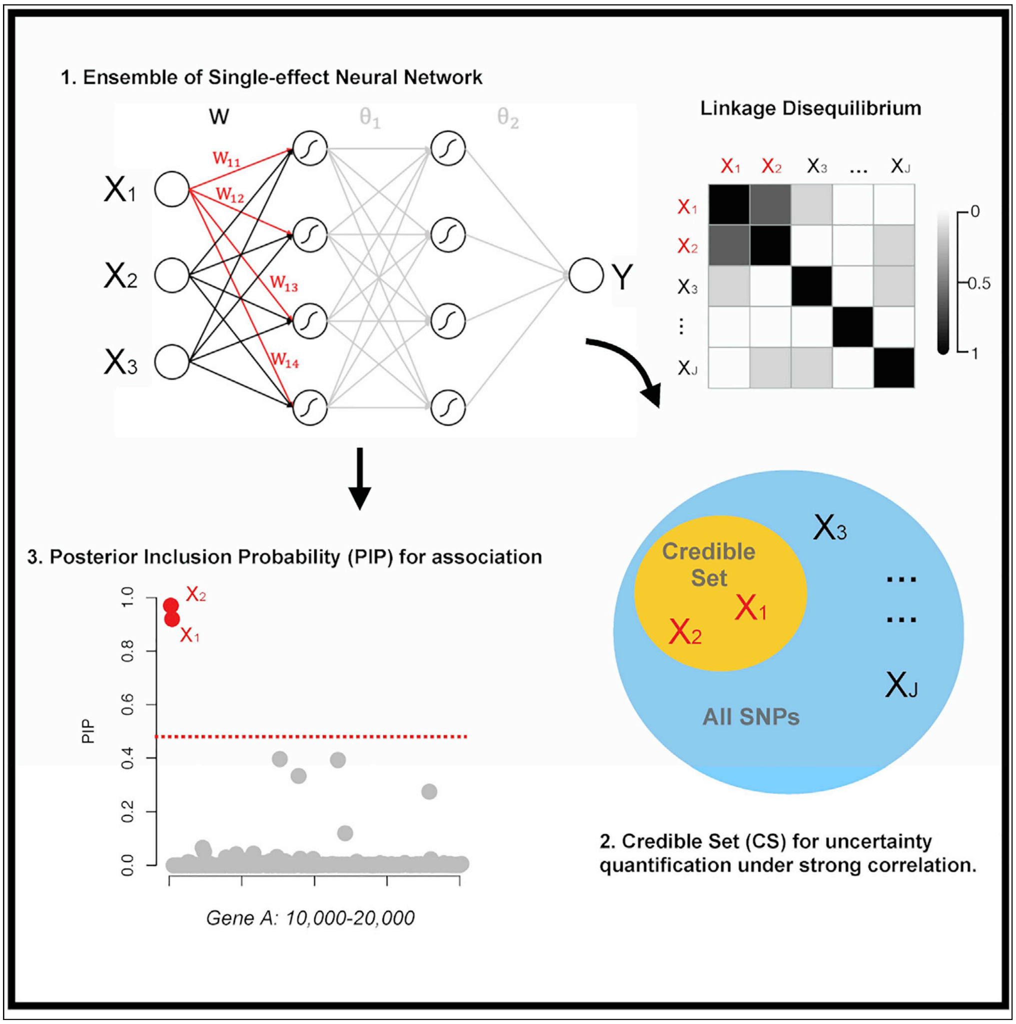
iScience

CellPress
OPEN ACCESS

Article

Uncertainty quantification in variable selection for genetic fine-mapping using bayesian neural networks



Wei Cheng, Sohini Ramachandran, Lorin Crawford

weicheng1@brown.edu (W.C.)
lcrawford@microsoft.com (L.C.)

Highlights

Nonlinear genetic fine-mapping using an ensemble of single-effect neural networks

Posterior quantification of uncertainty for variable selection via credible sets

Improved coverage of credible sets and power for variable selection in simulations

Data analyses reveal variants that nonlinearly contribute to phenotypic variation

Cheng et al., iScience *25*, 104553
July 15, 2022
https://doi.org/10.1016/j.isci.2022.104553

iScience

CellPress
OPEN ACCESS

Article

Uncertainty quantification in variable selection for genetic fine-mapping using bayesian neural networks

Wei Cheng,[1,2,3,*] Sohini Ramachandran,[1,2,3] and Lorin Crawford[3,4,5,6,*]

SUMMARY

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.

INTRODUCTION

Variable selection is a fundamental problem in high-dimensional statistical learning that arises in a wide range of application domains (George and McCulloch, 1993; Fan and Lv, 2010; Carbonetto and Stephens, 2012; Yamada et al., 2020). 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 (Manolio, 2010; Maller et al., 2012). This can further help with downstream clinical applications such as drug development.

Although many methods for variable selection have been developed in the literature (George and McCulloch, 1993; Fan and Lv, 2010; Carbonetto and Stephens, 2012; Yamada et al., 2020; Zou and Hastie, 2005; Tibshirani, 1996), some significant challenges still remain. One important challenge is assessing the uncertainty in which variables should be selected when they are highly correlated (Wang et al., 2020; Carbonetto and Stephens, 2012). 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 the interest and disregard the other (Wang et al., 2020). 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. (2020) 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 is relevant, but we are unsure as to which one." SuSiE uses an iterative Bayesian stepwise selection (IBSS) procedure where it 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 nonlinear 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 (Minamikawa et al., 2017; Crawford et al., 2017; Ramstein et al.,

[1]Department of Computer Science, Brown University, Providence, RI, USA

[2]Department of Ecology and Evolutionary Biology, Brown University, Providence, RI, USA

[3]Center for Computational Molecular Biology, Brown University, Providence, RI, USA

[4]Department of Biostatistics, Brown University, Providence, RI, USA

[5]Microsoft Research New England, Cambridge, MA, USA

[6]Lead contact

*Correspondence: weicheng1@brown.edu (W.C.), lcrawford@microsoft.com (L.C.)

https://doi.org/10.1016/j.isci.2022.104553

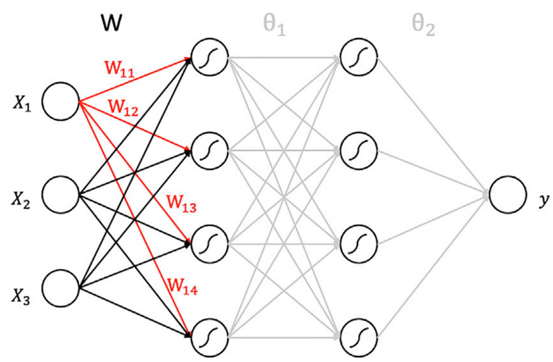


Figure 1. An example of a single-effect neural network (SNN) with only the first input variable having an effect on the outcome

2020; Li et al., 2013). 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 (Figure 1). We demonstrate our approach in a simulation study and on two real GWA datasets.

