Supplementary material for

Simultaneous inference of phenotype-associated genes and relevant tissues from GWAS data via Bayesian integration of multiple tissue-specific gene networks

Mengmeng Wu, Zhixiang Lin, Shining Ma, Ting Chen, Rui Jiang, Wing Hung Wong

Supplementary Text

Derivation of the equation (4)

According to the equation (2), the joint distribution of $\boldsymbol{z}$ is specified as:

$$p(\boldsymbol{z} \mid \boldsymbol{\Phi}, \mathbf{W}) = \frac{1}{T(\boldsymbol{\Phi})} \exp\left\{ \gamma \sum_{i=1}^{N} z_i + \sum_{k=1}^{K} \beta_k \sum_{1 \le i < j \le N} w_{ij}^{(k)} \mathrm{I}(\mathrm{z}_i = z_j) \right\}$$

Then,

$$p(z_i = 1 \mid \boldsymbol{z}_{-i}, \boldsymbol{\Phi}, \mathbf{W}) = \frac{p(z_i = 1, \boldsymbol{z}_{-i}, \boldsymbol{\Phi}, \mathbf{W})}{p(z_i = 1, \boldsymbol{z}_{-i}, \boldsymbol{\Phi}, \mathbf{W}) + p(z_i = 0, \boldsymbol{z}_{-i}, \boldsymbol{\Phi}, \mathbf{W})}$$

$$= \frac{\frac{1}{T(\boldsymbol{\Phi})} \exp\left\{ \gamma + \sum_{k=1}^{K} \beta_k \sum_{j \ne i} w_{ij}^{(k)} \mathrm{I}(\mathrm{z}_j = 1) + C \right\}}{\frac{1}{T(\boldsymbol{\Phi})} \exp\left\{ \gamma + \sum_{k=1}^{K} \beta_k \sum_{j \ne i} w_{ij}^{(k)} \mathrm{I}(\mathrm{z}_j = 1) + C \right\} + \frac{1}{T(\boldsymbol{\Phi})} \exp\left\{ \sum_{k=1}^{K} \beta_k \sum_{j \ne i} w_{ij}^{(k)} \mathrm{I}(\mathrm{z}_j = 0) + C \right\}}$$

$$= \frac{\exp\left\{ \gamma + \sum_{k=1}^{K} \beta_k \sum_{j \ne i} w_{ij}^{(k)} \mathrm{I}(\mathrm{z}_j = 1) \right\}}{\exp\left\{ \gamma + \sum_{k=1}^{K} \beta_k \sum_{j \ne i} w_{ij}^{(k)} \mathrm{I}(\mathrm{z}_j = 1) \right\} + \exp\left\{ \sum_{k=1}^{K} \beta_k \sum_{j \ne i} w_{ij}^{(k)} \mathrm{I}(\mathrm{z}_j = 0) \right\}}$$

$$= \frac{1}{1 + \exp\left\{ -\left(\gamma + \sum_{k=1}^{K} \beta_k \sum_{j \ne i} w_{ij}^{(k)} \left(\mathrm{I}(\mathrm{z}_j = 1) - \mathrm{I}(\mathrm{z}_j = 0) \right) \right) \right\}}$$

$$= \frac{1}{1 + \exp\left\{ -\left(\gamma + \sum_{k=1}^{K} \beta_k \sum_{j \ne i} w_{ij}^{(k)} \left(2 z_j - 1 \right) \right) \right\}}$$

$$= \frac{1}{1 + \exp\left\{ -\left(\gamma + \sum_{k=1}^{K} \beta_k x_{ik} \right) \right\}}$$

where C is an expression that is irrelevant to z_i, and $x_{ik} = \sum_{j \ne i} w_{ij}^{(k)} (2 z_j - 1)$. Then, we can easily obtain the equation (4).

Parameter initialization via a simple model

We resorted to a simple two-component mixture model of p-values for the initialization of α_0 and α_1, in which association status of genes were assumed to be independent. With a similar approach as described above, we introduced a hidden indicator z_i to gene i, indicting the association status of the gene and the phenotype of interest, and we use the same equation (1) to describe the conditional distributions of *p*-values given the hidden indicators. The distribution for the hidden indicators is specified without the MRF prior, as:

$$p(\boldsymbol{z}) = \prod_{i=1}^{N} p(z_i)$$
$$p(z_i = 1) = \pi_0$$

Parameters of this simple model include $\Phi_0 = \{\alpha_0, \alpha_1, \pi_0\}$ and can be estimated using the expectation maximization (EM) algorithm implemented as iterative alternation between the E-step and the M-step, as

E-step:

$$q_i = p(\mathrm{z}_i = 1 \mid p_i, \Phi_0) = \frac{\pi_0 \alpha_1 p_i^{\alpha_1 - 1}}{\pi_0 \alpha_1 p_i^{\alpha_1 - 1} + (1 - \pi_0)\alpha_0 p_i^{\alpha_0 - 1}}$$

M-step:

$$\pi = \frac{1}{N}\sum_{i=1}^{N} q_i, \alpha_1 = -\frac{\sum_{i=1}^{N} q_i}{\sum_{i=1}^{N} q_i \log p_i}, \alpha_0 = -\frac{\sum_{i=1}^{N} (1 - q_i)}{\sum_{i=1}^{N} (1 - q_i) \log p_i}$$

Empirically, the EM procedure converged rapidly, and we initialized parameters with $\alpha_0 = 1, \alpha_1 = 0.2, \pi_0 = 0.1$, which were found to work well in practice. The estimated parameters α_0 and α_1 were then served as the starting point for the MCMC sampling and were observed to speed up convergence as expected. The hidden indicator $\boldsymbol{z}$ was initialized by

$z_i \sim \text{Bernoulli}(q_i)$. The initialization procedures for other parameters are specified as: 1) parameters γ and β are initialized as zeros; 2) parameters γ and $\boldsymbol{I}$ are sampled given the other parameters according to equations (9) and (10)

Simulation studies for different genetic characteristics

Our model assumes that genetic characteristics of a phenotype could be described by three parameters α_0, α_1 and γ, among which the latest two determine statistical properties that a gene is associated with a phenotype and are of particular interest. Specifically, α_1 controls the shape of the distribution of *p*-values for genes associated with a phenotype, and γ controls the probability that a gene is associated with a phenotype without considering the contribution of gene networks. To study the performance of our model in different combinations of these genetic characteristics, we conducted similar simulation studies as the previous section, except that we varied α_1 and γ, where $\alpha_1 \in \{0.05, 0.1, 0.2\}$ and $\gamma \in \{-3, -2, -1\}$. First, we found that our method could correctly estimate these parameters (Supplementary Figure 2). We then compared the performance of our method under different settings and presented the result in Supplementary Figure 3. As expected, parameter γ determines the number of associated genes, with larger γ resulting in more associated genes, as shown in Supplementary Figure 3 (A). We then calculated the average improvement of performance in identifying associated genes under different settings, with the use of the *p*-value approach served as a baseline. As shown in Supplementary Figure 3 (B), the improvement is more pronounced when $\alpha_1 = 0.2$ than $\alpha_1 = 0.1$ and $\alpha_1 = 0.05$, implying more space for improvement when the association strength is weaker (i.e., larger value of α_1). This is reasonable because the statistical power is already

