Discovery of gene-gene co-perturbation through big data

Jiasheng Wang (1, 2, 3), Ying-Wooi Wan (1, 4), Rami Al-Ouran (5), Zhandong Liu (1, 2, 3)

1. Jan and Dan Duncan Neurological Research Institute at Texas Children's Hospital, Houston, TX 77030, USA.
2. Graduate Program in Quantitative and Computational Biosciences, Baylor College of Medicine, Houston, TX, 77030, USA
3. Department of Pediatrics, Baylor College of Medicine, Houston, TX 77030, USA.
4. Department of Molecular and Human Genetics, Baylor College of Medicine, Howard Hughes Medical Institute, Houston, TX 77030, USA.
5. Al Hussein Technical University, Amman, Jordan

Keywords: gene network; co-perturbation network; non-linear correlation; Simpson's paradox

Abstract:

Scientists have learned much about gene expression in the past few years. Integrating large-scale gene-expression data will significantly bolster researchers' understanding of biology and diseases. While we currently use co-expression models to discover gene-gene associations in RNA seq data, integrating RNA seq data generated from different experiments can have a "batch effect," which decreases data quality and makes it harder to glean definitive gene relationships from co-expression models. The biological relationship can also be context-dependent and non-linear, which make them difficult to detect using current co-expression approaches. Here, we propose a co-perturbation model that better identifies gene-gene associations in integrated data and reveals non-linear correlation between genes.

Introduction:

Discovering gene-gene relationships is crucial to understanding cell biology and can help to identify both potential gene-disease associations and diseases' underlying biological mechanisms[1][2]. Researchers have therefore used several approaches to generate different biological networks and pinpoint gene-gene relationships. For instance, the protein-protein interaction networks (PPI) elucidate the physical contact and functional association between two proteins[3][4]. Gene regulatory networks are composed of many genes that regulate a common function by interacting with each other[5][6]. Gene co-expression networks highlight the dependency of gene pairs with similar expression patterns under the same experimental conditions[7].

With the development of large biological data portals like Expression Atlas and Recount3, gene expression data from thousands of different studies can now be uniformly processed in high quality[8][9]. However, no reliable statistical model exists to analyze the integrated expression data from different experiments for two reasons. First, the sample size of each study is limited. Although scientists typically use a gene co-expression model to investigate gene correlation in expression data, this requires numerous samples and similar experimental conditions. In Recount3 data, 6,377 out of 8,677 SRA human studies used less than 20 samples. In other words, if a study has two groups, each group typically has ≤ 10 samples—if it contains multiple groups, this number is even smaller. Such small sample sizes

make it difficult to estimate gene correlations in a single study[10]. Second, all of these studies are conducted by different labs with different goals, so simply integrating samples from different experiments will have the batch effect[11]. To address this challenge, scientists currently choose either to combine samples from fewer studies or to apply tools for batch correction. Some research that combined samples across different experiments to form compendia and then analyze these compendia to reveal new biology, for example, either combined 20-100 small studies or had trouble with batch correction[12][13][14][15]. Although tools like limma and ComBat-seq help with batch correction, they rely entirely on prior information[16][17]. When combining only a few experiments, one can try to correct for experiment-specific noise, but this becomes more difficult when many experiments are combined[18]. For many biological experiments, samples in one group are usually experimental replicates and they have very similar gene expression, so when we calculate the gene-gene correlations between two small groups, it's equivalent to do the regression on only two points, as samples in the same group cluster together and may overlap with each other, which will always make the regressions significant. Finally, the correlations between different contrasts are very different, so it is difficult to combine all of them. The co-expression model is therefore insufficient to identify gene-gene associations [Sup animation1].

Here, we propose a new co-perturbation model. This model generates a gene co-perturbation network in which genes co-perturbed under many heterogeneous experimental conditions are connected. Our co-perturbation model is not vulnerable to the batch effect, can better implement integrated gene expression data, and can help to reveal non-linear gene relationships.

In this study, we applied our co-perturbation model to a huge data set (Recount3) and two small data sets (Recount2 Brain[19] and Mouse Alzheimer's Disease data) (https://doi.org/10.7303/syn2580853). We compared our co-perturbation model's performance with the well-known co-expression model WGCNA[20]. The results showed that the gene-gene associations our co-perturbation model identified are more reliable than WGCNA, because combining expression data from multiple experiments results in batch effect which ruins correlations. We also found that many of the gene-gene associations that WGCNA identified in pooled data were incorrect and that, in some of the gene-gene associations that did actually exist, the Simpson's Paradox obfuscated their correlation trends[21].