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 GWA studies. We also apply our method to real-world GWA datasets from the Wellcome Trust Case Control Consortium (WTCCC) and the Wellcome Trust Centre for Human Genetics.

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 $N = 5{,}000$ randomly sampled individuals of self-identified European ancestry in the UK Biobank (Bycroft et al., 2018). After quality control (Demetci et al., 2021), this dataset had 36,518 SNPs (see STAR Methods). To simulate fine-mapping applications, we used the NCBI's Reference Sequence (RefSeq) database in the UCSC Genome Browser (Pruitt et al., 2005) 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 kb window of the boundary to also include regulatory elements (see STAR Methods).

In this study, each gene is considered to be its own dataset with its own complex correlation structure (see Figure S1) 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

$$\mathbf{y} = \sum_{j \in \mathcal{C}} \mathbf{x}_j \beta_j 1(\mathbf{x}_j = 0 \text{ or } 2) + \sum_{j \in \mathcal{C}} \mathbf{x}_j \omega_j 1(\mathbf{x}_j = 1) + \mathbf{e}, \quad \mathbf{e} \sim \mathcal{N}\left(0, \sigma_y^2 \mathbf{I}\right) \qquad \text{(Equation 1)}$$

where $\mathcal{C}$ represents the set of causal SNPs, and $1(\bullet)$ is an indicator function. Here, β and ω are different effect sizes for heterozygotes and homozygotes, respectively. Both variables are randomly sampled from a standard normal distribution and rescaled according to their frequencies. The error term $\mathbf{e}$ 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 other fine-mapping approaches: SuSiE (Wang et al., 2020), DAP-G (Wen et al., 2016), CAVIAR (Hormozdiari et al., 2014), and FINEMAP (Benner et al., 2016). We run all competing methods under their default parameter settings. We set $L = 10$ for both SuSiE and our approach. For ESNN, we used a simple sparse architecture with 5 hidden neurons and tanh activation functions (see STAR Methods). Here,

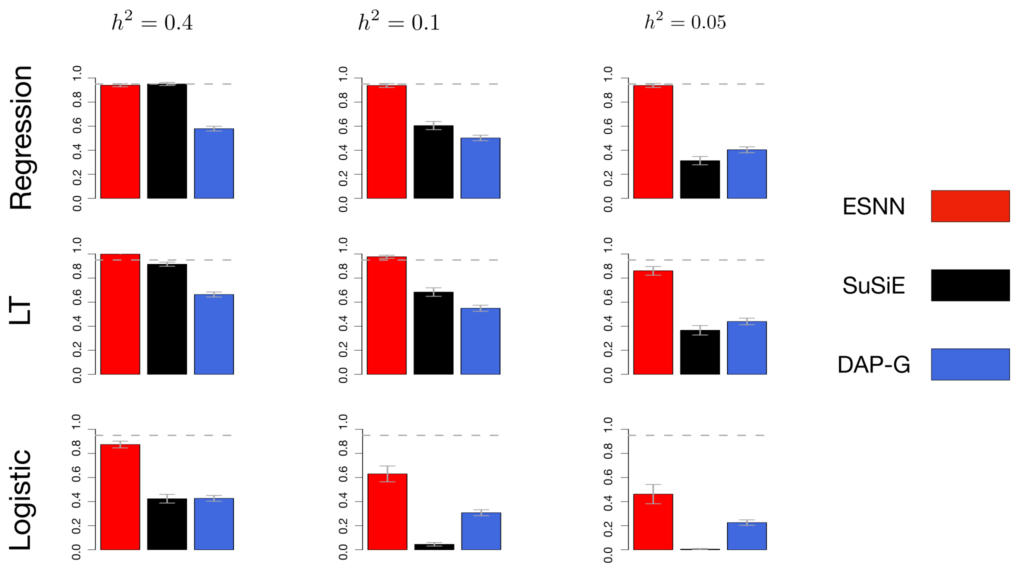


Figure 2. Comparisons of coverage for ESNN, SuSiE, and DAP-G in simulation studies under different levels of heritability

Results are based on 200 data replicates with standard errors represented by the grey bars.

we set the maximum number of epochs to be 30; the hyper-parameter $\boldsymbol{\pi}$ for the indicator variable $\boldsymbol{\gamma}$ is chosen from a uniform distribution; we fixed $\sigma_1^2 = 1$ for all L models, and during training, we take 100 Monte Carlo samples to evaluate the log likelihood (see STAR Methods). 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 performance, we consider three different metrics. The first two metrics focus on evaluating the credible sets. To our knowledge, since only SuSiE and DAP-G generate credible sets, we only compare ESNN with these two methods for these metrics (DAP-G produces "signal clusters," which follows a definition similar to credible sets in Definition 1; see Wang et al. (2020) for relevant discussion on this distinction). We begin by assessing the probability that each method creates a credible set containing at least 1 effect SNP (first row Figure 2). 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 and DAP-G drops, while the coverage of ESNN remains the same. The second metric we check is the average number of effect variables included in all credible sets (Figure S3). In practice, each method can report multiple credible sets. Therefore, this metric essentially helps evaluate the total number of effect variables discovered by each method. Overall, DAP-G and ESNN consistently outperform SuSiE, with DAP-G having the advantage. However, because the coverage of DAP-G is poor (Figure 2), this result effectively means that DAP-G generates a large number of credible sets with false-positive signals. For the final metric, we assess the ability of ESNN and the competing approaches to accurately prioritize causal variants according to the posterior inclusion probabilities (PIPs) that each method provides (see STAR Methods). Here, we use receiver operating characteristic (ROC) and precision-recall curves to compare their ability to rank true positives over false positives (Figures 3 and S4). As h^2 decreases, accuracy of the PIPs for all methods decreases, while our method is relatively better powered for all scenarios. Since SuSiE is the most comparable method to the ESNN model, we also highlight the scenarios (denoted by an asterisk) where the distribution of the area under the curve for our method is significantly larger than that for SuSiE (satisfying $P < 0.05$). Importantly, the PIPs from ESNN and SuSiE are calibrated similarly (Figure S2).

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

(Lee et al., 2011; Golan et al., 2014; Falconer, 1965). In the former, we simply use the genotypes from chromosome 1 of the $N = 5{,}000$ randomly sampled individuals from the UK Biobank to assume that

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

where, in addition to the previous notation, the binary traits follow a Bernoulli distribution with probability $\mathbf{p}$. 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 a 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 1 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 Equation (1). Then we consider a prevalence $k \in \{50\%, 10\%, 1\%\}$ and define case-control labels for each of the million individuals. Finally, we subsample 2,500 cases and 2,500 controls for the analysis.

