


# Modeling regulatory network topology improves genome-wide analyses of complex human traits

Xiang Zhu <sup>1,2,3✉</sup>, Zhana Duren<sup>3,4</sup> & Wing Hung Wong <sup>3,5✉</sup>

Genome-wide association studies (GWAS) have cataloged many significant associations between genetic variants and complex traits. However, most of these findings have unclear biological significance, because they often have small effects and occur in non-coding regions. Integration of GWAS with gene regulatory networks addresses both issues by aggregating weak genetic signals within regulatory programs. Here we develop a Bayesian framework that integrates GWAS summary statistics with regulatory networks to infer genetic enrichments and associations simultaneously. Our method improves upon existing approaches by explicitly modeling network topology to assess enrichments, and by automatically leveraging enrichments to identify associations. Applying this method to 18 human traits and 38 regulatory networks shows that genetic signals of complex traits are often enriched in interconnections specific to trait-relevant cell types or tissues. Prioritizing variants within enriched networks identifies known and previously undescribed trait-associated genes revealing biological and therapeutic insights.

<sup>1</sup>Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA. <sup>2</sup>Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA 16802, USA. <sup>3</sup>Department of Statistics, Stanford University, Stanford, CA 94305, USA. <sup>4</sup>Center for Human Genetics, Department of Genetics and Biochemistry, Clemson University, Greenwood, SC 29646, USA. <sup>5</sup>Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA 94305, USA. ✉email: [xiangzhu@psu.edu](mailto:xiangzhu@psu.edu); [whwong@stanford.edu](mailto:whwong@stanford.edu)

Genome-wide association studies (GWAS) have cataloged many significant and reproducible associations between common genetic variants, notably single-nucleotide polymorphisms (SNPs), and diverse human complex traits<sup>1</sup>. However, it remains challenging<sup>2</sup> to translate these findings into biological mechanisms and clinical applications, because most trait-associated variants have individually small effects and map to non-coding sequences.

One hypothesis is that non-coding variants cumulatively affect complex traits through cell type- or tissue-specific<sup>3</sup> gene regulation<sup>4</sup>. To test this hypothesis, large-scale epigenomic<sup>5,6</sup> and transcriptomic<sup>7–10</sup> data have been made available spanning diverse human cell types and tissues. Exploiting these data many studies have shown enrichments of trait-associated SNPs in chromatin regions<sup>11–13</sup> and genes<sup>14–16</sup> that are active in trait-relevant cell types or tissues. These studies overlap regulatory maps with GWAS data and often ignore functional interactions among loci within regulatory programs.

Gene regulatory networks<sup>17–20</sup> have proven useful in mining functional interactions of genes from genomic data. Transcriptional regulatory interactions, rather than gene expression alone, drive tissue specificity<sup>19</sup>. Further, context-specific regulatory networks have emerged as promising tools to dissect the genetics of complex traits<sup>21–23</sup>. Network-connectivity analyses in GWAS have shown that trait-associated genes are more interconnected than expected<sup>18</sup> and highly interconnected genes are enriched for trait heritability<sup>24</sup>. However, these analyses do not leverage observed enrichments to further enhance trait-associated gene discovery.

To unleash the potential of regulatory networks in GWAS, we develop a Bayesian framework for simultaneous genome-wide network enrichment and gene prioritization analysis. Through extensive simulations we show several advantages of the method such as flexibility in various genetic architectures, robustness to a wide range of model mis-specification and improved performance over existing methods. Applying the method to 18 human traits and 38 regulatory networks, we identify strong enrichments of genetic associations in network topology specific to trait-relevant cell types or tissues. By prioritizing variants within enriched networks we identify trait-associated genes that were not implicated by the same GWAS. Many of these previously undescribed genes have strong support from multiple lines of external evidence; some are further validated by follow-up GWAS of the same traits with increased sample sizes. Together, these results demonstrate the potential for our method to yield additional biological and therapeutic insights from existing data.

## Results

**Method overview.** Figure 1 shows the method schematic. In brief, we develop a model dissecting the total effect of a single SNP on a trait into effects of multiple (nearby and distal) genes through a regulatory network, and we combine it with a multiple-SNP regression likelihood<sup>25</sup> based on GWAS summary statistics to perform Bayesian inference.

Conceptually, we decompose the total effect of a common SNP on a complex trait into three components: a cis-regulatory component through nearby genes, a trans-regulatory component through distal genes that are regulated by genes near this SNP, and a remaining component due to other factors (Fig. 1a). Since common genetic variation contributes to complex traits primarily via gene regulation<sup>22</sup>, we find this decomposition a sensible approximation to the genetic basis of complex traits.

Despite various ways to model the regulatory components, here we use cell type- or tissue-specific regulatory networks<sup>18,20</sup> linking transcription factors (TFs) to target genes (TGs). Specifically, we define a regulatory network as a directed bipartite

graph with weighted edges from TFs to TGs (Fig. 1b). Given a TF-TG network, we use its topology to decompose the total effect of each SNP into effects of multiple interconnected genes. As shown in Fig. 1c, we approximate the effect of SNP  $j$  using a weighted sum of cis effects of three nearby genes (outside-network gene  $k$ , TG  $u$  and TF  $g$ ) and trans effects of three TGs ( $u$  and  $t$  on the same chromosome, and  $n$  on a different chromosome) that are directly regulated by TF  $g$  near SNP  $j$ . For identifiability we assume the SNP-gene ( $c_{jg}$ ) and TF-TG ( $v_{gt}$ ) weights in the decomposition are known, specified by existing omics data (Methods).

To implement this regulatory decomposition in GWAS, we formulate a network-induced prior for SNP-level effects ( $\beta$ ), and combine it with a regression likelihood<sup>25</sup> of  $\beta$  based on single-SNP association statistics from a GWAS (Fig. 1d) and linkage disequilibrium (LD) estimates from a reference panel with ancestry matching the GWAS (Fig. 1e). We refer to the resulting Bayesian framework (Fig. 1f) as Regression with Summary Statistics exploiting NETwork Topology (RSS-NET).

RSS-NET accomplishes two tasks simultaneously: (1) testing if a network is enriched for genetic associations (Fig. 1g); (2) identifying which genes within this network drive the enrichment (Fig. 1h). Specifically, RSS-NET estimates two independent enrichment parameters ( $\theta$  and  $\sigma^2$ ) that measure the extent to which, SNPs near network genes and regulatory elements (REs) have higher chances to be associated with the trait, and, SNPs near network edges have larger effect sizes, respectively. To assess network enrichment, RSS-NET computes a Bayes factor (BF) comparing the “enrichment model” ( $M_1: \theta > 0$  or  $\sigma^2 > 0$ ) against the “baseline model” ( $M_0: \theta = 0$  and  $\sigma^2 = 0$ ). To prioritize genes within enriched networks, RSS-NET contrasts posterior distributions of  $\beta$  estimated under  $M_0$  and  $M_1$ .

RSS-NET improves upon its predecessor RSS-E<sup>16</sup>. Specifically, RSS-NET exploits the full network topology, whereas RSS-E ignores the edge information. By explicitly modeling regulatory interconnections, RSS-NET outperforms RSS-E on both simulated and real data. Despite different treatments of network information, RSS-NET and RSS-E share computation schemes (Box 1; Supplementary Notes 1–3), allowing us to reuse the efficient algorithm of RSS-E. Software is available at <https://github.com/suwonglab/rss-net>.

### Box 1 | RSS-NET model fitting algorithm

**Input:** GWAS summary statistics ( $\hat{\beta}, \hat{S}$ ), LD estimates  $\hat{R}$ , network annotations  $\{\mathbf{a}, \mathbf{O}, \mathbf{W}\}$  and a grid of hyper-parameters  $\{\theta_0^{(h)}, \theta^{(h)}, \sigma_0^{(h)}, \sigma^{(h)}\}$ ,  $h = 1, \dots, H$ ; see Methods for details.

**Output:**  $\{\tau, \nu, \alpha\}$  such that  $\prod_{j=1}^p [\alpha_j \cdot \mathcal{N}(\beta_j; \nu_j, \tau_j^2) + (1 - \alpha_j) \cdot \delta_0(\beta_j)]$  is the closest mean-field approximation in Kullback-Leibler divergence to the exact conditional posterior of  $\beta$  given the hyper-parameters  $\{\theta_0, \theta, \sigma_0, \sigma\}$ .

**1. Initialize:** Set the initial values of  $\{\nu, \alpha\}$  randomly.

**2. Optimize:**

2a. Compute the prior parameters for each SNP  $j = 1, \dots, p$ :

$$\pi_j = 1 / [1 + 10^{-(\theta_0 + \alpha_j \theta)}], \quad \sigma_j^2 = \sigma_0^2 + \sigma^2 \cdot \sum_{g \in \mathbf{O}} w_{jg}^2$$

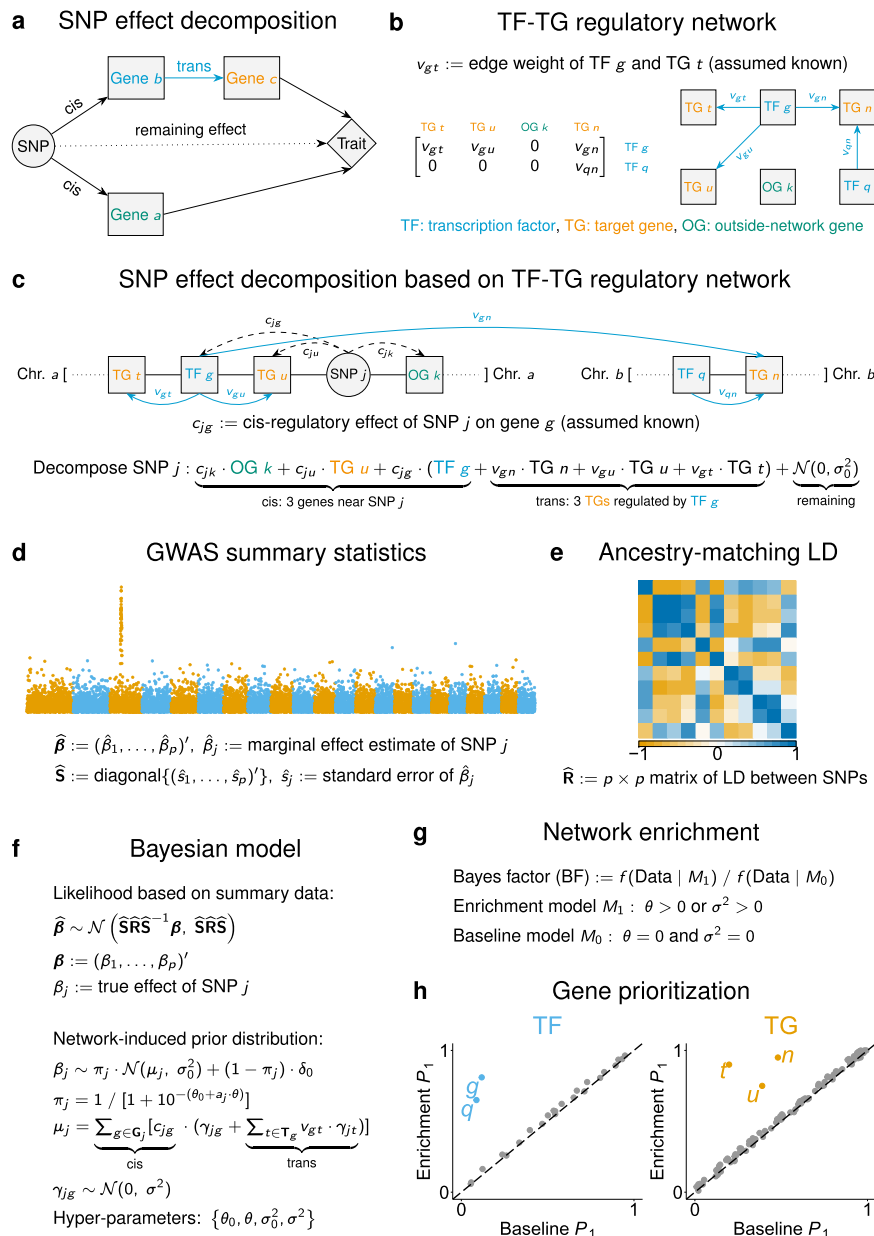
2b. Determine  $\tau$ .  $\tau_j = \hat{s}_j \sigma_j / \sqrt{\hat{s}_j^2 + \sigma_j^2}$  for SNP  $j = 1, \dots, p$ .

2c. Iterate through all SNPs to update  $\{\nu, \alpha\}$  as follows:

$$\nu_j = \tau_j^2 \cdot \left( \frac{\hat{\beta}_j - \sum_{i \neq j} \hat{r}_{ij} \alpha_i \nu_i}{\hat{s}_j^2 - \sum_{i \neq j} \hat{r}_{ij} \hat{s}_i \hat{s}_j} \right), \quad \frac{\alpha_j}{1 - \alpha_j} = \frac{\pi_j}{1 - \pi_j} \cdot \frac{\tau_j}{\sigma_j} \cdot \exp\left(\frac{\nu_j^2}{2\tau_j^2}\right)$$

2d. Repeat 2c until  $\{\nu, \alpha\}$  converge.

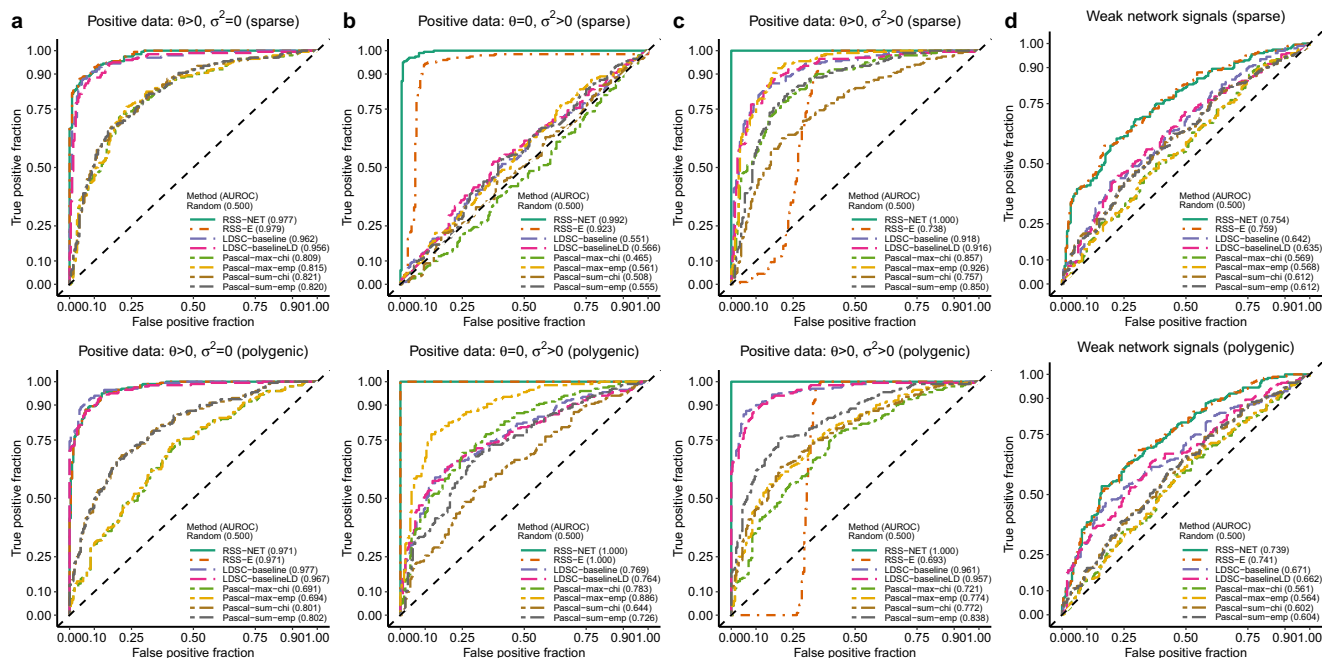
**3. Repeat:** Repeat 2 for each  $\{\theta_0^{(h)}, \theta^{(h)}, \sigma_0^{(h)}, \sigma^{(h)}\}$  in the grid to obtain the corresponding optimal  $\{\tau^{(h)}, \nu^{(h)}, \alpha^{(h)}\}$ ,  $h = 1, \dots, H$ .



