iScience, Volume *25*

Supplemental information

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

Wei Cheng, Sohini Ramachandran, and Lorin Crawford

Supplementary Figures

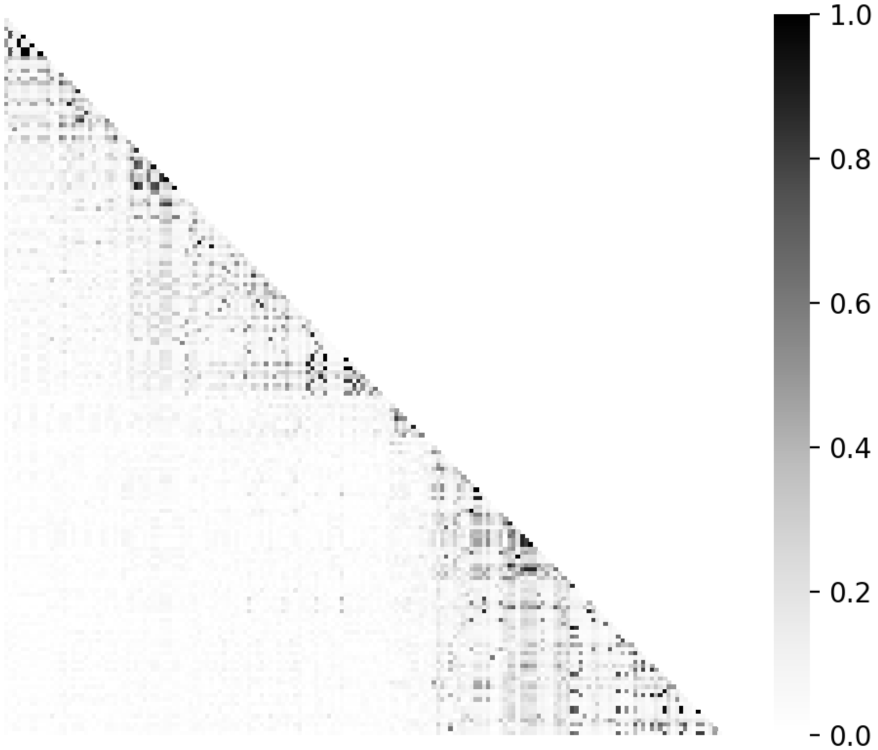


Figure S1. An example of absolute correlation matrix for genotype data, related to Figures 2 and 3.