Once again, we compare ESNN to SuSiE (Wang et al., 2020), DAP-G (Wen et al., 2016), CAVIAR (Hormozdiari et al., 2014), and FINEMAP (Benner et al., 2016) using coverage (Figure 2), the number of effect variables included in all credible sets per dataset (Figure S3), ROC curves (Figure 3), and precision-recall curves (Figure S4). 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 a Markov Chain Monte Carlo to estimate continuous liability scores as phenotypes (Felsenstein, 2005; Falconer, 1965; Curnow and Smith, 1975). Here, we use all the same parameter settings as in the regression simulation study, except that we set the learning rate for ESNN to be 0.01. Overall, performances follow a similar trend to the regression simulations such that ESNN consistently outperforms other methods. When disease prevalence is very low (e.g., $k = 1\%$), cases are assumed to come from the "tail" of the distribution. In this scenario, statistical models are generally better powered (Weissbrod et al., 2015). As the prevalence k becomes greater, such that the LT 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.

Fine-mapping in heterogeneous stock of mice

We applied ESNN and SuSiE to two continuous traits: high-density lipoprotein (HDL) and low-density lipoprotein (LDL) in a heterogeneous stock of mice dataset from the Wellcome Trust Centre for Human Genetics (Valdar et al., 2006). This dataset contains $J = 10{,}346$ SNPs with $N = 1594$ samples for HDL and $N = 1637$ samples for LDL (see STAR Methods). 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 can potentially be due to the criterion that we only include an SNN model into the ensemble if it increases the likelihood. This criterion demonstrated to ensure that a credible set generated by ESNN would have high coverage in simulations (Figure 2). 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 Figure S5. One SNP found by both methods, rs3090325 in LDL (Figure S5A), can be mapped to the *Smarca2* gene, which has been found to be associated with cholesterol regulation (Meaney, 2014). In HDL (Figure S5B), SNP *gnf04.147.942* can be mapped to the *Panc1* gene, which regulates pancreatic activity and has been shown to be linked with HDL (Mancuso et al., 2020). 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 (Yoo and Desiderio, 2003; Lee et al., 2006).

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 WTCCC 1 study (Wellcome Trust Case Control Consortium, 2007). This dataset has $N = 1963$ cases and $N = 2938$ controls for T1D, $N = 1924$ cases and $N = 2938$ controls for T2D, along with $J = 458{,}868$

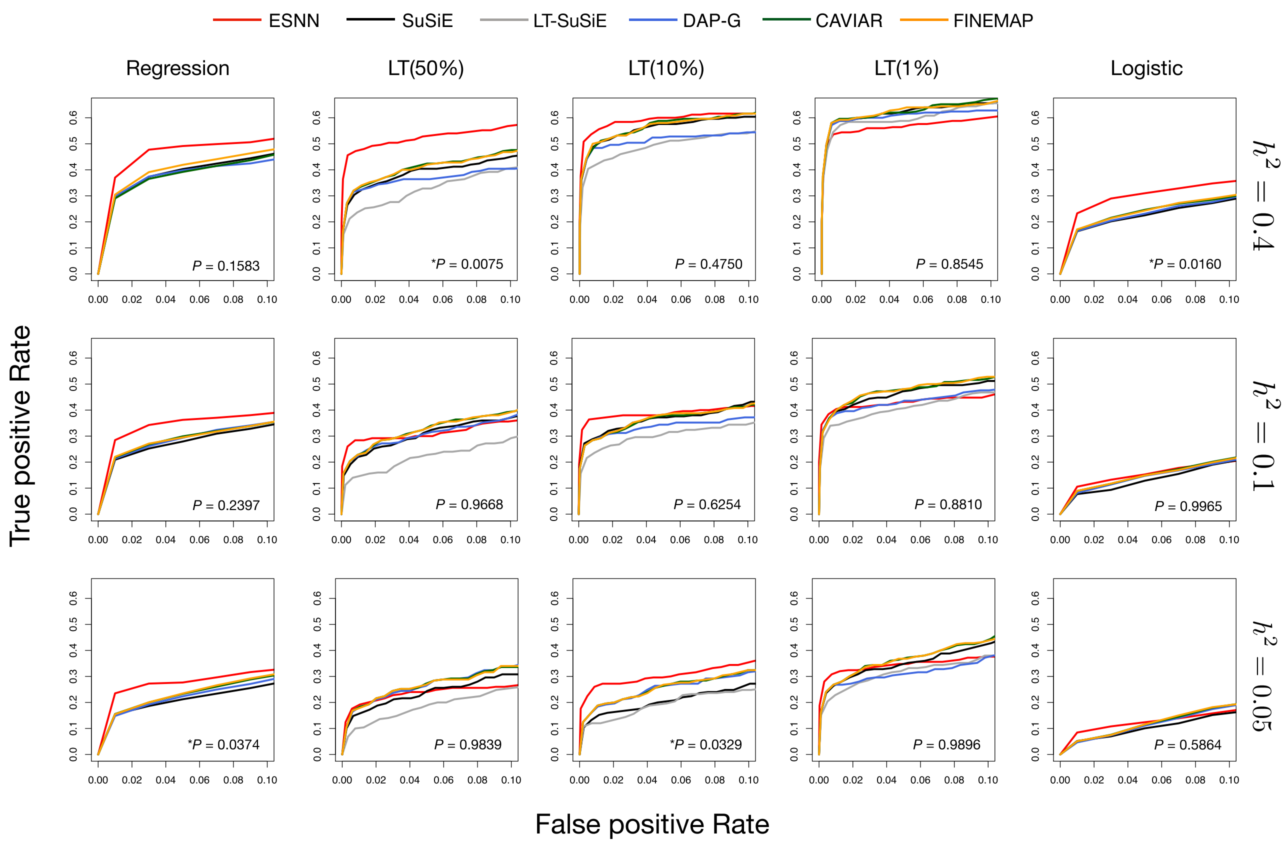


Figure 3. Receiver operating characteristic (ROC) curves for simulation studies of different scenarios

Listed in each panel are p values indicating the level of significant difference between results for ESNN and SuSiE according to their respective areas under the curve (AUCs) across the simulations. Asterisks (*) denote scenarios where ESNN is significantly better powered than SuSiE (i.e., satisfying $P < 0.05$). Results are based on 200 data replicates.

genotyped SNPs for each individual (see STAR Methods). Similarly, we run ESNN and SuSiE with a window size of 500 SNPs and use the same model settings as in the binary simulations.

