Aggregation of recount3 RNA-seq data improves inference of consensus and tissue-specific gene coexpression networks

Prashanthi Ravichandran, Princy Parsana, Rebecca Keener, et al.

Genome Res. published online July 17, 2025
Access the most recent version at doi:10.1101/gr.280808.125

Aggregation of recount3 RNA-seq data improves inference of consensus and tissue-specific gene co-expression networks

Prashanthi Ravichandran[1], Princy Parsana[2], Rebecca Keener[1], Kasper D. Hansen[1,3,4], Alexis Battle[*1,2,4,5,6]

1. Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA
2. Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA
3. Department of Biostatistics, Johns Hopkins School of Public Health, Baltimore, MD, USA
4. Department of Genetic Medicine, Johns Hopkins University, Baltimore, MD, USA
5. Malone Center for Engineering in Healthcare, Johns Hopkins University, Baltimore, MD, USA
6. Data Science and AI Institute, Johns Hopkins University, Baltimore, MD, USA

* Corresponding author

Alexis Battle, PhD
ajbattle@jhu.edu

Keywords

Gene co-expression networks (GCNs), Graphical lasso, recount3, Public RNA-seq data, SRA, GTEx, TCGA, Data aggregation, Context-specificity, Complex trait heritability, s-LDSC.

Abstract

Gene co-expression networks (GCNs) describe relationships among genes that maintain cellular identity and homeostasis. However, typical RNA-seq experiments often lack sufficient sample sizes for reliable GCN inference. Recount3, a dataset with 316,443 processed human RNA-seq samples, provides an opportunity to improve network reconstruction. However, GCN inference from public data is challenged by confounders and inconsistent labeling. To address this, we developed a pipeline to annotate samples based on cell-type composition. By comparing aggregation strategies, we found that regressing confounders within studies and prioritizing larger studies optimized network reconstruction. We applied these findings to infer three consensus networks (universal, cancer, non-cancer) and 27 context-specific networks. Central genes in consensus networks were enriched for evolutionarily constrained genes and ubiquitous biological pathways, while context-specific central nodes included tissue-specific transcription factors. The increased statistical power from data aggregation facilitated the derivation of variant annotations from context-specific networks, which were significantly enriched for complex-trait heritability independent of overlap with baseline functional genomic annotations. While data aggregation led to strictly increasing held-out log-likelihood, we observed diminishing marginal improvements, suggesting that integrating complementary modalities, such as Hi-C and ChIP-seq, could further refine network reconstruction. Our approach outlines best practices for GCN inference and highlights both the strengths and limitations of data aggregation.

Introduction

Critical cellular processes including the maintenance of cellular identity, homeostasis, and the cellular response to external stimuli are orchestrated through complex transcriptional co-regulation of multiple genes (Hartwell et al. 1999; Hasty, McMillen, and Collins 2002; Oltvai

and Barabási 2002; Alon 2003). Gene Co-expression Networks (GCNs) are a commonly used framework to describe gene-gene relationships and are comprised of nodes that represent genes and edges linking co-expressed genes (Stuart et al. 2003). A comprehensive catalogue of gene co-expression relationships has the potential to characterize genes with unknown functions (Schlitt et al. 2003), identify regulatory genes (Narang et al. 2015), determine changes in regulatory mechanisms that are key to cellular identity (Wang, Choi, and Roeder 2021), and prioritize genes that drive phenotypic variability (van Dam et al. 2017).

Despite the utility of GCNs in understanding biological systems, network inference is still a challenging problem and suffers from both false positive and negative edges (Diaz and Stumpf 2022). In particular, the typical sample size of most RNA-seq studies is orders of magnitude smaller than the number of gene pairs over which regulatory relationships are inferred, making network inference an underdetermined problem. Additionally, factors such as the stochastic nature of gene expression, experimental noise, missing data, and unobserved technical confounders make it difficult to avoid false positives or negatives.

Since the number of possible gene-gene interactions scales with the square of the number of genes examined, a potential solution to increase statistical power by reducing network complexity has been to utilize methods such as WGCNA that infer modules or groups of co-expressed genes that are regulated by one or more transcription factors rather than individual gene interactions (Segal et al. 2003). While this approach has been successful at decreasing the number of hypotheses tested, and thereby increasing statistical power (Wolf et al. 2014), it does not identify detailed network structure or distinguish between direct and indirect gene interactions. In contrast, network inference by graphical lasso (Friedman, Hastie, and Tibshirani 2008; Hastie, Tibshirani, and Friedman 2013) results in the identification of pairwise edges

reflecting direct effects, such that the absence of an edge implies the conditional independence of the genes when all other genes are observed. Further, the formulation of graphical lasso enables flexible penalization based on the number of edges found in the network which aids in the identification of a network structure that improves the discovery of true gene-gene interactions while reducing false positives (Huang, Lu, and Hsiao 2020). Here, we focus on improving the statistical power of network inference by significantly increasing the number of samples used in network inference, leveraging large-scale publicly available and uniformly processed RNA-seq data from recount3 (Wilks et al. 2021) which includes human RNA-seq samples from GTEx (The GTEx Consortium 2020), TCGA (Tomczak, Czerwińska, and Wiznerowicz 2015), and SRA (Katz et al. 2022; Kodama et al. 2012).

Public RNA-seq data resources such as recount3 offer an unprecedented opportunity to improve GCN inference, but also present major challenges due to heterogeneity in sample preparation, sequencing protocols, missing or inconsistent metadata, and variation in biological and technical factors across studies. To address these challenges, we set out to develop a data pre-processing pipeline that can identify and exclude outliers, harmonize expression measurements across studies, and group samples into biologically meaningful contexts appropriate for network reconstruction.

While statistical methods for confounder correction have been widely applied within individual studies (Parsana et al. 2019), it remains unclear how best to account for confounding and integrate data across multiple studies in a way that preserves true biological signal and enhances the quality of network inference. This motivated us to systematically evaluate strategies for confounder adjustment and data aggregation across studies. Finally, to assess the utility of these networks for downstream biological discovery, we examine the relevance of

inferred network features, such as node centrality, to known regulatory roles and their enrichment for complex trait heritability using stratified LD score regression (S-LDSC; Finucane et al., 2015). In summary, our study seeks to provide a carefully annotated RNA-seq data set, outline best practices for GCN inference by leveraging publicly available RNA-seq data, and a set of consensus and context-specific networks that will aid the scientific community in achieving the full potential of GCN inference in biomedical research.

Results

Manual annotation and clustering of RNA-seq data from recount3 identified 27 unique tissue contexts

We downloaded uniformly processed RNA-seq samples from humans using the recount3 R package (Wilks et al. 2021) comprised of experiments from three data sources, The Sequence Read Archive (SRA) (Kodama et al. 2012; Katz et al. 2022), Genotype-tissue Expression (GTEx version 8) (The GTEx Consortium 2020), and The Cancer Genome Atlas (TCGA) (Tomczak, Czerwińska, and Wiznerowicz 2015) and selected 1,747 projects that included 30 or more samples each. Following quality control **(Methods)**, 95,280 human bulk-RNA sequencing samples remained from 50 GTEx tissues (18,828 samples), 33 TCGA cancer types (11,091 samples), and 884 SRA studies (65,361 samples) **(Fig. 1 A)**. The number of genes present following data pre-processing varied by study and is summarized in **Supplemental Fig. S1**. The aggregated data includes samples from a wide array of tissues, cell types, and diseases. While GTEx and TCGA studies included metadata specifying the tissue of origin and disease status for all samples, SRA studies had inconsistent nomenclature. Therefore, to obtain reliable labels for SRA samples, we manually parsed sample descriptions to obtain sample characteristics corresponding to tissue type and disease status for 65,361 SRA samples **(Methods)**. Based on

curated annotations, 93.5% of TCGA samples and 30.4% of SRA samples were cancer. In contrast, all GTEx samples were non-cancer, as expected **(Fig. 1 B)**. Tissue labels with the greatest number of samples across all three data sources included blood, central nervous system, breast, skin, and lung **(Fig. 1 C)**. SRA included 224 distinct tissue labels derived from manual annotation that was not observed in GTEx or TCGA, and reflected a wide range of disease states including Type I Diabetes, Alzheimer's disease, bipolar disorder, arthritis, cancer, and infectious conditions **(Fig. 1 D)**. We grouped SRA samples based on their study accession IDs, GTEx samples by tissue, and TCGA samples by cancer code (**Methods**). To simplify terminology, we defined each group of samples from a data source as a single study. To leverage the extensive biological diversity in the data, we inferred two broad types of networks: consensus and tissue context-specific (context-specific). Our universal consensus network included all samples, regardless of tissue or disease. Our non-cancer consensus network included healthy samples and samples with disease status other than cancer. Finally, our cancer consensus network solely included cancerous samples. We restricted our context-specific networks to non-cancerous samples grouped by tissue context. We did not examine differential coregulation resulting from non-cancer disease, and regressed these effects from gene expression. Thus, by including SRA, recount3 provides an unprecedented opportunity to examine unique contexts that were not previously studied in GTEx and TCGA.

Across the three data sources, SRA, GTEx, and TCGA, we obtained 266 unique manually annotated tissue labels with a median sample size of 31 which was much lower than the number of protein-coding genes. Therefore, we used a study-pooling strategy based on related tissue contexts to increase power **(Methods)**. Specifically, in this work, we define a tissue context as a group of samples with similar cell-type composition, which we infer

computationally. Mapping manual annotations to tissue contexts **(Supplemental Table S1)** generated 48 tissue contexts across 63,193 non-cancerous samples for context-specific network analysis. In each context, to ensure that a tissue context represented samples with similar cell-type composition as estimated by xCell (Aran, Hu, and Butte 2017) deconvolution, we learned a joint lower-dimensional t-SNE embedding using cell-type deconvolution scores, and for 25 contexts with more than 500 samples before outlier exclusion, we detected and excluded outliers **(Supplemental Fig. S2)**. For the immune context, we observed that samples displayed extensive heterogeneity in cell-type composition. Thus we further separated this group into B cells, PBMCs/ T cells, and myeloid cells **(Supplemental Fig. S3)**. Finally, we examined the differences in the empirical covariance estimates obtained when we considered alternate thresholds to exclude outliers, and found that while we observed minor differences across covariance estimates obtained by varying the exclusion criterion, these differences were much smaller than the differences in empirical covariance matrices between unrelated contexts, and hence selected samples that were found to be outliers by fewer than two methods, to ensure we considered the maximum possible number of samples for each context, while minimizing outliers that may truly reflect different contexts. **(Supplemental Fig. S4)**.

In total, we obtained 27 contexts including 7 tissue contexts that were not present in GTEx: airway, eye, human embryonic stem cells, induced pluripotent stem cells, multipotent cells, myeloid cells, and PBMCs / T cells **(Fig. 1 E)**. The number of samples present in each tissue context following outlier exclusion varied from 8,797 (Blood) to 485 (Multipotent cells) **(Fig. 1 F)**. Of the tissue categories present in GTEx we were able to significantly increase the number of samples for several tissue contexts including blood, central nervous system, skin, and liver **(Fig. 1 F)**. These harmonized samples have increased context resolution, i.e. include novel

contexts which were not examined in GTEx, and increased sample size which can be used to improve the inference of consensus and context-specific networks.

Data aggregation improves the inference of consensus and context-specific GCNs

The median study-specific sample size across the three data sources was 44 for SRA, 309 for TCGA, and 285 for GTEx. Further, 766 of 884 SRA studies had a sample size lesser than 100. PC-based data correction has been used within a single study to reduce potential false positive gene regulatory associations (Leek et al. 2012; Parsana et al. 2019), but best practices for applying PC-based data correction in the context of aggregating multiple studies to infer GCNs have not been fully examined. First, we sought to determine whether data correction should be performed jointly across all samples, across samples belonging to a specific tissue context and multiple studies, or across samples from a specific tissue context and a specific study. We observed that PCs recapitulate different sample characteristics depending on the level at which data aggregation is performed. PCs calculated across all samples were driven predominantly by tissue labels followed by technical confounders (e.g. study and data source) **(Supplemental Fig. S5 A-D, Supplemental Fig. S6 A)**. PCs which were calculated across samples belonging to a single tissue context (blood) but multiple studies were predominantly driven by study and data source. Further, accounting for tissue heterogeneity enabled us to better model cancer status and disease annotations using PCs **(Supplemental Fig. S5 E-H, Supplemental Fig. S6 B)**. Finally, when we limited samples to a specific tissue context and a specific study (GTEx-skeletal muscle), we found that the top PCs were significantly associated with age, cause of death, and technical batch, consistent with the findings of Parsana et al. (Parsana et al. 2019) **(Supplemental Fig. S5 I-L, Supplemental Fig. S6 C)**. Thus regressing

PCs computed by accounting for tissue and cross-study heterogeneity from expression is integral to excluding technical effects and unwanted biological signals.

We examined the effect of PC-based data correction when inferring GCNs by aggregating data across diverse sources by comparing PC-based correction applied to either aggregate data, or individual studies followed by aggregation, where the number of PCs regressed is based on the permutation method described by Buja et. al (Buja and Eyuboglu 1992) and Leek et. al (Leek et al. 2012). Additionally, we compared the consequences of aggregating empirical covariance matrices inferred from PC-corrected data by either treating all studies irrespective of sample size equally (unweighted aggregation), or upweighting the covariance estimates from studies with a greater sample size (weighted aggregation). Thus, we used 4 paradigms: (1) Aggregating data before PC-based data correction followed by estimation of empirical covariance from residual expression, (2) PC-based data correction applied to individual studies followed by aggregation of residual expression and joint estimation of empirical covariance, (3) Unweighted aggregation of covariance matrices inferred from each study separately after study-specific PC-based correction and (4) Weighted aggregation of covariance matrices computed from individual studies following study-specific PC-based data correction, where the weights were the ratio of the study sample size to the total number of samples used in network reconstruction **(Fig. 2 A, Methods)**.