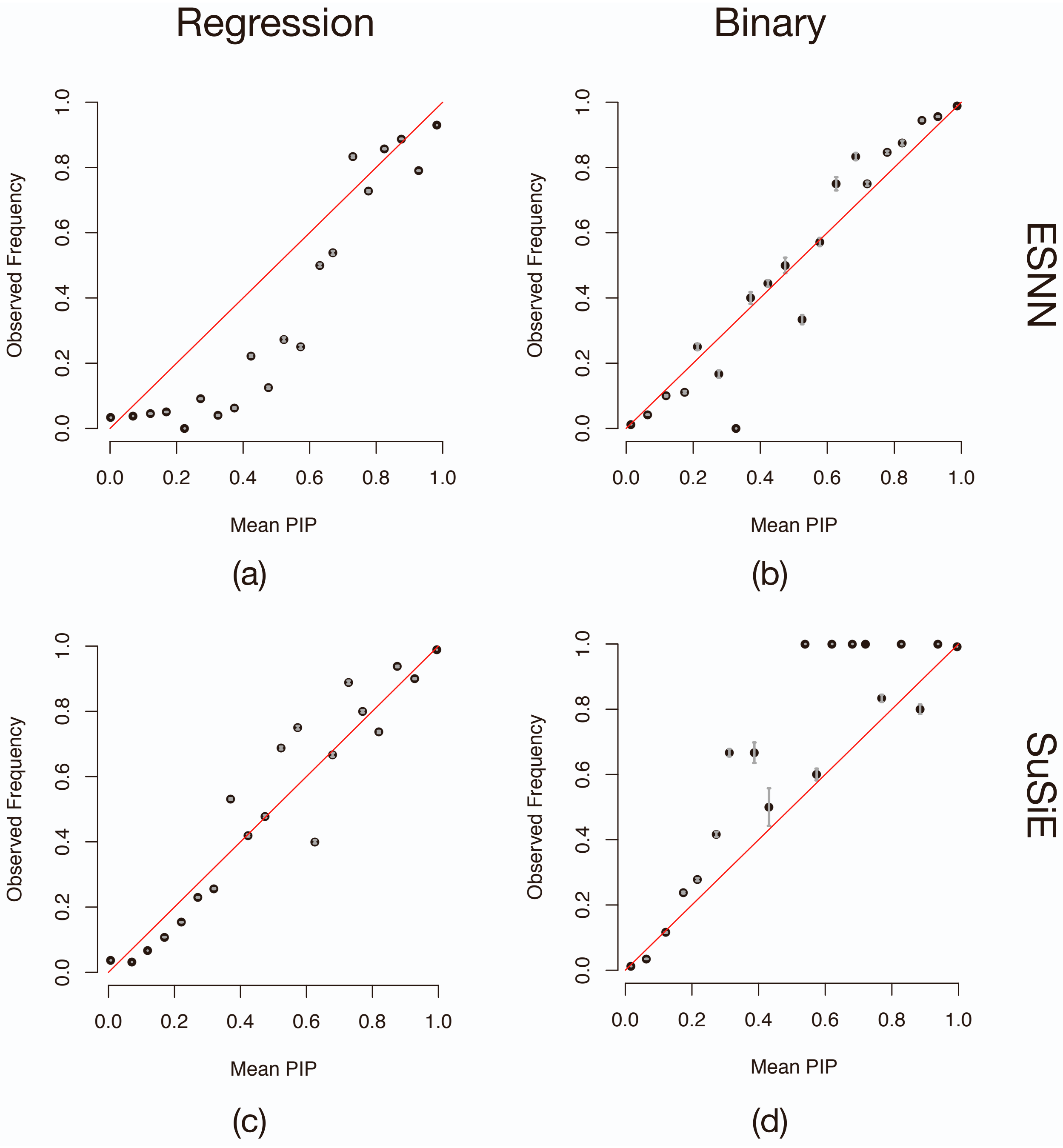


Figure S2. Assessments of posterior inclusion probability (PIP) calibration for ESNN and SuSiE, related to Figures 2 and 3. This experiment follows largely from previous work. Here, SNPs are grouped into bins according to their reported PIPs (using 20 equally spaced bins, from 0 to 1).The plots show 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). Grey 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.