ESNN identifies 32 and 19 credible sets for T1D and T2D, respectively, whereas SuSiE finds 67 and 30 sets for each trait, respectively. 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 Figure 4 where we plot the PIPs of SNPs computed by ESNN and SuSiE. In panel (a), we show a window near the human leukocyte antigen (HLA) region on chromosome 6, which has been well studied in the literature and found to be associated with T1D (Hu et al., 2015; Nejentsev et al., 2007; Erlich et al., 2008; Noble and Valdes, 2011). 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 (Noble and Valdes, 2011). 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 (Hu et al., 2010; Chu et al., 2010; Qin et al., 2010). 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 (Hu et al., 2010). Similar to SuSiE, these conclusions were previously made using linear models. We hypothesize that this region may contribute to T2D nonlinearly, and thus, the traditional hypothesis-testing methods would have missed this signal.

DISCUSSION

In this paper, we present the ESNN which generalizes the sum of single-effects regression framework by accounting for nonlinear genetic architecture and extending to non-continuous phenotypes. The ESNN approach

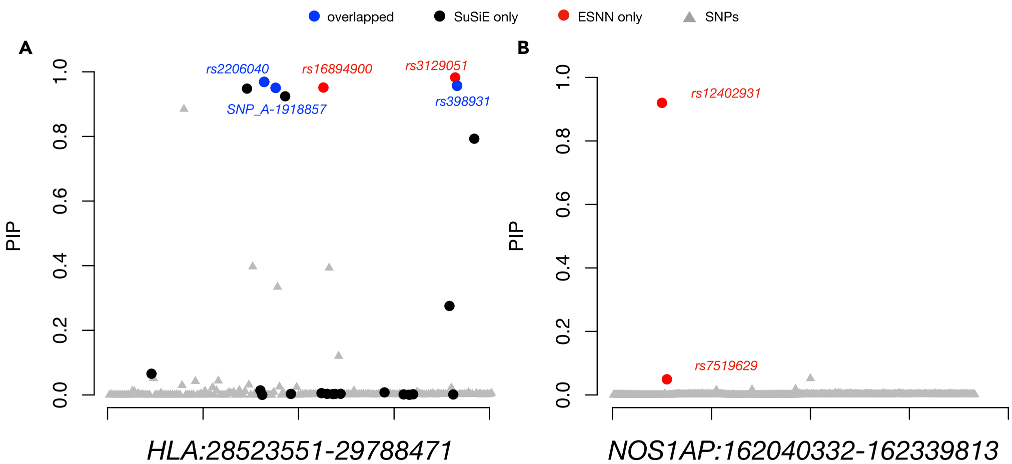


Figure 4. Posterior inclusion probabilities (PIP) of ESNN and SuSiE in the WTCCC analysis

(A) Highlighted region for type 1 diabetes (T1D). 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 type 2 diabetes (T2D).

provides PIPs and credible sets that can guide variable selection (see STAR Methods). 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 (see STAR Methods). 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.

LIMITATIONS OF THE STUDY

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.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - The sum of single-effects regression model
  - The ensemble of single-effect neural networks
  - Posterior inference via variational bayes
  - Details of the variational algorithm
  - Iterative bayesian stepwise selection
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.104553.

ACKNOWLEDGMENTS

S.R. is supported by US National Institutes of Health (NIH) grant R01 GM118652, NIH grant R35 GM139628, and National Science Foundation (NSF) CAREER award DBI1452622. L.C. is supported by a David & Lucile Packard Fellowship for Science and Engineering. Data from the UK Biobank Resource were made available under application numbers 22419 (S.R.) and 14649 (L.C.). This study also makes use of data generated by the Wellcome Trust Case Control Consortium (WTCCC). Funding for the WTCCC project was provided by the Wellcome Trust under award 076113, 085475, and 090355. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of any of the funders.

AUTHOR CONTRIBUTIONS

W.C. and L.C. conceived the methods. W.C. developed the software and carried out all analyses. W.C., S.R., and L.C. wrote and reviewed the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 23, 2022
Revised: May 9, 2022
Accepted: June 1, 2022
Published: July 15, 2022

REFERENCES

Barron, A.R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. IEEE Trans. Inf. Theor. *39*, 930–945. https://doi.org/10.1109/18.256500.

Benner, C., Spencer, C.C., Havulinna, A.S., Salomaa, V., Ripatti, S., and Pirinen, M. (2016). Finemap: efficient variable selection using summary data from genome-wide association studies. Bioinformatics *32*, 1493–1501. https://doi.org/10.1093/bioinformatics/btw018.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209. https://doi.org/10.1038/s41586-018-0579-z.

Carbonetto, P., and Stephens, M. (2012). Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. Bayesian Anal. *7*, 73–108. https://doi.org/10.1214/12-ba703.

Chen, Y., Gao, Q., Liang, F., and Wang, X. (2021). Nonlinear variable selection via deep neural networks. J. Comput. Graph. Stat. *30*, 484–492. https://doi.org/10.1080/10618600.2020.1814305.

Chu, A.Y., Coresh, J., Arking, D.E., Pankow, J.S., Tomaselli, G.F., Chakravarti, A., Post, W., Spooner, P.H., Spooner, P., Boerwinkle, E., Kao, W.H.L., and Kao, W. (2010). Nos1ap variant associated with incidence of type 2 diabetes in calcium channel blocker users in the atherosclerosis risk in communities (aric) study. Diabetologia *53*, 510–516. https://doi.org/10.1007/s00125-009-1608-0.

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

Curnow, R.N., and Smith, C. (1975). Multifactorial models for familial diseases in man. J. R. Stat. Soc. *138*, 131. https://doi.org/10.2307/2984646.

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

Erlich, H., Valdes, A.M., Noble, J., Carlson, J.A., Varney, M., Concannon, P., Mychaleckyj, J.C., Todd, J.A., Bonella, P., Fear, A.L., et al. (2008). Hla dr-dq haplotypes and genotypes and type 1 diabetes risk: analysis of the type 1 diabetes genetics consortium families. Diabetes *57*, 1084–1092. https://doi.org/10.2337/db07-1331.

Falconer, D.S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. Ann. Hum. Genet. *29*, 51–76. https://doi.org/10.1111/j.1469-1809.1965.tb00500.x.

Fan, J., and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. Stat. Sin. *20*, 101–148.

Felsenstein, J. (2005). Using the quantitative genetic threshold model for inferences between and within species. Phil. Trans. Biol. Sci. *360*, 1427–1434. https://doi.org/10.1098/rstb.2005.1669.