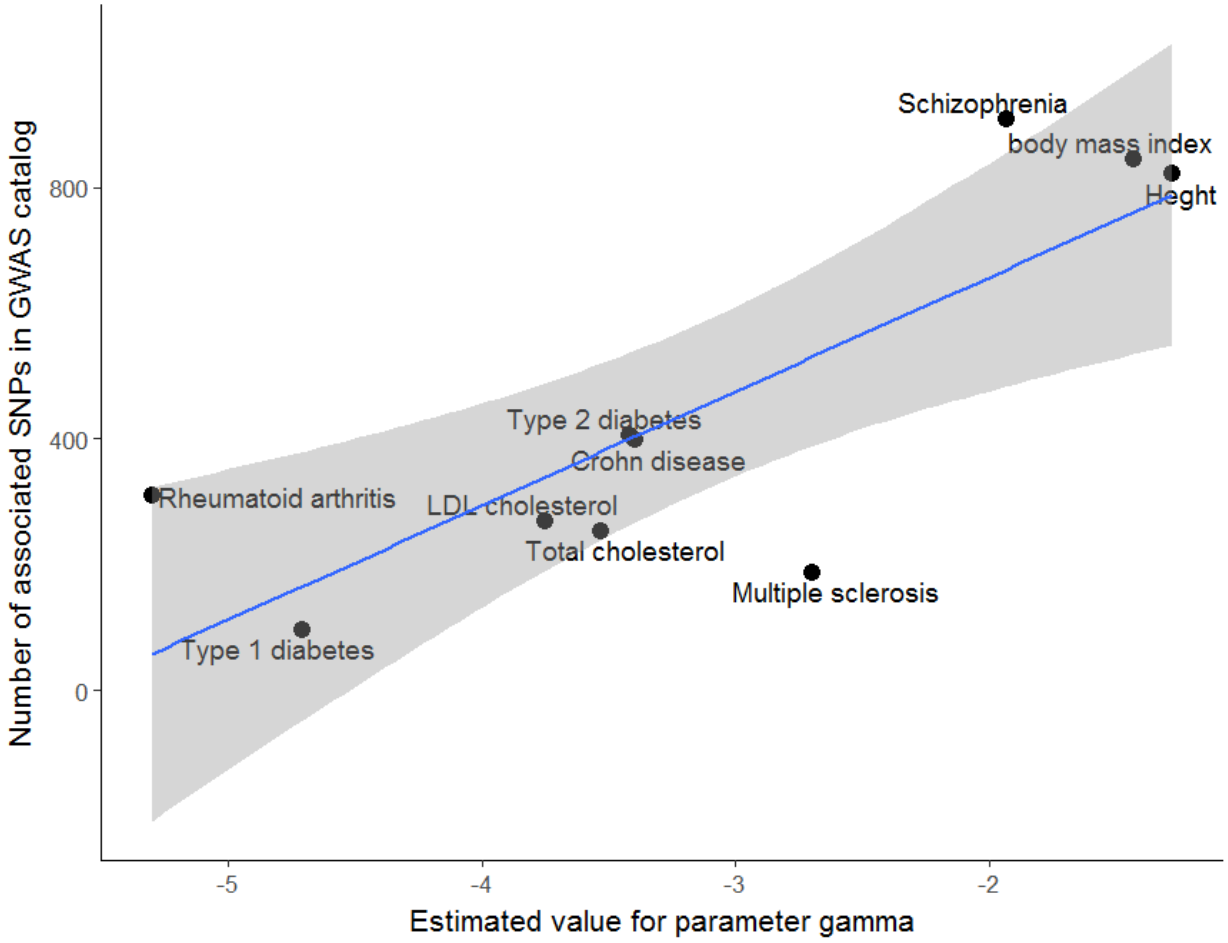
high for small values of α_1. At a fixed value of α_1, the improvement of performance increases with the increase of γ, because more genes are associated with the phenotype for larger γ, and hence the identification of associated genes become easier. As for the identification of relevant tissues, the power of our method increases with the increase of γ at a fixed α_1, as shown in Supplementary Figure 3 (C). This is also reasonable because larger γ means more associated genes, which makes it easier to estimate the effect sizes of different gene networks and identify corresponding relevant tissues. In summary, all the above evidence supports the effectiveness of our method under different genetic characteristics.

GO analysis of complex diseases

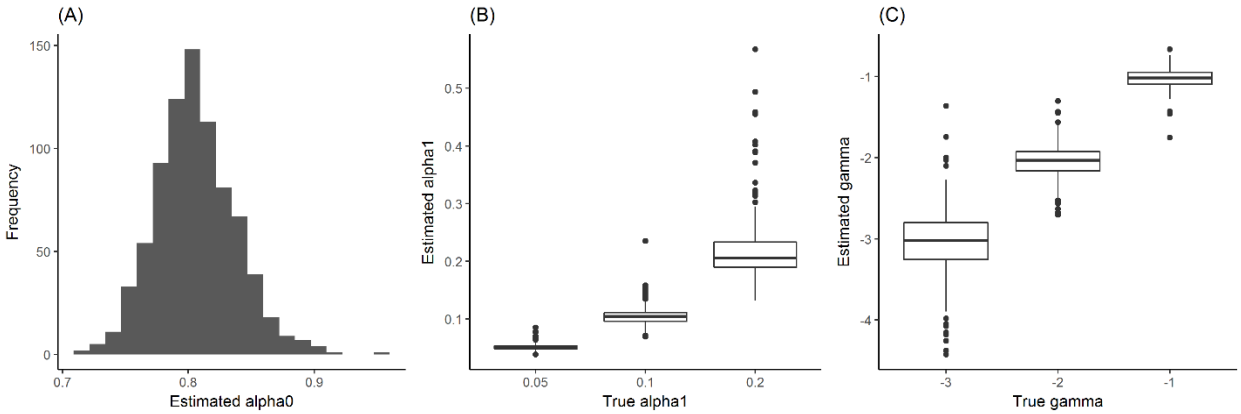
Using the same procedure as the one used in the main text, we performed GO enrichment analysis for Rheumatoid Arthritis, Crohn's Disease, Osteoporosis and Multiple Sclerosis and drew the corresponding figures, including Supplementary Figure 11-14. As shown in these figures, the prioritized genes given by SIGNET show stronger enrichment in some GOs while less enrichment in other GOs compared with GWAS only. In detail, we found that the GOs enhanced by SIGNET had more clear phenotype-associated biological meanings. For example, we observed that many immune-related GOs showed more significant enrichment for Rheumatoid Arthritis, Crohn's Disease and Multiple Sclerosis, and all of the three diseases were immune-related. For Osteoporosis, we found that skeletal system development were lifted by SIGNET, and it made sense because Osteoporosis was bone-related diseases. Therefore, SIGNET showed ability to improve discovery of phenotype-associated GOs.

Supplementary Figures

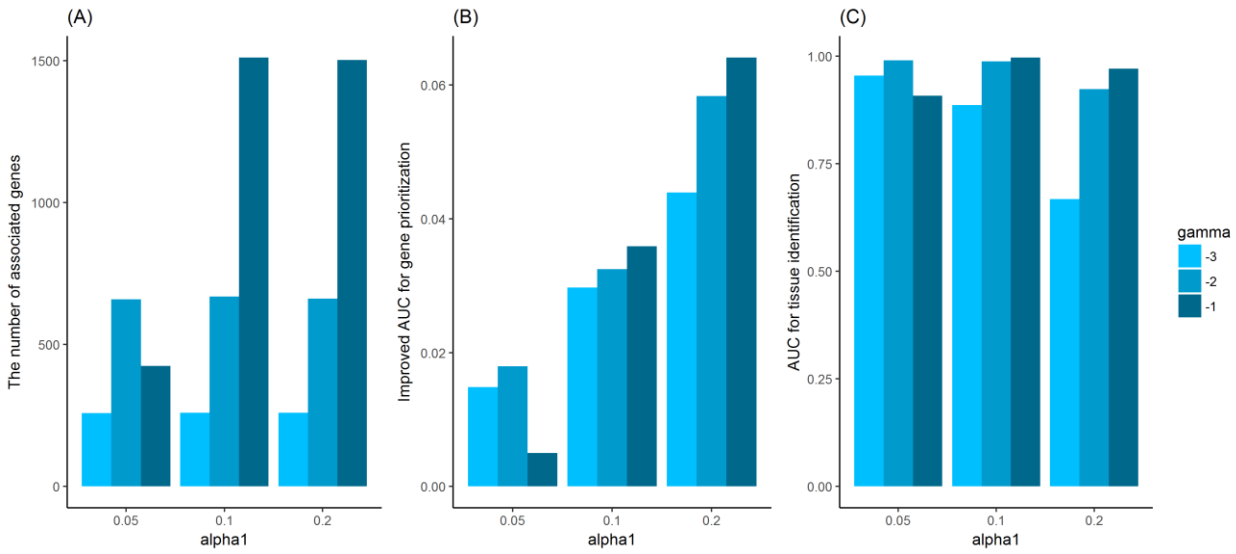
Supplementary Figure S1. The relationship between estimated values of γ and the number of associated SNPs. Each point represents a complex trait, x axis denotes the estimated value of gamma from the SIGNET, and y axis denotes the number of associated SNPs (from the GWAS catalog database). The blue line denotes the fitted line for linear regression and the shaded regions represents standard deviation.