After combining the experiments together, we found that many gene-pair expressions are context-dependent, so they concordantly regulated in some experiments and discordantly regulated in others. In these cases, co-expression models cannot identify non-linear correlation patterns due to heterogeneity. Although models like MMC can identify co-expression patterns even in heterogeneous expression data, they over-correct true biological signals by interpreting hub genes' biological effect as a global confounder and adjusting it[22]. Our co-perturbation model is not influenced by batch effect at all and thus can still identify those non-linearly correlated gene pairs. The Recount 3 Co-perturbation network can be accessed by http://coperturbationnetwork.liuzlab.org/

Results:

Co-perturbation workflow:

In order to identify gene-gene associations in multiple studies with relatively small sample sizes, we proposed a new approach called a co-perturbation model (Figure1). If a gene is differentially expressed between two sample groups (a contrast), we call it a DEG (differentially expressed gene). A contrast can result from experimental conditions, sample genotype, phenotype, or some unknown effect detected by

the DASC[23]. Every two DEGs identified in a contrast constitute a DEG pair co-perturbed under corresponding conditions. Our co-perturbation model identifies the DEG pairs that are significantly co-perturbed across multiple experiments.

The first step needed to generate the co-perturbation network is identifying DEGs in each experiment (Figure 1A). Second, these DEGs are tabulated in a binary DEG matrix in which rows represent the genes and columns represent the experiments (Figure 1B). Any two DEGs in the same experiment are treated as a DEG pair, and the number of experiments in which the pair are co-perturbed are recorded in the gene-by-gene co-perturbation matrix (Figure 1 C). A baseline model is necessary to identify which DEG pairs are significantly co-perturbed. Based on the true binary DEG matrix, we simulated 1,000 random binary DEG matrices, each of which had the same sum of rows and sum of columns to accurately mimic the DEG probabilities (Figure 1D). In other words, our null hypothesis H_0 is that (1) every single gene has the same chance of being a DEG as it does in the real data, and (2) each experiment provides the same number of DEGs as would actual data. Third, for each simulated binary DEG matrix, we calculate the simulation-based co-perturbation matrix (Figure 1E). Each gene pair i has a co-perturbation vector $V_{randco,i}$ indicating the number of experiments the gene pair i co-perturbed in random cases. The DEG pairs are considered significant when their $Pr(V_{randco,i} \geq Co_i)$ score is <0.01, where Co_i is the number of experiments that gene pair i co-perturbed in the real data (Figure 1F). However, 1000 simulations means that the p-value precision is 0.001, which may not be precise enough to accurately compare two gene pairs that are both significantly co-perturbed; p-values for those gene pairs may be 0. We therefore need to fit the random variable with a known distribution. We can treat co-perturbation in each experiment as a Bernoulli trial, where the co-perturbation vector $V_{randco,i}$ describes the probability of success in a sequence of Bernoulli trials. Since each experiment has a different co-perturbation probability for gene pair i, the $V_{randco,i}$ should follow a beta-binomial distribution. We therefore used beta-binomial fitting on each $V_{randco,i}$ to obtain a distribution Bb_i and get more precise p-values for each gene pair i, $p = \Pr(Bb_i \geq Co_i | H_{0,i})$ (Figure 1F, blue curve). Because there's a large number of gene pairs in the integrated data, and each gene pair is a null hypothesis testing, we next calculated the FDR to minimize the risk of type I errors in our final co-perturbation network. After setting an FDR threshold, we will connect significantly co-perturbed genes to construct the final co-perturbation network (Figure1 G, H).

We also examined the runtime generating the random binary matrix based on the null hypothesis H_0. Given the sparsity of the DEG matrix, the runtime is completely linearly aligned with the squared DEG matrix size, where the matrix size is defined by the number of genes multiplied by the number of experiments (Sup Fig1A). Given the matrix size, the runtime decreased with sparsity (Sup Fig1B). Thus, the runtime for even a huge binary matrix with 10,000 rows, 30,000 columns, and 0.99 sparsity is ≤ 1 minute.