**Fig. 1 Schematic of RSS-NET.** **a** Decomposition of the total effect of a common SNP on a complex trait through multiple nearby and distal genes. **b** Gene regulatory network defined as a weighted and directed bipartite graph linking TFs to TGs. **c** RSS-NET exploits the topology of a TF-TG network to decompose the total genetic effect into cis and trans-regulatory components. Both the SNP-gene ( $c_{jg}$ ) and TF-TG ( $v_{gt}$ ) weights in this decomposition are assumed known and are specified by existing omics data (Methods). In addition to TF-TG networks, RSS-NET also requires **d** GWAS summary statistics and **e** ancestry-matching LD estimates as input. **f** Bayesian hierarchical model underlying RSS-NET. An in-depth description is provided in Methods. **g** Given a network, RSS-NET produces a Bayes factor comparing the baseline ( $M_0$ ) and enrichment ( $M_1$ ) models to summarize the evidence for network enrichment. **h** RSS-NET prioritizes loci within an enriched network by computing  $P_1$ , the posterior probability that at least one SNP  $j$  in a locus is trait-associated ( $\beta_j \neq 0$ ). Differences between  $P_1$  under  $M_0$  and  $M_1$  reflect the influence of a regulatory network on genetic associations, highlighting previously undescribed trait-associated genes.

**Method comparison through simulations.** The key contribution of RSS-NET is a unified framework that leverages network topology to infer enrichments from whole-genome association statistics and prioritizes loci in light of inferred enrichments automatically. We are not aware of any published method with the same features. However, one could ignore topology and simply annotate SNPs based on their proximity to network genes and REs (Methods). For these SNP-level annotations there are methods to assess global enrichments or local associations on GWAS summary data. Here we use Pascal<sup>26</sup>, LDSC<sup>13,27</sup>, and RSS-E<sup>16</sup> to benchmark RSS-NET.

Given a network, we first simulated SNP effects ( $\beta$ ) from either RSS-NET or mis-specified models, and then combined them with real genotypes to simulate phenotypes from a genome-wide multiple-SNP model. We computed the single-SNP association statistics, on which we compared RSS-NET with other methods (Figs. 2–4; Supplementary Figs. 1–9). Since RSS-NET is model-based, we designed a large array of simulation scenarios for both correctly- and mis-specified  $\beta$ . To reduce computation of this large-scale design, we mainly used genotypes<sup>28</sup> of 348,965 genome-wide common SNPs and a whole-genome regulatory network inferred for human B cells (436 TFs, 3,018 TGs)<sup>20,29</sup>. We



**Fig. 2 Flexibility of RSS-NET to identify network-level enrichments from GWAS summary statistics.** We used a B cell-specific regulatory network and real genotypes of 348,965 genome-wide SNPs to simulate negative and positive individual-level data under two genetic architectures (“sparse” and “polygenic”). We simulated SNP effects ( $\beta$ ) for negative datasets from the baseline model ( $M_0$ :  $\theta = 0$  and  $\sigma^2 = 0$ ). We simulated  $\beta$  for positive datasets from the enrichment model ( $M_1$ :  $\theta > 0$  or  $\sigma^2 > 0$ ) for the target network under three scenarios: **a**  $\theta > 0$ ,  $\sigma^2 = 0$ ; **b**  $\theta = 0$ ,  $\sigma^2 > 0$ ; **c**  $\theta > 0$ ,  $\sigma^2 > 0$ . Using the simulated individual-level data we computed single-SNP association statistics, on which we compared RSS-NET with RSS-E<sup>16</sup>, LDSC-baseline<sup>13</sup>, LDSC-baselineLD<sup>27</sup>, and Pascal<sup>26</sup> using their default setups (Methods). Pascal includes two gene (“max”: maximum-of- $\chi^2$ ; “sum”: sum-of- $\chi^2$ ) and two pathway (“chi”:  $\chi^2$  approximation; “emp”: empirical sampling) scoring options. For each dataset, Pascal and LDSC methods produced  $P$ -values, whereas RSS-E and RSS-NET produced BFs; these statistics were used to rank the significance of enrichments. A false and true positive occurs if a method identifies enrichment of the target network in a negative and positive dataset respectively. Each panel displays the trade-off between false and true positives via receiver operating characteristics (ROC) curves for all methods in 200 negative and 200 positive datasets of a simulation scenario, and also reports the corresponding areas under ROC curves (AUROCs, higher value indicating better performance). Dashed diagonal lines denote random ROC curves (AUROC = 0.5). **d** RSS-NET, as well as other methods, does not perform well when the target network harbors weak genetic associations. Simulation details and additional results are provided in Supplementary Figs. 1, 2.

obtained similar results from simulations based on genotypes<sup>30</sup> of 1 million common SNPs<sup>31</sup> (Supplementary Fig. 9) or a different network (Supplementary Figs. 2 and 8).

We started with simulations where RSS-NET modeling assumptions were satisfied. We considered two genetic architectures: a sparse scenario with most SNPs being null and a polygenic scenario with most SNPs being trait-associated. For each architecture, we created negative datasets by simulating SNP effects ( $\beta$ ) from  $M_0$  and positive datasets by simulating  $\beta$  from three  $M_1$  patterns (only  $\theta > 0$ ; only  $\sigma^2 > 0$ ; both  $\theta > 0$  and  $\sigma^2 > 0$ ) of the target network, and applied the methods to detect  $M_1$  from all datasets (Fig. 2; Supplementary Figs. 1, 2). Existing methods tend to perform well in select settings. For example, Pascal and LDSC perform poorly when genetic signals are very sparse (Fig. 2b); RSS-E performs poorly when enrichment patterns are inconsistent with its modeling assumptions (Fig. 2c). Except for datasets with weak genetic signals on the network (Fig. 2d), RSS-NET performs consistently well in all scenarios. This is expected because the flexible model underlying RSS-NET can capture various genetic architectures and enrichment patterns. In practice, one rarely knows beforehand the correct architecture, which makes the flexibility of RSS-NET appealing.

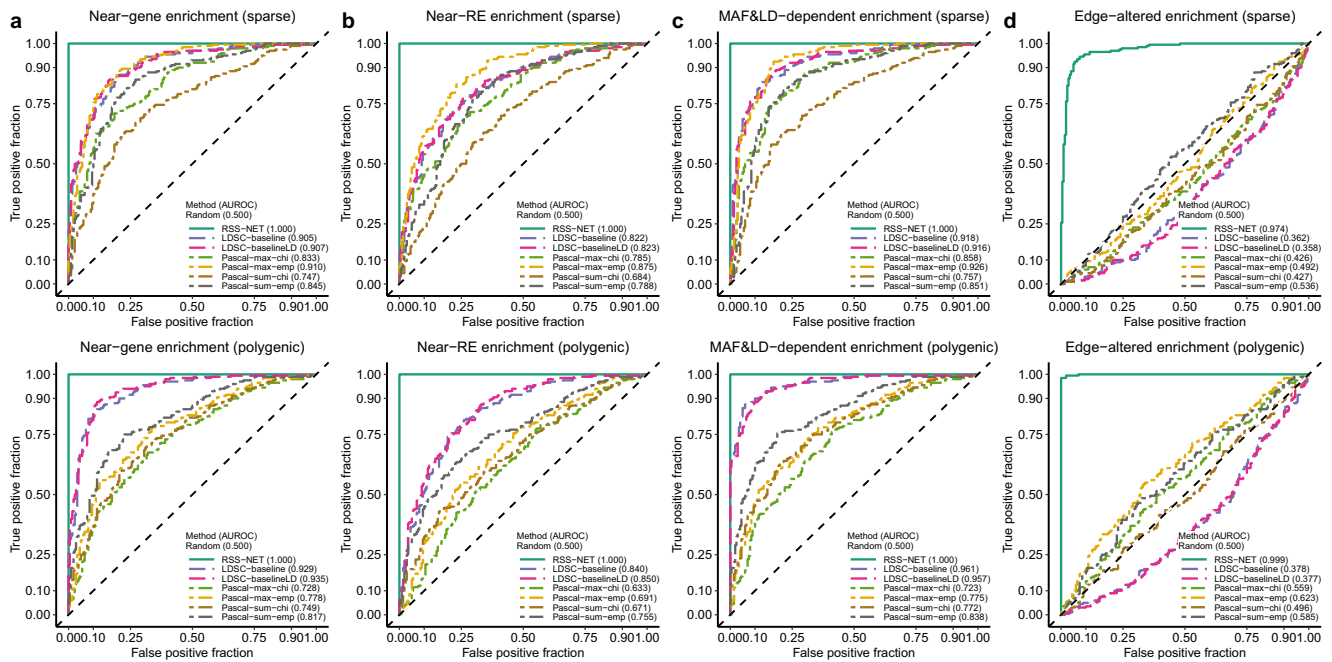
Genetic associations of complex traits are enriched in regulatory regions<sup>5,6</sup>. Since a regulatory network is a set of genes linked by REs, it is important to confirm that network enrichments identified by RSS-NET are not driven by general

regulatory enrichments. To this end, we simulated negative datasets with enriched associations in random SNPs that are near genes (Fig. 3a; Supplementary Fig. 3) or REs (Fig. 3b; Supplementary Fig. 4). The results show that RSS-NET is unlikely to yield false discoveries due to arbitrary enrichments in regulatory regions, and it is yet more powerful than other methods.

Minor allele frequency (MAF)- and LD-dependent genetic architectures are identified in complex traits<sup>27</sup>. To assess the impact of MAF- and LD-dependence on RSS-NET results, we simulated MAF- and LD-dependent SNP effects ( $\beta$ ) from an additive model of 10 MAF bins and 6 LD-related annotations<sup>27</sup>, which were then used to create negative datasets (Fig. 3c; Supplementary Fig. 5). Similarly, enrichments identified by RSS-NET are unlikely to be false positives induced by MAF- and LD-dependence.

Interconnections within regulatory programs play key roles in driving context specificity<sup>19</sup> and propagating disease risk<sup>22</sup>, but existing methods often ignore the edge information. In contrast, RSS-NET leverages the full topology of a given network. The topology-aware feature increases the potential of RSS-NET to identify the most relevant network for a trait among candidates that share many nodes but differ in edges. To illustrate this feature, we designed a scenario where a real target network and random candidates had the same nodes and edge counts, but different edges. We simulated positive and negative datasets where genetic associations were enriched in the target network





**Fig. 3 Robustness of RSS-NET to model mis-specification in enrichment analyses.** Here positive datasets were generated from  $M_1$  with  $\theta > 0$  and  $\sigma^2 > 0$  (Fig. 2c). Negative datasets were simulated from four scenarios where genetic associations were enriched in: **a** a random set of near-gene SNPs; **b** a random set of near-RE SNPs; **c** SNPs with MAF- and LD-dependent effects; **d** a random edge-altered network. By this design, RSS-NET was mis-specified in all four scenarios. Similar to positive datasets, the simulated false enrichments in all negative datasets manifested in both association proportion (more frequent) and magnitude (larger effect). RSS-E was excluded here because of its poor performance shown in Fig. 2c. The rest is the same as Fig. 2. Simulation details and additional results are provided in Supplementary Figs. 3–6.

and random candidates respectively, and then tested enrichment of the target network on all datasets. As expected, only RSS-NET can reliably distinguish true enrichments of the target network from enrichments of its edge-altered counterparts (Fig. 3d; Supplementary Fig. 6).

To benchmark its prioritization component, we compared RSS-NET with gene-based association modules in RSS-E<sup>16</sup> and Pascal<sup>26</sup> (Fig. 4; Supplementary Figs. 7–9). Consistent with previous work<sup>16</sup>, RSS methods outperform Pascal methods even without network enrichment (Fig. 4a). This is because RSS-NET and RSS-E exploit a multiple regression framework<sup>25</sup> to learn the genetic architecture from data of all genes and assess their effects jointly, whereas Pascal only uses data of a single gene to estimate its effect. Similar to enrichment simulations (Fig. 2), RSS-NET outperforms RSS-E in prioritizing genes across different architectures (Fig. 4b–d). This again highlights the flexibility of RSS-NET.

Finally, since RSS-NET uses network as is and most networks to date are algorithmically inferred, we performed simulations to assess the robustness of RSS-NET under noisy networks. Specifically, we simulated datasets from a real target network, created noisy networks by randomly removing edges from this real target, and then fed the noisy networks (rather than the real one) to RSS-NET. By exploiting retained true nodes and edges, RSS-NET produces reliable results in identifying both network enrichments and genetic associations, and unsurprisingly, its performance drops as the noise level increases (Supplementary Fig. 10).

