LAST EDITED ON 2020-03-12 BY X.Z.

MODELING REGULATORY NETWORK TOPOLOGY IMPROVES GENOME-WIDE ANALYSES OF COMPLEX HUMAN TRAITS

BY XIANG ZHU AND ZHANA DUREN AND WING HUNG WONG

Stanford University

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 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 networks specific to trait-relevant tissue or cell types. Prioritizing variants within enriched networks identifies known and new trait-associated genes revealing novel biological and therapeutic insights.

INTRODUCTION

Genome-wide association studies (GWAS) have catalogued many significant associations between common genetic variants, notably single-nucleotide polymorphisms (SNPs), and a full spectrum of human complex traits [1,2]. However, it remains challenging to translate most of these findings into biological mechanisms and clinical applications. In particular, most variants have small effects [3] and are often mapped to non-coding regions [4].

One possible interpretation is that non-coding variants cumulatively affect complex traits through gene regulation. To test this hypothesis, large-scale epigenomic [5,6] and transcriptomic [7,8] data have been made available spanning diverse human cell and tissue types. Exploiting these regulatory genomic data, many studies have shown enrichments of trait-associated SNPs in chromatin regions [6,9,10] and genes [11–13] that are active in trait-relevant tissue or cell types. These studies often incorporate regulatory information into effects of SNPs in a linear manner, and ignore potential functional interactions among loci within regulatory programs.

Gene regulatory networks [14–17] have proven useful in mining functional interactions of genes from genomic data. Further, context-specific regulatory

*Correspondence should be addressed to X.Z. (xiangzhu@stanford.edu) and W.H.W (whwong@stanford.edu).

networks are potentially informative to dissect the genetics of complex traits, since, through cellular interactions, trait-associated variants are likely to be topologically related[18]. Though promising, the full potential of regulatory networks is yet to be unleashed in GWAS. For example, recent connectivity analyses[15,19] identify enrichments of genetic signals across many traits and networks, but do not leverage observed enrichments to further enhance trait-associated gene discovery[14].

To further exploit regulatory networks in GWAS, we develop a Bayesian framework for simultaneous genome-wide enrichment and prioritization analysis. Through extensive simulations on the new method, we show its flexibility to various genetic architectures, its robustness to a wide range of model mis-specification, and its improved performance over existing methods. Applying the method to 18 human traits and 38 regulatory networks, we identify strong enrichments of genetic associations in networks that are specific to trait-relevant tissue or cell types. By prioritizing variants within the enriched networks we identify trait-associated genes that were not implicated by the same GWAS. Many of these putatively novel genes have strong support from multiple lines of external biological and clinical 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 novel biological and therapeutic insights from existing GWAS.

Results

Method overview. Figure 1 provides a schematic method overview. In brief, we develop a new prior distribution that dissects the total effect of a single variant on a trait into effects of multiple (nearby or distal) genes through a regulatory network, and then we combine this network-induced prior with a multiple-SNP regression likelihood based on single-SNP association summary statistics[20] to analyze regulatory networks and complex traits jointly. We refer to this integrative framework as RSS-NET (Methods).

RSS-NET specifies the likelihood for SNP-level effects (β) based on single-SNP effect estimates and their standard errors from GWAS (Fig. 1a), and linkage disequilibrium (LD) estimates from an external reference panel with ancestry matching the GWAS (Fig. 1b). For a given network (Fig. 1c), RSS-NET uses its topology (nodes and edges) to specify a prior that decomposes the total effect of each SNP (β) into effects of multiple interconnected genes. This prior contains two independent enrichment parameters: θ and σ^2, which measures the extent to which, SNPs near network nodes have increased likelihood to be associated with the trait, and, SNPs near network edges have larger effect sizes, respectively. See Methods for mathematical definitions.

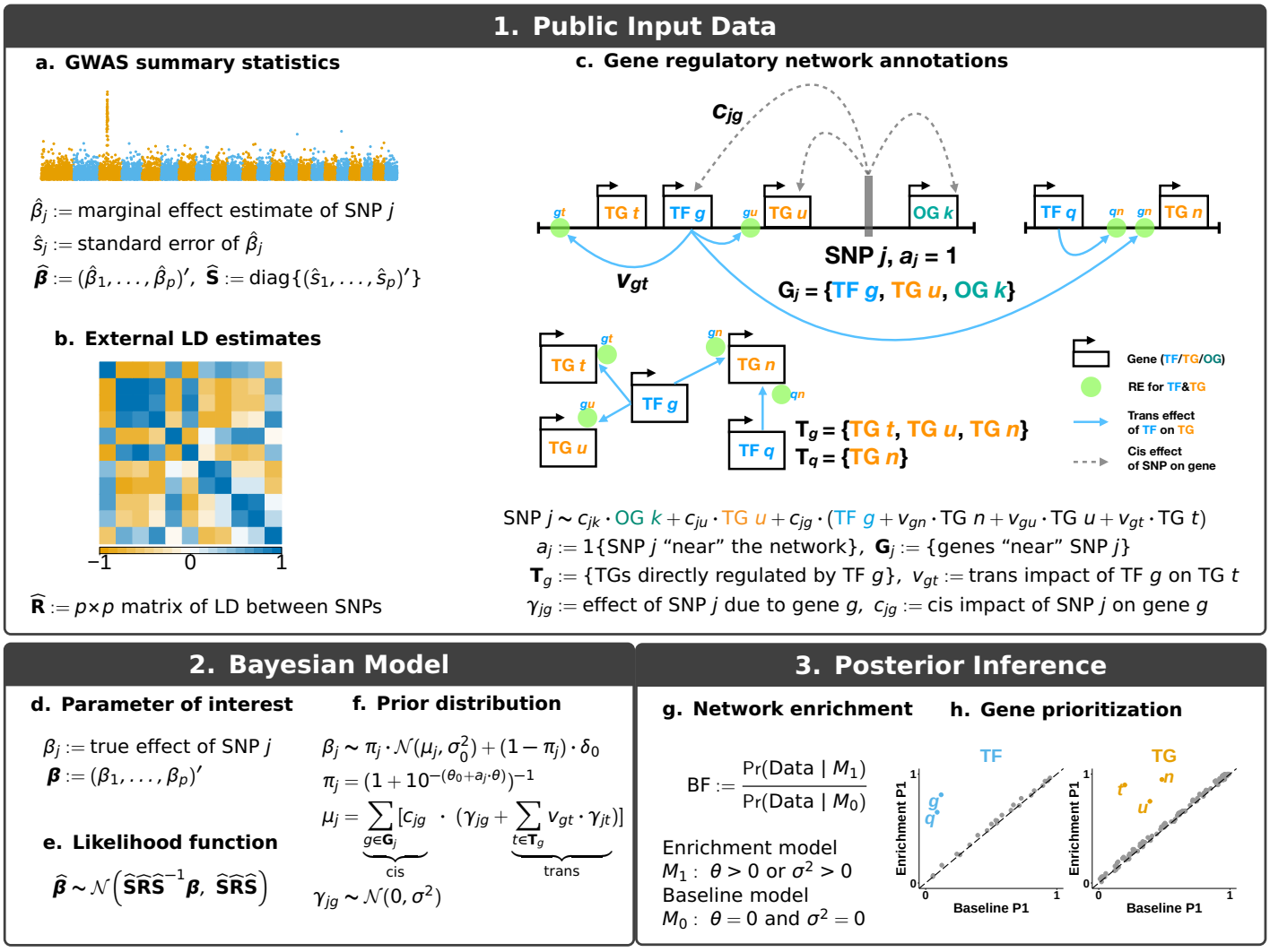

Fig 1: **Schematic overview of RSS-NET.** RSS-NET requires three types of input data: GWAS summary statistics (**a**), external LD estimates (**b**) and gene regulatory network annotations (**c**). Here a regulatory network is a bipartite graph that has two types of nodes, transcription factor (TF) and target gene (TG), and has directed edges from TFs to TGs through regulatory elements (REs). RSS-NET uses a regulatory network to decompose the total effect of each SNP into effects of multiple genes. For example, the expected total effect of SNP j shown in Panel **c** can be represented as a sum of cis effects of three nearby genes, outside-network gene (OG) k, TG u and TF g, and trans effects of three TGs (n, u, t) that are directly regulated by TF g. RSS-NET performs Bayesian hierarchical modeling (**d-f**) under two models about two enrichment parameters (θ for nodes; σ^2 for edges): the “baseline model” ($M_0 : \theta = 0$ and $\sigma^2 = 0$) that each SNP has equal chance of being associated with the trait ($\theta = 0$) and each trait-associated SNP has the same effect size distribution ($\sigma^2 = 0$), and, the “enrichment model” ($M_1 : \theta > 0$ or $\sigma^2 > 0$) that SNPs near network nodes are more often associated with the trait ($\theta > 0$) or SNPs near network edges have larger effect sizes ($\sigma^2 > 0$). To assess network enrichment, RSS-NET computes a Bayes factor (BF) comparing M_0 and M_1 (**g**). RSS-NET also automatically prioritizes loci within an enriched network by comparing the posterior distributions of genetic effects (β) under M_0 and M_1 (**h**). For each locus, RSS-NET summarizes the posterior of β as P_1, the posterior probability that at least one SNP in this locus is associated with the trait ($\beta_j \neq 0$). Differences between P_1 under M_0 and M_1 reflect the influence of a regulatory network on genetic associations, which can highlight new trait-associated genes (**h**).

RSS-NET provides a unified framework (Fig. 1d-f) for two tasks: (1) testing whether a network is enriched for genetic associations; (2) identifying which genes within this network drive the enrichment. To assess network enrichment (Fig. 1g), 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 (Fig. 1h) RSS-NET contrasts posterior distributions of genetic effects (β) under M_0 and M_1. RSS-NET outputs results for these two tasks simultaneously.

RSS-NET improves upon its predecessor RSS-E[13]. Indeed RSS-NET includes RSS-E as a special case where edge-enrichment $\sigma^2 = 0$ and only node-enrichment θ is learned from data. By estimating the additional parameter σ^2, RSS-NET is more flexible than RSS-E, and thus, RSS-NET consistently outperforms RSS-E in various simulation scenarios, and often yields better fit on real data. Despite different treatments of σ^2, RSS-NET and RSS-E share computation schemes (Supplementary Notes), which allows us to build RSS-NET on the efficient algorithm of RSS-E. Software is available at `https://github.com/suwonglab/rss-net`.

Method comparison based on simulations. The novelty of RSS-NET lies in its use of regulatory network topology to infer enrichments from whole-genome association statistics, and more importantly, its automatic prioritization of loci in light of inferred enrichments. We are not aware of any published method with the same features. However, one could ignore topology and simply create SNP-level annotations based on proximity to network nodes (Supplementary Notes). On the node-based annotations, there are methods to test global enrichments or local associations using GWAS summary data. Here we use Pascal[21], LDSC[10,22] and RSS-E[13] to benchmark RSS-NET through genome-wide simulations on real genotypes[23] (Methods).