To compare strategies, we split non-cancerous samples into two data splits, GTEx and SRA **(Supplemental Fig. S7)**, followed by network inference by graphical lasso on one of the two data splits and evaluation with the held-out split by computing the held-out log-likelihood. Details pertaining to the number of studies, samples and median PCs regressed for incremental data aggregation are provided in **Supplemental Table S2**. Additionally, we assessed the recapitulation of known biological pathways by computing the F1-score of finding canonical

gene-gene relationships compiled from KEGG (Kanehisa and Goto 2000), Biocarta, and Pathway Interaction Database (Schaefer et al. 2009) obtained using Enrichr (Chen et al. 2013; Kuleshov et al. 2016) **(Supplemental Table S3)**. Paradigms in which PC-based data correction preceded aggregation led to a strict increase in held-out log-likelihood and F1-scores of known gene relationships from canonical pathways with the addition of more studies **(Fig. 2 B, C, Supplemental Fig. S8, S9)**. Additionally, we find that the low F1-scores were attributed to lower recall while the precision remained high, in accordance with previous studies that utilize graphical lasso to infer GCNs (Parsana et al. 2019). Further, we observed an increase in the precision with data aggregation for a particular value of recall across all strategies which performed PC-based data correction on individual studies **(Supplemental Fig. S10, S11)**. This suggests that data aggregation resulted in GCNs with greater generalizability and recapitulated known biology better when technical confounders were estimated and regressed for individual studies. Since data aggregation led to a decrease in the number of edges found in the network for a particular value of the penalization parameter **(Supplemental Fig. S12)**, we tested whether estimating denser networks would result in higher held-out log-likelihood specifically when the networks were inferred over a greater number of samples, but instead observed that a larger number of edges led to overfitting and lower generalizability **(Supplemental Fig. S13)**. Further, the marginal improvement in network reconstruction diminished with the subsequent rounds of aggregation. While all methods that estimated technical confounders within each study performed similarly and were superior to estimating technical confounders across all samples when evaluated by held-out log-likelihood **(Fig. 2 D)**, we found that weighted aggregation of covariance matrices led to a slight improvement in the F1-score of the networks when compared

to canonical pathways **(Fig. 2 E)**, suggesting that this is the optimal strategy among the alternatives compared.

While our primary analysis concerns the recapitulation of individual, direct gene-gene edges through graphical lasso, which we expected to be heavily impacted by sample size, we also performed a limited evaluation of the effects of data aggregation with Weighted Gene Co-expression Network Analysis (WGCNA) (Langfelder and Horvath 2008), a popular method for identifying modules of co-expressed genes. We inferred WGCNA modules by sequentially aggregating GTEx tissues and SRA studies following PC correction **(Methods)**. We discovered that data aggregation led to an improvement in the recapitulation of known gene relationships from KEGG (Kanehisa and Goto 2000), Biocarta, and Pathway Interaction Database (Schaefer et al. 2009) for both GTEx and SRA. However, we note that this improvement is not strictly monotonic, particularly at the highest levels of data aggregation, in contrast to graphical lasso. This may be expected as module detection for WGCNA combines information across multiple gene pairs to identify modules, and has been shown to work reasonably well even at small sample sizes. WGCNA module detection simply may not benefit as much from increased sample sizes. Thus, we conclude that PC-correction followed by data aggregation can potentially lead to an improvement in network performance for multiple network inference methods **(Supplemental Fig. S14)**, but the improvement plateaus more quickly when inferring modules with WGCNA than when inferring direct edges with graphical lasso.

Finally, since graphical lasso can potentially identify edges reflecting direct gene-gene interactions, we evaluated the impact of data aggregation by weighted aggregation of covariance matrices which were estimated from SRA studies following PC correction on recapitulating known transcription factor (TF)-target gene relationships that were curated by Liska et. al (Liska

et al. 2022) based on experimental evidence such as ChIP-seq. We observed a modest improvement in the precision of predicting direct gene-gene relationships at higher values of recall with sequential data aggregation.While we do see improvement, the identification of direct gene interactions remains challenging in the presence of unknown technical and biological confounders (Kernfeld et al. 2024) **(Supplemental Fig. S15 A)**.

We inferred consensus GCNs across diverse tissues by weighted aggregation of covariance matrices estimated from residual expression and graphical lasso **(Methods)**. In addition to a universal consensus network which was inferred across 966 studies with sample size greater than or equal to 15, amounting to 95,276 samples across 48 tissue contexts, we constructed a non-cancer consensus network and a cancer consensus network. The non-cancer consensus network reflects data aggregated across 629 studies and 63,031 samples, and the cancer network reflects 386 studies and 29,967 samples **(Fig. 2 F)**. The number of genes which were included in the inference of the three consensus networks is summarized in **(Supplemental Table S4)**.We evaluated each consensus network to recapitulate previously reported gene-gene interaction using the F1 score. Across a range of network with a varying number of edges we obtained a higher estimate of the F1 score from the universal consensus network, followed by non-cancer and cancer consensus networks, which mirrors differences in sample size **(Fig. 2 G)**. We observed that the universal consensus and non-cancer consensus networks performed better than the cancer consensus network at recapitulating potential TF-target relationships obtained from TFLink (version 1.0) (Liska et al. 2022), however, this could be either due to differences in sample size or the lack of cancer samples in the reference **(Supplemental Fig. S15 B)**

We inferred networks across 27 tissue contexts and examined the impact of data aggregation on context-specific network reconstruction by considering GTEx samples (GTEx

only) or by aggregating across samples from all data sources for that tissue context. The number of samples, studies, and median study-specific sample size for each tissue context in either aggregation setting are provided in **Supplemental Fig. S16**. The number of genes over which context-specific networks was inferred is summarized in **(Supplemental Table S4)**. As in the consensus network inference, we used weighted aggregation of covariance matrices as the input to network inference by graphical lasso **(Methods**). To quantify the improvement in network reconstruction with data aggregation, we examined two tissue contexts with the largest sample size: blood and central nervous system (CNS). In both cases, we sequentially aggregated the blood or CNS SRA studies and computed the held-out log-likelihood utilizing context-matched GTEx samples. Similar to the results obtained from our consensus network analyses, we found that data aggregation led to a strict increase in held-out log-likelihood with additional studies and the greatest increase was observed while aggregating the first 20 studies (blood) and 15 studies (CNS) **(Methods, Fig. 2 H, I)**. Further, we examined the impact of data aggregation on inferring a context-specific network, specifically, if PC-based data correction sufficiently accounts for inter-study heterogeneity and minimizes the discovery of false positives, while improving the detection of true positive edges, when compared to alternate approaches to minimize false positives such as tuning the penalization parameter to be sufficiently large. Thus we compared a blood-specific network inferred across all available samples to a blood-specific network inferred from a single study with a large sample-size (GTEx) in recapitulating known gene-gene interactions. We found that increasing the penalization parameter to improve the precision led to lower recall, a limitation that was overcome by data aggregation, which led to higher values of precision even for higher recall values. Finally, we observed that the optimal F1 score obtained

by data aggregation exceeded the optimal F1 score obtained by tuning the penalization parameter **(Supplemental Fig. S17)**.

Orthogonally, we verified that our networks recapitulated known context-specific gene relationships by examining the enrichment of previously known tissue-specific protein-protein interactions (PPIs) from SNAP (Leskovec and Sosič 2016) in edges belonging to six tissue context-specific networks (adipose, blood, CNS, liver, lung, and skin) **(Methods)**. Details of the SNAP PPI terms that were mapped to specific tissue-contexts can be found in **Supplemental Table S5**. Data aggregation and increased sample size in the network significantly increased the estimated odds ratio of tissue-specific PPIs in the liver (Median GTEx OR = 7.85, Median Aggregate OR = 38.85, p = 0.03) and skin (Median GTEx OR = 21.83, Median Aggregated OR = 27.09, p = 0.05), and there was weak enrichment in adipose (Median GTEx OR = 18.49, Median Aggregated OR = 21.14, p = 0.06). In other tissue contexts, we observed a higher median odds ratio of PPI enrichment in networks inferred by data aggregation compared to GTEx-only networks **(Fig. 2 J)**. Finally, we observed that context-specific networks obtained from blood, skin, central nervous system, liver, lung, and adipose were able to recapitulate known TF-target relationships from TFLink to varying degrees. The tissue with the highest number of samples, blood, resulted in networks with the best performance in reproducing TF-Target relationships as indicated by precision and recall **(Supplemental Fig. S15)**. Thus, we demonstrated that data aggregation led to the improved inference of consensus networks that capture canonical gene interactions and tissue context-specific networks that capture tissue biology by observing better network generalizability and reproduction of known biological processes.

Central network nodes are evolutionarily constrained and include genes that are critical to tissue identity

The biological information captured by a GCN can be evaluated by comparing individual network edges, by examining whether there are edges between genes that are known to interact in a particular cellular pathway (Ha, Baladandayuthapani, and Do 2015), or by examining the properties of network nodes with a high number of connections or hubs (Pastor-Satorras, Rubi, and Diaz-Guilera 2003; Chou et al. 2014; Oh et al. 2015). Since eukaryotic transcriptional networks typically consist of a subset of genes, often transcription factors, that regulate many downstream target genes (Albert 2005), we chose specific values of the penalization parameter λ and number of resulting network edges such that the selected networks are approximately scale-free **(Methods, Supplemental Fig. S18; Supplemental Table S6, Supplemental Table S7)**. We computed different measures of centrality corresponding to each network node and tested for the enrichment of genes involved in GO terms that reflect ubiquitous or context-specific processes **(Supplemental Table S8)** among network nodes selected with progressively increasing thresholds for degree centrality against a background of all 18,882 protein-coding genes **(Methods)**. Central nodes from all three consensus networks were strongly enriched for genes involved in functions such as microtubule-based process, chromosome organization, and regulation of organelle organization **(Fig. 3 A, Supplemental Fig. S19)**. In contrast, we found that central nodes from blood context-specific networks were enriched for genes associated with platelet activation, leukocyte differentiation, and leukocyte chemotaxis to a greater extent than central genes derived from either consensus or a discordant context-specific network (CNS) **(Fig. 3 B)**. Further, these trends were reflected across multiple tissue contexts; context-specific networks corresponding to CNS **(Fig. 3 C)**, skin, liver, and lung were enriched

for genes associated with tissue-matched GO terms **(Supplemental Fig. S20)**. In addition, we examined the enrichment of central genes from tissue-specific gold standards obtained from HumanBase. We observed that network nodes with higher degree centrality were enriched for hub nodes (with a degree in the top 80^{th} percentile) of concordant tissue-specific reference networks, as evidenced by an odds ratio > 1 **(Supplemental Fig. S21)**. Thus, while central genes from consensus networks included genes involved in essential cellular processes, context-specific gene relationships are lost with global aggregation and are unlikely to be recovered by increasing the sample size.

Next, we evaluated whether hub genes were evolutionarily constrained and whether they have known roles in complex traits or diseases **(Methods)**. We first binned network nodes such that genes with a closeness centrality of 0 were assigned to Quintile 1, and those with at least 1 connecting edge were grouped based on estimated quantiles of closeness centrality. We then computed the excess overlap of a particular gene set among network nodes binned by closeness centrality, as the ratio of the fraction of genes from the set found among the network nodes in a particular bin to the fraction of genes from the gene set found among all network nodes **(Methods)** . Consensus network hub genes (Quintile 5) had a significantly higher excess overlap (>1) with evolutionarily constrained gene sets than peripheral nodes (Quintiles 1-3), using metrics including high pLI genes, high s_{het} genes, and high missense Z-score genes. These trends were similar in context-specific networks with some variation in the strength of the trend across tissues likely driven by sample size differences and differential power in inferring these networks **(Fig. 3 D, Supplemental Fig. S22)**. We then examined the enrichment of eQTL-deficient, ClinVar, OMIM, and FDA drug-targeted genes across quintiles of network centrality **(Supplemental Fig. S22, S23, Supplemental Table S9)**. We observed that peripheral genes

(Quintile 1) and central genes (Quintile 5) of consensus networks exhibited an excess overlap > 1 of eQTL deficient genes (those which lacked significant *cis*-regulatory variants), while genes with intermediate connectivity were depleted of eQTL deficient genes, while the six context-specific networks had widely variable trends. While central genes from consensus networks were weakly depleted for OMIM and FDA drug-targeted genes, we observed an excess overlap > 1 of hub genes belonging to context-specific networks. This could be explained by our earlier observations that central nodes from consensus networks were involved in essential and non-specific cellular processes while central nodes from context-specific networks were both specific and critical to tissue identity, thus, altering central consensus network hub nodes may have widespread and potentially deleterious off-target effects while context-specific hub nodes may identify targetable genes with tissue-specific effects. Finally, we observed that matched tissue-specific transcription factors (TFs) from Pierson et. al (Pierson et al. 2015) **(Supplemental Table S10)** had a significantly greater number of neighbors than 88 general TFs for context-specific networks corresponding to blood ($p < 1e-4$), lung ($p < 1e-4$), and skin ($p < 0.01$), while in other tissues, except for CNS, the median degree of tissue-specific TFs was greater than general TFs and non-transcription factors **(Methods)**. In contrast, while tissue-specific and general TFs had more neighbors in consensus networks when compared to non-transcription factors, we found no significant differences in the degree distribution of tissue-specific TFs to general TFs in either the universal or non-cancer consensus network **(Fig. 3 E)**. Therefore, while both consensus and context-specific central genes were enriched for genes under high-selection pressure, context-specific hub nodes were more likely to be OMIM genes, drug targets and tissue-specific TFs.

We examined similarities and differences between network architectures of the non-cancer consensus and cancer consensus networks based on shared and distinct hub genes, since central genes are likely to be more relevant to network functionality (Pastor-Satorras, Rubi, and Diaz-Guilera 2003), and the identification of hubs has led to the discovery of genes involved in cancer (Chou et al. 2014; Oh et al. 2015), tissue regeneration (Rodius et al. 2016), and other diseases (Keller et al. 2008; Presson et al. 2008). Specifically, we found 296 shared hubs between cancer (λ = 0.24, 7552 edges) and non-cancer (λ = 0.18, 7,355 edges) consensus networks, and 312 hubs which were specific to the cancer consensus network **(Supplemental Table S11)**. Cancer-specific hub genes were enriched for pathways such as ncRNA metabolic processing (GO:0034660, p = 5.2e-02) which is believed to play a role in metabolic reprogramming in cancer, DNA damage response (GO:0006974, p = 1.17e-12) which has been posited to play a role in cancer cell survival in non-optimal conditions, and response to ionizing radiation (GO:0010212, p = 6.1e-04). Further, we found that cancer-specific hub genes were enriched for a plethora of DNA repair and replication pathways including, double-strand break repair via homologous recombination (GO:0000724, p = 6.36e-06), recombinational repair (GO:0000725, p = 1.35e-06), interstrand cross-link repair (GO:0036297, p = 5.48e-03), and double-strand break repair via break-induced replication (GO:0000727, p = 1.36e-03) **(Supplemental File 1)**.