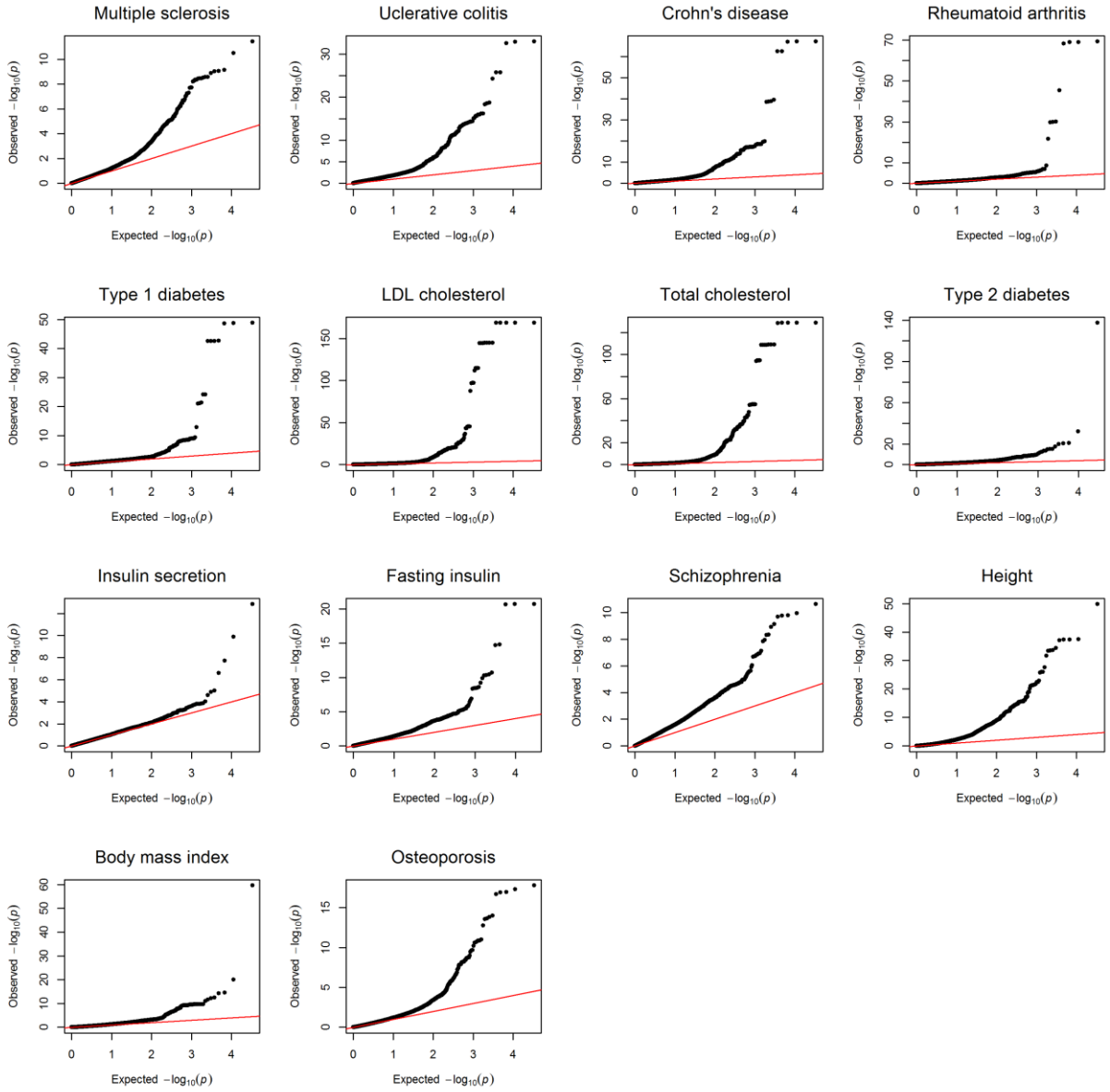
Supplementary Figure S2. Parameters estimation for different genetic characteristics in simulation studies. (A) Estimated values of α_0, with real value being 0.8. (B) Estimated values of α_1, with real values being 0.05, 0.1, and 0.2, respectively. (C) Estimated values of γ, with real values being -3, -2 and -1, respectively.



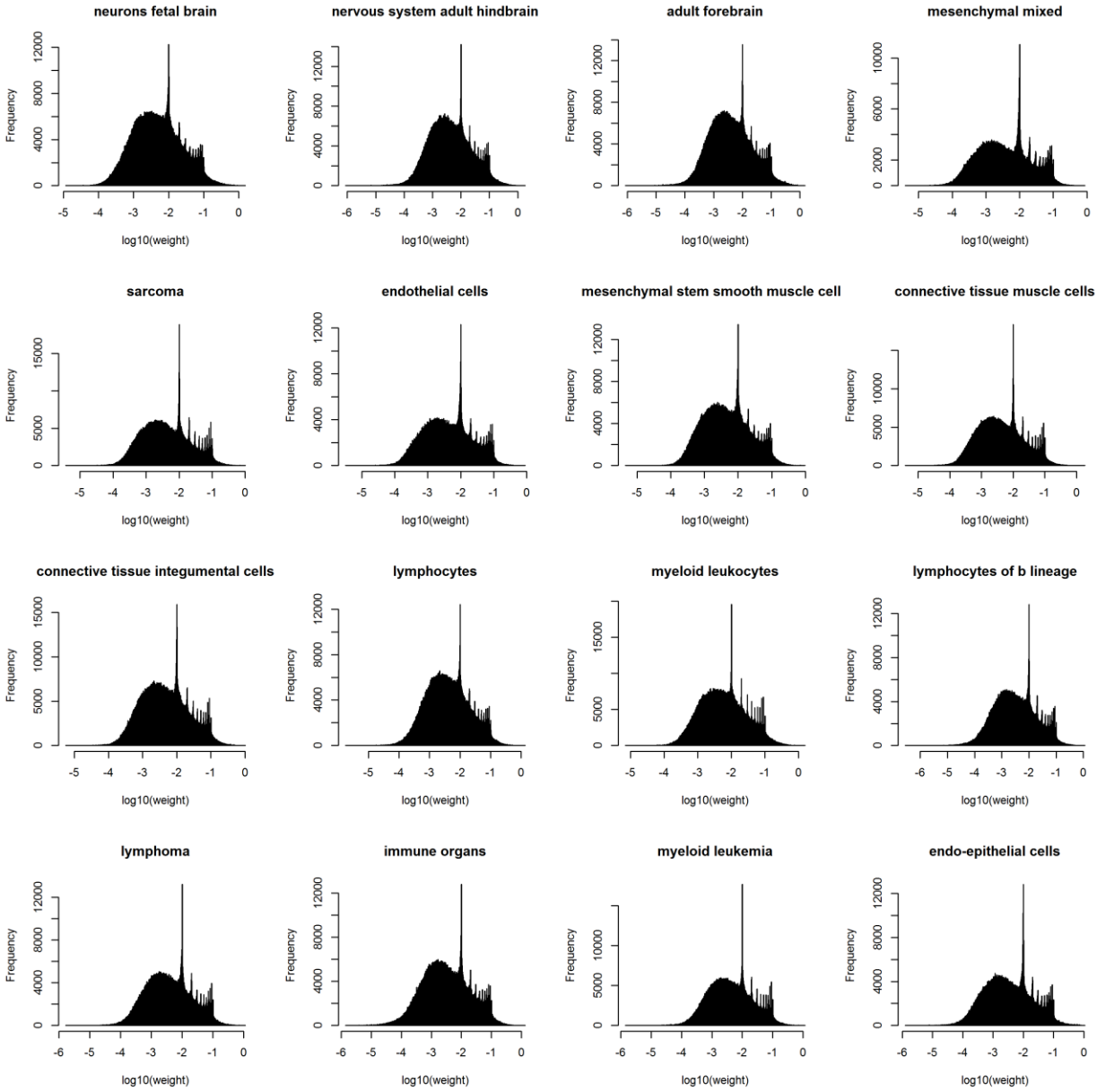
Supplementary Figure S3. Results for simulation studies with different genetic characteristics. (A) Numbers of associated genes under different simulation settings; (B) average improvement of SIGNET in AUC for gene prioritization compared with *p*-value under different simulation settings; (C) AUCs of SIGNET for tissue identification under different simulation settings.