We started with simulations where RSS-NET modeling assumptions were satisfied. Here we considered two genetic architectures: a sparse scenario with most SNPs being null ($\beta = 0$), and, a polygenic scenario with most SNPs being trait-associated ($\beta \neq 0$); see Supplementary Figure 1 for details. For each architecture, we simulated baseline datasets from M_0 and enrichment datasets from three patterns of M_1 (only $\theta > 0$; only $\sigma^2 > 0$; both $\theta > 0$ and $\sigma^2 > 0$), and used RSS-NET and existing methods to detect M_1 from all datasets. Figure 2 and Supplementary Figure 1 show the trade-off between false and true enrichment discoveries for each method. Existing methods tend to perform well in select settings. For example, Pascal and LDSC perform poorly when genetic signals are very sparse (Fig.2b) or weak (Supplementary Fig. 1); RSS-E performs poorly when enrichment patterns are inconsistent with

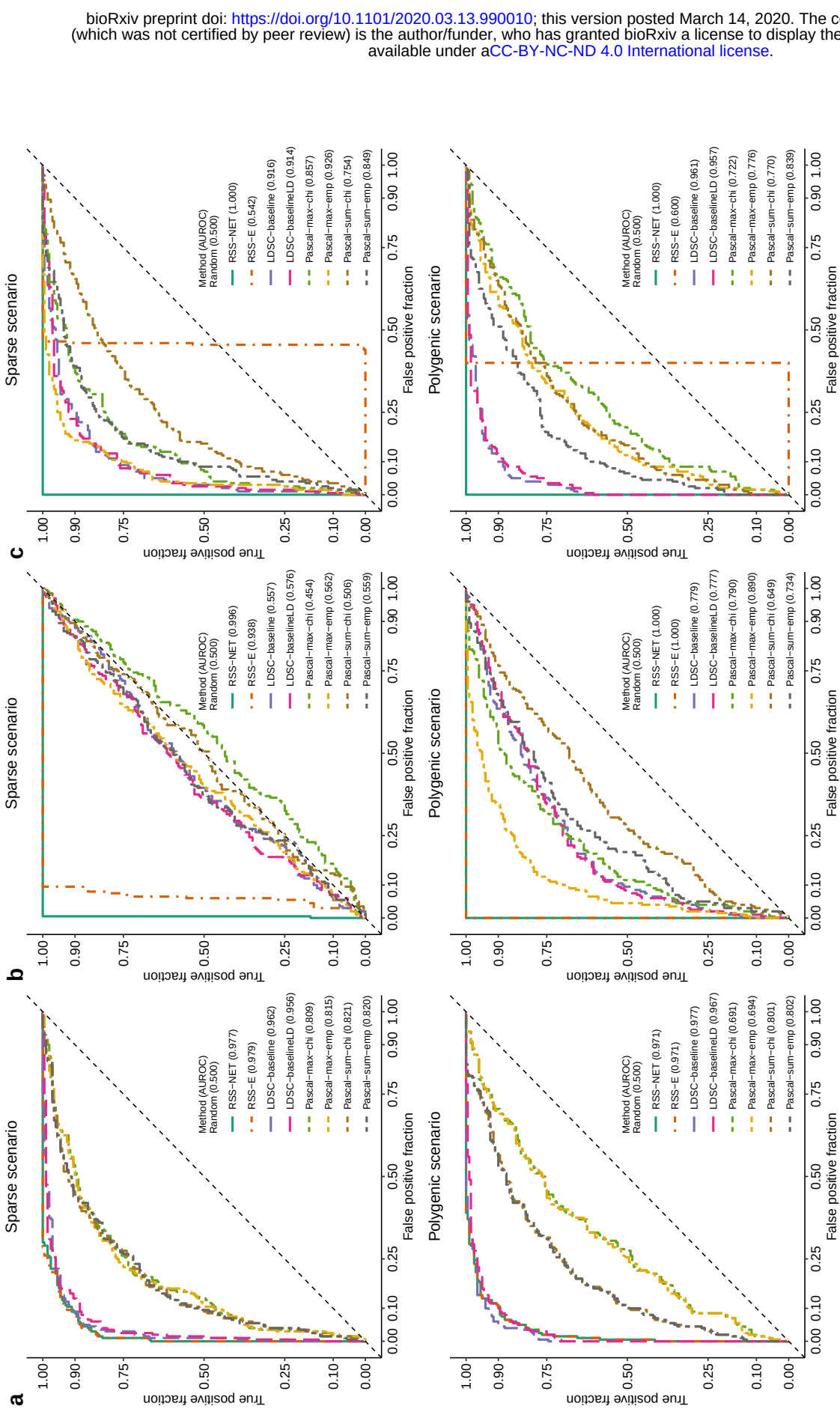

Fig 2: Flexibility of RSS-NET to identify network-based enrichments from GWAS summary statistics. We used real genotypes[23] of 348,965 genome-wide SNPs to simulate baseline and enrichment individual-level data under two genetic architectures ("sparse" and "polygenic"). The baseline data followed the baseline model (M_0: $\theta = 0$ and $\sigma^2 = 0$). The enrichment data followed the enrichment model (M_1: $\theta > 0$ or $\sigma^2 > 0$) for a target network under three scenarios: **a** $\theta > 0$, $\sigma^2 = 0$; **b** $\theta = 0$, $\sigma^2 > 0$; **c** $\theta > 0$, $\sigma^2 > 0$. We computed the corresponding single-SNP summary statistics, and, on the summary data, we compared RSS-NET with RSS-E[13], LDSC-baseline[10], LDSC-baselineLD[22] and Pascal[21] using their default setups. Pascal includes two gene scoring options: maximum-of-χ^2 (`-max`) and sum-of-χ^2 (`-sum`), and two pathway scoring options: χ^2 approximation (`-chi`) and empirical sampling (`-emp`). 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 baseline and enrichment dataset respectively. Each panel displays the trade-off between false and true enrichment discoveries, i.e. receiver operating characteristics (ROC) curves, for all methods in 200 baseline and 200 enrichment datasets of a given simulation scenario, and also reports the corresponding areas under ROC curve (AUROCs), where a higher value indicates better performance. The dashed diagonal line denotes random ROC curve, with AUROC being 0.5. Simulation details and additional results (with weaker enrichment signal) are provided in Supplementary Figure 1.

its modeling assumptions (Fig.2c). In contrast, RSS-NET performs consistently well in all scenarios. This is expected because RSS-NET models are sufficiently flexible to capture various genetic architectures and enrichment patterns. In practice, one rarely knows in advance the correct genetic or enrichment architecture. This makes the flexibility of RSS-NET appealing.

Genetic associations of complex traits are often enriched in regulatory regions[5,6,10–13,22]. Since a regulatory network is a set of genes linked by regulatory elements, it is important to confirm that network enrichments identified by RSS-NET are not driven by general regulatory enrichments. Hence, we performed simulations where baseline datasets had enriched associations in random near-gene (Fig. 3a; Supplementary Fig. 2) or regulatory SNPs (Fig. 3b; Supplementary Fig. 3). The results show that RSS-NET is unlikely to yield false discoveries due to arbitrary regulatory enrichments, and is yet more powerful than other methods.

Regulatory network edges play important roles in driving context specificity[16] and propagating disease risk[24], but existing methods largely focus on network nodes (genes). In contrast, RSS-NET leverages information from both edges and nodes. This 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, we designed a scenario where a real target network and random candidates had the same nodes and edge counts, but different edges. We simulated enrichment and baseline datasets where genetic associations were enriched in the target network 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. 3c; Supplementary Fig. 4).

To benchmark its prioritization component, we compared RSS-NET with gene-based association methods in RSS-E[13] and Pascal[21]. Figure 4 and Supplementary Figure 5 show the power of each method to identify genome-wide gene-level associations. Consistent with previous work[13], RSS-based methods substantially outperform Pascal methods even without network enrichment (Fig. 4a). This is because RSS-NET and RSS-E exploit a multiple regression framework[20] to learn the genetic architecture from data of all genes and assesses 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 methods in prioritizing genes across different enrichment patterns (Fig. 4b-d). This again highlights the flexibility of RSS-NET.

Finally, since RSS-NET uses a regulatory network as is, and, most networks to date are algorithmically inferred, we performed simulations to as-

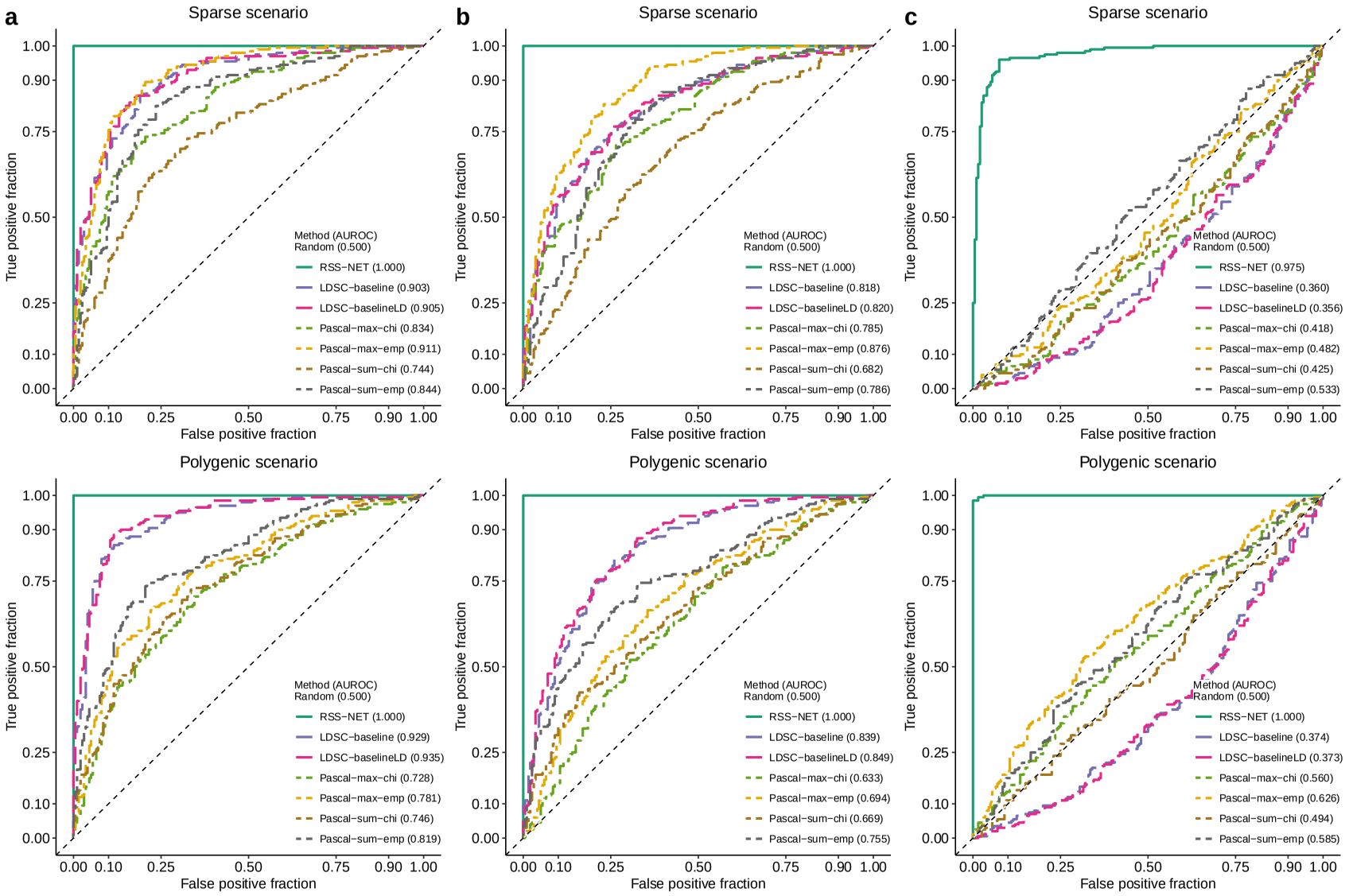

Fig 3: **Robustness of RSS-NET to model mis-specification in enrichment analysis.** Here the enrichment data were generated from a special case of M_1 with $\theta > 0$ and $\sigma^2 > 0$. The baseline data were simulated from three scenarios where genetic associations were enriched in: **a** a random set of near-gene SNPs; **b** a random set of near-RE SNPs; **c** a random edge-altered network. By this design, M_0 was mis-specified in all three scenarios. Similar to enrichment datasets, enrichments in all baseline datasets manifested in both association proportion (more frequent) and magnitude (larger effect). RSS-E was removed from this set of simulations because of its poor performance shown in Figure 2. The rest of this simulation study is the same as Figure 2. Simulation details and additional results (with weaker enrichment signal) are provided in Supplementary Figures 2-4.