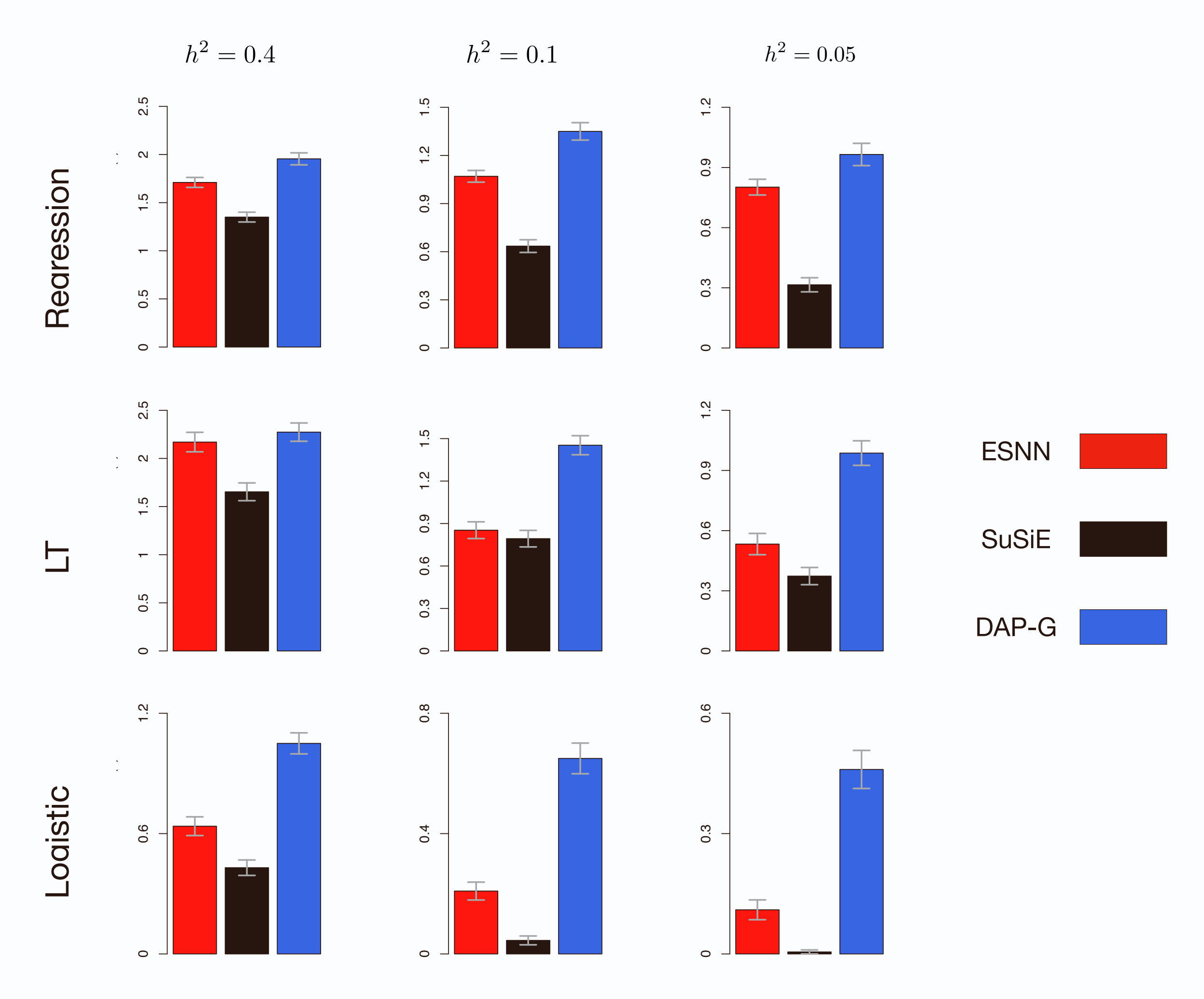


Figure S3. Comparisons of the number of covered effect variables per simulation run for ESNN, SuSiE, and DAP-G in simulation studies under different levels of heritability, related to Figures 2 and 3. Results are based on 200 data replicates with standard errors represented by the grey bars.

