization step (labeled OLS; orange). We compare gene-$\varepsilon$ with five existing methods: PEGASUS (brown) (183), VEGAS (teal) (154), the Bayesian approach RSS (black) (307), SKAT (green) (281), and MAGMA (peach) (56). Note that each was method implemented without using any covariates. (A, C) ROC curves show power versus false positive rate for each approach of sparse (1% enriched genes) and polygenic (10% enriched genes) architectures, respectively. Note that the upper limit of the x-axis has been truncated at 0.1. (B, D) Precision-Recall curves for each method applied to the simulations. Note that, in the sparse case (1% enriched genes), the top ranked genes are always true positives, and therefore the minimal recall is not 0. All results are based on 100 replicates (see Section A.1.2).
Figure A.16. (A, C) Receiver operating characteristic (ROC) and (B, D) precision-recall curves comparing the performance of gene-$\varepsilon$ and competing approaches in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer and with population stratification ($N = 10,000; h^2 = 0.6$). Here, the sample size $N = 10,000$ and the narrow-sense heritability of the simulated quantitative trait is $h^2 = 0.6$. In this simulation, traits were generated while using the top five principal components (PCs) of the genotype matrix as covariates. GWA summary statistics were computed by fitting a single-SNP univariate linear model (via ordinary least squares) without any control for the additional structure. Results for gene-$\varepsilon$ are shown with LASSO (blue), Elastic Net (EN; red), and Ridge Regression (RR; purple) regularizations. We also show the results of gene-$\varepsilon$ without regularization to illustrate the importance of the regularization step (labeled OLS; orange). We compare gene-$\varepsilon$ with five existing methods: PEGASUS (brown) (183), VEGAS (teal) (154), the Bayesian approach RSS (black) (307), SKAT (green) (281), and MAGMA (peach) (56). Note that each was method implemented without using any covariates. (A, C) ROC curves show power versus false positive rate for each approach of sparse (1% enriched genes) and polygenic (10% enriched genes) architectures, respectively. Note that the upper limit of the x-axis has been truncated at 0.1. (B, D) Precision-Recall curves for each method applied to the simulations. Note that, in the sparse case (1% enriched genes), the top ranked genes are always true positives, and therefore the minimal recall is not 0. All results are based on 100 replicates (see Section A.1.2).
Figure A.17. Scatter plots assessing how regularization on SNP-level summary statistics affects the ability to identify enriched genes in simulations ($h^2 = 0.2$). Here, the narrow-sense heritability of the simulated quantitative traits is $h^2 = 0.2$ and sample sizes are set to $N = 5,000$ in (A, B) and $N = 10,000$ in (C, D). In each case, standard GWA summary statistics were computed by fitting a single-SNP univariate linear model (via ordinary least squares). Results are shown comparing the $-\log_{10}$ transformed gene-level $P$-values derived by gene-$\epsilon$ with Elastic Net (EN) regularization on the y-axis and without regularization (labeled as OLS) on the x-axis. The horizontal and vertical dashed lines are marked at the Bonferroni-corrected threshold $P = 3.55 \times 10^{-5}$ corrected for the 1,408 genes on chromosome 1 from the UK Biobank genotype data. True positive causal genes used to generate the synthetic phenotypes are colored in red, while non-causal genes are given in grey. Genes in the top right quadrant are selected by both approaches. Genes in the top left and bottom right quadrants are uniquely identified by gene-$\epsilon$-EN and gene-$\epsilon$-OLS, respectively. To illustrate the importance of regularization on SNP-level summary statistics, we highlight the true positive genes only identified by gene-$\epsilon$-EN in blue. Each plot combines results from 100 simulated replicates (see Section A.1.2).
Figure A.18. Scatter plots assessing how regularization on SNP-level summary statistics affects the ability to identify enriched genes in simulations ($h^2 = 0.6$). Here, the narrow-sense heritability of the simulated quantitative traits is $h^2 = 0.6$ and sample sizes are set to $N = 5,000$ in (A, B) and $N = 10,000$ in (C, D). In each case, standard GWA summary statistics were computed by fitting a single-SNP univariate linear model (via ordinary least squares). Results are shown comparing the $-\log_{10}$ transformed gene-level $P$-values derived by gene-$\varepsilon$ with Elastic Net (EN) regularization on the y-axis and without regularization (labeled as OLS) on the x-axis. The horizontal and vertical dashed lines are marked at the Bonferonni-corrected threshold $P = 3.55 \times 10^{-5}$ corrected for the 1,408 genes on chromosome 1 from the UK Biobank genotype data. True positive causal genes used to generate the synthetic phenotypes are colored in red, while non-causal genes are given in grey. Genes in the top right quadrant are selected by both approaches. Genes in the top left and bottom right quadrants are uniquely identified by gene-$\varepsilon$-EN and gene-$\varepsilon$-OLS, respectively. To illustrate the importance of regularization on SNP-level summary statistics, we highlight the true positive genes only identified by gene-$\varepsilon$-EN in blue. Each plot combines results from 100 simulated replicates (see Section A.1.2).
Figure A.19. Scatter plots assessing how regularization on SNP-level summary statistics affects the ability to identify enriched genes in simulations with population stratification ($h^2 = 0.2$). Here, the narrow-sense heritability of the simulated quantitative traits is $h^2 = 0.2$ and sample sizes are set to $N = 5,000$ in (A, B) and $N = 10,000$ in (C, D). In this simulation, traits were generated while using the top five principal components (PCs) of the genotype matrix as covariates. GWA summary statistics were computed by fitting a single-SNP univariate linear model (via ordinary least squares) without any control for the additional structure. Results are shown comparing the -log$_{10}$ transformed gene-level $P$-values derived by gene-$\varepsilon$ with Elastic Net (EN) regularization on the y-axis and without regularization (labeled as OLS) on the x-axis. The horizontal and vertical dashed lines are marked at the Bonferonni-corrected threshold $P = 3.55 \times 10^{-5}$ corrected for the 1,408 genes on chromosome 1 from the UK Biobank genotype data. True positive causal genes used to generate the synthetic phenotypes are colored in red, while non-causal genes are given in grey. Genes in the top right quadrant are selected by both approaches. Genes in the top left and bottom right quadrants are uniquely identified by gene-$\varepsilon$-EN and gene-$\varepsilon$-OLS, respectively. To illustrate the importance of regularization on SNP-level summary statistics, we highlight the true positive genes only identified by gene-$\varepsilon$-EN in blue. Each plot combines results from 100 simulated replicates (see Section A.1.2).
Figure A.20. Scatter plots assessing how regularization on SNP-level summary statistics affects the ability to identify enriched genes in simulations with population stratification ($h^2 = 0.6$). Here, the narrow-sense heritability of the simulated quantitative traits is $h^2 = 0.6$ and sample sizes are set to $N = 5,000$ in (A, B) and $N = 10,000$ in (C, D). In this simulation, traits were generated while using the top five principal components (PCs) of the genotype matrix as covariates. GWA summary statistics were computed by fitting a single-SNP univariate linear model (via ordinary least squares) without any control for the additional structure. Results are shown comparing the $-\log_{10}$ transformed gene-level $P$-values derived by gene-$\varepsilon$ with Elastic Net (EN) regularization on the y-axis and without regularization (labeled as OLS) on the x-axis. The horizontal and vertical dashed lines are marked at the Bonferroni-corrected threshold $P = 3.55 \times 10^{-5}$ corrected for the 1,408 genes on chromosome 1 from the UK Biobank genotype data. True positive causal genes used to generate the synthetic phenotypes are colored in red, while non-causal genes are given in grey. Genes in the top right quadrant are selected by both approaches. Genes in the top left and bottom right quadrants are uniquely identified by gene-$\varepsilon$-EN and gene-$\varepsilon$-OLS, respectively. To illustrate the importance of regularization on SNP-level summary statistics, we highlight the true positive genes only identified by gene-$\varepsilon$-EN in blue. Each plot combines results from 100 simulated replicates (see Section A.1.2).
Figure A.21. Scatter plots assessing how regularization on SNP-level summary statistics affects the ability to identify enriched genes in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer ($h^2 = 0.2$). Here, the narrow-sense heritability of the simulated quantitative traits is $h^2 = 0.2$ and sample sizes are set to $N = 5,000$ in (A, B) and $N = 10,000$ in (C, D). In each case, standard GWA summary statistics were computed by fitting a single-SNP univariate linear model (via ordinary least squares). Results are shown comparing the $-\log_{10}$ transformed gene-level $P$-values derived by gene-$\varepsilon$ with Elastic Net (EN) regularization on the y-axis and without regularization (labeled as OLS) on the x-axis. The horizontal and vertical dashed lines are marked at the Bonferonni-corrected threshold $P = 2.61 \times 10^{-5}$ corrected for the 1,916 genes on chromosome 1 from the UK Biobank genotype data. True positive causal genes used to generate the synthetic phenotypes are colored in red, while non-causal genes are given in grey. Genes in the top right quadrant are selected by both approaches. Genes in the top left and bottom right quadrants are uniquely identified by gene-$\varepsilon$-EN and gene-$\varepsilon$-OLS, respectively. To illustrate the importance of regularization on SNP-level summary statistics, we highlight the true positive genes only identified by gene-$\varepsilon$-EN in blue. Each plot combines results from 100 simulated replicates (see Section A.1.2).
Figure A.22. Scatter plots assessing how regularization on SNP-level summary statistics affects the ability to identify enriched genes in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer ($h^2 = 0.6$). Here, the narrow-sense heritability of the simulated quantitative traits is $h^2 = 0.6$ and sample sizes are set to $N = 5,000$ in (A, B) and $N = 10,000$ in (C, D). In each case, standard GWA summary statistics were computed by fitting a single-SNP univariate linear model (via ordinary least squares). Results are shown comparing the $-\log_{10}$ transformed gene-level $P$-values derived by gene-$\varepsilon$ with Elastic Net (EN) regularization on the y-axis and without regularization (labeled as OLS) on the x-axis. The horizontal and vertical dashed lines are marked at the Bonferonni-corrected threshold $P = 2.61 \times 10^{-5}$ corrected for the 1,916 genes on chromosome 1 from the UK Biobank genotype data. True positive causal genes used to generate the synthetic phenotypes are colored in red, while non-causal genes are given in grey. Genes in the top right quadrant are selected by both approaches. Genes in the top left and bottom right quadrants are uniquely identified by gene-$\varepsilon$-EN and gene-$\varepsilon$-OLS, respectively. To illustrate the importance of regularization on SNP-level summary statistics, we highlight the true positive genes only identified by gene-$\varepsilon$-EN in blue. Each plot combines results from 100 simulated replicates (see Section A.1.2).
Figure A.23. Scatter plots assessing how regularization on SNP-level summary statistics affects the ability to identify enriched genes in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer and with population stratification ($h^2 = 0.2$). Here, the narrow-sense heritability of the simulated quantitative traits is $h^2 = 0.2$ and sample sizes are set to $N = 5,000$ in (A, B) and $N = 10,000$ in (C, D). In this simulation, traits were generated while using the top five principal components (PCs) of the genotype matrix as covariates. GWA summary statistics were computed by fitting a single-SNP univariate linear model (via ordinary least squares) without any control for the additional structure. Results are shown comparing the $-\log_{10}$ transformed gene-level $P$-values derived by gene-$\varepsilon$ with Elastic Net (EN) regularization on the y-axis and without regularization (labeled as OLS) on the x-axis. The horizontal and vertical dashed lines are marked at the Bonferonni-corrected threshold $P = 2.61 \times 10^{-5}$ corrected for the 1,916 genes on chromosome 1 from the UK Biobank genotype data. True positive causal genes used to generate the synthetic phenotypes are colored in red, while non-causal genes are given in grey. Genes in the top right quadrant are selected by both approaches. Genes in the top left and bottom right quadrants are uniquely identified by gene-$\varepsilon$-EN and gene-$\varepsilon$-OLS, respectively. To illustrate the importance of regularization on SNP-level summary statistics, we highlight the true positive genes only identified by gene-$\varepsilon$-EN in blue. Each plot combines results from 100 simulated replicates (see Section A.1.2).
Figure A.24. Scatter plots assessing how regularization on SNP-level summary statistics affects the ability to identify enriched genes in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer and with population stratification ($h^2 = 0.6$). Here, the narrow-sense heritability of the simulated quantitative traits is $h^2 = 0.6$ and sample sizes are set to $N = 5,000$ in (A, B) and $N = 10,000$ in (C, D). In this simulation, traits were generated while using the top five principal components (PCs) of the genotype matrix as covariates. GWA summary statistics were computed by fitting a single-SNP univariate linear model (via ordinary least squares) without any control for the additional structure. Results are shown comparing the -log$_{10}$ transformed gene-level $P$-values derived by gene-$\varepsilon$ with Elastic Net (EN) regularization on the y-axis and without regularization (labeled as OLS) on the x-axis. The horizontal and vertical dashed lines are marked at the Bonferonni-corrected threshold $P = 2.61 \times 10^{-5}$ corrected for the 1,916 genes on chromosome 1 from the UK Biobank genotype data. True positive causal genes used to generate the synthetic phenotypes are colored in red, while non-causal genes are given in grey. Genes in the top right quadrant are selected by both approaches. Genes in the top left and bottom right quadrants are uniquely identified by gene-$\varepsilon$-EN and gene-$\varepsilon$-OLS, respectively. To illustrate the importance of regularization on SNP-level summary statistics, we highlight the true positive genes only identified by gene-$\varepsilon$-EN in blue. Each plot combines results from 100 simulated replicates (see Section A.1.2).
Figure A.25. Gene-level association results from applying gene-ε to body height (panels A and C) and mean platelet volume (MPV; panels B and D), assayed in European-ancestry individuals in the UK Biobank with UCSC RefSeq gene boundaries augmented by a 50 kilobase (kb) buffer. Body height has been estimated to have a narrow-sense heritability $h^2$ in the range of 0.45 to 0.80 (288; 294; 279; 96; 218; 284; 78; 227; 167; 258); while, MPV has a narrow-sense heritability between 0.50 and 0.70 (198; 96; 218). Manhattan plots of gene-ε gene-level association $P$-values using Elastic Net regularized effect sizes for MPV. The purple dashed line indicates a log-transformed $P$ correcting for 17,680 autosomal genes analyzed). We color code all significant genes identified by gene-ε in orange, and annotate genes overlapping with the database of Genotypes and Phenotypes (dbGaP). In (C) and (D), we conduct gene set enrichment analysis using Enrichr (34; 129) to identify dbGaP categories enriched for significant gene-level associations reported by gene-ε. We highlight categories with Q-values (i.e., false discovery rates) less than 0.05 and annotate corresponding genes in the Manhattan plots in (A) and (B), respectively. For height, the most enriched dbGaP category is “Body Height”, with 5 of the genes identified by gene-ε appearing in this category. For MPV, the four significant dbGaP categories are “Platelet Count”, “Behcet Syndrome”, “Psoriasis”, and “Face” — all of which have been connected to trait (1; 26; 4; 142; 180).
Figure A.26. Gene-level association results from applying gene-ε to body mass index (BMI), assayed in European-ancestry individuals in the UK Biobank. BMI has been estimated to have a narrow-sense heritability $h^2$ ranging from 0.25 to 0.60 (253; 294; 290; 96; 218; 78; 204; 227; 205; 258). Manhattan plots of gene-ε gene-level association $P$-values using Elastic Net regularized effect sizes when gene boundaries are defined by (A) using UCSC annotations directly, and (B) augmenting the gene boundaries by adding SNPs within a ±50kb buffer. The purple dashed line indicates a log-transformed Bonferroni-corrected significance threshold ($P = 3.49 \times 10^{-6}$ and $P = 2.83 \times 10^{-6}$ correcting for the 14,322 and 17,680 autosomal genes analyzed, respectively). We color code all significant genes identified by gene-ε in orange, and annotate genes previously associated with BMI in the database of Genotypes and Phenotypes (dbGaP). In (C) and (D), we conduct gene set enrichment analysis using Enrichr (34; 129) to identify dbGaP categories enriched for significant gene-level associations reported by gene-ε in (A) and (B), respectively. While many of the scored categories are biologically related to BMI (e.g., “Body Mass Index”, “Adiposity”, and “Arteries”) (68; 130; 156; 226), none of them had $Q$-values (i.e., false discovery rates) less than 0.05.
Figure A.27. Gene-level association results from applying gene-ε to mean corpuscular volume (MCV), assayed in European-ancestry individuals in the UK Biobank. MCV has been estimated to have a narrow-sense heritability $h^2$ in the range of 0.20 to 0.60 (76; 252; 96; 218). Manhattan plots of gene-ε gene-level association $P$-values using Elastic Net regularized effect sizes when gene boundaries are defined by (A) using UCSC annotations directly, and (B) augmenting the gene boundaries by adding SNPs within a ±50kb buffer. The purple dashed line indicates a log-transformed Bonferroni-corrected significance threshold ($P$ = 3.49 × 10^{-6} and $P$ = 2.83 × 10^{-6} correcting for the 14,322 and 17,680 autosomal genes analyzed, respectively). We color code all significant genes identified by gene-ε in orange, and annotate genes previously associated with MCV in the database of Genotypes and Phenotypes (dbGaP). In (C) and (D), we conduct gene set enrichment analysis using Enrichr (34; 129) to identify dbGaP categories enriched for significant gene-level associations reported by gene-ε. We highlight categories with $Q$-values (i.e., false discovery rates) less than 0.05 and annotate corresponding genes in the Manhattan plots in (A) and (B), respectively. The dbGaP categories significantly enriched for gene-level associations with MCV included “Transferrin”, “Erythrocyte Indices”, “Hematocrit”, “Narcolepsy”, and “Iron” — all of which have been connected to trait (73; 3; 12; 33; 225; 75; 145; 4).
Figure A.28. Gene-level association results from applying gene-ε to platelet count (PLC), assayed in European-ancestry individuals in the UK Biobank. PLC has been estimated to have a narrow-sense heritability \( h^2 \) ranging from 0.55 to 0.80 (198; 96; 218). Manhattan plots of gene-ε gene-level association \( P \)-values using Elastic Net regularized effect sizes when gene boundaries are defined by (A) using UCSC annotations directly, and (B) augmenting the gene boundaries by adding SNPs within a ±50kb buffer. The purple dashed line indicates a log-transformed Bonferroni-corrected significance threshold (\( P = 3.49 \times 10^{-6} \) and \( P = 2.83 \times 10^{-5} \) correcting for the 14,322 and 17,680 autosomal genes analyzed, respectively). We color code all significant genes identified by gene-ε in orange, and annotate genes previously associated with PLC in the database of Genotypes and Phenotypes (dbGaP). In (C) and (D), we conduct gene set enrichment analysis using Enrichr (34; 129) to identify dbGaP categories enriched for significant gene-level associations reported by gene-ε. We highlight categories with \( Q \)-values (i.e., false discovery rates) less than 0.05 and annotate corresponding genes in the Manhattan plots in (A) and (B), respectively. The most significant dbGaP category is “Platelet Count” for both SNP-to-gene annotation schemes. The other significant dbGaP category was “Smoking” which has been previously connected to PLC (92; 116; 180).
Figure A.29. Gene-level association results from applying gene-$\varepsilon$ to waist-hip ratio (WHR), assayed in European-ancestry individuals in the UK Biobank. WHR has been estimated to have a narrow-sense heritability $h^2$ ranging from 0.10 to 0.25 (253; 294; 96; 284; 157; 205). Manhattan plots of gene-$\varepsilon$ gene-level association $P$-values using Elastic Net regularized effect sizes when gene boundaries are defined by (A) using UCSC annotations directly, and (B) augmenting the gene boundaries by adding SNPs within a ±50kb buffer. The purple dashed line indicates a log-transformed Bonferroni-corrected significance threshold ($P = 3.49 \times 10^{-6}$ and $P = 2.83 \times 10^{-6}$ correcting for the 14,322 and 17,680 autosomal genes analyzed, respectively). We color code all significant genes identified by gene-$\varepsilon$ in orange, and annotate genes previously associated with WHR in the database of Genotypes and Phenotypes (dbGaP). In (C) and (D), we conduct gene set enrichment analysis using Enrichr (34; 129) to identify dbGaP categories enriched for significant gene-level associations reported by gene-$\varepsilon$ in (A) and (B), respectively. While many of the scored categories are biologically related to WHR (e.g., “Body Mass Index”, “Adiposity”, and “Inflammatory Bowel Diseases”) (219; 66), none of them had $Q$-values (i.e., false discovery rates) less than 0.05.
### A.3 Supplementary Tables

<table>
<thead>
<tr>
<th>Causal Genes</th>
<th>Metric</th>
<th>gene-ε Approaches</th>
<th>Other Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>OLS</td>
<td>RR</td>
</tr>
<tr>
<td>1%</td>
<td>Power</td>
<td>0.572 (0.149)</td>
<td>0.358 (0.131)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.415 (0.164)</td>
<td>0.652 (0.314)</td>
</tr>
<tr>
<td>10%</td>
<td>Power</td>
<td><strong>0.072 (0.102)</strong></td>
<td>0.044 (0.026)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.187 (0.184)</td>
<td>0.468 (0.310)</td>
</tr>
</tbody>
</table>

Table A.1. Empirical power and false discovery rates (FDR) for detecting enriched genes (genes containing at least one causal SNP) after correcting for multiple hypothesis testing in simulations ($N = 5,000; h^2 = 0.2$). We computed standard GWA SNP-level effect sizes (estimated using ordinary least squares) as input to each method listed. We show the power of gene-ε to identify enriched genes under the Bonferonni-corrected threshold $P = 3.55 \times 10^{-5}$, corrected for 1,408 genes simulated using chromosome 1 from the UK Biobank genotype data (see Section A.1.2). Results for gene-ε are shown with LASSO, Elastic Net (EN), and Ridge Regression (RR) regularizations. We also show the power of gene-ε without regularization to illustrate the importance of this step (OLS). Additionally, we compare the performance gene-ε with five existing methods: PEGASUS (183), VEGAS (154), RSS (307), SKAT (281), and MAGMA (56). The last is a Bayesian method and is evaluated based on the “median probability criterion” (i.e., posterior enrichment probability of a gene is greater than 0.5). All results are based on 100 replicates and standard deviations of the estimates across runs are given in the parentheses. Approaches with the greatest power are bolded in purple, while methods with the lowest FDR is bolded in blue.
<table>
<thead>
<tr>
<th>Causal Genes</th>
<th>Metric</th>
<th>OLS</th>
<th>RR</th>
<th>EN</th>
<th>LASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>Power</td>
<td>0.686 (0.130)</td>
<td>0.497 (0.120)</td>
<td>0.404 (0.125)</td>
<td>0.406 (0.130)</td>
</tr>
<tr>
<td>FDR</td>
<td>0.512 (0.146)</td>
<td>0.356 (0.207)</td>
<td>0.012 (0.055)</td>
<td><strong>0.005 (0.036)</strong></td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td>Power</td>
<td>0.131 (0.031)</td>
<td><strong>0.155 (0.045)</strong></td>
<td>0.018 (0.013)</td>
<td>0.021 (0.016)</td>
</tr>
<tr>
<td>FDR</td>
<td>0.169 (0.082)</td>
<td>0.678 (0.150)</td>
<td>0.255 (0.281)</td>
<td>0.271 (0.306)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Causal Genes</th>
<th>Metric</th>
<th>PEGASUS</th>
<th>VEGAS</th>
<th>RSS</th>
<th>SKAT</th>
<th>MAGMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>Power</td>
<td>0.678 (0.119)</td>
<td><strong>0.742 (0.123)</strong></td>
<td>0.609 (0.119)</td>
<td>0.569 (0.130)</td>
<td>0.737 (0.123)</td>
</tr>
<tr>
<td>FDR</td>
<td>0.527 (0.143)</td>
<td>0.536 (0.131)</td>
<td>0.088 (0.102)</td>
<td>0.487 (0.153)</td>
<td>0.571 (0.126)</td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td>Power</td>
<td>0.139 (0.027)</td>
<td>0.106 (0.026)</td>
<td>0.060 (0.017)</td>
<td>0.105 (0.021)</td>
<td>0.152 (0.027)</td>
</tr>
<tr>
<td>FDR</td>
<td>0.169 (0.080)</td>
<td>0.232 (0.098)</td>
<td><strong>0.065 (0.089)</strong></td>
<td>0.172 (0.088)</td>
<td>0.191 (0.075)</td>
<td></td>
</tr>
</tbody>
</table>

Table A.2. Empirical power and false discovery rates (FDR) for detecting enriched genes (genes containing at least one causal SNP) after correcting for multiple hypothesis testing in simulations ($N = 10,000; h^2 = 0.2$). We computed standard GWA SNP-level effect sizes (estimated using ordinary least squares) as input to each method listed. We show the power of gene-$\varepsilon$ to identify enriched genes under the Bonferroni-corrected threshold $P = 3.55 \times 10^{-5}$, corrected for 1,408 genes simulated using chromosome 1 from the UK Biobank genotype data (see Section A.1.2). Results for gene-$\varepsilon$ are shown with LASSO, Elastic Net (EN), and Ridge Regression (RR) regularizations. We also show the power of gene-$\varepsilon$ without regularization to illustrate the importance of this step (OLS). Additionally, we compare the performance gene-$\varepsilon$ with five existing methods: PEGASUS (183), VEGAS (154), RSS (307), SKAT (281), and MAGMA (56). The last is a Bayesian method and is evaluated based on the “median probability criterion” (i.e., posterior enrichment probability of a gene is greater than 0.5). All results are based on 100 replicates and standard deviations of the estimates across runs are given in the parentheses. Approaches with the greatest power are bolded in purple, while methods with the lowest FDR is bolded in blue.
Table A.3. Empirical power and false discovery rates (FDR) for detecting enriched genes (genes containing at least one causal SNP) after correcting for multiple hypothesis testing in simulations ($N = 5,000; h^2 = 0.6$). We computed standard GWA SNP-level effect sizes (estimated using ordinary least squares) as input to each method listed. We show the power of gene-$\varepsilon$ to identify enriched genes under the Bonferroni-corrected threshold $P = 3.55 \times 10^{-5}$, corrected for 1,408 genes simulated using chromosome 1 from the UK Biobank genotype data (see Section A.1.2). Results for gene-$\varepsilon$ are shown with LASSO, Elastic Net (EN), and Ridge Regression (RR) regularizations. We also show the power of gene-$\varepsilon$ without regularization to illustrate the importance of this step (OLS). Additionally, we compare the performance gene-$\varepsilon$ with five existing methods: PEGASUS (183), VEGAS (154), RSS (307), SKAT (281), and MAGMA (56). The last is a Bayesian method and is evaluated based on the “median probability criterion” (i.e., posterior enrichment probability of a gene is greater than 0.5). All results are based on 100 replicates and standard deviations of the estimates across runs are given in the parentheses. Approaches with the greatest power are bolded in purple, while methods with the lowest FDR is bolded in blue.
Table A.4. Empirical power and false discovery rates (FDR) for detecting enriched genes (genes containing at least one causal SNP) after correcting for multiple hypothesis testing in simulations ($N = 10,000; h^2 = 0.6$). We computed standard GWA SNP-level effect sizes (estimated using ordinary least squares) as input to each method listed. We show the power of gene-$\varepsilon$ to identify enriched genes under the Bonferroni-corrected threshold $P = 3.55 \times 10^{-5}$, corrected for 1,408 genes simulated using chromosome 1 from the UK Biobank genotype data (see Section A.1.2). Results for gene-$\varepsilon$ are shown with LASSO, Elastic Net (EN), and Ridge Regression (RR) regularizations. We also show the power of gene-$\varepsilon$ without regularization to illustrate the importance of this step (OLS). Additionally, we compare the performance gene-$\varepsilon$ with five existing methods: PEGASUS (183), VEGAS (154), RSS (307), SKAT (281), and MAGMA (56). The last is a Bayesian method and is evaluated based on the “median probability criterion” (i.e., posterior enrichment probability of a gene is greater than 0.5). All results are based on 100 replicates and standard deviations of the estimates across runs are given in the parentheses. Approaches with the greatest power are bolded in purple, while methods with the lowest FDR is bolded in blue.
<table>
<thead>
<tr>
<th>Causal Genes</th>
<th>Metric</th>
<th>gene-ε Approaches</th>
<th>OLS</th>
<th>RR</th>
<th>EN</th>
<th>LASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>Power</td>
<td>0.524 (0.136)</td>
<td>0.340 (0.134)</td>
<td>0.324 (0.119)</td>
<td>0.339 (0.127)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.773 (0.102)</td>
<td>0.467 (0.325)</td>
<td><strong>0.004 (0.032)</strong></td>
<td>0.014 (0.058)</td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td>Power</td>
<td>0.044 (0.024)</td>
<td>0.053 (0.017)</td>
<td>0.013 (0.019)</td>
<td>0.011 (0.015)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.715 (0.136)</td>
<td>0.560 (0.179)</td>
<td><strong>0.434 (0.388)</strong></td>
<td>0.463 (0.395)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Causal Genes</th>
<th>Metric</th>
<th>Other Methods</th>
<th>PEGASUS</th>
<th>VEGAS</th>
<th>RSS</th>
<th>SKAT</th>
<th>MAGMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>Power</td>
<td>0.558 (0.126)</td>
<td>0.611 (0.134)</td>
<td>0.490 (0.137)</td>
<td>0.467 (0.132)</td>
<td><strong>0.615 (0.140)</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.780 (0.096)</td>
<td>0.778 (0.090)</td>
<td>0.334 (0.147)</td>
<td>0.783 (0.092)</td>
<td>0.789 (0.091)</td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td>Power</td>
<td><strong>0.064 (0.023)</strong></td>
<td>0.044 (0.019)</td>
<td>0.014 (0.010)</td>
<td>0.051 (0.017)</td>
<td>0.063 (0.021)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.661 (0.129)</td>
<td>0.749 (0.114)</td>
<td>0.563 (0.257)</td>
<td>0.679 (0.119)</td>
<td>0.683 (0.135)</td>
<td></td>
</tr>
</tbody>
</table>

Table A.5. Empirical power and false discovery rates (FDR) for detecting enriched genes (genes containing at least one causal SNP) after correcting for multiple hypothesis testing in simulations with population stratification ($N = 5,000; h^2 = 0.2$). In this simulation, traits were generated while using the top five principal components (PCs) of the genotype matrix as covariates. GWA summary statistics were computed by fitting a single-SNP univariate linear model (via ordinary least squares) without any control for the additional structure. We show the power of gene-ε to identify enriched genes under the Bonferonni-corrected threshold $P = 3.55 \times 10^{-5}$, corrected for 1,408 genes simulated using chromosome 1 from the UK Biobank genotype data (see Section A.1.2). Results for gene-ε are shown with LASSO, Elastic Net (EN), and Ridge Regression (RR) regularizations. We also show the power of gene-ε without regularization to illustrate the importance of this step (OLS). Additionally, we compare the performance of gene-ε with five existing methods: PEGASUS (183), VEGAS (154), RSS (307), SKAT (281), and MAGMA (56). The last is a Bayesian method and is evaluated based on the “median probability criterion” (i.e., posterior enrichment probability of a gene is greater than 0.5). All results are based on 100 replicates and standard deviations of the estimates across runs are given in the parentheses. Approaches with the greatest power are bolded in purple, while methods with the lowest FDR is bolded in blue.
Table A.6. Empirical power and false discovery rates (FDR) for detecting enriched genes (genes containing at least one causal SNP) after correcting for multiple hypothesis testing in simulations with population stratification ($N = 10,000$; $h^2 = 0.2$). In this simulation, traits were generated while using the top five principal components (PCs) of the genotype matrix as covariates. GWA summary statistics were computed by fitting a single-SNP univariate linear model (via ordinary least squares) without any control for the additional structure. We show the power of gene-$\varepsilon$ to identify enriched genes under the Bonferroni-corrected threshold $P = 3.55 \times 10^{-5}$, corrected for 1,408 genes simulated using chromosome 1 from the UK Biobank genotype data (see Section A.1.2). Results for gene-$\varepsilon$ are shown with LASSO, Elastic Net (EN), and Ridge Regression (RR) regularizations. We also show the power of gene-$\varepsilon$ without regularization to illustrate the importance of this step (OLS). Additionally, we compare the performance gene-$\varepsilon$ with five existing methods: PEGASUS (183), VEGAS (154), RSS (307), SKAT (281), and MAGMA (56). The last is a Bayesian method and is evaluated based on the “median probability criterion” (i.e., posterior enrichment probability of a gene is greater than 0.5). All results are based on 100 replicates and standard deviations of the estimates across runs are given in the parentheses. Approaches with the greatest power are bolded in purple, while methods with the lowest FDR is bolded in blue.