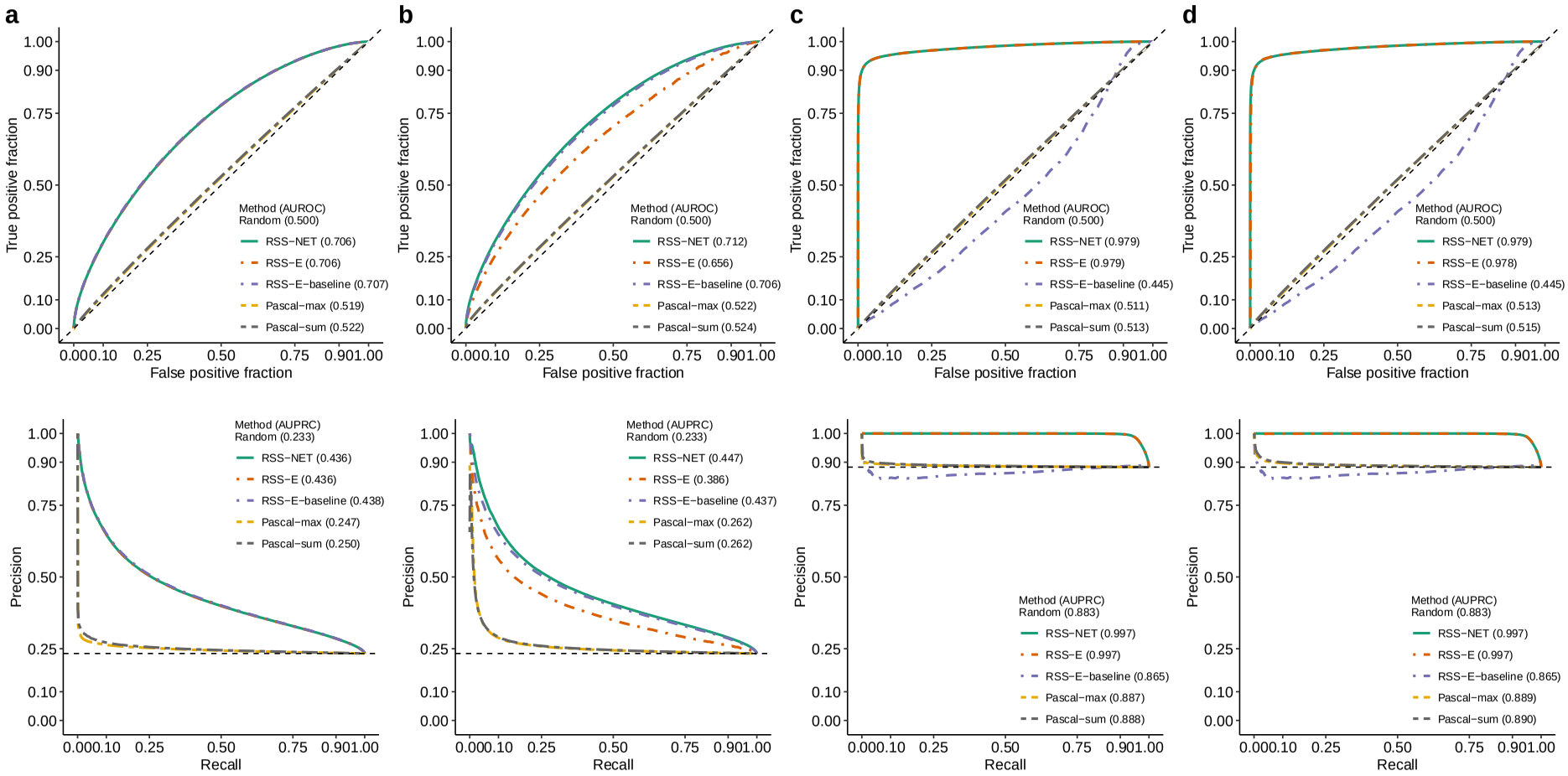

Fig 4: **Power of RSS-NET to identify gene-based associations from GWAS summary statistics.** We used real genotypes of 348,965 genome-wide SNPs[23] to simulate individual-level GWAS data under four scenarios: **a** $\theta = 0$, $\sigma^2 = 0$; **b** $\theta = 0$, $\sigma^2 > 0$; **c** $\theta > 0$, $\sigma^2 = 0$; **d** $\theta > 0$, $\sigma^2 > 0$. We computed the corresponding single-SNP summary statistics, and, on the summary data, we compared RSS-NET with gene-based association methods available in RSS-E[13] and Pascal[21]. 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-χ^2 (`-max`) and sum-of-χ^2 (`-sum`). Each scenario contains 200 datasets, each with 16,954 autosomal protein-coding genes. For each dataset, we defined a gene as "trait-associated" if at least one SNP within 100 kb of the transcribed region of this gene had nonzero effect. For each gene in each dataset, RSS methods produced posterior probabilities that the gene was trait-associated, whereas Pascal methods produced association p-values; these statistics were used to rank the significance of gene-level associations. The upper half of each panel displays ROC curves and AUROCs for all methods, with dashed diagonal lines indicating random performance (AUROC=0.5). The lower half 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 AUROC and AUPRC, higher value indicates better performance. Simulation details and additional results are provided in Supplementary Figure 5.

sess 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 used the noisy networks, rather than the real one, in RSS-NET analyses. 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. 6).

In conclusion, RSS-NET is flexible to perform well in various genetic architectures and enrichment patterns, is robust to a wide range of model misspecification, and 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[17] whole-genome regulatory networks for 38 tissue or cell types, using public paired data of gene expression and chromatin accessibility (Methods; Supplementary Table 1). Clustering analysis showed that networks recapitulated context similarity, with immune cells and brain regions grouping together as two single units (Fig. 5a; Supplementary Fig. 7).

On these 38 networks, we then applied RSS-NET to analyze 1.1 million common SNPs[25] for 18 traits, using GWAS summary statistics from 20,883 to 253,288 European ancestry individuals (Supplementary Table 2). For each trait-network pair we computed a BF assessing network enrichment. Full results of 684 trait-network pairs are available online (Methods).

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 edge, and then 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 log10 BF increase: 13.94, one-sided $p = 5.1 \times 10^{-15}$), and, 16 out of 18 traits had multiple networks more enriched than the near-gene control (minimum 5; one-sided Wilcoxon $p = 1.2 \times 10^{-4}$; Supplementary Table 3). Consistent with simulations (Fig. 3a-b), these results indicate that network enrichments identified by RSS-NET are unlikely due to generic regulatory 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 associated the BFs with

three network features and did not observe any correlation (proportion of SNPs in a network: $r = -3.0 \times 10^{-2}$, $p = 0.49$; 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 (e.g., promoter), we computed the correlation between BFs and proportions of SNPs falling into a network and each of 73 functional categories[10,22], 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 Tables 4-5). Together, these results suggest that observed network enrichments are unlikely to be driven by known features and the resulting model mis-specification.

For each trait-network pair, we also computed BFs comparing the baseline (M_0) against three disjoint models where enrichments were contributed by (1) only nodes ($M_{11} : \theta > 0, \sigma^2 = 0$); (2) only edges ($M_{12} : \theta = 0, \sigma^2 > 0$); (3) both nodes and 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 $p = 1.2 \times 10^{-45}$), corroborating the "omnigenic" hypothesis[24,26] that genetic signals of complex traits tend to be spread across the genome via regulatory interconnections. When stratifying results by traits, however, we observed that enrichment patterns could vary considerably (Fig. 5b; Supplementary Table 6). For type 2 diabetes (T2D), two of five networks passing the near-gene enrichment control showed the strongest support for node-only enrichment (M_{11}). Many networks showed the strongest support for edge-only enrichment (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 the restricted models (M_{11}, M_{12}, M_{13}) as special cases, we used BFs based on M_1 in this study. Collectively, these results highlight the heterogeneity of network enrichments across complex 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 = 2.07×10^{13}), bowel (BF = 8.02×10^{12}) and adipose (BF = 4.73×10^{12}), consistent with the roles of obesity-related genes in insulin biology and energy metabolism. Networks of immune cells showed enrichments for rheumatoid arthritis (RA, BF = 2.95×10^{60}), inflammatory bowel disease (IBD, BF = 5.07×10^{35}) and Alzheimer's disease (BF = 8.31×10^{26}). Networks of cardiac and other muscle tissues showed enrichments for coronary artery disease (CAD, BF = 9.78×10^{28}), atrial fibrillation (AF, BF = 8.55×10^{14}), and heart rate (BF = 2.43×10^{7}). Other examples include brain network with neuroticism (BF = 2.12×10^{19}), and, liver network with high-

density lipoprotein (HDL, BF = 2.81×10^{21}) and low-density lipoprotein (LDL, BF = 7.66×10^{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 = 3.95×10^{13}), LDL (BF = 5.18×10^{30}) and T2D (BF = 1.49×10^{77}). This result supports a recent mouse study[27] 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[28]. Other examples include adipose network with CAD[29] (BF = 1.67×10^{29}), liver network with Alzheimer's disease[13,30] (BF = 1.09×10^{20}) and monocyte network with AF[31,32] (BF = 4.84×10^{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 (Fig. 5c; Supplementary Table 7). Among 153 trait pairs, 29 of them were significantly correlated ($p < 0.05/153$). Reassuringly, subtypes of the same disease showed strongly correlated enrichments, as in IBD subtypes ($r = 0.96$, $p = 1.3 \times 10^{-20}$) and CAD subtypes ($r = 0.90$, $p = 3.3 \times 10^{-14}$). The results also recapitulated known genetic correlations including RA with IBD[33] ($r = 0.79$, $p = 5.3 \times 10^{-9}$), and, neuroticism with schizophrenia[34] ($r = 0.73$, $p = 1.6 \times 10^{-7}$). Network enrichments of CAD were correlated with enrichments of its established risk factors such as heart rate[35] ($r = 0.75$, $p = 5.1 \times 10^{-8}$) and BMI[36] ($r = 0.71$, $p = 5.1 \times 10^{-7}$), and its associated traits such as AF[37] ($r = 0.65$, $p = 9.2 \times 10^{-6}$) and height[38] ($r = 0.64$, $p = 1.6 \times 10^{-5}$). Network enrichments of Alzheimer's disease were strongly correlated with 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 recent data linking Alzheimer's disease to lipid metabolism[39] and immune activation[40]. The results show the potential of GWAS to highlight trait similarities via regulatory networks, complementing previous work via SNPs[41], heritabilities[42] and pathways[13].