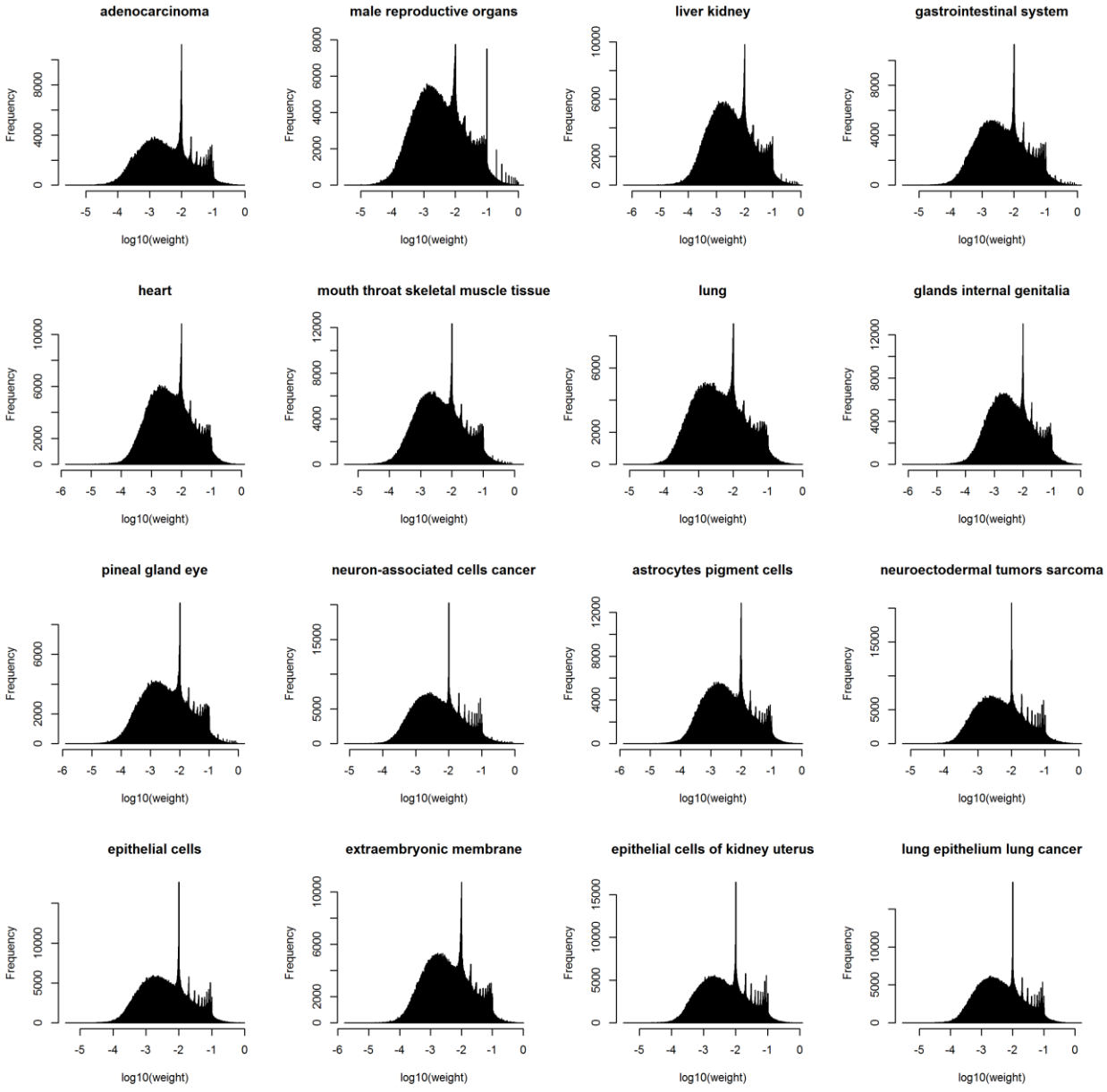
Supplementary Figure S4. QQ plots of the gene-level *p*-values of the 14 complex traits analyzed in the main txt. In each subplot, x axis and y axis represents quantiles of $-\log_{10}(\text{p-value})$ under uniform distribution and observed empirical distribution. The red line denotes $y = x$ and the back line denotes the quantile-quantile (QQ) plot.



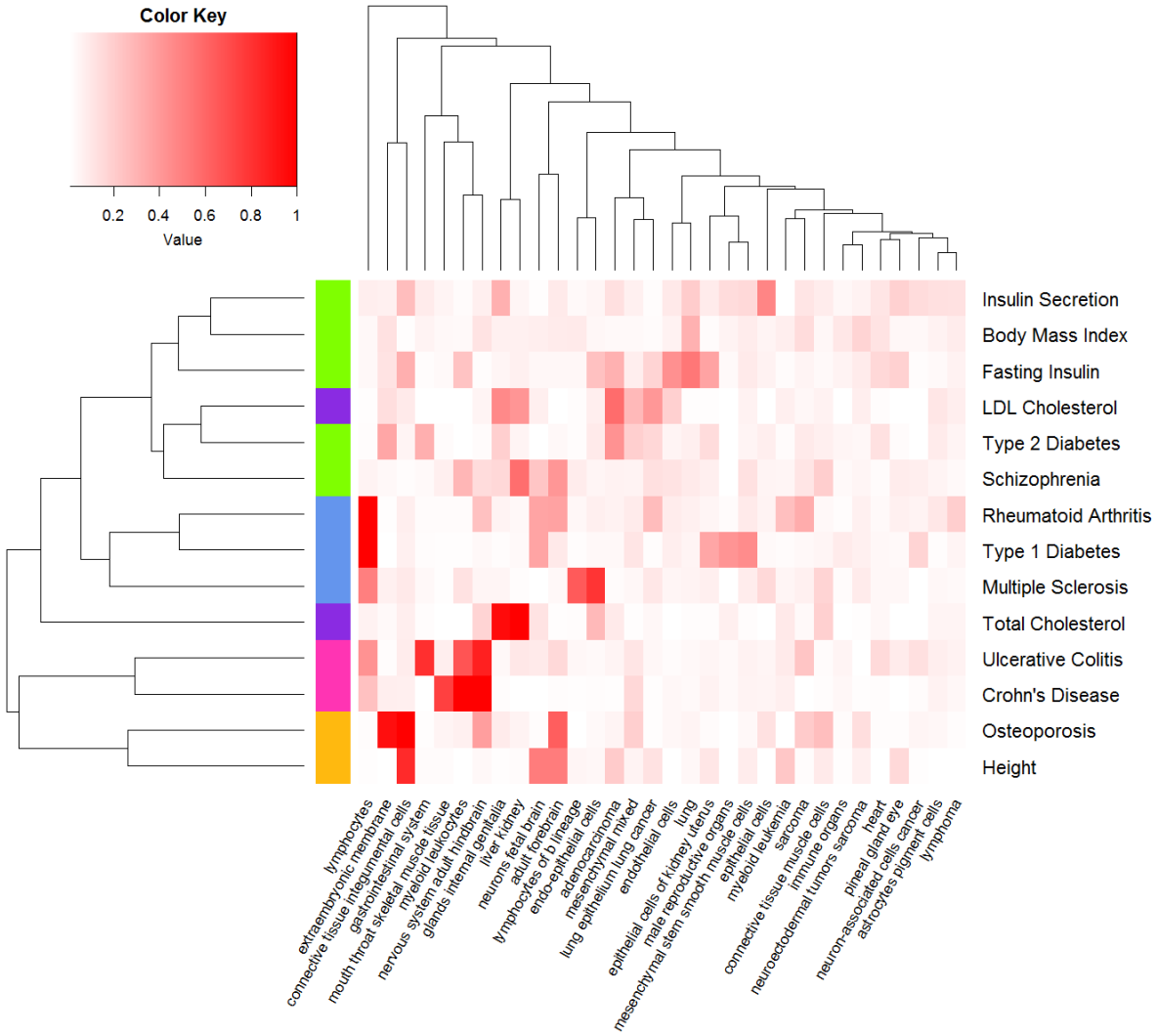
Supplementary Figure S5. Distributions of edge weights across the 32 tissue-specific gene networks. In each subplot, x axis denotes $\log_{10}(\text{weight})$ and y axis denotes corresponding frequency.



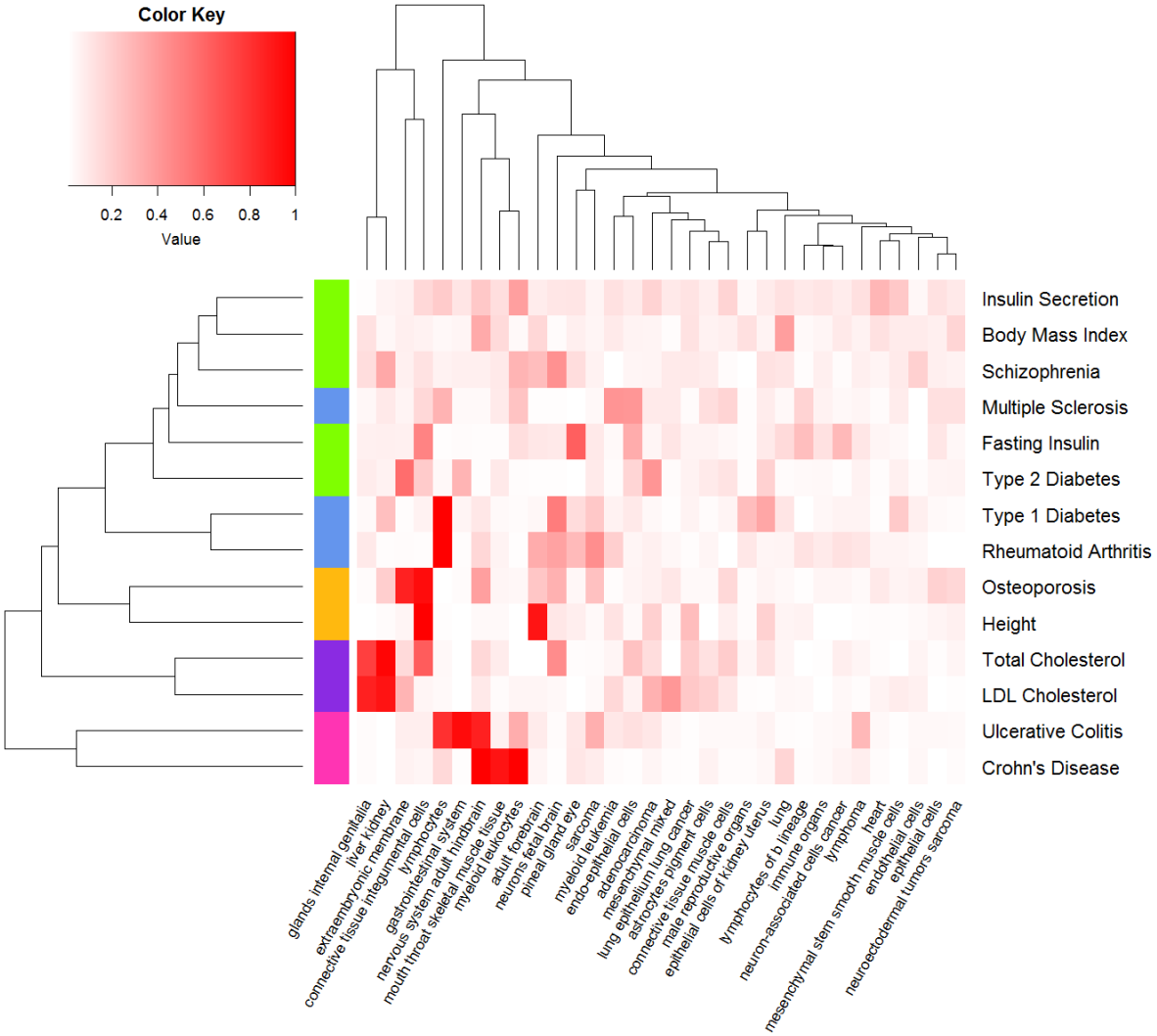
Supplementary Figure S5. Continued.