George, E.I., and McCulloch, R.E. (1993). Variable selection via Gibbs sampling. J. Am. Stat. Assoc. *88*, 881–889. https://doi.org/10.1080/01621459.1993.10476353.

Ghosh, S., Yao, J., and Doshi-Velez, F. (2019). Model selection in bayesian neural networks via horseshoe priors. J. Mach. Learn. Res. *20*, 1–46.

Golan, D., Lander, E.S., and Rosset, S. (2014). Measuring missing heritability: inferring the contribution of common variants. Proc. Natl. Acad. Sci. USA *111*, E5272–E5281. https://doi.org/10.1073/pnas.1419064111.

Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. *06*, 107–116. https://doi.org/10.1142/s0218488598000094.

Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B., and Eskin, E. (2014). Identifying causal variants at loci with multiple signals of association. Genetics *198*, 497–508. https://doi.org/10.1534/genetics.114.167908.

Hu, C., Wang, C., Zhang, R., Ng, M.C., Bao, Y., So, W., So, W.Y., Ma, R., Ma, R.C., Ma, X., et al. (2010). Association of genetic variants of nos1ap with type 2 diabetes in a Chinese population. Diabetologia *53*, 290–298. https://doi.org/10.1007/s00125-009-1594-2.

Hu, X., Deutsch, A.J., Lenz, T.L., Onengut-Gumuscu, S., Han, B., Chen, W.-M., Howson, J.M.M., Todd, J.A., de Bakker, P.I.W., Rich, S.S., and Raychaudhuri, S. (2015). Additive and interaction effects at three amino acid positions in hla-dq and hla-dr molecules drive type 1 diabetes risk. Nat. Genet. *47*, 898–905. https://doi.org/10.1038/ng.3353.

Jang, E., Gu, S., and Poole, B. (2017). Categorical reparametrization with gumbel-softmax. In Proceedings International Conference on Learning Representations (ICLR).

Kingma, D.P., and Welling, M. (2014). Auto-encoding variational bayes. In 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, Y. Bengio and Y. LeCun, eds..

Lee, K.-Y., Kim, S.-J., Cha, Y.-S., So, J.-R., Park, J.-S., Kang, K.-S., and Chon, T.-W.

(2006). Effect of exercise on hepatic gene expression in an obese mouse model using cdna microarrays. Obesity *14*, 1294–1302. https://doi.org/10.1038/oby.2006.147.

Lee, S.H., Wray, N.R., Goddard, M.E., and Visscher, P.M. (2011). Estimating missing heritability for disease from genome-wide association studies. Am. J. Hum. Genet. *88*, 294–305. https://doi.org/10.1016/j.ajhg.2011.02.002.

Leshno, M., Lin, V.Y., Pinkus, A., and Schocken, S. (1993). Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. Neural Netw. *6*, 861–867. https://doi.org/10.1016/s0893-6080(05)80131-5.

Li, J., Glessner, J.T., Zhang, H., Hou, C., Wei, Z., Bradfield, J.P., Mentch, F.D., Guo, Y., Kim, C., Xia, Q., et al. (2013). Gwas of blood cell traits identifies novel associated loci and epistatic interactions in caucasian and african-american children. Hum. Mol. Genet. *22*, 1457–1464. https://doi.org/10.1093/hmg/dds534.

Maddison, C.J., Mnih, A., and Teh, Y.W. (2017). The concrete distribution: a continuous relaxation of discrete random variables. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings.

Maller, J.B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J.M.M., Auton, A., Myers, S., Morris, A., et al. (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. Nat. Genet. *44*, 1294–1301. https://doi.org/10.1038/ng.2435.

Mancuso, E., Mannino, G.C., Fuoco, A., Leo, A., Citraro, R., Averta, C., Spiga, R., Russo, E., De Sarro, G., Andreozzi, F., and Sesti, G. (2020). Hdl (high-density lipoprotein) and apoa-1 (apolipoprotein a-1) potentially modulate pancreatic α-cell glucagon secretion. Arterioscler. Thromb. Vasc. Biol. *40*, 2941–2952. https://doi.org/10.1161/atvbaha.120.314640.

Manolio, T.A. (2010). Genomewide association studies and assessment of the risk of disease. N. Engl. J. Med. *363*, 166–176. https://doi.org/10.1056/nejmra0905980.

Meaney, S. (2014). Epigenetic regulation of cholesterol homeostasis. Front. Genet. *5*, 311. https://doi.org/10.3389/fgene.2014.00311.

Minamikawa, M.F., Nonaka, K., Kaminuma, E., Kajiya-Kanegae, H., Onogi, A., Goto, S., Yoshioka, T., Imai, A., Hamada, H., Hayashi, T., et al. (2017). Genome-wide association study and genomic prediction in citrus: potential of genomics-assisted breeding for fruit quality traits. Sci. Rep. *7*, 4721. https://doi.org/10.1038/s41598-017-05100-x.

Nejentsev, S., Howson, J.M.M., Walker, N.M., Szeszko, J., Field, S.F., Stevens, H.E., Reynolds, P., Hardy, M., King, E., Masters, J., et al. (2007). Localization of type 1 diabetes susceptibility to the mhc class i genes hla-b and hla-a. Nature *450*, 887–892. https://doi.org/10.1038/nature06406.

Noble, J.A., and Valdes, A.M. (2011). Genetics of the hla region in the prediction of type 1 diabetes. Curr. Diabetes Rep. *11*, 533–542. https://doi.org/10.1007/s11892-011-0223-x.

Pickrell, J.K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am. J. Hum. Genet. *94*, 559–573. https://doi.org/10.1016/j.ajhg.2014.03.004.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2004). Ncbi reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. *33*, D501–D504. https://doi.org/10.1093/nar/gki025.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575. https://doi.org/10.1086/519795.

Qin, W., Zhang, R., Hu, C., Wang, C.-r., Lu, J.-y., Yu, W.-h., Bao, Y.-q., Xiang, K.-s., and Jia, W.-p. (2010). A variation in nos1ap gene is associated with repaglinide efficacy on insulin resistance in type 2 diabetes of Chinese. Acta. Pharmacol. Sin. *31*, 450–454. https://doi.org/10.1038/aps.2010.25.

Ramstein, G.P., Larsson, S.J., Cook, J.P., Edwards, J.W., Ersoz, E.S., Flint-Garcia, S., Gardner, C.A., Holland, J.B., Lorenz, A.J., McMullen, M.D., et al. (2020). Dominance effects and functional enrichments improve prediction of agronomic traits in hybrid maize. Genetics *215*, 215–230. https://doi.org/10.1534/genetics.120.303025.