Enrichment-informed prioritization of network genes. A key feature of RSS-NET, inherited from RSS-E[13], is that inferred network enrichments automatically contribute to prioritize genetic associations of network genes. Specifically, for each locus RSS-NET produces P_1^{base}, P_1^{near} and P_1^{net}, the posterior probability 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 (Method). When multiple networks are enriched, RSS-NET produces P_1^{bma} by averaging P_1^{net} over all networks passing the near-gene control, weighted by their BFs (Method). This allows us to assess

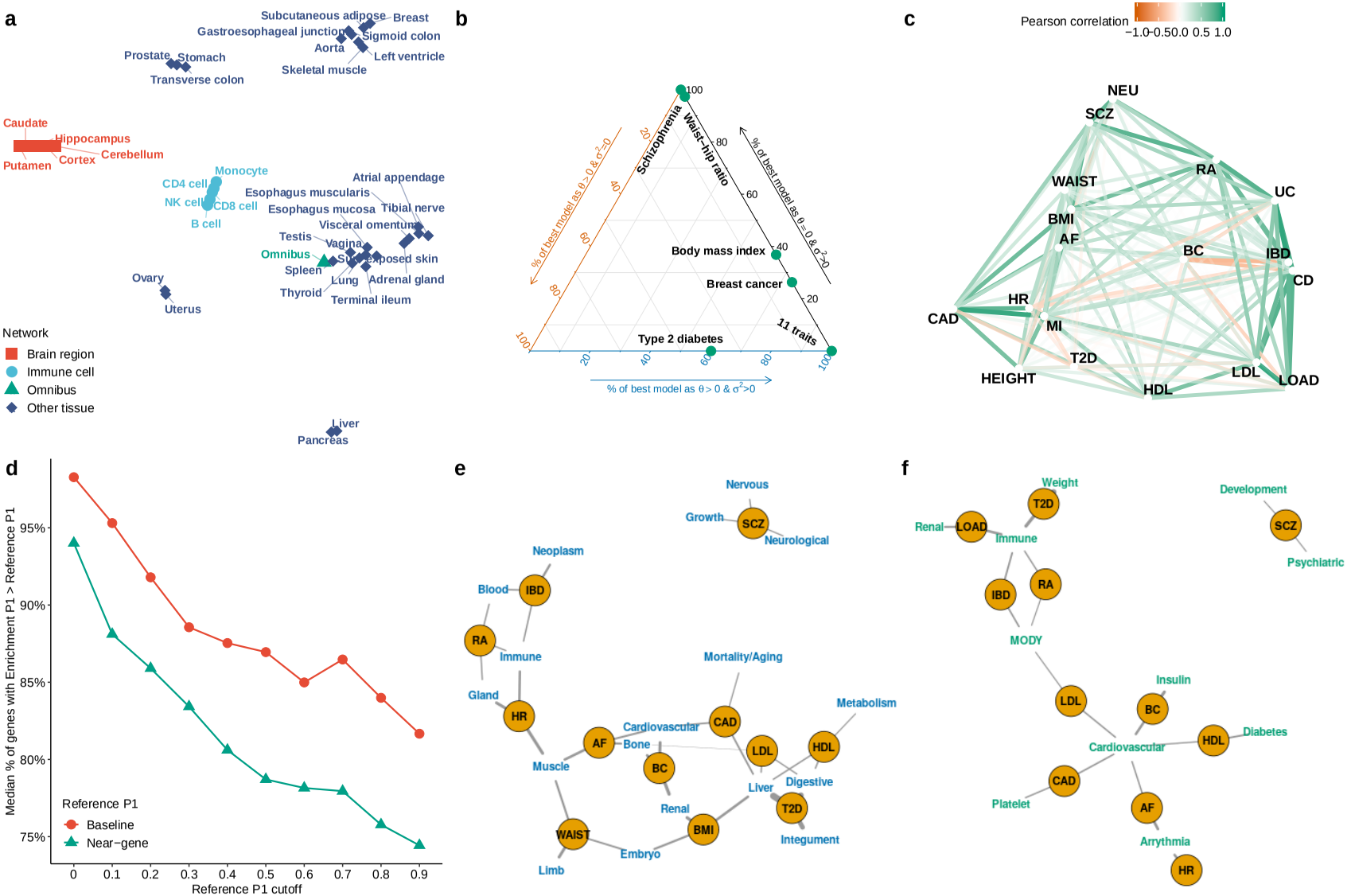

Fig 5: **RSS-NET analyses of GWAS summary statistics for 18 complex traits across 38 regulatory networks. a** Clustering of 38 regulatory networks based on t-distributed stochastic neighbour embedding. Analysis details and additional results are provided in Supplementary Figure 7. **b** Ternary diagram showing, for each trait, percentages of the "best" enrichment model (showing the largest BF) as the node-only ($M_{11}: \theta > 0, \sigma^2 = 0$), edge-only ($M_{12}: \theta = 0, \sigma^2 > 0$) and node-and-edge ($M_{13}: \theta > 0, \sigma^2 > 0$) model across networks. See Supplementary Table 6 for numerical values. Shown are 16 traits that had multiple networks more enriched than the near-gene control. **c** Network plot of BF correlations among 18 traits across 38 networks, where traits with stronger correlations are located closer and connected by darker edges. See Supplementary Table for correlation estimates. **d** 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 that had multiple networks more enriched than the near-gene control. See Supplementary Table 8 for numerical values. **e-f** Overlap of RSS-NET prioritized genes ($P_1^{\text{bma}} \geq 0.9$) with genes implicated in knockout mouse phenotypes[43] (**e**) and human Mendelian diseases[44,45] (**f**). An edge indicates that a category of knockout mouse or Mendelian genes is significantly enriched for genes prioritized for a GWAS trait (FDR ≤ 0.1). Thicker edges correspond to stronger enrichment odds ratios. To simplify visualization, only top-ranked categories are shown for each trait (**e**: 3; **f**: 2). See Supplementary Tables 12-13 for full results. Abbreviations of GWAS traits are defined in Supplementary Table 2.

genetic associations in light of enrichment without having to select a single enriched network. Differences in estimates based on enrichment (P_1^{net} or P_1^{bma}) and reference (P_1^{base} or P_1^{near}) reflect the enrichment impact 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^{bma} than reference estimates (P_1^{base} or P_1^{near}), among genes with reference P_1 passing a given cutoff (Fig. 5d). When using P_1^{base} as reference, we observed high proportions of genes with $P_1^{bma} > P_1^{base}$ (median: $82-98\%$) across a wide range of P_1^{base}-cutoffs ($0-0.9$), and as expected, the improvement decreased as the reference cutoff increased. When using P_1^{near} as reference, we observed less genes with improved P_1 than using P_1^{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 $P_1^{bma} > P_1^{near}$ remained high (median: $74-94\%$) nonetheless. Similar patterns occurred when we repeated the analysis with P_1^{net} across 512 trait-network pairs (Supplementary Table 8). 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 $P_1^{net} > P_1^{near}$ in a network is often positively correlated with its enrichment BF ($r: 0.28-0.91$; Supplementary Table 9). When a gene belongs to multiple networks, its highest P_1^{net} often occurs in the top-enriched networks. We illustrate this coherent pattern with *MT1G*, a liver-active[7] gene that was prioritized for HDL by RSS-NET and also implicated in a recent multi-ancestry genome-wide sleep-SNP interaction analysis of HDL[46]. Although *MT1G* belongs to regulatory networks of 18 contexts, only the top enrichment in liver (BF $= 2.81 \times 10^{21}$) informs a strong association between *MT1G* and HDL ($P_1^{net} = 0.98$), and remaining networks with weaker enrichments yield minimal improvement ($P_1^{base} = 0.10$, $P_1^{net}: 0.14-0.19$). Additional examples are shown in Figure 6.

RSS-NET recapitulates many genes previously implicated in the same GWAS. For each analyzed dataset we downloaded the corresponding genes from the GWAS Catalog[47] and computed the proportion of these genes that had high P_1^{bma}. With a stringent cutoff 0.9, we observed a significant overlap (median across traits: 69%; median Fisher exact $p = 1.24 \times 10^{-26}$; Supplementary Table 10). Reassuringly, many recapitulated genes are well-established for the traits (Supplementary Table 11), 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 new associations are

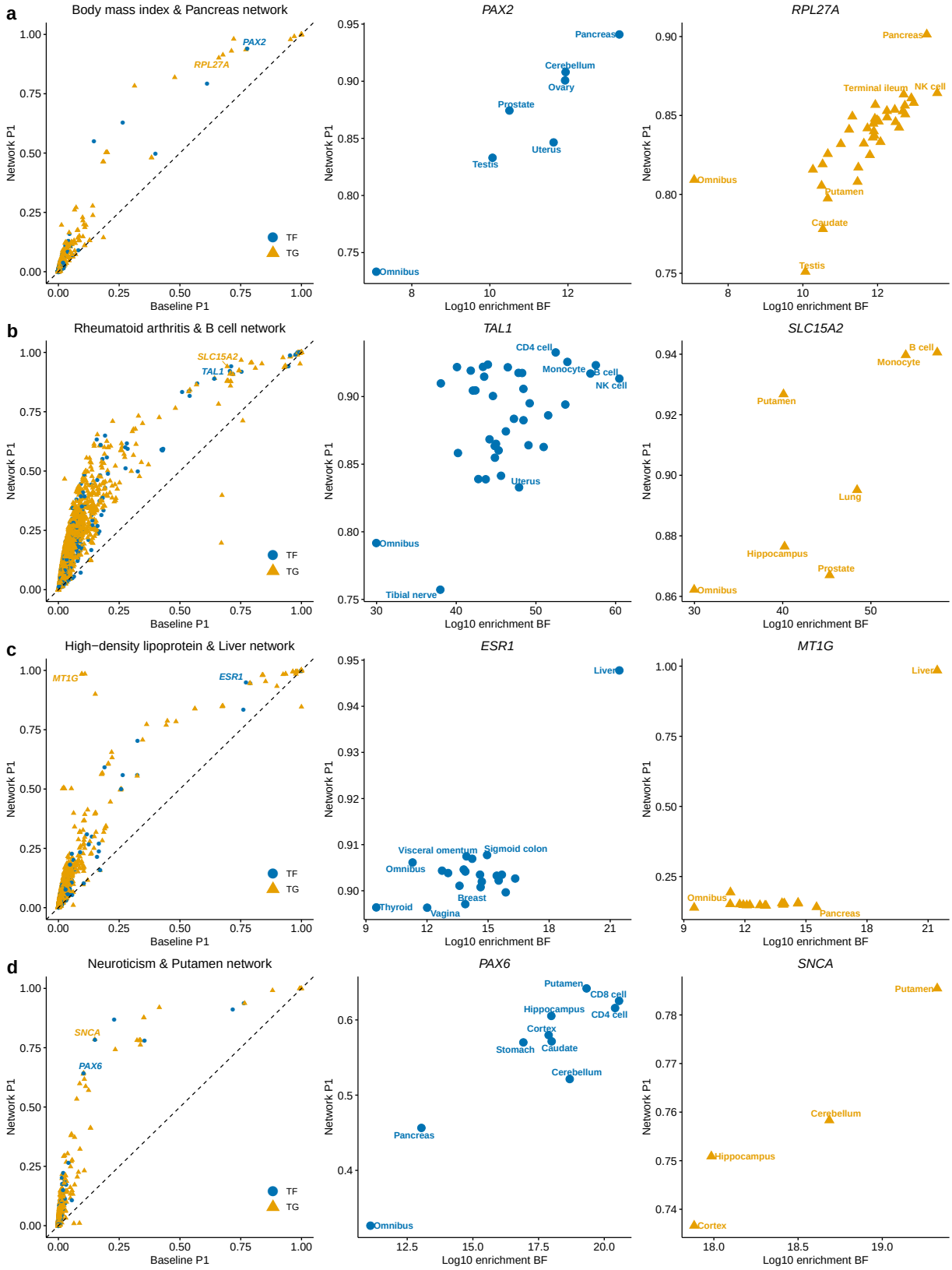

Fig 6: **RSS-NET gene prioritization results of select trait-network pairs.** In the left column, each dot represents a member gene of a given network. In the center and right columns, each dot represents a network to which a select gene belongs. Numerical values are available online (Methods).