To examine whether network properties were shared between related tissues we compared the overlap of hub, we considered all hub genes found in at least two tissue contexts (N=1,956). Details of the penalization parameter selected for tissue-context specific networks used in this analysis are summarized in **Supplemental Table S12**. Grouping tissue contexts based on hub genes using non-negative matrix factorization with eight latent factors **(Methods)**

led to the grouping of related tissues such as blood and PBMCs/T cells (Factor 7) **(Fig. 3 F, Supplemental Table S13)**. As expected, we found that hub genes that led to the grouping of blood and PBMCs/T cells were enriched for defense response (GO:0006952, p = 2.7-24) and cytokine production (GO:0001816, p = 9.9e-07) **(Supplemental File 2)**. Further, Factor 6 which led to the grouping of hESCs, iPSCs, and multipotent cells comprised of hub genes which were enriched for gastrulation (GO:0007369, p = 5.9e-04) and circulatory system development (GO:0072359, p = 6.2e-03) **(Supplemental File 3)**.

Genes with high network centrality are proximal to variants enriched for complex trait heritability

Previous work by Kim et al. (Kim et al. 2019) reported that network topology annotations did not contribute to heritability once the LDSC baseline model (Finucane et al. 2015) was included. We examined whether data aggregation would increase the utility of network features for heritability analysis independently of baseline functional annotations. For each network we calculated centrality annotations by computing 6 different centrality measures including, degree, maximum weight, strength, closeness, betweeness, and PageRank. The correlation between the different centrality measures is summarized for select networks in **Supplemental Fig. S24**. We meta-analyzed estimates of heritability enrichment, the ratio of the proportion of heritability explained by SNPs belonging to an annotation to the proportion of SNPs in the annotation, and τ^*, an estimate of the heritability of SNPs unique to the annotation (Finucane et al. 2015), using a random effects model to obtain a summary of effect sizes estimated for a set of 42 independent traits considered by Kim et al. **(Methods, Supplemental Table S14)**. We estimated both heritability enrichment and τ^* by either conditioning an annotation corresponding to whether a

variant was located in a 100 kilobase window of all protein-coding genes (all-genes annotation), or conditioning on 97 functional annotations such as known enhancer and promoter regions which are included in the baseline-LD model and the all-genes annotation (all-genes + baseline). Similar to the results found by Kim et al., we observed a significant estimate τ^* corresponding to our consensus network-derived annotations when conditioning on just the all-genes annotation, however, when conditioning on the baseline-LD model, the τ^* observed for consensus network-derived annotations were no longer significant **(Fig. 4 A)**. While we found no significant differences in enrichment across a varying number of edges present in the network when conditioned on the all-genes annotation, we observed that τ^* decreased as the number of network edges decreased. There were no significant differences in either enrichment or τ^* with the number of edges found in the network when conditioning on the baseline-LD annotations **(Supplemental Fig. S25)**. Further, our observations were not dependent on the traits studied and remained consistent when we applied s-LDSC to 219 UKBB traits **(Supplemental Table S15; Supplemental Fig. S26).**

A possible explanation for the lack of heritability enrichment signal unique to network annotations is the redundancy between the baseline LD annotations and network topology annotations. Therefore, we hypothesized that context-specific data aggregation could prioritize variants enriched for heritability of concordant traits independent of baseline annotations. We applied s-LDSC to network centrality annotations derived from networks inferred only from GTEx blood samples (blood GTEx), networks inferred by aggregating recount3 blood samples (blood), and as a control, networks inferred by aggregating all samples (universal consensus), for a subset of 9 blood-related traits from the 42 independent traits (Crohn's disease, rheumatoid arthritis, ulcerative colitis, eosinophil count, platelet count, red blood cell count, red blood cell

width, white blood cell count, and eczema) **(Supplemental Table S16)**. As with the consensus networks, tissue-specific networks displayed similar trends in heritability estimates with the number of edges found in the network **(Supplemental Fig. S27, S28)**. When we conditioned on the baseline-LD annotations, we observed that annotations derived from the blood consensus networks had a significant τ^* across all centrality annotations, while blood GTEx networks had a significant τ^* for strength, degree, maximum weight and PageRank **(Fig. 4 B)**. In contrast, we did not observe a significant τ^* corresponding to annotations derived from the universal consensus network. We examined the generalizability of our results by conducting a similar experiment in CNS samples, another tissue with a large sample size. We applied s-LDSC to annotations derived from CNS networks inferred from GTEx samples (CNS GTEx), CNS networks inferred by aggregating samples from recount3 (CNS consensus), and the universal consensus networks for CNS-related traits which included waist-hip ratio adjusted BMI from the earlier set of 42 traits, as well as 7 traits from the Psychiatric Genomics Consortium (Alzheimer's, epilepsy, Parkinson's, bipolar disorder, smoking cessation, schizophrenia, major depressive disorder, and number of alcoholic drinks per week) **(Supplemental Table S17)**. We found that annotations derived from both universal consensus and CNS-specific networks led to significant non-zero τ^* when conditioning on the baseline-LD model. While we note that significant non-zero τ^* was observed for the consensus networks for the chosen set of CNS traits in contrast to the 42 independent traits, possibly due to power, study quality, and other attributes of the GWAS, we found that annotations from tissue-specific networks led to significantly higher estimates of τ^* and outperformed consensus networks **(Fig. 4 C)** for all centrality measures except closeness. Further, for betweenness, maximum weight, and PageRank centrality, CNS consensus networks

outperformed CNS GTEx networks, similar to the results in blood, demonstrating context-specific data aggregation results in network annotations that are enriched for trait heritability across tissue contexts. Across both sets of blood- and CNS-related traits, we found that PageRank centrality-derived annotations, which captured both the number of connections that a node has in addition to the centrality of its neighbors to determine the importance of a connection, performed consistently well. We conclude that context-specific aggregation results in the identification of the central genes of the network which are enriched for the heritability of concordant traits, and an increased sample size leads to a greater heritability enrichment signal.

Discussion

GCNs aid in determining changes in regulatory mechanisms that are key to cellular identity and prioritizing genes that drive phenotypic variability. However, conventional network analyses are often too underpowered to reliably discover gene-gene relationships and are compromised by spurious false positives and false negatives that result from limited power, noise, and unobserved technical confounders. We leveraged publicly available RNA-seq data from recount3 and manually curated tissue/cell type annotations to improve the inference of consensus and context-specific GCNs. Utilizing data splits, we demonstrated that accounting for confounders within individual studies followed by weighted aggregation of empirical covariance matrices led to the best improvement in network characteristics with data aggregation across multiple paradigms.

We then inferred three consensus networks (universal, non-cancer, and cancer networks) that recapitulated ubiquitous biological processes. Further, we aggregated data belonging to individual tissue contexts to infer 27 tissue context-specific networks that were enriched for

matched tissue-specific PPIs and shared similarities across related tissues. All networks and sample annotations are made publicly available as a resource for future studies.

Central genes from both consensus and context-specific networks were enriched for high PLI, and high Phi genes, indicating that hub genes are enriched for genes under high selective pressure. Context-specific hub genes were enriched for FDA-approved drug targets and OMIM genes while central genes from consensus networks which were inferred over a greater number of samples were depleted for both categories. Thus, context-specific information was lost by global aggregation, cannot be recovered by data aggregation or increased sample sizes, and is important to identifying drug targets and disease mechanisms. While the central genes of the network as determined by global data aggregation in the consensus network did not explain trait heritability independent of known functional annotations in the baseline-LD model, we found that context-specific data aggregation prioritized variants enriched for concordant trait heritability that did not overlap with previously known functional annotations. Thus topological properties of genes from context-specific GCNs hold significant promise as a functional annotation for identifying genetic variation that contributes to complex trait heritability.

A commonly used approach to identify genes associated with complex traits is to use colocalization analysis between GWAS and eQTL studies, however, often only about half of the signals colocalize with an eQTL (Mostafavi et al. 2023). Recent work by Mostafavi et al (Mostafavi et al. 2023) demonstrated that genes driving GWAS signals were often genes with complex context-dependent regulatory architecture and were depleted for eQTL variants. This has raised a call in the computational genomics community for orthogonal approaches to identify genes involved in complex traits. We found that annotations derived from context-specific GCNs are informative of trait heritability independent of context-agnostic functional annotations. This

suggests that tissue- and context-specific network centrality and other network properties could be used to help prioritize genes near GWAS loci (Zhu, Duren, and Wong 2021) or supplement eQTL data.

One of the major challenges in network inference remains the presence of unobserved technical confounders and undesirable biological signals which leads to spurious network edges and precludes causality claims. While PC-based data correction has been extensively utilized to reduce false positives resulting from confounding, recent work by Cote et. al (Cote, Young, and Huckins 2022) suggests that PC-based data correction, and related methods such as PEER (Stegle et al. 2012) and CONFETI (Ju et al. 2017), may over-correct expression data and remove biological co-expression of potential interest. Correcting or modeling confounders is essential to network accuracy, so tuning parameters such as the number of latent factors to correct as well as exploring alternative methods will continue to be important. Alternate approaches to handle confounding and infer causal regulatory relationships include instrumental variable analysis through the construction of local genetic instruments as outlined by Lujik et al. (Luijk et al. 2018), however, since central network nodes are evolutionarily constrained and tightly regulated, it can be challenging to construct well-tracking genetic instruments for central genes,. Publicly available RNA-seq data including recount3, the extensive annotations we provide, and recent work which illustrated genotype calling using RNA-seq data (Deelen et al. 2015) could improve our ability to detect context-specific cis-regulatory effects, the reconstruction of local genetic instruments, and hence causal regulatory network inference.

Future directions aimed at improving GCN inference could leverage our extensively annotated sample characteristics and data aggregation strategies with complementary strategies including sharing information between related contexts (Omranian et al. 2016) to increase the

effective sample size, introducing constraints or priors corresponding to known regulatory relationships (Hellstern et al. 2021), and using alternate statistical measures of expression similarity that capture nonlinear associations between genes (Margolin et al. 2006). Additionally, heuristic algorithms such as the one proposed by Opgen-Rhein et al. (Opgen-Rhein and Strimmer 2007) could be utilized to enrich our current networks with directionality information. Finally, while we studied tissue-contexts, we provide annotations of disease status which can be utilized to infer disease-specific GCNs.

Our finding that marginal improvement in network reconstruction decreases with continued data aggregation suggests that simply addressing statistical considerations due to sample size may have limitations for improving GCNs. Including orthogonal sources of information such as gene-enhancer associations inferred from Hi-C data (Babaei et al. 2015), transcription factor binding sites from ChIP-seq data (Zhou et al. 2016), and regulatory information derived from other epigenetic data types, in addition to gene expression quantified by RNA-seq in both bulk and single-cell studies, might result in a more accurate understanding of the shared regulatory architecture between genes. An example of integrating orthogonal information sources is found in recent work by Fu, X. et al. (Fu et al. 2025), which leverages paired scATAC-seq and snRNA-seq data across 213 fetal and adult cell types, along with prior knowledge of TF motifs to computationally infer local chromatin environments. The resulting model can then be used to predict gene expression in new cell types and TF regulators of downstream genes. This method and related approaches based on epigenetic data and TF motif information to infer targets of well-characterized TFs can be viewed as complementary to our approach, which is able to capture regulatory and co-expression relationships genome-wide based on large-scale RNA-seq data. Future methods may unify these approaches and incorporate

both sources of information. Additionally, experimental protocols such as Perturb-Seq (Dixit et al. 2016), which quantifies the transcriptional changes mediated by genetic manipulations on genes, processes, and states, could provide a new avenue for network inference and suggest causal mechanisms and edge directionality (Ota et al. 2025).

In conclusion, the growing availability of publicly shared RNA-seq data presents a valuable opportunity to improve gene GCN inference through data aggregation. While GCNs offer insights distinct from eQTL analyses and hold promise for uncovering the regulatory mechanisms underlying complex traits, their utility has been limited by challenges in data integration and variability across studies. In this work, we addressed these challenges by developing an approach that accounts for latent confounding and study-specific variability, enabling the construction of context-specific GCNs enriched for complex-trait heritability beyond baseline functional annotations.

Methods

Data pre-processing and quality control

We downloaded uniformly processed human RNA-seq samples using the recount3 R package (version 1.0.7; Wilks et al., 2021), running on R version 4.0.2 (R Core Team 2020) under Rocky Linux 8.8, and selected 1,747 projects that included 30 or more samples each **(Fig. 1A)**. Before normalization, we excluded samples with zero expression across all genes and genes that had zero expression across all samples in a project. We used in-built functions from the recount3 package to compute the RPKM transformed count matrix, selected genes that were protein-coding, autosomal, and unambiguously mapped to the reference genome (Battle 2019; Saha and Battle 2018), and generated the $log_2(RPKM + 1)$ count matrix for each project.

Following preliminary processing, we applied a unique data processing pipeline based on the data source. For projects belonging to GTEx, we excluded duplicates (labeled as STUDY_NA), and samples derived from the chronic myelogenous leukemia (CML) cell line. We grouped samples by the tissue of origin to obtain 50 groups and 18,828 samples.

For TCGA, we excluded 67 samples missing sample type and 39 samples lacking patient ID. Replicates were identified as samples sharing both patient ID and sample type, and their gene expression was aggregated using the median. We grouped TCGA samples by cancer code, resulting in 33 groups and 11,091 samples.

For SRA, we first removed samples obtained via size fractionation. Replicates were defined by identical experiment accession numbers and aggregated using the median expression. Further, we excluded 89,101 samples where more than 50 percent of genes showed zero expression, which we used as a threshold to eliminate likely microRNA or degraded samples. To further restrict the dataset to bulk RNA-seq, we used predicted experiment-type labels from recount3 and excluded samples predicted as single-cell or small RNA-seq. If predicted labels were unavailable, we excluded samples with keywords such as “single cell,” “scRNA,” “snRNA,” or “single nucleus” in the study abstract. Additionally, we retained only those samples whose library selection metadata contained either cDNA or RT-PCR, when available. We also excluded studies which are known scRNA-seq experiments including: SRP096986, SRP135684, SRP166966, SRP200058, and SRP063998. For the remaining studies, we performed a text-based analysis to obtain the Study, Tissue, Organ, Biopsy, Cell, Disease, Source and Description from the metadata `sample_description` field. We then manually annotated 10,179 unique combinations of these fields to obtain tissue, cancer status, and disease type. In this manner, we

were able to obtain labels for 65,361 samples which we grouped on the basis of the study accession IDs to form 884 SRA studies **(Supplemental File 4)**.