Servin, B., and Stephens, M. (2005). Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet. *3*, e114. https://doi.org/10.1371/journal.pgen.0030114.eor.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc. B *58*, 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Valdar, W., Solberg, L.C., Gauguier, D., Burnett, S., Klenerman, P., Cookson, W.O., Taylor, M.S., Rawlins, J.N.P., Mott, R., and Flint, J. (2006). Genome-wide genetic association of complex traits in heterogeneous stock mice. Nat. Genet. *38*, 879–887. https://doi.org/10.1038/ng1840.

Wang, G., Sarkar, A., Carbonetto, P., and Stephens, M. (2020). A simple new approach to variable selection in regression, with application to genetic fine mapping. J. Roy. Stat. Soc. B *82*, 1273–1300. https://doi.org/10.1111/rssb.12388.

Weissbrod, O., Lippert, C., Geiger, D., and Heckerman, D. (2015). Accurate liability estimation improves power in ascertained case-control studies. Nat. Methods *12*, 332–334. https://doi.org/10.1038/nmeth.3285.

Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14, 000 cases of seven common diseases and 3, 000 shared controls. Nature *447*, 661–678. https://doi.org/10.1038/nature05911.

Wen, X., Lee, Y., Luca, F., and Pique-Regi, R. (2016). Efficient integrative multi-snp association analysis via deterministic approximation of posteriors. Am. J. Hum. Genet. *98*, 1114–1129. https://doi.org/10.1016/j.ajhg.2016.03.029.

Yamada, Y., Lindenbaum, O., Negahban, S., and Kluger, Y. (2020). Feature selection using stochastic gates. In International Conference on Machine Learning (PMLR), pp. 10648–10659.

Yoo, J.-Y., and Desiderio, S. (2003). Innate and acquired immunity intersect in a global view of the acute-phase response. Proc. Natl. Acad. Sci. USA *100*, 1157–1162. https://doi.org/10.1073/pnas.0336385100.

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

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. J. Roy. Stat. Soc. B *67*, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Mouse Resources	Valdar et al. (2006)	http://mtweb.cs.ucl.ac.uk/mus/www/mouse/index.shtml
WTCCC 1 Study	Wellcome Trust Case Control Consortium (2007)	www.wtccc.org.uk
UK Biobank	Bycroft et al. (2018)	https://www.ukbiobank.ac.uk
Software and algorithms		
ESNN	This Study	https://github.com/ramachandran-lab/ESNN
SuSiE	Wang et al. (2020)	https://github.com/stephenslab/susieR
DAP-G	Wen et al. (2016)	https://github.com/xqwen/dap
CAVIAR	Hormozdiari et al. (2014)	http://genetics.cs.ucla.edu/caviar
FINEMAP	Benner et al. (2016)	http://www.christianbenner.com

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Lorin Crawford (lcrawford@microsoft.com).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyzes existing, publicly available data. These accession URLs for the datasets are listed in the Key resources table.
- 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.
- Any additional information required to reanalyze the data reported in this paper is available from the Lead contact upon request

METHOD DETAILS

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 (Servin and Stephens, 2007; Pickrell, 2014) assumes that exactly one of J input variables has a non-zero coefficient. More specifically,

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\mathbf{b}+\mathbf{e}, \quad \mathbf{e}\sim\mathcal{N}\left(0,\sigma_y^2\mathbf{I}\right) \\ \mathbf{b} &= b\boldsymbol{\gamma}, b\sim\mathcal{N}\left(0,\sigma_1^2\right), \quad \boldsymbol{\gamma}\sim\mathrm{Mult}(1,\boldsymbol{\pi}) \end{aligned} \quad \text{(Equation 3)}$$

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; $\boldsymbol{\gamma}$ is a binary indicator that determines which regression coefficient is to be non-zero; and $\mathrm{Mult}(m,\boldsymbol{\pi})$ denotes the multinomial distribution with m samples drawn with class probability distribution $\boldsymbol{\pi}$. For simplicity, we will consider a uniform prior such that $\boldsymbol{\pi} = (1/J,\ldots,1/J)$. 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 $\mathrm{PIP}_j\equiv\Pr[\widehat{b}_j\neq 0 \mid \mathbf{y},\mathbf{X}]$.

In the context of statistical genetics, the original SER model only assumes one causal SNP. However, we know that in many real-world applications, it is desired to have a method that flexibly allows for many variants to have an effect on trait architecture (Carbonetto and Stephens, 2012; Demetci et al., 2021). The SuSiE framework is based on an extension of summing over L-multiple SER models (Wang et al., 2020). Here, the main idea is to construct an overall effect vector $\mathbf{b}$ from multiple single-effect coefficients $\mathbf{b}_1, \ldots, \mathbf{b}_L$ via the following

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}, \quad \mathbf{e} \sim \mathcal{N}\left(0, \sigma_y^2 \mathbf{I}\right),$$
$$\mathbf{b} = \sum_{l=1}^{L} \mathbf{b}^{(l)}, \quad \mathbf{b}^{(l)} = b^{(l)} \boldsymbol{\gamma}^{(l)}, \quad b^{(l)} \sim \mathcal{N}\left(0, \sigma_1^2\right), \quad \boldsymbol{\gamma}^{(l)} \sim \mathrm{Mult}(1, \boldsymbol{\pi}) \tag{Equation 4}$$

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 $\mathbf{b}^{(l)}$ using the residuals from the model $\mathbf{y} - \sum_{l' \neq l} \mathbf{X}\mathbf{b}^{(l')}$. At the end of training, the SuSiE model provides L estimated coefficient vectors $\widehat{\mathbf{b}}$ and L corresponding PIP vectors $\boldsymbol{\alpha}^{(l)} = \{\Pr[b_1^{(l)} \neq 0 \mid \mathbf{y}, \mathbf{X}], \ldots, \Pr[b_J^{(l)} \neq 0 \mid \mathbf{y}, \mathbf{X}]\}$. Computation of a final inclusion probability assumes that effects are independent across the L different models and is computed as

$$\mathrm{PIP}_j \equiv \Pr\left[\widehat{b}_j \neq 0 | \mathbf{y}, \mathbf{X}\right] \approx 1 - \prod_{l=1}^{L}\left(1 - \alpha_j^{(l)}\right) \tag{Equation 5}$$

A key component of SuSiE is that it uses these PIPs to naturally construct credible sets. Effectively, a level ρ credible set $\mathcal{S}(\boldsymbol{\alpha}, \rho)$ can be estimated by simply sorting variables in descending order and then including variables into the set until their cumulative probability exceeds ρ (Wang et al., 2020). Below we give the rigorous definition for credible sets.