Trait	Gene (Role)	P_1^{base}	P_1^{near}	P_1^{bma}	P_1^{net} (Network, BF)	Mouse trait	Therapeutic/clinical evidence
BMI	*PAX2* (TF)	0.78	0.80	0.94	0.94 (Pancreas, 2.07×10^{13})	Eye, Renal	FSGS7, PAPRS
	FLT3 (TG)	0.61	0.70	0.85	0.85 (Cerebellum, 8.70×10^{11})	Growth, Immune	Acute myeloid leukemia
WAIST	*LAMB1* (TG)	0.97	0.97	0.98	0.98 (Esophagus, 6.78×10^{239})	Neuron, NS	Lissencephaly 5
BC	*KCTD1* (TG)	0.89	0.93	0.98	0.98 (Heart, 8.08×10^{7})	CS	Scalp-ear-nipple syndrome
	CASP8 (TG)	0.71	0.72	0.94	0.94 (Aorta, 8.27×10^{8})	Growth, Immune	HCC, Glionitrin A
RA	*AIRE* (TF)	0.54	0.61	0.84	0.84 (B cell, 3.31×10^{57})	Immune	APS1
IBD	*LPP* (TG)	0.98	0.94	0.99	0.99 (Monocyte, 6.28×10^{31})	Cellular	Acute myeloid leukemia
	FOXP1 (TF)	0.84	0.78	0.95	0.95 (NK cell, 5.07×10^{35})	Immune, Neuron	Language impairment
	CCND3 (TG)	0.81	0.89	0.95	0.95 (NK cell, 5.07×10^{35})	Immune	
HDL	*ALOX5* (TG)	0.97	0.97	0.99	0.99 (Monocyte, 4.75×10^{15})	Immune, Metab.	Atherosclerosis
	GPAM (TG)	0.92	0.95	0.98	0.98 (Liver, 2.81×10^{21})	Liver, Metab.	
	NR0B2 (TG)	0.84	0.93	0.98	0.98 (Liver, 2.81×10^{21})	Growth, Metab.	Early-onset obesity
LDL	*CERS2* (TG)	0.99	0.99	1.00	1.00 (NK cell, 5.18×10^{30})	Liver, Metab.	
	ABCA1 (TG)	0.98	0.98	0.99	0.99 (Liver, 7.66×10^{27})	Liver, Metab.	Tangier disease, Probucol
	ABCB11 (TG)	0.68	0.72	0.88	0.88 (Liver, 7.66×10^{27})	Liver, Metab.	Cholestasis BRI2, PFI2
	DLG4 (TG)	0.69	0.59	0.85	0.85 (NK cell, 5.18×10^{30})	Metab., NS	Tat-NR2B9c
	SOX17 (TF)	0.52	0.65	0.82	0.84 (CD8, 5.86×10^{28})	Liver, Metab.	Vesicoureteral reflux 3
CAD	*TGFB1* (TG)	0.92	0.99	0.99	0.99 (Adipose, 1.67×10^{29})	CS, Growth	Camurati-Engelmann disease
	FN1 (TG)	0.58	0.79	0.91	0.92 (GEJ, 9.78×10^{28})	CS, Metab.	GFND2, SMDCF
	CDH13 (TG)	0.31	0.55	0.77	0.82 (Heart, 1.93×10^{28})	CS, Metab.	
	EDNRA (TG)	0.57	0.79	0.80	0.82 (Aorta, 1.09×10^{27})	CS, Muscle	Ambrisentan, Macitentan
AF	*SCN5A* (TG)	0.87	0.92	1.00	1.00 (Heart, 6.89×10^{12})	CS, Muscle	Brugada syndrome 1, FAF 10
	ENPEP (TG)	0.50	0.76	0.92	0.94 (Uterus, 2.71×10^{11})		QGC-001
	ATXN1 (TG)	0.45	0.62	0.90	0.90 (Colon, 7.54×10^{14})	Muscle, NS	Spinocerebellar ataxia 1
	MYOT (TG)	0.55	0.66	0.86	0.87 (Muscle, 8.55×10^{14})		Spheroid body myopathy, MFM3
SCZ	*FOXP1* (TF)	1.00	1.00	1.00	1.00 (Colon, 1.20×10^{144})	Growth, Neuron	Mental retardation
	BCL11A (TG)	1.00	1.00	1.00	1.00 (Spleen, 1.44×10^{141})	Immune, NS	Dias-Logan syndrome
	SLC25A12 (TG)	0.79	0.81	0.88	0.88 (Muscle, 4.99×10^{127})	Neuron, NS	EIEE39
NEU	*TCF4* (TF)	0.72	0.88	0.95	0.95 (CD8, 3.66×10^{20})	Immune, NS	Pitt-Hopkins syndrome
	RAPSN (TG)	0.77	0.88	0.93	0.93 (Muscle, 8.20×10^{17})	Muscle, NS	CMS11
	MEF2C (TF)	0.15	0.40	0.83	0.83 (Ileum, 8.56×10^{22})	Growth, Neuron	Mental retardation 20
	SNCA (TG)	0.15	0.32	0.78	0.79 (Putamen, 2.12×10^{19})	Neuron, NS	DLB, Parkinson 1, 4, BIIB054
	PAX6 (TF)	0.10	0.22	0.62	0.64 (Putamen, 2.12×10^{19})	NS, Vision	Optic nerve hypoplasia
	PCLO (TG)	0.06	0.17	0.63	0.63 (Ileum, 8.56×10^{22})	Growth, NS	Pontocerebellar hypoplasia 3

TABLE 1

Examples of RSS-NET highlighted genes that were not reported in GWAS of the same data ($p \geq 5 \times 10^{-8}$) but were implicated in later GWAS with increased sample sizes ($p < 5 \times 10^{-8}$). The "mouse trait" column is based on the Mouse Genome Informatics[43]. The "therapeutic/clinical evidence" column is based on the Online Mendelian Inheritance in Man[44] and Therapeutic Target Database[48]. Click blue links to view details online. Drugs are highlighted in yellow. Abbreviations of GWAS traits are defined in Supplementary Table 2. GEJ: gastroesophageal junction; CS: cardiovascular system; DS: digestive/alimentary system; Metab.: metabolism; NS: nervous system.

potentially real we exploited 15 analyzed traits that each had an updated GWAS with larger sample size. In each case we obtained newly mapped genes from the GWAS Catalog[47] 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 Fisher exact $p = 1.93 \times 10^{-5}$; Supplementary Table 10), showing the potential of RSS-NET to identify trait-associated 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. A particular example is *NR0B2*, a liver-active[7] gene prioritized for HDL (BF = 2.81×10^{21}, $P_1^{\text{base}} = 0.84$, $P_1^{\text{net}} = 0.98$), which was not identified by standard GWAS[49] of the same data (minimum single-SNP $p = 1.4 \times 10^{-7}$ within 100kb, $n = 99{,}900$). *NR0B2* is associated with various mouse lipid traits[50–52] and human obesity[53], and was later identified in a GWAS of HDL[54] with larger sample size ($p = 9.7 \times 10^{-16}$, $n = 187{,}056$). Table 1 lists additional examples.

Biological and clinical relevance of prioritized genes. Despite significant overlaps with GWAS-implicated genes, a large fraction of RSS-NET prioritized genes ($P_1^{\text{bma}} \geq 0.9$) were not identified by GWAS (median: 70%; Supplementary Table 10). To systematically assess their relevance, we cross-referenced these genes with multiple orthogonal databases.

Mouse phenomics provides important resources to study genetics of human traits[55]. Here we evaluated overlap between RSS-NET prioritized genes and genes implicated in 27 categories of knockout mouse phenotypes[43]. Network-informed genes ($P_1^{\text{bma}} \geq 0.9$) were significantly enriched in 128 mouse-human trait pairs (FDR ≤ 0.1; Supplementary Table 12). 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. 5e). Schizophrenia-associated genes 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 associated with heart conditions (OR: 2.45 – 2.92). The biliary system showed strong enrichments of genes associated with lipids, BMI, CAD and T2D (OR: 2.16 – 10.78). The cross-species phenotypically matched enrichments strengthen the biological relevance of RSS-NET results.

Mendelian disease-causing genes have been recognized as an vital contributor to complex traits[56,57]. Here we quantified overlap between RSS-NET prioritized genes and genes causing 19 categories[45] of Mendelian disor-