<table>
<thead>
<tr>
<th>Causal Genes</th>
<th>Metric</th>
<th>OLS</th>
<th>RR</th>
<th>EN</th>
<th>LASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>Power</td>
<td>0.658 (0.134)</td>
<td>0.506 (0.123)</td>
<td>0.471 (0.130)</td>
<td>0.487 (0.136)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.763 (0.093)</td>
<td>0.233 (0.189)</td>
<td><strong>0.005 (0.031)</strong></td>
<td>0.020 (0.060)</td>
</tr>
<tr>
<td>10%</td>
<td>Power</td>
<td>0.086 (0.050)</td>
<td>0.120 (0.038)</td>
<td>0.024 (0.021)</td>
<td>0.021 (0.024)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.618 (0.144)</td>
<td>0.617 (0.188)</td>
<td><strong>0.236 (0.269)</strong></td>
<td>0.253 (0.311)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Causal Genes</th>
<th>Metric</th>
<th>PEGASUS</th>
<th>VEGAS</th>
<th>RSS</th>
<th>SKAT</th>
<th>MAGMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>Power</td>
<td>0.694 (0.116)</td>
<td>0.756 (0.121)</td>
<td>0.611 (0.119)</td>
<td>0.583 (0.120)</td>
<td><strong>0.760 (0.117)</strong></td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.793 (0.072)</td>
<td>0.792 (0.068)</td>
<td>0.325 (0.116)</td>
<td>0.775 (0.084)</td>
<td>0.807 (0.064)</td>
</tr>
<tr>
<td>10%</td>
<td>Power</td>
<td><strong>0.154 (0.029)</strong></td>
<td>0.126 (0.029)</td>
<td>0.061 (0.017)</td>
<td>0.121 (0.024)</td>
<td>0.176 (0.031)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.536 (0.117)</td>
<td>0.605 (0.104)</td>
<td>0.273 (0.109)</td>
<td>0.534 (0.122)</td>
<td>0.549 (0.102)</td>
</tr>
</tbody>
</table>
Table A.7. Empirical power and false discovery rates (FDR) for detecting enriched genes (genes containing at least one causal SNP) after correcting for multiple hypothesis testing in simulations with population stratification \((N = 5,000; h^2 = 0.6)\). In this simulation, traits were generated while using the top five principal components (PCs) of the genotype matrix as covariates. GWA summary statistics were computed by fitting a single-SNP univariate linear model (via ordinary least squares) without any control for the additional structure. We show the power of gene-\(\varepsilon\) to identify enriched genes under the Bonferroni-corrected threshold \(P = 3.55 \times 10^{-5}\), corrected for 1,408 genes simulated using chromosome 1 from the UK Biobank genotype data (see Section A.1.2). Results for gene-\(\varepsilon\) are shown with LASSO, Elastic Net (EN), and Ridge Regression (RR) regularizations. We also show the power of gene-\(\varepsilon\) without regularization to illustrate the importance of this step (OLS). Additionally, we compare the performance gene-\(\varepsilon\) with five existing methods: PEGASUS (183), VEGAS (154), RSS (307), SKAT (281), and MAGMA (56). The last is a Bayesian method and is evaluated based on the “median probability criterion” (i.e., posterior enrichment probability of a gene is greater than 0.5). All results are based on 100 replicates and standard deviations of the estimates across runs are given in the parentheses. Approaches with the greatest power are bolded in purple, while methods with the lowest FDR is bolded in blue.
Table A.8. Empirical power and false discovery rates (FDR) for detecting enriched genes (genes containing at least one causal SNP) after correcting for multiple hypothesis testing in simulations with population stratification ($N = 10,000; h^2 = 0.6$). In this simulation, traits were generated while using the top five principal components (PCs) of the genotype matrix as covariates. GWA summary statistics were computed by fitting a single-SNP univariate linear model (via ordinary least squares) without any control for the additional structure. We show the power of gene-ε to identify enriched genes under the Bonferroni-corrected threshold $P = 3.55 \times 10^{-5}$, corrected for 1,408 genes simulated using chromosome 1 from the UK Biobank genotype data (see Section A.1.2). Results for gene-ε are shown with LASSO, Elastic Net (EN), and Ridge Regression (RR) regularizations. We also show the power of gene-ε without regularization to illustrate the importance of this step (OLS). Additionally, we compare the performance of gene-ε with five existing methods: PEGASUS (183), VEGAS (154), RSS (307), SKAT (281), and MAGMA (56). The last is a Bayesian method and is evaluated based on the “median probability criterion” (i.e., posterior enrichment probability of a gene is greater than 0.5). All results are based on 100 replicates and standard deviations of the estimates across runs are given in the parentheses. Approaches with the greatest power are bolded in purple, while methods with the lowest FDR is bolded in blue.
<table>
<thead>
<tr>
<th>Causal Genes</th>
<th>Metric</th>
<th>OLS</th>
<th>RR</th>
<th>EN</th>
<th>LASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>Power</td>
<td>0.487 (0.118)</td>
<td>0.482 (0.167)</td>
<td>0.126 (0.063)</td>
<td>0.136 (0.063)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.407 (0.178)</td>
<td>0.770 (0.259)</td>
<td><strong>0.000 (0.000)</strong></td>
<td>0.006 (0.054)</td>
</tr>
<tr>
<td>10%</td>
<td>Power</td>
<td>0.019 (0.029)</td>
<td><strong>0.070 (0.055)</strong></td>
<td>0.005 (0.005)</td>
<td>0.006 (0.005)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.077 (0.185)</td>
<td>0.622 (0.208)</td>
<td>0.581 (0.352)</td>
<td>0.447 (0.334)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Causal Genes</th>
<th>Metric</th>
<th>PEGASUS</th>
<th>VEGAS</th>
<th>RSS</th>
<th>SKAT</th>
<th>MAGMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>Power</td>
<td>0.457 (0.110)</td>
<td><strong>0.563 (0.104)</strong></td>
<td>0.540 (0.110)</td>
<td>0.388 (0.106)</td>
<td>0.514 (0.124)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.366 (0.192)</td>
<td>0.393 (0.177)</td>
<td>0.090 (0.124)</td>
<td>0.329 (0.212)</td>
<td>0.430 (0.183)</td>
</tr>
<tr>
<td>10%</td>
<td>Power</td>
<td>0.014 (0.007)</td>
<td>0.008 (0.004)</td>
<td>0.006 (0.002)</td>
<td>0.012 (0.006)</td>
<td>0.010 (0.006)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.060 (0.158)</td>
<td>0.178 (0.290)</td>
<td><strong>0.016 (0.088)</strong></td>
<td>0.073 (0.177)</td>
<td>0.121 (0.237)</td>
</tr>
</tbody>
</table>

Table A.9. Empirical power and false discovery rates (FDR) for detecting enriched genes (genes containing at least one causal SNP) after correcting for multiple hypothesis testing in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer ($N = 5,000; h^2 = 0.2$). We computed standard GWA SNP-level effect sizes (estimated using ordinary least squares) as input to each method listed. We show the power of gene-$\epsilon$ to identify enriched genes under the Bonferroni-corrected threshold $P = 2.61 \times 10^{-5}$, corrected for 1,916 genes simulated using chromosome 1 from the UK Biobank genotype data (see Section A.1.2). Results for gene-$\epsilon$ are shown with LASSO, Elastic Net (EN), and Ridge Regression (RR) regularizations. We also show the power of gene-$\epsilon$ without regularization to illustrate the importance of this step (OLS). Additionally, we compare the performance gene-$\epsilon$ with five existing methods: PEGASUS (183), VEGAS (154), RSS (307), SKAT (281), and MAGMA (56). The last is a Bayesian method and is evaluated based on the “median probability criterion” (i.e., posterior enrichment probability of a gene is greater than 0.5). All results are based on 100 replicates and standard deviations of the estimates across runs are given in the parentheses. Approaches with the greatest power are bolded in purple, while methods with the lowest FDR is bolded in blue.
<table>
<thead>
<tr>
<th>Causal Genes</th>
<th>Metric</th>
<th>OLS</th>
<th>RR</th>
<th>EN</th>
<th>LASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1% Power</td>
<td>0.602 (0.115)</td>
<td>0.388 (0.119)</td>
<td>0.220 (0.086)</td>
<td>0.226 (0.088)</td>
<td></td>
</tr>
<tr>
<td>FDR</td>
<td>0.523 (0.151)</td>
<td>0.229 (0.230)</td>
<td>0.000 (0.000)</td>
<td>0.000 (0.000)</td>
<td></td>
</tr>
<tr>
<td>10% Power</td>
<td>0.106 (0.055)</td>
<td>0.221 (0.057)</td>
<td>0.011 (0.019)</td>
<td>0.004 (0.004)</td>
<td></td>
</tr>
<tr>
<td>FDR</td>
<td>0.183 (0.131)</td>
<td>0.746 (0.094)</td>
<td>0.374 (0.398)</td>
<td>0.396 (0.494)</td>
<td></td>
</tr>
</tbody>
</table>

Table A.10. Empirical power and false discovery rates (FDR) for detecting enriched genes (genes containing at least one causal SNP) after correcting for multiple hypothesis testing in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer (N = 10,000; h^2 = 0.2). We computed standard GWA SNP-level effect sizes (estimated using ordinary least squares) as input to each method listed. We show the power of gene-ε to identify enriched genes under the Bonferroni-corrected threshold P = 2.61×10^{-5}, corrected for 1,916 genes simulated using chromosome 1 from the UK Biobank genotype data (see Section A.1.2). Results for gene-ε are shown with LASSO, Elastic Net (EN), and Ridge Regression (RR) regularizations. We also show the power of gene-ε without regularization to illustrate the importance of this step (OLS). Additionally, we compare the performance gene-ε with five existing methods: PEGASUS (183), VEGAS (154), RSS (307), SKAT (281), and MAGMA (56). The last is a Bayesian method and is evaluated based on the “median probability criterion” (i.e., posterior enrichment probability of a gene is greater than 0.5). All results are based on 100 replicates and standard deviations of the estimates across runs are given in the parentheses. Approaches with the greatest power are bolded in purple, while methods with the lowest FDR is bolded in blue.
<table>
<thead>
<tr>
<th>Causal Genes</th>
<th>Metric</th>
<th>OLS</th>
<th>RR</th>
<th>EN</th>
<th>LASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>Power</td>
<td>0.707 (0.111)</td>
<td>0.567 (0.107)</td>
<td>0.624 (0.115)</td>
<td>0.638 (0.123)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.591 (0.132)</td>
<td>0.260 (0.161)</td>
<td>0.004 (0.022)</td>
<td>0.006 (0.030)</td>
</tr>
<tr>
<td>10%</td>
<td>Power</td>
<td>0.101 (0.030)</td>
<td><strong>0.172 (0.060)</strong></td>
<td>0.013 (0.012)</td>
<td>0.010 (0.007)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.119 (0.101)</td>
<td>0.365 (0.160)</td>
<td>0.060 (0.192)</td>
<td><strong>0.056 (0.183)</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Causal Genes</th>
<th>Metric</th>
<th>PEGASUS</th>
<th>VEGAS</th>
<th>RSS</th>
<th>SKAT</th>
<th>MAGMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>Power</td>
<td>0.719 (0.108)</td>
<td><strong>0.787 (0.106)</strong></td>
<td>0.744 (0.107)</td>
<td>0.636 (0.114)</td>
<td>0.753 (0.109)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.605 (0.127)</td>
<td>0.626 (0.116)</td>
<td>0.116 (0.126)</td>
<td>0.575 (0.142)</td>
<td>0.677 (0.108)</td>
</tr>
<tr>
<td>10%</td>
<td>Power</td>
<td>0.117 (0.019)</td>
<td>0.065 (0.015)</td>
<td>0.054 (0.012)</td>
<td>0.093 (0.016)</td>
<td>0.118 (0.017)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.132 (0.101)</td>
<td>0.226 (0.154)</td>
<td>0.091 (0.128)</td>
<td>0.116 (0.106)</td>
<td>0.162 (0.106)</td>
</tr>
</tbody>
</table>

Table A.11. Empirical power and false discovery rates (FDR) for detecting enriched genes (genes containing at least one causal SNP) after correcting for multiple hypothesis testing in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer ($N = 5,000; h^2 = 0.6$). We computed standard GWA SNP-level effect sizes (estimated using ordinary least squares) as input to each method listed. We show the power of gene-$\varepsilon$ to identify enriched genes under the Bonferroni-corrected threshold $P = 2.61 \times 10^{-5}$, corrected for 1,916 genes simulated using chromosome 1 from the UK Biobank genotype data (see Section A.1.2). Results for gene-$\varepsilon$ are shown with LASSO, Elastic Net (EN), and Ridge Regression (RR) regularizations. We also show the power of gene-$\varepsilon$ without regularization to illustrate the importance of this step (OLS). Additionally, we compare the performance gene-$\varepsilon$ with five existing methods: PEGASUS (183), VEGAS (154), RSS (307), SKAT (281), and MAGMA (56). The last is a Bayesian method and is evaluated based on the “median probability criterion” (i.e., posterior enrichment probability of a gene is greater than 0.5). All results are based on 100 replicates and standard deviations of the estimates across runs are given in the parentheses. Approaches with the greatest power are bolded in purple, while methods with the lowest FDR is bolded in blue.
<table>
<thead>
<tr>
<th>Causal Genes</th>
<th>Metric</th>
<th>OLS</th>
<th>RR</th>
<th>EN</th>
<th>LASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>Power</td>
<td>0.784 (0.103)</td>
<td>0.617 (0.114)</td>
<td>0.667 (0.115)</td>
<td>0.673 (0.118)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.691 (0.096)</td>
<td>0.219 (0.157)</td>
<td><strong>0.006 (0.023)</strong></td>
<td>0.007 (0.028)</td>
</tr>
<tr>
<td>10%</td>
<td>Power</td>
<td>0.224 (0.048)</td>
<td>0.100 (0.083)</td>
<td>0.024 (0.015)</td>
<td>0.020 (0.012)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.199 (0.088)</td>
<td>0.053 (0.134)</td>
<td><strong>0.007 (0.040)</strong></td>
<td>0.008 (0.041)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Causal Genes</th>
<th>Metric</th>
<th>PEGASUS</th>
<th>VEGAS</th>
<th>RSS</th>
<th>SKAT</th>
<th>MAGMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>Power</td>
<td>0.799 (0.099)</td>
<td><strong>0.843 (0.087)</strong></td>
<td>0.820 (0.087)</td>
<td>0.733 (0.104)</td>
<td>0.826 (0.088)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.705 (0.095)</td>
<td>0.726 (0.078)</td>
<td>0.118 (0.113)</td>
<td>0.680 (0.098)</td>
<td>0.766 (0.071)</td>
</tr>
<tr>
<td>10%</td>
<td>Power</td>
<td>0.261 (0.027)</td>
<td>0.213 (0.026)</td>
<td>0.159 (0.018)</td>
<td>0.217 (0.026)</td>
<td><strong>0.304 (0.028)</strong></td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.229 (0.082)</td>
<td>0.278 (0.094)</td>
<td>0.089 (0.066)</td>
<td>0.204 (0.082)</td>
<td>0.260 (0.081)</td>
</tr>
</tbody>
</table>

Table A.12. Empirical power and false discovery rates (FDR) for detecting enriched genes (genes containing at least one causal SNP) after correcting for multiple hypothesis testing in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer ($N = 10,000; h^2 = 0.6$). We computed standard GWA SNP-level effect sizes (estimated using ordinary least squares) as input to each method listed. We show the power of gene-$\varepsilon$ to identify enriched genes under the Bonferonni-corrected threshold $P = 2.61 \times 10^{-5}$, corrected for 1,916 genes simulated using chromosome 1 from the UK Biobank genotype data (see Section A.1.2). Results for gene-$\varepsilon$ are shown with LASSO, Elastic Net (EN), and Ridge Regression (RR) regularizations. We also show the power of gene-$\varepsilon$ without regularization to illustrate the importance of this step (OLS). Additionally, we compare the performance gene-$\varepsilon$ with five existing methods: PEGASUS (183), VEGAS (154), RSS (307), SKAT (281), and MAGMA (56). The last is a Bayesian method and is evaluated based on the “median probability criterion” (i.e., posterior enrichment probability of a gene is greater than 0.5). All results are based on 100 replicates and standard deviations of the estimates across runs are given in the parentheses. Approaches with the greatest power are bolded in purple, while methods with the lowest FDR is bolded in blue.
gene-ε Approaches

<table>
<thead>
<tr>
<th>Causal Genes</th>
<th>Metric</th>
<th>OLS</th>
<th>RR</th>
<th>EN</th>
<th>LASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>Power</td>
<td>0.431 (0.117)</td>
<td>0.427 (0.137)</td>
<td>0.155 (0.073)</td>
<td>0.158 (0.073)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.838 (0.128)</td>
<td>0.577 (0.286)</td>
<td><strong>0.004 (0.035)</strong></td>
<td>0.014 (0.081)</td>
</tr>
<tr>
<td>10%</td>
<td>Power</td>
<td>0.026 (0.089)</td>
<td><strong>0.079 (0.030)</strong></td>
<td>0.027 (0.035)</td>
<td>0.009 (0.021)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.903 (0.091)</td>
<td><strong>0.571 (0.125)</strong></td>
<td>0.624 (0.392)</td>
<td>0.705 (0.430)</td>
</tr>
</tbody>
</table>

Other Methods

<table>
<thead>
<tr>
<th>Causal Genes</th>
<th>Metric</th>
<th>PEGASUS</th>
<th>VEGAS</th>
<th>RSS</th>
<th>SKAT</th>
<th>MAGMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>Power</td>
<td>0.463 (0.116)</td>
<td><strong>0.556 (0.117)</strong></td>
<td>0.520 (0.110)</td>
<td>0.387 (0.110)</td>
<td>0.500 (0.110)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.847 (0.115)</td>
<td>0.828 (0.114)</td>
<td>0.608 (0.133)</td>
<td>0.858 (0.115)</td>
<td>0.846 (0.105)</td>
</tr>
<tr>
<td>10%</td>
<td>Power</td>
<td>0.027 (0.013)</td>
<td>0.016 (0.011)</td>
<td>0.006 (0.006)</td>
<td>0.022 (0.012)</td>
<td>0.020 (0.011)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.868 (0.094)</td>
<td>0.921 (0.081)</td>
<td>0.898 (0.103)</td>
<td>0.886 (0.091)</td>
<td>0.896 (0.091)</td>
</tr>
</tbody>
</table>

Table A.13. Empirical power and false discovery rates (FDR) for detecting enriched genes (genes containing at least one causal SNP) after correcting for multiple hypothesis testing in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer and with population stratification \((N = 5,000; \ h^2 = 0.2)\). In this simulation, traits were generated while using the top five principal components (PCs) of the genotype matrix as covariates. GWA summary statistics were computed by fitting a single-SNP univariate linear model (via ordinary least squares) without any control for the additional structure. We show the power of gene-ε to identify enriched genes under the Bonferonni-corrected threshold \(P = 2.61 \times 10^{-5}\), corrected for 1,916 genes simulated using chromosome 1 from the UK Biobank genotype data (see Section A.1.2). Results for gene-ε are shown with LASSO, Elastic Net (EN), and Ridge Regression (RR) regularizations. We also show the power of gene-ε without regularization to illustrate the importance of this step (OLS). Additionally, we compare the performance gene-ε with five existing methods: PEGASUS (183), VEGAS (154), RSS (307), SKAT (281), and MAGMA (56). The last is a Bayesian method and is evaluated based on the “median probability criterion” (i.e., posterior enrichment probability of a gene is greater than 0.5). All results are based on 100 replicates and standard deviations of the estimates across runs are given in the parentheses. Approaches with the greatest power are bolded in purple, while methods with the lowest FDR is bolded in blue.
Table A.14. Empirical power and false discovery rates (FDR) for detecting enriched genes (genes containing at least one causal SNP) after correcting for multiple hypothesis testing in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer and with population stratification ($N = 10,000; h^2 = 0.2$). In this simulation, traits were generated while using the top five principal components (PCs) of the genotype matrix as covariates. GWA summary statistics were computed by fitting a single-SNP univariate linear model (via ordinary least squares) without any control for the additional structure. We show the power of gene-ε to identify enriched genes under the Bonferroni-corrected threshold $P = 2.61 \times 10^{-5}$, corrected for 1,916 genes simulated using chromosome 1 from the UK Biobank genotype data (see Section A.1.2). Results for gene-ε are shown with LASSO, Elastic Net (EN), and Ridge Regression (RR) regularizations. We also show the power of gene-ε without regularization to illustrate the importance of this step (OLS). Additionally, we compare the performance gene-ε with five existing methods: PEGASUS (183), VEGAS (154), RSS (307), SKAT (281), and MAGMA (56). The last is a Bayesian method and is evaluated based on the “median probability criterion” (i.e., posterior enrichment probability of a gene is greater than 0.5). All results are based on 100 replicates and standard deviations of the estimates across runs are given in the parentheses. Approaches with the greatest power are bolded in purple, while methods with the lowest FDR is bolded in blue.
### gene-ε Approaches

<table>
<thead>
<tr>
<th>Causal Genes</th>
<th>Metric</th>
<th>OLS</th>
<th>RR</th>
<th>EN</th>
<th>LASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1% Power</td>
<td>0.687 (0.119)</td>
<td>0.636 (0.116)</td>
<td>0.696 (0.122)</td>
<td>0.700 (0.114)</td>
<td></td>
</tr>
<tr>
<td>1% FDR</td>
<td>0.820 (0.097)</td>
<td>0.324 (0.162)</td>
<td>0.011 (0.043)</td>
<td>0.019 (0.052)</td>
<td></td>
</tr>
<tr>
<td>10% Power</td>
<td>0.069 (0.033)</td>
<td><strong>0.155 (0.030)</strong></td>
<td>0.008 (0.004)</td>
<td>0.006 (0.003)</td>
<td></td>
</tr>
<tr>
<td>10% FDR</td>
<td>0.702 (0.171)</td>
<td>0.269 (0.112)</td>
<td><strong>0.042 (0.134)</strong></td>
<td>0.207 (0.361)</td>
<td></td>
</tr>
</tbody>
</table>

### Other Methods

<table>
<thead>
<tr>
<th>Causal Genes</th>
<th>Metric</th>
<th>PEGASUS</th>
<th>VEGAS</th>
<th>RSS</th>
<th>SKAT</th>
<th>MAGMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1% Power</td>
<td>0.712 (0.108)</td>
<td><strong>0.781 (0.101)</strong></td>
<td>0.745 (0.113)</td>
<td>0.649 (0.113)</td>
<td>0.751 (0.111)</td>
<td></td>
</tr>
<tr>
<td>1% FDR</td>
<td>0.847 (0.076)</td>
<td>0.837 (0.074)</td>
<td>0.513 (0.112)</td>
<td>0.840 (0.083)</td>
<td>0.849 (0.069)</td>
<td></td>
</tr>
<tr>
<td>10% Power</td>
<td>0.133 (0.019)</td>
<td>0.077 (0.017)</td>
<td>0.053 (0.013)</td>
<td>0.109 (0.017)</td>
<td>0.137 (0.023)</td>
<td></td>
</tr>
<tr>
<td>10% FDR</td>
<td>0.614 (0.179)</td>
<td>0.719 (0.149)</td>
<td>0.530 (0.129)</td>
<td>0.634 (0.169)</td>
<td>0.614 (0.161)</td>
<td></td>
</tr>
</tbody>
</table>

Table A.15. Empirical power and false discovery rates (FDR) for detecting enriched genes (genes containing at least one causal SNP) after correcting for multiple hypothesis testing in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer and with population stratification ($N = 5,000$; $h^2 = 0.6$). In this simulation, traits were generated while using the top five principal components (PCs) of the genotype matrix as covariates. GWA summary statistics were computed by fitting a single-SNP univariate linear model (via ordinary least squares) without any control for the additional structure. We show the power of gene-ε to identify enriched genes under the Bonferonni-corrected threshold $P = 2.61 \times 10^{-5}$, corrected for 1,916 genes simulated using chromosome 1 from the UK Biobank genotype data (see Section A.1.2). Results for gene-ε are shown with LASSO, Elastic Net (EN), and Ridge Regression (RR) regularizations. We also show the power of gene-ε without regularization to illustrate the importance of this step (OLS). Additionally, we compare the performance gene-ε with five existing methods: PEGASUS (183), VEGAS (154), RSS (307), SKAT (281), and MAGMA (56). The last is a Bayesian method and is evaluated based on the “median probability criterion” (i.e., posterior enrichment probability of a gene is greater than 0.5). All results are based on 100 replicates and standard deviations of the estimates across runs are given in the parentheses. Approaches with the greatest power are bolded in purple, while methods with the lowest FDR is bolded in blue.
Approaches

<table>
<thead>
<tr>
<th>Causal Genes</th>
<th>Metric</th>
<th>OLS</th>
<th>RR</th>
<th>EN</th>
<th>LASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>Power</td>
<td>0.762 (0.103)</td>
<td>0.682 (0.118)</td>
<td>0.752 (0.119)</td>
<td>0.764 (0.111)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.828 (0.072)</td>
<td>0.326 (0.147)</td>
<td>0.015 (0.041)</td>
<td>0.035 (0.060)</td>
</tr>
<tr>
<td>10%</td>
<td>Power</td>
<td>0.055 (0.059)</td>
<td>0.151 (0.045)</td>
<td>0.037 (0.020)</td>
<td>0.033 (0.022)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.729 (0.159)</td>
<td>0.029 (0.036)</td>
<td>0.007 (0.032)</td>
<td>0.066 (0.178)</td>
</tr>
</tbody>
</table>

Other Methods

<table>
<thead>
<tr>
<th>Causal Genes</th>
<th>Metric</th>
<th>PEGASUS</th>
<th>VEGAS</th>
<th>RSS</th>
<th>SKAT</th>
<th>MAGMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>Power</td>
<td>0.802 (0.104)</td>
<td><strong>0.842 (0.089)</strong></td>
<td>0.811 (0.096)</td>
<td>0.733 (0.111)</td>
<td>0.825 (0.092)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.865 (0.054)</td>
<td>0.862 (0.050)</td>
<td>0.542 (0.100)</td>
<td>0.854 (0.061)</td>
<td>0.881 (0.042)</td>
</tr>
<tr>
<td>10%</td>
<td>Power</td>
<td>0.284 (0.029)</td>
<td>0.228 (0.029)</td>
<td>0.158 (0.019)</td>
<td>0.235 (0.026)</td>
<td><strong>0.324 (0.032)</strong></td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.543 (0.100)</td>
<td>0.589 (0.097)</td>
<td>0.336 (0.075)</td>
<td>0.534 (0.113)</td>
<td>0.546 (0.088)</td>
</tr>
</tbody>
</table>

Table A.16. Empirical power and false discovery rates (FDR) for detecting enriched genes (genes containing at least one causal SNP) after correcting for multiple hypothesis testing in simulations with gene boundaries augmented by a 50 kilobase (kb) buffer and with population stratification \((N = 10,000; \ h^2 = 0.6)\). In this simulation, traits were generated while using the top five principal components (PCs) of the genotype matrix as covariates. GWA summary statistics were computed by fitting a single-SNP univariate linear model (via ordinary least squares) without any control for the additional structure. We show the power of gene-ε to identify enriched genes under the Bonferroni-corrected threshold \(P = 2.61 \times 10^{-5}\), corrected for 1,916 genes simulated using chromosome 1 from the UK Biobank genotype data (see Section A.1.2). Results for gene-ε are shown with LASSO, Elastic Net (EN), and Ridge Regression (RR) regularizations. We also show the power of gene-ε without regularization to illustrate the importance of this step (OLS). Additionally, we compare the performance gene-ε with five existing methods: PEGASUS (183), VEGAS (154), RSS (307), SKAT (281), and MAGMA (56). The last is a Bayesian method and is evaluated based on the “median probability criterion” (i.e., posterior enrichment probability of a gene is greater than 0.5). All results are based on 100 replicates and standard deviations of the estimates across runs are given in the parentheses. Approaches with the greatest power are bolded in purple, while methods with the lowest FDR is bolded in blue.
Table A.17. Empirical type I error estimates using different gene-ε approaches. Here, quantitative traits are simulated with just noise randomly drawn from standard normal distributions. This represents the scenario in which all SNPs are non-causal and satisfy the conventional null hypothesis $H_0: \beta_j = 0$. GWA summary statistics were computed by fitting a single-SNP univariate linear model (via ordinary least squares). Each table entry lists the mean type I error rate estimates for the four gene-ε modeling approaches — which is computed as the proportion of $P$-values under some significance level $\alpha$. Empirical size for the analyses used significance levels of $\alpha = 0.05$, $0.01$, $0.001$, and $2.61 \times 10^{-5}$ (the Bonferroni-corrected threshold), respectively. Sample sizes of the individual-level data (used to derive the summary statistics), were set to $N = 5,000$ and 10,000 observations. These results are based on 100 simulated datasets and the standard errors across the replicated are included in the parentheses. Overall, gene-ε controls the type I error rate for reasonably sized datasets, and can be slightly conservative when the sample size is small and the GWA summary statistics are less precise/more inflated.