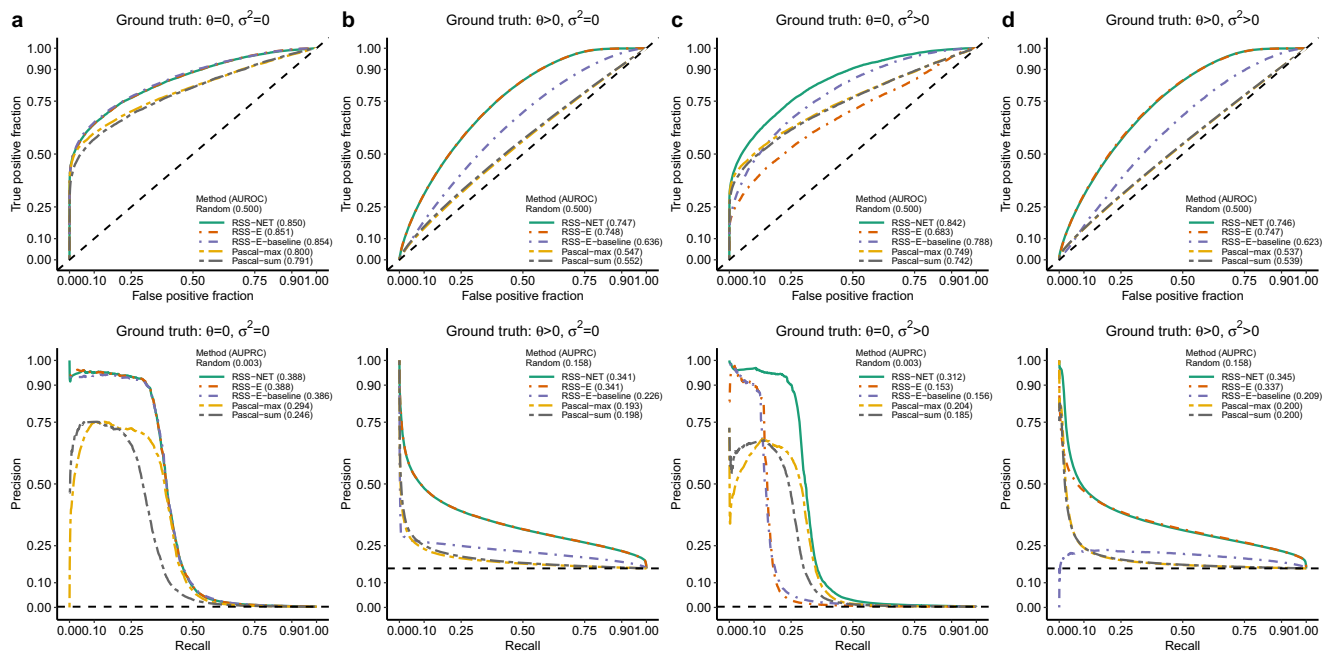
In conclusion, RSS-NET is adaptive to various genetic architectures and enrichment patterns, it is robust to a wide range of model mis-specification, and it outperforms existing related methods. To further investigate its real-world utility, we applied RSS-NET to analyze 18 complex traits and 38 regulatory networks.

**Enrichment analyses of 38 networks across 18 traits.** We first inferred<sup>20</sup> whole-genome regulatory networks for 38 human cell types and tissues (Methods; Supplementary Data 1) from public data<sup>29</sup> of paired expression and chromatin accessibility (PECA). On average each network has 431 TFs, 3,298 TGs, and 93,764 weighted TF-TG edges. Clustering showed that networks recapitulated context similarity, with immune cells and brain regions grouping together as two units (Fig. 5a; Supplementary Fig. 11).

As a validation, we assessed the pairwise similarity between the 38 PECA-based networks and 394 human cell type- and tissue-specific regulatory networks<sup>18</sup> reconstructed from independent cap analysis of gene expression (CAGE) data<sup>7,8</sup>. Reassuringly, PECA- and CAGE-based networks often reached maximum overlap when they were derived from biosamples of matched cell types or tissues (Fig. 5b; Supplementary Fig. 12), showing that the context specificity of regulatory networks is replicable.

On the 38 networks, we applied RSS-NET to analyze 1.1 million common SNPs<sup>31</sup> for 18 traits, using GWAS summary statistics from 20,883 to 253,288 European-ancestry individuals (Supplementary Table 1) and LD estimates<sup>16</sup> from the European panel of 1000 Genomes Project<sup>30</sup>. For each trait-network pair we computed a BF assessing network enrichment. Full results of 684 trait-network pairs are available online (Data availability).

To check whether observed enrichments could be driven by general regulatory enrichments, we created a “near-gene” control network with 18,334 protein-coding autosomal genes as nodes and no edges, and analyzed this control with RSS-NET on the same GWAS data. For most traits, the near-gene control has substantially weaker enrichment than the actual networks. In particular, 512 out of 684 trait-network pairs (one-sided binomial  $P = 2.2 \times 10^{-40}$ ) showed stronger enrichments than their near-gene counterparts (average  $\log_{10}$  BF increase: 13.94; one-sided  $t P = 5.1 \times 10^{-15}$ ), and, 16 out of 18 traits had multiple networks more enriched than the near-gene control (minimum: 5;



**Fig. 4** Power of RSS-NET to identify gene-level associations from GWAS summary statistics. We used a B cell-specific regulatory network and real genotypes of 348,965 genome-wide SNPs to simulate individual-level GWAS data under four scenarios: **a**  $\theta = 0, \sigma^2 = 0$ ; **b**  $\theta > 0, \sigma^2 = 0$ ; **c**  $\theta = 0, \sigma^2 > 0$ ; **d**  $\theta > 0, \sigma^2 > 0$ . Using the simulated individual-level data we computed single-SNP association statistics, on which we compared RSS-NET with gene-level association components of RSS-E<sup>16</sup> and Pascal<sup>26</sup>. RSS-E is a special case of RSS-NET assuming  $\sigma^2 = 0$ , and RSS-E-baseline is a special case of RSS-E assuming  $\theta = 0$ . Pascal includes two gene scoring options: maximum-of- $\chi^2$  (“max”) and sum-of- $\chi^2$  (“sum”). Given a network, Pascal and RSS-E-baseline do not leverage any network information, RSS-E ignores the edge information, and RSS-NET exploits the full topology. Each scenario contains 200 datasets and each dataset contains 16,954 autosomal protein-coding genes for testing. We defined a gene as “trait-associated” if at least one SNP *j* within 100 kb of the transcribed region of this gene had non-zero effect ( $\beta_j \neq 0$ ). For each gene in each dataset, RSS methods produced posterior probabilities that the gene was trait-associated ( $P_i$ ), whereas Pascal methods produced association *P*-values; these statistics were used to rank the significance of gene-level associations. The first row of each panel displays ROC curves and AUOCs for all methods, with dashed diagonal lines indicating random performance (AUOC = 0.5). The second row of each panel displays precision-recall (PRC) curves and areas under PRC curves (AUPRCs) for all methods, with dashed horizontal lines indicating random performance. For both AUOC and AUPRC, higher value indicates better performance. Simulation details and additional results are provided in Supplementary Figs. 7, 8.

one-sided Wilcoxon  $P = 1.2 \times 10^{-4}$ ). In contrast, LDSC and Pascal methods identified fewer trait-network pairs passing the near-gene enrichment control (LDSC maximum: 389, one-sided  $\chi^2 P = 1.7 \times 10^{-12}$ ; Pascal maximum: 69,  $P = 2.0 \times 10^{-129}$ ; Supplementary Table 2). Consistent with simulations (Fig. 3a, b), these results indicate that network enrichments identified by RSS-NET are unlikely driven by arbitrary enrichments harbored in the vicinity of genes.

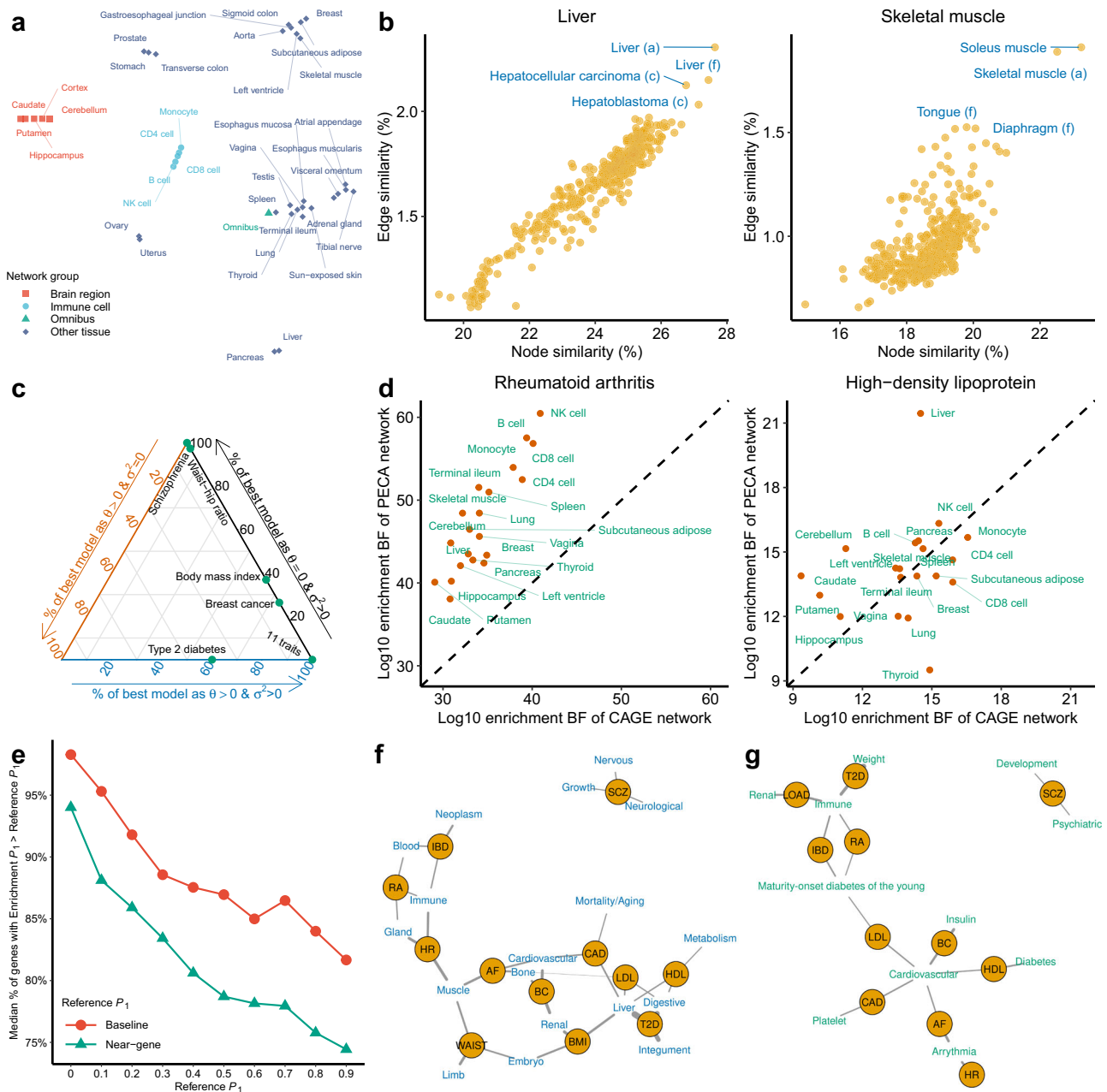
Among 512 trait-network pairs passing the near-gene enrichment control, we further examined whether the observed enrichments could be confounded by network properties or genomic annotations. We did not observe any correlation between BFs and three network features (proportion of SNPs in a network: Pearson  $R = -3.0 \times 10^{-2}$ , two-sided  $P = 0.49$ ; node counts:  $R = -5.4 \times 10^{-2}$ ,  $P = 0.23$ ; edge counts:  $R = -9.2 \times 10^{-3}$ ,  $P = 0.84$ ). To check confounding effects of genomic annotations, we computed the correlation between BFs and proportions of SNPs falling into both a network and each of 73 functional categories<sup>27</sup>, and we did not find any significant correlation ( $-0.13 < R < -0.01$ ,  $P > 0.05/73$ ). Similar patterns hold for all 684 trait-network pairs (Supplementary Table 3 and Data 2). Together, the results suggest that observed enrichments are unlikely driven by generic network or genome features.

For each trait-network pair, we also computed BFs comparing the baseline ( $M_0$ ) against three disjoint models where enrichment ( $M_1$ ) was contributed by (1) network genes and REs only ( $M_{11}$ :  $\theta > 0, \sigma^2 = 0$ ); (2) TF-TG edges only ( $M_{12}$ :  $\theta = 0, \sigma^2 > 0$ ); (3) network genes, REs and TF-TG edges ( $M_{13}$ :  $\theta > 0, \sigma^2 > 0$ ). We

found that  $M_{13}$  was the most supported model by data (with the largest BF) for 411 out of 512 trait-network pairs (one-sided binomial  $P = 1.2 \times 10^{-45}$ ), highlighting the key role of TF-TG edges in driving enrichments. To further confirm this finding, we repeated RSS-NET analyses by fixing all TF-TG edge weights as zero ( $v_{tg} = 0$ ) and we observed substantially weaker enrichments (average log<sub>10</sub> BF decrease: 30.46; one-sided  $t P = 8.6 \times 10^{-35}$ ; Supplementary Fig. 13). Altogether the results corroborate the “omnigenic model” that genetic signals of complex traits are distributed across the genome via regulatory interconnections<sup>22</sup>.

Enrichment patterns varied considerably among traits (Fig. 5c; Supplementary Table 4). For type 2 diabetes (T2D), two of five networks passing the near-gene enrichment control showed the strongest support for  $M_{11}$ . Many networks showed the strongest support for  $M_{12}$  in breast cancer (10), body mass index (BMI, 14), waist-hip ratio (37), and schizophrenia (38). Since one rarely knows the true enrichment patterns a priori, and  $M_1$  includes  $\{M_{11}, M_{12}, M_{13}\}$  as special cases, we used  $M_1$ -based BFs throughout this study. Collectively, these results highlight the heterogeneity of network enrichments across traits, which can be potentially learned from data by flexible approaches like RSS-NET.