One might think that the more experiments in which a DEG pair is co-perturbed, the more likely it is to be significant. However, gene pairs that were co-perturbed more often aren't always more significant than those that were co-perturbed less often. In *Sup Fig2A*, we show that a gene pair co-perturbed in 7 contrasts had at least a 37% chance of being co-perturbed in ≥7 contrasts in the baseline model, but the gene pair co-perturbed in 6 contrasts had only a 0.4% chance of reach that number of co-perturbations in 1,000 simulations. Therefore, some genes with a very low DEG probability that co-perturbed in a relatively small number of experiments can still be very significant. This is why we chose to calculate a

local co-perturbation probability for each gene pair rather than develop an overall co-perturbation cutoff for all gene pairs. Still, the assumption that the more experiments in which a DEG pair is co-perturbed, the more likely it is to be significant is generally correct, as we did find that most significantly co-perturbed genes have a high co-perturbation frequency [Sup Fig2B].

Co-perturbation model reveals non-linearly correlated genes:

We first applied our co-perturbation model to Recount3 data, which contains more than 8,000 human studies. For each study, we extracted label information to classify the samples into different groups. One study may contain n groups, which provides C_n^2 contrasts (inter-group experiments). Overall, Recount3 provides ~12,000 contrasts and 27,000 genes to generate the co-perturbation network.

The co-perturbation model identified 6,000,000 significantly co-perturbed gene pairs among more than 360,000,000 gene pairs. We combined all of the samples and calculated the significantly co-perturbed gene pairs' Pearson's correlation score in the integrated data and non-linear score among co-perturbed contrasts [Method]. Our main takeaway was that the non-linear score is negatively related to the Pearson's correlation score for co-perturbed gene pairs (Figure2 A). Compared with the gene pairs identified by WGCNA, which has a higher Pearson's correlation score (Figure2 B), our co-perturbation model identified gene pairs in both high and low correlation score regimes. The reason these co-perturbed gene pairs have a low Pearson's correlation is because they are non-linearly correlated among different contrasts (Figure2 A), which explains why a co-expression model like WGCNA cannot identify such gene-gene association. The non-linear correlation means that two genes are concordantly expressed in some contrasts and discordantly expressed in others. Figure 2C contains a schematic diagram in which samples depicted in red and orange are positively correlated while samples depicted in blue and dark blue are negatively correlated. Although we can discern their correlation separately, the overall correlation is very weak after combining all the samples, so pooling these data together would only ruin the co-expression pattern. We also show two real examples in which the gene pairs were significantly co-perturbed in many contrasts but were not always concordantly expressed together (Figure2 D, E). According to the heatmap, these two gene pairs are concordantly co-expressed in half of the contrasts and discordantly co-expressed in the other half. However, in the pooled data, we can barely discern a linear correlation between two genes and the Pearson's correlation score is close to 0 (Figure2 F, G). The Pearson's correlation between FLT1 and LGALS1 is 0.0843, and the correlation between PREX1 and LIMCH1 is 0.0156. The co-expression model cannot identify either of the gene pairs although they are actually significantly co-perturbed in 167 (FDR = 3.05×10^{-12}) and 190 (FDR = 8.54×10^{-20}) contrasts, respectively. We separately plotted the samples that are positively co-expressed (Figure 2H and I) and negatively co-expressed (Figure 2J and K). Although the color we used to distinguish samples from different contrasts may be inadequate due to the large number of contrasts, we can still see the positive trend in Figure 2H and I and the negative trend in Figure 2J and K. Due to the context-dependent heterogeneity, gene pair correlations can be non-linear and hard to capture in the pooled data with traditional co-expression methodology. But our model successfully identified these gene-gene associations using co-perturbed contrasts rather than co-expression patterns.