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>gene-ε Approach</th>
<th>$\alpha = 0.05$</th>
<th>$\alpha = 0.01$</th>
<th>$\alpha = 0.001$</th>
<th>$\alpha = 2.61 \times 10^{-5}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N = 5,000$</td>
<td>OLS</td>
<td>0.0481 (0.0103)</td>
<td>0.0091 (0.0038)</td>
<td>0.0008 (0.0010)</td>
<td>0.0000 (0.0001)</td>
</tr>
<tr>
<td></td>
<td>Ridge Regression</td>
<td>0.0082 (0.0024)</td>
<td>0.0065 (0.0020)</td>
<td>0.0056 (0.0018)</td>
<td>0.0000 (0.0003)</td>
</tr>
<tr>
<td></td>
<td>Elastic Net</td>
<td>0.0035 (0.0094)</td>
<td>0.0013 (0.0045)</td>
<td>0.0004 (0.0016)</td>
<td>0.0000 (0.0001)</td>
</tr>
<tr>
<td></td>
<td>LASSO</td>
<td>0.0043 (0.0093)</td>
<td>0.0015 (0.0043)</td>
<td>0.0004 (0.0013)</td>
<td>0.0000 (0.0001)</td>
</tr>
<tr>
<td>$N = 10,000$</td>
<td>OLS</td>
<td>0.0486 (0.0109)</td>
<td>0.0095 (0.0034)</td>
<td>0.0008 (0.0008)</td>
<td>0.0000 (0.0000)</td>
</tr>
<tr>
<td></td>
<td>Ridge Regression</td>
<td>0.0067 (0.0029)</td>
<td>0.0050 (0.0031)</td>
<td>0.0044 (0.0033)</td>
<td>0.0000 (0.0003)</td>
</tr>
<tr>
<td></td>
<td>Elastic Net</td>
<td>0.0009 (0.0028)</td>
<td>0.0004 (0.0009)</td>
<td>0.0000 (0.0002)</td>
<td>0.0000 (0.0001)</td>
</tr>
<tr>
<td></td>
<td>LASSO</td>
<td>0.0007 (0.0026)</td>
<td>0.0002 (0.0009)</td>
<td>0.0000 (0.0002)</td>
<td>0.0000 (0.0001)</td>
</tr>
<tr>
<td>gene-ε Approach</td>
<td>Trait</td>
<td># Mix. Comp.</td>
<td>% Associated SNPs</td>
<td>% Causal SNPs</td>
<td>ε-genic Threshold (σ_ε^2 = σ_2^2)</td>
</tr>
<tr>
<td>-----------------</td>
<td>-------</td>
<td>--------------</td>
<td>-------------------</td>
<td>---------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>Elastic Net</td>
<td>Height</td>
<td>8</td>
<td>10.88%</td>
<td>1.39%</td>
<td>3.46 × 10^{-5}</td>
</tr>
<tr>
<td></td>
<td>BMI</td>
<td>6</td>
<td>12.61%</td>
<td>6.23%</td>
<td>5.18 × 10^{-5}</td>
</tr>
<tr>
<td></td>
<td>MCV</td>
<td>8</td>
<td>13.38%</td>
<td>0.32%</td>
<td>6.15 × 10^{-5}</td>
</tr>
<tr>
<td></td>
<td>MPV</td>
<td>9</td>
<td>11.49%</td>
<td>0.21%</td>
<td>7.05 × 10^{-5}</td>
</tr>
<tr>
<td></td>
<td>PLC</td>
<td>8</td>
<td>13.20%</td>
<td>0.45%</td>
<td>6.56 × 10^{-5}</td>
</tr>
<tr>
<td></td>
<td>WHR</td>
<td>6</td>
<td>13.33%</td>
<td>6.28%</td>
<td>5.01 × 10^{-5}</td>
</tr>
<tr>
<td>OLS</td>
<td>Height</td>
<td>4</td>
<td>48.00%</td>
<td>7.90%</td>
<td>4.16 × 10^{-5}</td>
</tr>
<tr>
<td></td>
<td>BMI</td>
<td>3</td>
<td>48.74%</td>
<td>23.28%</td>
<td>4.39 × 10^{-5}</td>
</tr>
<tr>
<td></td>
<td>MCV</td>
<td>9</td>
<td>35.87%</td>
<td>1.67%</td>
<td>6.04 × 10^{-5}</td>
</tr>
<tr>
<td></td>
<td>MPV</td>
<td>9</td>
<td>35.94%</td>
<td>2.21%</td>
<td>6.70 × 10^{-5}</td>
</tr>
<tr>
<td></td>
<td>PLC</td>
<td>7</td>
<td>40.42%</td>
<td>2.45%</td>
<td>5.96 × 10^{-5}</td>
</tr>
<tr>
<td></td>
<td>WHR</td>
<td>2</td>
<td>99.99%</td>
<td>44.51%</td>
<td>1.55 × 10^{-5}</td>
</tr>
</tbody>
</table>

Table A.18. Characterization of the genetic architectures of six traits assayed in European-ancestry individuals in the UK Biobank. Here, we report the way difference regularization makes when gene-ε characterizes ε-genic effects in complex traits. Results are shown for Elastic Net (which is highlighted in the main text). We also show results when no shrinkage is applied to illustrate the importance of this step (denoted by OLS). In the former case, we regress the GWA SNP-level effect size estimates onto chromosome-specific LD matrices to derive a regularized set of summary statistics $\hat{\beta}$. gene-ε assumes a reformulated null distribution of SNP-level effects $\tilde{\beta}_j \sim \mathcal{N}(0, \sigma^2_{\epsilon})$, where $\sigma^2_{\epsilon}$ is the SNP-level null threshold and represents the maximum proportion of phenotypic variance explained (PVE) by a spurious or non-associated SNP. We used an EM-algorithm with 100 iterations to fit $K$-mixture Gaussian models over the regularized effect sizes to estimate $\sigma^2_{\epsilon}$. Here, each mixture component had distinctively smaller variances ($\sigma^2_{1} > \cdots > \sigma^2_{K}$; with the $K$-th component fixed at $\sigma^2_{K} = 0$), and the number of total mixture components $K$ was chosen based on a grid of values where the best model yielded the highest Bayesian Information Criterion (BIC). We assume associated SNPs appear in the first component, non-associated SNPs appear in the last component, and null SNPs with spurious effects fell in between (i.e., $\sigma^2_{\epsilon} = \sigma^2_{2}$). Thus, a SNP is considered to have some level of association with a trait if $E[\tilde{\beta}^2] > \sigma^2_{K} = 0$; while a SNP is considered “causal” if $E[\tilde{\beta}^2] > \sigma^2_{2}$. Column 3 gives the $K$ used for each trait. Column 4 and 5 detail the percentage of associated and causal SNPs, respectively. The last column gives the mean threshold for ε-genic effects across the chromosomes.
Table A.19. Significant genes for body height in the UK Biobank analysis using gene-ε-EN. Here, we analyze 17,680 genes from \( N = 349,468 \) individuals of European-ancestry. This file gives the gene-ε gene-level association \( P \)-values using Elastic Net regularized effect sizes when gene boundaries are defined by (page 1) using UCSC annotations directly, and (page 2) augmenting the gene boundaries by adding SNPs within a ±50kb buffer. Significance was determined by using a Bonferroni-corrected \( P \)-value threshold (in our analyses, \( P = 0.05/14322 \) autosomal genes = \( 3.49 \times 10^{-6} \) and \( P = 0.05/17680 \) autosomal genes = \( 2.83 \times 10^{-6} \), respectively). The columns of tables on both pages provide: (1) chromosome position; (2) gene name; (3) gene-ε-EN gene \( P \)-value; (4) gene-specific heritability estimates; (5) whether or not an association between gene and trait is listed in the GWAS catalog (marked as “yes” or “no”); (6-7) the starting and ending position of the gene’s genomic position; (8) number of SNPs within a gene that were included in analysis; (9) the most significant SNP according to GWA summary statistics; (10) the \( P \)-value of the most significant SNP; and, on the first page, (11) the corresponding gene-level posterior enrichment probability as found by RSS for comparison. Note that an “NA” in column (11) occurs wherever the MCMC for RSS failed to converge. Highlighted rows represent enriched genes whose top SNP is not marginally significant according to a genome-wide Bonferroni-corrected threshold (\( P = 4.67 \times 10^{-8} \) correcting for 1,070,306 SNPs analyzed). (XLSX)

Table A.20. Significant genes for body mass index (BMI) in the UK Biobank analysis using gene-ε-EN. Here, we analyze 17,680 genes from \( N = 349,468 \) individuals of European-ancestry. This file gives the gene-ε gene-level association \( P \)-values using Elastic Net regularized effect sizes when gene boundaries are defined by (page 1) using UCSC annotations directly, and (page 2) augmenting the gene boundaries by adding SNPs within a ±50kb buffer. Significance was determined by using a Bonferroni-corrected \( P \)-value threshold (in our analyses, \( P = 0.05/14322 \) autosomal genes = \( 3.49 \times 10^{-6} \) and \( P = 0.05/17680 \) autosomal genes = \( 2.83 \times 10^{-6} \), respectively). The columns of tables on both pages provide: (1) chromosome position; (2) gene name; (3) gene-ε-EN gene \( P \)-value; (4) gene-specific heritability estimates; (5) whether or not an association between gene and trait is listed in the GWAS catalog (marked as “yes” or “no”); (6-7) the starting and ending position of the gene’s genomic position; (8) number of SNPs within a gene that were included in analysis; (9) the most significant SNP according to GWA summary statistics; (10) the \( P \)-value of the most significant SNP; and, on the first page, (11) the corresponding gene-level posterior enrichment probability as found by RSS for comparison. Note that an “NA” in column (11) occurs wherever the MCMC for RSS failed to converge. Highlighted rows represent enriched genes whose top SNP is not marginally significant according to a genome-wide Bonferroni-corrected threshold (\( P = 4.67 \times 10^{-8} \) correcting for 1,070,306 SNPs analyzed). (XLSX)
Table A.21. Significant genes for mean corpuscular volume (MCV) in the UK Biobank analysis using gene-ε-EN. Here, we analyze 17,680 genes from $N = 349,468$ individuals of European-ancestry. This file gives the gene-ε gene-level association $P$-values using Elastic Net regularized effect sizes when gene boundaries are defined by (page 1) using UCSC annotations directly, and (page 2) augmenting the gene boundaries by adding SNPs within a $±50$kb buffer. Significance was determined by using a Bonferroni-corrected $P$-value threshold (in our analyses, $P = 0.05/14322$ autosomal genes = $3.49 \times 10^{-6}$ and $P = 0.05/17680$ autosomal genes = $2.83 \times 10^{-6}$, respectively). The columns of tables on both pages provide: (1) chromosome position; (2) gene name; (3) gene-ε-EN gene $P$-value; (4) gene-specific heritability estimates; (5) whether or not an association between gene and trait is listed in the GWAS catalog (marked as “yes” or “no”); (6-7) the starting and ending position of the gene’s genomic position; (8) number of SNPs within a gene that were included in analysis; (9) the most significant SNP according to GWA summary statistics; (10) the $P$-value of the most significant SNP; and, on the first page, (11) the corresponding gene-level posterior enrichment probability as found by RSS for comparison. Note that an “NA” in column (11) occurs wherever the MCMC for RSS failed to converge. Highlighted rows represent enriched genes whose top SNP is not marginally significant according to a genome-wide Bonferroni-corrected threshold ($P = 4.67 \times 10^{-8}$ correcting for 1,070,306 SNPs analyzed). (XLSX)

Table A.22. Significant genes for mean platelet volume (MPV) in the UK Biobank analysis using gene-ε-EN. Here, we analyze 17,680 genes from $N = 349,468$ individuals of European-ancestry. This file gives the gene-ε gene-level association $P$-values using Elastic Net regularized effect sizes when gene boundaries are defined by (page 1) using UCSC annotations directly, and (page 2) augmenting the gene boundaries by adding SNPs within a $±50$kb buffer. Significance was determined by using a Bonferroni-corrected $P$-value threshold (in our analyses, $P = 0.05/14322$ autosomal genes = $3.49 \times 10^{-6}$ and $P = 0.05/17680$ autosomal genes = $2.83 \times 10^{-6}$, respectively). The columns of tables on both pages provide: (1) chromosome position; (2) gene name; (3) gene-ε-EN gene $P$-value; (4) gene-specific heritability estimates; (5) whether or not an association between gene and trait is listed in the GWAS catalog (marked as “yes” or “no”); (6-7) the starting and ending position of the gene’s genomic position; (8) number of SNPs within a gene that were included in analysis; (9) the most significant SNP according to GWA summary statistics; (10) the $P$-value of the most significant SNP; and, on the first page, (11) the corresponding gene-level posterior enrichment probability as found by RSS for comparison. Note that an “NA” in column (11) occurs wherever the MCMC for RSS failed to converge. Highlighted rows represent enriched genes whose top SNP is not marginally significant according to a genome-wide Bonferroni-corrected threshold ($P = 4.67 \times 10^{-8}$ correcting for 1,070,306 SNPs analyzed). (XLSX)
Table A.23. Significant genes for platelet count (PLC) in the UK Biobank analysis using gene-ε-EN. Here, we analyze 17,680 genes from $N = 349,468$ individuals of European-ancestry. This file gives the gene-ε gene-level association $P$-values using Elastic Net regularized effect sizes when gene boundaries are defined by (page 1) using UCSC annotations directly, and (page 2) augmenting the gene boundaries by adding SNPs within a $±50$kb buffer. Significance was determined by using a Bonferroni-corrected $P$-value threshold (in our analyses, $P = 0.05/14322$ autosomal genes $= 3.49 \times 10^{-6}$ and $P = 0.05/17680$ autosomal genes $= 2.83 \times 10^{-6}$, respectively). The columns of tables on both pages provide: (1) chromosome position; (2) gene name; (3) gene-ε-EN gene $P$-value; (4) gene-specific heritability estimates; (5) whether or not an association between gene and trait is listed in the GWAS catalog (marked as “yes” or “no”); (6-7) the starting and ending position of the gene’s genomic position; (8) number of SNPs within a gene that were included in analysis; (9) the most significant SNP according to GWA summary statistics; (10) the $P$-value of the most significant SNP; and, on the first page, (11) the corresponding gene-level posterior enrichment probability as found by RSS for comparison. Note that an “NA” in column (11) occurs wherever the MCMC for RSS failed to converge. Highlighted rows represent enriched genes whose top SNP is not marginally significant according to a genome-wide Bonferroni-corrected threshold ($P = 4.67 \times 10^{-8}$ correcting for 1,070,306 SNPs analyzed). (XLSX)

Table A.24. Significant genes for waist-hip ratio (WHR) in the UK Biobank analysis using gene-ε-EN. Here, we analyze 17,680 genes from $N = 349,468$ individuals of European-ancestry. This file gives the gene-ε gene-level association $P$-values using Elastic Net regularized effect sizes when gene boundaries are defined by (page 1) using UCSC annotations directly, and (page 2) augmenting the gene boundaries by adding SNPs within a $±50$kb buffer. Significance was determined by using a Bonferroni-corrected $P$-value threshold (in our analyses, $P = 0.05/14322$ autosomal genes $= 3.49 \times 10^{-6}$ and $P = 0.05/17680$ autosomal genes $= 2.83 \times 10^{-6}$, respectively). The columns of tables on both pages provide: (1) chromosome position; (2) gene name; (3) gene-ε-EN gene $P$-value; (4) gene-specific heritability estimates; (5) whether or not an association between gene and trait is listed in the GWAS catalog (marked as “yes” or “no”); (6-7) the starting and ending position of the gene’s genomic position; (8) number of SNPs within a gene that were included in analysis; (9) the most significant SNP according to GWA summary statistics; (10) the $P$-value of the most significant SNP; and, on the first page, (11) the corresponding gene-level posterior enrichment probability as found by RSS for comparison. Note that an “NA” in column (11) occurs wherever the MCMC for RSS failed to converge. Highlighted rows represent enriched genes whose top SNP is not marginally significant according to a genome-wide Bonferroni-corrected threshold ($P = 4.67 \times 10^{-8}$ correcting for 1,070,306 SNPs analyzed). (XLSX)
Table A.25. Characterization of the genetic architectures of six traits assayed in European-ancestry individuals in the UK Biobank (using un-imputed genotypes). Here, we report the way different regularizations in gene-ε characterize ε-genic effects in complex traits. Results are shown for Elastic Net (which is highlighted in the main text), as well as for LASSO and Ridge Regression. We also show results when no shrinkage is applied to illustrate the importance of this step (denoted by OLS). In the three former cases, we regress the GWA SNP-level effect size estimates onto chromosome-specific LD matrices to derive a regularized set of summary statistics $\bar{\beta}$. gene-ε assumes a reformulated null distribution of SNP-level effects $\bar{\beta}_j \sim \mathcal{N}(0, \sigma^2_\varepsilon)$, where $\sigma^2_\varepsilon$ is the SNP-level null threshold and represents the maximum proportion of phenotypic variance explained (PVE) by a spurious or non-associated SNP. We used an EM-algorithm with 100 iterations to fit $K$-mixture Gaussian models over the regularized effect sizes to estimate $\sigma^2_\varepsilon$. Here, each mixture component had distinctively smaller variances ($\sigma^2_1 > \cdots > \sigma^2_K$; with the $K$-th component fixed at $\sigma^2_K = 0$), and the number of total mixture components $K$ was chosen based on a grid of values where the best model yielded the highest Bayesian Information Criterion (BIC). We assume associated SNPs appear in the first component, non-associated SNPs appear in the last component, and null SNPs with spurious effects fell in between (i.e., $\sigma^2_1 = \sigma^2_2$). Thus, a SNP is considered to have some level of association with a trait if $\mathbb{E}[^2] > 0$; while a SNP is considered “causal” if $\mathbb{E}[^2] > 0$. Column 3 gives the $K$ used for each trait. Column 4 and 5 detail the percentage of associated and causal SNPs, respectively. The last column gives the mean threshold for ε-genic effects across the chromosomes.

<table>
<thead>
<tr>
<th>gene-ε Approach</th>
<th>Trait</th>
<th># Mix. Comp.</th>
<th>% Associated SNPs</th>
<th>% Causal SNPs</th>
<th>ε-genic Threshold ($\sigma^2_\varepsilon$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elastic Net</td>
<td>Height</td>
<td>7</td>
<td>11.65%</td>
<td>1.27%</td>
<td>$1.90 \times 10^{-5}$</td>
</tr>
<tr>
<td></td>
<td>BMI</td>
<td>6</td>
<td>16.61%</td>
<td>7.60%</td>
<td>$2.73 \times 10^{-5}$</td>
</tr>
<tr>
<td></td>
<td>MCV</td>
<td>8</td>
<td>11.60%</td>
<td>0.21%</td>
<td>$3.48 \times 10^{-5}$</td>
</tr>
<tr>
<td></td>
<td>MPV</td>
<td>8</td>
<td>5.47%</td>
<td>0.20%</td>
<td>$3.37 \times 10^{-5}$</td>
</tr>
<tr>
<td></td>
<td>PLC</td>
<td>7</td>
<td>10.33%</td>
<td>0.39%</td>
<td>$3.68 \times 10^{-5}$</td>
</tr>
<tr>
<td></td>
<td>WHR</td>
<td>8</td>
<td>14.38%</td>
<td>7.20%</td>
<td>$3.67 \times 10^{-5}$</td>
</tr>
<tr>
<td>LASSO</td>
<td>Height</td>
<td>8</td>
<td>10.66%</td>
<td>0.67%</td>
<td>$2.29 \times 10^{-5}$</td>
</tr>
<tr>
<td></td>
<td>BMI</td>
<td>6</td>
<td>16.57%</td>
<td>7.35%</td>
<td>$2.71 \times 10^{-5}$</td>
</tr>
<tr>
<td></td>
<td>MCV</td>
<td>7</td>
<td>9.94%</td>
<td>0.23%</td>
<td>$3.56 \times 10^{-5}$</td>
</tr>
<tr>
<td></td>
<td>MPV</td>
<td>9</td>
<td>4.51%</td>
<td>0.20%</td>
<td>$3.32 \times 10^{-5}$</td>
</tr>
<tr>
<td></td>
<td>PLC</td>
<td>8</td>
<td>8.57%</td>
<td>0.26%</td>
<td>$3.80 \times 10^{-4}$</td>
</tr>
<tr>
<td></td>
<td>WHR</td>
<td>6</td>
<td>15.27%</td>
<td>7.05%</td>
<td>$3.29 \times 10^{-4}$</td>
</tr>
<tr>
<td>Ridge Regression</td>
<td>Height</td>
<td>5</td>
<td>47.73%</td>
<td>13.93%</td>
<td>$5.67 \times 10^{-7}$</td>
</tr>
<tr>
<td></td>
<td>BMI</td>
<td>5</td>
<td>38.53%</td>
<td>17.98%</td>
<td>$1.49 \times 10^{-7}$</td>
</tr>
<tr>
<td></td>
<td>MCV</td>
<td>5</td>
<td>45.28%</td>
<td>9.79%</td>
<td>$8.76 \times 10^{-6}$</td>
</tr>
<tr>
<td></td>
<td>MPV</td>
<td>9</td>
<td>34.42%</td>
<td>2.69%</td>
<td>$1.13 \times 10^{-7}$</td>
</tr>
<tr>
<td></td>
<td>PLC</td>
<td>5</td>
<td>46.14%</td>
<td>8.59%</td>
<td>$8.71 \times 10^{-6}$</td>
</tr>
<tr>
<td></td>
<td>WHR</td>
<td>5</td>
<td>42.09%</td>
<td>19.12%</td>
<td>$1.31 \times 10^{-7}$</td>
</tr>
<tr>
<td>OLS</td>
<td>Height</td>
<td>5</td>
<td>49.08%</td>
<td>6.36%</td>
<td>$3.08 \times 10^{-5}$</td>
</tr>
<tr>
<td></td>
<td>BMI</td>
<td>3</td>
<td>54.54%</td>
<td>25.25%</td>
<td>$3.31 \times 10^{-5}$</td>
</tr>
<tr>
<td></td>
<td>MCV</td>
<td>9</td>
<td>34.39%</td>
<td>1.96%</td>
<td>$5.55 \times 10^{-5}$</td>
</tr>
<tr>
<td></td>
<td>MPV</td>
<td>9</td>
<td>36.07%</td>
<td>2.16%</td>
<td>$6.74 \times 10^{-5}$</td>
</tr>
<tr>
<td></td>
<td>PLC</td>
<td>9</td>
<td>36.95%</td>
<td>1.84%</td>
<td>$6.28 \times 10^{-5}$</td>
</tr>
<tr>
<td></td>
<td>WHR</td>
<td>3</td>
<td>50.03%</td>
<td>23.81%</td>
<td>$3.49 \times 10^{-5}$</td>
</tr>
<tr>
<td>Trait</td>
<td>OLS</td>
<td>Ridge Regression</td>
<td>LASSO</td>
<td>Elastic Net</td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>-----</td>
<td>------------------</td>
<td>-------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>Height</td>
<td>501</td>
<td>8</td>
<td>65</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td>BMI</td>
<td>640</td>
<td>8</td>
<td>42</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>MCV</td>
<td>318</td>
<td>10</td>
<td>62</td>
<td>78</td>
<td></td>
</tr>
<tr>
<td>MPV</td>
<td>326</td>
<td>29</td>
<td>62</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td>PLC</td>
<td>289</td>
<td>15</td>
<td>54</td>
<td>52</td>
<td></td>
</tr>
<tr>
<td>WHR</td>
<td>677</td>
<td>6</td>
<td>49</td>
<td>22</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trait</th>
<th>% Sig. Gene Overlap w/ Elastic Net</th>
<th>% Sig. Gene Overlap w/ Elastic Net</th>
<th>% Sig. Gene Overlap w/ Elastic Net</th>
<th>% Sig. Gene Overlap w/ Elastic Net</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>7.78%</td>
<td>37.50%</td>
<td>69.23%</td>
<td>—</td>
</tr>
<tr>
<td>BMI</td>
<td>2.19%</td>
<td>12.50%</td>
<td>64.29%</td>
<td>—</td>
</tr>
<tr>
<td>MCV</td>
<td>12.89%</td>
<td>60.00%</td>
<td>70.97%</td>
<td>—</td>
</tr>
<tr>
<td>MPV</td>
<td>13.80%</td>
<td>51.70%</td>
<td>83.87%</td>
<td>—</td>
</tr>
<tr>
<td>PLC</td>
<td>11.42%</td>
<td>46.67%</td>
<td>75.93%</td>
<td>—</td>
</tr>
<tr>
<td>WHR</td>
<td>1.18%</td>
<td>0.00%</td>
<td>20.41%</td>
<td>—</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trait</th>
<th># Enriched dbGaP Categories</th>
<th># Enriched dbGaP Categories</th>
<th># Enriched dbGaP Categories</th>
<th># Enriched dbGaP Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>BMI</td>
<td>33</td>
<td>16</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MCV</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>MPV</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>PLC</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>WHR</td>
<td>23</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trait</th>
<th>% Enriched dbGaP Overlap w/ Elastic Net</th>
<th>% Enriched dbGaP Overlap w/ Elastic Net</th>
<th>% Enriched dbGaP Overlap w/ Elastic Net</th>
<th>% Enriched dbGaP Overlap w/ Elastic Net</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>100.00% (Body Height)</td>
<td>100.00% (Body Height)</td>
<td>0.00%</td>
<td>—</td>
</tr>
<tr>
<td>BMI</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>—</td>
</tr>
<tr>
<td>MCV</td>
<td>100.00% (Erythrocyte Indices)</td>
<td>33.33% (Erythrocyte Indices)</td>
<td>100.00% (Erythrocyte Indices)</td>
<td>—</td>
</tr>
<tr>
<td>MPV</td>
<td>16.67% (Platelet Count)</td>
<td>100.00% (Platelet Count)</td>
<td>100.00% (Platelet Count)</td>
<td>—</td>
</tr>
<tr>
<td>PLC</td>
<td>50.00% (Platelet Count)</td>
<td>33.33% (Platelet Count)</td>
<td>100.00% (Platelet Count)</td>
<td>—</td>
</tr>
<tr>
<td>WHR</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>—</td>
</tr>
</tbody>
</table>

Table A.26. Comparison of the different gene-ε approaches on the six quantitative traits assayed in European-ancestry individuals from the UK Biobank un-imputed genotyped data. Traits include: height; body mass index (BMI); mean corpuscular volume (MCV); mean platelet volume (MPV); platelet count (PLC); and waist-hip ratio (WHR). Here, we list the number of significant genes found when using gene-ε with various regularization strategies, as well as the number of dbGAP categories enriched for significant genes identified by gene-ε. We also assess how well these results overlap with the gene-ε -EN findings that were reported in the main text. Significant genes were determined by using a Bonferroni-corrected $P$-value threshold (in our analyses, $P = 0.05/13029$ autosomal genes = $3.84 \times 10^{-6}$). Enriched dbGAP categories were those with Enrichr $Q$-values (i.e., false discovery rates) less than 0.05.
<table>
<thead>
<tr>
<th>Trait</th>
<th>OLS</th>
<th>Ridge Regression</th>
<th>LASSO</th>
<th>Elastic Net</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>859</td>
<td>21</td>
<td>90</td>
<td>71</td>
</tr>
<tr>
<td>BMI</td>
<td>770</td>
<td>9</td>
<td>14</td>
<td>73</td>
</tr>
<tr>
<td>MCV</td>
<td>564</td>
<td>90</td>
<td>86</td>
<td>104</td>
</tr>
<tr>
<td>MPV</td>
<td>595</td>
<td>119</td>
<td>83</td>
<td>80</td>
</tr>
<tr>
<td>PLC</td>
<td>517</td>
<td>73</td>
<td>75</td>
<td>69</td>
</tr>
<tr>
<td>WHR</td>
<td>721</td>
<td>4</td>
<td>25</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trait</th>
<th>% Sig. Gene Overlap w/ Elastic Net</th>
<th># Enriched dbGaP Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>6.82%</td>
<td>3</td>
</tr>
<tr>
<td>BMI</td>
<td>1.69%</td>
<td>30</td>
</tr>
<tr>
<td>MCV</td>
<td>13.83%</td>
<td>2</td>
</tr>
<tr>
<td>MPV</td>
<td>12.77%</td>
<td>9</td>
</tr>
<tr>
<td>PLC</td>
<td>11.99%</td>
<td>5</td>
</tr>
<tr>
<td>WHR</td>
<td>0.28%</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trait</th>
<th>% Enriched dbGaP Overlap w/ Elastic Net</th>
<th># Enriched dbGaP Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>33.33% (Body Height)</td>
<td>3</td>
</tr>
<tr>
<td>BMI</td>
<td>0.00%</td>
<td>30</td>
</tr>
<tr>
<td>MCV</td>
<td>50.00% (Erythrocyte Indices)</td>
<td>2</td>
</tr>
<tr>
<td>MPV</td>
<td>11.11% (Platelet Count)</td>
<td>9</td>
</tr>
<tr>
<td>PLC</td>
<td>20.00% (Platelet Count)</td>
<td>5</td>
</tr>
<tr>
<td>WHR</td>
<td>0.00%</td>
<td>10</td>
</tr>
</tbody>
</table>

Table A.27. Comparison of the different gene-ε approaches on the six quantitative traits assayed in European-ancestry individuals from the UK Biobank un-imputed genotyped data with gene boundaries augmented by a 50 kilobase (kb) buffer. Traits include: height; body mass index (BMI); mean corpuscular volume (MCV); mean platelet volume (MPV); platelet count (PLC); and waist-hip ratio (WHR). Here, we list the number of significant genes found when using gene-ε with various regularization strategies, as well as the number of dbGAP categories enriched for significant genes identified by gene-ε. We also assess how well these results overlap with the gene-ε -EN findings that were reported in the main text. Significant genes were determined by using a Bonferroni-corrected P-value threshold (in our analyses, P = 0.05/17680 autosomal genes = 2.83 × 10⁻⁶). Enriched dbGAP categories were those with Enrichr Q-values (i.e., false discovery rates) less than 0.05.
A.4 gene-ε Shows Robustness to Different Prevalence under Case Control GWAS simulations.