Trait	Gene (Role)	P_1^{base}	P_1^{near}	P_1^{bma}	P_1^{net} (Network, BF)	Mouse trait	Therapeutic/clinical evidence
BMI	*NEXN* (TG)	0.71	0.79	0.89	0.90 (Muscle, 9.31×10^{12})	CS, Muscle	Cardiomyopathy D1CC, H20
	CDX2 (TF)	0.61	0.70	0.83	0.86 (NK cell, 3.95×10^{13})	DS, Growth	
WAIST	*BSCL2* (TG)	0.80	0.68	0.87	0.87 (Esophagus, 6.78×10^{239})	Adipose, Growth	Lipodystrophy CG2
	FOXP2 (TF)	0.56	0.59	0.73	0.73 (Esophagus, 6.78×10^{239})	Growth, NS	Speech-language disorder 1
BC	*ADSL* (TG)	0.76	0.80	0.91	0.92 (Aorta, 8.27×10^{8})	CS, Eye	Adenylosuccinase deficiency
	SYNE1 (TG)	0.57	0.63	0.89	0.90 (Esophagus, 6.30×10^{7})	Growth, Muscle	AMCM, EDMD4, SCAR8
RA	*TAL1* (TF)	0.71	0.79	0.91	0.93 (CD4, 3.02×10^{52})	Immune, Tumor	Acute lymphocytic leukemia
	FHIT (TG)	0.30	0.60	0.90	0.91 (CD4, 3.02×10^{52})	Immune, Tumor	
	FLT3 (TG)	0.33	0.57	0.73	0.73 (B cell, 3.31×10^{57})	Immune, Tumor	Acute myeloid leukemia
IBD	*FHIT* (TG)	0.63	0.87	0.95	0.95 (CD4, 5.32×10^{33})	Immune, Tumor	
	GATA3 (TF)	0.85	0.83	0.94	0.94 (NK cell, 5.07×10^{35})	Immune, Renal	Barakat syndrome
	RORA (TF)	0.66	0.78	0.87	0.90 (B cell, 1.49×10^{32})	Immune, NS	IDDECA
	NFKB2 (TF)	0.74	0.85	0.84	0.88 (B cell, 1.49×10^{32})	Immune	CVID10, DIMS-0150
	LRBA (TG)	0.42	0.58	0.72	0.72 (NK cell, 5.07×10^{35})	Immune	Immunodeficiency CV8
	DOCK2 (TG)	0.38	0.53	0.71	0.71 (NK cell, 5.07×10^{35})	Immune	Immunodeficiency 40
HDL	*MT1G* (TG)	0.10	0.09	0.98	0.98 (Liver, 2.81×10^{21})	CS, Metab.	
	RETSAT (TG)	0.79	0.80	0.95	0.95 (Liver, 2.81×10^{21})	Adipose, Metab.	
	ESR1 (TF)	0.77	0.82	0.95	0.95 (Liver, 2.81×10^{21})	CS, Metab.	Myocardial infarction
	HCAR3 (TG)	0.85	0.85	0.92	0.92 (Monocyte, 4.75×10^{15})	Metab.	ARI-3037MO
	TNNC1 (TG)	0.48	0.45	0.78	0.78 (Liver, 2.81×10^{21})	CS, Muscle	CMD1Z, CMH13, Levosimendan
LDL	*RAF1* (TG)	0.79	0.83	0.90	0.90 (Aorta, 3.71×10^{27})	CS, Immune	CMD1NN, Semapimod
	APOA1 (TG)	0.70	0.76	0.90	0.90 (Liver, 7.66×10^{27})	CS, Metab.	Amyloidosis, HDL deficiency
	ACADVL (TG)	0.69	0.59	0.85	0.85 (NK cell, 5.18×10^{30})	Liver, Metab.	VLCAD deficiency
T2D	*ITGB6* (TG)	0.75	0.99	0.99	0.99 (Ileum, 4.52×10^{62})	Immune, Metab.	AI1H
HR	*TKT* (TG)	0.65	0.67	0.92	0.93 (Aorta, 2.43×10^{7})	CS, Growth	SDDHD
CAD	*OSM* (TG)	0.56	0.78	0.86	0.86 (Aorta, 1.09×10^{27})	Immune, Metab.	GSK2330811
	TRIB1 (TG)	0.43	0.68	0.85	0.85 (Adipose, 1.67×10^{29})	Adipose, Metab.	
	TAB2 (TG)	0.19	0.43	0.61	0.61 (CD8, 1.13×10^{25})	CS	Congenital heart defects
AF	*TPMT* (TG)	0.88	0.93	0.99	0.99 (Ileum, 4.43×10^{13})	Metab.	THPM1
	RUNX1 (TF)	0.44	0.60	0.88	0.89 (Heart, 2.15×10^{14})	CS, Immune	Acute myeloid leukemia, FPDMM
	CSF3 (TG)	0.56	0.72	0.88	0.88 (Muscle, 8.55×10^{14})	Blood, Immune	Interleukin-3
LOAD	*CASP2* (TG)	0.99	1.00	1.00	1.00 (CD8, 8.31×10^{26})	Cellular, NS	Caspase-2
	TTR (TG)	0.64	0.92	0.94	0.94 (Pancreas, 3.53×10^{20})	Metab.	FAP, Inotersen , Patisiran
SCZ	*RORA* (TF)	1.00	1.00	1.00	1.00 (Cortex, 5.39×10^{128})	Neuron, NS	IDDECA
	ERBB4 (TG)	1.00	1.00	1.00	1.00 (Putamen, 7.22×10^{116})	Neuron, NS	ALS19
	NFIB (TF)	0.97	0.97	0.98	0.98 (Cortex, 5.39×10^{128})	NS	MACID
	GRIK2 (TG)	0.90	0.94	0.97	0.97 (Cerebellum, 3.15×10^{129})	Neuron, NS	Mental retardation 6
	SYT1 (TG)	0.84	0.89	0.93	0.93 (Cerebellum, 3.15×10^{129})	Neuron, NS	Baker-Gordon syndrome
	ESR1 (TF)	0.80	0.84	0.93	0.93 (Colon, 1.07×10^{141})	Neuron, NS	Migraine
	NTRK2 (TG)	0.78	0.84	0.91	0.91 (Cerebellum, 3.15×10^{129})	Neuron, NS	EIEE58
	LRRK2 (TG)	0.73	0.78	0.86	0.86 (Monocyte, 5.85×10^{131})	Neuron, NS	Parkinson 8, DNL151 , DNL201
	C9orf72 (TG)	0.74	0.78	0.83	0.83 (Spleen, 1.44×10^{141})	Neuron, NS	FTDALS1
	SNCA (TG)	0.60	0.66	0.74	0.74 (Cerebellum, 3.15×10^{129})	Neuron, NS	DLB, Parkinson 1, 4
NEU	*LMBRD1* (TG)	0.42	0.66	0.94	0.94 (Ileum, 8.56×10^{22})	Metab.	MAHCF
	PRKCQ (TG)	0.36	0.56	0.90	0.91 (Spleen, 2.13×10^{19})	Immune, NS	
	ATP1A2 (TG)	0.33	0.39	0.76	0.78 (Putamen, 2.12×10^{19})	Neuron, NS	AHC1, FHM2

TABLE 2

Examples of RSS-NET highlighted genes that have not reached genome-wide significance in the GWAS Catalog[47] ($p \geq 5 \times 10^{-8}$) at the time of analysis. The rest is the same as Table 1.

ders[44]. Leveraging regulatory networks ($P_1^{bma} \geq 0.9$), we observed 47 significantly enriched Mendelian-complex trait pairs (FDR ≤ 0.1; 44 for $P_1^{near} \geq 0.9$; 31 for $P_1^{base} \geq 0.9$; Supplementary Table 13), among which the top-ranked ones were often phenotypically matched (Fig. 5f). Schizophrenia-associated genes were strongly enriched in Mendelian development and psychiatric disorders (OR: 2.22 – 2.23). Genes prioritized for atrial fibrillation 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 associated with 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 included Alzheimer's disease with immune dysregulation[40] (OR = 7.32) and breast cancer with insulin disorders[58] (OR = 9.71). The results corroborate that Mendelian and complex traits are not dichotomous, but rather exist on a continuum.

Human genetics has proven valuable in therapeutic development for validating molecular targets[59] and improving success rates[60]. To evaluate their potential in drug discovery, we examined whether RSS-NET prioritized genes are pharmacologically active targets with known clinical indications[48]. We identified genes with perfectly matched drug indications and GWAS traits. The most illustrative identical match is *EDNRA*, a gene that is prioritized for CAD ($P_1^{base} = 0.57$, $P_1^{net} = 0.82$ for aorta network), and is also a successful target of approved drugs for cardiovascular diseases (Table 1). We identified genes with closely related drug indications and GWAS traits. For example, gene *TTR* is prioritized for Alzheimer ($P_1^{base} = 0.64$, $P_1^{bma} = 0.94$), and is 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, gene *HCAR3* is prioritized for HDL ($P_1^{base} = 0.85$, $P_1^{bma} = 0.92$), and is 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). We also found mismatches between drug indications and GWAS traits, which could suggest drug repurposing opportunities[61]. For example, gene *CSF3* is prioritized for AF ($P_1^{base} = 0.56$, $P_1^{bma} = 0.88$), and is also a successful target of an approved drug for aplastic anemia (AA). Since *CSF3* is associated with various blood cell traits in mouse[62] and human[63], and inflammation plays a role in both AA and AF etiology[32,64], it is tempting to assess effects of the approved AA drug on AF. Overall, further evaluations are required to mechanistically understand the prioritized therapeutic genes, but the findings could be a useful basis for future studies.

Discussion

We have presented RSS-NET, a new topology-aware method for integrative analysis of regulatory networks and GWAS summary data. We have demonstrated the improvement of RSS-NET over existing methods through a wide variety of simulations, and illustrated its potential to yield novel insights via extensive analyses of 38 networks and 18 traits. With multi-omics integration becoming a routine in modern GWAS, we expect that researchers will find RSS-NET and its open-sourced software useful.

Compared with existing integrative approaches, RSS-NET has several key strengths. First, unlike many methods that require loci passing a significance threshold[9,14,65], 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 (θ) and sizes (σ^2) of SNP-level associations, and thus bypasses the issue of converting SNP-level GWAS summary data to gene-level data[14,15,21]. Third, RSS-NET inherits from RSS-E[13] 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[10,12,15,19,22]. Fourth, compared with RSS-E[13], RSS-NET makes more flexible modeling assumptions, and thus is more adaptive to unknown genetic and enrichment architectures.

RSS-NET provides a new view of complex trait genetics through the lens of regulatory topology. Complementing previous connectivity analyses[14–16,19,24], RSS-NET highlights a consistent pattern where genetic signals of complex traits often distribute across genome via the regulatory topology. RSS-NET further leverages topology enrichments to enhance trait-associated gene discovery. The topology awareness of RSS-NET relies on a novel model that decomposes effect size of a single SNP into effects of multiple (cis or trans) genes through a regulatory network. Other than similar perspective in a recent theory paper[26], we are not aware of any published work implementing and evaluating the topology-aware model in practice.

RSS-NET depends critically on the quality of input regulatory networks. The more accurate networks are, the better performance RSS-NET achieves. Currently our understanding of regulatory networks remains incomplete, and most of available networks are algorithmically constructed[14–17]. Artifact nodes and edges of 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. As more accurate networks become available, 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 single causal variants within prioritized loci. For this task we recommend using off-the-shelf fine-mapping methods [66]. Second, RSS-NET analyzes a single 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 size in each network, weighted by network relevance for a disease. Third, RSS-NET does not leverage known genomic annotations, either at the level of SNPs [10,19,22] or genes [11–13]. Although our mis-specification simulations and near-gene control analyses have confirmed that RSS-NET is robust to generic enrichments of known features, accounting for known annotations can help interpret observed network enrichments [19]. Our preliminary experiments, however, showed that incorporating additional networks or annotations in RSS-NET increased computation costs. Hence, we view the development of more efficient multi-network, multi-annotation methods as an important direction for future work.

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, accessed January 3, 2019). Simulations used 348,965 genome-wide SNPs [23] (https://www.wtccc.org.uk), and data analyses used 1,289,786 genome-wide HapMap3 [25] SNPs (https://data.broadinstitute.org/alkesgroup/LDSCORE/w_hm3.snplist.bz2, accessed November 27, 2018). This study also excluded SNPs on sex chromosomes, SNPs with minor allele frequency less than 1%, and SNPs in the human leukocyte antigen region.

GWAS summary statistics and LD estimates. The GWAS summary statistics [49,67–79] (Supplementary Table 2) and LD estimates [80] used in the present study were processed in the same way as those in our previous work [13]. Data download links are provided in Supplementary Notes.

Gene regulatory networks. We inferred 38 regulatory networks from paired high-throughput sequencing data of gene expression (e.g., RNA-seq) and chromatin accessibility (e.g., DNAse-seq or ATAC-seq), using a regression-based method [17]. We first constructed an "omnibus" network from paired data

of all available biological samples, and then reorganized this "omnibus" network in light of regulatory elements (REs) identified[81] in each context to generate context-specific networks for 5 immune cell types, 5 brain regions and 27 non-brain tissues. The network-construction software is available at https://github.com/suwonglab/PECA. The 38 networks are available at https://github.com/suwonglab/rss-net.