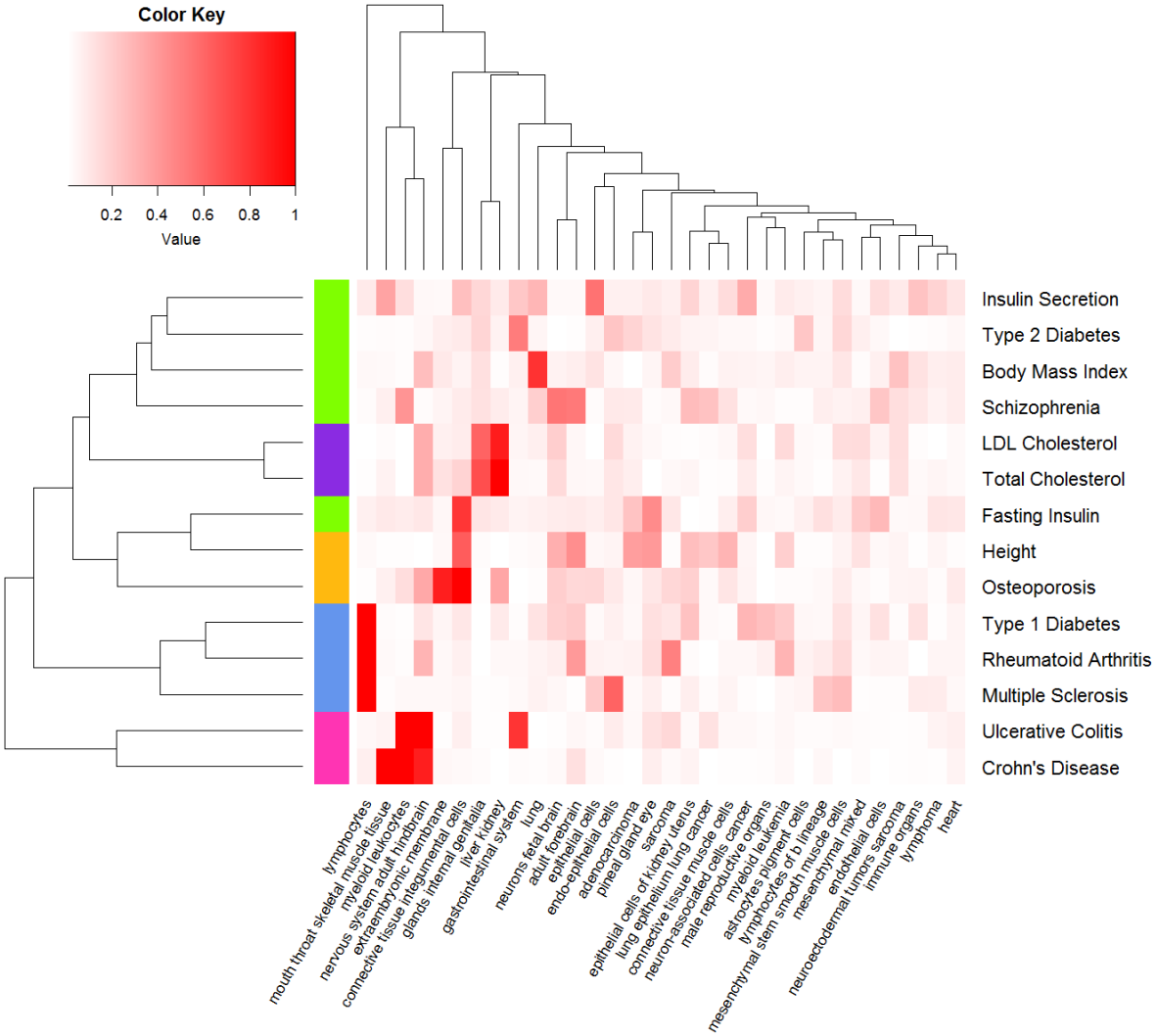
Supplementary Figure S6. Cluster analysis of the 14 complex traits by their PIPs across the 32 tissues on the 32 filtered tissue-specific gene regulatory networks (threshold: 0.0001). Each column denotes a tissue, and each row represents the vector of PIPs across the 32 tissues for a complex trait. The from-white-to-red color represents the value of PIP from low to high. Cluster assignments for phenotypes are based on the result of cluster analysis on original networks (Figure 4 in the main text).



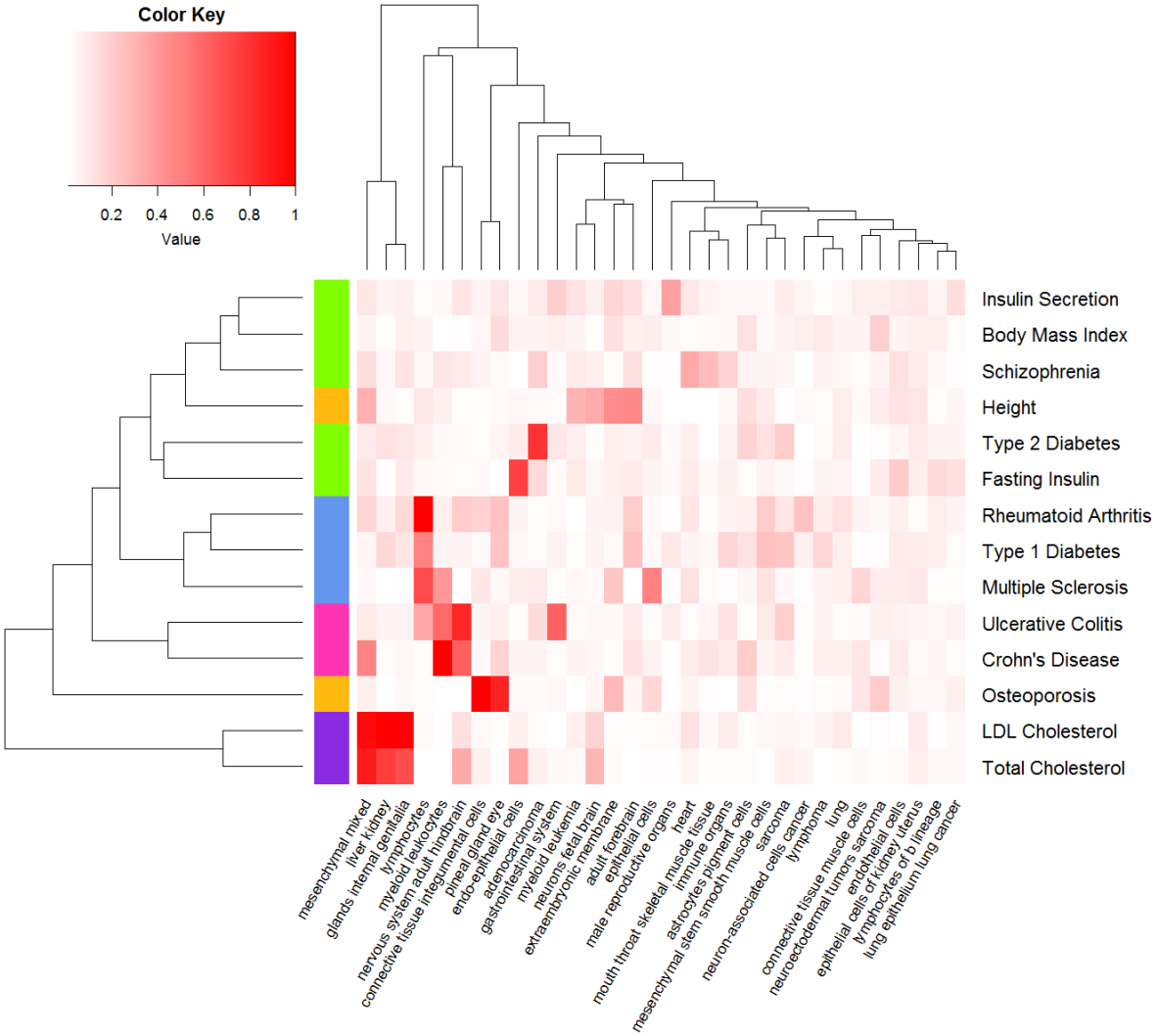
Supplementary Figure S7. Cluster analysis of the 14 complex traits by their PIPs across the 32 tissues on the 32 filtered tissue-specific gene regulatory networks (threshold: 0.001). Each column denotes a tissue, and each row represents the vector of PIPs across the 32 tissues for a complex trait. The from-white-to-red color represents the value of PIP from low to high. Cluster assignments for phenotypes are based on the result of cluster analysis on original networks (Figure 4 in the main text).