In this section, we explore the utility of gene-ε on case-control data. As preliminary results, we demonstrate the performance using simulations that generated with a liability threshold (LT) model (138; 86; 69) that is the same as the simulation scheme in Chapter 3. Another noticeable difference is that we directly run shrinkage regression on the raw data (e.g. genotype matrix $X$ and phenotype $y$) instead of on summary statistics. This is because the relationship between summary statistics and LD matrix for case-control studies can no longer be linked to the multiple regression solution since logistic regression doesn’t have a closed-form solution.

From Fig.A.30 and A.31, we can see that GWAS performance can be effected by different prevalence $k$. And GWAS are generally under-powered when looking at precision and recall curves (right hand side of the figures). Compare with GWAS, gene level method such as gene-ε and SKAT show more promising results (Fig.A.32 and Fig.A.33). For all the simulation scenarios, gene-ε outperforms SKAT similarly to the quantitative trait simulations in Chapter I. What’s more interesting is that when the trait is sparse (e.g. sparsity = 0.01) and heritability is high (e.g. $h^2 = 0.6$), gene-ε shows robust power regardless of $k$ (Bottom row of Fig.A.32)).
Figure A.30. Receiver Operating Characteristic (ROC) curves and Precision and Recall curves of GWAS for sparse case-control simulations.
Figure A.31. Receiver Operating Characteristic (ROC) curves and Precision and Recall curves of GWAS for sparse case-control simulations.
Figure A.32. Receiver Operating Characteristic (ROC) curves and Precision and Recall curves of gene level association methods for sparse case-control simulations.
Figure A.33. Receiver Operating Characteristic (ROC) curves and Precision and Recall curves of gene level association methods for polygenic case-control simulations.
Appendix B

Supporting Information to
“Multi-scale Inference of Genetic Trait Architecture using Biologically Annotated Neural Networks”

B.1 Overview of Partially Connected Bayesian Neural Networks

Biologically annotated neural networks (BANNs) are feedforward Bayesian models with partially connected architectures that are inspired by the hierarchical nature of biological enrichment analyses in GWA studies. The BANNs software takes in one of two data types from genome-wide
association (GWA) studies: (i) individual-level data $\mathcal{D} = \{X, y\}$ where $X$ is an $N \times J$ matrix of genotypes with $J$ denoting the number of single nucleotide polymorphisms (SNPs) encoded as \{0, 1, 2\} copies of a reference allele at each locus and $y$ is an $N$-dimensional vector of quantitative traits (see Fig. 1.2 in the main text); or (ii) GWA summary statistics $\mathcal{D} = \{R, \hat{\theta}\}$ where $R$ is a $J \times J$ empirical linkage disequilibrium (LD) matrix of pairwise correlations between SNPs and $\hat{\theta}$ are marginal effect size estimates for each SNP computed using ordinary least squares (OLS) (see Fig. B.1). In either setting, the BANNs software also requires a predefined list of SNP-set annotations $\{S_1, \ldots, S_G\}$ to construct partially connected network layers that represent different scales of genomic units. In this section, we review the hierarchical probabilistic specification of the BANNs framework for individual data; however, note that extensions to summary statistics is straightforward and only requires substituting the genotypes $X$ for the LD matrix $R$ and substituting the phenotypes $y$ for the OLS effect sizes $\hat{\theta}$.

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

$$y = \sum_{g=1}^{G} h(X_g \theta_g + 1 b^{(1)}_g) w_g + 1 b^{(2)}$$

where $X_g = [x_1, \ldots, x_{|S_g|}]$ is the subset of SNPs annotated for SNP-set $g$; $\theta_g = (\theta_1, \ldots, \theta_{|S_g|})$ are the corresponding inner layer weights; $h(\bullet)$ denotes the nonlinear activations defined for neurons in the hidden layer; $w = (w_1, \ldots, w_G)$ are the weights for the $G$-predefined SNP-sets in the hidden layer; $b^{(1)} = (b^{(1)}_1, \ldots, b^{(1)}_G)$ and $b^{(2)}$ are deterministic biases that are produced during the network training phase in the input and hidden layers, respectively; and $1$ 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 \(\text{ReLU}_\alpha\), where $h(x) = x$ if $x > 0$ and $0.01x$ otherwise. Throughout this Supporting Information, we will equivalently write Eq. (B.1) in matrix notation
\[
\mathbf{y} = \mathbf{H}(\mathbf{\theta}) \mathbf{w} + \mathbf{1}^{(2)},
\]

where \( \mathbf{H}(\mathbf{\theta}) = [h(\mathbf{X}_1 \mathbf{\theta}_1 + \mathbf{1}^{(1)}), \ldots, h(\mathbf{X}_G \mathbf{\theta}_G + \mathbf{1}^{(1)}\mathbf{G})] \) denotes the matrix of nonlinear neurons in the hidden layer which are empirically computed given estimates the input layer weights. The hierarchical structure of the joint likelihood can be seen as a nonlinear take on classical integrative and structural regression models frequently used in GWA analyses (265; 25; 28; 291; 251; 307; 121).

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

\[
\theta_j \sim \pi_{\theta} \sum_{k=1}^{K} \eta_{\theta_k} \mathcal{N}(0, \sigma_{\theta_k}^2) + (1 - \pi_{\theta})\delta_0
\]

where \( \delta_0 \) is a point mass at zero; \( \sigma_{\theta}^2 = (\sigma_{\theta_1}^2, \ldots, \sigma_{\theta_K}^2) \) are variance of the \( K \) nonzero mixture components; \( \eta_{\theta} = (\eta_{\theta_1}, \ldots, \eta_{\theta_K}) \) represents the marginal (unconditional) probability that a randomly selected SNP belongs to the \( k \)-th mixture component such that \( \sum_k \eta_{\theta_k} = 1 \); and \( \pi_{\theta} \) denotes the total proportion of SNPs that have a nonzero effect on the trait of interest. Notice that we write the mixture prior in this form as a way to simplify updates in the algorithm for posterior inference. For reference, one can think of the delta mass as a normal distribution with fixed variance set to zero. Intuitively, specifying a larger \( K \) allows the neural network to learn general SNP effect size distributions spanning over a diverse class of trait architectures. For example, one can take a non-parametric approach and allow \( K \to \infty \) such that Eq. (B.2) mirrors a Dirichlet process Gaussian mixture (297). For results in the main text, we follow previous work and fix \( K = 3 \) (179? ; 155). This corresponds to the general hypothesis that SNPs can have large, moderate, and small nonzero effects on phenotypic variation (39). For inference on the hidden layer, we assume that enriched SNP-sets contain at least one SNP with a nonzero effect. This simpler criterion is formulated by
plac[ing a spike and slab prior on the hidden weights

\[ w_g \sim \pi_w \mathcal{N}(0, \sigma_w^2) + (1 - \pi_w) \delta_0. \quad (B.3) \]

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

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

\[ \log(\pi_\theta) \sim \mathcal{U}(-\log(J), \log(1)), \quad \log(\pi_w) \sim \mathcal{U}(-\log(G), \log(1)). \quad (B.4) \]

Following previous work, relatively uninformative uniform priors are assumed over \( \log \pi_\theta \) and \( \log \pi_w \) to reflect our lack of knowledge \textit{a priori} about the proportion on associated SNP and SNP-sets with nonzero weights (91; 305; 306). To facilitate posterior computation and interpretable inference, we also introduce two vectors of binary indicator variables \( \gamma_\theta = (\gamma_{\theta_1}, \ldots, \gamma_{\theta_J}) \in \{0, 1\}^J \) and \( \gamma_w = (\gamma_{w_1}, \ldots, \gamma_{w_G}) \in \{0, 1\}^G \) where we implicitly assume \textit{a priori} that

\[ \Pr[\gamma_{\theta_j} = 1] = \Pr[\theta_j \neq 0] = \pi_\theta, \quad \Pr[\gamma_{w_g} = 1] = \Pr[w_g \neq 0] = \pi_w. \quad (B.5) \]

Alternatively, we say \( \gamma_{\theta_j} \) and \( \gamma_{w_g} \) take values of 1 when weights \( \theta_j \) and \( w_g \) are drawn from the normal “slabs” of Eqs. (B.2) and (B.3), respectively; they take values of 0 otherwise. In the main text, we refer to these indicators as inclusion probabilities (79) and we use the marginal posterior means of these quantities as general summaries of evidence that SNPs and SNP-sets are statistically associated with phenotypic variation.

### B.2 Variational Expectation-Maximization (EM) Algorithm

We modify a previously developed variational expectation-maximization (EM) algorithm to estimate the posterior distribution of parameters in the BANNs framework. The derivations in this section largely follow those developed in previous work (16; 27; 30; 297; 307). As mentioned in the main
text, the overall goal of variational inference is to approximate the true posterior distribution for network parameters with a similar distribution from an approximating family \((115; 14; 188; 89; 17)\). The EM algorithm we use aims to minimize the Kullback-Leibler divergence between the exact and approximate posterior distributions, respectively. To begin, we assign exchangeable uniform hyper-priors over a grid of values on the log-scale for \(\pi_\theta\) and \(\pi_w\) \(\text{(305)}\). We then run the EM algorithm while iterating through each combination of these values. In the E-step, we use co-ordinate ascent to update the free parameters of the approximate variational posterior. In the M-step, we derive updates for the model hyper-parameters by solving for the roots of their gradients. Finally, in the last step, we empirically compute (approximate) posterior values for the network connection weights \((\theta, w)\) and their corresponding inclusion probabilities by marginalizing over the different model combinations for \(\pi_\theta\) and \(\pi_w\) with normalized importance weights \((27; 30)\). A complete overview of the algorithm is given below. Again, note that extensions to summary statistics is straightforward and only requires substituting the genotypes \(X\) for an LD matrix \(R\) and substituting the phenotypes \(y\) for estimated OLS effect sizes \(\hat{\theta}\) (see Material and Methods in the main text).

Given the formulation of the BANNs model and the partially connected neural network architecture, the weights in the second layer are conditionally independent of the weights in the input layer given the activations (or outputs) from the first layer. This means that we can break up the model fitting procedure into two integrative parts and assess two different lower bounds for the input and hidden layer weights, respectively, to ensure convergence. Specifically estimates on the SNP-level are first maximized with respect to the trait of interest; while, parameters corresponding to the SNP-set level are maximized with respect to the observed trait. The software code iterates between the “inner” lower bound and the “outer” lower bound each step of the algorithm until convergence. Iterations in the algorithm are terminated when either one of two stopping criteria are met: \((i)\) the difference between the lower bound of two consecutive updates are within some small range (specified by tolerance argument \(\epsilon\)), 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.
B.2.1 Input Layer (SNP-Level) Updates

For the SNP-level effects in the input layer, we aim to find a distribution $q(\theta, \gamma_\theta)$ that approximates the true posterior distribution $p(\theta, \gamma_\theta | D)$, where $\theta = (\theta_1, \ldots, \theta_G)$ and $D$ is used to denote the individual-level data and all relevant hyper-parameters. The similarity between these two distributions is maximized by minimizing the Kullback-Leibler (KL) divergence between them. This is formulated by

$$\text{KL}(q(\theta, \gamma_\theta) \parallel p(\theta, \gamma_\theta | D)) = \int \log \left[ \frac{q(\theta, \gamma_\theta)}{p(\theta, \gamma_\theta | D)} \right] q(\theta, \gamma_\theta) d\theta d\gamma_\theta.$$  

(B.6)

Once again, to facilitate posterior inference on this layer, we follow previous work (297) by introducing another vector of binary indicator variables $\varphi_{jk} \in \{0, 1\}$ to indicate which of the $K$-normal components $\theta_j$ belongs to in the prior specified in Eq. (B.2) such that $\text{Pr}[\varphi_{jk} = 1] = \eta_{\theta k}$. This leads to the following natural expression for the variational mixture distribution on each of the individual SNP-level weights

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

(B.7)

where, in addition to previous notation, we assume $q[\varphi_{jk} = 1] = \alpha_{jk}$ for the binary indicators. Additionally, we will let $\phi_j = (\alpha_{jk}, m_{jk}, s_{jk}^2)_{k=1}^K$ denote a collection of free parameters and $\phi = (\phi_1, \ldots, \phi_J)$ will be used to compute the approximations (260; 190). The basic idea behind the variational approximation is to formulate a lower bound to the marginal likelihood, then to iteratively adjust the free parameters in $\phi$ so that this bound becomes as tight as possible (27; 30; 17). Finding the “best” factorized variational distribution amounts to finding the free parameters $\phi$ that make the Kullback-Leibler divergence in Eq. (B.6) as small as possible. Taking moments with respect to the specific class of variational distributions in Eq. (B.7) yields the following analytical
expression for the lower bound on the inner layer (or SNP-level)

$$\text{LB}(\pi_\theta, \sigma^2_\theta, \tau^2_\theta) = \frac{-N}{2} \log(2\pi\tau^2_\theta) - \frac{1}{2\tau^2_\theta} \|y - X\beta_\theta\|^2_2 - \frac{1}{2} \sum^J_{j=1}(X^TX)_{jj} \mathbb{V}[\theta_j]$$

$$- \sum^J_{j=1} \sum^K_{k=1} \alpha_{jk} \log \left( \frac{\alpha_{jk}}{\pi_\theta} \right) + \frac{1}{2} \sum^J_{j=1} \sum^K_{k=1} \alpha_{jk} \left[ 1 + \log \left( \frac{s^2_{jk}}{\sigma^2_\theta \sigma^2_{\theta k}} \right) - \frac{s^2_{jk} + m^2_{jk}}{\sigma^2_\theta \sigma^2_{\theta k}} \right]$$

(B.8)

where $\tau^2_\theta \approx \mathbb{V}[y - X\beta_\theta]$ approximates the variance of residual training error in the input layer; $\| \cdot \|_2$ is the Euclidean norm; $\beta_\theta$ is a $J$-dimensional estimate of the posterior mean for $\theta$ with individual elements $\beta_{\theta j} = \sum^K_{k=1} \alpha_{jk} m_{jk}$; the term $(X^TX)_{jj}$ is the $j$-th diagonal component of the matrix $(X^TX)$; and $\mathbb{V}[\theta_j] = \sum^K_{k=1} \alpha_{jk} (m^2_{jk} + s^2_{jk}) - (\sum^K_{k=1} \alpha_{jk} m_{jk})^2$ is the variance of the $j$-th weight under the approximating distribution in Eq. (B.7). We now describe the expectation and maximization steps of the approximate EM algorithm below (see Software Details in Supporting Information, Section B.10). Again, the derivations in this section largely follow those developed in previous work and other theoretical derivations can be found in those corresponding sources (16; 27; 30; 297; 307).

1. **E-Step: Update the Variational Free Parameters.** As done in Carbonetto and Stephens (2012) (27), for the E-step of the algorithm, we take partial derivatives of the variational lower bound in Eq. (B.8) with respect to the free parameters and conditioned on hyperparameters $(\pi_\theta, \sigma^2_\theta, \tau^2_\theta)$. Next, we set these partial derivatives to zero and solve for $m_{jk}$, $s^2_{jk}$, and $\alpha_{jk}$ which yields

$$m_{jk} = \frac{s^2_{jk}}{\tau^2_\theta} \left[ (X^Ty)_{j} - \sum_{i \neq j} (X^TX)_{ji} \beta_{\theta l} \right]$$

(B.9)

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

(B.10)

$$\alpha_{jk} = \text{Sigmoid} \left( \frac{1}{2} \right)$$

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

2. **M-Step: Update the Variance Hyper-Parameters.** In the M-step of the algorithm, we fix the newly estimated values of the variational free parameters $\phi$ from the E-step and maximize the lower bound in Eq. (B.8) with respect to $\sigma^2_{\theta k}$ and $\tau^2_{\theta}$. This done in the usual way of solving for roots $\sigma^2_{\theta k}$ and $\tau^2_{\theta}$ of the gradient which yields (30)

$$
\sigma^2_{\theta k} = \left[ \sum_{j=1}^{J} \alpha_{jk} (m^2_{jk} + s^2_{jk}) \right] / \left( \tau^2_{\theta} \sum_{j=1}^{J} \alpha_{jk} \right) \quad \text{(B.12)}
$$

$$
\tau^2_{\theta} = \left( N + \sum_{j=1}^{J} \sum_{k=1}^{K} \alpha_{jk} \right)^{-1} \left[ \|y - X\beta_{\theta}\|^2_2 + \sum_{j=1}^{J} (X^t X)_{jj} V[\theta_j] + \sum_{j=1}^{J} \sum_{k=1}^{K} \alpha_{jk} (m^2_{jk} + s^2_{jk}) \right] \quad \text{(B.13)}
$$

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

Following previous work (91; 27; 305; 30; 306; 307), we incorporate Eq. (B.4) and account for our lack of *a priori* knowledge about the “correct” proportion of associated SNPs with nonzero effects by placing an exchangeable uniform hyper-prior distribution over an $L$-valued grid of possible values where $\{\pi_{\theta}^{(1)}, \ldots, \pi_{\theta}^{(L)}\} \in [1/J, 1]$. We then use the lower bound to the likelihood in Eq. (B.8) to approximate the posterior distribution of $\pi_{\theta}$. Formally, we approximate $\Pr[\pi_{\theta} = \pi_{\theta}^{(l)} \mid y, X]$ with the normalized importance weights

$$
\lambda_{\theta}^{(l)} = \frac{\text{LB}(\pi_{\theta}^{(l)}, \sigma^2_{\theta}, \tau^2_{\theta})}{\sum_{l'=1}^{L} \text{LB}(\pi_{\theta}^{(l')}, \sigma^2_{\theta}, \tau^2_{\theta})}. \quad \text{(B.14)}
$$

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

$$
\text{PIP}(j) \equiv \Pr[\gamma_{\theta j} = 1 \mid y, X] \approx \sum_{l=1}^{L} \lambda_{\theta}^{(l)} \Pr[\gamma_{\theta j} = 1 \mid y, X, \pi_{\theta}^{(l)}, \sigma^2_{\theta}, \tau^2_{\theta}]. \quad \text{(B.15)}
$$
This final step can be viewed as an analogy to Bayesian model averaging where marginal distributions are estimated via a weighted average of conditional distributions multiplied by importance weights (100; 27; 30).

### B.2.2 Outer Layer (SNP-Set Level) Updates

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

$$q(w_g, \gamma_{wg}; \psi_g) = \begin{cases} 
\alpha_g \mathcal{N}(m_g, s^2_g) & \text{if } \gamma_{wg} = 1 \\
(1 - \alpha_g) \delta_0 & \text{if } \gamma_{wg} = 0
\end{cases} \quad (B.16)$$

where $\psi_g = (\alpha_g, m_g, s^2_g)$ is used to describe a new set free parameters for the $g$-th SNP-set. Once again, our goal is to find the “best” factorized variational distribution amounts with free parameters $\psi$ that minimize the Kullback-Leibler divergence between the exact and approximate posteriors.

The specific class of variational distributions in Eq. (B.16) yields the following analytical expression for the lower bound on the outer layer (or SNP-set level)

$$LB(\pi_w, \sigma^2_w, \tau^2_w | \theta) = -\frac{N}{2} \log(2\pi\tau^2_w) - \frac{1}{2\tau^2_w} \| y - H(\theta)\beta_w \|^2_2 - \frac{1}{2} \sum_{g=1}^G \{ H(\theta)^T H(\theta) \}_{gg} \mathcal{V}[w_g]$$

$$- \sum_{g=1}^G \alpha_g \log \left( \frac{\alpha_g}{\pi_w} \right) - \sum_{g=1}^G (1 - \alpha_g) \log \left( \frac{1 - \alpha_g}{1 - \pi_w} \right)$$

$$+ \frac{1}{2} \sum_{g=1}^G \alpha_g \left[ 1 + \log \left( \frac{s^2_g}{\sigma^2_{w^2}} \right) - \frac{m^2_g + s^2_g}{\sigma^2_{w^2}} \right] \quad (B.17)$$
where, similar to the input layer updates, $\tau^2_w \approx \mathbb{V}[y - H(\theta)\beta_w]$ estimates the variance of residual training error in the outer layer; the term $\beta_w$ is a $G$-dimensional estimate of the posterior mean for $w$ with elements $\beta_{wg} = \alpha_g m_g$ for the $g$-th SNP-set; the matrix $H(\theta)^\top H(\theta)$ is deterministically computed given posterior estimates of the weights $\theta$ from the input layer, and $\mathbb{V}[w_g] = \alpha_g (m_g^2 + s_g^2) + \alpha^2 m^2_g$ is the variance of the $g$-th weight under the approximating distributional family in Eq. (B.16). We describe the explicit expectation and maximization steps of the approximate EM algorithm for the outer layer below.

1. **E-Step: Update the Variational Free Parameters.** In the E-step of the algorithm, we this time take the partial derivatives of the lower bound in Eq. (B.17) with respect to the free parameters in $\psi$ and set them equal to zero. Solving for $m_g$, $s_g^2$, and $\alpha_g$ yields the following updates

$$m_g = \frac{s_g^2}{\tau^2_w} \left[ (H(\theta)^\top y)_g - \sum_{l \neq g} \{H(\theta)^\top H(\theta)\}_{lg} \beta_{wl} \right] \quad (B.18)$$

$$s_g^2 = \tau^2_w \left[ \{H(\theta)^\top H(\theta)\}_{gg} + \frac{1}{\sigma^2_w} \right]^{-1} \quad (B.19)$$

$$\alpha_g = \text{Sigmoid} \left( \log \left( \frac{\pi_w}{1 - \pi_w} \right) + \log \left( \frac{s_g}{\sigma_w \tau_w} \right) + \frac{m_g^2}{2s_g^2} \right) \quad (B.20)$$

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

2. **M-Step: Update the Variance Hyper-Parameters.** In the M-step of the algorithm, we fix values of the variational free parameters and maximize the lower bound in Eq. (B.17) with respect to $\sigma^2_w$ and $\tau^2_w$. Once again, this is done by solving for the roots $\sigma^2_w$ and $\tau^2_w$ of the gradient which then yields the following updates

$$\sigma^2_w = \left[ \sum_{g=1}^G \alpha_g (m_g^2 + s_g^2) \right] / \left( \tau^2_w \sum_{g=1}^G \alpha_g \right) \quad (B.21)$$
\[
\tau^2_w = \left( N + \sum_{g=1}^{G} \alpha_g \right)^{-1} \left[ \| y - H(\theta)\beta_0 \|_2^2 + \sum_{g=1}^{G} \{ H(\theta)^\top H(\theta) \}_{gg} \gamma[w_g] + \frac{1}{\sigma^2_w} \sum_{g=1}^{G} \alpha_g (m^2 + s^2) \right]
\]

(B.22)

where, again, \(N\) is equal to the dimensionality of the phenotypic response vector \(y\).

Similar to the algorithmic updates in the input layer, we account for our lack of \(a \text{ priori}\) knowledge about the “correct” proportion of enriched SNP-sets by placing another exchangeable uniform hyper-prior distribution over an \(L\)-valued grid of possible values where \(\{ \pi_w^{(1)}, \ldots, \pi_w^{(L)} \} \in [1/G, 1]\). Here, we now use the variational lower bound in Eq. (B.17) to approximate the posterior distribution of \(\pi_w\). As a final step in the model fitting procedure, we again conduct a Bayesian model averaging-like procedure by integrating over the different grid combinations for \(\pi_w\) and computing marginal posterior inclusion probabilities for each of the \(G\)-annotated SNP-sets as the following

\[
\text{PIP}(g) \equiv \Pr[\gamma_{wg} = 1 \mid y, X, \theta] \approx \sum_{l=1}^{L} \lambda_{w}^{(l)} \Pr[\gamma_{wg} = 1 \mid y, X, \theta, \pi_w^{(l)}, \sigma^2_w, \tau^2_w].
\]

(B.23)

where each importance weight \(\lambda_{w}^{(l)}\) takes on a form similar to the normalized ratio described in Eq. (B.13).

### B.3 Accounting for Non-Additive Genetic Effects

As mentioned in the main text, the BANNs framework models the proportion of phenotypic variance that is explained by sparse genetic effects (both additive and non-additive) (305). This primarily done through the inclusion of the nonlinear Leaky ReLU activation function \(h(\bullet)\) in the hidden layer (?). In other areas of statistical genetics, similar nonlinear functions have been used to model non-additive effects that contribute to phenotypic variation (57; 58; 178; 239; 269; 38; 270). For example, it has been shown that the Taylor series expansion of the Gaussian kernel function enumerates higher-order interaction terms between SNPs (46; 113; 51; 52), thus alleviating potential combinatorial concerns with exhaustive searches (49). The nonlinear ReLU function family shares
this same property (245; 246; 181). To see this, consider the general ReLU function for SNPs in
the \( g \)-th SNP-set which is defined as

\[
h(X_g \theta_g) = \max \left\{ 0, \sum_{j=1}^{\|S_g\|} x_j \theta_j + 1 b_g^{(1)} \right\}.
\]  

(B.24)

Under this formulation, it becomes clear that the effect that any one SNP has in determining the
output of the ReLU activation jointly depends on the effects of all other variants in the SNP-set —
therefore, capturing the dependence or interaction between the inputs in the function. The SNP-set
specific bias terms \( b_g^{(1)} \) allow each node in the hidden layer to change slope for different combinations
of genotypes and provides more flexible estimation of broad-sense heritability. Theoretically, as
more nodes and hidden layers are added to the network architecture, the model will have an
even greater ability to account for non-additive genetic effects (acting similarly to classic Gaussian
process regression methods) (? ). Through our simulation studies, we demonstrate the capability
to accurately prioritize/rank associated SNPs and enriched SNP-sets in the BANNs framework,
even in the presence of pairwise SNP-by-SNP interactions and population structure.

### B.4 Estimating Phenotypic Variance Explained (PVE)

As described in the main text, we are able to provide an estimate of phenotypic variance explained
(PVE) within the BANNs framework as the total proportion of phenotypic variance that is explained
by sparse genetic effects (both additive and non-additive) (305). Given the true values of the neural
network parameters, we define this proportion on the SNP-level in the inner layer and SNP-set level
in the outer layer as the following

\[
PVE(\theta) \approx \frac{\mathbb{V}[X\theta]}{\mathbb{V}[y]}, \quad PVE(w) \approx \frac{\mathbb{V}[H(\theta)w]}{\mathbb{V}[y]},
\]  

(B.25)

where, as a reminder, \( \mathbb{V}[\cdot] \) is the variance function and \( H(\theta) = [h(X_1 \theta_1 + 1 b^{(1)}_1), \ldots, h(X_G \theta_G + 1 b^{(1)}_G)] \) denotes the matrix of deterministic nonlinear neurons in the hidden layer given estimates
of the input layer weights. In practice, we estimate PVE using posterior values of the network parameters derived from the variational EM algorithm described in the previous section. Specifically, after averaging over the grid of different models, we use the (approximate) marginal posterior means $\beta_\theta$ and $\beta_w$ for the input and outer layer weights from Eqs. (B.8) and (B.16), respectively. We also approximate the variance of residual error that is observed during the training phase of both layers with estimates of $\tau_\theta^2$ and $\tau_w^2$ from Eqs. (B.13) and (B.22). This yields the following empirical estimate for the PVE of complex traits

$$PVE(\theta) \approx \frac{\nabla[X\beta_\theta]}{\nabla[X\beta_\theta] + \tau_\theta^2}, \quad PVE(w) \approx \frac{\nabla[H(\beta_\theta)\beta_w]}{\nabla[H(\beta_\theta)\beta_w] + \tau_w^2},$$

where the matrix hidden neurons is empirically estimated as $H(\beta_\theta) = [h(X_1\beta_{\theta 1} + b_{11}), \ldots, h(X_G\beta_{\theta G} + b_{1G}^{(1)})]$. Note that this formula is similar to the traditional form used for estimating PVE, except here we also consider the contribution of non-additive genetic effects through the nonlinear Leaky ReLU activation function $h(\bullet)$ (B.1). Through various simulations, we demonstrate the ability to accurately estimate PVE in the BANNs framework under additive sparse architectures (see Figs. B.26 and B.27). We underestimate PVE in both polygenic traits and traits with pairwise SNP-by-SNP interactions, which we believe is caused by either (i) a misestimation of the approximate posterior mean for network weights or (ii) an over-estimation of the residual error variance during the variational EM algorithm. Similar observations have been noted when using variational inference (27; 17; 297; 84). Results from other work also suggest that the sparsity assumption on the SNP-level effects can lead to the underestimation of the PVE (91; 305).

### B.5 Ablation Test and Analysis

To investigate how choices in the model setup contribute to variable selection, we performed an “ablation analysis” where we modified parts of the BANNs framework independently and observed their direct effect on model performance. We considered two different modifications to our model: (1) removing the activation function and training a fully linear hierarchical model, and (2) removing
the approximate Bayesian model averaging approach and updating the probabilities $\pi_\theta$ and $\pi_w$ as additional parameters in the variational EM algorithm. In the normal BANNs setup, we initialize $L$ different models with varying priors for inclusion probabilities specified over a grid \( \{\pi_\theta^{(1)}, \ldots, \pi_\theta^{(L)}\} \in [1/J, 1] \) and \( \{\pi_w^{(1)}, \ldots, \pi_w^{(L)}\} \in [1/G, 1] \), respectively. However, in the case of the latter ablation modification, we initialize $\pi_\theta = 1/J$ and $\pi_w = 1/G$ as an analogy to the “single causal variant” assumption frequently used in fine mapping (262). Next, we update their values in the M-step of the algorithm according to the following analytic expressions

\[
\frac{\pi_\theta}{1 - \pi_\theta} = \frac{\sum_j \sum_k \alpha_{jk}}{\sum_j \sum_k (1 - \alpha_{jk})}, \quad \frac{\pi_w}{1 - \pi_w} = \frac{\sum_g \alpha_g}{\sum_g (1 - \alpha_g)}, \tag{B.27}
\]

Results presented in the main text and Supporting Information (see Fig. B.25) are shown using simulations with the self-identified “white British” ancestry cohort from the UK Biobank on synthetic traits that have broad-sense heritability $H^2 = 0.6$. Each plot combines results from 100 simulated replicates (see Section B.9 for details on the setup for the simulation study).