Simpson's paradox in co-expression model:

Simpson's paradox states that combining different groups of data together results in an overall correlation that is the opposite of each group's true individual correlation[24]. When the co-perturbation

model showed that two genes were negatively correlated in every contrast, the co-expression model showed that the overall correlation was positive (Figure3 A). We used two examples to demonstrate that these two gene pairs are negatively correlated according to the log2FC heatmap in Figure 3B and C, but the overall correlation between TUBA1B and ZKSCAN1 and between ATAD3A and IGBP1 is ~0.62. For both gene pairs, the Pearson's correlation is positive, but if we look at the contrasts independently, they are both negatively correlated in all contrasts (Figure3 D, E). Therefore, while the co-expression model may identify a significant connection between these genes, the genes' relationship can be completely reversed due to the Simpson's paradox. One explanation for this is that the batch effect between different studies causes data bias, so that when integrating a large number of studies without batch correction, the correlation detected becomes the biased batch trend and not the gene signals. Overall, more than 40,000 gene pairs—90% of which are protein coding genes—are subject to the Simpson's paradox.

Co-perturbation Network on Recount3 Data:

For the purpose of analyzing and learning the biology behind the Recount3 co-perturbation network, we used the recursive Louvain module detection method[25] on the Recount3 co-perturbation network and identified 79 modules. Then, we generated a binary heatmap (Figure 4A) with rows of contrasts and columns of genes. The top color bar represents modules in the Recount3 co-perturbation network while the left color bar indicates contrasts from different experiments. The black color indicates that the gene in the corresponding contrast is a DEG. We observed that genes in each module are co-perturbed in a set of common contrasts, which come from various studies. This result means that genes co-perturbed together are closely connected to each other in the network. It also begs the question whether, if a group of genes were co-perturbed by the same set of contrasts, they share the same biological functions. To answer this, we further investigated each module's biological functions using the GO enrichment test and comparing these modules' common significant GO terms (Figure 4B). Multiple modules share some functions, like immune receptor activity and GTPase regulator activity, while other functions, like GDP binding and phosphatase activity, are unique to only one of the modules (Sup Figure 3A). So, we needed to confirm that these enriched GO terms derived from our co-perturbation network are biologically meaningful and that a random network could not have the same quantity of enriched GO terms. We therefore compared our co-perturbation network's GO term interaction density with a random network that has the same genes and interaction density. We found that the co-perturbation network has a considerably higher GO term interaction density than the random network (Figure 4C), which indicates that the gene-gene associations in the co-perturbation network are more likely to be biologically meaningful and not randomly assigned.

Validation of Recount3 Co-perturbated genes:

We used WGCNA to identify co-expressed genes and compare our co-perturbation model's performance with co-expression's performance on big data. We validated the results by comparing the protein-protein interaction between co-perturbed gene pairs and co-expressed gene pairs according to the STRING database[26]. We found that a higher percentage of gene pairs (Chi-Square p-value < 10^{-20}) in the co-perturbation network could be validated in the PPI database than could the co-expressed gene pairs (Figure 4D). PPI can detect about 15.6% of 3 million co-perturbed gene pairs and validate only 4.3% of 5 million co-expressed gene pairs. To ensure that the gene-gene associations we identified are generalizable—in other words, that the gene pairs found in Recount3 are also meaningful in another

database—we used the gene pairs identified in Recount3 as training results and compared their correlations in a testing data set, which is GTEx data[27]. We chose GTEx data because it provides a more homogenous background and ensures that all the samples are healthy human samples. We hypothesized that the gene-gene associations that Recount3 identified would be reproduced in the GTEx data. We found that in the Recount3 training data, the co-expressed gene pairs that WGCNA identified have a higher correlation score than co-perturbed gene pairs (Figure4 E). However, when we put them in GTEx testing data, we found that co-perturbed gene pairs tend to have a higher correlation score and that co-expressed gene pairs demonstrate a background correlation distribution in testing data (Figure 4F). These results suggest that our co-perturbation model does a better job finding the biologically meaningful gene pairs in the integrated data because other independent databases can more easily validate co-perturbed gene pairs.

Co-perturbation network on Recount2 Brain and MAD:

In addition to the Recount3 co-perturbation network, we also applied our co-perturbation model to two smaller data sets: (1) the recount brain project and (2) mouse Alzheimer's disease (MAD). The recount brain project consists of 78 qualified studies with 93 total contrasts. These studies are all human brain-tissue-specific and contain ~1,300 samples and 15,000 genes. The MAD dataset consists of 70 unique mouse studies and 138 contrasts and has ~1,000 samples and 10,000 genes.