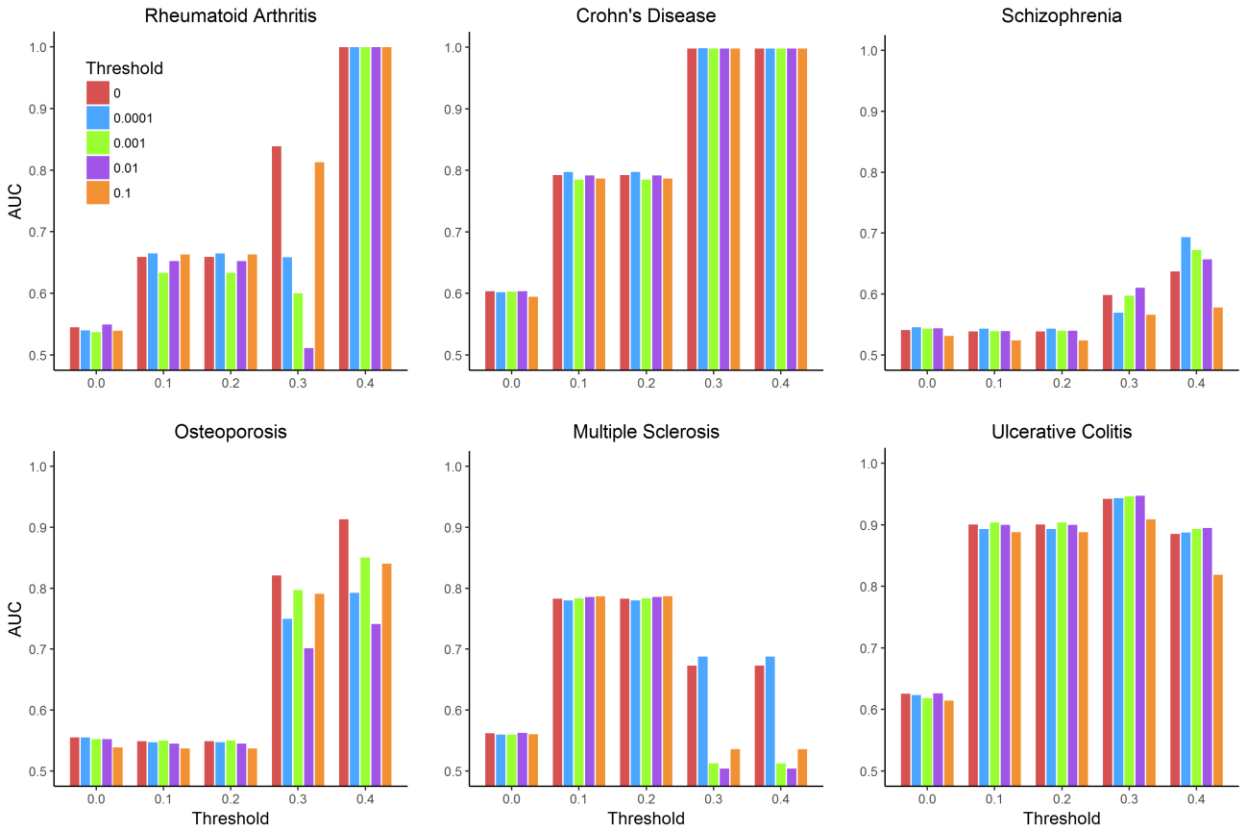
Supplementary Figure S8. Cluster analysis of the 14 complex traits by their PIPs across the 32 tissues on the 32 filtered tissue-specific gene regulatory networks (threshold: 0.01). Each column denotes a tissue, and each row represents the vector of PIPs across the 32 tissues for a complex trait. The from-white-to-red color represents the value of PIP from low to high. Cluster assignments for phenotypes are based on the result of cluster analysis on original networks (Figure 4 in the main text).



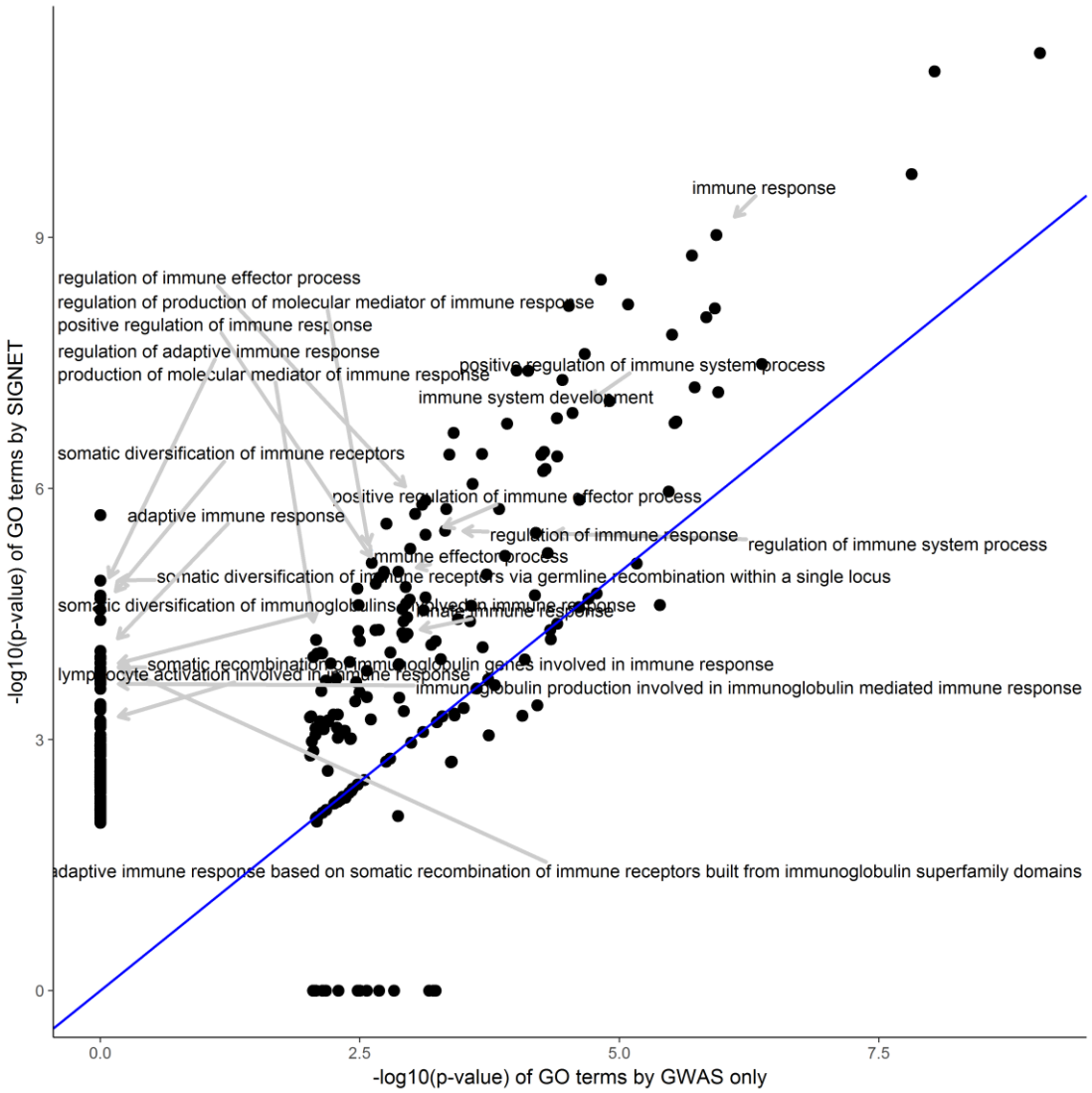
Supplementary Figure S9. Cluster analysis of the 14 complex traits by their PIPs across the 32 tissues on the 32 filtered tissue-specific gene regulatory networks (threshold: 0.1). Each column denotes a tissue, and each row represents the vector of PIPs across the 32 tissues for a complex trait. The from-white-to-red color represents the value of PIP from low to high. Cluster assignments for phenotypes are based on the result of cluster analysis on original networks (Figure 4 in the main text).