### B.6 Data Quality Control Procedures for Stock of Mice

Some of the real data analysis results in this work made use of GWA data from the Wellcome Trust Centre for Human Genetics (http://mtweb.cs.ucl.ac.uk/mus/www/mouse/index.shtml). This study contains $N = 1,814$ heterogenous stock of mice from 85 families (all descending from eight inbred progenitor strains) (249), and 131 quantitative traits that are classified into 6 broad categories including behavior, diabetes, asthma, immunology, haematology, and biochemistry (http://mtweb.cs.ucl.ac.uk/mus/www/GSCAN/index.shtml/index.old.shtml). In the main text, we focused on six specific phenotypes from these categories including: body mass index (BMI) (Obesity.BMI), body weight (Glucose.BodyWeight), percentage of CD8+ cells (Imm.PctCD8), mean corpuscular hemoglobin (MCH) (Haem.MCH), high-density lipoprotein content (Biochem.HDL), and low-density lipoprotein content (Biochem.LDL). All phenotypes were previously corrected for sex, age, body weight, season, year, and cage effects (249). For individuals with missing genotypes, we imputed
values by the mean genotype of that SNP in their corresponding family. Only polymorphic SNPs with minor allele frequency above 5% were kept for the analyses. This left a total of $J = 10,227$ autosomal SNPs that were available for all mice. For annotations, we used the Mouse Genome Informatics database (http://www.informatics.jax.org) to map SNPs to the closest neighboring gene(s). Here, pseudogenes, quantitative trait loci (QTL), and genes with only 1 annotated SNP within their boundary were excluded from the analyses. Unannotated SNPs located within the same genomic region were labeled as being within the “intergenic region” between two genes. Altogether, a total of $G = 1,925$ SNP-sets were analyzed.

**B.7 Data Quality Control Procedures for Framingham Heart Study**

The other real data analysis results made use of human GWA data from the Framingham Heart Study (https://www.ncbi.nlm.nih.gov/gap) (228). This study originally contains $N = 6,950$ individuals and $J = 394,174$ SNPs. For quality control on these data, we removed (i) SNPs with minor allele frequency less than 2.5%, (ii) SNPs not in Hardy-Weinberg Equilibrium (Fisher’s exact test $P > 1 \times 10^{-4}$), and (iii) proximal SNPs in high linkage disequilibrium (using the flag --indep-pairwise 50 5 0.8 with PLINK 1.9 (197)). This resulted in a final dataset containing $J = 372,131$ SNPs, where any missing values for a given SNP were imputed by using the estimated mean genotype of that SNP. Next, we used the NCBI’s Reference Sequence (RefSeq) database in the UCSC Genome Browser (195) to annotate SNPs with appropriate genes. Recall that in the real data analysis, we define genes with boundaries in two ways: (a) we use the UCSC gene boundary definitions directly, or (b) we augment the gene boundaries by adding SNPs within a ±500 kilobase (kb) buffer to account for possible regulatory elements. Genes with only 1 SNP within their boundary were excluded from either analysis. Unannotated SNPs located within the same genomic region were labeled as being within the “intergenic region” between two genes. Altogether, a total of $G = 18,364$ SNP-sets were analyzed—which included 8,658 intergenic SNP-sets and 9,706 annotated genes—using the UCSC boundaries. When including the 500kb buffer, a total of $G =$
35,871 SNP-sets were analyzed.

**B.8 Data Quality Control Procedures for UK Biobank**

The simulation results and additional lipoprotein study presented in the main text made use of imputed data released from the UK Biobank (24). Quality control procedures for these data are as follows. First, we only studied individuals who self-identified as being of European ancestry. From this cohort, we further excluded individuals identified by the UK Biobank to have high heterozygosity, excessive relatedness, or aneuploidy (1,550 individuals removed). We also removed individuals whose kinship coefficient was greater than 0.0442 (i.e., close relatives). Next, we removed (i) monomorphic SNPs, (ii) SNPs with minor allele frequency less than 2.5%, (iii) SNPs not in Hardy-Weinberg Equilibrium (Fisher’s exact test \( P > 1 \times 10^{-6} \)), (iv) SNPs with missingness greater than 1%, and (v) SNPs in high linkage disequilibrium (using the flag \(--\text{indep-pairwise} \ 5 \ 5 \ 0.9\) with PLINK 1.9 (197)). After all QC steps, we had a final dataset of \( N = 349,414 \) individuals and \( J = 1,070,306 \) SNPs. Next, we used the NCBI’s Reference Sequence (RefSeq) database in the UCSC Genome Browser (195) to annotate SNPs with appropriate genes. Again, in the real data analysis, we define genes with boundaries in two ways: (a) we use the UCSC gene boundary definitions directly, or (b) we augment the gene boundaries by adding SNPs within a ±500 kilobase (kb) buffer to account for possible regulatory elements. Genes with only 1 SNP within their boundary were excluded from either analysis. Unannotated SNPs located within the same genomic region were labeled as being within the “intergenic region” between two genes. Altogether, a total of \( G = 28,644 \) SNP-sets were kept for analysis using the UCSC boundaries and a total of \( G = 35,849 \) SNP-sets were kept for analysis when including the 500kb buffer.

**B.9 Simulation Setup and Scenarios**

In our simulation studies, we used the following general simulation scheme to generate quantitative traits using real genotype data on chromosome 1 from ten thousand randomly sampled individuals
of European ancestry in the UK Biobank (24). We consider two different data compositions. In
the first, we simulate synthetic traits only using individuals who self-identify as being of “white
British” ancestry. In the second, we simulate phenotypes by random subsampling 3,000 individuals
who self-identify as being of “white British” ancestry, 3,000 individuals who self-identify as being of
“white Irish” ancestry, and 4,000 individuals who identify as being of “any other white background”.
Note that the latter composition introduces additional, yet cryptic, population structure into the
problem. In the main text and Supporting information (i.e., Figs. 1.2-1.3, Figs. B.2-B.27, and
Tables B.1-B.8), we refer to these datasets as the “British” and “European” cohorts, respectively.

The setup to generate synthetic traits follows mostly from previous studies (49; 51; 52; 39). We
will denote this genotype matrix as $X$, with $x_j$ denoting the genotypic vector for the $j$-th SNP.
Following quality control procedures detailed in the previous section, our simulations included $J =$
36,518 SNPs distributed across genome. Again, we used the NCBI’s RefSeq database in the UCSC
Genome Browser to assign SNPs to genes which resulted in 1,408 genes to be used in the simulation
study. We also consider the unannotated SNPs between two genes to be located within intergenic
regions. Altogether, a total of $G = 2,816$ SNP-sets were analyzed.

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

$$y = \sum_{c \in C} x_c \theta_c + W\phi + \epsilon.$$  

(B.28)

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

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

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

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

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

B.10 Software Details

Source code for the BANNs framework is freely available at https://github.com/lcrawlab/BANNs and is licensed under the GNU General Public License (version 3.0). We have released two versions of the BANNs software: one implemented within Python 3 (release version 3.7.7) and other within R (compatible with versions 3.3.2 through 3.6.3). The BANNs GitHub repository includes example data, documentation, and instructions for how to execute the code within both coding languages. Results in the main text and Supporting Information are based on the Python 3 implementation which depends on the pandas library (version 1.0.1) (170) for automatically creating partial neural network architectures based on the biological annotations provided by the user; the NumPy (version 1.18-19) (259) and Numba (version 0.48.0) (131) packages for efficient matrix operations; and the multiprocessing library (version 2.6) (169) for parallelizing posterior computation over multiple threads and providing faster execution. Training, estimation of the network parameters, and optimization was done by using an Adam optimizer (123) in TensorFlow (version 1.5). While the software can be run directly using the source code, it can also be installed as a package through pip with the command: pip3 install BANNs. All dependencies are also automatically installed with the package.

The R implementation uses the dplyr package (version 0.8.5) (272) for automatically creating partial neural network architectures based on the biological annotations provided by the user; the Matrix package (version 1.2-18) (8) for efficient matrix operations; and the doParallel (version 1.0.15) (45), foreach (version 1.4.8) (173), iterators (version 1.0.12) (2), and standard parallel packages for parallelized execution of the variational expectation-maximization algorithm. Similarly, the R implementation of the software can be run by directly downloading the source code or it
can be installed using `devtools` (273) with the commands: `devtools::install("lcrawlab/BANNs")` and `library(BANNs)`.

**Software Details for Competing Approaches.** In this work, comparisons to SNP-level association mapping methods were made using software for CAVIAR (version 2.0.0; `http://genetics.cs.ucla.edu/caviar/`), FINEMAP (version 1.4; `http://www.christianbenner.com`), and SuSiE (version 0.9.0; `https://github.com/stephenslab/susieR`). Comparisons to SNP-set mapping methods were made using software for GBJ (version 0.5.3; `https://cran.r-project.org/web/packages/GBJ/`), GSEA (`https://www.nr.no/en/projects/software-genomics`), MAGMA (version 1.07b; `https://ctg.cnrc.nl/software/magma`), PEGASUS (version 1.3.0; `https://github.com/ramachandran-lab/PEGASUS`), RSS (version 1.0.0; `https://github.com/stephenslab/rss`), and SKAT (version 1.3.2.1; `https://www.hsph.harvard.edu/skat`), which are also publicly available. All software for competing methods were fit using the default settings, unless otherwise stated in the main text and Supporting Information.
B.11 Pseudocode for Biologically Annotated Neural Networks

Algorithm 1 BANNs Model with Individual Level Data

1: Input genotype data $X$, continuous trait $y$, and annotations $\{S_1, \ldots, S_G\}$.
2: Choose the number of models $L$, number of maximum iterations $T$, and tolerance parameter $\epsilon$.
3: Set up the $L$-grid of possible values $\{\pi^{(1)}_g, \ldots, \pi^{(L)}_g\} \in [1/J, 1]$ and $\{\pi^{(1)}_w, \ldots, \pi^{(L)}_w\} \in [1/G, 1]$ for the inner and outer layer, respectively.
4: Randomly initialize variational parameters $\{\alpha_{jk}, m_{jk}, s_{jk}^2\}_{Kk=1}^K$, $\{\sigma^2_{\theta k}\}_{Kk=1}^K$, and $\tau^2_{gh}$ for the inner layer.
5: Randomly initialize variational parameters $\{\alpha_g, m_g, s^2_g\}, \sigma^2_w$, and $\tau^2_w$ for the outer layer.
6: for each $\pi^{(l)}_g \in \{\pi^{(1)}_g, \ldots, \pi^{(L)}_g\}$ and $\pi^{(l)}_w \in \{\pi^{(1)}_w, \ldots, \pi^{(L)}_w\}$ do
7:     Compute inner lower bound $LB_{inner_{new}}$.
8:     for $t = 1 \rightarrow T$ do \hspace{1cm} \triangleright Inner Layer Updates
9:         Set $LB_{inner} = LB_{inner_{new}}$.
10:        Update variational parameters $\{\alpha_{jk}, m_{jk}, s_{jk}^2\}_{Kk=1}^K$ for $j = 1, \ldots, J$ SNPs. \hspace{1cm} \triangleright E-Step
11:       Update hyper-parameters parameters $\{\sigma^2_{\theta k}\}_{Kk=1}^K$ and $\tau^2_{gh}$.
12:       Update lower bound $LB_{inner_{new}}$.
13:       if $LB_{inner_{new}} - LB_{inner} \leq \epsilon$ then
14:           Save $LB_{inner} = LB_{inner_{new}}$.
15:           Break
16:     end if
17: end for
18: Compute hidden layer neurons $H(\theta)$.
19: Compute outer lower bound $LB_{outer_{new}}$.
20: for $t = 1 \rightarrow T$ do \hspace{1cm} \triangleright Outer Layer Updates
21:     Set $LB_{outer} = LB_{outer_{new}}$.
22:     Update variational parameters $\{\alpha_g, m_g, s^2_g\}$ for $g = 1, \ldots, G$ SNP-sets. \hspace{1cm} \triangleright E-Step
23:     Update hyper-parameters parameters $\sigma^2_w$ and $\tau^2_w$.
24:     Update lower bound $LB_{outer_{new}}$.
25:     if $LB_{outer_{new}} - LB_{outer} \leq \epsilon$ then
26:         Save $LB_{outer} = LB_{outer_{new}}$.
27:         Break
28:     end if
29: end for
30: end for
31: Compute normalized importance weights $\lambda^{(l)}_g$ and $\lambda^{(l)}_w$ for $l = 1, \ldots, L$ models.
32: Compute (marginal) posterior means $\beta_\theta$ and $\beta_w$ for network weights $\theta$ and $w$, respectively.
33: Compute (marginal) posterior inclusion probabilities $PIP(\theta)$ and $PIP(w)$.
34: Compute the phenotypic variance explained by the input and hidden layers $PVE(\theta)$ and $PVE(w)$.
35: Return $\{\beta_\theta, \beta_w, PIP(\theta), PIP(w), PVE(\theta), PVE(w)\}$. 
Algorithm 2 BANN-SS Model with GWA Summary Statistics

1: Input LD matrix \( R \), OLS effect size estimates \( \hat{\theta} \), and annotations \( \{S_1, \ldots, S_G\} \).  
2: Choose the number of models \( L \), number of maximum iterations \( T \), and tolerance parameter \( \epsilon \).  
3: Set up the \( L \)-grid of possible values \( \{\pi_\theta^{(1)}, \ldots, \pi_\theta^{(L)}\} \in [1/J, 1] \) and \( \{\pi_w^{(1)}, \ldots, \pi_w^{(L)}\} \in [1/G, 1] \) for the inner and outer layer, respectively.  
4: Randomly initialize variational parameters \( \{\alpha_{jk}, m_{jk}, s_{jk}^2\}_{k=1}^K \) and \( \sigma^2_{\theta k}, \tau^2_{\theta} \) for the inner layer.  
5: Randomly initialize variational parameters \( \{\alpha_g, m_g, s_g^2\}, \sigma^2_w, \tau^2_w \) for the outer layer. 
6: for each \( \pi_\theta^{(l)} \in \{\pi_\theta^{(1)}, \ldots, \pi_\theta^{(L)}\} \) and \( \pi_w^{(l)} \in \{\pi_w^{(1)}, \ldots, \pi_w^{(L)}\} \) do 
7: Compute inner lower bound \( \text{LB}_{\text{inner}}^{\text{new}} \).  
8: for \( t = 1 \rightarrow T \) do  
9: \( \text{Set LB}_{\text{inner}} = \text{LB}_{\text{inner}}^{\text{new}}. \)  
10: Update variational parameters \( \{\alpha_{jk}, m_{jk}, s_{jk}^2\}_{k=1}^K \) for \( j = 1, \ldots, J \) SNPs.  
11: \( \text{Update hyper-parameters parameters } \{\sigma^2_{\theta k}\}_{k=1}^K \text{ and } \tau^2_{\theta}. \)  
12: \( \text{Update lower bound } \text{LB}_{\text{inner}}^{\text{new}}. \)  
13: if \( \text{LB}_{\text{inner}}^{\text{new}} - \text{LB}_{\text{inner}} \leq \epsilon \) then  
14: \( \text{Save } \text{LB}_{\text{inner}} = \text{LB}_{\text{inner}}^{\text{new}}. \)  
15: \( \text{Break} \)  
16: end if  
17: end for  
18: Compute hidden layer neurons \( H(\theta) \).  
19: Compute outer lower bound \( \text{LB}_{\text{outer}}^{\text{new}}. \)  
20: for \( t = 1 \rightarrow T \) do  
21: \( \text{Set LB}_{\text{outer}} = \text{LB}_{\text{outer}}^{\text{new}}. \)  
22: Update variational parameters \( \{\alpha_g, m_g, s_g^2\} \) for \( g = 1, \ldots, G \) SNP-sets.  
23: Update hyper-parameters parameters \( \sigma^2_w, \tau^2_w \).  
24: Update lower bound \( \text{LB}_{\text{outer}}^{\text{new}}. \)  
25: if \( \text{LB}_{\text{outer}}^{\text{new}} - \text{LB}_{\text{outer}} \leq \epsilon \) then  
26: \( \text{Save } \text{LB}_{\text{outer}} = \text{LB}_{\text{outer}}^{\text{new}}. \)  
27: \( \text{Break} \)  
28: end if  
29: end for  
30: end for 
31: Compute normalized importance weights \( \lambda_\theta^{(l)} \) and \( \lambda_w^{(l)} \) for \( l = 1, \ldots, L \) models.  
32: Compute (marginal) posterior means \( \beta_\theta \) and \( \beta_w \) for network weights \( \theta \) and \( w \), respectively.  
33: Compute (marginal) posterior inclusion probabilities PIP(\( \theta \)) and PIP(\( w \)).  
34: Compute the phenotypic variance explained by the input and hidden layers PVE(\( \theta \)) and PVE(\( w \)).  
35: Return \{\( \beta_\theta, \beta_w, \text{PIP}(\theta), \text{PIP}(w), \text{PVE}(\theta), \text{PVE}(w) \}\).
B.12 Supplementary Figures

<table>
<thead>
<tr>
<th>SNP-Set</th>
<th>Chr</th>
<th>Start</th>
<th>End</th>
<th>SNP Sum. Stat.</th>
</tr>
</thead>
<tbody>
<tr>
<td>S_1</td>
<td>1</td>
<td>69090</td>
<td>70008</td>
<td>( \theta_1, \theta_2, \theta_3 )</td>
</tr>
<tr>
<td>S_2</td>
<td>1</td>
<td>3676581</td>
<td>368597</td>
<td>( \theta_4, \theta_5, \theta_6 )</td>
</tr>
<tr>
<td>S_a</td>
<td>6</td>
<td>29200</td>
<td>351355</td>
<td>( \theta_{101}, \theta_{102}, \theta_{103} )</td>
</tr>
<tr>
<td>S_{a+1}</td>
<td>6</td>
<td>391751</td>
<td>411443</td>
<td>( \theta_{104}, \theta_{105}, \theta_{106} )</td>
</tr>
<tr>
<td>S_{a-1}</td>
<td>22</td>
<td>51195513</td>
<td>51237934</td>
<td>( \theta_{107}, \theta_{108}, \theta_{109} )</td>
</tr>
<tr>
<td>S_a</td>
<td>22</td>
<td>51205919</td>
<td>51222087</td>
<td>( \theta_{110}, \theta_{111}, \theta_{112} )</td>
</tr>
</tbody>
</table>

Figure B.1. Biologically annotated neural networks also take in GWA summary statistics (BANN-SS) for multiscale genotype-phenotype by specifying a partially connected architecture based on the hierarchical nature of enrichment studies. (a) The BANN-SS framework requires a \( J \)-dimensional vector of SNP-level GWA marginal effect size (OLS) estimates \( \theta = (\theta_1, \ldots, \theta_J) \); an empirical \( J \times J \) linkage disequilibrium (LD) matrix \( R = [r_{ij}] \), where \( r_{ij} = |r(x_j, x_i), \ldots, r(x_j, x_J)| \) is a vector of correlation coefficients between the \( j \)-th SNP and all other SNPs in the study; and a list of \( G \)-predefined SNP-sets \( \{S_1, \ldots, S_G\} \). 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 (195). (b) A partially connected Bayesian neural network is constructed based on the annotated SNP groups. In the first hidden layer, only SNPs within the boundary of a gene are connected to the same node. Similarly, SNPs within the same intergenic region between genes are connected to the same node. Completing this specification for all SNPs gives the hidden layer the natural interpretation of being the “SNP-set” layer. (c) The hierarchical nature of the network is represented as nonlinear regression model. The corresponding weights in both the SNP (\( \theta \)) and SNP-set (\( \omega \)) layers are treated as random variables with biologically motivated sparse prior distributions. Posterior inclusion probabilities \( \text{PIP}(j) \equiv \text{Pr}[\theta_j \neq 0 | y, X] \) and \( \text{PIP}(g) \equiv \text{Pr}[\omega_g \neq 0 | y, X, \theta_g] \) summarize associations at the SNP and SNP-set level, respectively. The BANN-SS framework uses the same variational inference procedure that is used when we have access to individual-level data.
Figure B.2. Receiver operating characteristic (ROC) curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations (British cohort). Here, quantitative traits are simulated to have broad-sense heritability of \( H^2 = 0.2 \) with only contributions from additive effects (i.e., \( \rho = 1 \)). We show power versus false positive rate for two different trait architectures: (a, b) sparse where only 1% of SNP-sets are enriched for the trait; and (c, d) polygenic where 10% of SNP-sets are enriched. We then set the number of causal SNPs with non-zero effects to be 1% and 10% of all SNPs located within the selected enriched SNP-sets, respectively. To derive results, the full genotype matrix and phenotypic vector are given to the BANNs model and all competing methods that require individual-level data. For the BANN-SS model and other competing methods that take GWA summary statistics, we compute standard GWA SNP-level effect sizes and \( P \)-values (estimated using ordinary least squares). (a, c) Competing SNP-level mapping approaches include: CAVIAR (104), SuSiE (262), and FINEMAP (11). The software for SuSiE requires an input \( \ell \) 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 (307), PEGASUS (183), GBJ (237), SKAT (281), GSEA (101), and MAGMA (56). Note that the upper limit of the x-axis has been truncated at 0.1. All results are based on 100 replicates (see Supporting Information, Section B.9).
Figure B.3. Receiver operating characteristic (ROC) curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations with population structure (European cohort). Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with only contributions from additive effects (i.e., $\rho = 1$). In these simulations, traits were generated while using the top ten principal components (PCs) of the genotype matrix as covariates. We show power versus false positive rate for two different trait architectures: (a, b) sparse where only 1% of SNP-sets are enriched for the trait; and (c, d) polygenic where 10% of SNP-sets are enriched. We then set the number of causal SNPs with non-zero effects to be 1% and 10% of all SNPs located within the selected enriched SNP-sets, respectively. To derive results, the full genotype matrix and phenotypic vector are given to the BANNs model and all competing methods that require individual-level data. For the BANN-SS model and other competing methods that take GWA summary statistics, we compute standard GWA SNP-level effect sizes and $P$-values (estimated using ordinary least squares). (a, c) Competing SNP-level mapping approaches include: CAVIAR (104), SuSiE (262), and FINEMAP (11). The software for SuSiE requires an input $\ell$ 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 (307), PEGASUS (183), GBJ (237), SKAT (281), GSEA (101), and MAGMA (56). Note that the upper limit of the x-axis has been truncated at 0.1. All results are based on 100 replicates (see Supporting Information, Section B.9).
Figure B.4. Receiver operating characteristic (ROC) curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations with population structure (European cohort). Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.6$ with only contributions from additive effects (i.e., $\rho = 1$). In these simulations, traits were generated while using the top ten principal components (PCs) of the genotype matrix as covariates. We show power versus false positive rate for two different trait architectures: (a, b) sparse where only 1% of SNP-sets are enriched for the trait; and (c, d) polygenic where 10% of SNP-sets are enriched. We then set the number of causal SNPs with non-zero effects to be 1% and 10% of all SNPs located within the selected enriched SNP-sets, respectively. To derive results, the full genotype matrix and phenotypic vector are given to the BANNs model and all competing methods that require individual-level data. For the BANN-SS model and other competing methods that take GWA summary statistics, we compute standard GWA SNP-level effect sizes and $P$-values (estimated using ordinary least squares). (a, c) Competing SNP-level mapping approaches include: CAVIAR (104), SuSiE (262), and FINEMAP (11). The software for SuSiE requires an input $\ell$ 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 (307), PEGASUS (183), GBJ (237), SKAT (281), GSEA (101), and MAGMA (56). Note that the upper limit of the x-axis has been truncated at 0.1. All results are based on 100 replicates (see Supporting Information, Section B.9).
Figure B.5. Receiver operating characteristic (ROC) curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations (British cohort). Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with equal contributions from additive effects and epistatic interactions (i.e., $\rho = 0.5$). We show power versus false positive rate for two different trait architectures: (a, b) sparse where only 1% of SNP-sets are enriched for the trait; and (c, d) polygenic where 10% of SNP-sets are enriched. We then set the number of causal SNPs with non-zero effects to be 1% and 10% of all SNPs located within the enriched SNP-sets, respectively. To derive results, the full genotype matrix and phenotypic vector are given to the BANNs model and all competing methods that require individual-level data. For the BANN-SS model and other competing methods that take GWA summary statistics, we compute standard GWA SNP-level effect sizes and $P$-values (estimated using ordinary least squares). (a, c) Competing SNP-level mapping approaches include: CAVIAR (104), SuSiE (262), and FINEMAP (11). The software for SuSiE requires an input $\ell$ 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 (307), PEGASUS (183), GBJ (237), SKAT (281), GSEA (101), and MAGMA (56). Note that the upper limit of the x-axis has been truncated at 0.1. All results are based on 100 replicates (see Supporting Information, Section B.9).
Figure B.6. Receiver operating characteristic (ROC) curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations (British cohort). Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.6$ with equal contributions from additive effects and epistatic interactions (i.e., $\rho = 0.5$). We show power versus false positive rate for two different trait architectures: (a, b) sparse where only 1% of SNP-sets are enriched for the trait; and (c, d) polygenic where 10% of SNP-sets are enriched. We then set the number of causal SNPs with non-zero effects to be 1% and 10% of all SNPs located within the enriched SNP-sets, respectively. To derive results, the full genotype matrix and phenotypic vector are given to the BANNs model and all competing methods that require individual-level data. For the BANN-SS model and other competing methods that take GWA summary statistics, we compute standard GWA SNP-level effect sizes and $P$-values (estimated using ordinary least squares). (a, c) Competing SNP-level mapping approaches include: CAVIAR (104), SuSiE (262), and FINEMAP (11). The software for SuSiE requires an input $\ell$ 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 (307), PEGASUS (183), GBJ (237), SKAT (281), GSEA (101), and MAGMA (56). Note that the upper limit of the x-axis has been truncated at 0.1. All results are based on 100 replicates (see Supporting Information, Section B.9).
Figure B.7. Receiver operating characteristic (ROC) curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations with population structure (European cohort). Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with equal contributions from additive effects and epistatic interactions (i.e., $\rho = 0.5$). In these simulations, traits were generated while using the top ten principal components (PCs) of the genotype matrix as covariates. We show power versus false positive rate for two different trait architectures: (a, b) sparse where only 1% of SNP-sets are enriched for the trait; and (c, d) polygenic where 10% of SNP-sets are enriched. We then set the number of causal SNPs with non-zero effects to be 1% and 10% of all SNPs located within the enriched SNP-sets, respectively. To derive results, the full genotype matrix and phenotypic vector are given to the BANNs model and all competing methods that require individual-level data. For the BANN-SS model and other competing methods that take GWA summary statistics, we compute standard GWA SNP-level effect sizes and $P$-values (estimated using ordinary least squares). (a, c) Competing SNP-level mapping approaches include: CAVIAR (104), SuSiE (262), and FINEMAP (11). The software for SuSiE requires an input $\ell$ 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 (307), PEGASUS (183), GBJ (237), SKAT (281), GSEA (101), and MAGMA (56). Note that the upper limit of the x-axis has been truncated at 0.1. All results are based on 100 replicates (see Supporting Information, Section B.9).
Figure B.8. Receiver operating characteristic (ROC) curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations with population structure (European cohort). Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.6$ with equal contributions from additive effects and epistatic interactions (i.e., $\rho = 0.5$). In these simulations, traits were generated while using the top ten principal components (PCs) of the genotype matrix as covariates. We show power versus false positive rate for two different trait architectures: (a, b) sparse where only 1% of SNP-sets are enriched for the trait; and (c, d) polygenic where 10% of SNP-sets are enriched. We then set the number of causal SNPs with non-zero effects to be 1% and 10% of all SNPs located within the enriched SNP-sets, respectively. To derive results, the full genotype matrix and phenotypic vector are given to the BANNs model and all competing methods that require individual-level data. For the BANN-SS model and other competing methods that take GWA summary statistics, we compute standard GWA SNP-level effect sizes and $P$-values (estimated using ordinary least squares). (a, c) Competing SNP-level mapping approaches include: CAVIAR (104), SuSiE (262), and FINEMAP (11). The software for SuSiE requires an input $\ell$ 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 (307), PEGASUS (183), GBJ (237), SKAT (281), GSEA (101), and MAGMA (56). Note that the upper limit of the x-axis has been truncated at 0.1. All results are based on 100 replicates (see Supporting Information, Section B.9).
Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with only contributions from additive effects (i.e., $\rho = 1$). We show precision versus recall for two different trait architectures: (a, b) sparse where only 1% of SNP-sets are enriched for the trait; and (c, d) polygenic where 10% of SNP-sets are enriched. We then set the number of causal SNPs with non-zero effects to be 1% and 10% of all SNPs located within the selected enriched SNP-sets, respectively. To derive results, the full genotype matrix and phenotypic vector are given to the BANNs model and all competing methods that require individual-level data. For the BANN-SS model and other competing methods that take GWA summary statistics, we compute standard GWA SNP-level effect sizes and $P$-values (estimated using ordinary least squares). (a, c) Competing SNP-level mapping approaches include: CAVIAR (104), SuSiE (262), and FINEMAP (11). The software for SuSiE requires an input $\ell$ 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 (307), PEGASUS (183), GBJ (237), SKAT (281), GSEA (101), and MAGMA (56). Note that, for traits with sparse architectures, the top ranked SNPs and SNP-sets are always true positives, and therefore the minimal recall is not 0. All results are based on 100 replicates (see Supporting Information, Section B.9).
Figure B.10. Precision-recall curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations (British cohort). Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.6$ with only contributions from additive effects (i.e., $\rho = 1$). We show precision versus recall for two different trait architectures: (a, b) sparse where only 1% of SNP-sets are enriched for the trait; and (c, d) polygenic where 10% of SNP-sets are enriched. We then set the number of causal SNPs with non-zero effects to be 1% and 10% of all SNPs located within the selected enriched SNP-sets, respectively. To derive results, the full genotype matrix and phenotypic vector are given to the BANNs model and all competing methods that require individual-level data. For the BANN-SS model and other competing methods that take GWA summary statistics, we compute standard GWA SNP-level effect sizes and $P$-values (estimated using ordinary least squares). (a, c) Competing SNP-level mapping approaches include: CAVIAR (104), SuSiE (262), and FINEMAP (11). The software for SuSiE requires an input $\ell$ 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 (307), PEGASUS (183), GBJ (237), SKAT (281), GSEA (101), and MAGMA (56). Note that, for traits with sparse architectures, the top ranked SNPs and SNP-sets are always true positives, and therefore the minimal recall is not 0. All results are based on 100 replicates (see Supporting Information, Section B.9).
Figure B.11. Precision-recall curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations with population structure (European cohort). Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with only contributions from additive effects (i.e., $\rho = 1$). In these simulations, traits were generated while using the top ten principal components (PCs) of the genotype matrix as covariates. We show precision versus recall for two different trait architectures: (a, b) sparse where only 1% of SNP-sets are enriched for the trait; and (c, d) polygenic where 10% of SNP-sets are enriched. We then set the number of causal SNPs with non-zero effects to be 1% and 10% of all SNPs located within the selected enriched SNP-sets, respectively.