For simplicity we formulate a regulatory network as a bipartite graph $\{\mathbf{V}_{\mathrm{TF}}, \mathbf{V}_{\mathrm{TG}}, \mathbf{E}_{\mathrm{TF}\to\mathrm{TG}}\}$, where $\mathbf{V}_{\mathrm{TF}}$ denotes the node set of transcription factors (TFs), $\mathbf{V}_{\mathrm{TG}}$ denotes the node set of target genes (TGs), and $\mathbf{E}_{\mathrm{TF}\to\mathrm{TG}}$ denotes the set of directed TF-to-TG edges, summarizing how TFs regulate TGs through REs (but not vice versa). Each edge has a weight between 0 and 1, measuring the relative regulation strength of a TF on a TG. Each network file contains a list of REs, TFs, TGs, TF-to-TG edges and weights. On average each network has 431 TFs, 3,298 TGs and 93,764 TF-to-TG edges. Additional information of networks is provided in Supplementary Notes and Tables 14-16.

The sequencing data used for network construction were obtained from ENCODE[5,82] data portal (https://www.encodeproject.org, accessed December 14, 2018) and GTEx[7] data portal (https://gtexportal.org, accessed July 13, 2019). Details are provided in Supplementary Table 1.

External databases for cross-reference. To validate and interpret RSS-NET gene-based results, we used the following external databases (accessed November 28, 2019): GWAS Catalog[47] (https://www.ebi.ac.uk/gwas/), Mouse Genome Informatics[43] (http://www.informatics.jax.org/), phenotype-specific Mendelian gene sets[45] (https://github.com/bogdanlab/gene_sets/), Online Mendelian Inheritance in Man[44] (https://www.omim.org/), Therapeutic Target Database[48] (http://db.idrblab.net/ttd/).

Network-induced effect size distribution. We model the total effect of SNP j on a given trait, β_j, as

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

where π_j denotes the probability that SNP j is associated with the trait ($\beta_j \neq 0$), $\{\mu_j, \sigma_0^2\}$ characterize the center and variability of a trait-associated SNP j's effect size, and δ_0 indicates point mass at zero ($\beta_j = 0$).

We model the trait-association probability π_j as

$$\log_{10}\left(\frac{\pi_j}{1-\pi_j}\right) = \theta_0 + a_j \cdot \theta, \tag{2}$$

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 inside a network is trait-associated, and a_j indicates whether SNP j is inside a network. Following previous analyses[12,13,19], we let $a_j = 1$ if SNP j is within 100 kb of any element (TG, RE, or TF) in a given network. The idea of (2) is that if a tissue or cell type plays an important role in a trait then genetic associations may tend to occur more often in SNPs involved in the network of this context than expected by chance.

We model the mean effect size μ_j as

$$\mu_j = \sum_{g \in \mathbf{O}_j} w_{jg} \cdot \gamma_{jg} \tag{3}$$

where $\mathbf{O}_j$ 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 γ_{jg} denotes the effect of SNP j on a trait due to gene g. In this study we use a pre-defined regulatory network to specify $\{\mathbf{O}_j, w_{jg}\}$:

$$\mu_j = \underbrace{\sum_{g \in \mathbf{G}_j} [c_{jg}}_{\text{cis}} \cdot (\gamma_{jg} + \underbrace{\sum_{t \in \mathbf{T}_g} v_{gt} \cdot \gamma_{jt})}_{\text{trans}}], \tag{4}$$

where $\mathbf{G}_j$ is the set of all genes within 1 Mb cis window of SNP j, c_{jg} measures the relative impact of a cis SNP j on gene g, $\mathbf{T}_g$ is the set of all genes that are directly regulated by TF g in trans in a given network ($\mathbf{T}_g = \emptyset$ if gene g is not a TF), and v_{gt} measures the relative impact of a TF g on its target gene t. We use pre-defined network edges and weights[17] to specify the trans interconnection $\mathbf{T}_g$ and impact v_{gt} respectively. We use context-matching *cis*-eQTL data[7,8,83] to specify the cis impact c_{jg}; see Supplementary Notes and Tables 17-18 for details. The idea of (3)-(4) is that the true effect of a SNP may fan out through some regulatory network of multiple (nearby or distal) genes to affect the trait[24,26].

We model the random effect of SNP j due to gene g, γ_{jg}, as

$$\gamma_{jg} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \tag{5}$$

where the SNP-level subscript j in γ_{jg} ensures the exchangeability of β_j in (1); see Supplementary Notes. The constant variance σ^2 in (5) is chosen for computational convenience. (One could potentially improve (5) by letting σ^2 depend on functional annotations[10,22] of SNP j and/or context-specific expression[11–13] of gene g, though possibly at higher computational cost.)

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 $\boldsymbol{\beta} := (\beta_1, \ldots, \beta_p)'$ in (1), but we can infer them from GWAS summary statistics and LD estimates. Specifically, we perform Bayesian inference

for β by combining the network-based effect size prior (1)–(5) with the "Regression with Summary Statistics" (RSS) likelihood[20]:

$$\widehat{\beta} \sim \mathcal{N}(\widehat{\mathbf{S}}\widehat{\mathbf{R}}\widehat{\mathbf{S}}^{-1}\beta,\ \widehat{\mathbf{S}}\widehat{\mathbf{R}}\widehat{\mathbf{S}}), \tag{6}$$

where $\widehat{\beta} := (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$ is a $p \times 1$ vector, $\widehat{\mathbf{S}} := \mathrm{diag}(\widehat{\mathbf{s}})$ is a $p \times p$ diagonal matrix with diagonal elements being $\widehat{\mathbf{s}} := (\hat{s}_1, \ldots, \hat{s}_p)'$, $\hat{\beta}_j$ and $\hat{s}_j$ are estimated single-SNP effect size of each SNP j and its standard error from the GWAS, and $\widehat{\mathbf{R}}$ is the $p \times p$ LD matrix estimated from an external reference panel with ancestry matching the GWAS.

RSS-NET, defined by the hierarchical model (1)–(6), consists of four unknown hyper-parameters: $\{\theta_0, \theta, \sigma_0^2, \sigma^2\}$. To specify hyper-priors, we first introduce two free parameters $\{\eta, \rho\} \in [0,1]$ to re-parameterize $\{\sigma_0^2, \sigma^2\}$:

$$\sigma_0^2 = \eta \cdot (1-\rho) \cdot \left(\sum_{j=1}^{p} \frac{\pi_j}{n\hat{s}_j^2} \right)^{-1}, \quad \sigma^2 = \eta \cdot \rho \cdot \left(\sum_{j=1}^{p} \frac{\pi_j \cdot \sum_{g \in \mathbf{O}_j} w_{jg}^2}{n\hat{s}_j^2} \right)^{-1}, \tag{7}$$

where, roughly, η represents the proportion of the total phenotypic variation explained by p SNPs, and ρ represents the proportion of total genetic variation explained by network annotations $\{\mathbf{O}_j, w_{jg}\}$. Because $n\hat{s}_j^2$ is roughly the ratio of phenotype variance to genotype variance, (7) ensures that genetic effect sizes (β) do not rely on sample size n, and have the same measurement unit as the trait. See Supplementary Notes for the derivation of (7). We then place independent uniform grid priors on $\{\theta_0, \theta, \eta, \rho\}$ (Supplementary Table 19). We verify that RSS-NET results are robust to grid choice (Supplementary Fig. 8). (If one had specific information about $\{\theta_0, \theta, \eta, \rho\}$ in a given setting then this could be incorporated here.)

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

$$\mathrm{BF} = \frac{p(\widehat{\beta} \mid \widehat{\mathbf{S}}, \widehat{\mathbf{R}}, \mathbf{a}, \mathbf{O}, \mathbf{W}, M_1)}{p(\widehat{\beta} \mid \widehat{\mathbf{S}}, \widehat{\mathbf{R}}, \mathbf{a}, \mathbf{O}, \mathbf{W}, M_0)}, \tag{8}$$

where $p(\cdot)$ denotes probability densities, $\mathbf{a}$ is defined in (2), $\{\mathbf{O}, \mathbf{W}\}$ are defined in (3), M_1 denotes the enrichment model where $\theta > 0$ or $\sigma^2 > 0$, and M_0 denotes the baseline model where $\theta = 0$ 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 Notes for details of computing BF. To compute BFs used in Figure 5b, we replace M_1 in (8) 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.

Locus association. To identify 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:

$$P_1 = 1 - \Pr(\beta_j = 0, \forall j \in \text{locus} \mid \mathbf{D}, \text{model}), \tag{9}$$

where $\mathbf{D}$ is a shorthand for the input data of RSS-NET including GWAS summary statistics $\{\widehat{\beta}, \widehat{\mathbf{S}}\}$, LD estimates $\widehat{\mathbf{R}}$ and network annotations $\{\mathbf{a}, \mathbf{O}, \mathbf{W}\}$. See Supplementary Notes for details of computing P_1. For a locus, P_1^{base}, P_1^{near} and P_1^{net} correspond to P_1 evaluated under the baseline model M_0, the enrichment model M_1 for the near-gene control network with all genes as nodes and no edges, and M_1 for a given network. In this study a locus is defined as the transcribed region of a gene plus 100 kb upstream and downstream.

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

$$P_1^{\text{bma}} = \frac{\sum_{k=1}^{K} P_1^{\text{net}}(k) \cdot \text{BF}(k)}{\sum_{k=1}^{K} \text{BF}(k)}, \tag{10}$$

where $P_1^{\text{net}}(k)$ and $\text{BF}(k)$ are enrichment P_1 and BF for network $k = 1, \ldots, K$. The ability to average across models in (10) is an advantage the Bayesian approach, because it allows us to assess associations in light of the network enrichment without having to select a single enrichment model.

Computation time. The total computational time of RSS-NET to analyze a pair of trait and network is determined by the number of 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 computational time. However, to give a specific example, we finished the analysis of 1.1 million 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).

Simulation overview. Using genotypes of 348,965 genome-wide autosomal SNPs from 1,458 individuals[23], we simulated enrichment datasets under M_1 for the B cell regulatory network[5,17,82] (Fig.s 2-4; Supplementary Fig.s 1-6), and simulated baseline datasets in the following scenarios: (1) M_0 (Figs. 2, 4; Supplementary Fig.s 1, 5, 6); (2) random near-gene SNPs were enriched for associations (Fig 3a; Supplementary Fig. 2); (3) random near-RE SNPs were enriched for associations (Fig 3b; Supplementary Fig. 3); (4) edge-altered B cell networks were enriched for associations (Fig 3c; Supplementary Fig.

4). We matched enrichment and baseline datasets by the number of trait-associated SNPs and the proportion of phenotypic variation explained by all SNPs. On the simulated datasets we assessed enrichments of the B cell network (Fig.s 2-3; Supplementary Fig.s 1-4) and tested gene-based associations (Fig. 4; Supplementary Fig. 5). The only exception is the noisy network simulations (Supplementary Fig. 6) where we analyzed random subsets of the B cell network. Simulation details are provided in Supplementary Figures 1-6.

This study used the following software packages in simulations: 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 (version 1.0.0, https://github.com/bulik/ldsc, accessed November 27, 2018). See Supplementary Notes for details.

Code availability. The RSS-NET software is available at https://github.com/suwonglab/rss-net. Tutorials of installing and using RSS-NET are provided in 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. All other codes are specified in Methods and Supplementary Notes.