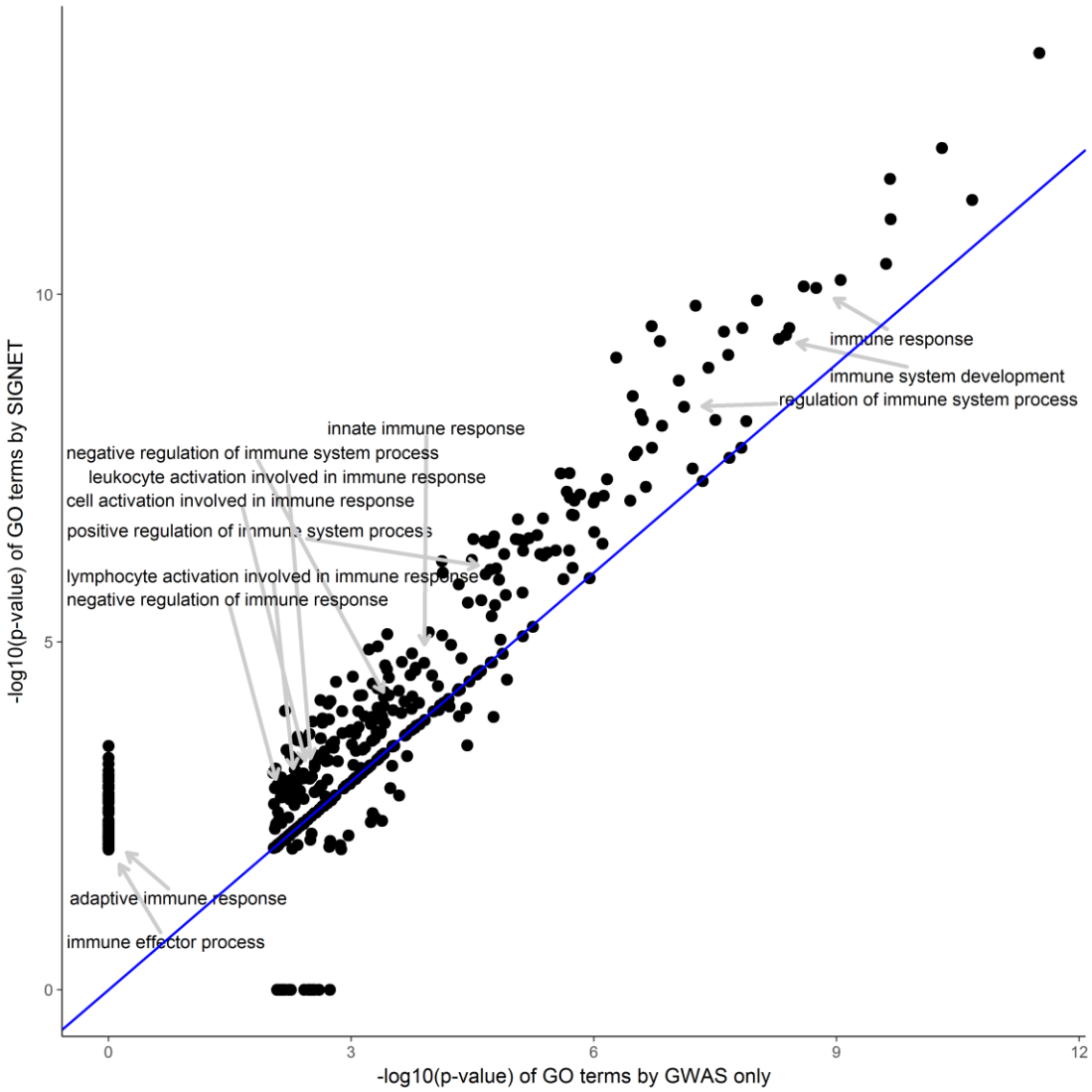
Supplementary Figure S10. The influence of network edge filtering on gene prioritization. For each one of the six complex diseases, including rheumatoid arthritis; Crohn's disease; schizophrenia; osteoporosis; multiple sclerosis and ulcerative colitis, we extracted corresponding disease genes with evidence scores from DisGeNET. The AUCs of SIGNET on different networks filtered by different thresholds are computed under different thresholds for the evidence score. In each subplot, the x axis denotes the threshold of the evidence score, and the y axis indicates AUC.



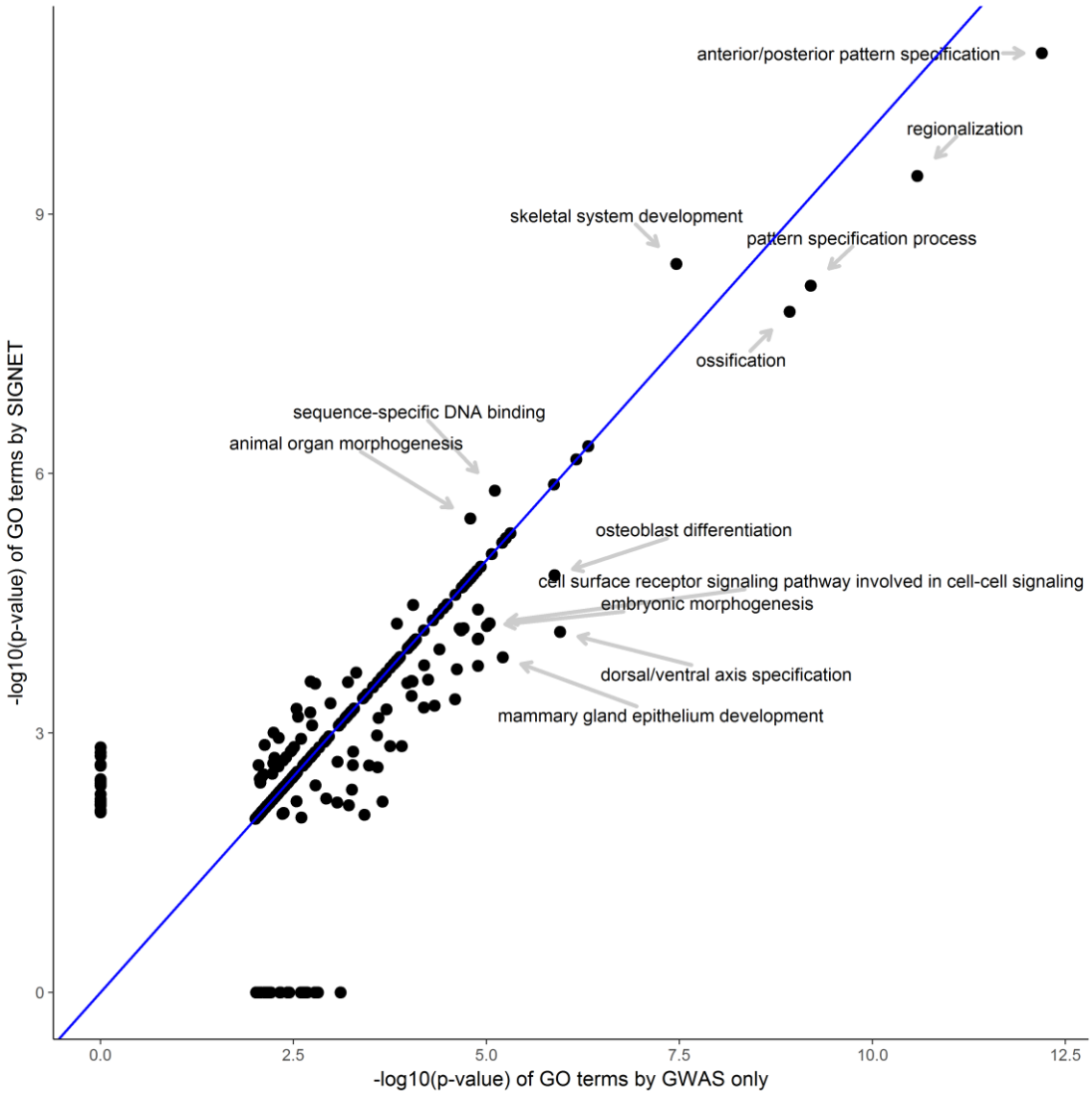
Supplementary Figure S11. GO analysis for Rheumatoid Arthritis. Using the top 23 genes (global FDR ≤ 0.05) ranked by *p*-value (or GWAS only) and SIGNET, we conducted gene ontology (GO) enrichment analysis and compared the significance of each GO term given by the two methods. Each point represents a GO term, and x axis and y axis denotes the $-\log_{10}(\text{p-value})$ obtained by *p*-value and SIGNET.