In Recount2 brain data, our co-perturbation model identified 52,002 significantly co-perturbed gene pairs (4,555 genes). The Louvain module detection method found 9 modules and for each module, we can clearly see that genes are co-perturbed in a group of common contrasts (Figure 5A). Because these studies are all brain-region-specific, some gene modules shared many GO terms in common (Figure 5). These GO terms are related to nerves with diverse channel activities (Sup Figure 3B) such as ion channel activity. To further validate the biology behind the co-perturbation network, we compared the Rcount2 brain co-perturbation network with a random network with the same genes but different connections. The result showed that the co-perturbation network has a higher GO term interaction density than the random network (Figure 5C). We also visualized the network in Figure 5D, in which the color of the genes correspond to their module color.

Around 30% of the gene pairs recognized by co-perturbation model were found to have protein-protein interactions according to the STRING database while only 22% of the gene pairs detected by WGCNA were validated by PPI data (Chi-Square p-value < 10^{-20}) (Figure 5E). We also validated our gene pairs' RNA seq data by GTEx brain data (Figure 5F). Regarding background distribution, most gene pairs in the brain GTEx had a low correlation score, while gene pairs identified by WGCNA in Recount2 Brain had similar background distribution as the brain GTEx. In the co-perturbation model, however, a bimodal distribution showed that though many gene pairs still had a low correlation score, and that half of the co-perturbed gene pairs have a high correlation score in GTEx brain data. Gene pairs identified by the co-perturbation model in Recount2 Brain are generally better validated than those identified by WGCNA in both the GTEx data and STRING database.

In the second small case study, by using the co-perturbation model on the MAD dataset, we successfully identified 201,454 significantly co-perturbed gene pairs (7,199 genes). The Louvain module detection approach found 10 modules in the network, and genes co-perturbed in common contrasts were clustered together (Sup 4A). Because the MAD studies are more domain-specific—they are all related to neurodegenerative diseases in mice—the GO term enrichment results showed that some modules share

many common GO terms (Sup 4B). These enriched GO terms were also related to channel activities and channel bindings (Sup 3C). We performed the GO term interaction density test, which demonstrated that the MAD co-perturbation network still has a higher interaction density than the random network. This in turn suggests that there is agreement between the co-perturbation network and GO terms (Sup 4C). Finally, we tried to visualize the MAD co-perturbation network (Sup 4D).

Our PPI validation showed that co-expressed genes from WGCNA had a higher PPI rate than co-perturbed genes (Chi-Square p-value < 10^{-20}), but both gene groups' validation percentage was very close (Sup 4F). Also, we not only used GTEx data to perform RNA validation, but also borrowed disease oriented ROSMAP data[28]. We found that co-expressed genes from WGCNA could not be accurately validated in either the GTEx or ROSMAP data because the correlation distribution still follows the background distribution in all independent datasets. The co-perturbed genes' correlation distribution in MAD, however, had a higher average correlation score (Sup 4E). Thus, we still believe that co-perturbed gene pairs are better validated than co-expressed gene pairs.

Linear and non-linear subnetwork

To further validate the Recount3 co-perturbation network, we compared two subnetworks with the Recount3 co-expression and STRING PPI networks. One is a linear local module containing 11 genes, all of which are closely connected in both the co-perturbation network and WGCNA (Fig6 A). Although not all of these connections were validated by PPI, this local module is still biologically significant because we found that many of the genes—including KIF11, CENPE, ASPM and KNL1—are related to microcephaly. Also, we think WGCNA can identify these gene-gene associations only because this subnetwork is highly significant and the associations are all linearly correlated (Figure 6B), meaning that there is little noise interfering with the co-expression model's ability to identify these correlations even in the pooled data. In the second local module, however, the non-linear score between genes is much higher than in the first module, which created a more difficult and complicated gene-gene association for the co-expression model to recognize (Fig6 C). Unsurprisingly, WGCNA only detected one connection—between CD9 and PTDFRN—in this module because the Pearson's correlation between other gene pairs was low and the correlation pattern between these gene pairs could hardly be seen (Fig6 D). Based on the STRING database, only some of these gene pairs have been shown to have protein-protein interactions, whereas our co-perturbation model reveals more novel and potential gene-gene associations. Although CD9 (TSPAN29) and TSPAN18 do not interact on the protein level, both of them belong to the transmembrane 4 superfamily, and research shows that CD9 and TSPAN18 co-express in chick embryos during later spinal cord development[29]. Our co-perturbation model also reveals a connection between PTGFR (prostaglandin F2 receptor) and PTGFRN (prostaglandin F2 receptor inhibitor), where the two genes shared related functions. Other research has also demonstrated that these two genes have different expressions under different experimental conditions[30]. We believe that, due to the non-linear relationship between these genes, their associations were hard to detect using the co-expression model and scientists therefore tended to neglect them. Because our co-perturbation model is better equipped for this job, we now have the power to better discover and investigate these novel gene-gene associations.