Definition 1 (Wang et al., (2020))

In the context of a multiple-regression model, a level ρ credible set is defined to be a subset of variables that has probability ρ 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 (Carbonetto and Stephens, 2012; Tibshirani, 1996; Zou and Hastie, 2005) where the effect sizes for an arbitrarily selected subset of correlated variables will be penalized while the others are retained.

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 (Leshno et al., 1993; Barron, 1993). For simplicity, we will focus on multi-layer perceptrons 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

$$g(\boldsymbol{\mu}) = \mathbf{f} \quad = \mathbf{Z}_K \boldsymbol{\Theta}_K + \epsilon_K, \quad \ldots, \quad \mathbf{z}_k = h(\mathbf{Z}_{k-1}\boldsymbol{\Theta}_{k-1} + \epsilon_{k-1}), \quad \ldots, \quad \mathbf{z}_1 = h(\mathbf{X}\mathbf{W} + \epsilon_0) \tag{Equation 6}$$

where, in expectation, the response variable is related to the input data by $\mathbb{E}[\mathbf{y} \mid \mathbf{X}] = \boldsymbol{\mu}$; $\mathbf{f}$ 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 $\mathbf{y}$ is continuous or the logit if $\mathbf{y}$ is binary; $\mathbf{Z}_k$ denotes the matrix of nonlinear neurons from the k-th hidden layer with corresponding weight matrix $\boldsymbol{\Theta}_k$; ε_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 $\mathbf{W}$ 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 $\mathbf{W}$. Let H_k represent the number of neurons in the k-th hidden layer such that $\mathbf{W}$ 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 $\mathbf{w}_{j\bullet}$ denote the j-th row of the weight matrix $\mathbf{W}$. We place a grouped "single-effect" shrinkage prior on the input weights

$$\mathbf{W} = \mathbf{A} \circ \boldsymbol{\Gamma}, \quad \mathbf{a} \sim \mathcal{N}(0, \sigma_1^2 \mathbf{I}), \quad \boldsymbol{\gamma} \sim \text{Mult}(1, \boldsymbol{\pi}) \qquad \text{(Equation 7)}$$

where $\boldsymbol{\Gamma}$ is a matrix that is H_1 copies of the binary vector $\boldsymbol{\gamma}$, $\mathbf{a}$ is a H_1-dimensional row-vector of continuous weights in $\mathbf{A} = [\mathbf{a}_{1\bullet}, \ldots, \mathbf{a}_{J\bullet}]$, 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 (Chen et al., 2021; Ghosh et al., 2019), except that the binary indicator variable $\boldsymbol{\gamma}$ 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 Figure 1). 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 $\boldsymbol{\gamma}$ 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, ESNN maybe specified as a summation of L-latent nonlinear models of the form

$$\mathbf{f}^{(l)} = \mathbf{Z}_K^{(l)} \boldsymbol{\Theta}_K^{(l)} + \epsilon_K^{(l)}, \quad \ldots, \quad \mathbf{z}_k^{(l)} = h\left(\mathbf{Z}_{k-1}^{(l)} \boldsymbol{\Theta}_{k-1} + \epsilon_{k-1}^{(l)}\right), \quad \ldots, \quad \mathbf{z}_1^{(l)} = h\left(\mathbf{X}\mathbf{W}^{(l)} + \epsilon_0^{(l)}\right) \qquad \text{(Equation 8)}$$

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

$$\mathbf{W}^{(l)} = \mathbf{A}^{(l)} \circ \boldsymbol{\Gamma}^{(l)}, \quad \mathbf{a}^{(l)} \sim \mathcal{N}(0, \sigma_1^2 \mathbf{I}), \quad \boldsymbol{\gamma}^{(l)} \sim \text{Mult}(1, \boldsymbol{\pi}) \qquad \text{(Equation 9)}$$

Notice that at the end of training, each l-th neural network will also yield an estimated set of input layer weights $\widehat{\mathbf{W}}$ and a corresponding set of inclusion probabilities $\boldsymbol{\alpha}^{(l)} = \{\Pr[\mathbf{w}_{1\bullet}^{(l)} \neq 0 \mid \mathbf{y}, \mathbf{X}], \ldots, \Pr[\mathbf{w}_{J\bullet}^{(l)} \neq 0 \mid \mathbf{y}, \mathbf{X}]\}$ 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 $\mathcal{S}(\boldsymbol{\alpha}, \rho)$ in the same way as SuSiE by defining the overall posterior inclusion probabilities as

$$\text{PIP}_j \equiv \Pr\left[\widehat{\mathbf{w}}_{j\bullet} \neq 0 \mid \mathbf{y}, \mathbf{X}\right] \approx 1 - \prod_{l=1}^{L} \left(1 - \alpha_j^{(l)}\right) \qquad \text{(Equation 10)}$$

which we use to determine variable significance. A motivating example of the benefits of the ESNN model can be found in Figure S6.

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 (Equations 8 and 9) and replace the intractable true posterior distribution over the parameters $p(\mathbf{W}_{1:L}, \boldsymbol{\Gamma}_{1:L} \mid \mathcal{D})$ with an approximating family of distributions $q(\mathbf{W}_{1:L}, \boldsymbol{\Gamma}_{1:L}; \varphi_{1:L})$—where we use shorthand $1:L = 1, \ldots, L$ to represent the L models in the ensemble, $\varphi_{1:L}$ represent the collection of free parameters in the approximations, and $\mathcal{D}$ 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)

$$\mathcal{L}(\varphi_{1:L}) = \mathbb{E}_q\left[\log p(\mathbf{y} \mid \mathbf{W}_{1:L}, \boldsymbol{\Gamma}_{1:L}, \mathcal{D})\right] + \text{KL}\left(q(\mathbf{W}_{1:L}, \boldsymbol{\Gamma}_{1:L}; \varphi_{1:L}) \| \, p(\mathbf{W}_{1:L}, \boldsymbol{\Gamma}_{1:L})\right) \qquad \text{(Equation 11)}$$

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

Algorithm 1. Training Algorithm for the ESNN Framework

Input genotype data X and phenotypic vector **y**.
Choose the number of models L, number of maximum iterations T, and credible set level ρ.
Randomly initialize variational parameters $\varphi_1, \ldots, \varphi_L$ for the L models.
Initialize the models $l = 1$ and iterations $t = 1$.
while $l \leq L$ and $t \leq T$ **do**
  Fix hyper-parameters $\varphi_1, \ldots, \varphi_{l-1}$.
  Sample $\mathrm{W}^{(l)}, \boldsymbol{\Gamma}^{(l)} \sim q(\varphi_l)$ using re-parameterization trick.
  Compute the approximate log likelihood $\mathcal{L}(\varphi_1, \ldots, \varphi_L; \mathcal{D})$.
  Compute the gradients for only φ_l using the approximate log likelihood.
  Update φ_l using the gradients with optimizers.
  Compute PIPs and credible sets for the l-th model.
  if $\lambda_l > 1$ **then** ◁ "Purity" Check
    $l = l + 1$
    if y is continuous **then**
      $\mathrm{y} = \mathrm{y} - \sum_{m=1}^{l-1} \boldsymbol{f}^{(m)}$ ◁ IBSS Procedure
    end if
  end if
  $t = t + 1$
end while
Compute (marginal) posterior inclusion probabilities (PIP) for each variable.
Determine credible sets $\mathcal{S}(\boldsymbol{\alpha}, \rho)$.
Return $\{\mathrm{PIP}, \mathcal{S}(\boldsymbol{\alpha}, \rho)\}$.