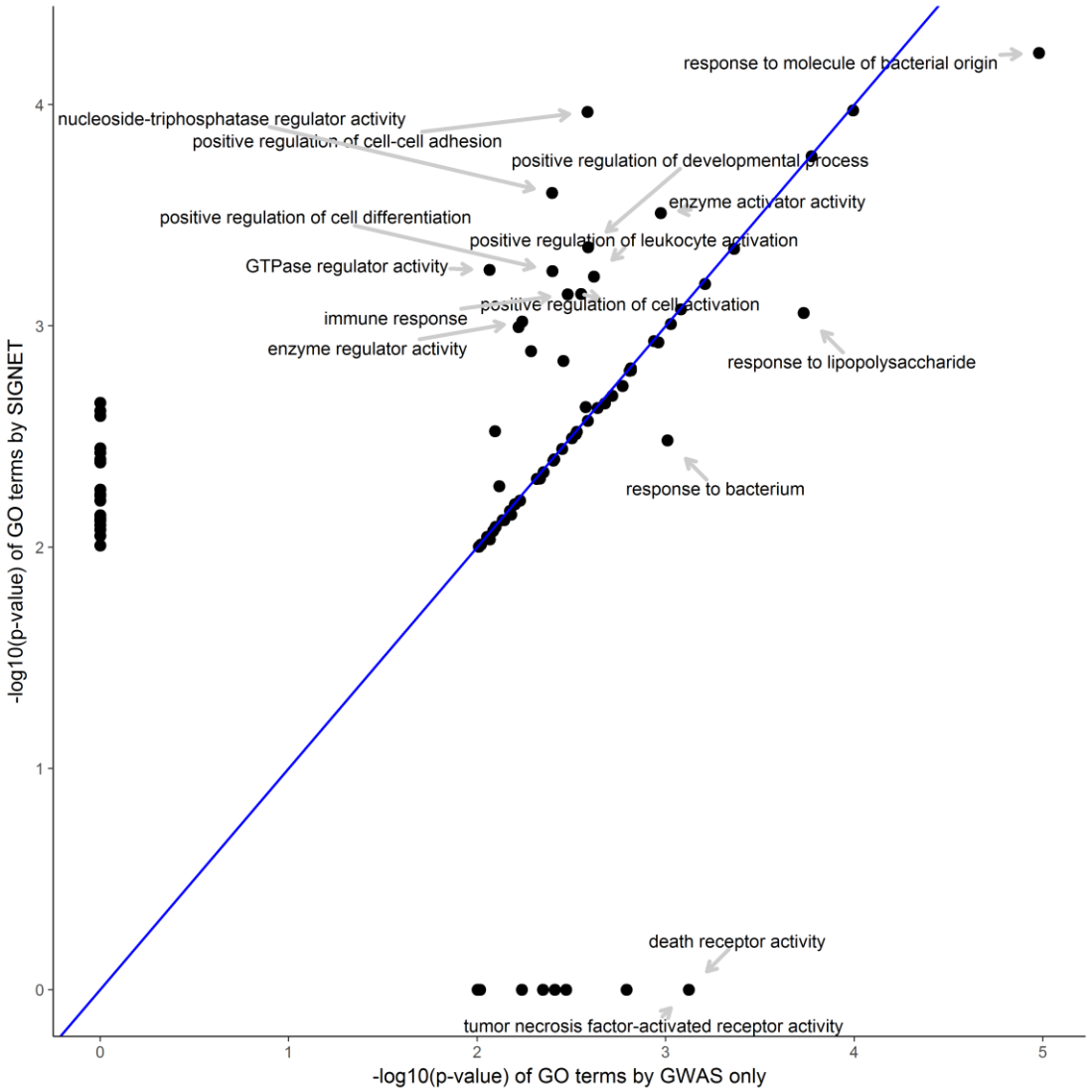
Supplementary Figure S12. GO analysis for Crohn's Disease. Using the top 204 genes (global FDR ≤ 0.05) ranked by *p*-value (or GWAS only) and SIGNET, we conducted gene ontology (GO) enrichment analysis and compared the significance of each GO term given by the two methods. Each point represents a GO term, and x axis and y axis denotes the $-\log_{10}(\text{p-value})$ obtained by *p*-value and SIGNET.



Supplementary Figure S13. GO analysis for Osteoporosis. Using the top 115 genes (global FDR $\leq$ 0.05) ranked by *p*-value (or GWAS only) and SIGNET, we conducted gene ontology (GO) enrichment analysis and compared the significance of each GO term given by the two methods. Each point represents a GO term, and x axis and y axis denotes the $-\log_{10}(\text{p-value})$ obtained by *p*-value and SIGNET.



Supplementary Figure S14. GO analysis for Multiple Sclerosis. Using the top 115 genes (global FDR $\leq$ 0.05) ranked by *p*-value (or GWAS only) and SIGNET, we conducted gene ontology (GO) enrichment analysis and compared the significance of each GO term given by the two methods. Each point represents a GO term, and x axis and y axis denotes the $-\log_{10}(\text{p-value})$ obtained by *p*-value and SIGNET.



Supplementary Tables

Supplementary Table S1. Performance comparison between different algorithms on Rheumatoid Arthritis. K denotes the number of top ranked genes. Each entry denotes the number of associated genes (retrieved from the DisGeNet database) in the top k genes given by each algorithm. The largest numbers for each k are bolded.

k	p-value	SIGNET (single)	NetWAS	SIGNET
100	37	37	15	**44**
200	54	53	32	**63**
300	67	70	52	**87**
400	93	92	65	**109**
500	109	111	86	**128**
600	123	121	98	**141**
700	136	131	107	**158**
800	145	139	117	**169**
900	153	151	127	**178**
1,000	171	165	134	**192**

Supplementary Table S2. Performance comparison between different algorithms on Crohn Disease. K denotes the number of top ranked genes. Each entry denotes the number of associated genes (retrieved from the DisGeNet database) in the top k genes given by each algorithm. The largest numbers for each k are bolded.

k	p-value	SIGNET (single)	NetWAS	SIGNET
100	49	50	14	**50**
200	75	75	24	**77**
300	84	84	27	**90**
400	89	89	35	**100**
500	97	99	37	**107**
600	106	107	51	**118**
700	111	112	60	**121**
800	114	117	69	**130**
900	120	122	79	**139**
1,000	127	129	83	**141**

Supplementary Table S3. Performance comparison between different algorithms on Schizophrenia. K denotes the number of top ranked genes. Each entry denotes the number of associated genes (retrieved from the DisGeNet database) in the top k genes given by each algorithm. The largest numbers for each k are bolded.

k	p-value	SIGNET (single)	NetWAS	SIGNET
100	25	25	12	**27**
200	41	42	21	**41**
300	52	54	32	**54**
400	65	68	42	**70**
500	78	83	53	**83**
600	87	90	68	**93**
700	102	101	73	**108**
800	114	117	86	**121**
900	126	129	102	**134**
1,000	135	140	116	**141**

Supplementary Table S4. Performance comparison between different algorithms on Osteoporosis. K denotes the number of top ranked genes. Each entry denotes the number of associated genes (retrieved from the DisGeNet database) in the top k genes given by each algorithm. The largest numbers for each k are bolded.

k	p-value	SIGNET (single)	NetWAS	SIGNET
100	**18**	**18**	8	17
200	21	**22**	9	**22**
300	26	26	12	**27**
400	**30**	**30**	15	**30**
500	32	32	18	**36**
600	37	37	22	**39**
700	**41**	40	27	**41**
800	42	42	32	**43**
900	47	**48**	36	**48**
1,000	50	50	38	**52**

Supplementary Table S5. Performance comparison between different algorithms on Multiple Sclerosis. K denotes the number of top ranked genes. Each entry denotes the number of associated genes (retrieved from the DisGeNet database) in the top k genes given by each algorithm. The largest numbers for each k are bolded.

k	p-value	SIGNET (single)	NetWAS	SIGNET
100	14	14	13	**16**
200	25	25	30	**31**
300	37	37	37	**39**
400	**47**	**47**	46	**47**
500	55	55	55	**62**
600	60	60	66	**68**
700	65	65	**80**	72
800	69	69	**84**	77
900	76	74	**95**	89
1,000	82	81	**103**	96

Supplementary Table S6. Performance comparison between different algorithms on Ulcerative Colitis. K denotes the number of top ranked genes. Each entry denotes the number of associated genes (retrieved from the DisGeNet database) in the top k genes given by each algorithm. The largest numbers for each k are bolded.

k	p-value	SIGNET (single)	NetWAS	SIGNET
100	**20**	**20**	10	**20**
200	37	37	23	**39**
300	52	52	30	**54**
400	61	60	41	**72**
500	71	71	50	**83**
600	82	81	56	**95**
700	90	89	63	**104**
800	98	95	70	**110**
900	102	101	76	**116**
1,000	106	106	84	**127**