To derive results, the full genotype matrix and phenotypic vector are given to the BANNs model and all competing methods that require individual-level data. For the BANN-SS model and other competing methods that take GWA summary statistics, we compute standard GWA SNP-level effect sizes and $P$-values (estimated using ordinary least squares). (a, c) Competing SNP-level mapping approaches include: CAVIAR (104), SuSiE (262), and FINEMAP (11). The software for SuSiE requires an input $\ell$ 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 (307), PEGASUS (183), GBJ (237), SKAT (281), GSEA (101), and MAGMA (56). Note that, for traits with sparse architectures, the top ranked SNPs and SNP-sets are always true positives, and therefore the minimal recall is not 0. All results are based on 100 replicates (see Supporting Information, Section B.9).
Figure B.12. Precision-recall curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations with population structure (European cohort). Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.6$ with only contributions from additive effects (i.e., $\rho = 1$). In these simulations, traits were generated while using the top ten principal components (PCs) of the genotype matrix as covariates. We show precision versus recall for two different trait architectures: (a, b) sparse where only 1% of SNP-sets are enriched for the trait; and (c, d) polygenic where 10% of SNP-sets are enriched. We then set the number of causal SNPs with non-zero effects to be 1% and 10% of all SNPs located within the selected enriched SNP-sets, respectively. To derive results, the full genotype matrix and phenotypic vector are given to the BANNs model and all competing methods that require individual-level data. For the BANN-SS model and other competing methods that take GWA summary statistics, we compute standard GWA SNP-level effect sizes and P-values (estimated using ordinary least squares). (a, c) Competing SNP-level mapping approaches include: CAVIAR (104), SuSiE (262), and FINEMAP (11). The software for SuSiE requires an input $\ell$ 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 (307), PEGASUS (183), GBJ (237), SKAT (281), GSEA (101), and MAGMA (56). Note that, for traits with sparse architectures, the top ranked SNPs and SNP-sets are always true positives, and therefore the minimal recall is not 0. All results are based on 100 replicates (see Supporting Information, Section B.9).
Figure B.13. Precision-recall curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations (British cohort). Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with equal contributions from additive effects and epistatic interactions (i.e., $\rho = 0.5$). We show precision versus recall for two different trait architectures: (a, b) sparse where only 1% of SNP-sets are enriched for the trait; and (c, d) polygenic where 10% of SNP-sets are enriched. We then set the number of causal SNPs with non-zero effects to be 1% and 10% of all SNPs located within the selected enriched SNP-sets, respectively. To derive results, the full genotype matrix and phenotypic vector are given to the BANNs model and all competing methods that require individual-level data. For the BANN-SS model and other competing methods that take GWA summary statistics, we compute standard GWA SNP-level effect sizes and $P$-values (estimated using ordinary least squares). (a, c) Competing SNP-level mapping approaches include: CAVIAR (104), SuSiE (262), and FINEMAP (11). The software for SuSiE requires an input $\ell$ 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 (307), PEGASUS (183), GBJ (237), SKAT (281), GSEA (101), and MAGMA (56). Note that, for traits with sparse architectures, the top ranked SNPs and SNP-sets are always true positives, and therefore the minimal recall is not 0. All results are based on 100 replicates (see Supporting Information, Section B.9).
Figure B.14. Precision-recall curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations (British cohort). Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.6$ with equal contributions from additive effects and epistatic interactions (i.e., $\rho = 0.5$). We show precision versus recall for two different trait architectures: (a, b) sparse where only 1% of SNP-sets are enriched for the trait; and (c, d) polygenic where 10% of SNP-sets are enriched. We then set the number of causal SNPs with non-zero effects to be 1% and 10% of all SNPs located within the selected enriched SNP-sets, respectively. To derive results, the full genotype matrix and phenotypic vector are given to the BANNs model and all competing methods that require individual-level data. For the BANN-SS model and other competing methods that take GWA summary statistics, we compute standard GWA SNP-level effect sizes and $P$-values (estimated using ordinary least squares). (a, c) Competing SNP-level mapping approaches include: CAVIAR (104), SuSiE (262), and FINEMAP (11). The software for SuSiE requires an input $\ell$ 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 (307), PEGASUS (183), GBJ (237), SKAT (281), GSEA (101), and MAGMA (56). Note that, for traits with sparse architectures, the top ranked SNPs and SNP-sets are always true positives, and therefore the minimal recall is not 0. All results are based on 100 replicates (see Supporting Information, Section B.9).
Figure B.15. Precision-recall curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations with population structure (European cohort). Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with equal contributions from additive effects and epistatic interactions (i.e., $\rho = 0.5$). In these simulations, traits were generated while using the top ten principal components (PCs) of the genotype matrix as covariates. We show precision versus recall for two different trait architectures: (a, b) sparse where only 1% of SNP-sets are enriched for the trait; and (c, d) polygenic where 10% of SNP-sets are enriched. We then set the number of causal SNPs with non-zero effects to be 1% and 10% of all SNPs located within the selected enriched SNP-sets, respectively. To derive results, the full genotype matrix and phenotypic vector are given to the BANNs model and all competing methods that require individual-level data. For the BANN-SS model and other competing methods that take GWA summary statistics, we compute standard GWA SNP-level effect sizes and $P$-values (estimated using ordinary least squares). (a, c) Competing SNP-level mapping approaches include: CAVIAR (104), SuSiE (262), and FINEMAP (11). The software for SuSiE requires an input $\ell$ 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 (307), PEGASUS (183), GBJ (237), SKAT (281), GSEA (101), and MAGMA (56). Note that, for traits with sparse architectures, the top ranked SNPs and SNP-sets are always true positives, and therefore the minimal recall is not 0. All results are based on 100 replicates (see Supporting Information, Section B.9).
Figure B.16. Precision-recall curves comparing the performance of the BANNs (red) and BANN-SS (black) models with competing SNP and SNP-set mapping approaches in simulations with population structure (European cohort). Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with equal contributions from additive effects and epistatic interactions (i.e., $\rho = 0.5$). In these simulations, traits were generated while using the top ten principal components (PCs) of the genotype matrix as covariates. We show precision versus recall for two different trait architectures: (a, b) sparse where only 1% of SNP-sets are enriched for the trait; and (c, d) polygenic where 10% of SNP-sets are enriched. We then set the number of causal SNPs with non-zero effects to be 1% and 10% of all SNPs located within the selected enriched SNP-sets, respectively. To derive results, the full genotype matrix and phenotypic vector are given to the BANNs model and all competing methods that require individual-level data. For the BANN-SS model and other competing methods that take GWA summary statistics, we compute standard GWA SNP-level effect sizes and $P$-values (estimated using ordinary least squares). (a, c) Competing SNP-level mapping approaches include: CAVIAR (104), SuSiE (262), and FINEMAP (11). The software for SuSiE requires an input $\ell$ 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 (307), PEGASUS (183), GBJ (237), SKAT (281), GSEA (101), and MAGMA (56). Note that, for traits with sparse architectures, the top ranked SNPs and SNP-sets are always true positives, and therefore the minimal recall is not 0. All results are based on 100 replicates (see Supporting Information, Section B.9).
Figure B.17. Scatter plots comparing how the integrative neural network training procedure enables the ability to identify associated SNPs and enriched SNP-sets in simulations (British cohort). Quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with only contributions from additive effects set (i.e., $\rho = 1$). We consider two different trait architectures: (a, b) sparse where only 1% of SNP-sets are enriched for the trait; and (c, d) polygenic where 10% of SNP-sets are enriched. We set the number of causal SNPs with non-zero effects to be $\sim 1\%$ and $\sim 10\%$ of all SNPs located within the enriched SNP-sets, respectively. Results are shown comparing the posterior inclusion probabilities (PIPs) derived by the BANNs model fit with individual-level data on the x-axis and (a, c) SuSiE (262) and (b, d) RSS (307) 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) (6). True positive causal variants used to generate the synthetic phenotypes are colored in red, while non-causal variants are given in grey. SNPs and SNP-sets in the top right quadrant are selected by both approaches; while, elements in the bottom right and top left quadrants are uniquely identified by BANNs and SuSiE/RSS, respectively. Each plot combines results from 100 simulated replicates (see Section B.9).
Figure B.18. Scatter plots comparing how the integrative neural network training procedure enables the ability to identify associated SNPs and enriched SNP-sets in simulations with population structure (European cohort). Quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with only contributions from additive effects set (i.e., $\rho = 1$). We consider two different trait architectures: (a, b) sparse where only 1% of SNP-sets are enriched for the trait; and (c, d) polygenic where 10% of SNP-sets are enriched. We set the number of causal SNPs with non-zero effects to be 1% and 10% of all SNPs located within the enriched SNP-sets, respectively. In these simulations, traits were generated while also using the top ten principal components (PCs) of the genotype matrix as covariates. Results are shown comparing the posterior inclusion probabilities (PIPs) derived by the BANNs model fit with individual-level data on the x-axis and (a, c) SuSiE (262) and (b, d) RSS (307) 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) (6). True positive causal variants used to generate the synthetic phenotypes are colored in red, while non-causal variants are given in grey. SNPs and SNP-sets in the top right quadrant are selected by both approaches; while, elements in the bottom right and top left quadrants are uniquely identified by BANNs and SuSiE/RSS, respectively. Each plot combines results from 100 simulated replicates (see Section B.9).
Figure B.19. Scatter plots comparing how the integrative neural network training procedure enables the ability to identify associated SNPs and enriched SNP-sets in simulations with population structure (European cohort). Quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.6$ with only contributions from additive effects set (i.e., $\rho = 1$). We consider two different trait architectures: (a, b) sparse where only 1% of SNP-sets are enriched for the trait; and (c, d) polygenic where 10% of SNP-sets are enriched. We set the number of causal SNPs with non-zero effects to be 1% and 10% of all SNPs located within the enriched SNP-sets, respectively. In these simulations, traits were generated while also using the top ten principal components (PCs) of the genotype matrix as covariates. Results are shown comparing the posterior inclusion probabilities (PIPs) derived by the BANNs model fit with individual-level data on the x-axis and (a, c) SuSiE (262) and (b, d) RSS (307) 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) (6). True positive causal variants used to generate the synthetic phenotypes are colored in red, while non-causal variants are given in grey. SNPs and SNP-sets in the top right quadrant are selected by both approaches; while, elements in the bottom right and top left quadrants are uniquely identified by BANNs and SuSiE/RSS, respectively. Each plot combines results from 100 simulated replicates (see Section B.9).
Figure B.20. Scatter plots comparing how the integrative neural network training procedure enables the ability to identify associated SNPs and enriched SNP-sets in simulations (British cohort). Quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with equal contributions from additive effects and epistatic interactions (i.e., $\rho = 0.5$). We consider two different trait architectures: (a, b) sparse where only 1% of SNP-sets are enriched for the trait; and (c, d) polygenic where 10% of SNP-sets are enriched. We set the number of causal SNPs with non-zero effects to be 1% and 10% of all SNPs located within the enriched SNP-sets, respectively. Results are shown comparing the posterior inclusion probabilities (PIPs) derived by the BANNs model fit with individual-level data on the x-axis and (a, c) SuSiE (262) and (b, d) RSS (307) 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) (6). True positive causal variants used to generate the synthetic phenotypes are colored in red, while non-causal variants are given in grey. SNPs and SNP-sets in the top right quadrant are selected by both approaches; while, elements in the bottom right and top left quadrants are uniquely identified by BANNs and SuSiE/RSS, respectively. Each plot combines results from 100 simulated replicates (see Section B.9).
Figure B.21. Scatter plots comparing how the integrative neural network training procedure enables the ability to identify associated SNPs and enriched SNP-sets in simulations (British cohort). Quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.6$ with equal contributions from additive effects and epistatic interactions (i.e., $\rho = 0.5$). We consider two different trait architectures: (a, b) sparse where only 1% of SNP-sets are enriched for the trait; and (c, d) polygenic where 10% of SNP-sets are enriched. We set the number of causal SNPs with non-zero effects to be 1% and 10% of all SNPs located within the enriched SNP-sets, respectively. Results are shown comparing the posterior inclusion probabilities (PIPs) derived by the BANNs model fit with individual-level data on the x-axis and (a, c) SuSiE (262) and (b, d) RSS (307) 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) (6). True positive causal variants used to generate the synthetic phenotypes are colored in red, while non-causal variants are given in grey. SNPs and SNP-sets in the top right quadrant are selected by both approaches; while, elements in the bottom right and top left quadrants are uniquely identified by BANNs and SuSiE/RSS, respectively. Each plot combines results from 100 simulated replicates (see Section B.9).
Figure B.22. Scatter plots comparing how the integrative neural network training procedure enables the ability to identify associated SNPs and enriched SNP-sets in simulations with population structure (European cohort). Quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with equal contributions from additive effects and epistatic interactions (i.e., $\rho = 0.5$). We consider two different trait architectures: (a, b) sparse where only 1% of SNP-sets are enriched for the trait; and (c, d) polygenic where 10% of SNP-sets are enriched. We set the number of causal SNPs with non-zero effects to be 1% and 10% of all SNPs located within the enriched SNP-sets, respectively. In these simulations, traits were generated while also using the top ten principal components (PCs) of the genotype matrix as covariates. Results are shown comparing the posterior inclusion probabilities (PIPs) derived by the BANNs model fit with individual-level data on the x-axis and (a, c) SuSiE (262) and (b, d) RSS (307) 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) (6). True positive causal variants used to generate the synthetic phenotypes are colored in red, while non-causal variants are given in grey. SNPs and SNP-sets in the top right quadrant are selected by both approaches; while, elements in the bottom right and top left quadrants are uniquely identified by BANNs and SuSiE/RSS, respectively. Each plot combines results from 100 simulated replicates (see Section B.9).
Figure B.23. Scatter plots comparing how the integrative neural network training procedure enables the ability to identify associated SNPs and enriched SNP-sets in simulations with population structure (European cohort). Quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.6$ with equal contributions from additive effects and epistatic interactions (i.e., $\rho = 0.5$). We consider two different trait architectures: (a, b) sparse where only 1% of SNP-sets are enriched for the trait; and (c, d) polygenic where 10% of SNP-sets are enriched. We set the number of causal SNPs with non-zero effects to be 1% and 10% of all SNPs located within the enriched SNP-sets, respectively. In these simulations, traits were generated while also using the top ten principal components (PCs) of the genotype matrix as covariates. Results are shown comparing the posterior inclusion probabilities (PIPs) derived by the BANNs model fit with individual-level data on the x-axis and (a, c) SuSiE (262) and (b, d) RSS (307) 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) (6). True positive causal variants used to generate the synthetic phenotypes are colored in red, while non-causal variants are given in grey. SNPs and SNP-sets in the top right quadrant are selected by both approaches; while, elements in the bottom right and top left quadrants are uniquely identified by BANNs and SuSiE/RSS, respectively. Each plot combines results from 100 simulated replicates (see Section B.9).
Figure B.24. Assessments of posterior inclusion probability (PIP) calibration for both SNP-level associations and enrichment of SNP-sets. This experiment follows largely from previous work (91; 262). Here, SNPs and SNP-sets across simulations are grouped into bins according to their reported PIPs (using 20 equally spaced bins, from 0 to 1). The plots show the average PIP for each bin against the proportion of causal SNPs or SNP-sets in that bin. A well calibrated method should produce points near the x-axis = y-axis line (i.e., the diagonal red lines). Gray error bars show ±2 standard errors. Panel (a, b) shows the comparison of BANNs SNP layer with SuSiE (262), and (c, d) shows the comparison of BANNs SNP-set layer with RSS (307). While the inclusion probabilities are not perfectly calibrated for any of the methods, the empirical power and false discovery rate (FDR) above the “median probability criterion” (i.e., PIPs for SNPs and SNP-sets greater than 0.5) (6) are still reasonably well controlled (see Tables B.1-B.8). We hypothesize that these calibration results are due both to consequences of both variational inference and the level of polygenicity with which we simulated synthetic phenotypes.
Figure B.25. Receiver operating characteristic (ROC) curves comparing the performance of the BANNs models with different modifications via an ablation test. To investigate how choices in the model setup contribute to variable selection, we performed an “ablation analysis” where we modified parts of the BANNs framework independently and observed their direct effect on model performance (see Section B.5). We considered two different modifications to our model: (1) removing the activation function and training a fully linear hierarchical model, and (2) removing the approximate Bayesian model averaging approach and updating the probabilities $\pi_\theta$ and $\pi_w$ as additional parameters in the variational EM algorithm. In the normal BANNs setup, we initialize $L$ different models with varying priors for inclusion probabilities specified over a grid $\{\pi^{(1)}_\theta, \ldots, \pi^{(L)}_\theta\} \in [1/J, 1]$ and $\{\pi^{(1)}_w, \ldots, \pi^{(L)}_w\} \in [1/G, 1]$, respectively. However, in the case of the latter ablation modification, we initialize $\pi_\theta = 1/J$ and $\pi_w = 1/G$ as an analogy to the “single causal variant” assumption frequently used in fine mapping (262). Next, we update their values in the M-step of the algorithm according to the following analytic expressions: (a, c) $\pi_\theta/1 - \pi_\theta = \sum_j \sum_k \alpha_{jk}/\sum_j \sum_k (1 - \alpha_{jk})$, and (b, d) $\pi_w/1 - \pi_w = \sum_g \alpha_g/\sum_g (1 - \alpha_g)$. Results here are shown using simulations with the self-identified “white British” ancestry cohort from the UK Biobank on synthetic traits that have broad-sense heritability $H^2 = 0.6$. Each plot combines results from 100 simulated replicates (see Section B.9).
Figure B.26. Boxplots depicting the ability of the BANNs and BANN-SS models to estimate the phenotypic variation explained (PVE) by SNPs (pink) and SNP-sets (blue) for complex traits in simulations. In this work, we define PVE as the total proportion of phenotypic variance that is explained by sparse genetic effects (both additive and non-additive) (305). Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with different levels of contributions from additive effects and epistatic interactions. We consider two different trait architectures: (a, b) sparse where only 1% of SNP-sets are enriched for the trait; and (c, d) polygenic where 10% of SNP-sets are enriched. Panels (a, c) show heritability estimates on simulations with genetic data from individuals who self-identify as being of “white British” ancestry in the UK Biobank; while, panels (b, d) show heritability estimates on simulations with genetic data from individuals who more broadly identify as being of European ancestry. True heritability values are shown as the dashed grey horizontal lines. The root mean square error (RMSE) between the BANNs model estimates of the PVE and the true values are also provided.
Figure B.27. Boxplots depicting the ability of the BANNs and BANN-SS models to estimate the phenotypic variation explained (PVE) by SNPs (pink) and SNP-sets (blue) for complex traits in simulations. In this work, we define PVE as the total proportion of phenotypic variance that is explained by sparse genetic effects (both additive and non-additive) \( H^2 = 0.6 \) with different levels of contributions from additive effects and epistatic interactions. We consider two different trait architectures: (a, b) sparse where only 1% of SNP-sets are enriched for the trait; and (c, d) polygenic where 10% of SNP-sets are enriched. Panels (a, c) show heritability estimates on simulations with genetic data from individuals who self-identify as being of “white British” ancestry in the UK Biobank; while, panels (b, d) show heritability estimates on simulations with genetic data from individuals who more broadly identify as being of European ancestry. True heritability values are shown as the dashed grey horizontal lines. The root mean square error (RMSE) between the BANNs model estimates of the PVE and the true values are also provided.
Figure B.28. Manhattan plots of variant-level fine mapping results for six traits in heterogenous stock of mice from the Wellcome Trust Centre for Human Genetics. Traits are grouped based on their category and include: (a) body mass index (BMI) and body weight, (b) percentage of CD8+ cells and mean corpuscular hemoglobin (MCH), and (c) high-density and low-density lipoprotein (HDL and LDL, respectively) cholesterol. Posterior inclusion probabilities (PIP) for the input layer weights are derived from the BANNs model fit on individual-level data and are plotted for each SNP against their genomic positions. Chromosomes are shown in alternating colors for clarity. The black dashed line is marked at 0.5 and represents the “median probability model (MPM)” threshold (6). Here, we only color code SNPs that had a PIP greater than 1% in either trait. SNPs with PIPs exceeding 1% in both traits are marked by a star and denoted as falling in the “overlap” category. BANNs estimated the following PVEs on the SNP and SNP-set levels for these traits, respectively: (i) 0.09 and 0.08 for BMI, (ii) 0.39 and 0.40 for body weight, (iii) 0.51 and 0.48 for percentage of CD8+ cells, (iv) 0.34 and 0.32 for MCH, (v) 0.34 and 0.28 for HDL, and (vi) 0.15 and 0.15 for LDL.
Figure B.29. Gene set enrichment analyses using the significant SNP-sets identified by BANNs for high-density and low-density lipoprotein (HDL and LDL, respectively) traits in the Framingham Heart Study (228). Here, SNP-set annotations are based on gene boundaries defined by the NCBI’s RefSeq database in the UCSC Genome Browser (195). Unannotated SNPs located within the same genomic region were labeled as being within the “intergenic region” between two genes. Posterior inclusion probabilities (PIP) for the input and hidden layer weights are derived by fitting the BANNs model on ≥0.5 (i.e., the “median probability model” threshold (6)). We take these significant SNP-sets and conduct “gene set enrichment analysis” using Enrichr (34; 129) to identify the categories they overrepresent in (a, b) the database of Genotypes and Phenotypes (dbGaP) and (c, d) the GWAS Catalog (2019). Nearly all enriched categories are related with (a, c) HDL and (b, d) LDL, respectively. Note that in LDL, the BANNs framework identified the gene APOB as having a high PIP = 0.976. There have been hypotheses connecting LDL to cognitive traits (35; 264), and APOB has been shown to be related to cerebrospinal fluid and memory (192; 120; 107). Therefore, we argue that results in panel (d) are also relevant.
<table>
<thead>
<tr>
<th>p-value</th>
<th>Odds.ratio</th>
<th># of sig. genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.906e-02</td>
<td>1.678e-01</td>
<td>14.9</td>
</tr>
<tr>
<td>3.428e-02</td>
<td>4.571e-02</td>
<td>33.1</td>
</tr>
<tr>
<td>2.297e-02</td>
<td>4.571e-02</td>
<td>50.0</td>
</tr>
<tr>
<td>3.627e-04</td>
<td>5.199e-03</td>
<td>93.5</td>
</tr>
<tr>
<td>1.313e-04</td>
<td>5.199e-03</td>
<td>158</td>
</tr>
<tr>
<td>1.874e-02</td>
<td>1.678e-01</td>
<td>58.7</td>
</tr>
<tr>
<td>7.688e-02</td>
<td>1.678e-01</td>
<td>13.3</td>
</tr>
<tr>
<td>2.035e-03</td>
<td>4.681e-02</td>
<td>33.8</td>
</tr>
<tr>
<td>3.495e-03</td>
<td>3.146e-02</td>
<td>370</td>
</tr>
<tr>
<td>7.297e-02</td>
<td>1.678e-01</td>
<td>14.0</td>
</tr>
</tbody>
</table>

In this analysis, each gene boundary annotation is modified by adding SNPs within a specific range and the database of Genotypes and Phenotypes (dbGaP) and the GWAS Catalog (2019). Nearly all enriched categories are related with LDL, respectively.
Figure B.31. Manhattan plot of variant-level association mapping results for high-density and low-density lipoprotein (HDL and LDL, respectively) traits in ten thousand randomly sampled individuals of European ancestry from the UK Biobank (24). Posterior inclusion probabilities (PIP) for the neural network weights are derived from the BANNs model fit on individual-level data and are plotted for each SNP against their genomic positions. Chromosomes are shown in alternating colors for clarity. The black dashed line is marked at 0.5 and represents the “median probability model” threshold (6). 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 (195). Unannotated SNPs located within the same genomic region were labeled as being within the “intergenic region” between two genes. These regions are labeled as Gene1-Gene2 in the legend. Gene set enrichment analyses for these SNP-sets can be found in Fig. B.31. Stars (★) denote SNPs and SNP-sets that replicate findings from our analyses of HDL and LDL in the Framingham Heart Study (See Fig. 1.4 in the main text).
Figure B.32. Gene set enrichment analyses using the significant SNP-sets identified by BANNs for high-density and low-density lipoprotein (HDL and LDL, respectively) traits in ten thousand randomly sampled individuals of European ancestry from the UK Biobank (24). Here, SNP-set annotations are based on gene boundaries defined by the NCBI’s RefSeq database in the UCSC Genome Browser (195). Unannotated SNPs located within the same genomic region were labeled as being within the “intergenic region” between two genes. Posterior inclusion probabilities (PIP) for the input and hidden layer weights are derived by fitting the BANNs model on individual-level data. A SNP-set is considered significant if it has a PIP(g) ≥ 0.5 (i.e., the “median probability model” threshold (6)). We take these significant SNP-sets and conduct “gene set enrichment analysis” using Enrichr (34; 129) to identify the categories they overrepresent in (a, b) the database of Genotypes and Phenotypes (dbGaP) and (c, d) the GWAS Catalog (2019). Nearly all enriched categories are related with (a, c) HDL and (b, d) LDL, respectively. Note that in LDL, the BANNs framework again identifies the gene APOB as having a high PIP (replicating the finding in the Framingham Heart Study). There have been hypotheses connecting LDL to cognitive traits (35; 264), and APOB has been shown to be related to cerebrospinal fluid and memory (192; 120; 107). Therefore, we argue that results in panel (b) are also relevant (a similar argument can be made for Fig. B.33).
Figure B.33. Gene set enrichment analyses using the significant SNP-sets identified by BANNs for high-density and low-density lipoprotein (HDL and LDL, respectively) traits in ten thousand randomly sampled individuals of European ancestry from the UK Biobank (24). Here, SNP-set annotations are based on gene boundaries defined by the NCBI’s RefSeq database in the UCSC Genome Browser (195). Unannotated SNPs located within the same genomic region were labeled as being within the “intergenic region” between two genes. In this analysis, each gene boundary annotation is modified by adding SNPs within a ±500 kilobase (kb) buffer to account for possible regulatory elements. Posterior inclusion probabilities (PIP) for the input and hidden layer weights are derived by fitting the BANNs model on individual-level data. A SNP-set is considered significant if it has a PIP($\geq$ 0.5 (i.e., the “median probability model” threshold (6)). We take these significant SNP-sets and conduct “gene set enrichment analysis” using Enrichr (34; 129) to identify the categories they overrepresent in (a, b) the database of Genotypes and Phenotypes (dbGaP) and (c, d) the GWAS Catalog (2019). Note that for panel (a), BANNs did not find many enriched SNP-sets with PIPs meeting the “median probability model” threshold and so we used a lower SNP-set threshold (PIP $\geq$ 0.1) to enable Enrichr to find associated dbGaP categories.
## B.13 Supplementary Tables

<table>
<thead>
<tr>
<th>Trait Type</th>
<th>Metric</th>
<th>BANN</th>
<th>BANN-SS</th>
<th>SuSiE (High)</th>
<th>SuSiE (Low)</th>
<th>CAVIAR</th>
<th>FINEMAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sparse</td>
<td>Power</td>
<td>0.614 (0.106)</td>
<td>0.609 (0.106)</td>
<td>0.608 (0.087)</td>
<td>0.424 (0.103)</td>
<td>0.597 (0.167)</td>
<td>0.529 (0.156)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.211 (0.092)</td>
<td>0.210 (0.104)</td>
<td><strong>0.207 (0.052)</strong></td>
<td>0.512 (0.088)</td>
<td>0.331 (0.155)</td>
<td>0.346 (0.048)</td>
</tr>
<tr>
<td>Polygenic</td>
<td>Power</td>
<td>0.119 (0.039)</td>
<td><strong>0.120 (0.053)</strong></td>
<td>0.103 (0.021)</td>
<td>0.072 (0.037)</td>
<td>0.098 (0.032)</td>
<td>0.101 (0.044)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.224 (0.098)</td>
<td>0.227 (0.107)</td>
<td><strong>0.217 (0.034)</strong></td>
<td>0.491 (0.051)</td>
<td>0.244 (0.086)</td>
<td>0.257 (0.103)</td>
</tr>
<tr>
<td>Trait Type</td>
<td>Metric</td>
<td>BANN</td>
<td>BANN-SS</td>
<td>RSS</td>
<td>PEGASUS</td>
<td>SKAT</td>
<td>MAGMA</td>
</tr>
<tr>
<td>------------</td>
<td>--------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>Sparse</td>
<td>Power</td>
<td>0.608 (0.132)</td>
<td>0.602 (0.161)</td>
<td>0.609 (0.127)</td>
<td>0.672 (0.118)</td>
<td>0.543 (0.142)</td>
<td><strong>0.737 (0.123)</strong></td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td><strong>0.052 (0.107)</strong></td>
<td>0.067 (0.098)</td>
<td>0.088 (0.103)</td>
<td>0.514 (0.137)</td>
<td>0.466 (0.156)</td>
<td>0.571 (0.126)</td>
</tr>
<tr>
<td>Polygenic</td>
<td>Power</td>
<td>0.078 (0.027)</td>
<td>0.081 (0.041)</td>
<td>0.073 (0.024)</td>
<td>0.148 (0.032)</td>
<td>0.112 (0.042)</td>
<td><strong>0.152 (0.039)</strong></td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.166 (0.151)</td>
<td>0.163 (0.126)</td>
<td><strong>0.065 (0.112)</strong></td>
<td>0.179 (0.098)</td>
<td>0.181 (0.101)</td>
<td>0.191 (0.018)</td>
</tr>
</tbody>
</table>

Table B.1. Comparing the empirical power and false discovery rates (FDR) of the BANNs framework against competing SNP and SNP-set mapping approaches in simulations. Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with only contributions from additive effects set (i.e., $\rho = 1$). We consider two different trait architectures: sparse where only 1% of SNP-sets are enriched for the trait; and polygenic where 10% of SNP-sets are enriched. We set the number of causal SNPs with non-zero effects to be 1% and 10% of all SNPs located within the enriched SNP-sets, respectively. (Top) Competing SNP-level mapping approaches include: CAVIAR (104), SuSiE (262), and FINEMAP (11). The software for SuSiE requires an input $\ell$ which fixes the maximum number of causal SNPs in the model. We display results when this input number is high ($\ell = 3000$) and when this input number is low ($\ell = 10$). (Bottom) Competing SNP-set mapping approaches include: RSS (307), PEGASUS (183), GBJ (237), SKAT (281), GSEA (101), and MAGMA (56). Results for the BANN, BANN-SS, and other Bayesian methods are evaluated based on the “median probability criterion” (i.e., PIPs for SNPs and SNP-sets greater than 0.5) (6). Results for the frequentist approaches are based on Bonferroni-corrected thresholds for multiple hypothesis testing ($P = 0.05/36518 = 1.37 \times 10^{-6}$ at the SNP-level and $P = 0.05/2816 = 1.78 \times 10^{-5}$ at the SNP-set level, respectively). All results are based on 100 replicates and standard deviations of the estimates across runs are given in the parentheses. Approaches with the greatest power are bolded in purple, while methods with the lowest FDR is bolded in blue.
### SNP-Level Approaches