Example of gene function prediction by Recount3 co-perturbation network

Interferon-induced protein with tetratricopeptide repeats (IFIT) genes are the well-known interferon-stimulated genes (ISG)[31]. This family contains four genes: IFIT1, IFIT2, IFIT3 and IFIT5. In our Recount3 co-perturbation network, we successfully identified all four IFIT genes and ISG15 in the same sub-network (Fig 7A). We also found that several other genes in this subnetwork shared similar disease-related functions. According to the OMIM (Online Mendelian Inheritance in Man) database, ISG15, IRF9, and STAT1 are related to immunodeficiency. Research has shown that these three genes belong to the ISG15 conjugation pathway, in which IRF9 interacts with phosphorylated STAT1 and STAT2 to form the ISGF3 complex; it then recognizes the interferon-sensitive response elements that promote ISG15 and its conjugation enzymes to induce ISG15 expression[32]. This same study confirmed that ISG15 plays an important role in the host antiviral response[32]. We also found that the IFIH1 and DDX58 genes are associated with Singleton-Merten syndrome, a multi-system innate immune disorder[33][34]. Reactome shows that they play a key role in DDX58/IFIH1-mediated induction of the interferon-alpha/beta pathway, which is also part of the innate immune system[35]. Further, SP110 mutations associate with hepatic venoocclusive disease with immunodeficiency, a rare form of severe combined immune deficiency[36]. Genes SAMD9 and SAMD9L in this sub-network are both related to monosomy 7 myelodysplasia and leukemia syndrome 1. Though these two genes are not immunodeficiency related, they are antiviral factors that play crucial roles in the innate immune system's defense against poxviruses[37].

Based on the findings detailed above, we are optimistic that we will be able to predict known genes' unknown gene functions. In the IFIT family subnetwork, ClinVar records the PARP10 gene as a pathogenic gene not yet linked to any specific disease phenotype[38]. But, because many genes in this subnetwork are immunodeficiency-related, we believe that PARP10 has a higher likelihood of being associated with immunodeficiency. As we expected, multiple studies have confirmed PARP10's critical role in regulating innate immune responses[39] and host-pathogen interactions[40].

Hub genes in Recount3 co-perturbation network are more essential

The Recount3 co-perturbation network consists of significantly co-perturbed gene pairs. Interestingly, we found that when we increase the FDR threshold, the network becomes increasingly scale-free (Sup Fig 5A), which means that a small number of genes connects with many neighbors while a large number of genes connect with very few neighbors. We analyzed these hub genes in the scale-free co-perturbation network and found that they are more likely to be disease- and lethality-related.

We ranked each gene based on its number of links and correlated this with the disease label from the OMIM database. We found that the likelihood of a gene being disease-related is positively correlated with the gene's degree in the network – the number of neighbors a gene has (Figure 7B). According to the scale-free co-perturbation network, more than half the genes contain fewer than 8 links and 13% of them are disease-related genes. And while only 2% of the genes have over 512 links, ~30% of them are disease related. We borrowed the mutation information from GnomAD (Genome Aggregation Database) and analyzed the relationship between the genes' degrees and mutation intolerance (Figure 7C). The result showed that the hub genes are less likely to tolerate loss-of-function (LoF) mutations. Around 80% of genes with fewer than 4 links tolerate LoF alleles. In other words, losing one gene copy is likely not lethal. In contrast, ~40% of genes with more than 512 links cannot tolerate LoF alleles. Therefore, we

believe that the central genes in the Recount3 co-perturbation network are more functional-essential genes. They have a stronger connection with the disease and a greater impact on vitality.