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

$$q\left(\mathrm{W}^{(l)}, \boldsymbol{\Gamma}^{(l)}\right) = q\left(\mathrm{A}^{(l)}\right) q\left(\boldsymbol{\Gamma}^{(l)}\right), \quad q\left(\mathrm{a}^{(l)}\right) = \mathcal{N}\left(\mathrm{m}, \tau_1^2 \mathrm{I}\right), \quad q\left(\boldsymbol{\gamma}^{(l)}\right) = \mathrm{Mult}(1, \boldsymbol{\kappa}) \qquad \text{(Equation 12)}$$

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 below for theoretical details and corresponding pseudocode in Algorithm 1). 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, ESNN can be applied for both types of data analyses.

Details of 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

$$q(\mathrm{a}) = \mathcal{N}\left(\mathrm{m}, \tau_1^2 \mathrm{I}\right) \quad \Leftrightarrow \quad \mathrm{a} = \mathrm{m} + \tau_1 \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1) \qquad \text{(Equation 13)}$$

This technique has been shown to successfully reduce the variance of gradients (Kingma and Welling, 2014) and stabilizes the training process. Next, we assume that the indicator variables $\boldsymbol{\gamma}^{(l)}$ 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 (Jang et al., 2017; Maddison et al., 2017),

$$\tilde{\gamma}_j \sim \frac{\exp\left(\log(\alpha_j) + \vartheta_j\right)/\delta}{\sum_{j'} \exp\left(\log(\alpha_{j'}) + \vartheta_{j'}\right)/\delta}, \quad \vartheta_j = -\log\left(-\log(v_j)\right), \quad v_j \sim \mathcal{U}(0,1), \tag{Equation 14}$$

where $\tilde{\gamma}_j$ are the approximate samples for γ, τ is a temperature parameter, and v_j uniformly sampled random variable. As $\tau \rightarrow 0$, 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 α_j is also important for the ESNN model as it directly influences the performance of variable selection. Importantly, α_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 α can be very small and hinder convergence during training. This problem is commonly known as "vanishing gradients" (Hochreiter, 1998). For our work, we found that simply scaling up the learning rate when updating α_j works well in practice. Note that the Kullback-Leibler (KL) divergence term in the approximate likelihood can be decomposed as the following

$$\begin{aligned} \mathrm{KL}\left(q(\mathbf{W}, \boldsymbol{\Gamma}; \varphi) \| p(\mathbf{W}, \boldsymbol{\Gamma})\right) = & \sum_{j=1}^{J} q(\gamma_j = 1; \varphi) \mathrm{KL}\left(q(\mathbf{a}_{j\bullet} | \gamma_j = 1, \varphi) \| p(\mathbf{a}_{j\bullet} | \gamma_j = 1)\right) \\ & + \mathrm{KL}\left(q(\gamma_j = 1; \varphi) \| p(\gamma_j = 1)\right) \end{aligned} \tag{Equation 15}$$

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

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 (Wang et al., 2020). It is worth noting that, when the model is uncertain about which variables to choose (e.g., when there are no significant effect variables), α 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 (Wang et al., 2020). 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 φ are not optimal. In the scenario where the model is unable to find correct effect variable, regressing out φ 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

$$\lambda^{(l)} = \frac{\mathcal{L}(\varphi_1, \ldots, \varphi_{l-1}, \varphi_l, \varphi_{l+1}, \ldots, \varphi_L)}{\mathcal{L}(\varphi_1, \ldots, \varphi_{l-1}, \varphi_{l+1}, \ldots, \varphi_L)} \tag{Equation 16}$$

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.

QUANTIFICATION AND STATISTICAL ANALYSIS

Our study made use of three real datasets. The simulation results made use of imputed data released from the UK Biobank (Bycroft et al., 2018). Quality control for these data were carried out using the following procedure. 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 exact test $P > 1 \times 10^{-6}$), (*iv*) SNPs with missingness greater than 1 %, and (*v*) SNPs in high linkage disequilibrium (using the flag –indep-pairwise 50 5 0.9 with PLINK 1.9 (Purcell et al., 2007)). After all QC steps, we had a final dataset of $N = 349{,}414$ individuals from which we could downsample and $J = 36{,}518$ SNPs on the first chromosome. Next, we used the NCBI's Reference Sequence (RefSeq) database in the UCSC Genome Browser (Pruitt et al., 2005) to annotate SNPs with appropriate genes. We defined genes using the UCSC gene boundary and augmenting those boundaries by adding SNPs within a $\pm$ 500 kilobase (kb) buffer to account for possible regulatory elements. Genes with only one SNP within their boundary were excluded.

One part of the analysis results in this work made use of GWA data from the Wellcome Trust Centre for Human Genetics. This study contains a total of $N = 1{,}814$ heterogeneous stock of mice from 85 families (all descending from eight inbred progenitor strains) (Valdar et al., 2006), and 131 quantitative traits that are classified into six broad categories including behavior, diabetes, asthma, immunology, haematology, and biochemistry. Here, we focused on two specific phenotypes from these categories including: high-density lipoprotein content (Biochem.HDL) and low-density lipoprotein content (Biochem.LDL). Both phenotypes were corrected for sex, age, body weight, season, year, and cage effects. For individuals with missing genotypes, we imputed values by the mean genotype of that SNP in their corresponding mouse family. Only polymorphic SNPs with minor allele frequency above 5% were kept for the analyses. This left a total of $J = 10{,}346$ autosomal SNPs that were available for all mice.

The second part of the data analysis used data from the Wellcome Trust Case Control Consortium (WTCCC) one study (Wellcome Trust Case Control Consortium, 2007) which consists of about 14,000 cases of seven common diseases, including 1,963 cases of type 1 diabetes (T1D) and 1,924 cases of type 2 diabetes (T2D), as well as 2,938 shared controls. We selected a total of 458,868 shared single nucleotide polymorphisms (SNPs) following a previous study (Zhou et al., 2013).