We grouped GTEx samples by tissue of origin, SRA samples by study accession ID, and TCGA samples by cancer accession code, and have referred to each individual group of samples as a "study" to simplify nomenclature. We did not distinguish on the basis of disease state or cancer status while organizing the data, until we proceeded to compute the inputs to network inference. The number of genes retained after filtering varied by study, particularly in smaller SRA projects with limited sample size. These gene count distributions are summarized in **Supplemental Fig. S1**.

Identifying tissue type and cancer status

Wilks et. al (Wilks et al. 2021) demonstrated that for human bulk RNA-seq data, tissue or cell type of origin is the dominant source of gene expression variation, with clear tissue-specific clustering visible in the top four principal components. Building on this we manually refined annotated labels **(Supplemental File 4)** and grouped 95,484 samples and 5,999 genes (with non-zero variance) using t-SNE dimensionality reduction and clustering. Because cancerous and non-cancerous samples did not separate distinctly in the t-SNE space, we used manually annotated labels to restrict the analysis to 63,193 non-cancerous samples.

To maximize sample size, we merged similar tissues into broader tissue contexts **(Supplemental Table S1)**, defined as a groups of samples with similar cell-type composition. Recognizing that gene expression can also vary due to technical factors such as batch effects and library size, we estimated the relative enrichment of 64 stromal and immune cell types using `xCell` gene signatures (Aran, Hu, and Butte 2017). Enrichment vectors for each sample were

then used for visualization and clustering via t-SNE, allowing us to group samples by shared cellular composition into distinct tissue contexts.

For 24 contexts with more than 500 samples, we removed outliers using six different outlier detection methods **(Supplemental Methods: Outlier calling methods to refine sample selection, Supplemental Fig. S2)**. We examined whether selecting outliers based on any metric, including outliers labeled by 1, 2, or 3 metrics significantly impacted our estimate of the empirical covariance matrix from the data. We compared the frobenius norm of the differences in empirical covariance estimated for a particular tissue context with varying stringency on the inclusion criterion to the difference in the empirical covariance matrix estimated across contexts. We observed that the differences in the covariance estimates for varying the exclusion criterion were in general much smaller than the differences between two unrelated contexts, and hence we included samples within a tissue-specific context if it was labeled an outlier by 2 or fewer metrics to maximize the number of samples available for downstream analysis **(Supplemental Fig. S4)**.

While this works with simpler tissue categories such as blood, skeletal muscle or colon where the samples have relatively homogenous cell type compositions, for the immune system, which comprises samples from the distinct myeloid, lymphoid and innate immune systems, these outlier metrics failed due to large intersample variation. Therefore for immune cell types, we first performed K-means clustering with 3 centroids using the in-built `kmeans()` function. We annotated the three resulting clusters based on the representation of manually annotated labels in each cluster as either B cells, myeloid cells (including monocytes and macrophages) and PBMCs w/ T cells **(Supplemental Fig. S3)**. For each cluster, we performed outlier detection using all six methods outlined above and excluded samples that were found to be outliers in any two methods.

Data correction and aggregation

Principal component (PC) based correction methods can account for technical and biological artifacts that confound gene expression measurements and reduce false positives in gene network inference (Parsana et al. 2019). However, these methods have been applied to one experiment and not across multiple experiments from disparate sources. We systematically compared 4 strategies of data aggregation. In the first approach (aggregating data), expression data were pooled across all studies of interest, and confounders were estimated jointly from the combined dataset. This strategy reflects a global correction model in which confounding structure is assumed to be shared across datasets.

In the second approach (aggregating data adjusted for confounding), confounder correction was performed independently within each study before combining the corrected data. This addresses the possibility that latent structure differs across studies. The third and fourth approaches involved aggregating covariance matrices rather than gene expression matrices. In the unweighted aggregation of covariance matrices strategy, we treated each study equally and computed the average of study-level covariance matrices after within-study correction. In contrast, the weighted aggregation of covariance matrices strategy assumed that larger studies yield more reliable estimates, and we weighted each study's contribution by its sample size. Both approaches avoid direct normalization across heterogeneous datasets and instead pool structure at the level of gene–gene covariance. Complete implementation details for all four strategies are provided in the **Supplemental Methods: Detailed procedures for data correction and aggregation strategies**.

Network reconstruction with graphical lasso

Following the computation of aggregate covariance matrices using the strategies outlined in **Supplemental Methods: Data correction and aggregation**, we inferred gene regulatory relationships using graphical lasso (Friedman, Hastie, and Tibshirani 2008). The desired network structure is obtained by identifying the precision matrix, $\Theta = \Sigma^{-1}$ that maximizes the penalized log-likelihood given by Eqn. 1, where C is the estimated covariance matrix.

$$\widehat{\Theta} = argmin_{\theta} \; tr(C.\Theta) - log|\Theta| + \lambda||\Theta||_1 \qquad \text{Eqn. 1}$$

We estimated the precision matrix Θ across a range of λ between 0.04 to 1.00 in intervals of 0.02 using the R package `QUIC` (version 1.1.1). For genes p and q, an edge connecting them exists if the corresponding entry in the precision matrix is non-zero as given by Eqn. 2

$$\widehat{N}_{p,q} = \begin{cases} 1, & if \; |\widehat{\Theta}_{p,q}| > 0 \\ 0, & otherwise \end{cases} \qquad \text{Eqn. 2}$$

Network evaluation to determine the optimal aggregation strategy

To evaluate the impact of different data correction and aggregation strategies on network quality, we applied all four approaches (see **Supplemental Methods: Detailed procedures for data correction and aggregation strategies**) to non-cancerous samples from two partitions: GTEx and SRA. GTEx samples were organized by tissue type, excluding tissues with fewer than 15 samples (e.g., Kidney Medulla), resulting in 49 tissue-level groups. For the SRA non-cancer partition, we selected 566 studies with non-cancer annotations and at least 15 samples per study.

To assess how increasing sample size affects network performance, we constructed aggregate networks by sequentially adding studies or tissues in order of increasing sample size.

For SRA, we evaluated networks at 1, 100, 200, 300, 400, 500, and all 566 studies. For GTEx, we tested 1, 10, 20, 30, 40, and 49 tissues. Each network was inferred using graphical lasso over a grid of penalization parameters between 0.04 and 1.00. **Supplemental Table S2** summarizes the number of samples, studies, median sample size, and principal components regressed at each aggregation level.

To compare network quality across aggregation strategies and sample sizes, we used two evaluation criteria: (1) held-out log-likelihood on independent datasets to measure generalizability and (2) biological validity based on (a) enrichment for known pathway co-membership and (b) transcription factor–target interactions. Precision, recall, and F1-score were computed for each network using standardized reference sets. Full computational details for all evaluation metrics are provided in the **Supplemental Methods: Detailed evaluation metrics to determine the impact of data aggregation**.

Network inference with WGCNA

To evaluate the effect of data aggregation on module-based gene co-expression networks, we used the `WGCNA` R package (Langfelder and Horvath 2008) version 1.72-5, (Langfelder and Horvath 2008). We constructed consensus networks by aggregating samples incrementally across three settings: (1) GTEx tissues in batches of 1, 3, 10, 30, and all 49 tissues; (2) non-cancerous SRA studies in batches of 1, 10, 20, 30, 100, 300, and all 566 studies; and (3) SRA studies annotated as blood in batches of 1, 20, 40, 60, and 65. For each aggregation level, we applied principal component correction to the individual studies or tissues before combining them. We retained only genes that were present in all datasets being aggregated. After performing standard WGCNA preprocessing, we estimated the adjacency matrix for each individual dataset using a power parameter selected to ensure approximate scale-free topology. Consensus networks were

then constructed by aggregating topological overlap matrices (TOMs) across studies. Modules were identified by hierarchical clustering followed by dynamic tree cutting from the WGCNA package, and similar modules were merged based on eigengene correlation. For each inferred network, we evaluated functional relevance by computing precision, recall, and F1-score based on known gene-gene interactions obtained by curating pathway information from KEGG, Biocarta and Pathway Interaction Database from Enrichr and selected those pathways that were annotated as canonical pathways by MSigDB.. Full computational procedures and parameter thresholds are provided in the **Supplemental Methods: Detailed procedure for WGCNA-based aggregation and evaluation**.

Inference of consensus and tissue context-specific networks

We inferred consensus networks across samples from disparate tissues and cell types to capture shared biological pathways across contexts. Since weighted covariance aggregation yielded the best performance among the four data correction and aggregation strategies, we grouped SRA samples by study accession ID, GTEx samples by tissue of origin, and TCGA samples by cancer code. We inferred the network structure over a range of penalization parameters from 0.08 to 1.00. We constructed three types of consensus networks (universal, non-cancer, and cancer), each defined by different sample inclusion criteria (see **Supplemental Methods: Construction of consensus networks** for details).

We inferred context-specific networks in 27 contexts with 500 or more samples. Details of the number of samples, studies, and median sample size across studies for each context are provided in **Supplemental Fig. S16**. For 20 contexts including adipose, B cells, blood, breast, cardiac, central nervous system, colon, esophagus, fibroblasts, intestine, kidney, liver, lung,

nervous system, pancreas, prostate, skeletal muscle, skin, stomach, and vascular we inferred networks either using GTEx sample only or by aggregating context-specific samples from recount3 which included GTEx. For each context-specific network, we first performed PC-based data correction within each study followed by covariance estimation and aggregation by weighting the covariance matrix with the proportion of study-specific sample size to the total number of samples as detailed in **Supplemental Methods: Detailed procedures for data correction and aggregation strategies**. We then inferred GCNs using graphical lasso using Eqn. 1 and Eqn. 2.

Evaluating the impact of data aggregation on the inference of consensus and context-specific networks

To evaluate whether aggregating samples across studies improves the performance of consensus and context specific networks, we computed the precision, recall, and F1 score of observing known gene co-regulatory relationships annotated in canonical biological pathways compiled across KEGG, Biocarta and the Pathway Interaction Database, as detailed in **Supplemental Methods: Detailed evaluation metrics to determine the impact of data aggregation**. For consensus networks we compared F1 scores across the universal, non-cancer, and cancer networks over a range of network sizes from 5,000 to 500,000 edges.

To test context-specific network performance we focused on the two largest contexts: blood and CNS. For each of these contexts, we sequentially aggregated SRA studies by increasing sample size and inferred networks. We then evaluated the generalizability of the networks inferred across varying numbers of samples by computing the held-out log-likelihood

using the corresponding GTEx tissues as held-out test sets **(Supplemental Methods: Computation of held-out log-likelihood for content-specific networks)**.

Additionally, for 6 contexts (adipose, blood, CNS, liver, lung, and skin) we compared context specific networks constructed solely from GTEx to those derived through data aggregation by assessing the enrichment of tissue specific protein-protein interactions from the SNAP database. Odds ratios for enrichment were computed using Fisher's exact test. Additional statistical procedures, including two types of permutation testing and a Wilcoxon rank sum test for comparing aggregated versus GTEx only networks, are described in the **Supplemental Methods: Determining enrichment of known tissue-specific PPI interactions**.

Sparsity parameter selection for consensus and context-specific networks

For each of the three consensus networks and 27 context specific networks, we evaluated the degree distribution of inferred networks across a range of penalization parameters. Specifically, we identified scale free networks as those where the degree distribution followed a power law with an estimated scaling exponent between 2 and 3 and a coefficient of determination (R^2) greater than 0.8. Additional methodological details, including model fitting and confidence interval calculations, are provided in the **Supplemental Methods: Detailed description of sparsity parameter selection**. Summary statistics and selected network parameters are reported in **Supplemental Tables S6** and **S7**.

Computing gene centrality measures based on network structure

We computed measures of network connectivity for each gene in the network with the `igraph` R package (version 1.3.5) (Csardi and Nepusz 2006). We used the absolute value of entries in the precision matrix (**Methods: Network reconstruction with graphical lasso)** to

define edge weights between genes and normalized these values by the maximum edge weight across the entire graph. Using these weights, we calculated gene-level centrality using several standard metrics: betweenness, closeness, degree, strength, maximum edge weight, and PageRank. For each gene, we used distance-based weights to compute betweenness and closeness centrality, while normalized edge weights were used for the remaining centrality measures. Details of how each centrality metric was computed, including mathematical formulas and algorithms used, are provided in the **Supplemental Methods: Computation of centrality measures**. Correlations between centrality measures across select networks are shown in **Supplemental Fig. S24**.

Enrichment of specific pathways among central genes in consensus and context-specific networks

To assess whether highly connected genes in inferred networks were enriched for known biological functions, we developed a general framework to test for the enrichment of functional gene sets among central genes. For each network, we estimated node degree (see **Methods: Computing gene centrality measures based on network structure**) and assessed functional enrichment at a series of increasing degree thresholds **(Supplemental Methods: Functional enrichment analysis of central genes)**. For the consensus networks we first selected the value of the penalization parameter such that the resulting network had ~ 7,000 edges. We then determined the odds ratio of finding genes belonging to ubiquitous biological processes among network nodes selected by successively increasing degree thresholds and compared the results to those obtained from blood and CNS-specific networks with a similar number of edges. Full

details of GO terms corresponding to ubiquitous biological processes is provided in **Supplemental Table S8**.

We extended this analysis to context-specific networks by selecting GO terms relevant to each context and testing for enrichment among high-degree nodes. For each of six contexts (blood, CNS, skin, lung, liver, and adipose), we selected one network inferred solely from GTEx samples and one inferred from aggregated samples (including GTEx) with ~7,000 edges. For example, in the blood context, we tested enrichment for GO terms such as leukocyte migration, leukocyte activation, and blood coagulation among high-degree nodes. As negative controls, we compared the observed enrichment in the blood-specific networks to that found in an unrelated context-specific network inferred across aggregated samples, as well as the universal and non-cancer consensus networks. A complete list of GO terms tested for each context is provided in **Supplemental Table S8**.