Top-ranked enrichments recapitulated many trait-context links reported in previous GWAS. Genetic associations with BMI were enriched in the networks of pancreas (BF =  $2.07 \times 10^{13}$ ), bowel (BF =  $8.02 \times 10^{12}$ ), and adipose (BF =  $4.73 \times 10^{12}$ ), consistent with the roles of obesity-related genes in insulin biology and energy metabolism. Networks of immune cells showed



**Fig. 5** RSS-NET analyses of 18 complex traits and 38 regulatory networks. **a** Clustering of 38 regulatory networks based on *t*-distributed stochastic neighbour embedding. Details are provided in Supplementary Fig. 11. **b** Similarity between a given tissue-specific PECA-based network and 394 CAGE-based networks for various cell types and tissues (a: adult samples; c: cell lines; f: fetal samples). The similarity between a PECA- and CAGE-based network is summarized by Jaccard indices of their node sets (x-axis) and edge sets (y-axis). To simplify visualization, only labels of top four CAGE-based networks with the highest edge similarity are shown for each PECA-based network. See Supplementary Fig. 12 for additional results. **c** Ternary diagram showing, for each trait, percentages of the “best” enrichment model (with the largest BF) as  $M_{11}: \theta > 0, \sigma^2 = 0$ ,  $M_{12}: \theta = 0, \sigma^2 > 0$  and  $M_{13}: \theta > 0, \sigma^2 > 0$  across networks. See Supplementary Table 4 for numerical values. Shown are 16 traits having multiple networks more enriched than the near-gene control. **d** Comparison of context-matched PECA-based (y-axis) and CAGE-based (x-axis) network enrichments on the same GWAS. Dashed lines have slope 1 and intercept 0. See Supplementary Fig. 14 for additional results. **e** Median proportion of genes with  $p_1^{bma}$  higher than reference estimates ( $p_1^{base}$  or  $p_1^{near}$ ), among genes with reference estimates higher than a given cutoff. Medians are evaluated among 16 traits in **c**. See Supplementary Table 5 for numerical values. Overlap of RSS-NET prioritized genes ( $p_1^{bma} \geq 0.9$ ) with genes implicated in **f** knockout mouse phenotypes<sup>47</sup> and **g** human Mendelian diseases<sup>49,50</sup>. An edge indicates that a category of knockout mouse or Mendelian genes is significantly enriched for genes prioritized for a GWAS trait (FDR  $\leq 0.1$ ). Thicker edges correspond to stronger enrichments. To simplify visualization, only top-ranked categories are shown for each trait (**f**; **g**; 2). See Supplementary Data 4, 5 for full results. Trait abbreviations are defined in Supplementary Table 1.

enrichments for rheumatoid arthritis (RA,  $BF = 2.95 \times 10^{60}$ ), inflammatory bowel disease (IBD,  $BF = 5.07 \times 10^{35}$ ) and Alzheimer's disease ( $BF = 8.31 \times 10^{26}$ ). Networks of cardiac and other muscle tissues showed enrichments for coronary artery disease (CAD,  $BF = 9.78 \times 10^{28}$ ), atrial fibrillation (AF,  $BF = 8.55 \times 10^{14}$ ), and heart rate ( $BF = 2.43 \times 10^7$ ). Other examples are brain network with neuroticism ( $BF = 2.12 \times 10^{19}$ ), and, liver network with high- and low-density lipoprotein (HDL,  $BF = 2.81 \times 10^{21}$ ; LDL,  $BF = 7.66 \times 10^{27}$ ).

Some top-ranked enrichments were not identified in the original GWAS, but they are biologically relevant. For example, natural killer (NK) cell network showed the strongest enrichment among 38 networks for BMI ( $BF = 3.95 \times 10^{13}$ ), LDL ( $BF = 5.18 \times 10^{30}$ ), and T2D ( $BF = 1.49 \times 10^{77}$ ). This result supports a recent mouse study<sup>32</sup> revealing the role of NK cell in obesity-induced inflammation and insulin resistance, and adds to the considerable evidence unifying metabolism and immunity in many pathological states<sup>33</sup>. Other examples include adipose network with CAD<sup>34</sup> ( $BF = 1.67 \times 10^{29}$ ), liver network with Alzheimer's disease<sup>16,35</sup> ( $BF = 1.09 \times 10^{20}$ ) and monocyte network with AF<sup>36,37</sup> ( $BF = 4.84 \times 10^{12}$ ).

Some networks show enrichments in multiple traits. To assess network co-enrichments among traits, we tested correlations for all trait pairs using their BFs of 38 networks (Supplementary Data 3). In total 29 of 153 trait pairs had significant correlations (two-sided Pearson  $P < 0.05/153$ ). Reassuringly, subtypes of the same disease showed strongly correlated enrichments, as in IBD ( $R = 0.96$ ,  $P = 1.3 \times 10^{-20}$ ) and CAD subtypes ( $R = 0.90$ ,  $P = 3.3 \times 10^{-14}$ ). The results also recapitulated known genetic correlations including RA with IBD ( $R = 0.79$ ,  $P = 5.3 \times 10^{-9}$ ) and neuroticism with schizophrenia ( $R = 0.73$ ,  $P = 1.6 \times 10^{-7}$ ). Network enrichments of CAD were correlated with network enrichments of known CAD risk factors such as heart rate ( $R = 0.75$ ,  $P = 5.1 \times 10^{-8}$ ), BMI ( $R = 0.71$ ,  $P = 5.1 \times 10^{-7}$ ), AF ( $R = 0.65$ ,  $P = 9.2 \times 10^{-6}$ ) and height ( $R = 0.64$ ,  $P = 1.6 \times 10^{-5}$ ). Network enrichments of Alzheimer's disease were strongly correlated with network enrichments of LDL ( $R = 0.90$ ,  $P = 2.6 \times 10^{-14}$ ) and IBD ( $R = 0.78$ ,  $P = 8.3 \times 10^{-9}$ ), consistent with roles of lipid metabolism and inflammation in Alzheimer's disease<sup>35</sup>. Genetic correlations among traits are not predictive of correlations based on network enrichments (Pearson  $R = 0.12$ , two-sided  $P = 0.18$ ), suggesting the additional explanatory power from regulatory networks to reveal trait similarities in GWAS.

To show that RSS-NET can be applied more generally, we analyzed the CAGE-based networks<sup>18</sup> of 20 cell types and tissues that were present in 38 PECA-based networks (Fig. 5d; Supplementary Fig. 14). PECA-based networks often produced larger BFs than their CAGE-based counterparts on the same GWAS data (average  $\log_{10}$  BF increase: 17.36; one-sided  $t$   $P = 1.4 \times 10^{-11}$ ), suggesting that PECA-based networks are more enriched in genetic signals. Reassuringly, PECA- and CAGE-based networks consistently highlighted known trait-context links (e.g., immune cells and autoimmune diseases, muscle tissues and heart diseases). For some traits PECA-based networks produced more informative results. For example, CAGE-based analysis of HDL showed a broad enrichment pattern across cell types and tissues (which is consistent with previous connectivity analysis<sup>18</sup> of the same data), whereas PECA-based analysis identified liver as the top-enriched context by a wide margin. Although not our main focus, these results highlight the potential for RSS-NET to systematically evaluate different network inferences in GWAS.

**Enrichment-informed prioritization of network genes.** A key feature of RSS-NET is that inferred network enrichments automatically contribute to prioritization of network genes (Method).

Specifically, for each locus RSS-NET produces  $P_1^{\text{base}}$ ,  $P_1^{\text{near}}$  and  $P_1^{\text{net}}$ , the posterior probabilities that at least one SNP in the locus is associated with the trait, assuming  $M_0$ ,  $M_1$  for the near-gene control network and  $M_1$  for a given network, respectively. When multiple networks are enriched, RSS-NET produces  $P_1^{\text{bma}}$  by averaging  $P_1^{\text{net}}$  over all networks passing the near-gene control, weighted by their BFs. This allows us to assess genetic associations in light of enrichment without having to select a single enriched network. Differences between enrichment estimates ( $P_1^{\text{net}}$  or  $P_1^{\text{bma}}$ ) and reference estimates ( $P_1^{\text{base}}$  or  $P_1^{\text{near}}$ ) reflect the impact of network on a locus.

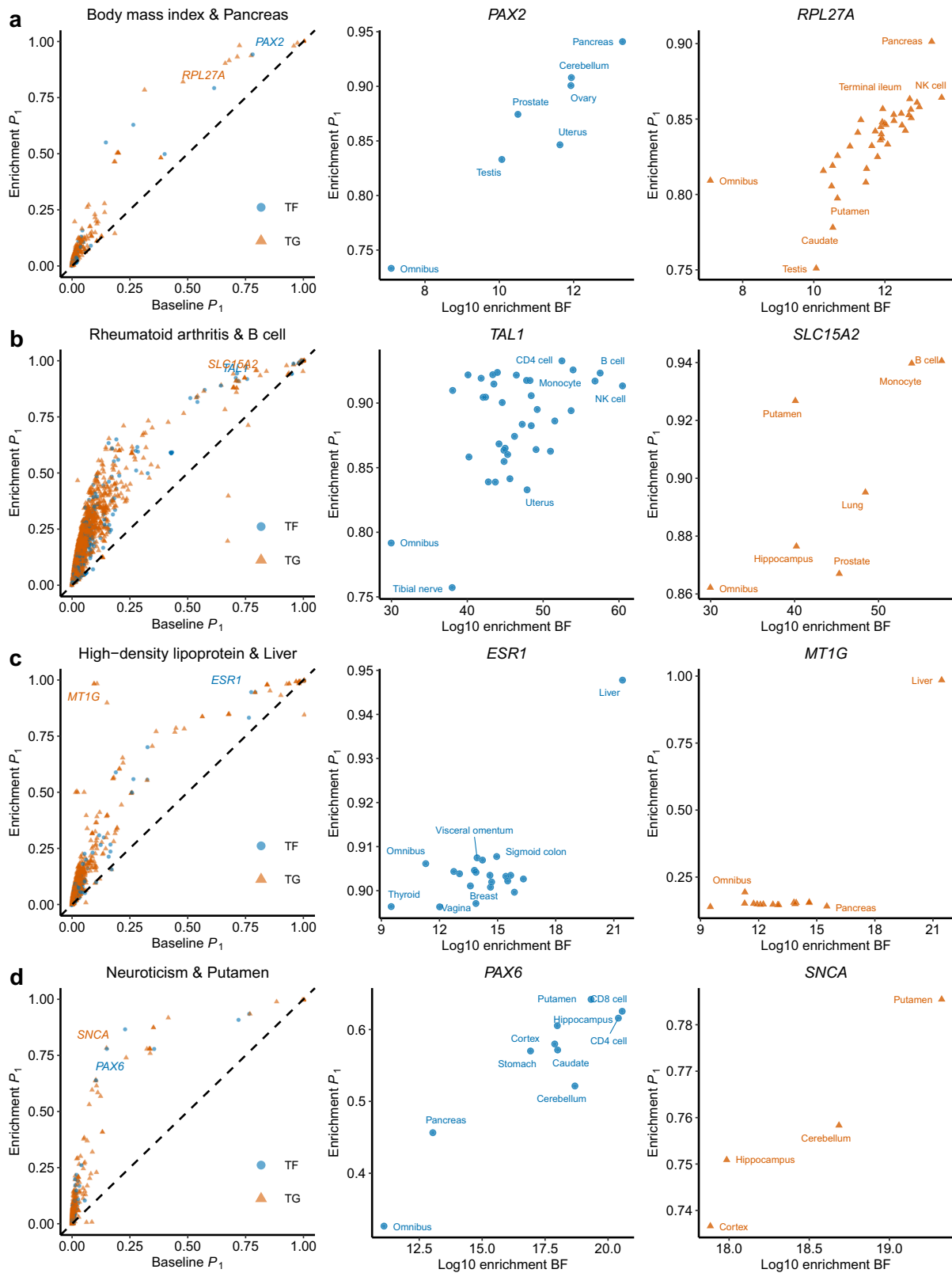
RSS-NET enhances genetic association detection by leveraging inferred enrichments. To quantify this improvement, for each trait we calculated the proportion of genes with higher  $P_1^{\text{bma}}$  than reference estimates ( $P_1^{\text{base}}$  or  $P_1^{\text{near}}$ ), among genes with reference  $P_1$  passing a given cutoff (Fig. 5e). When using  $P_1^{\text{base}}$  as reference, we observed high proportions of genes with  $P_1^{\text{bma}} > P_1^{\text{base}}$  (median: 82–98%) across a wide range of  $P_1^{\text{base}}$ -cutoffs (0–0.9), and as expected, the improvement decreased as the reference cutoff increased. When using  $P_1^{\text{near}}$  as reference, we observed less genes with improved  $P_1^{\text{bma}}$  than using  $P_1^{\text{base}}$  (one-sided Wilcoxon  $P = 9.8 \times 10^{-4}$ ), suggesting the observed improvement might be partially due to general near-gene enrichments, but proportions of genes with  $P_1^{\text{bma}} > P_1^{\text{near}}$  remained high (median: 74–94%) nonetheless. Similar patterns occurred when we repeated the analysis with  $P_1^{\text{net}}$  across 512 trait-network pairs (Supplementary Table 5). Together the results demonstrate the strong influence of network enrichments on nominating additional trait-associated genes.

RSS-NET tends to promote more genes in networks with stronger enrichments. For each trait, the proportion of genes with  $P_1^{\text{net}} > P_1^{\text{near}}$  in a network is often positively correlated with the network enrichment BF ( $R: 0.28$ – $0.91$ ; Supplementary Table 6). When a gene belongs to multiple networks, the highest  $P_1^{\text{net}}$  often occurs in the top-enriched networks (Fig. 6). We illustrate this coherent pattern with *MTIG*, a liver-active<sup>9</sup> gene prioritized for HDL by RSS-NET and also implicated in a recent multi-ancestry genome-wide interaction analysis of HDL<sup>38</sup>. Although *MTIG* belongs to regulatory networks of 18 contexts, only the top enrichment in liver informs a strong association between *MTIG* and HDL ( $P_1^{\text{net}} = 0.98$ ), and remaining networks with weaker enrichments yield minimal improvement ( $P_1^{\text{base}} = 0.10$ ,  $P_1^{\text{net}} : 0.14$ – $0.19$ ).

RSS-NET recapitulates many genes implicated in the same GWAS. For each analyzed dataset we downloaded the GWAS-implicated genes from the GWAS Catalog<sup>1</sup> and computed the proportion of these genes with high  $P_1^{\text{bma}}$ . With a stringent cutoff  $P_1^{\text{bma}} \geq 0.9$ , we observed a significant overlap (median across traits: 69%; median two-sided Fisher exact  $P = 1.2 \times 10^{-26}$ ; Supplementary Table 7). Reassuringly, many recapitulated genes are well-established for the traits (Supplementary Table 8), such as *CACNA1C* for schizophrenia, *TCF7L2* for T2D, *APOB* for lipids, and *STAT4* for autoimmune diseases.

RSS-NET also uncovers putative associations that were not reported in the same GWAS. To demonstrate that many of these previously undescribed associations are potentially real, we exploited 15 analyzed traits that each had a updated GWAS with larger sample size. In each case, we obtained newly implicated genes from the GWAS Catalog<sup>1</sup> and computed the proportion of these genes that were identified by RSS-NET ( $P_1^{\text{bma}} \geq 0.9$ ). The overlap proportions remained significant (median: 12%; median two-sided Fisher exact  $P = 1.9 \times 10^{-5}$ ; Supplementary Table 7), showing the potential of RSS-NET to identify trait-associated





**Fig. 6** RSS-NET gene prioritization results of select trait-network pairs. Shown are four trait-network pairs: **a** body mass index and pancreas; **b** rheumatoid arthritis and B cell; **c** high-density lipoprotein cholesterol and liver; **d** neuroticism and putamen. In the first column of each panel, each point represents a member gene of a given network (blue circle: TF; orange triangle: TG). Dashed lines have slope 1 and intercept 0. In the second and third columns, each point represents a cell type- or tissue-specific network to which a select gene belongs. Numerical values of  $P_1$  and BF are available online (Data availability) and are provided as a Source Data file.