Data availability. Network files used in this study are available at https://github.com/suwonglab/rss-net. Analysis results of this study are available at https://xiangzhu.github.io/rss-peca. All other data are specified in Methods and Supplementary Notes.

Author contributions. X.Z. and W.H.W. conceived the study. X.Z. developed the methods and implemented the software. X.Z. 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.

Acknowledgments. This study is supported by Stein Fellowship to X.Z. and NIH grants P50HG007735 and R01HG010359 to W.H.W. This study uses data generated by the Wellcome Trust Case Control Consortium, 1000 Genomes Project, ENCODE Consortium, GTEx Project, DICE Project, eQTLGen Consortium, and multiple GWAS consortia. We thank them for making their data publicly available. Detailed acknowledgments are in Supplementary Notes.

References.

[1] Visscher, P. M. *et al.* 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics* **101**, 5–22 (2017).

[2] Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics* **20**, 467–484 (2019).

[3] Manolio, T. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).

[4] Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences* **106**, 9362–9367 (2009).

[5] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

[6] Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

[7] GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

[8] Schmiedel, B. J. *et al.* Impact of genetic polymorphisms on human immune cell gene expression. *Cell* **175**, 1701–1715 (2018).

[9] Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature Genetics* **45**, 124–130 (2013).

[10] Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* **47**, 1228–1235 (2015).

[11] Calderon, D. *et al.* Inferring relevant cell types for complex traits by using single-cell gene expression. *American Journal of Human Genetics* **101**, 686–699 (2017).

[12] Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nature Genetics* **50**, 621–629 (2018).

[13] Zhu, X. & Stephens, M. Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. *Nature Communications* **9**, 4361 (2018).

[14] Greene, C. S. *et al.* Understanding multicellular function and disease with human tissue-specific networks. *Nature Genetics* **47**, 569–576 (2015).

[15] Marbach, D. *et al.* Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nature Methods* **13**, 366–370 (2016).

[16] Sonawane, A. R. *et al.* Understanding tissue-specific gene regulation. *Cell Reports* **21**, 1077–1088 (2017).

[17] Duren, Z., Chen, X., Jiang, R., Wang, Y. & Wong, W. H. Modeling gene regulation from paired expression and chromatin accessibility data. *Proceedings of the National Academy of Sciences* **114**, E4914–E4923 (2017).

[18] 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. *Nature Genetics* **44**, 841–847 (2012).

[19] Kim, S. S. *et al.* Genes with high network connectivity are enriched for disease heritability. *American Journal of Human Genetics* **104**, 896–913 (2019).

[20] Zhu, X. & Stephens, M. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *Annals of Applied Statistics* **11**, 1561–1592 (2017).

[21] 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 Computational Biology* **12**, e1004714 (2016).

[22] Gazal, S. *et al.* Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nature Genetics* **49**, 1421–1427 (2017).

[23] 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).

682 [24] Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from poly-
683 genic to omnigenic. *Cell* **169**, 1177–1186 (2017).
684 [25] International HapMap 3 Consortium. Integrating common and rare genetic variation in
685 diverse human populations. *Nature* **467**, 52–58 (2010).
686 [26] Liu, X., Li, Y. I. & Pritchard, J. K. Trans effects on gene expression can drive omnigenic
687 inheritance. *Cell* **177**, 1022–1034 (2019).
688 [27] Lee, B.-C. *et al.* Adipose natural killer cells regulate adipose tissue macrophages to
689 promote insulin resistance in obesity. *Cell Metabolism* **23**, 685–698 (2016).
690 [28] Hotamisligil, G. S. Inflammation, metaflammation and immunometabolic disorders. *Na-*
691 *ture* **542**, 177–185 (2017).
692 [29] Oikonomou, E. K. & Antoniades, C. The role of adipose tissue in cardiovascular health
693 and disease. *Nature Reviews Cardiology* **16**, 83–99 (2019).
694 [30] Kang, J. & Rivest, S. Lipid metabolism and neuroinflammation in Alzheimer's disease:
695 a role for liver X receptors. *Endocrine Reviews* **33**, 715–746 (2012).
696 [31] Shahid, F., Lip, G. Y. & Shantsila, E. Role of monocytes in heart failure and atrial fibril-
697 lation. *Journal of the American Heart Association* **7**, e007849 (2018).
698 [32] Aviles, R. J. *et al.* Inflammation as a risk factor for atrial fibrillation. *Circulation* **108**,
699 3006–3010 (2003).
700 [33] Cotsapas, C. *et al.* Pervasive sharing of genetic effects in autoimmune disease. *PLoS*
701 *Genetics* **7**, e1002254 (2011).
702 [34] Lo, M. *et al.* Genome-wide analyses for personality traits identify six genomic loci and
703 show correlations with psychiatric disorders. *Nature Genetics* **49**, 152–156 (2017).
704 [35] Ferrari, R. & Fox, K. Heart rate reduction in coronary artery disease and heart failure.
705 *Nature Reviews Cardiology* **13**, 493–501 (2016).
706 [36] Dale, C. E. *et al.* Causal associations of adiposity and body fat distribution with coronary
707 heart disease, stroke subtypes, and type 2 diabetes mellitus: a Mendelian randomization
708 analysis. *Circulation* **135**, 2373–2388 (2017).
709 [37] Odutayo, A. *et al.* Atrial fibrillation and risks of cardiovascular disease, renal disease,
710 and death: systematic review and meta-analysis. *BMJ* **354**, i4482 (2016).
711 [38] Nelson, C. P. *et al.* Genetically determined height and coronary artery disease. *New*
712 *England Journal of Medicine* **372**, 1608–1618 (2015).
713 [39] Di, G. P. & Kim, T. Linking lipids to Alzheimer's disease: cholesterol and beyond. *Nature*
714 *Reviews Neuroscience* **12**, 284–296 (2011).
715 [40] Heppner, F., Ransohoff, R. & Becher, B. Immune attack: the role of inflammation in
716 Alzheimer disease. *Nature Reviews Neuroscience* **16**, 358–372 (2015).
717 [41] Pickrell, J. *et al.* Detection and interpretation of shared genetic influences on 42 human
718 traits. *Nature Genetics* **48**, 709–717 (2016).
719 [42] Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and
720 traits. *Nature Genetics* **47**, 1236–1241 (2015).
721 [43] Bult, C. *et al.* Mouse Genome Database (MGD) 2019. *Nucleic Acids Research* **47**, D801–
722 D806 (2019).
723 [44] Amberger, J., Bocchini, C., Scott, A. & Hamosh, A. OMIM.org: leveraging knowledge
724 across phenotype-gene relationships. *Nucleic Acids Research* **47**, D1038–D1043 (2019).
725 [45] Freund, M. K. *et al.* Phenotype-specific enrichment of Mendelian disorder genes near
726 GWAS regions across 62 complex traits. *The American Journal of Human Genetics* **103**,
727 535–552 (2018).
728 [46] Noordam, R. *et al.* Multi-ancestry sleep-by-SNP interaction analysis in 126,926 indi-
729 viduals reveals lipid loci stratified by sleep duration. *Nature Communications* **10**, 1–13
730 (2019).
731 [47] Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide associa-

tion studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research* **47**, D1005–D1012 (2019).

[48] Wang, Y. *et al.* Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. *Nucleic acids research* **48**, D1031–D1041 (2020).

[49] Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).

[50] Kerr, T. A. *et al.* Loss of nuclear receptor SHP impairs but does not eliminate negative feedback regulation of bile acid synthesis. *Developmental Cell* **2**, 713–720 (2002).

[51] Wang, L. *et al.* Redundant pathways for negative feedback regulation of bile acid production. *Developmental Cell* **2**, 721–731 (2002).

[52] Hartman, H. B., Lai, K. & Evans, M. J. Loss of small heterodimer partner expression in the liver protects against dyslipidemia. *Journal of Lipid Research* **50**, 193–203 (2009).

[53] Nishigori, H. *et al.* Mutations in the small heterodimer partner gene are associated with mild obesity in Japanese subjects. *Proceedings of the National Academy of Sciences* **98**, 575–580 (2001).

[54] Willer, C. *et al.* Discovery and refinement of loci associated with lipid levels. *Nature Genetics* **45**, 1274–1283 (2013).

[55] Brown, S. *et al.* High-throughput mouse phenomics for characterizing mammalian gene function. *Nature Reviews Genetics* **19**, 357–370 (2018).

[56] Lupski, J., Belmont, J., Boerwinkle, E. & Gibbs, R. Clan genomics and the complex architecture of human disease. *Cell* **147**, 32–43 (2011).

[57] Blair, D. R. *et al.* A nondegenerate code of deleterious variants in mendelian loci contributes to complex disease risk. *Cell* **155**, 70–80 (2013).

[58] Bruning, P. F. *et al.* Insulin resistance and breast-cancer risk. *International Journal of Cancer* **52**, 511–516 (1992).

[59] Plenge, R., Scolnick, E. & Altshuler, D. Validating therapeutic targets through human genetics. *Nature Reviews Drug Discovery* **12**, 581–594 (2013).

[60] Nelson, M. *et al.* The support of human genetic evidence for approved drug indications. *Nature Genetics* **47**, 856–860 (2015).

[61] Sanseau, P. *et al.* Use of genome-wide association studies for drug repositioning. *Nature Biotechnology* **30**, 317–320 (2012).

[62] 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).

[63] 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).

[64] Barrett, A. & Sloand, E. Autoimmune mechanisms in the pathophysiology of myelodysplastic syndromes and their clinical relevance. *Haematologica* **94**, 449–451 (2009).

[65] Wang, Q. *et al.* A Bayesian framework that integrates multi-omics data and gene networks predicts risk genes from schizophrenia GWAS data. *Nature Neuroscience* **22**, 691–699 (2019).

[66] Schaid, D., Chen, W. & Larson, N. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics* **19**, 491–504 (2018).

[67] Christophersen, I. E. *et al.* Large-scale analyses of common and rare variants identify 12 new loci associated with atrial fibrillation. *Nature Genetics* **49**, 946–952 (2017).

[68] Michailidou, K. *et al.* Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nature Genetics* **45**, 353–61 (2013).

[69] Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **515**, 197–206 (2015).

[70] Nikpay, M. *et al.* A comprehensive 1000 genomes–based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics* **47**, 1121–1130 (2015).

[71] Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics* **46**, 1173–1186 (2014).

[72] Den Hoed, M. *et al.* Identification of heart rate-associated loci and their effects on cardiac conduction and rhythm disorders. *Nature Genetics* **45**, 621–631 (2013).

[73] Liu, J. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature Genetics* **47**, 979–986 (2015).

[74] Lambert, J.-C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics* **45**, 1452–1458 (2013).

[75] Okbay, A. *et al.* Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nature Genetics* **48**, 624–633 (2016).

[76] Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).

[77] Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).

[78] Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics* **44**, 981–990 (2012).

[79] Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187–196 (2015).

[80] 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

[81] Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biology* **9**, R137 (2008).

[82] Luo, Y. *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Research* **48**, D882–D889 (2020).

[83] Võsa, U. *et al.* Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. *bioRxiv* 447367 (2018).

XIANG ZHU
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
390 JANE STANFORD WAY
STANFORD, CA 94305
USA

ZHANA DUREN AND WING HUNG WONG
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
390 JANE STANFORD WAY
STANFORD, CA 94305
USA
AND
DEPARTMENT OF BIOMEDICAL DATA SCIENCE
STANFORD UNIVERSITY
1265 WELCH ROAD
STANFORD, CA 94305
USA