In a separate analysis, we compared inferred context specific networks to gold standard tissue specific networks from HumanBase (Greene et al. 2015). For each inferred network, we assigned nodes to degree-based bins and tested whether high-degree genes in our networks were enriched for hub genes (defined as those above the 80th percentile in degree) in the corresponding HumanBase network. Enrichment was tested across bins to assess the degree to which centrality in inferred networks aligns with regulatory importance in an external reference. Detailed procedures for these enrichment analyses are described in **Supplemental Methods: Enrichment of HumanBase hub nodes among central network nodes**.

Excess overlap of genes grouped by centrality measures with known evolutionarily constrained and functionally prioritized gene sets

We grouped genes for each consensus network such that the first bin includes genes with a closeness centrality of 0, the second bin contains nodes with closeness centrality in the lowest 25th percentile, the third bin includes genes with closeness centrality in the 25th - 50th percentile, the fourth bin includes genes with closeness centrality in the 50th - 75th percentile and the fifth and final bin includes genes with closeness centrality greater than the 75th percentile.

For each bin, we computed the excess overlap with evolutionarily constrained and functionally prioritized gene sets as defined in Kim et al. (2019). We quantified enrichment by comparing genes in bin i, G^{i}_{b} with each reference gene set j, G^{j}_{r}, relative to the set of all genes in the network G_{tot}, using the following equations.

$$P_{d} = \frac{|G^{j}_{r} \cap G^{i}_{b}|}{|G^{i}_{b}|} \quad \text{Eqn. 3}$$

$$P_{tot} = \frac{|G^{j}_{r} \cap G_{tot}|}{|G_{tot}|} \quad \text{Eqn. 4}$$

$$excess\ overlap\ (G^{i}_{b}, G^{j}_{r}) = \frac{P_{d}}{P_{tot}} \quad \text{Eqn. 5}$$

$$SE\ excess\ overlap\ (G^{i}_{b}, G^{j}_{r}) = \sqrt{\frac{P_{d}(1-P_{d})}{|G^{i}_{b}|}} / P_{tot} \quad \text{Eqn. 6}$$

Details on the reference gene sets are provided in the **Supplemental Methods:Reference gene sets used for overlap analysis** and in **Supplemental Table S9**.

Identifying biological processes associated with shared and distinct hub genes from non-cancer and cancer consensus network

In accordance with the scale-free criterion we selected the non-cancer network inferred using a penalization parameter $\lambda = 0.18$ resulting in a network with 7,355 edges and the cancer consensus network corresponding to the penalization parameter $\lambda = 0.24$ and a network with 7,552 edges to compare the biological processes which are represented by shared and distinct network hubs. First, we defined hub genes as network nodes with a closeness centrality in the 90th percentile independently for each consensus network, such that the non-cancer hub genes are given by H_N and the cancer hub genes are given by H_C. We then identified hub genes shared between cancer and non-cancer consensus networks $H_S = H_N \cap H_C$, non-cancer specific hub genes $H_{NS} = H_N \cap H_C'$, and cancer-specific hub genes as $H_{CS} = H_C \cap H_N'$ (**Supplemental Table S11**). We then used the `GOTermFinder` tool (Boyle et al. 2004) to identify GO terms that are shared by the genes in the sets H_S, H_{NS}, and H_{CS}.

Examining the differences in the degree distribution of tissue-specific vs. general transcription factors in consensus and context-specific networks

We selected context-specific networks inferred from CNS, blood, cardiac, skin, lung, skeletal muscle, and pancreas samples from recount3 which each had approximately 7,000 edges (**Supplemental Table S12**). Additionally, we selected values of the penalization parameter λ which yielded universal and non-cancer consensus networks with ~ 7,000 edges . We referred to Pierson et al. (Pierson et al. 2015) to obtain a list of tissue-specific and general transcription factors which are provided in **Supplemental Table S10**. To select a background set of non-TFs,

we first obtained the intersection of genes which were included in each network considered, then we excluded both general and tissue-specific TFs from this list, and randomly selected 100 of these genes. The selected background is provided in **Supplemental Table S10**. For each network, we compared the degree distribution of tissue-specific and general transcription factors to non-TFs and the degree distribution of tissue-specific to general transcription factors using the Wilcoxon rank-sum test. Across all tests, we adjusted for multiple hypothesis correction using the Holm-Bonferroni method.

Non-negative matrix factorization to determine shared co-regulatory relationships in similar tissues

We selected context-specific networks with ~7,000 edges across 27 contexts for which we inferred GCNs by aggregating samples assigned to the context from recount3 (**Supplemental Table S12**). Further, the selected networks were in accordance with the scale-free selection criterion detailed in **Methods: Sparsity parameter selection for consensus and context-specific networks**. For each network, we identified hub genes as network nodes with a degree centrality in the 95th percentile. Across the 27 contexts, we found 3,682 unique hubs. We subset to hubs that are present in at least 2 contexts which resulted in 1,956 hubs. We then used the R package `RcppML` (version 0.3.7) to perform non-negative matrix factorization to learn 8 underlying factors to group similar patterns of hub genes. Specifically, H is a binary matrix of dimensions $N_{Hubs} \times N_c$, where N_c is 27 the number of contexts. $H_{i,j} = 1$ when the i^{th} gene is a hub gene in context j and $H_{i,j} = 0$ otherwise. We then obtain matrices W of dimensions $N_{Hubs} \times 8$, and C of dimensions $N_c \times 8$ by solving the following optimization function,

$$min\ ||H\ -\ WC^{T}||\ \text{such that}\ W\ \geq 0,\ C\ \geq\ 0 \qquad \text{Eqn. 25}$$

Finally, we chose a solution after 10 iterations that resulted in the greatest sparsity, i.e. smallest values of $||W||_1$ and $||C||_1$. To interpret the resulting context cluster, we examined the genes that contributed the most to the corresponding factor. Specifically, for a given factor, $W[:, p]$, we first obtained genes with loadings in the 80% percentile. Then for each gene we computed the maximum difference between the loading of the gene on factor p to all other factors and selected genes where this difference is $\geq 5\ \times 10^{-5}$. Thus we obtained a set of factor-specific hub genes G_p (**Supplemental Table S13**). We then used the `GOTermFinder` tool (Boyle et al. 2004) to identify GO terms that were more likely to be present in the set G_p than a background of all protein-coding genes by estimating the odds ratio and p-value by applying the hypergeometric test.

Stratified LD-Score regression to quantify the heritability enrichment of variants proximal to central network genes in consensus and context-specific networks

We applied stratified LD-score regression (s-LDSC) (Finucane et al. 2015) to assess whether SNPs near central genes in inferred networks contributed disproportionately to complex trait heritability. For the following networks, universal consensus, non-cancer consensus, blood-specific (inferred from only GTEx data and across aggregated data from recount3), and CNS-specific (inferred from only GTEx data and across aggregated data from recount3), we obtained annotations corresponding to 6 different centrality measures, degree, betweenness, closeness, strength, and PageRank **(Methods: Computing gene centrality measures based on network structure)** by transforming the centrality scores to lie between 0 and 1. Each SNP was

annotated based on the centrality of genes within 100 kb, taking the maximum value when multiple genes were nearby. These annotations were tested for heritability enrichment and standardized effect (τ_c *) using s-LDSC under two models, (1) one including only an all-genes annotation, and (2) another including the all-genes annotation and 97 annotations from the baseline-LD v2.2 model (Gazal et al. 2017; Hujoel et al. 2019; Villar et al. 2015; Marnetto et al. 2018). Trait sets **(Supplemental Methods: Description of trait sets used in sLDSC)** included 42 curated traits (Kim et al., 2019, **Supplemental Tables S14**), 219 UK Biobank traits **(Supplemental Tables S15)**, 9 blood-related traits **(Supplemental Tables S16)**, and 9 CNS-related traits **(Supplemental Tables S17)**. All networks used in the analysis satisfied the scale-free degree criterion (see **Methods: Sparsity parameter selection**). Final enrichment and τ_c * estimates were summarized using random-effects meta-analysis by using the function `meta.summaries()` from the R package `rmeta` (version 3.0). Full procedural details and model specifications are provided in **Supplemental Methods: Stratified LD Score Regression Procedures**.

Data access

Scripts to reproduce the analysis and figures included in this manuscript are available in the Supplemental Code and can also be accessed at

https://github.com/prashanthi-ravichandran/recount3_networks

Networks and annotations used in this project can be obtained from:

https://zenodo.org/records/10480999

Competing interest statement

A.B. is a consultant for Third Rock Ventures, LLC, a shareholder in Alphabet, Inc, and a founder of CellCipher, Inc.

Acknowledgements

The authors would like to thank helpful discussions from Battle lab members throughout the course of this work, particularly Joshua Weinstock and Eric Kernfeld for code review and manuscript feedback. Alexis Battle was supported by NIH/NIGMS R35GM139580. Kaspar Hansen and Alexis Battle were supported by NIH/NIGMS R01GM121459.

Authors' contributions

P.R., P.P.., and A.B. conceived the project. P.R., P.P., and A.B. designed the analyses. P.R. and P.P. performed the analyses. R.K. and K.H. contributed feedback to experimental design and interpretation. P.R., R.K., and A.B. organized and wrote the paper with input from all authors.

References

Albert, Réka. 2005. “Scale-Free Networks in Cell Biology.” *Journal of Cell Science*. https://doi.org/10.1242/jcs.02714.

Alon, U. 2003. “Biological Networks: The Tinkerer as an Engineer.” *Science* 301 (5641): 1866–67.

Aran, Dvir, Zicheng Hu, and Atul J. Butte. 2017. “xCell: Digitally Portraying the Tissue Cellular Heterogeneity Landscape.” *Genome Biology* 18 (1): 220.

Babaei, Sepideh, Ahmed Mahfouz, Marc Hulsman, Boudewijn P. F. Lelieveldt, Jeroen de Ridder, and Marcel Reinders. 2015. “Hi-C Chromatin Interaction Networks Predict Co-Expression in the Mouse Cortex.” *PLoS Computational Biology* 11 (5): e1004221.

Battle, Alexis. 2019. “Pre-Computed Cross-Mappability Resources for Human Genomes (hg19 and GRCh38),” March. https://doi.org/10.6084/m9.figshare.c.4297352.v4.

Boyle, Elizabeth I., Shuai Weng, Jeremy Gollub, Heng Jin, David Botstein, J. Michael Cherry, and Gavin Sherlock. 2004. “GO::TermFinder--Open Source Software for Accessing Gene Ontology Information and Finding Significantly Enriched Gene Ontology Terms Associated with a List of Genes.” *Bioinformatics* 20 (18): 3710–15.

Buja, A., and N. Eyuboglu. 1992. “Remarks on Parallel Analysis.” *Multivariate Behavioral Research* 27 (4): 509–40.

Chen, Edward Y., Christopher M. Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R. Clark, and Avi Ma’ayan. 2013. “Enrichr: Interactive and Collaborative HTML5 Gene List Enrichment Analysis Tool.” *BMC Bioinformatics* 14 (April):128.

Chou, Wei-Chun, An-Lin Cheng, Marco Brotto, and Chun-Yu Chuang. 2014. “Visual Gene-Network Analysis Reveals the Cancer Gene Co-Expression in Human Endometrial Cancer.” *BMC Genomics* 15 (April):300.

Cote, Alanna C., Hannah E. Young, and Laura M. Huckins. 2022. “Comparison of Confound Adjustment Methods in the Construction of Gene Co-Expression Networks.” *Genome Biology* 23 (1): 44.

Csardi, Gabor, and Tamas Nepusz. 2006. “The Igraph Software Package for Complex Network Research.” *InterJournal*. https://igraph.org.

Dam, Sipko van, Urmo Võsa, Adriaan van der Graaf, Lude Franke, and João Pedro de Magalhães. 2017. “Gene Co-Expression Analysis for Functional Classification and Gene–disease Predictions.” *Briefings in Bioinformatics* 19 (4): 575–92.

Deelen, Patrick, Daria V. Zhernakova, Mark de Haan, Marijke van der Sijde, Marc Jan Bonder, Juha Karjalainen, K. Joeri van der Velde, et al. 2015. “Calling Genotypes from Public RNA-Sequencing Data Enables Identification of Genetic Variants That Affect Gene-Expression Levels.” *Genome Medicine* 7 (1): 30.

Diaz, Léo P. M., and Michael P. H. Stumpf. 2022. “Gaining Confidence in Inferred Networks.” *Scientific Reports* 12. https://doi.org/10.1038/s41598-022-05402-9.

Dixit, Atray, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, Nemanja D. Marjanovic, et al. 2016. “Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens.” *Cell* 167 (7): 1853–66.e17.

Finucane, Hilary K., Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, et al. 2015. “Partitioning Heritability by Functional Annotation Using Genome-Wide Association Summary Statistics.” *Nature Genetics* 47 (11): 1228–35.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2008. “Sparse Inverse Covariance Estimation with the Graphical Lasso.” *Biostatistics* 9 (3): 432–41.

Fu, Xi, Shentong Mo, Alejandro Buendia, Anouchka P. Laurent, Anqi Shao, Maria del Mar Alvarez-Torres, Tianji Yu, et al. 2025. “A Foundation Model of Transcription across Human Cell Types.” *Nature* 637 (8047): 965–73.

Gazal, Steven, Hilary K. Finucane, Nicholas A. Furlotte, Po-Ru Loh, Pier Francesco Palamara, Xuanyao

Liu, Armin Schoech, et al. 2017. “Linkage Disequilibrium-Dependent Architecture of Human Complex Traits Shows Action of Negative Selection.” *Nature Genetics* 49 (10): 1421–27.

Greene, Casey S., Arjun Krishnan, Aaron K. Wong, Emanuela Ricciotti, Rene A. Zelaya, Daniel S. Himmelstein, Ran Zhang, et al. 2015. “Understanding Multicellular Function and Disease with Human Tissue-Specific Networks.” *Nature Genetics* 47 (6): 569–76.

Ha, Min Jin, Veerabhadran Baladandayuthapani, and Kim-Anh Do. 2015. “DINGO: Differential Network Analysis in Genomics.” *Bioinformatics* 31 (21): 3413–20.