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

### SNP-Set Level Approaches

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

Table B.2. Comparing the empirical power and false discovery rates (FDR) of the BANNs framework against competing SNP and SNP-set mapping approaches in simulations. Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.6$ with only contributions from additive effects set (i.e., $\rho = 1$). We consider two different trait architectures: sparse where only 1% of SNP-sets are enriched for the trait; and polygenic where 10% of SNP-sets are enriched. We set the number of causal SNPs with non-zero effects to be 1% and 10% of all SNPs located within the enriched SNP-sets, respectively. (Top) Competing SNP-level mapping approaches include: CAVIAR (104), SuSiE (262), and FINEMAP (11). The software for SuSiE requires an input $\ell$ which fixes the maximum number of causal SNPs in the model. We display results when this input number is high ($\ell = 3000$) and when this input number is low ($\ell = 10$). (Bottom) Competing SNP-set mapping approaches include: RSS (307), PEGASUS (183), GBJ (237), SKAT (281), GSEA (101), and MAGMA (56). Results for the BANN, BANN-SS, and other Bayesian methods are evaluated based on the “median probability criterion” (i.e., PIPs for SNPs and SNP-sets greater than 0.5) (6). Results for the frequentist approaches are based on Bonferroni-corrected thresholds for multiple hypothesis testing ($P = 0.05/36518 = 1.37 \times 10^{-6}$ at the SNP-level and $P = 0.05/2816 = 1.78 \times 10^{-5}$ at the SNP-set level, respectively). All results are based on 100 replicates and standard deviations of the estimates across runs are given in the parentheses. Approaches with the greatest power are bolded in purple, while methods with the lowest FDR is bolded in blue.
<table>
<thead>
<tr>
<th>Trait Type</th>
<th>Metric</th>
<th>BANN</th>
<th>BANN-SS</th>
<th>SuSiE (High)</th>
<th>SuSiE (Low)</th>
<th>CAVIAR</th>
<th>FINEMAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sparse</td>
<td>Power</td>
<td>(0.621\ (0.132))</td>
<td>0.613 (0.100)</td>
<td>0.602 (0.084)</td>
<td>0.591 (0.122)</td>
<td>0.585 (0.094)</td>
<td>0.569 (0.117)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.219 (0.094)</td>
<td>0.214 (0.104)</td>
<td>(0.173 \ (0.099))</td>
<td>0.191 (0.056)</td>
<td>0.184 (0.080)</td>
<td>0.221 (0.103)</td>
</tr>
<tr>
<td>Polygenic</td>
<td>Power</td>
<td>0.120 (0.103)</td>
<td>0.117 (0.059)</td>
<td>(0.127 \ (0.088))</td>
<td>0.081 (0.070)</td>
<td>0.109 (0.041)</td>
<td>0.091 (0.022)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.194 (0.090)</td>
<td>0.223 (0.103)</td>
<td>0.206 (0.061)</td>
<td>0.274 (0.092)</td>
<td>0.256 (0.106)</td>
<td>0.321 (0.092)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trait Type</th>
<th>Metric</th>
<th>BANN</th>
<th>BANN-SS</th>
<th>RSS</th>
<th>PEGASUS</th>
<th>SKAT</th>
<th>MAGMA</th>
<th>GSEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sparse</td>
<td>Power</td>
<td>(0.616 \ (0.112))</td>
<td>0.602 (0.126)</td>
<td>0.610 (0.101)</td>
<td>0.573 (0.156)</td>
<td>0.536 (0.124)</td>
<td>0.549 (0.119)</td>
<td>0.498 (0.090)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.129 (0.093)</td>
<td>0.182 (0.081)</td>
<td>(0.097 \ (0.107))</td>
<td>0.521 (0.133)</td>
<td>0.512 (0.231)</td>
<td>0.617 (0.216)</td>
<td>0.633 (0.105)</td>
</tr>
<tr>
<td>Polygenic</td>
<td>Power</td>
<td>(0.102 \ (0.108))</td>
<td>0.096 (0.104)</td>
<td>(0.102 \ (0.117))</td>
<td>0.099 (0.018)</td>
<td>0.063 (0.102)</td>
<td>0.078 (0.053)</td>
<td>0.057 (0.022)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.228 (0.084)</td>
<td>0.186 (0.092)</td>
<td>0.231 (0.097)</td>
<td>0.443 (0.201)</td>
<td>0.502 (0.310)</td>
<td>0.491 (0.199)</td>
<td>0.388 (0.178)</td>
</tr>
</tbody>
</table>

Table B.3. Comparing the empirical power and false discovery rates (FDR) of the BANNs framework against competing SNP and SNP-set mapping approaches in simulations with population structure (European cohort). Here, quantitative traits are simulated to have broad-sense heritability of \(H^2 = 0.2\) with only contributions from additive effects set (i.e., \(\rho = 1\)). We consider two different trait architectures: sparse where only 1% of SNP-sets are enriched for the trait; and polygenic where 10% of SNP-sets are enriched. We set the number of causal SNPs with non-zero effects to be 1% and 10% of all SNPs located within the enriched SNP-sets, respectively. (Top) Competing SNP-level mapping approaches include: CAVIAR (104), SuSiE (262), and FINEMAP (11). The software for SuSiE requires an input \(\ell\) which fixes the maximum number of causal SNPs in the model. We display results when this input number is high (\(\ell = 3000\)) and when this input number is low (\(\ell = 10\)). (Bottom) Competing SNP-set mapping approaches include: RSS (307), PEGASUS (183), GBJ (237), SKAT (281), GSEA (101), and MAGMA (56). Results for the BANN, BANN-SS, and other Bayesian methods are evaluated based on the “median probability criterion” (i.e., PIPs for SNPs and SNP-sets greater than 0.5) (6). Results for the frequentist approaches are based on Bonferroni-corrected thresholds for multiple hypothesis testing (\(P = 0.05/36518 = 1.37 \times 10^{-6}\) at the SNP-level and \(P = 0.05/2816 = 1.78 \times 10^{-5}\) at the SNP-set level, respectively). All results are based on 100 replicates and standard deviations of the estimates across runs are given in the parentheses. Approaches with the greatest power are bolded in purple, while methods with the lowest FDR is bolded in blue.
### SNP-Level Approaches

<table>
<thead>
<tr>
<th>Trait Type</th>
<th>Metric</th>
<th>BANN</th>
<th>BANN-SS</th>
<th>SuSiE (High)</th>
<th>SuSiE (Low)</th>
<th>CAVIAR</th>
<th>FINEMAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sparse</td>
<td>Power</td>
<td>0.822 (0.113)</td>
<td>0.818 (0.092)</td>
<td>0.811 (0.071)</td>
<td>0.623 (0.103)</td>
<td>0.787 (0.109)</td>
<td>0.762 (0.121)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.118 (0.074)</td>
<td>0.201 (0.104)</td>
<td>0.123 (0.099)</td>
<td>0.461 (0.056)</td>
<td>0.227 (0.080)</td>
<td>0.203 (0.103)</td>
</tr>
<tr>
<td>Polygenic</td>
<td>Power</td>
<td>0.335 (0.042)</td>
<td>0.317 (0.059)</td>
<td>0.326 (0.088)</td>
<td>0.122 (0.070)</td>
<td>0.301 (0.041)</td>
<td>0.298 (0.031)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.262 (0.090)</td>
<td>0.341 (0.103)</td>
<td>0.228 (0.061)</td>
<td>0.388 (0.092)</td>
<td>0.211 (0.106)</td>
<td>0.294 (0.092)</td>
</tr>
</tbody>
</table>

### SNP-Set Level Approaches

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

Table B.4. Comparing the empirical power and false discovery rates (FDR) of the BANNs framework against competing SNP and SNP-set mapping approaches in simulations with population structure (European cohort). Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.6$ with only contributions from additive effects set (i.e., $\rho = 1$). We consider two different trait architectures: sparse where only 1% of SNP-sets are enriched for the trait; and polygenic where 10% of all SNPs are enriched. We set the number of causal SNPs with non-zero effects to be 1% and 10% of all SNPs located within the enriched SNP-sets, respectively. **(Top)** Competing SNP-level mapping approaches include: CAVIAR (104), SuSiE (262), and FINEMAP (11). The software for SuSiE requires an input $\ell$ which fixes the maximum number of causal SNPs in the model. We display results when this input number is high ($\ell = 3000$) and when this input number is low ($\ell = 10$). **(Bottom)** Competing SNP-set mapping approaches include: RSS (307), PEGASUS (183), GBJ (237), SKAT (281), GSEA (101), and MAGMA (56). Results for the BANN, BANN-SS, and other Bayesian methods are evaluated based on the “median probability criterion” (i.e., PIPs for SNPs and SNP-sets greater than 0.5) (6). Results for the frequentist approaches are based on Bonferroni-corrected thresholds for multiple hypothesis testing ($P = 0.05 / 36518 = 1.37 \times 10^{-6}$ at the SNP-level and $P = 0.05 / 2816 = 1.78 \times 10^{-5}$ at the SNP-set level, respectively). All results are based on 100 replicates and standard deviations of the estimates across runs are given in the parentheses. Approaches with the greatest power are bolded in purple, while methods with the lowest FDR is bolded in blue.
Table B.5. Comparing the empirical power and false discovery rates (FDR) of the BANNs framework against competing SNP and SNP-set mapping approaches in simulations. Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with contributions from both additive and epistatic effects set (i.e., $\rho = 0.5$). We consider two different trait architectures: sparse where only 1% of SNP-sets are enriched for the trait; and polygenic where 10% of SNP-sets are enriched. We set the number of causal SNPs with non-zero effects to be 1% and 10% of all SNPs located within the enriched SNP-sets, respectively. (Top) Competing SNP-level mapping approaches include: CAVIAR (104), SuSiE (262), and FINEMAP (11). The software for SuSiE requires an input $\ell$ which fixes the maximum number of causal SNPs in the model. We display results when this input number is high ($\ell = 3000$) and when this input number is low ($\ell = 10$). (Bottom) Competing SNP-set mapping approaches include: RSS (307), PEGASUS (183), GBJ (237), SKAT (281), GSEA (101), and MAGMA (56). Results for the BANN, BANN-SS, and other Bayesian methods are evaluated based on the “median probability criterion” (i.e., PIPs for SNPs and SNP-sets greater than 0.5) (6). Results for the frequentist approaches are based on Bonferroni-corrected thresholds for multiple hypothesis testing ($P = 0.05/36518 = 1.37 \times 10^{-6}$ at the SNP-level and $P = 0.05/2816 = 1.78 \times 10^{-5}$ at the SNP-set level, respectively). All results are based on 100 replicates and standard deviations of the estimates across runs are given in the parentheses. Approaches with the greatest power are bolded in purple, while methods with the lowest FDR is bolded in blue.
### SNP-Level Approaches

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

### SNP-Set Level Approaches

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

Table B.6. Comparing the empirical power and false discovery rates (FDR) of the BANNs framework against competing SNP and SNP-set mapping approaches in simulations. Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.6$ with contributions from both additive and epistatic effects set (i.e., $\rho = 0.5$). We consider two different trait architectures: sparse where only 1% of SNP-sets are enriched for the trait; and polygenic where 10% of SNP-sets are enriched. We set the number of causal SNPs with non-zero effects to be 1% and 10% of all SNPs located within the enriched SNP-sets, respectively. (Top) Competing SNP-level mapping approaches include: CAVIAR (104), SuSiE (262), and FINEMAP (11). The software for SuSiE requires an input $\ell$ which fixes the maximum number of causal SNPs in the model. We display results when this input number is high ($\ell = 3000$) and when this input number is low ($\ell = 10$). (Bottom) Competing SNP-set mapping approaches include: RSS (307), PEGASUS (183), GBJ (237), SKAT (281), GSEA (101), and MAGMA (56). Results for the BANN, BANN-SS, and other Bayesian methods are evaluated based on the “median probability criterion” (i.e., PIPs for SNPs and SNP-sets greater than 0.5) (6). Results for the frequentist approaches are based on Bonferroni-corrected thresholds for multiple hypothesis testing ($P = 0.05/36518 = 1.37 \times 10^{-6}$ at the SNP-level and $P = 0.05/2816 = 1.78 \times 10^{-5}$ at the SNP-set level, respectively). All results are based on 100 replicates and standard deviations of the estimates across runs are given in the parentheses. Approaches with the greatest power are bolded in purple, while methods with the lowest FDR is bolded in blue.
### SNP-Level Approaches

<table>
<thead>
<tr>
<th>Trait Type</th>
<th>Metric</th>
<th>BANN</th>
<th>BANN-SS</th>
<th>SuSiE (High)</th>
<th>SuSiE (Low)</th>
<th>CAVIAR</th>
<th>FINEMAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sparse</td>
<td>Power</td>
<td>0.582 (0.114)</td>
<td>0.491 (0.092)</td>
<td>0.526 (0.071)</td>
<td>0.453 (0.103)</td>
<td>0.474 (0.109)</td>
<td>0.491 (0.121)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.118 (0.094)</td>
<td>0.201 (0.104)</td>
<td>0.123 (0.099)</td>
<td>0.461 (0.056)</td>
<td>0.227 (0.080)</td>
<td>0.203 (0.103)</td>
</tr>
<tr>
<td>Polygenic</td>
<td>Power</td>
<td>0.112 (0.103)</td>
<td>0.087 (0.059)</td>
<td>0.094 (0.088)</td>
<td>0.042 (0.070)</td>
<td>0.072 (0.041)</td>
<td>0.064 (0.022)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.262 (0.090)</td>
<td>0.341 (0.103)</td>
<td>0.228 (0.061)</td>
<td>0.388 (0.092)</td>
<td>0.211 (0.106)</td>
<td>0.294 (0.092)</td>
</tr>
</tbody>
</table>

### SNP-Set Level Approaches

<table>
<thead>
<tr>
<th>Trait Type</th>
<th>Metric</th>
<th>BANN</th>
<th>BANN-SS</th>
<th>RSS</th>
<th>PEGASUS</th>
<th>SKAT</th>
<th>MAGMA</th>
<th>GSEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sparse</td>
<td>Power</td>
<td>0.544 (0.076)</td>
<td>0.506 (0.053)</td>
<td>0.509 (0.101)</td>
<td>0.498 (0.122)</td>
<td>0.388 (0.124)</td>
<td>0.431 (0.119)</td>
<td>0.320 (0.090)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.132 (0.111)</td>
<td>0.196 (0.097)</td>
<td>0.130 (0.104)</td>
<td>0.564 (0.204)</td>
<td>0.551 (0.120)</td>
<td>0.427 (0.118)</td>
<td>0.547 (0.221)</td>
</tr>
<tr>
<td>Polygenic</td>
<td>Power</td>
<td>0.076 (0.123)</td>
<td>0.047 (0.104)</td>
<td>0.051 (0.087)</td>
<td>0.048 (0.018)</td>
<td>0.036 (0.102)</td>
<td>0.043 (0.053)</td>
<td>0.022 (0.022)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.241 (0.052)</td>
<td>0.238 (0.092)</td>
<td>0.212 (0.097)</td>
<td>0.397 (0.201)</td>
<td>0.484 (0.310)</td>
<td>0.563 (0.199)</td>
<td>0.488 (0.178)</td>
</tr>
</tbody>
</table>

Table B.7. Comparing the empirical power and false discovery rates (FDR) of the BANNs framework against competing SNP and SNP-set mapping approaches in simulations with population structure (European cohort). Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.2$ with contributions from both additive and epistatic effects set (i.e., $\rho = 0.5$). We consider two different trait architectures: sparse where only 1% of SNP-sets are enriched for the trait; and polygenic where 10% of SNP-sets are enriched. We set the number of causal SNPs with non-zero effects to be 1% and 10% of all SNPs located within the enriched SNP-sets, respectively. **(Top)** Competing SNP-level mapping approaches include: CAVIAR (104), SuSiE (262), and FINEMAP (11). The software for SuSiE requires an input $\ell$ which fixes the maximum number of causal SNPs in the model. We display results when this input number is high ($\ell = 3000$) and when this input number is low ($\ell = 10$). **(Bottom)** Competing SNP-set mapping approaches include: RSS (307), PEGASUS (183), GBJ (237), SKAT (281), GSEA (101), and MAGMA (56). Results for the BANN, BANN-SS, and other Bayesian methods are evaluated based on the “median probability criterion” (i.e., PIPs for SNPs and SNP-sets greater than 0.5) (6). Results for the frequentist approaches are based on Bonferroni-corrected thresholds for multiple hypothesis testing ($P = 0.05 / 36518 = 1.37 \times 10^{-6}$ at the SNP-level and $P = 0.05 / 2816 = 1.78 \times 10^{-5}$ at the SNP-set level, respectively). All results are based on 100 replicates and standard deviations of the estimates across runs are given in the parentheses. Approaches with the greatest power are bolded in purple, while methods with the lowest FDR is bolded in blue.
### SNP-Level Approaches

<table>
<thead>
<tr>
<th>Trait Type</th>
<th>Metric</th>
<th>BANN</th>
<th>BANN-SS</th>
<th>SuSiE (High)</th>
<th>SuSiE (Low)</th>
<th>CAVIAR</th>
<th>FINEMAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sparse</td>
<td>Power</td>
<td>0.817 (0.126)</td>
<td>0.798 (0.117)</td>
<td>0.792 (0.092)</td>
<td>0.563 (0.104)</td>
<td>0.752 (0.134)</td>
<td>0.726 (0.128)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.182 (0.038)</td>
<td>0.191 (0.045)</td>
<td>0.346 (0.057)</td>
<td>0.467 (0.075)</td>
<td>0.237 (0.076)</td>
<td>0.282 (0.084)</td>
</tr>
<tr>
<td>Polygenic</td>
<td>Power</td>
<td>0.348 (0.109)</td>
<td>0.319 (0.094)</td>
<td>0.305 (0.081)</td>
<td>0.211 (0.039)</td>
<td>0.327 (0.093)</td>
<td>0.338 (0.053)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.239 (0.047)</td>
<td>0.221 (0.038)</td>
<td>0.224 (0.042)</td>
<td>0.385 (0.035)</td>
<td>0.309 (0.041)</td>
<td>0.325 (0.091)</td>
</tr>
</tbody>
</table>

### SNP-Set Level Approaches

<table>
<thead>
<tr>
<th>Trait Type</th>
<th>Metric</th>
<th>BANN</th>
<th>BANN-SS</th>
<th>RSS</th>
<th>PEGASUS</th>
<th>SKAT</th>
<th>MAGMA</th>
<th>GSEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sparse</td>
<td>Power</td>
<td>0.745 (0.093)</td>
<td>0.698 (0.102)</td>
<td>0.702 (0.089)</td>
<td>0.634 (0.104)</td>
<td>0.521 (0.063)</td>
<td>0.561 (0.097)</td>
<td>0.481 (0.128)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.146 (0.099)</td>
<td>0.118 (0.097)</td>
<td>0.152 (0.107)</td>
<td>0.311 (0.132)</td>
<td>0.275 (0.126)</td>
<td>0.262 (0.088)</td>
<td>0.288 (0.101)</td>
</tr>
<tr>
<td>Polygenic</td>
<td>Power</td>
<td>0.326 (0.123)</td>
<td>0.273 (0.104)</td>
<td>0.282 (0.094)</td>
<td>0.382 (0.077)</td>
<td>0.216 (0.122)</td>
<td>0.262 (0.085)</td>
<td>0.117 (0.035)</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.254 (0.052)</td>
<td>0.266 (0.092)</td>
<td>0.261 (0.097)</td>
<td>0.461 (0.201)</td>
<td>0.392 (0.310)</td>
<td>0.280 (0.199)</td>
<td>0.457 (0.178)</td>
</tr>
</tbody>
</table>

Table B.8. Comparing the empirical power and false discovery rates (FDR) of the BANNs framework against competing SNP and SNP-set mapping approaches in simulations with population structure (European cohort). Here, quantitative traits are simulated to have broad-sense heritability of $H^2 = 0.6$ with contributions from both additive and epistatic effects set (i.e., $\rho = 0.5$). We consider two different trait architectures: sparse where only 1% of SNP-sets are enriched for the trait; and polygenic where 10% of SNP-sets are enriched. We set the number of causal SNPs with non-zero effects to be 1% and 10% of all SNPs located within the enriched SNP-sets, respectively. **(Top)** Competing SNP-level mapping approaches include: CAVIAR (104), SuSiE (262), and FINEMAP (11). The software for SuSiE requires an input $\ell$ which fixes the maximum number of causal SNPs in the model. We display results when this input number is high ($\ell = 3000$) and when this input number is low ($\ell = 10$). **(Bottom)** Competing SNP-set mapping approaches include: RSS (307), PEGASUS (183), GBJ (237), SKAT (281), GSEA (101), and MAGMA (56). Results for the BANN, BANN-SS, and other Bayesian methods are evaluated based on the “median probability criterion” (i.e., PIPs for SNPs and SNP-sets greater than 0.5) (6). Results for the frequentist approaches are based on Bonferroni-corrected thresholds for multiple hypothesis testing ($P = 0.05/36518 = 1.37 \times 10^{-6}$ at the SNP-level and $P = 0.05/2816 = 1.78 \times 10^{-5}$ at the SNP-set level, respectively). All results are based on 100 replicates and standard deviations of the estimates across runs are given in the parentheses. Approaches with the greatest power are bolded in purple, while methods with the lowest FDR is bolded in blue.
Simulation Parameters | Average Run Time (seconds)
--- | ---
| **SNPs** | **Samples Sizes** | BANN | SuSiE (low) | SuSiE (high) | CAVIAR | FINEMAP |
| 2500 | 1000 | 3.34 | 1.89 | 4.22 | 8.21 | 56.99 |
| 2000 | | 6.71 | 2.87 | 8.72 | 8.21 | 56.99 |
| 4000 | | 10.82 | 8.42 | 13.63 | 8.21 | 56.99 |
| 5000 | 1000 | 7.42 | 2.49 | 7.12 | 31.48 | 102.58 |
| 2000 | | 13.21 | 5.04 | 21.84 | 31.48 | 102.58 |
| 4000 | | 21.34 | 9.45 | 32.81 | 31.48 | 102.58 |
| 10000 | 1000 | 31.39 | 3.52 | 52.24 | 118.98 | 145.51 |
| 2000 | | 127.18 | 10.22 | 159.97 | 118.98 | 145.51 |
| 4000 | | 318.81 | 22.62 | 754.63 | 118.98 | 145.51 |

Table B.9. Computational time for running Bayesian annotated neural networks (BANNs) and other SNP-level association mapping approaches, as a function of the total number SNPs analyzed and the number of samples in the data. Methods compared include: BANNs, CAVIAR (104), SuSiE (262), and FINEMAP (11). Each table entry represents the average computation time (in seconds) it takes each approach to analyze a dataset of the size indicated. Run times were measured on an Intel i5-8259U CPU with base frequency of 2.30GHz, turbo frequency of 3.80GHz, and memory 16GB 2133 MHz LPDDR3. Here, we used 4 cores for parallelization when applicable. The software for SuSiE requires an input $\ell$ which fixes the maximum number of causal SNPs in the model. We display results when this input parameter is high ($\ell = 3000$) and when this input parameter is low ($\ell = 10$). Note that we implemented BANNs using the Python 3 version of the software, and the timing for its variational algorithm includes inference on both SNPs and SNP-sets. CAVIAR and FINEMAP are set up to work with GWA summary statistics, so their inputs (and timing) are the same irrespective of the sample size.
<table>
<thead>
<tr>
<th>SNP-Sets</th>
<th>SNPs per SNP-set</th>
<th>BANN</th>
<th>RSS</th>
<th>PEGASUS</th>
<th>GBJ</th>
<th>SKAT</th>
<th>MAGMA</th>
<th>GSEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td>10</td>
<td>12.58</td>
<td>13.12</td>
<td>2.41</td>
<td>2.68</td>
<td>2.13</td>
<td>0.03</td>
<td>2.48</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>44.32</td>
<td>58.21</td>
<td>2.13</td>
<td>5.18</td>
<td>3.82</td>
<td>0.08</td>
<td>4.68</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>189.44</td>
<td>224.62</td>
<td>2.22</td>
<td>9.64</td>
<td>6.47</td>
<td>0.18</td>
<td>8.51</td>
</tr>
<tr>
<td>500</td>
<td>10</td>
<td>48.92</td>
<td>59.31</td>
<td>5.11</td>
<td>5.37</td>
<td>5.23</td>
<td>0.09</td>
<td>5.31</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>223.14</td>
<td>244.07</td>
<td>5.02</td>
<td>11.26</td>
<td>9.22</td>
<td>0.21</td>
<td>11.12</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>965.48</td>
<td>1026.12</td>
<td>5.72</td>
<td>27.91</td>
<td>14.84</td>
<td>0.24</td>
<td>20.36</td>
</tr>
<tr>
<td>1000</td>
<td>10</td>
<td>194.62</td>
<td>249.57</td>
<td>8.67</td>
<td>12.27</td>
<td>11.31</td>
<td>0.72</td>
<td>11.41</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>1213.19</td>
<td>2176.33</td>
<td>8.93</td>
<td>27.62</td>
<td>18.16</td>
<td>1.48</td>
<td>24.93</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>6823.31</td>
<td>14495.72</td>
<td>10.21</td>
<td>61.37</td>
<td>30.83</td>
<td>4.26</td>
<td>60.82</td>
</tr>
</tbody>
</table>

Table B.10. Computational time for running Bayesian annotated neural networks (BANNs) and other SNP-set level enrichment approaches, as a function of the total number SNP-sets analyzed and the number of SNPs within each SNP-set. Methods compared include: BANNs, RSS (307), PEGASUS (183), GBJ (237), SKAT (281), GSEA (101), and MAGMA (56). Here, we simulated 10 datasets for each pair of parameter values (number of SNP-sets analyzed and number of SNPs within each SNP-set). Sample size was held constant at $n = 10,000$ individuals. Each table entry represents the average computation time (in seconds) it takes each approach to analyze a dataset of the size indicated. Run times were measured on an Intel i5-8259U CPU with base frequency of 2.30GHz, turbo frequency of 3.80GHz, and memory 16GB 2133 MHz LPDDR3. Here, we used 4 cores for parallelization when applicable. Note that PEGASUS, GBJ, SKAT, and MAGMA are score-based methods and, thus, are expected to take the least amount of time to run. Both the BANNs framework and RSS are regression-based methods. The increased computational burden of these approaches results from its need to do (approximate) Bayesian posterior inference; however, the sparse and partially connected architecture of the BANNs model allows it to scale more favorably for larger dimensional datasets. Note that we implemented BANNs using the Python 3 version of the software, and the timing for its variational algorithm includes inference on both SNPs and SNP-sets.
Table B.11. SNP and SNP-set results for body mass index (BMI) in the heterogenous stock of mice from the Wellcome Trust Centre for Human Genetics. We analyze $J \approx 10,000$ SNPs and $G = 1,925$ SNP-sets from $N = 1,814$ mice—with specific numbers varying slightly depending on the quality control procedure for each phenotype (Supporting Information, Section B.6). Here, SNP-set annotations are based on gene boundaries defined by the Mouse Genome Informatics database (see URLs listed in the main text). Unannotated SNPs located within the same genomic region were labeled as being within the “intergenic region” between two genes. This file gives the posterior inclusion probabilities (PIPs) for the input and hidden layer neural network weights after fitting the BANNs model on the individual-level data. We assess significance for both SNPs and SNP-sets according to the “median probability model” threshold (6) (i.e., $\text{PIP} \geq 0.5$). Page #1 provides the variant-level association mapping results with columns corresponding to: (1) chromosome; (2) SNP ID; (3) chromosomal position in base-pair (bp) coordinates; (4) SNP PIP; and (5) SuSiE PIP, which corresponds to SNP-level posterior inclusion probabilities computed by SuSiE (262). Page #2 provides the SNP-set level enrichment results with columns corresponding to: (1) chromosome; (2) SNP-set ID; (3-4) the starting and ending position of the SNP-set chromosomal boundaries; (5) SNP-set PIP; (6) RSS PIP, which corresponds to the posterior inclusion probabilities computed by RSS (307); (7) the number of SNPs that have been annotated within each SNP-set; (8) the “top” associated SNP within each SNP-set; (9) the PIP of each top SNP.

Table B.12. SNP and SNP-set results for body weight in the heterogenous stock of mice from the Wellcome Trust Centre for Human Genetics. We analyze $J \approx 10,000$ SNPs and $G = 1,925$ SNP-sets from $N = 1,814$ mice—with specific numbers varying slightly depending on the quality control procedure for each phenotype (Supporting Information, Section B.6). Here, SNP-set annotations are based on gene boundaries defined by the Mouse Genome Informatics database (see URLs in the main text). Unannotated SNPs located within the same genomic region were labeled as being within the “intergenic region” between two genes. This file gives the posterior inclusion probabilities (PIPs) for the input and hidden layer neural network weights after fitting the BANNs model on the individual-level data. We assess significance for both SNPs and SNP-sets according to the “median probability model” threshold (6) (i.e., $\text{PIP} \geq 0.5$). Page #1 provides the variant-level association mapping results with columns corresponding to: (1) chromosome; (2) SNP ID; (3) chromosomal position in base-pair (bp) coordinates; (4) SNP PIP; and (5) SuSiE PIP, which corresponds to SNP-level posterior inclusion probabilities computed by SuSiE (262). Page #2 provides the SNP-set level enrichment results with columns corresponding to: (1) chromosome; (2) SNP-set ID; (3-4) the starting and ending position of the SNP-set chromosomal boundaries; (5) SNP-set PIP; (6) RSS PIP, which corresponds to the posterior inclusion probabilities computed by RSS (307); (7) the number of SNPs that have been annotated within each SNP-set; (8) the “top” associated SNP within each SNP-set; (9) the PIP of each top SNP.
Table B.13. SNP and SNP-set results for percentage of CD8+ cells in the heterogenous stock of mice from the Wellcome Trust Centre for Human Genetics. We analyze $J \approx 10,000$ SNPs and $G = 1,925$ SNP-sets from $N = 1,814$ mice—with specific numbers varying slightly depending on the quality control procedure for each phenotype (Supporting Information, Section B.6). Here, SNP-set annotations are based on gene boundaries defined by the Mouse Genome Informatics database (see URLs listed in the main text). Unannotated SNPs located within the same genomic region were labeled as being within the “intergenic region” between two genes. This file gives the posterior inclusion probabilities (PIPs) for the input and hidden layer neural network weights after fitting the BANNs model on the individual-level data. We assess significance for both SNPs and SNP-sets according to the “median probability model” threshold (6) (i.e., PIP $\geq 0.5$). Page #1 provides the variant-level association mapping results with columns corresponding to: (1) chromosome; (2) SNP ID; (3) chromosomal position in base-pair (bp) coordinates; (4) SNP PIP; and (5) SuSiE PIP, which corresponds to SNP-level posterior inclusion probabilities computed by SuSiE (262). Page #2 provides the SNP-set level enrichment results with columns corresponding to: (1) chromosome; (2) SNP-set ID; (3-4) the starting and ending position of the SNP-set chromosomal boundaries; (5) SNP-set PIP; (6) RSS PIP, which corresponds to the posterior inclusion probabilities computed by RSS (307); (7) the number of SNPs that have been annotated within each SNP-set; (8) the “top” associated SNP within each SNP-set; (9) the PIP of each top SNP.