**Table 1** Examples of RSS-NET highlighted genes that were not reported in GWAS of the same data but were implicated in later GWAS with increased sample sizes (genome-wide significance threshold: single-SNP association  $P < 5 \times 10^{-8}$ ).

Trait	Gene (Role)	$P_1^{\text{base}}$	$P_1^{\text{near}}$	$P_1^{\text{bma}}$	$P_1^{\text{net}}$ (Network, BF)	Mouse trait	Therapeutic and clinical evidence
BMI	<i>PAX2</i> (TF)	0.78	0.80	0.94	0.94 (Pancreas, $2.07 \times 10^{13}$ )	Eye, Renal	Ocular and renal anomalies
	<i>FLT3</i> (TG)	0.61	0.70	0.85	0.85 (Cerebellum, $8.70 \times 10^{11}$ )	Growth, Immune	Acute myeloid leukemia
WAIST	<i>LAMB1</i> (TG)	0.97	0.97	0.98	0.98 (Esophagus, $6.78 \times 10^{239}$ )	Neuron, NS	Lissencephaly-5
BC	<i>KCTD1</i> (TG)	0.89	0.93	0.98	0.98 (Heart, $8.08 \times 10^7$ )	CS	Scalp-ear-nipple syndrome
	<i>CASP8</i> (TG)	0.71	0.72	0.94	0.94 (Aorta, $8.27 \times 10^8$ )	Growth, Immune	Hepatoma, Glionitrin A <sup>+</sup>
RA	<i>AIRE</i> (TF)	0.54	0.61	0.84	0.84 (B cell, $3.31 \times 10^{57}$ )	Immune	APS1
IBD	<i>LPP</i> (TG)	0.98	0.94	0.99	0.99 (Monocyte, $6.28 \times 10^{31}$ )	Cellular	Acute myeloid leukemia
	<i>FOXP1</i> (TF)	0.84	0.78	0.95	0.95 (NK cell, $5.07 \times 10^{35}$ )	Immune, Neuron	Language impairment
HDL	<i>CCND3</i> (TG)	0.81	0.89	0.95	0.95 (NK cell, $5.07 \times 10^{35}$ )	Immune	
	<i>ALOX5</i> (TG)	0.97	0.97	0.99	0.99 (Monocyte, $4.75 \times 10^{15}$ )	Immune, Metab.	Atherosclerosis
	<i>GPAM</i> (TG)	0.92	0.95	0.98	0.98 (Liver, $2.81 \times 10^{21}$ )	Liver, Metab.	
LDL	<i>NROB2</i> (TG)	0.84	0.93	0.98	0.98 (Liver, $2.81 \times 10^{21}$ )	Growth, Metab.	Early-onset obesity
	<i>CERS2</i> (TG)	0.99	0.99	1.00	1.00 (NK cell, $5.18 \times 10^{30}$ )	Liver, Metab.	
CAD	<i>ABCA1</i> (TG)	0.98	0.98	0.99	0.99 (Liver, $7.66 \times 10^{27}$ )	Liver, Metab.	Tangier disease, Probuco <sup>l</sup>
	<i>ABCB11</i> (TG)	0.68	0.72	0.88	0.88 (Liver, $7.66 \times 10^{27}$ )	Liver, Metab.	Cholestasis
	<i>DLG4</i> (TG)	0.69	0.59	0.85	0.85 (NK cell, $5.18 \times 10^{30}$ )	Metab., NS	Tat-NR2B9c <sup>+</sup>
	<i>SOX17</i> (TF)	0.52	0.65	0.82	0.84 (CD8, $5.86 \times 10^{28}$ )	Liver, Metab.	Vesicoureteral reflux-3
	<i>TGFB1</i> (TG)	0.92	0.99	0.99	0.99 (Adipose, $1.67 \times 10^{29}$ )	CS, Growth	Camurati-Engelmann disease
	<i>FN1</i> (TG)	0.58	0.79	0.91	0.92 (GEJ, $9.78 \times 10^{28}$ )	CS, Metab.	GFND2, SMDCF
	<i>CDH13</i> (TG)	0.31	0.55	0.77	0.82 (Heart, $1.93 \times 10^{28}$ )	CS, Metab.	
AF	<i>EDNRA</i> (TG)	0.57	0.79	0.80	0.82 (Aorta, $1.09 \times 10^{27}$ )	CS, Muscle	Ambrisentan <sup>+</sup> , Macitentan <sup>+</sup>
	<i>SCNSA</i> (TG)	0.87	0.92	1.00	1.00 (Heart, $6.89 \times 10^{12}$ )	CS, Muscle	Brugada syndrome-1, ATRF10
	<i>ENPEP</i> (TG)	0.50	0.76	0.92	0.94 (Uterus, $2.71 \times 10^{11}$ )		QGC-001 <sup>+</sup>
SCZ	<i>ATXN1</i> (TG)	0.45	0.62	0.90	0.90 (Colon, $7.54 \times 10^{14}$ )	Muscle, NS	Spinocerebellar ataxia-1
	<i>MYOT</i> (TG)	0.55	0.66	0.86	0.87 (Muscle, $8.55 \times 10^{14}$ )		Myofibrillar myopathy
	<i>FOXP1</i> (TF)	1.00	1.00	1.00	1.00 (Colon, $1.20 \times 10^{144}$ )	Growth, Neuron	Language impairment
NEU	<i>BCL11A</i> (TG)	1.00	1.00	1.00	1.00 (Spleen, $1.44 \times 10^{141}$ )	Immune, NS	Dias-Logan syndrome
	<i>SLC25A12</i> (TG)	0.79	0.81	0.88	0.88 (Muscle, $4.99 \times 10^{127}$ )	Neuron, NS	DEE39
	<i>TCF4</i> (TF)	0.72	0.88	0.95	0.95 (CD8, $3.66 \times 10^{20}$ )	Immune, NS	Pitt-Hopkins syndrome
	<i>RAPSN</i> (TG)	0.77	0.88	0.93	0.93 (Muscle, $8.20 \times 10^{17}$ )	Muscle, NS	Congenital myasthenic syndrome-11
PCL0 (TG)	<i>MEF2C</i> (TF)	0.15	0.40	0.83	0.83 (Ileum, $8.56 \times 10^{22}$ )	Growth, Neuron	Mental retardation-20
	<i>SNCA</i> (TG)	0.15	0.32	0.78	0.79 (Putamen, $2.12 \times 10^{19}$ )	Neuron, NS	Parkinsonism, BliB054 <sup>+</sup>
	<i>PAX6</i> (TF)	0.10	0.22	0.62	0.64 (Putamen, $2.12 \times 10^{19}$ )	NS, Vision	Optic nerve hypoplasia
	<i>PCL0</i> (TG)	0.06	0.17	0.63	0.63 (Ileum, $8.56 \times 10^{22}$ )	Growth, NS	Pontocerebellar hypoplasia-3

The "mouse trait" column is based on the Mouse Genome Informatics<sup>47</sup>. The "therapeutic/clinical evidence" column is based on the Online Mendelian Inheritance in Man<sup>50</sup> and Therapeutic Target Database<sup>53</sup>. Drugs are identified with an asterisk ("\*"). Trait abbreviations are defined in Supplementary Table 1. GEJ: gastroesophageal junction. CS: cardiovascular system. DS: digestive/alimentary system. Metab.: metabolism. NS: nervous system. APS1: autoimmune polyendocrinopathy syndrome-1. GFND2: glomerulopathy with fibronectin deposits-2. SMDCF: corner fracture type of spondylometaphyseal dysplasia. ATRF10: familial atrial fibrillation-10. DEE39: developmental and epileptic encephalopathy-39.

genes that can be validated by later GWAS with additional samples. Among these validated genes, many are strongly supported by multiple lines of external evidence (Table 1). A particular example is *NROB2*, a liver-active<sup>9</sup> gene prioritized for HDL by RSS-NET ( $P_1^{\text{base}} = 0.84$ ,  $P_1^{\text{net}} = 0.98$ ) but not identified by standard GWAS<sup>39</sup> of the same data (minimum single-SNP association  $P = 1.4 \times 10^{-7}$  within 100 kb,  $n = 99,900$ ). *NROB2* was associated with mouse lipid traits<sup>40–42</sup> and human obesity<sup>43</sup>, and identified in a later GWAS of HDL<sup>44</sup> with doubled sample size ( $P = 9.7 \times 10^{-16}$ ,  $n = 187,056$ ).

**Biological and clinical relevance of prioritized genes.** Besides looking up overlaps with GWAS publications, we cross-referenced RSS-NET prioritized genes ( $P_1^{\text{bma}} \geq 0.9$ ) with multiple orthogonal databases to systematically assess their biological and therapeutic themes.

Mouse phenomics provides important resources to study genetics of human traits<sup>45</sup>. Here we evaluated overlap between RSS-NET prioritized genes and genes implicated in 27 categories of knockout mouse phenotypes<sup>46</sup>. Network-informed genes ( $P_1^{\text{bma}} \geq 0.9$ ) were significantly enriched in 128 mouse-human trait pairs (FDR  $\leq 0.1$ ; Supplementary Data 4). Fewer significant pairs were identified without network information (119 for  $P_1^{\text{near}} \geq 0.9$ ; 80 for  $P_1^{\text{base}} \geq 0.9$ ). For many human traits, top

enrichments of network-prioritized genes occurred in closely related mouse phenotypes (Fig. 5f). Genes prioritized for schizophrenia were strongly enriched in nervous, neurological and growth phenotypes (OR: 1.77–2.04). Genes prioritized for autoimmune diseases were strongly enriched in immune and hematopoietic phenotypes (OR: 2.05–2.35). The cardiovascular system showed strong enrichments of genes prioritized for heart conditions (OR: 2.45–2.92). The biliary system showed strong enrichments of genes prioritized for lipids, BMI, CAD, and T2D (OR: 2.16–10.78). The phenotypically matched cross-species enrichments strengthen the biological relevance of RSS-NET results.

Genes causing Mendelian diseases often contribute to complex traits<sup>47</sup>. Here we quantified overlap between RSS-NET prioritized genes and genes causing 19 categories<sup>48</sup> of Mendelian disorders<sup>49</sup>. Leveraging regulatory networks ( $P_1^{\text{bma}} \geq 0.9$ ), we observed 47 significantly enriched Mendelian-complex trait pairs (FDR  $\leq 0.1$ ; 44 for  $P_1^{\text{near}} \geq 0.9$ ; 31 for  $P_1^{\text{base}} \geq 0.9$ ; Supplementary Data 5), among which the top-ranked ones were often phenotypically matched (Fig. 5g). Genes prioritized for schizophrenia were strongly enriched in Mendelian development and psychiatric disorders (OR: 2.22–2.23). Genes prioritized for AF and heart rate were strongly enriched in arrhythmia (OR: 7.16–8.28). Genes prioritized for autoimmune diseases were strongly enriched in

monogenic immune dysregulation (OR: 3.11–4.32). Monogenic cardiovascular diseases showed strong enrichments of genes prioritized for lipids and heart conditions (OR: 2.69–3.70). We also identified pairs where Mendelian and complex traits seemed unrelated but were indeed linked. Examples include Alzheimer's disease with immune dysregulation<sup>35</sup> (OR = 7.32) and breast cancer with insulin disorders<sup>50</sup> (OR = 9.71). The results corroborate the continuum between Mendelian and complex traits.

Human genetics has proven valuable in therapeutic development<sup>51</sup>. To evaluate their potential in drug discovery, we examined whether RSS-NET prioritized genes are pharmacologically active and clinically relevant<sup>52</sup>. We identified genes with drug indications matching GWAS traits. One identical match is *EDNRA*, a gene that is prioritized for CAD ( $P_1^{\text{base}} = 0.57$ ,  $P_1^{\text{net}} = 0.82$  in aorta) and also a successful target of approved drugs for cardiovascular diseases (Table 1). We identified genes with drug indications closely related to GWAS traits. For example, *TTR* is prioritized for Alzheimer ( $P_1^{\text{base}} = 0.64$ ,  $P_1^{\text{bma}} = 0.94$ ) and also a successful target of approved drugs for amyloidosis (Table 2). For early-stage development, overlaps between drug indications and GWAS traits may provide additional genetic confidence. For example, *HCAR3* is prioritized for HDL ( $P_1^{\text{base}} = 0.85$ ,  $P_1^{\text{bma}} = 0.92$ ) and also a clinical trial target for lipid metabolism disorders (Table 2). Other examples include *CASP8* with cancer, *NFKB2* with IBD, and *DLG4* with stroke (Tables 1, 2). For some genes we found mismatches between drug indications and GWAS traits, which could suggest drug repurposing opportunities<sup>53</sup>. For example, *CSF3* is prioritized for AF ( $P_1^{\text{base}} = 0.56$ ,  $P_1^{\text{bma}} = 0.88$ ) and also a successful target of an approved drug for aplastic anemia (AA). Since *CSF3* is associated with various blood cell traits in mouse<sup>54</sup> and human<sup>55</sup>, and inflammation plays a role in both AA and AF etiology<sup>36,37,56</sup>, it is tempting to assess effects of the approved AA drug on AF. Mechanistic evaluations are required to understand the prioritized therapeutic genes, but they could form a useful basis for future studies.

## Discussion

We present RSS-NET, a topology-aware method for integrative analysis of regulatory networks and GWAS summary data. We demonstrate the improvement of RSS-NET over existing methods through extensive simulations, and illustrate its potential to yield biological and therapeutic insights via analyses of 38 networks and 18 traits. With multi-omics integration becoming a routine in GWAS, we expect that researchers will find RSS-NET useful.

Compared with existing integrative approaches, RSS-NET has several key strengths. First, unlike many methods that require loci passing a significance threshold<sup>11,12,17</sup>, RSS-NET uses data from genome-wide common variants. This potentially allows RSS-NET to identify subtle enrichments even in studies with few significant hits. Second, RSS-NET models enrichments directly as increased rates ( $\theta$ ) and sizes ( $\sigma^2$ ) of SNP-level associations, and thus bypasses the issue of converting SNP-level summary data to gene-level statistics<sup>17,18,26</sup>. Third, RSS-NET inherits from RSS-E<sup>16</sup> an important feature that inferred enrichments automatically highlight which network genes are most likely to be trait-associated. This prioritization component, though useful, is missing in current polygenic analyses<sup>13,15,24,27</sup>. Fourth, by making flexible modeling assumptions, RSS-NET is adaptive to unknown genetic architectures.

RSS-NET allows us to study complex trait genetics through the lens of regulatory topology. Complementing previous connectivity analyses<sup>17–19,24</sup>, RSS-NET highlights a consistent pattern that genetic signals of complex traits often distribute across genome via regulatory topology. RSS-NET further leverages topology enrichments to enhance trait-associated gene discovery.

The topology awareness of RSS-NET in both enrichment and prioritization analyses is enabled by a model that decomposes the effect of a single SNP into effects of multiple (cis or trans) genes through a regulatory network.