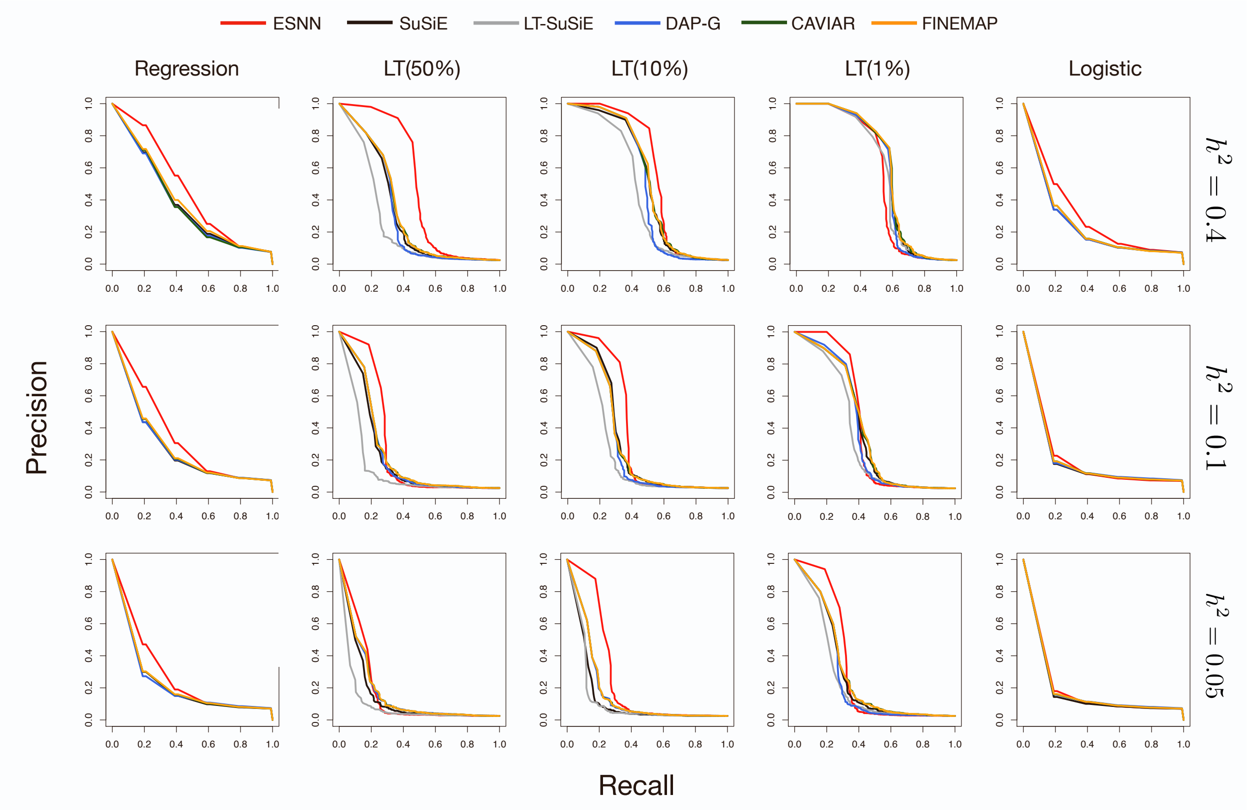


Figure S4. Precision and recall curves for simulation studies of different scenarios, related to Figures 2 and 3. Results are based on 200 data replicates.