Table B.14. SNP and SNP-set results for high-density lipoprotein (HDL) cholesterol in the heterogenous stock of mice from the Wellcome Trust Centre for Human Genetics. We analyze $J \approx 10,000$ SNPs and $G = 1,925$ SNP-sets from $N = 1,814$ mice—with specific numbers varying slightly depending on the quality control procedure for each phenotype (Supporting Information, Section B.6). Here, SNP-set annotations are based on gene boundaries defined by the Mouse Genome Informatics database (see URLs listed in the main text). Unannotated SNPs located within the same genomic region were labeled as being within the “intergenic region” between two genes. This file gives the posterior inclusion probabilities (PIPs) for the input and hidden layer neural network weights after fitting the BANNs model on the individual-level data. We assess significance for both SNPs and SNP-sets according to the “median probability model” threshold (6) (i.e., PIP $\geq 0.5$). Page #1 provides the variant-level association mapping results with columns corresponding to: (1) chromosome; (2) SNP ID; (3) chromosomal position in base-pair (bp) coordinates; (4) SNP PIP; and (5) SuSiE PIP, which corresponds to SNP-level posterior inclusion probabilities computed by SuSiE (262). Page #2 provides the SNP-set level enrichment results with columns corresponding to: (1) chromosome; (2) SNP-set ID; (3-4) the starting and ending position of the SNP-set chromosomal boundaries; (5) SNP-set PIP; (6) RSS PIP, which corresponds to the posterior inclusion probabilities computed by RSS (307); (7) the number of SNPs that have been annotated within each SNP-set; (8) the “top” associated SNP within each SNP-set; (9) the PIP of each top SNP.
Table B.15. SNP and SNP-set results for low-density lipoprotein (LDL) cholesterol in the heterogenous stock of mice from the Wellcome Trust Centre for Human Genetics. We analyze $J \approx 10,000$ SNPs and $G = 1,925$ SNP-sets from $N = 1,814$ mice—with specific numbers varying slightly depending on the quality control procedure for each phenotype (Supporting Information, Section B.6). Here, SNP-set annotations are based on gene boundaries defined by the Mouse Genome Informatics database (see URLs listed in the main text). Unannotated SNPs located within the same genomic region were labeled as being within the “intergenic region” between two genes. This file gives the posterior inclusion probabilities (PIPs) for the input and hidden layer neural network weights after fitting the BANNs model on the individual-level data. We assess significance for both SNPs and SNP-sets according to the “median probability model” threshold (6) (i.e., $\text{PIP} \geq 0.5$). Page #1 provides the variant-level association mapping results with columns corresponding to: (1) chromosome; (2) SNP ID; (3) chromosomal position in base-pair (bp) coordinates; (4) SNP PIP; and (5) SuSiE PIP, which corresponds to SNP-level posterior inclusion probabilities computed by SuSiE (262). Page #2 provides the SNP-set level enrichment results with columns corresponding to: (1) chromosome; (2) SNP-set ID; (3-4) the starting and ending position of the SNP-set chromosomal boundaries; (5) SNP-set PIP; (6) RSS PIP, which corresponds to the posterior inclusion probabilities computed by RSS (307); (7) the number of SNPs that have been annotated within each SNP-set; (8) the “top” associated SNP within each SNP-set; (9) the PIP of each top SNP.

Table B.16. SNP and SNP-set results for mean corpuscular hemoglobin (MCH) in the heterogenous stock of mice from the Wellcome Trust Centre for Human Genetics. We analyze $J \approx 10,000$ SNPs and $G = 1,925$ SNP-sets from $N = 1,814$ mice—with specific numbers varying slightly depending on the quality control procedure for each phenotype (Supporting Information, Section B.6). Here, SNP-set annotations are based on gene boundaries defined by the Mouse Genome Informatics database (see URLs listed in the main text). Unannotated SNPs located within the same genomic region were labeled as being within the “intergenic region” between two genes. This file gives the posterior inclusion probabilities (PIPs) for the input and hidden layer neural network weights after fitting the BANNs model on the individual-level data. We assess significance for both SNPs and SNP-sets according to the “median probability model” threshold (6) (i.e., $\text{PIP} \geq 0.5$). Page #1 provides the variant-level association mapping results with columns corresponding to: (1) chromosome; (2) SNP ID; (3) chromosomal position in base-pair (bp) coordinates; (4) SNP PIP; and (5) SuSiE PIP, which corresponds to SNP-level posterior inclusion probabilities computed by SuSiE (262). Page #2 provides the SNP-set level enrichment results with columns corresponding to: (1) chromosome; (2) SNP-set ID; (3-4) the starting and ending position of the SNP-set chromosomal boundaries; (5) SNP-set PIP; (6) RSS PIP, which corresponds to the posterior inclusion probabilities computed by RSS (307); (7) the number of SNPs that have been annotated within each SNP-set; (8) the “top” associated SNP within each SNP-set; (9) the PIP of each top SNP.
Table B.17. Notable enriched SNP-sets after applying the BANNs framework to high-density and low-density lipoprotein (HDL and LDL, respectively) traits in the Framingham Heart Study (228) where each SNP-set annotation has been augmented with a ±500 kilobase (kb) buffer to account for possible regulatory elements. Here, SNP-set annotations are based on gene boundaries defined by the NCBI’s RefSeq database in the UCSC Genome Browser (195). Unannotated SNPs located within the same genomic region were labeled as being within the “intergenic region” between two genes. These regions are labeled as Gene1-Gene2 in the table. Posterior inclusion probabilities (PIP) for the input and hidden layer weights are derived by fitting the BANNs model on individual-level data. A SNP-set is considered enriched if it has a PIP($g$) ≥ 0.5 (i.e., the “median probability model” threshold (6)). We report the “top” associated SNP within each region and its corresponding PIP($j$). We also report the corresponding SNP and SNP-set level results after running SuSiE (262) and RSS (307) on these same traits, respectively. The last column details references and literature sources that have previously suggested some level of association or enrichment between the each genomic region and the traits of interest. See Tables B.18 and B.19 for the complete list of SNP and SNP-set level results. ♣: SNPs and SNP-sets replicated in an independent analysis of ten thousand randomly sampled individuals of European ancestry from the UK Biobank (24).

<table>
<thead>
<tr>
<th>Trait</th>
<th>SNP-Set</th>
<th>Chr</th>
<th>PIP($g$)</th>
<th>Rank</th>
<th>RSS PIP</th>
<th>RSS Rank</th>
<th>Top SNP</th>
<th>PIP($j$)</th>
<th>SuSiE PIP</th>
<th>SuSiE Rank</th>
<th>Ref(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDL</td>
<td>C10orf213-TCEA3</td>
<td>1</td>
<td>0.998</td>
<td>2</td>
<td>0.121</td>
<td>162</td>
<td>rs176141</td>
<td>0.298</td>
<td>0.120</td>
<td>19</td>
<td>(125; 61)</td>
</tr>
<tr>
<td></td>
<td>ST18-FAM150A</td>
<td>8</td>
<td>0.9169</td>
<td>4</td>
<td>0.959</td>
<td>4</td>
<td>rs11990693</td>
<td>0.999</td>
<td>0.277</td>
<td>8</td>
<td>(242)</td>
</tr>
<tr>
<td></td>
<td>LPL-SLC18A1</td>
<td>8</td>
<td>0.734</td>
<td>5</td>
<td>0.999</td>
<td>2</td>
<td>rs17482753</td>
<td>0.999</td>
<td>0.191</td>
<td>11</td>
<td>(43; 274)</td>
</tr>
<tr>
<td>LDL</td>
<td>AQP9-LIPC</td>
<td>15</td>
<td>0.913</td>
<td>1</td>
<td>0.198</td>
<td>29</td>
<td>rs4775041</td>
<td>0.422</td>
<td>0.238</td>
<td>4</td>
<td>(247; 244)</td>
</tr>
<tr>
<td></td>
<td>TRMU-CELSR1</td>
<td>22</td>
<td>0.678</td>
<td>2</td>
<td>0.895</td>
<td>3</td>
<td>rs5768939</td>
<td>0.379</td>
<td>0.185</td>
<td>5</td>
<td>(213; 275; 187; 267)</td>
</tr>
<tr>
<td></td>
<td>C10orf92-C10orf93</td>
<td>10</td>
<td>0.537</td>
<td>3</td>
<td>0.125</td>
<td>157</td>
<td>rs10781583</td>
<td>0.215</td>
<td>0.0377</td>
<td>26</td>
<td>—</td>
</tr>
</tbody>
</table>

Table B.17. Notable enriched SNP-sets after applying the BANNs framework to high-density and low-density lipoprotein (HDL and LDL, respectively) traits in the Framingham Heart Study (228) where each SNP-set annotation has been augmented with a ±500 kilobase (kb) buffer to account for possible regulatory elements. Here, SNP-set annotations are based on gene boundaries defined by the NCBI’s RefSeq database in the UCSC Genome Browser (195). Unannotated SNPs located within the same genomic region were labeled as being within the “intergenic region” between two genes. These regions are labeled as Gene1-Gene2 in the table. Posterior inclusion probabilities (PIP) for the input and hidden layer weights are derived by fitting the BANNs model on individual-level data. A SNP-set is considered enriched if it has a PIP($g$) ≥ 0.5 (i.e., the “median probability model” threshold (6)). We report the “top” associated SNP within each region and its corresponding PIP($j$). We also report the corresponding SNP and SNP-set level results after running SuSiE (262) and RSS (307) on these same traits, respectively. The last column details references and literature sources that have previously suggested some level of association or enrichment between the each genomic region and the traits of interest. See Tables B.18 and B.19 for the complete list of SNP and SNP-set level results. ♣: SNPs and SNP-sets replicated in an independent analysis of ten thousand randomly sampled individuals of European ancestry from the UK Biobank (24).
Table B.18. SNP and SNP-set results for high-density lipoprotein (HDL) cholesterol in individuals assayed within the Framingham Heart Study. We analyze \( J = 394,174 \) SNPs and \( G = 18,364 \) SNP-sets from \( N = 6,950 \) people. Here, SNP-set annotations are based on gene boundaries defined by the NCBI's RefSeq database in the UCSC Genome Browser (195). Unannotated SNPs located within the same genomic region were labeled as being within the “intergenic region” between two genes. This file gives the posterior inclusion probabilities (PIPs) for the input and hidden layer neural network weights after fitting the BANNs model on the individual-level data. We assess significance for both SNPs and SNP-sets according to the “median probability model” threshold (6) (i.e., \( PIP \geq 0.5 \)). Page #1 provides the variant-level association mapping results with columns corresponding to: (1) chromosome; (2) SNP ID; (3) chromosomal position in base-pair (bp) coordinates; (4) SNP PIP; and (5) SuSiE PIP, which corresponds to SNP-level posterior inclusion probabilities computed by SuSiE (262). Page #2 provides the SNP-set level enrichment results with columns corresponding to: (1) chromosome; (2) SNP-set ID; (3-4) the starting and ending position of the SNP-set chromosomal boundaries; (5) SNP-set PIP; (6) RSS PIP, which corresponds to the posterior inclusion probabilities computed by RSS (307); (7) the number of SNPs that have been annotated within each SNP-set; (8) the “top” associated SNP within each SNP-set; (9) the PIP of each top SNP. Pages #3 and #4 provide similar results based on analyses where each SNP-set annotation has been augmented with a \( \pm 500 \) kilobase (kb) buffer to account for possible regulatory elements.

Table B.19. SNP and SNP-set results for low-density lipoprotein (LDL) cholesterol in individuals assayed within the Framingham Heart Study. We analyze \( J = 394,174 \) SNPs and \( G = 18,364 \) SNP-sets from \( N = 6,950 \) people. Here, SNP-set annotations are based on gene boundaries defined by the NCBI's RefSeq database in the UCSC Genome Browser (195). Unannotated SNPs located within the same genomic region were labeled as being within the “intergenic region” between two genes. This file gives the posterior inclusion probabilities (PIPs) for the input and hidden layer neural network weights after fitting the BANNs model on the individual-level data. We assess significance for both SNPs and SNP-sets according to the “median probability model” threshold (6) (i.e., \( PIP \geq 0.5 \)). Page #1 provides the variant-level association mapping results with columns corresponding to: (1) chromosome; (2) SNP ID; (3) chromosomal position in base-pair (bp) coordinates; (4) SNP PIP; and (5) SuSiE PIP, which corresponds to SNP-level posterior inclusion probabilities computed by SuSiE (262). Page #2 provides the SNP-set level enrichment results with columns corresponding to: (1) chromosome; (2) SNP-set ID; (3-4) the starting and ending position of the SNP-set chromosomal boundaries; (5) SNP-set PIP; (6) RSS PIP, which corresponds to the posterior inclusion probabilities computed by RSS (307); (7) the number of SNPs that have been annotated within each SNP-set; (8) the “top” associated SNP within each SNP-set; (9) the PIP of each top SNP. Pages #3 and #4 provide similar results based on analyses where each SNP-set annotation has been augmented with a \( \pm 500 \) kilobase (kb) buffer to account for possible regulatory elements.
Table B.20. Complete summary of the results after applying BANNs, SuSiE (262), and RSS (307) to high-density lipoprotein (HDL) and low-density lipoprotein (LDL) cholesterol in both individuals assayed within the Framingham Heart Study and ten thousand randomly sampled individuals of European ancestry from the UK Biobank. The first page compares the overlap of significant SNPs and SNP-sets found by each method according to the “median probability model” threshold (6) (i.e., PIP $\geq 0.5$) in the Framingham Heart Study. The second page lists how many SNPs and SNP-sets were replicated for each method when analyzing the independent dataset from the UK Biobank. Results are based on defining gene boundaries in two ways: (a) we use the UCSC gene boundary definitions directly, and (b) we augment the gene boundaries by adding SNPs within a $\pm 500$ kilobase (kb) buffer to account for possible regulatory elements.

Table B.21. SNP and SNP-set results for high-density lipoprotein (HDL) cholesterol in ten thousand randomly sampled individuals of European ancestry from the UK Biobank. We analyze the same $J = 394,174$ SNPs and $G = 18,364$ SNP-sets used in the Framingham Heart Study analyses. Here, SNP-set annotations are based on gene boundaries defined by the NCBI’s RefSeq database in the UCSC Genome Browser (195). Unannotated SNPs located within the same genomic region were labeled as being within the “intergenic region” between two genes. This file gives the posterior inclusion probabilities (PIPs) for the input and hidden layer neural network weights after fitting the BANNs model on the individual-level data. We assess significance for both SNPs and SNP-sets according to the “median probability model” threshold (6) (i.e., PIP $\geq 0.5$). Page #1 provides the variant-level association mapping results with columns corresponding to: (1) chromosome; (2) SNP ID; (3) chromosomal position in base-pair (bp) coordinates; (4) SNP PIP; and (5) SuSiE PIP, which corresponds to SNP-level posterior inclusion probabilities computed by SuSiE (262). Page #2 provides the SNP-set level enrichment results with columns corresponding to: (1) chromosome; (2) SNP-set ID; (3-4) the starting and ending position of the SNP-set chromosomal boundaries; (5) SNP-set PIP; (6) RSS PIP, which corresponds to the posterior inclusion probabilities computed by RSS (307); (7) the number of SNPs that have been annotated within each SNP-set; (8) the “top” associated SNP within each SNP-set; (9) the PIP of each top SNP. Pages #3 and #4 provide similar results based on analyses where each SNP-set annotation has been augmented with a $\pm 500$ kilobase (kb) buffer to account for possible regulatory elements.
Table B.22. SNP and SNP-set results for low-density lipoprotein (LDL) cholesterol in ten thousand randomly sampled individuals of European ancestry from the UK Biobank. We analyze the same $J = 394,174$ SNPs and $G = 18,364$ SNP-sets used in the Framingham Heart Study analyses. Here, SNP-set annotations are based on gene boundaries defined by the NCBI’s RefSeq database in the UCSC Genome Browser (195). Unannotated SNPs located within the same genomic region were labeled as being within the “intergenic region” between two genes. This file gives the posterior inclusion probabilities (PIPs) for the input and hidden layer neural network weights after fitting the BANNs model on the individual-level data. We assess significance for both SNPs and SNP-sets according to the “median probability model” threshold (6) (i.e., $\text{PIP} \geq 0.5$). Page #1 provides the variant-level association mapping results with columns corresponding to: (1) chromosome; (2) SNP ID; (3) chromosomal position in base-pair (bp) coordinates; (4) SNP PIP; and (5) SuSiE PIP, which corresponds to SNP-level posterior inclusion probabilities computed by SuSiE (262). Page #2 provides the SNP-set level enrichment results with columns corresponding to: (1) chromosome; (2) SNP-set ID; (3-4) the starting and ending position of the SNP-set chromosomal boundaries; (5) SNP-set PIP; (6) RSS PIP, which corresponds to the posterior inclusion probabilities computed by RSS (307); (7) the number of SNPs that have been annotated within each SNP-set; (8) the “top” associated SNP within each SNP-set; (9) the PIP of each top SNP. Pages #3 and #4 provide similar results based on analyses where each SNP-set annotation has been augmented with a $\pm 500$ kilobase (kb) buffer to account for possible regulatory elements.
Appendix C

Supporting Information to

“Uncertainty Quantification in Variable Selection for Genetic Fine-Mapping using Bayesian Neural Networks’

C.1 Minor Details on the Variational Algorithm

To find the expectation of the log-likelihood during posterior inference, we use Monte Carlo samples and a local re-parameterization trick to compute gradients. More specifically, when assuming Gaussian distributions for the variational approximating families

\[ q(a) = \mathcal{N}(m, \tau^2 I) \iff a = m + \tau \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1) \] (C.1)
This technique has been shown to successfully reduce the variance of gradients (124) and stabilizes the training process. Next, we assume that the indicator variables $\gamma^{(l)}$ are sampled from a categorical distribution. We adopt a continuous relaxation technique for re-parameterizing these variables by sampling them from a Gumbel-Softmax distribution which is specified as the following (111; 162),

$$
\tilde{\gamma}_j \sim \frac{\exp(\log(\alpha_j) + \vartheta_j)/\delta}{\sum_{j'} \exp(\log(\alpha_{j'}) + \vartheta_{j'})/\delta}, \quad \vartheta_j = -\log(-\log(v_j)), \quad v_j \sim U(0, 1),
$$

(C.2)

where $\tilde{\gamma}_j$ are the approximate samples for $\gamma$, $\tau$ is a temperature parameter, and $v_j$ uniformly sampled random variable. As $\tau \to 0$, samples $\tilde{\gamma}_j$ will become closer to the desired vector where only one entry is one and the rest are zeros. In our experiments, we choose $\tau > 0.1$ for numerical stability.

The convergence of the inclusion probabilities $\alpha_j$ is also important for our model as it directly influences the performance of variable selection. Importantly, $\alpha_j$ appears very early in the computational pipeline since they are defined for the weights in first hidden layer. As a result, the gradients for $\alpha$ can be very small and hinder convergence during training. This problem is commonly known as “vanishing gradients” (98). For our work, we found that simply scaling up the learning rate when updating $\alpha_j$ works well in practice. Note that the Kullback-Leibler (KL) divergence term in the approximate likelihood can be decomposed as the following

$$
\text{KL}(q(W, \Gamma; \phi) \| p(W, \Gamma)) = \sum_{j=1}^J q(\gamma_j = 1; \phi) \text{KL}(q(a_{j\bullet} \mid \gamma_j = 1, \phi) \| p(a_{j\bullet} \mid \gamma_j = 1))
$$

$$
+ \text{KL}(q(\gamma_j = 1; \phi) \| p(\gamma_j = 1))
$$

(C.3)

where the KL divergence for the $J \times H_1$ weights $W = A \circ \Gamma$ is between two normal distributions with $A = [a_{1\bullet}, \ldots, a_{J\bullet}]$ and $a_{j\bullet}$ being an $H_1$-dimensional row-vector; while the KL divergence for the indicator variables, where $\Gamma$ is a matrix that is $H_1$ copies of the $J$-dimensional binary vector $\gamma$, is taken between two discrete multinomial distributions. Importantly, these terms have closed-form solutions with which gradients can be computed.
C.2 Pseudocode for ESNN

Algorithm 3 Training Algorithm for the ESNN Framework

1: Input genotype data $X$ and phenotypic vector $y$.
2: Choose the number of models $L$, number of maximum iterations $T$, and credible set level $\rho$.
3: Randomly initialize variational parameters $\phi_1, \ldots, \phi_L$ for the $L$ models.
4: Initialize the models $l = 1$ and iterations $t = 1$.
5: while $l \leq L$ and $t \leq T$ do
6: Fix hyper-parameters $\phi_1, \ldots, \phi_{l-1}$.
7: Sample $W^{(l)}, \Gamma^{(l)} \sim q(\phi_l)$ using re-parameterization trick.
8: Compute the approximate log-likelihood $L(\phi_1, \ldots, \phi_L; D)$.
9: Compute the gradients for only $\phi_l$ using the approximate log-likelihood.
10: Update $\phi_l$ using the gradients with optimizers.
11: Compute PIPs and credible sets for the $l$-th model.
12: if $\lambda_l > 1$ then $\triangleright$ “Purity” Check
13: $l = l + 1$
14: if $y$ is continuous then
15: $y = y - \sum_{m=1}^{l-1} f^{(m)}$ $\triangleright$ IBSS Procedure
16: end if
17: end if
18: $t = t + 1$
19: end while

Compute (marginal) posterior inclusion probabilities (PIP) for each variable. Determine credible sets $S(\alpha, \rho)$.

20: Return $\{\text{PIP}, S(\alpha, \rho)\}$. 
C.3 Supplementary Figures

Figure C.1. An example of a single-effect neural network (SNN) with only the first input variable having an effect on the outcome.
Figure C.2. Toy example demonstrating importance of accounting for nonlinearity when performing variable selection. To demonstrate that linear models lose power when non-additive variation exists, we generate two simulated datasets. (a) In the first case, we simulate $x_1 \sim N(0, 1)$ and then generate responses under $y = \cos(X_1) + e$ where $e \sim N(0, 1)$. The real data are plotted in grey points. We next run a univariate linear model (red line) and SuSiE (blue line) (263) on this dataset. Here, we perform variable selection by ranking the resulting p-values and posterior inclusion probabilities (PIP) for the respective approaches. These results show that neither method selects $x_1$ as being significant. We then run the ESNN (orange line) on these data and it successfully captures the signal. (b) In the second simulation example, we mimic a genome-wide association (GWA) study. Here, we use single nucleotide polymorphisms (SNPs) with values taking on $\{0, 1, 2\}$ based on copies of a reference allele where 0 and 2 represent “homozygotes” and 1 represents “heterozygotes”. We then simulate the phenotype $y$ by assuming the heterozygote has a significant effect. Similar relationships have been shown in the literature (241; 211). Similar to the first simulation case, linear methods fail to capture the causal effect while the ESNN is robust to the non-additive architecture. These two toy examples illustrate the importance of accounting for nonlinearity in variable selection methods.
Figure C.3. An example of absolute correlation matrix for genotype data.
Figure C.4. Assessments of posterior inclusion probability (PIP) calibration for ESNN and SuSiE. This experiment follows largely from previous work (263). Here, SNPs are grouped into bins according to their reported PIPs (using 20 equally spaced bins, from 0 to 1). The plots show the average PIP for each bin against the proportion of causal SNPs or SNP-sets in that bin. A well calibrated method should produce points near the x-axis = y-axis line (i.e., the diagonal red lines). Gray error bars show ±2 standard errors. Panel (a, b) shows the comparison of ESNN and panels (c, d) shows the comparison of SuSiE for continuous and binary traits, respectively.
Figure C.5. Precision and recall curves for simulation studies of different scenarios. Results are based on 200 data replicates.
Figure C.6. Posterior inclusion probabilities (PIP) of ESNN and SuSiE for the heterogeneous stock of mice dataset from the Wellcome Trust Centre for Human Genetics (250). (a) Highlighted region for low-density lipoprotein (LDL). Significant SNPs found only by ESNN (included in the credible sets), only by SuSiE, and by both methods are color coded in red, black, and blue, respectively. (b) Highlighted region for high-density lipoprotein (HDL).
Bibliography


[43] Susanne M Clee, Hanfang Zhang, Nagat Bissada, Li Miao, Ewa Ehrenborg, Pascale Benlian, Garry X Shen, Aubie Angel, Renee C LeBoeuf, and Michael R Hayden. Relationship between


[104] Farhad Hormozdiari, Martijn van de Bunt, Ayellet V. Segrè, Xiao Li, Jong Wha J. Joo, Michael Bilow, Jae Hoon Sul, Sriram Sankararaman, Bogdan Pasaniuc, and Eleazar Eskin.


[107] Li Hui, Mei Han, Xiang Dong Du, Bao Hua Zhang, Shu Chang He, Tian Nan Shao, and Guang Zhong Yin. Serum apob levels in depressive patients: associated with cognitive deficits. Scientific Rep, 7(1):39992, 2017. doi: 10.1038/srep39992. URL https://doi.org/10.1038/srep39992.


[125] Derek Klarin, Scott M. Damrauer, Kelly Cho, Yan V. Sun, Tanya M. Teslovich, Jacqueline Honerlaw, David R. Gagnon, Scott L. DuVall, Jin Li, Gina M. Peloso, Mark Chaffin, Aeron M. Small, Jie Huang, Hua Tang, Julie A. Lynch, Yuk-Lam Ho, Dajiang J. Liu, Connor A. Emdin, Alexander H. Li, Jennifer E. Huffman, Jennifer S. Lee, Pradeep Natarajan,


[129] Maxim V Kuleshov, Matthew R Jones, Andrew D Rouillard, Nicolas F Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L Jenkins, Kathleen M Jagodnik, Alexander


population-specific haplotype structure and linking correlated alleles of possible functional


[163] Julian B Maller, Gilean McVean, Jake Byrnes, Damjan Vukcevic, Kimmo Palin, Zhan Su,
Joanna MM Howson, Adam Auton, Simon Myers, Andrew Morris, et al. Bayesian refinement

[164] Elettra Mancuso, Gaia Chiara Mannino, Anastasia Fuoco, Antonio Leo, Rita Citraro, Car-
olina Averta, Rosangela Spiga, Emilio Russo, Giovambattista De Sarro, Francesco Andreozzi,
et al. Hdl (high-density lipoprotein) and apoa-1 (apolipoprotein a-1) potentially modulate
pancreatic α-cell glucagon secretion. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 40


[166] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J
Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, Judy H Cho,
Alan E Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N Rotimi,
Montgomery Slatkin, David Valle, Alice S Whitlemore, Michael Boehnke, Andrew G Clark,
Evan E Eichler, Greg Gibson, Jonathan L Haines, Trudy F C Mackay, Steven A McCarroll,
and Peter M Visscher. Finding the missing heritability of complex diseases. *Nature*, 461
pubmed/19812666.

[167] Eirini Marouli, Mariaelisa Graff, Carolina Medina-Gomez, Ken Sin Lo, Andrew R Wood,
Troels R Kjaer, Rebecca S Fine, Yingchang Lu, Claudia Schurmann, Heather M Highland,
Sina Rueger, Gudmar Thorleifsson, Anne E Justice, David Lamparter, Kathleen E Stirrups,
Valerie Turcot, Kristin L Young, Thomas W Winkler, Tonu Esko, Tugce Karaderi, Adam E


URL https://doi.org/10.1038/nature01266.


[199] Wen Qin, Rong Zhang, Cheng Hu, Cong-rong Wang, Jing-yi Lu, Wei-hui Yu, Yu-qian Bao, Kun-san Xiang, and Wei-ping Jia. A variation in nos1ap gene is associated with repaglinide


[212] Richa Saxena, Benjamin F Voight, Valeriya Lyssenko, Noël P Burtt, Paul I W de Bakker, Hong Chen, Jeffrey J Roix, Sekar Kathiresan, Joel N Hirschhorn, Mark J Daly, Thomas E Hughes, Leif Groop, David Altshuler, Peter Almgren, Jose C Florez, Joanne Meyer, Kristin Ardlie, Kristina Bengtsson Boström, Bo Isomaa, Guillaume Lettre, Ulf Lindblad, Helen N Lyon, Olle Melander, Christopher Newton-Cheh, Peter Nilsson, Marju Orho-Melander, Lennart Råstam, Elizabeth K Speliotes, Marja-Riitta Taskinen, Tiinamaija Tuomi, Candace Guiducci, Anna Berglund, Joyce Carlson, Lauren Gianniny, Rachel Hackett, Liselotte Hall, Johan Holmkvist, Esa Laurila, Marketa Sjögren, Maria Sterner, Aarti Surti, Margareta Svensson, Malin Svensson, Ryan Tewhey, Brendan Blumenstiel, Melissa Parkin, Matthew Defelice, Rachel Barry, Wendy Brodeur, Jody Camarata, Nancy Chia, Mary Fava, John Gibbons, Bob Handsaker, Claire Healy, Kieu Nguyen, Casey Gates, Carrie Sougnez, Diane Gage, Marcia Nizzari,


[219] Dmitry Shungin, Thomas W Winkler, Damien C Croteau-Chonka, Teresa Ferreira, Adam E Locke, Reedik Mägi, Rona J Strawbridge, Tune H Pers, Krista Fischer, Anne E Justice,