RSS-NET depends critically on the quality of input networks. The more accurate networks are, the better performance RSS-NET achieves. Currently, our understanding of regulatory networks remains incomplete, and most of the available networks are algorithmically inferred<sup>17–20</sup>. Artifacts in inferred networks can bias RSS-NET results; however, our simulations confirm the robustness of RSS-NET when input networks are not severely deviated from ground truth. The modular design of RSS-NET enables systematic assessment of various networks in the same GWAS and provides interpretable performance metrics, as illustrated in our comparison of PECA- and CAGE-based networks. As more accurate networks become available in diverse cellular contexts, the performance of RSS-NET will be markedly enhanced.

Like any method, RSS-NET has several limitations in its current form. First, despite its prioritization feature, RSS-NET does not attempt to pinpoint associations to causal SNPs within prioritized loci. For this task, we recommend off-the-shelf fine-mapping methods<sup>57</sup>. Second, the computation time of RSS-NET increases as the total number of analyzed SNPs increases, and thus our simulations and analyses focused on 0.35–1.19 million genome-wide common SNPs<sup>28,31</sup>. Relaxing the complexity will allow RSS-NET to analyze more SNPs jointly. Third, RSS-NET uses a simple method to derive SNP-gene relevance ( $c_{jg}$ ) from expression quantitative trait loci (eQTL). A more principled approach would be applying the RSS likelihood<sup>25</sup> to eQTL summary data (as we did in GWAS) and using the estimated SNP effects to specify  $c_{jg}$ . However, our initial assessments indicated that the model-based approach was limited by the small sample sizes of current eQTL studies<sup>9,10</sup>. With eQTL studies reaching large sample sizes<sup>58</sup> comparable to current GWAS<sup>1</sup>, this approach may improve  $c_{jg}$  specification in RSS-NET. Fourth, RSS-NET analyzes one network at a time. Since a complex disease typically manifests in various sites, multiple cellular networks are likely to mediate disease risk jointly. To extend RSS-NET to incorporate multiple networks, an intuitive idea would be representing the total effect of a SNP as an average of its effect in each network, weighted by network relevance for a disease. Fifth, RSS-NET does not include known SNP-level<sup>13,24,27</sup> or gene-level<sup>14–16</sup> annotations. Although our mis-specification simulations and near-gene control analyses confirm that RSS-NET is robust to generic enrichments of known features, accounting for known annotations can help interpret observed network enrichments<sup>24</sup>. Our preliminary experiments showed that incorporating additional networks or annotations in RSS-NET increased computation costs. Hence, we view developing computationally efficient multi-network, multi-annotation methods as an important area for future work.

In summary, improved understanding of complex trait genetics requires biologically informed models beyond the standard one employed in GWAS. By modeling context-specific regulatory topology, RSS-NET is a step forward in this direction.

## Methods

**Gene and SNP information.** This study used genes and SNPs from the human genome assembly GRCh37. This study used 18,334 protein-coding autosomal genes ([http://ftp.ensembl.org/pub/grch37/release-94/gtf/homo\\_sapiens](http://ftp.ensembl.org/pub/grch37/release-94/gtf/homo_sapiens), accessed January 3, 2019). Simulations used 348,965 genome-wide SNPs<sup>28</sup> (<https://www.wtccc.org.uk>), and data analyses used 1,289,786 genome-wide HapMap3<sup>31</sup> SNPs ([https://data.broadinstitute.org/alkesgroup/LDSCORE/w\\_hm3.snplist.bz2](https://data.broadinstitute.org/alkesgroup/LDSCORE/w_hm3.snplist.bz2), accessed November 27, 2018). As discussed later, these SNP sets were chosen to reduce computation. This study excluded SNPs on sex chromosomes, SNPs with MAF less than 1%, and SNPs in the human leukocyte antigen region.

**Table 2** Examples of RSS-NET highlighted genes that have not reached genome-wide significance in the GWAS Catalog<sup>1</sup> at the time of analysis.

Trait	Gene (Role)	$P_1^{\text{base}}$	$P_1^{\text{near}}$	$P_1^{\text{bma}}$	$P_1^{\text{net}}$ (Network, BF)	Mouse trait	Therapeutic and clinical evidence
BMI	NEXN (TG)	0.71	0.79	0.89	0.90 (Muscle, $9.31 \times 10^{12}$ )	CS, Muscle	Cardiomyopathy
	CDX2 (TF)	0.61	0.70	0.83	0.86 (NK cell, $3.95 \times 10^{13}$ )	DS, Growth	
WAIST	BSCL2 (TG)	0.80	0.68	0.87	0.87 (Esophagus, $6.78 \times 10^{239}$ )	Adipose, Growth	Berardinelli-Seip syndrome
	FOXP2 (TF)	0.56	0.59	0.73	0.73 (Esophagus, $6.78 \times 10^{239}$ )	Growth, NS	Speech-language disorder-1
BC	ADSL (TG)	0.76	0.80	0.91	0.92 (Aorta, $8.27 \times 10^8$ )	CS, Eye	Adenylosuccinase deficiency
	SYNE1 (TG)	0.57	0.63	0.89	0.90 (Esophagus, $6.30 \times 10^7$ )	Growth, Muscle	AMC3, EDMD4, SCAR8
RA	TAL1 (TF)	0.71	0.79	0.91	0.93 (CD4, $3.02 \times 10^{52}$ )	Immune, Tumor	Acute lymphocytic leukemia
	FHIT (TG)	0.30	0.60	0.90	0.91 (CD4, $3.02 \times 10^{52}$ )	Immune, Tumor	
IBD	FLT3 (TG)	0.33	0.57	0.73	0.73 (B cell, $3.31 \times 10^{57}$ )	Immune, Tumor	Acute myeloid leukemia
	FHIT (TG)	0.63	0.87	0.95	0.95 (CD4, $5.32 \times 10^{33}$ )	Immune, Tumor	
HDL	GATA3 (TF)	0.85	0.83	0.94	0.94 (NK cell, $5.07 \times 10^{35}$ )	Immune, Renal	Barakat syndrome
	RORA (TF)	0.66	0.78	0.87	0.90 (B cell, $1.49 \times 10^{32}$ )	Immune, NS	Intellectual disability
	NFKB2 (TF)	0.74	0.85	0.84	0.88 (B cell, $1.49 \times 10^{32}$ )	Immune	Immunodeficiency, DIMS-0150*
	LRBA (TG)	0.42	0.58	0.72	0.72 (NK cell, $5.07 \times 10^{35}$ )	Immune	Immunodeficiency
	DOCK2 (TG)	0.38	0.53	0.71	0.71 (NK cell, $5.07 \times 10^{35}$ )	Immune	Immunodeficiency
LDL	MT1G (TG)	0.10	0.09	0.98	0.98 (Liver, $2.81 \times 10^{21}$ )	CS, Metab.	
	RETSAT (TG)	0.79	0.80	0.95	0.95 (Liver, $2.81 \times 10^{21}$ )	Adipose, Metab.	
	ESR1 (TF)	0.77	0.82	0.95	0.95 (Liver, $2.81 \times 10^{21}$ )	CS, Metab.	Myocardial infarction
AF	HCAR3 (TG)	0.85	0.85	0.92	0.92 (Monocyte, $4.75 \times 10^{15}$ )	Metab.	ARI-3037MO*
	TNNC1 (TG)	0.48	0.45	0.78	0.78 (Liver, $2.81 \times 10^{21}$ )	CS, Muscle	Cardiomyopathy, Levosimendan*
	RAFI (TG)	0.79	0.83	0.90	0.90 (Aorta, $3.71 \times 10^{27}$ )	CS, Immune	Cardiomyopathy, Semapimod*
LDL	APOA1 (TG)	0.70	0.76	0.90	0.90 (Liver, $7.66 \times 10^{27}$ )	CS, Metab.	Amyloidosis, HDL deficiency
	ACADVL (TG)	0.69	0.59	0.85	0.85 (NK cell, $5.18 \times 10^{30}$ )	Liver, Metab.	VLCAD deficiency
T2D	ITGB6 (TG)	0.75	0.99	0.99	0.99 (Ileum, $4.52 \times 10^{62}$ )	Immune, Metab.	Amelogenesis imperfecta type IH
HR	TKT (TG)	0.65	0.67	0.92	0.93 (Aorta, $2.43 \times 10^7$ )	CS, Growth	SDDHD
CAD	OSM (TG)	0.56	0.78	0.86	0.86 (Aorta, $1.09 \times 10^{27}$ )	Immune, Metab.	GSK2330811*
	TRIB1 (TG)	0.43	0.68	0.85	0.85 (Adipose, $1.67 \times 10^{29}$ )	Adipose, Metab.	
AF	TAB2 (TG)	0.19	0.43	0.61	0.61 (CD8, $1.13 \times 10^{25}$ )	CS	Congenital heart defects
	TPMT (TG)	0.88	0.93	0.99	0.99 (Ileum, $4.43 \times 10^{13}$ )	Metab.	Poor metabolism of thiopurines-1
	RUNX1 (TF)	0.44	0.60	0.88	0.89 (Heart, $2.15 \times 10^{14}$ )	CS, Immune	Acute myeloid leukemia, FPDMM
LOAD	CSF3 (TG)	0.56	0.72	0.88	0.88 (Muscle, $8.55 \times 10^{14}$ )	Blood, Immune	Interleukin-3*
	CASP2 (TG)	0.99	1.00	1.00	1.00 (CD8, $8.31 \times 10^{26}$ )	Cellular, NS	Caspase-2*
SCZ	TTR (TG)	0.64	0.92	0.94	0.94 (Pancreas, $3.53 \times 10^{20}$ )	Metab.	Amyloidosis, Inotersen*, Patisiran*
	RORA (TF)	1.00	1.00	1.00	1.00 (Cortex, $5.39 \times 10^{128}$ )	Neuron, NS	Intellectual disability
NEU	ERBB4 (TG)	1.00	1.00	1.00	1.00 (Putamen, $7.22 \times 10^{116}$ )	Neuron, NS	Amyotrophic lateral sclerosis-19
	NFIB (TF)	0.97	0.97	0.98	0.98 (Cortex, $5.39 \times 10^{128}$ )	NS	MACID
	GRIK2 (TG)	0.90	0.94	0.97	0.97 (Cerebellum, $3.15 \times 10^{129}$ )	Neuron, NS	Mental retardation
	SYT1 (TG)	0.84	0.89	0.93	0.93 (Cerebellum, $3.15 \times 10^{129}$ )	Neuron, NS	Baker-Gordon syndrome
	ESR1 (TF)	0.80	0.84	0.93	0.93 (Colon, $1.07 \times 10^{141}$ )	Neuron, NS	Migraine
	NTRK2 (TG)	0.78	0.84	0.91	0.91 (Cerebellum, $3.15 \times 10^{129}$ )	Neuron, NS	DEE58
	LRRK2 (TG)	0.73	0.78	0.86	0.86 (Monocyte, $5.85 \times 10^{131}$ )	Neuron, NS	Parkinsonism, DNL151*, DNL201*
	C9orf72 (TG)	0.74	0.78	0.83	0.83 (Spleen, $1.44 \times 10^{141}$ )	Neuron, NS	FTDALS1
	SNCA (TG)	0.60	0.66	0.74	0.74 (Cerebellum, $3.15 \times 10^{129}$ )	Neuron, NS	Parkinsonism, BIIB054*
	LMBRD1 (TG)	0.42	0.66	0.94	0.94 (Ileum, $8.56 \times 10^{22}$ )	Metab.	MAHCF
NEU	PRKCQ (TG)	0.36	0.56	0.90	0.91 (Spleen, $2.13 \times 10^{19}$ )	Immune, NS	
	ATPIA2 (TG)	0.33	0.39	0.76	0.78 (Putamen, $2.12 \times 10^{19}$ )	Neuron, NS	AHC1, FHM2

AMC3: myogenic-type arthrogyrosis multiplex congenita-3. EDMD4: Emery-Dreifuss muscular dystrophy-4. SCAR8: autosomal recessive spinocerebellar ataxia-8. VLCAD: very long-chain acyl-CoA dehydrogenase. SDDHD: short stature, developmental delay, and congenital heart defects. FPDMM: familial platelet disorder with associated myeloid malignancy. MACID: acquired macrocephaly with impaired intellectual development. FTDALS1: frontotemporal dementia and/or amyotrophic lateral sclerosis. MAHCF: methylmalonic aciduria and homocystinuria of the cblF type. AHC1: alternating hemiplegia of childhood-1. FHM2: familial hemiplegic migraine-2. The remaining abbreviations are the same as in Table 1.

**Gene regulatory networks.** In this study a regulatory network is a directed bipartite graph  $\{V_{\text{TF}}, V_{\text{TG}}, E_{\text{TF} \rightarrow \text{TG}}\}$ , where  $V_{\text{TF}}$  and  $V_{\text{TG}}$  denote the node sets of TFs and TGs respectively, and  $E_{\text{TF} \rightarrow \text{TG}}$  denotes the set of TF-to-TG edges, summarizing how TFs regulate TGs through REs (Fig. 1b; Supplementary Note 4). Each edge has a weight between 0 and 1, measuring the relative regulation strength of a TF on a TG.

We inferred 38 regulatory networks from context-matched sequencing data of gene expression (e.g., RNA-seq) and chromatin accessibility (e.g., DNase-seq or ATAC-seq). We obtained these PECA data from ENCODE<sup>29</sup> (<https://www.encodeproject.org>, accessed December 14, 2018) and GTEx<sup>9</sup> (<https://gtexportal.org>, accessed July 13, 2019); see Supplementary Data 1. The network-construction software and TF-motif information are available at <https://github.com/suwonglab/PECA>. The 38 networks are available at <https://github.com/suwonglab/rss-net>, with descriptive statistics provided in Supplementary Tables 9–11.

We first constructed an “omnibus” network from PECA data of 201 biosamples across 80 cell types and tissues, using a regression-based method<sup>20</sup>. In brief, by modeling the distribution of TG expression levels conditional on RE accessibility levels and TF expression levels, we estimated a regression coefficient for each TF-TG pair. We selected a TF-TG pair as the network edge if this estimated coefficient was significantly non-zero, and divided the estimate by the maximum of estimates for all TF-TG pairs to set a (0, 1)-scale edge weight. We also estimated a regression coefficient for each RE-TG pair, which reflected the regulating strengths of REs on TGs and was later used to construct context-specific networks, i.e.,  $\{I_{ij}\}$  in Eq. (1). Here we defined REs as open chromatin peaks called from accessibility sequencing data by MACS2<sup>59</sup> (<https://github.com/macs3-project/MACS>, accessed July 12, 2018).