Hartwell, L. H., J. J. Hopfield, S. Leibler, and A. W. Murray. 1999. “From Molecular to Modular Cell Biology.” *Nature* 402 (6761 Suppl): C47–52.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2013. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

Hasty, Jeff, David McMillen, and J. J. Collins. 2002. “Engineered Gene Circuits.” *Nature* 420 (6912): 224–30.

Hellstern, Michael, Jing Ma, Kun Yue, and Ali Shojaie. 2021. “Netgsa: Fast Computation and Interactive Visualization for Topology-Based Pathway Enrichment Analysis.” *PLoS Computational Biology* 17 (6): e1008979.

Huang, Yu-Jyun, Tzu-Pin Lu, and Chuhsing Kate Hsiao. 2020. “Application of Graphical Lasso in Estimating Network Structure in Gene Set.” *Annals of Translational Medicine* 8 (23). https://doi.org/10.21037/atm-20-6490.

Hujoel, M. L. A., S. Gazal, F. Hormozdiari, B. van de Geijn, and A. L. Price. 2019. “Disease Heritability Enrichment of Regulatory Elements Is Concentrated in Elements with Ancient Sequence Age and Conserved Function across Species.” *American Journal of Human Genetics* 104 (4). https://doi.org/10.1016/j.ajhg.2019.02.008.

Ju, Jin Hyun, Sushila A. Shenoy, Ronald G. Crystal, and Jason G. Mezey. 2017. “An Independent Component Analysis Confounding Factor Correction Framework for Identifying Broad Impact Expression Quantitative Trait Loci.” *PLoS Computational Biology* 13 (5): e1005537.

Kanehisa, M., and S. Goto. 2000. “KEGG: Kyoto Encyclopedia of Genes and Genomes.” *Nucleic Acids Research* 28 (1): 27–30.

Katz, Kenneth, Oleg Shutov, Richard Lapoint, Michael Kimelman, J. Rodney Brister, and Christopher O’Sullivan. 2022. “The Sequence Read Archive: A Decade More of Explosive Growth.” *Nucleic Acids Research* 50 (D1): D387–90.

Keller, Mark P., Younjeong Choi, Ping Wang, Dawn Belt Davis, Mary E. Rabaglia, Angie T. Oler, Donald S. Stapleton, et al. 2008. “A Gene Expression Network Model of Type 2 Diabetes Links Cell Cycle Regulation in Islets with Diabetes Susceptibility.” *Genome Research* 18 (5): 706–16.

Kernfeld, Eric, Rebecca Kenner, Alexis Battle, and Patrick Cahan. 2024. “Transcriptome Data Are Insufficient to Control False Discoveries in Regulatory Network Inference.” *Cell Systems* 15 (8): 709–24.e13.

Kim, Samuel S., Chengzhen Dai, Farhad Hormozdiari, Bryce van de Geijn, Steven Gazal, Yongjin Park, Luke O’Connor, et al. 2019. “Genes with High Network Connectivity Are Enriched for Disease Heritability.” *American Journal of Human Genetics* 105 (6): 1302.

Kodama, Y., M. Shumway, R. Leinonen, and on behalf of the International Nucleotide Sequence Database Collaboration. 2012. “The Sequence Read Archive: Explosive Growth of Sequencing Data.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkr854.

Kuleshov, Maxim V., Matthew R. Jones, Andrew D. Rouillard, Nicolas F. Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, et al. 2016. “Enrichr: A Comprehensive Gene Set Enrichment Analysis Web Server 2016 Update.” *Nucleic Acids Research* 44 (W1): W90–97.

Langfelder, Peter, and Steve Horvath. 2008. “WGCNA: An R Package for Weighted Correlation Network Analysis.” *BMC Bioinformatics* 9 (1): 1–13.

Leek, Jeffrey T., W. Evan Johnson, Hilary S. Parker, Andrew E. Jaffe, and John D. Storey. 2012. “The Sva Package for Removing Batch Effects and Other Unwanted Variation in High-Throughput Experiments.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bts034.

Leskovec, Jure, and Rok Sosič. 2016. "SNAP: A General Purpose Network Analysis and Graph Mining Library." *ACM Transactions on Intelligent Systems and Technology* 8 (1). https://doi.org/10.1145/2898361.

Liska, Orsolya, Balázs Bohár, András Hidas, Tamás Korcsmáros, Balázs Papp, Dávid Fazekas, and Eszter Ari. 2022. "TFLink: An Integrated Gateway to Access Transcription Factor–target Gene Interactions for Multiple Species." *Database: The Journal of Biological Databases and Curation* 2022 (September):baac083.

Luijk, René, Koen F. Dekkers, Maarten van Iterson, Wibowo Arindrarto, Annique Claringbould, Paul Hop, Dorret I. Boomsma, et al. 2018. "Genome-Wide Identification of Directed Gene Networks Using Large-Scale Population Genomics Data." *Nature Communications* 9 (1): 3097.

Margolin, Adam A., Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. 2006. "ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context." *BMC Bioinformatics* 7 (1): 1–15.

Marnetto, Davide, Federica Mantica, Ivan Molineris, Elena Grassi, Igor Pesando, and Paolo Provero. 2018. "Evolutionary Rewiring of Human Regulatory Networks by Waves of Genome Expansion." *American Journal of Human Genetics* 102 (2): 207–18.

Mostafavi, Hakhamanesh, Jeffrey P. Spence, Sahin Naqvi, and Jonathan K. Pritchard. 2023. "Systematic Differences in Discovery of Genetic Effects on Gene Expression and Complex Traits." *Nature Genetics* 55 (11): 1866–75.

Narang, Vipin, Muhamad Azfar Ramli, Amit Singhal, Pavanish Kumar, Gennaro de Libero, Michael Poidinger, and Christopher Monterola. 2015. "Automated Identification of Core Regulatory Genes in Human Gene Regulatory Networks." *PLoS Computational Biology* 11 (9): e1004504.

Oh, Eun-Yeong, Stephen M. Christensen, Sindhu Ghanta, Jong Cheol Jeong, Octavian Bucur, Benjamin Glass, Laleh Montaser-Kouhsari, et al. 2015. "Extensive Rewiring of Epithelial-Stromal Co-Expression Networks in Breast Cancer." *Genome Biology* 16 (1): 128.

Oltvai, Zoltán N., and Albert-László Barabási. 2002. "Systems Biology. Life's Complexity Pyramid." *Science* 298 (5594): 763–64.

Omranian, Nooshin, Jeanne M. O. Eloundou-Mbebi, Bernd Mueller-Roeber, and Zoran Nikoloski. 2016. "Gene Regulatory Network Inference Using Fused LASSO on Multiple Data Sets." *Scientific Reports* 6 (February):20533.

Opgen-Rhein, Rainer, and Korbinian Strimmer. 2007. "From Correlation to Causation Networks: A Simple Approximate Learning Algorithm and Its Application to High-Dimensional Plant Gene Expression Data." *BMC Systems Biology* 1 (August):37.

Ota, Mineto, Jeffrey P. Spence, Tony Zeng, Emma Dann, Alexander Marson, and Jonathan K. Pritchard. 2025. "Causal Modeling of Gene Effects from Regulators to Programs to Traits: Integration of Genetic Associations and Perturb-Seq." *bioRxiv*. https://doi.org/10.1101/2025.01.22.634424.

Parsana, Princy, Claire Ruberman, Andrew E. Jaffe, Michael C. Schatz, Alexis Battle, and Jeffrey T. Leek. 2019. "Addressing Confounding Artifacts in Reconstruction of Gene Co-Expression Networks." *Genome Biology* 20 (1): 94.

Pastor-Satorras, Romualdo, Miguel Rubi, and Albert Diaz-Guilera. 2003. *Statistical Mechanics of Complex Networks*. Springer Science & Business Media.

Pierson, Emma, GTEx Consortium, Daphne Koller, Alexis Battle, Sara Mostafavi, Kristin G. Ardlie, Gad Getz, et al. 2015. "Sharing and Specificity of Co-Expression Networks across 35 Human Tissues." *PLoS Computational Biology* 11 (5): e1004220.

Presson, Angela P., Eric M. Sobel, Jeanette C. Papp, Charlyn J. Suarez, Toni Whistler, Mangalathu S. Rajeevan, Suzanne D. Vernon, and Steve Horvath. 2008. "Integrated Weighted Gene Co-Expression Network Analysis with an Application to Chronic Fatigue Syndrome." *BMC Systems Biology* 2 (November):95.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing* (version 4.0.2).

Rodius, Sophie, Anna Fournier, Lou Götz, Robin Liechti, Isaac Crespo, Susanne Merz, Petr V. Nazarov, et al. 2016. "Analysis of the Dynamic Co-Expression Network of Heart Regeneration in the

Zebrafish." *Scientific Reports* 6 (May):26822.
Saha, Ashis, and Alexis Battle. 2018. "False Positives in Trans-eQTL and Co-Expression Analyses Arising from RNA-Sequencing Alignment Errors." *F1000Research* 7 (November):1860.
Schaefer, Carl F., Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H. Buetow. 2009. "PID: The Pathway Interaction Database." *Nucleic Acids Research* 37 (Database issue): D674–79.
Schlitt, Thomas, Kimmo Palin, Johan Rung, Sabine Dietmann, Michael Lappe, Esko Ukkonen, and Alvis Brazma. 2003. "From Gene Networks to Gene Function." *Genome Research* 13 (12): 2568–76.
Segal, E., M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. 2003. "Module Networks: Identifying Regulatory Modules and Their Condition-Specific Regulators from Gene Expression Data." *Nature Genetics* 34 (2). https://doi.org/10.1038/ng1165.
Stegle, Oliver, Leopold Parts, Matias Piipari, John Winn, and Richard Durbin. 2012. "Using Probabilistic Estimation of Expression Residuals (PEER) to Obtain Increased Power and Interpretability of Gene Expression Analyses." *Nature Protocols* 7 (3): 500–507.
Stuart, Joshua M., Eran Segal, Daphne Koller, and Stuart K. Kim. 2003. "A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules." *Science*, October. https://doi.org/10.1126/science.1087447.
The GTEx Consortium. 2020. "The GTEx Consortium Atlas of Genetic Regulatory Effects across Human Tissues." *Science*. https://doi.org/10.1126/science.aaz1776.
Tomczak, Katarzyna, Patrycja Czerwińska, and Maciej Wiznerowicz. 2015. "The Cancer Genome Atlas (TCGA): An Immeasurable Source of Knowledge." *Contemporary Oncology* 19 (1A): A68–77.
Villar, Diego, Camille Berthelot, Sarah Aldridge, Tim F. Rayner, Margus Lukk, Miguel Pignatelli, Thomas J. Park, et al. 2015. "Enhancer Evolution across 20 Mammalian Species." *Cell* 160 (3): 554–66.
Wang, Xuran, David Choi, and Kathryn Roeder. 2021. "Constructing Local Cell-Specific Networks from Single-Cell Data." *Proceedings of the National Academy of Sciences of the United States of America* 118 (51). https://doi.org/10.1073/pnas.2113178118.
Wilks, Christopher, Shijie C. Zheng, Feng Yong Chen, Rone Charles, Brad Solomon, Jonathan P. Ling, Eddie Luidy Imada, et al. 2021. "recount3: Summaries and Queries for Large-Scale RNA-Seq Expression and Splicing." *Genome Biology* 22 (1): 323.
Wolf, Denise M., Marc E. Lenburg, Christina Yau, Aaron Boudreau, and Laura J. van 't Veer. 2014. "Gene Co-Expression Modules as Clinically Relevant Hallmarks of Breast Cancer Diversity." *PloS One* 9 (2). https://doi.org/10.1371/journal.pone.0088309.
Zhou, Ke-Ren, Shun Liu, Wen-Ju Sun, Ling-Ling Zheng, Hui Zhou, Jian-Hua Yang, and Liang-Hu Qu. 2016. "ChIPBase v2.0: Decoding Transcriptional Regulatory Networks of Non-Coding RNAs and Protein-Coding Genes from ChIP-Seq Data." *Nucleic Acids Research* 45 (D1): D43–50.
Zhu, Xiang, Zhana Duren, and Wing Hung Wong. 2021. "Modeling Regulatory Network Topology Improves Genome-Wide Analyses of Complex Human Traits." *Nature Communications* 12 (1): 1–15.