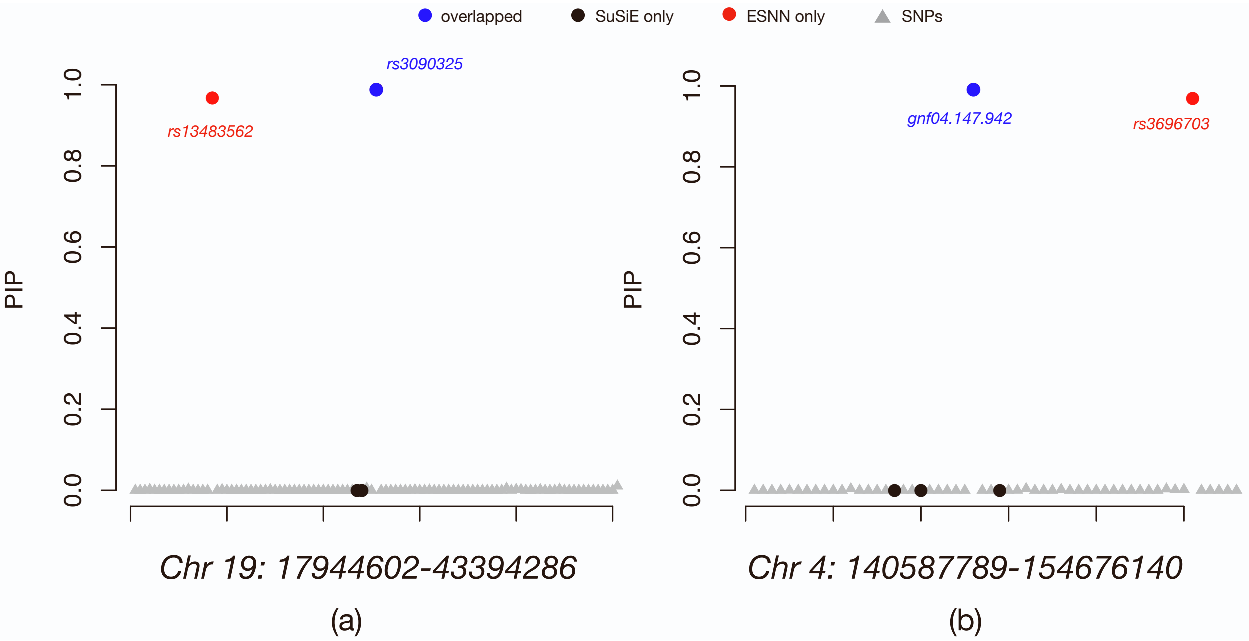
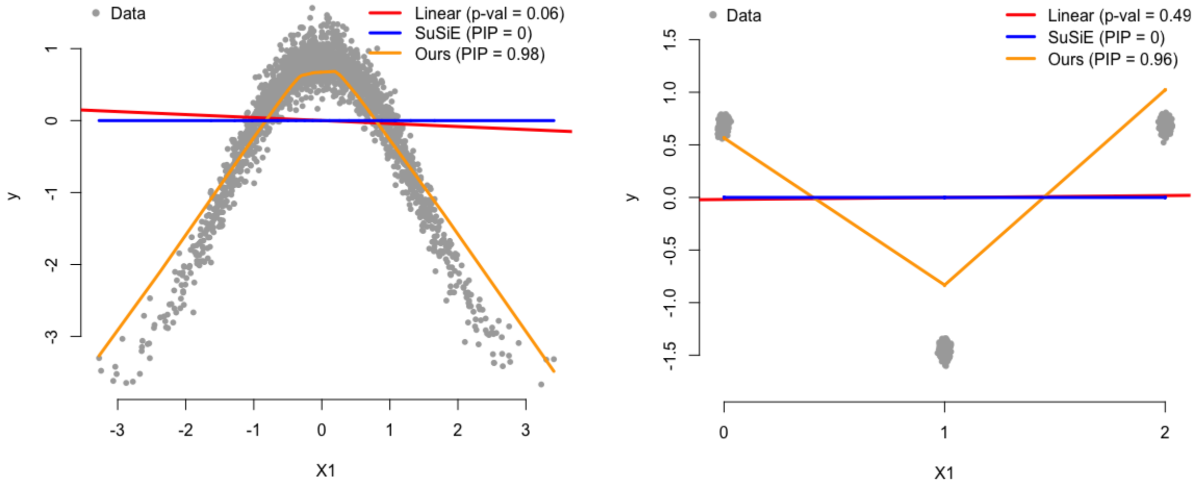


Figure S5. Posterior inclusion probabilities (PIP) of ESNN and SuSiE for the heterogenous stock of mice dataset from the Wellcome Trust Centre for Human Genetics, related to Figure 4. (a) Highlighted region for low-density lipoprotein (LDL). Significant SNPs found only by ESNN (included in the credible 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).



(a)

(b)

Figure S6. Toy example demonstrating importance of accounting for nonlinearity when performing variable selection, related to Figure 1. To demonstrate that linear models lose power when non-additive variation exists, we generate two simulated datasets. **(a)** In the first case, we simulate $\mathbf{x}_1 \sim \mathcal{N}(0, 1)$ and then generate responses under $\mathbf{y} = \cos(\mathbf{x}_1) + \mathbf{e}$ where $\mathbf{e} \sim \mathcal{N}(0, 1)$. The real data are plotted in grey points. We next run a univariate linear model (red line) and SuSiE (blue line) 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 $\mathbf{x}_1$ as being significant. We then run the ESNN model (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 $\mathbf{y}$ by assuming the heterozygote has a significant effect. Similar relationships have been shown in the literature. Similar to the first simulation case, linear methods fail to capture the causal effect while the ESNN approach is robust to the non-additive architecture. These two toy examples illustrate the importance of accounting for nonlinearity in variable selection methods.