With the omnibus network in place, we then constructed context-specific networks for 5 immune cell types, 5 brain regions and 27 non-brain tissues. For each context (tissue or cell type), we computed a trans-regulation score (TRS)



between TF  $g$  and TG  $t$ :

$$\text{TRS}_{gt} = 2^{|\mathcal{R}_{gt}|} \cdot \sqrt{\widehat{\text{TF}}_g \cdot \widehat{\text{TG}}_t \cdot \sum_i (\widehat{\text{RE}}_i \cdot B_{gi} \cdot I_{it})}, \quad (1)$$

where  $R_{gt}$  is the correlation of TF  $g$  and TG  $t$  expression levels across all contexts;  $\{\widehat{\text{TF}}_g, \widehat{\text{TG}}_t, \widehat{\text{RE}}_i\}$  are normalized context-specific expression (TF  $g$ , TG  $t$ ) and accessibility (RE  $i$ ) levels ( $\widehat{y} = y^2/y_{\text{med}}$ , where  $y$  denotes the actual accessibility or expression level in a given context, and  $y_{\text{med}}$  denotes median level across all contexts);  $B_{gi}$  reflects the motif binding strength of TF  $g$  on RE  $i$ , defined as the sum of motif position weight matrix-based log-odds probabilities of all binding sites on RE  $i$  and calculated by HOMER<sup>60</sup> (<http://homer.ucsd.edu/homer/>, accessed July 12, 2018); and  $I_{it}$  reflects the overall regulating strength of RE  $i$  on TG  $t$ , provided by the omnibus network. TRS naturally ranks and selects context-specific TF-TG edges because a larger value of  $\text{TRS}_{gt}$  indicates a stronger regulating strength of TF  $g$  on TG  $t$  in the given context. We set (0,1)-scale TF-TG edge weights by computing  $\log_2(1 + \text{TRS}_{gt})/\max_{i,j}\{\log_2(1 + \text{TRS}_{ij})\}$ .

To validate PECA-based networks and illustrate RSS-NET as a generally applicable tool, we also analyzed 394 cell type- and tissue-specific TF-TG circuits<sup>18</sup> inferred from independent CAGE data<sup>7,8</sup> (<http://regulatorycircuits.org/>, accessed May 8, 2019). When evaluating the similarity between PECA- and CAGE-based networks (Fig. 5b; Supplementary Fig. 12), we used their full node and edge sets to compute Jaccard indices. When running RSS-NET on context-matched PECA- and CAGE-based networks (Fig. 5d; Supplementary Fig. 14), we selected top-ranked CAGE-based edges to match PECA-based edge counts (Supplementary Table 10) and normalized CAGE-based edge weights ( $\widehat{x} = \min\{1, x^{1/6}\}$ , where  $x$  denotes original weight) to match the scale of PECA-based edge weights (Supplementary Table 11).

**External databases for cross-reference.** To validate and interpret RSS-NET results, we used the following external databases (accessed November 28, 2019): GWAS Catalog<sup>1</sup> (<https://www.ebi.ac.uk/gwas/>), Mouse Genome Informatics<sup>46</sup> (<http://www.informatics.jax.org/>), Mendelian gene sets<sup>48</sup> ([https://github.com/bogdanlab/gene\\_sets/](https://github.com/bogdanlab/gene_sets/)), Online Mendelian Inheritance in Man<sup>49</sup> (<https://www.omim.org/>), Therapeutic Target Database<sup>52</sup> (<http://db.idrblab.net/ttd/>).

When quantifying overlaps between RSS-NET prioritized genes and mouse or Mendelian genes, we used all genes for each GWAS trait. We repeated the overlap analysis under the same significance cutoff (FDR  $\leq 0.1$ ) after excluding genes implicated in the same or later GWAS (Supplementary Table 7). Since GWAS-implicated genes overlap significantly with phenotypically-matched mouse and Mendelian genes (median two-sided Fisher exact  $P = 7.1 \times 10^{-7}$ ), we identified fewer discoveries as expected (mouse-human pairs: 26, Mendelian-complex pairs: 4; Supplementary Data 4–5), but we obtained consistent effect sizes nonetheless (mouse  $R = 0.78$ , two-sided  $P = 8.6 \times 10^{-73}$ ; Mendelian  $R = 0.89$ ,  $P = 9.0 \times 10^{-74}$ ; Supplementary Fig. 15).

**Network-induced effect size distribution.** We model the total effect of SNP  $j$  on a given trait  $\beta_j$  as

$$\beta_j \sim \pi_j \cdot \mathcal{N}(\mu_j, \sigma_0^2) + (1 - \pi_j) \cdot \delta_0, \quad (2)$$

where  $\pi_j$  denotes the probability that SNP  $j$  is associated with the trait ( $\beta_j \neq 0$ ),  $\mathcal{N}(\mu_j, \sigma_0^2)$  denotes a normal distribution with mean  $\mu_j$  and variance  $\sigma_0^2$  specifying the effect size of a trait-associated SNP  $j$ , and  $\delta_0$  denotes point mass at zero ( $\beta_j = 0$ ).

We model the trait-association probability  $\pi_j$  as

$$\log_{10} \left( \frac{\pi_j}{1 - \pi_j} \right) = \theta_0 + a_j \cdot \theta, \quad (3)$$

where  $\theta_0 < 0$  captures the genome-wide background proportion of trait-associated SNPs,  $\theta > 0$  reflects the increase in probability, on the log10-odds scale, that a SNP near network genes and REs is trait-associated, and  $a_j$  reflects the proximity of SNP  $j$  to a network. Following previous analyses<sup>15,16,24</sup>, we let  $a_j = 1$  if SNP  $j$  is within 100 kb of any member gene (TF, TG) or RE for a given network. Equation (3) suggests that if a cell type or tissue plays an important role in a trait then genetic associations may occur more often in SNPs involved in the corresponding network genes and REs than expected by chance.

We model the mean effect size  $\mu_j$  as

$$\mu_j = \sum_{g \in \mathcal{O}_j} w_{jg} \cdot \gamma_{jg}, \quad (4)$$

where  $\mathcal{O}_j$  is the set of all nearby or distal genes contributing to the total effect of SNP  $j$ ,  $w_{jg}$  measures the relevance between SNP  $j$  and gene  $g$ , and  $\gamma_{jg}$  denotes the effect of SNP  $j$  on a trait due to gene  $g$ . Equation (4) provides a general decomposition of total SNP effect into gene effects through  $\{\mathcal{O}_j, w_{jg}\}$ .

Here we use a TF-TG network to specify  $\{\mathcal{O}_j, w_{jg}\}$  in Eq. (4):

$$\mu_j = \underbrace{\sum_{g \in \mathcal{G}_j} [c_{jg} \cdot (\gamma_{jg} + \sum_{t \in \mathcal{T}_g} v_{gt} \cdot \gamma_{jt})]}_{\text{cis}} \quad (5)$$

where  $\mathcal{G}_j$  is the set of all genes within 1 Mb window of SNP  $j$  (a standard window size

used in cis-eQTL studies<sup>9,10,58</sup>),  $c_{jg}$  measures the relative impact of a SNP  $j$  on gene  $g$ ,  $\mathcal{T}_g$  is the set of all genes directly regulated by TF  $g$  in a given network ( $\mathcal{T}_g$  is empty if gene  $g$  is not a TF), and  $v_{gt}$  measures the relative impact of a TF  $g$  on its TG  $t$ . Since a genome-wide analysis typically involves many SNPs and genes, we fix  $\{\mathcal{T}_g, v_{gt}, c_{jg}\}$  to ensure the identifiability of Eq. (5). We use inferred edges and weights of a context-specific TF-TG network<sup>20,29</sup> to specify  $\mathcal{T}_g$  and  $v_{gt}$  respectively. We use context-matched cis-eQTL<sup>9,10,58</sup> to specify  $c_{jg}$  (Supplementary Note 5 and Tables 12, 13). Equation (5) suggests that the total effect of a SNP may fan out through some regulatory network of multiple (nearby or distal) genes to affect the trait<sup>22</sup>.

We model the random effect  $\gamma_{jg}$  of SNP  $j$  due to gene  $g$  as

$$\gamma_{jg} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \quad (6)$$

where the SNP-level subscript  $j$  in  $\gamma_{jg}$  ensures the exchangeability of  $\beta_j$  in Eq. (2); see Supplementary Note 6. Equation (6) uses a constant  $\sigma^2$  for computational convenience. Equation (6) could be modified by letting  $\sigma^2$  depend on functional annotations<sup>13,27</sup> of SNP  $j$  and context-specific expression<sup>14–16</sup> of gene  $g$ , though possibly at higher computational cost.

Equations (2), (4), and (6) implies a variance decomposition for SNP effect:

$$\text{Var}(\beta_j) = \pi_j \cdot (\sigma_0^2 + \sigma^2 \cdot \sum_{g \in \mathcal{O}_j} w_{jg}^2). \quad (7)$$

We hypothesize that Eq. (7) may provide an alternative approach to heritability analyses<sup>13,24,27</sup> and we plan to investigate it elsewhere.

**Bayesian hierarchical modeling.** Consider a GWAS with  $n$  unrelated individuals measured on  $p$  SNPs. In practice we do not know the true SNP-level effects  $\beta := (\beta_1, \dots, \beta_p)'$  in Eq. (2), but we can infer them from GWAS summary statistics and LD estimates. Specifically, we perform Bayesian inference for  $\beta$  by combining the network-based prior defined by Eqs. (2)–(6) with the RSS likelihood<sup>25</sup>:

$$\widehat{\beta} \sim \mathcal{N}(\widehat{\text{SR}}^{-1} \beta, \widehat{\text{SR}}), \quad (8)$$

where  $\widehat{\beta} := (\widehat{\beta}_1, \dots, \widehat{\beta}_p)'$ ,  $\widehat{\text{S}} := \text{diag}(\widehat{\text{s}})$  is a  $p \times p$  diagonal matrix with  $\widehat{\text{s}} := (\widehat{s}_1, \dots, \widehat{s}_p)'$ ,  $\{\widehat{\beta}_j, \widehat{s}_j\}$  are estimated single-SNP effect size of each SNP  $j$  and its standard error from the GWAS, and  $\widehat{\text{R}} := [\widehat{r}_{ij}]$  is the  $p \times p$  LD matrix estimated from a reference panel with ancestry matching the GWAS.

RSS-NET, defined by Eqs. (2)–(6), and (8), consists of four unknown hyper-parameters  $\{\theta_0, \theta, \sigma_0^2, \sigma^2\}$ . To specify hyper-priors, we first introduce two free parameters  $\{\eta, \rho\}$  to re-parameterize  $\{\sigma_0^2, \sigma^2\}$ :

$$\sigma_0^2 = \eta \cdot (1 - \rho) \cdot \left( \sum_{j=1}^p \frac{\pi_j}{n s_j^2} \right)^{-1}, \quad \sigma^2 = \eta \cdot \rho \cdot \left( \sum_{j=1}^p \frac{\pi_j \cdot \sum_{g \in \mathcal{O}_j} w_{jg}^2}{n s_j^2} \right)^{-1}, \quad (9)$$

where, roughly,  $\eta$  represents the proportion of the total phenotypic variation explained by  $p$  SNPs, and  $\rho$  represents the proportion of total genetic variation explained by network annotations  $\{\mathcal{O}_j, w_{jg}\}$ . Because  $n s_j^2$  approximates the ratio of phenotype variance to genotype variance, Eq. (9) ensures that SNP effects ( $\beta$ ) do not rely on sample size  $n$  and have the same measurement unit as the trait. See Supplementary Note 7 for derivation of Eq. (9).

We then place independent uniform grid priors on  $\{\theta_0, \theta, \eta, \rho\}$  (Supplementary Table 14). These simple hyper-priors produce accurate posterior estimates for hyper-parameters in simulations (Supplementary Fig. 16). RSS-NET results are robust to grid choice on both simulated and real data (Supplementary Figs. 17–18). (If one had specific information about  $\{\theta_0, \theta, \eta, \rho\}$  in a given setting then this could be incorporated in the hyper-priors.)

**Network enrichment.** To assess whether a regulatory network is enriched for genetic associations with a trait, we evaluate a Bayes factor (BF):

$$\text{BF} = \frac{f(\widehat{\beta} | \widehat{\text{S}}, \widehat{\text{R}}, \mathbf{a}, \mathbf{O}, \mathbf{W}, M_1)}{f(\widehat{\beta} | \widehat{\text{S}}, \widehat{\text{R}}, \mathbf{a}, \mathbf{O}, \mathbf{W}, M_0)}, \quad (10)$$

where  $f(\cdot)$  denotes probability densities,  $\mathbf{a}$  is defined in Eq. (3),  $\{\mathbf{O}, \mathbf{W}\}$  are defined in Eq. (4),  $M_1$  denotes the enrichment model with  $\theta > 0$  or  $\sigma^2 > 0$ , and  $M_0$  denotes the baseline model with  $\theta = 0$  and  $\sigma^2 = 0$ . The observed data are BF times more likely under  $M_1$  than under  $M_0$ , and so the larger the BF, the stronger evidence for network enrichment. See Supplementary Note 2 for computation details. To compute BFs used in Fig. 5c, we replace  $M_1$  in Eq. (10) with three restricted enrichment models ( $M_{11}, M_{12}, M_{13}$ ). Unless otherwise specified, all BFs reported in this work are based on  $M_1$ .

Given a BF cutoff, false positive rates vary considerably across genetic architectures and enrichment patterns in simulations (Supplementary Table 15). As the genetic basis of most complex traits remains unknown, we find it impractical to fix some significance threshold. Instead we recommend an adaptive approach. Specifically, for a given GWAS we run RSS-NET on a near-gene control network containing all genes as nodes and no edges (i.e.,  $a_j = 1$  for all SNPs within 100 kb of any gene and  $v_{gt} = 0$  for all TF-TG pairs), and we use the resulting BF as the enrichment threshold in this GWAS. Our analyses show three advantages of this approach. First, it is adaptive to study heterogeneity such as trait differences and

sample sizes (Supplementary Table 1). Second, it accounts for generic enrichments of genetic signals residing near genes. Third, it facilitates comparisons with non-Bayesian methods based on  $P$ -values (Supplementary Table 2).

**Locus association.** To identify the association between a locus and a trait, we compute  $P_1$ , the posterior probability that at least one SNP in the locus is associated with the trait:

$$P_1 = 1 - \Pr(\beta_j = 0, \forall j \in \text{locus} \mid \mathbf{D}, \text{model}), \quad (11)$$

where  $\mathbf{D}$  is a shorthand for the input data of RSS-NET including GWAS summary statistics  $(\hat{\beta}, \hat{S})$ , LD estimates  $\hat{\mathbf{R}}$  and network annotations  $\{\mathbf{a}, \mathbf{O}, \mathbf{W}\}$ . See Supplementary Note 3 for computation details. For a locus,  $P_1^{\text{base}}$ ,  $P_1^{\text{near}}$ , and  $P_1^{\text{net}}$  correspond to  $P_1$  evaluated under the baseline model  $M_0$ , the enrichment model  $M_1$  for the near-gene control network, and  $M_1$  for a given TF-TG network. In this study, we defined a locus as the transcribed region of a gene plus 100 kb up and downstream, and we used “locus” and “gene” interchangeably.