Figure Legends

Figure 1: Overview of data pre-processing and annotations (A) Gene expression data was RPKM normalized and log-transformed along with gene-specific and sample-specific filters. Based on the data source, normalized gene expression was processed to merge replicates and exclude miRNA and scRNA seq samples. **(B)** Number of samples which were annotated to be non-cancer and cancer based on available metadata across GTEx, SRA, and TCGA. 63 SRA samples did not have an associated annotation corresponding to cancer status. **(C)** Top 10 tissue labels by sample size across all three data sources: SRA, GTEx, and TCGA. **(D)** Top 20 diseases

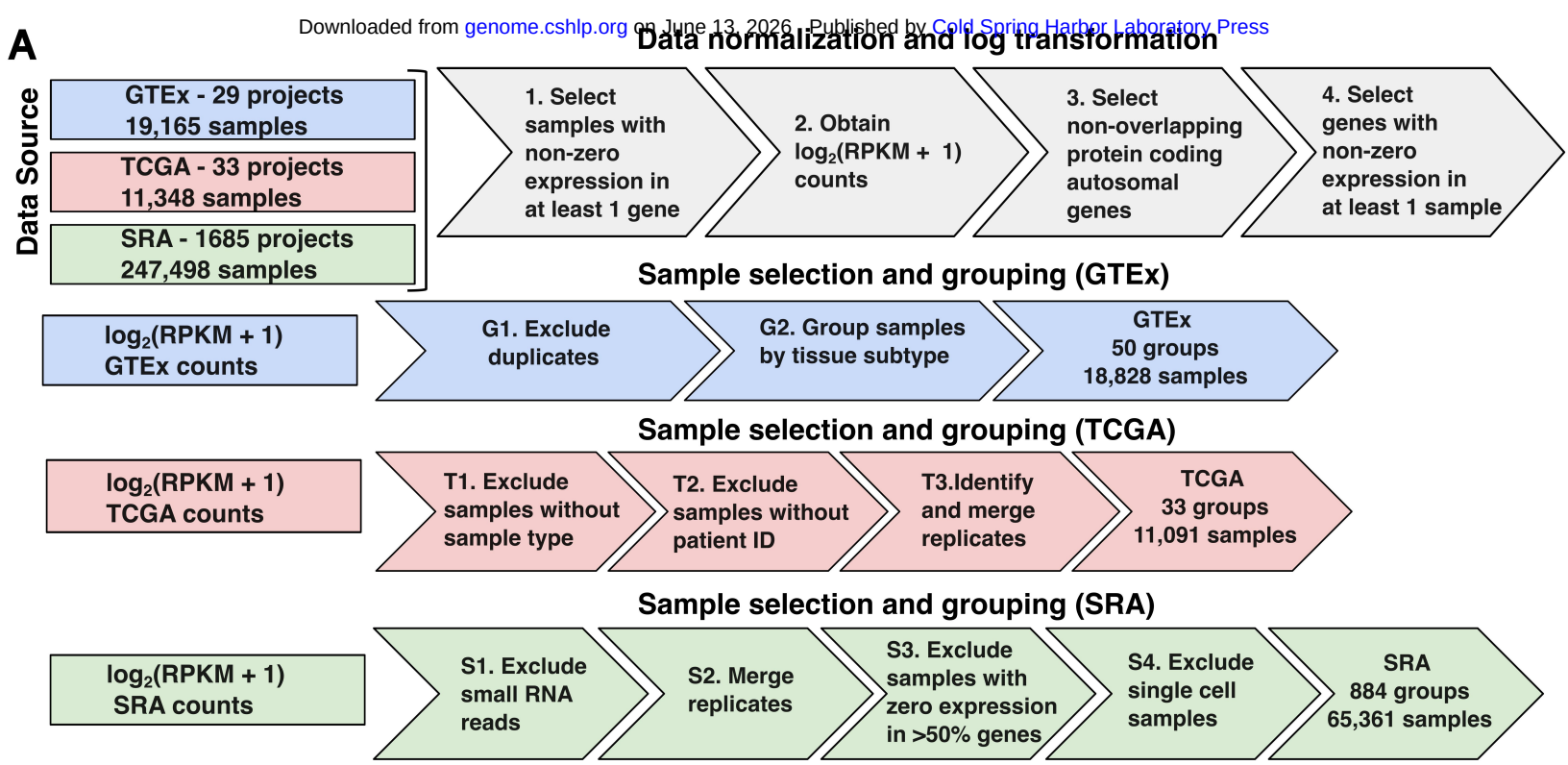
by sample size found in SRA that are not cancer. **(E)** t-SNE projection of xCell deconvolution scores of 63,193 non-cancerous samples colored by the tissue of origin. **(F)** Increase in the sample size of 27 tissue contexts by using SRA samples compared to GTEx only. SRA studies included 7 novel contexts which were not available in GTEx.

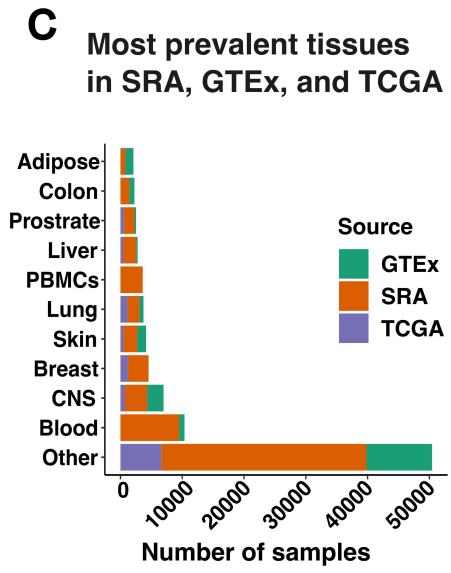
Figure 2: Comparison of aggregation strategies to optimize network reconstruction (A) Outline of strategies to compare data correction before and after aggregation and weighted and unweighted aggregation of single tissue/ study covariance matrices included (1) Aggregating data before PC-based data correction followed by estimation of empirical covariance from residual expression (Aggregating data, orange), (2) PC-based data correction applied to individual studies followed by aggregation of residual expression and joint estimation of empirical covariance (Aggregating data adjusted for confounding, brick red), (3) Unweighted aggregation of covariance matrices inferred from each study separately after study-specific PC-based correction (Unweighted covariance aggregation, purple) and (4) Weighted aggregation of covariance matrices computed from individual studies following study-specific PC-based data correction (Weighted covariance aggregation, magenta). **(B)** Held-out log-likelihood of networks inferred by sequentially aggregating either 10 GTEx studies or 100 SRA studies at a time vs. the log of the number of edges ($|E|$) found in networks obtained by varying the penalization parameter λ **(C)** F1-score of networks inferred by sequentially aggregating either 10 GTEx studies or 100 SRA studies at a time vs. the log of the number of edges ($|E|$) found in networks obtained by varying the penalization parameter λ when compared to canonical pathways compiled from KEGG, Biocarta, and Pathway Interaction Database. **(D)** Comparison of held-out log-likelihood corresponding to networks inferred over 49 GTEx studies or 566 SRA studies vs. the log of the number of edges ($|E|$) found in networks obtained by varying the penalization parameter λ using four different aggregation strategies including aggregating data, aggregating data adjusted for confounding, unweighted, and weighted aggregation of covariance matrices. **(E)** Comparison of F1-scores of obtaining edges corresponding to canonical pathways from KEGG, Biocarta, and Pathway Interaction Database in networks inferred over 49 GTEx studies or 566 SRA studies vs. the log of the number of edges ($|E|$) found in networks obtained by varying the penalization parameter λ using four different aggregation strategies including aggregating data, aggregating data adjusted for confounding, unweighted, and weighted aggregation of covariance matrices. **(F)** Total number of samples, number of individual studies, and the median sample size of each study which were used in the inference of universal consensus, non-cancer consensus, and cancer consensus networks. **(G)** Comparison of F1-scores of obtaining edges corresponding to canonical pathways in the three consensus networks, universal, non-cancer, and cancer, across networks with number of edges ($|E|$) varying between 5×10^3 to 5×10^6 edges. **(H)** Log-likelihood of GTEx blood samples based on networks inferred by sequentially aggregating SRA blood studies five at a time for number of edges ranging from 10^3 to 10^5 edges. **(I)** Log-likelihood of GTEx CNS samples based on networks inferred by sequentially aggregating SRA CNS studies five at a time for number of edges ranging from 10^3 to 10^5 edges. **(J)** Odds ratio of finding edges corresponding to tissue-specific protein-protein interactions (PPIs) derived from SNAP in tissue-context-specific networks inferred using all available samples vs. only samples found in GTEx for six tissue contexts.

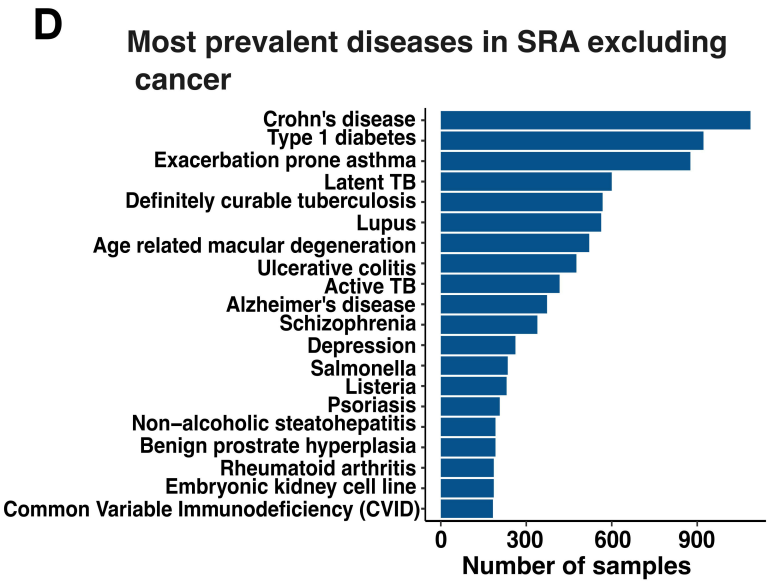
Figure 3: Properties of central network nodes of consensus and context-specific networks (A-C) Enrichment of genes involved in GO processes among network genes selected with

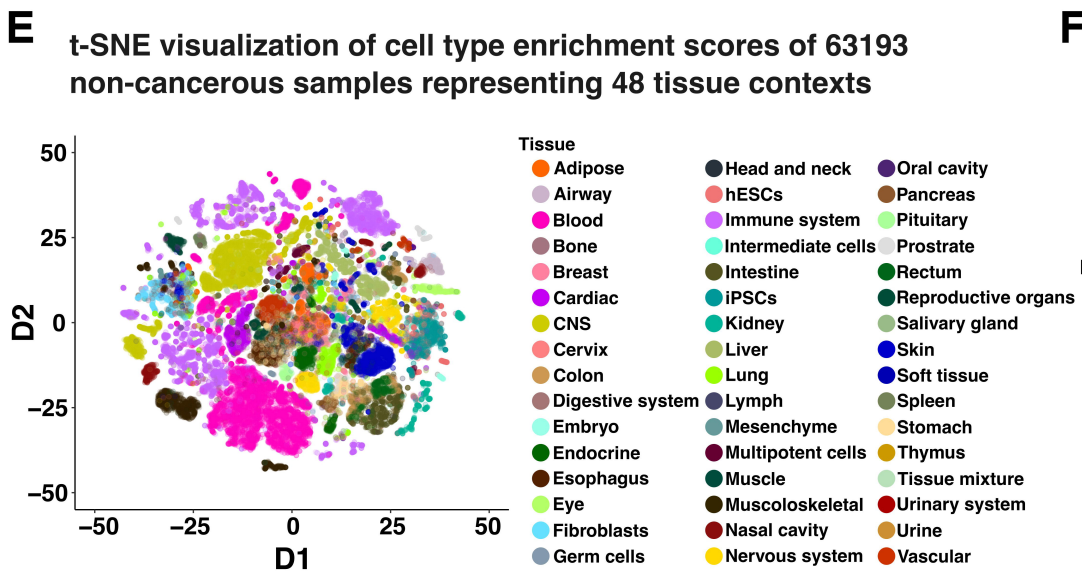
increasing thresholds of node degree in three consensus networks, universal (λ = 0.18, 7087 edges), non-cancer (λ = 0.18, 7355 edges), and cancer (λ = 0.24, 7552 edges), as well as two context-specific networks blood (λ = 0.24, 7283 edges) and CNS (λ = 0.28, 8430 edges). Tissue-context-specific networks were inferred only using non-cancerous samples. Blood (GTEx) or CNS (GTEx) networks were inferred using samples only found in GTEx while Blood and CNS networks were inferred using samples from GTEx and SRA. **(D)** Distribution of the excess overlap of evolutionarily conserved gene sets (Methods) for network nodes binned by the number of neighbors (degree) corresponding to universal consensus networks (λ = 0.14, 0.16, 0.18, 0.20), non-cancer consensus network (λ = 0.14, 0.16, 0.18, 0.20), cancer consensus networks (λ = 0.20, 0.22, 0.24, 0.26), blood network (λ = 0.18, 0.20, 0.22, 0.24, 0.26), and CNS network (λ = 0.24, 0.26, 0.28, 0.30, 0.32). Quintile 1 reflects nodes with no neighbors. Nodes with non-zero neighbors are split based on the degree quartile they belong to (Quintiles 2-5). We evaluated the excess overlap of 3,104 loss-of-function (LoF) genes with pLI > 0.9,2,853 genes with a S_{het} > 0.1, 588 genes with a Phi-score > 0.95, and 1,440 genes strongly depleted for missense mutations (high missense z-score). **(E)** The degree distribution of network nodes that are tissue-specific transcription factors (TFs) in blood (52 TFs), lung (58 TFs), skin (10 TFs), pancreas (16 TFs), cardiac (17 TFs), muscle (7 TFs), CNS (51 TFs), general transcription factors (88 TFs), and protein-coding genes which are not transcription factors in universal consensus (λ = 0.18, 7,087 edges), non-cancer consensus (λ = 0.18, 7,355 edges), skin (λ = 0.26, 7,567 edges), skeletal muscle (λ = 0.26, 6,254 edges), pancreas (λ = 0.32, 7,615 edges), lung (λ = 0.30, 6,349 edges), CNS (λ = 0.30, 6,316 edges), cardiac (λ = 0.30, 6,481 edges), blood (λ = 0.24, 7,283 edges). Pairs with no significance reported were not statistically distinct (p > 0.1). **(F)** Factor weights were obtained by non-negative matrix factorization of the presence of hub genes in tissue-specific networks with ~ 7,000 edges. Details of the penalization parameter λ and the number of edges of selected networks for each tissue context are provided in **Supplemental Table S7**.

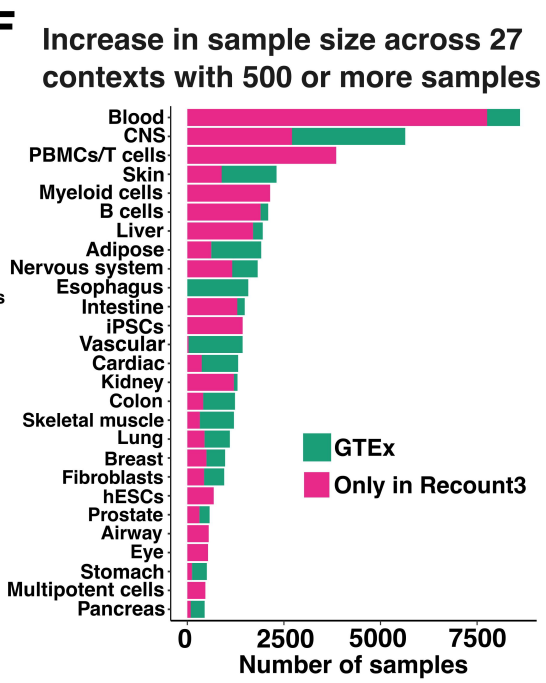
Figure 4: Heritability enrichment of network annotations Mean and standard deviation of heritability enrichment and the coefficient τ^*, an estimate of the heritability of SNPs unique to the annotation. All genes: whether a variant was located in a 100 kilobase window of all protein-coding genes (Translucent), All genes + baseline: all-genes annotation in addition to 97 functional annotations such as known enhancer and promoter regions (Opaque). **(A)** Meta-analysis of 42 independent traits for six centrality measures obtained from the universal consensus network and non-cancer consensus networks corresponding to values of the penalization parameter λ between 0.14 - 0.20. **(B)** Meta-analysis of 9 blood-related traits including, Crohn's disease, Ulcerative colitis, Rheumatoid arthritis, Allergy Eczema, Eosinophil count, Red blood cell count, White blood cell count, Red blood cell width, and Platelet count for network annotations from Blood GTEx (λ = 0.24 - 0.32), Blood consensus (λ = 0.18 - 0.26), and Universal consensus network (λ = 0.14 - 0.20). **(C)** Meta-analysis of 9 CNS-related traits including, Alzheimer's disease, Epilepsy, Parkinson's disease, Bipolar disorder, Smoking cessation, Waist-hip-ratio adjusted BMI, Schizophrenia, Major depressive disorder, and Number of alcoholic drinks per week, for network annotations corresponding to 6 centrality measures derived from CNS GTEx (λ = 0.26 - 0.32), CNS (λ = 0.20 - 0.28), and Universal consensus network (λ = 0.14 - 0.20).

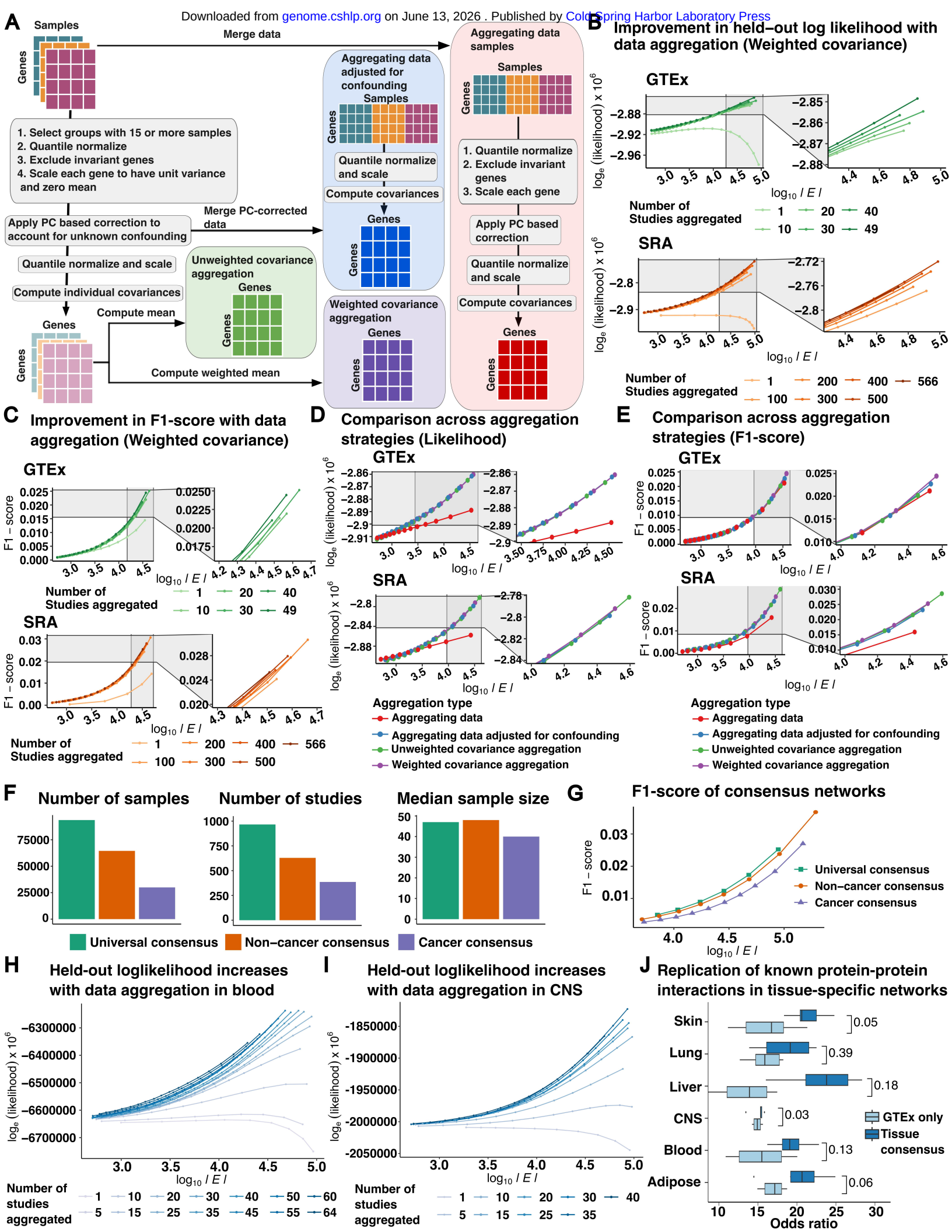


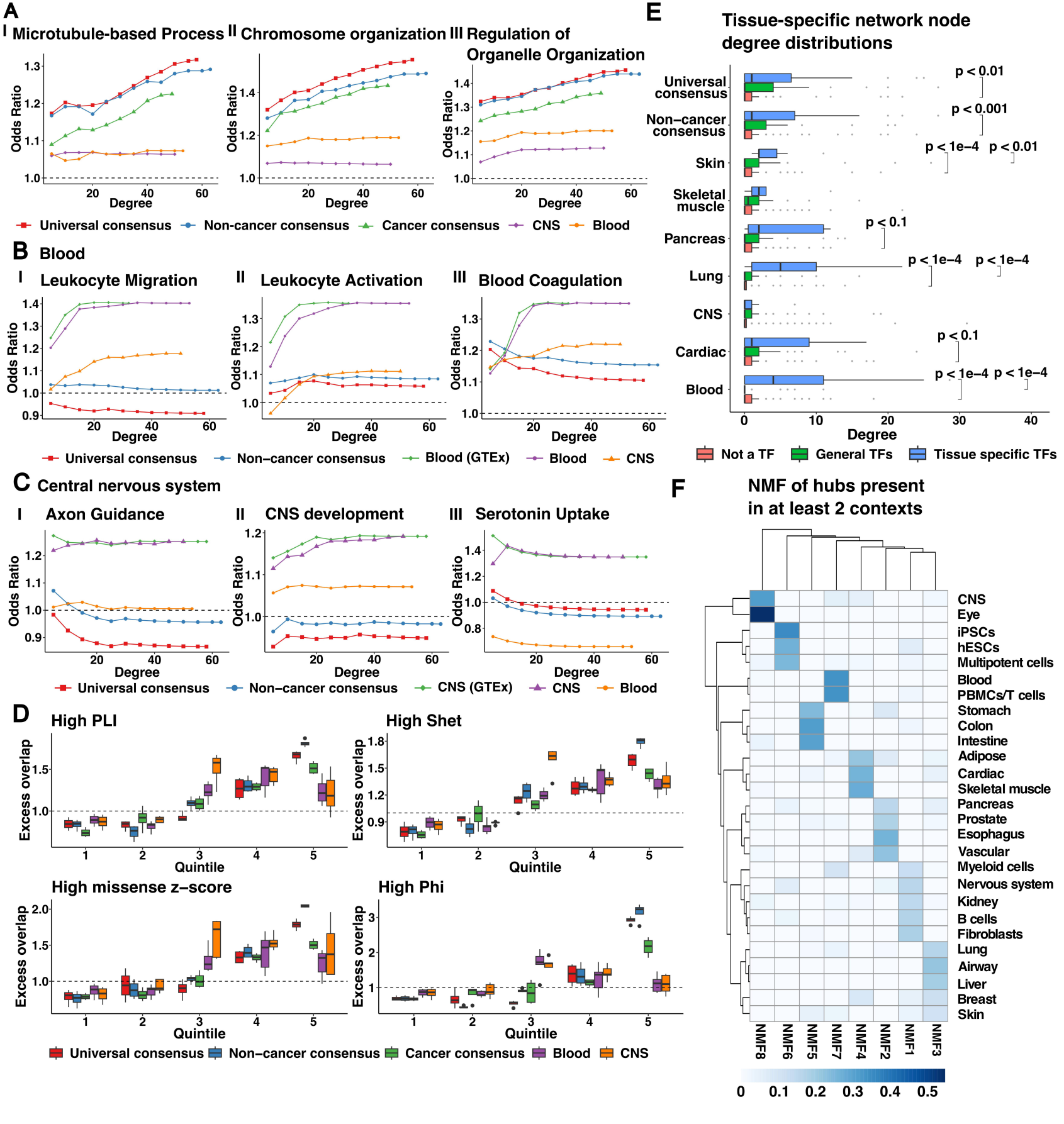


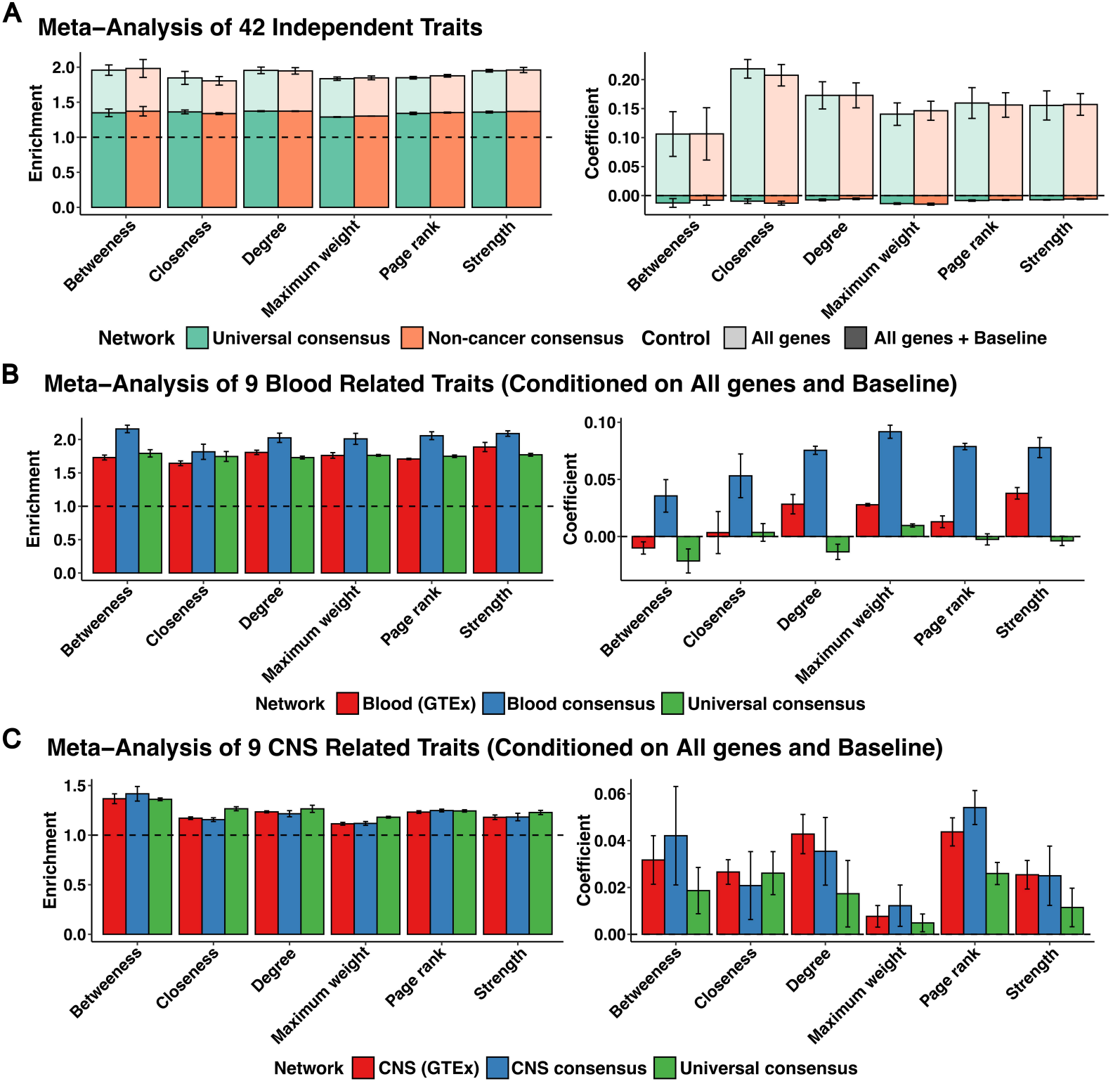









A
I Microtubule-based Process
II Chromosome organization
III Regulation of Organelle Organization
Odds Ratio
Degree
Universal consensus
Non-cancer consensus
Cancer consensus
CNS
Blood
B Blood
I Leukocyte Migration
II Leukocyte Activation
III Blood Coagulation
Blood (GTEx)
C Central nervous system
I Axon Guidance
II CNS development
III Serotonin Uptake
CNS (GTEx)
D
High PLI
High Shet
High missense z-score
High Phi
Excess overlap
Quintile
E
Tissue-specific network node degree distributions
Universal consensus
Non-cancer consensus
Skin
Skeletal muscle
Pancreas
Lung
CNS
Cardiac
Blood
p < 0.01
p < 0.001
p < 1e-4
p < 0.1
Not a TF
General TFs
Tissue specific TFs
F
NMF of hubs present in at least 2 contexts
CNS
Eye
iPSCs
hESCs
Multipotent cells
Blood
PBMCs/T cells
Stomach
Colon
Intestine
Adipose
Cardiac
Skeletal muscle
Pancreas
Prostate
Esophagus
Vascular
Myeloid cells
Nervous system
Kidney
B cells
Fibroblasts
Lung
Airway
Liver
Breast
Skin
NMF8
NMF6
NMF5
NMF7
NMF4
NMF2
NMF1
NMF3
0
0.1
0.2
0.3
0.4
0.5


A
Meta−Analysis of 42 Independent Traits
Enrichment
Coefficient
Betweeness
Closeness
Degree
Maximum weight
Page rank
Strength
Network
Universal consensus
Non-cancer consensus
Control
All genes
All genes + Baseline
B
Meta−Analysis of 9 Blood Related Traits (Conditioned on All genes and Baseline)
Network
Blood (GTEx)
Blood consensus
Universal consensus
C
Meta−Analysis of 9 CNS Related Traits (Conditioned on All genes and Baseline)
Network
CNS (GTEx)
CNS consensus
Universal consensus