For  $K$  networks with enrichments stronger than the near-gene control, we use Bayesian model averaging (BMA) to compute  $P_1^{\text{bma}}$  for each locus:

$$P_1^{\text{bma}} = \frac{\sum_{k=1}^K P_1^{\text{net}}(k) \cdot \text{BF}(k)}{\sum_{k=1}^K \text{BF}(k)}, \quad (12)$$

where  $P_1^{\text{net}}(k)$  and  $\text{BF}(k)$  are enrichment  $P_1$  and BF for network  $k$ . The ability to average across networks in Eq. (12) is an advantage of our Bayesian framework, because it allows us to assess associations in light of network enrichment without having to select a single enriched network.

In this study we used  $P_1 \geq 0.9$  as the significance cutoff, yielding a median false positive rate  $1.24 \times 10^{-4}$  and a median false discovery rate  $6.43 \times 10^{-2}$  in simulations (Supplementary Tables 16, 17). We also highlighted genes with  $P_1^{\text{net}} > P_1^{\text{near}}$  (Fig. 6 and Tables 1, 2), because they showcase the influence of context-specific regulatory topology on prioritizing genetic associations.

**Computation time.** The total computation time of RSS-NET to analyze a pair of trait and network is determined by the number of genome-wide SNPs analyzed, the size of hyper-parameter grid, and the number of variational iterations till convergence, all of which can vary considerably among studies. It is thus hard to make general statements about computation time. However, to give a specific example, we finished the analysis of 1,032,214 HapMap3 SNPs and liver network for HDL within 12 hours in a standard computer cluster (60 nodes, 8 CPUs, and 32 Gb memory per node).

The number of genome-wide SNPs analyzed ( $p$ ) affects the computation time of RSS-NET in two distinct ways. First, the per-iteration complexity of RSS-NET is linear with  $p$  (Box 1; Supplementary Note 1). Second, a large  $p$  defines a large optimization problem, often requiring many iterations to converge. To quantify the impact of  $p$  on computation time, we simulated datasets from different sets of genome-wide SNPs, analyzed them with RSS-NET on identical computers, and compared the computation time (Supplementary Fig. 9). When  $p$  increased from 348,965 to 1,030,397, on average the total computation time was four times longer (one-sided Wilcoxon  $P = 8.0 \times 10^{-132}$ ).

**Simulation overview.** To assess the network-induced model for SNP effects ( $\beta$ ) in RSS-NET, we simulated a large array of correctly- and mis-specified  $\beta$  for a given target network. Specifically, we generated “positive” datasets where the underlying  $\beta$  was simulated from  $M_1$  for the target network, and “negative” datasets where  $\beta$  was simulated from either  $M_0$  or the following scenarios: (1) random enrichments of near-gene SNPs; (2) random enrichments of near-RE SNPs; (3) MAF- and LD-dependent effect sizes; (4)  $M_1$  for edge-altered copies of the target network. For a fair comparison in each scenario, we matched positive and negative datasets by both the number of trait-associated SNPs and the proportion of phenotypic variation explained by all SNPs. See Supplementary Figs. 1–9 for details.

We combined the simulated  $\beta$  with genotypes of 348,965 genome-wide SNPs from 1,458 individuals<sup>28</sup> to simulate phenotypes using an additive multiple-SNP model with Gaussian noise. We performed the standard single-SNP analysis of simulated individual-level datasets to generate GWAS summary statistics, on which we compared RSS-NET with external methods.

**External software for benchmarking.** To benchmark RSS-NET this study used the following software: RSS-E (<https://github.com/stephenslab/rss>, accessed October 19, 2018), Pascal (<https://www2.unil.ch/cbg/index.php?title=Pascal>, accessed October 5, 2017) and LDSC with two sets of baseline annotations as covariates (version 1.0.0, <https://github.com/bulik/ldsc>; baseline model v1.1, [https://data.broadinstitute.org/alkesgroup/LDSCORE/1000G\\_Phase3\\_baseline\\_v1.1\\_ldscores.tgz](https://data.broadinstitute.org/alkesgroup/LDSCORE/1000G_Phase3_baseline_v1.1_ldscores.tgz); baselineLD model v2.1, [https://data.broadinstitute.org/alkesgroup/LDSCORE/1000G\\_Phase3\\_baselineLD\\_v2.1\\_ldscores.tgz](https://data.broadinstitute.org/alkesgroup/LDSCORE/1000G_Phase3_baselineLD_v2.1_ldscores.tgz); accessed November 27, 2018). Versions of all packages and files were up-to-date at the time of analysis.

Given a context-specific TF-TG network, RSS-E and LDSC methods use the same binary SNP-level annotations  $\{a_j\}$  defined in Eq. (3). The interface design of Pascal does not allow direct usage of  $\{a_j\}$ . Here we supplied Pascal program with a GMT file containing all member genes of a network and set SNP-to-gene window sizes as 100 kb (“-up = 100000 -down = 100000”). In this study all external

methods were used with their default setups, which did not include the edge information of a network.

RSS-E outputs the same statistics as RSS-NET (BF and  $P_1$ ). Pascal implements two gene scoring methods (maximum-of- $\chi^2$  and sum-of- $\chi^2$ ) to produce gene-based association  $P$ -values. Given gene scores, Pascal provides two gene set scoring options ( $\chi^2$  approximation and empirical sampling) to produce enrichment  $P$ -values. LDSC methods output enrichment  $P$ -values and coefficient  $Z$ -scores, yielding consistent results in our simulations (LDSC-baseline:  $R = 0.98$ , two-sided  $P = 1.2 \times 10^{-67}$ ; LDSC-baselineLD:  $R = 0.98$ ,  $P = 9.1 \times 10^{-63}$ ; Supplementary Fig. 19). Due to the higher power shown in simulations (LDSC-baseline: average AUROC increase = 0.012, one-sided  $t P = 4.0 \times 10^{-3}$ ; LDSC-baseline LD: average AUROC increase = 0.023, one-sided  $t P = 1.5 \times 10^{-5}$ ), we used enrichment  $P$ -values from LDSC in this study.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

The 38 network files are available at <https://github.com/suwonglab/rss-net> (<https://doi.org/10.5281/zenodo.4553387>). Analysis results of 38 networks and 18 traits are available at <https://suwonglab.github.io/rss-net/results>. Links and identifiers of other data are specified in Methods, Supplementary Notes 5 and 8. Source data are provided with this paper.

## Code availability

The RSS-NET software is available at <https://github.com/suwonglab/rss-net> (<https://doi.org/10.5281/zenodo.4553387>). Tutorials of installing and using RSS-NET are available at <https://suwonglab.github.io/rss-net>. Results of this study were generated from MATLAB version 9.3.0.713579 (R2017b), on a Linux system with Intel E5-2650V2 2.6 GHz and E5-2640V4 2.4 GHz processors. Links and identifiers of other codes are specified in Methods.

Received: 8 December 2020; Accepted: 3 March 2021;

Published online: 14 May 2021

## References

- Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
- Tam, V. et al. Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 467–484 (2019).
- Heikselman, I. & Yeger-Lotem, E. Mechanisms of tissue and cell-type specificity in heritable traits and diseases. *Nat. Rev. Genet.* **21**, 137–150 (2020).
- French, J. & Edwards, S. The role of noncoding variants in heritable disease. *Trends Genet.* **36**, 880–891 (2020).
- The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
- Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
- Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
- The FANTOM Consortium and the RIKEN PMI and CLST (DGT). A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
- GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
- Schmiedel, B. J. et al. Impact of genetic polymorphisms on human immune cell gene expression. *Cell* **175**, 1701–1715 (2018).
- Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
- Trynka, G. et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* **45**, 124–130 (2013).
- Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
- Calderon, D. et al. Inferring relevant cell types for complex traits by using single-cell gene expression. *Am. J. Hum. Genet.* **101**, 686–699 (2017).
- Finucane, H. K. et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
- Zhu, X. & Stephens, M. Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. *Nat. Commun.* **9**, 4361 (2018).
- Greene, C. S. et al. Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* **47**, 569–576 (2015).
- Marbach, D. et al. Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat. Methods* **13**, 366–370 (2016).

19. Sonawane, A. R. et al. Understanding tissue-specific gene regulation. *Cell Rep.* **21**, 1077–1088 (2017).
20. Duren, Z., Chen, X., Jiang, R., Wang, Y. & Wong, W. H. Modeling gene regulation from paired expression and chromatin accessibility data. *Proc. Natl. Acad. Sci. USA* **114**, E4914–E4923 (2017).
21. Califano, A., Butte, A. J., Friend, S., Ideker, T. & Schadt, E. Leveraging models of cell regulation and GWAS data in integrative network-based association studies. *Nat. Genet.* **44**, 841–847 (2012).
22. Liu, X., Li, Y. I. & Pritchard, J. K. Trans effects on gene expression can drive omnigenic inheritance. *Cell* **177**, 1022–1034 (2019).
23. Li, W., Duren, Z., Jiang, R. & Wong, W. H. A method for scoring the cell type-specific impacts of noncoding variants in personal genomes. *Proc. Natl. Acad. Sci. USA* **117**, 21364–21372 (2020).
24. Kim, S. S. et al. Genes with high network connectivity are enriched for disease heritability. *Am. J. Hum. Genet.* **104**, 896–913 (2019).
25. Zhu, X. & Stephens, M. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *Ann. Appl. Stat.* **11**, 1561–1592 (2017).
26. Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z. & Bergmann, S. Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. *PLoS Comput. Biol.* **12**, e1004714 (2016).
27. Gazal, S. et al. Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).
28. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
29. Luo, Y. et al. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res.* **48**, D882–D889 (2020).
30. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
31. International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
32. Lee, B.-C. et al. Adipose natural killer cells regulate adipose tissue macrophages to promote insulin resistance in obesity. *Cell Metab.* **23**, 685–698 (2016).
33. Hotamisligil, G. S. Inflammation, metaflammation and immunometabolic disorders. *Nature* **542**, 177–185 (2017).
34. Oikonomou, E. K. & Antoniades, C. The role of adipose tissue in cardiovascular health and disease. *Nat. Rev. Cardiol.* **16**, 83–99 (2019).
35. Kang, J. & Rivest, S. Lipid metabolism and neuroinflammation in Alzheimer's disease: a role for liver X receptors. *Endocr. Rev.* **33**, 715–746 (2012).
36. Shahid, F., Lip, G. Y. & Shantsila, E. Role of monocytes in heart failure and atrial fibrillation. *J. Am. Heart Assoc.* **7**, e007849 (2018).
37. Aviles, R. J. et al. Inflammation as a risk factor for atrial fibrillation. *Circulation* **108**, 3006–3010 (2003).
38. Noordam, R. et al. Multi-ancestry sleep-by-SNP interaction analysis in 126,926 individuals reveals lipid loci stratified by sleep duration. *Nat. Commun.* **10**, 5121 (2019).
39. Teslovich, T. M. et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
40. Kerr, T. A. et al. Loss of nuclear receptor SHP impairs but does not eliminate negative feedback regulation of bile acid synthesis. *Dev. Cell* **2**, 713–720 (2002).
41. Wang, L. et al. Redundant pathways for negative feedback regulation of bile acid production. *Dev. Cell* **2**, 721–731 (2002).
42. Hartman, H. B., Lai, K. & Evans, M. J. Loss of small heterodimer partner expression in the liver protects against dyslipidemia. *J. Lipid Res.* **50**, 193–203 (2009).
43. Nishigori, H. et al. Mutations in the small heterodimer partner gene are associated with mild obesity in Japanese subjects. *Proc. Natl. Acad. Sci. USA* **98**, 575–580 (2001).
44. Willer, C. et al. Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).
45. Brown, S. et al. High-throughput mouse phenomics for characterizing mammalian gene function. *Nat. Rev. Genet.* **19**, 357–370 (2018).
46. Bult, C. et al. Mouse Genome Database (MGD) 2019. *Nucleic Acids Res.* **47**, D801–D806 (2019).
47. Blair, D. R. et al. A nondegenerate code of deleterious variants in Mendelian loci contributes to complex disease risk. *Cell* **155**, 70–80 (2013).
48. Freund, M. K. et al. Phenotype-specific enrichment of Mendelian disorder genes near GWAS regions across 62 complex traits. *Am. J. Hum. Genet.* **103**, 535–552 (2018).
49. Amberger, J., Bocchini, C., Scott, A. & Hamosh, A. OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res.* **47**, D1038–D1043 (2019).
50. Bruning, P. F. et al. Insulin resistance and breast-cancer risk. *Int. J. Cancer* **52**, 511–516 (1992).
51. Plenge, R., Scolnick, E. & Altshuler, D. Validating therapeutic targets through human genetics. *Nat. Rev. Drug Discov.* **12**, 581–594 (2013).
52. Wang, Y. et al. Therapeutic Target Database 2020: enriched resource for facilitating research and early development of targeted therapeutics. *Nucleic Acids Res.* **48**, D1031–D1041 (2020).
53. Sanseau, P. et al. Use of genome-wide association studies for drug repositioning. *Nat. Biotechnol.* **30**, 317–320 (2012).
54. Lieschke, G. et al. Mice lacking granulocyte colony-stimulating factor have chronic neutropenia, granulocyte and macrophage progenitor cell deficiency, and impaired neutrophil mobilization. *Blood* **84**, 1737–1746 (1994).
55. Astle, W. J. et al. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* **167**, 1415–1429 (2016).
56. Barrett, A. & Sloand, E. Autoimmune mechanisms in the pathophysiology of myelodysplastic syndromes and their clinical relevance. *Haematologica* **94**, 449–451 (2009).
57. Schaid, D., Chen, W. & Larson, N. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).
58. Vösa, U. et al. Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. Preprint at <https://doi.org/10.1101/447367> (2018).
59. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
60. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).

### Acknowledgements

This study is supported by Stein Fellowship to X.Z. and NIH grants P50HG007735 and R01HG010359 to W.H.W. This study uses computational resources provided by the Stanford Research Computing Center. This study uses data generated by the WTCCC, 1000 Genomes, ENCODE, GTEx, DICE, eQTLGen, and multiple GWAS consortia (Supplementary Note 8). We thank Xin He for helpful comments on a draft manuscript.

### Author contributions

X.Z. and W.H.W. conceived the study. X.Z. developed the methods and implemented the software. X.Z. designed and conducted the simulation experiments. Z.D. provided the 38 regulatory networks. X.Z. performed the data analyses. X.Z. prepared the supplementary materials and online resources. X.Z. wrote the manuscript. X.Z. and W.H.W. revised the manuscript.

### Competing interests

The authors declare no competing interests.

### Additional information


**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-021-22588-0>.

**Correspondence** and requests for materials should be addressed to X.Z. or W.H.W.

**Peer review information** *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021