Discussion:

When analyzing integrated data from numerous experiments, the batch effect is a substantial obstacle that is difficult to overcome using a reliable statistical algorithm. We compared co-expression distribution between batch-corrected and non-batch-corrected GTEx data (Sup Fig 6A, B) and found that, for some tissues including brain, kidney, and blood, the peak is completely biased before batch correction. Although the distributions in other tissues have their peaks in low correlation regime, the density is still different, and there are more gene pairs with correlations close to zero after batch correlation. Thus, batch effect is significant in pooled GTEx data and can influence correlation distribution. We analyzed the correlation distribution and observed a biased peak in all three pooled data we used (Sup Fig 6C); these results are similar to the GTEx data before batch correction. Therefore, we believe that the batch effect is likely to occur when integrating expression data from numerous studies.

Another challenge is that gene-gene associations can be heterogeneous and vary under different experimental conditions. Thus, even after batch corrections, a combination of concordant changes and discordant changes between two genes will lead to a disordered and non-linear relationship. Such non-linear correlation usually has a low correlation score which co-expression models will struggle to identify. In addition, the correlations for the same gene pair vary with different contrasts; this makes combining these correlations unreliable. Even some significant correlations are not trustable. For instance, samples in each group are biological replicates so that they have very similar gene expression, and regression on such two groups are always significant as it's equivalent to do regression on two samples rather than two groups. Our co-perturbation model therefore focuses more on the number of contrasts amongst two genes differentially expressed together—we binarize the expression at the beginning to combat this issue. The large number of nonlinear gene-gene associations also suggests that the biological relationship between genes is not absolute but, rather, context-dependent. Finally, our co-perturbation model can successfully recognize correlation trends and is not confused by Simpson's paradox.

We demonstrated the co-perturbation model's power on one large dataset (Recount3) and two small but domain-specific datasets (Recount2 Brain and MAD). Data from Recount3 and Recount2 Brain are human data while MAD is mouse data. In these applications, the co-expression model discards almost all weakly correlated gene pairs and is more likely to identify highly correlated gene pairs in the pooled data. Other independent databases are less likely to validate those same gene pairs, which indicates that high correlation in pooled data can be false and actually caused by the batch effect. Our co-perturbation model, however, not only recognized highly correlated gene pairs in the pooled data, but also genes that have low correlation scores because they are non-linearly correlated in the integrated data. Thus, gene pairs with low correlation scores in the integrated data can still be biologically significant and are worthy of scientists' attention. On the whole, our co-perturbation model identifies reliable gene pairs in both high and low correlation regions.

Compared with the modules identified in Recount3, the GO enrichment test suggests that modules in the Recount2 Brain and in MAD share more common GO terms. While Recount2 Brain and MAD are domain-specific studies, there are far fewer unique genes in the final co-perturbation network than in the Recount3 co-perturbation network. And those genes are all detected in human brain-region-based or mouse-neurodegenerative-disease studies, so it is unsurprising that the different modules share many similar functions. GO term interaction density tests also proved that co-perturbation is biologically significant.

We used the several subnetworks to show that functionally related genes are connected to each other. Globally, we also showed that the hub genes in the scale-free co-perturbation network tend to be disease-related genes and functional essential genes. Thus, the co-perturbation network is not only statistically meaningful but also biologically meaningful in that it may reveal new biology.

With the current model, treating millions of gene pairs is incredibly time-consuming. Though generating a simulation-based binary DEG matrix is fast, fitting beta-binomial distribution for every single gene pair is not. Because calculating the random variable of a gene pair that will be co-perturbed is an NP (nondeterministic polynomial) problem, we used a simulation and fitting approach in the model. Now that the Recount3 co-perturbation network is available to the community, we hope and expect that it will drive new scientific insights, reveal novel gene connections, and predict unknown gene functions.

Methods:

All computational analysis was carried out in the R language (version 3.5.2 - 4.1.0).

Data collection and processing:

We obtained three datasets from 3 public databases including Recount3 (https://rna.recount.bio/), Recount2 Brain data (recount-brain project, Ramzara et al., bioRxiv, 2019), and MAD (https://doi.org/10.7303/syn2580853).

We collected 4,473 human experiments with a sample size between 6 to 40 from the Recount3 portal. We successfully extracted group label information in 3,293 experiments. For the remaining experiments, we used the DASC tool to identify hidden labels in 458 experiments. According to the label information, we found 19,681 total contrasts of which 18,997 contrasts were from manual labels and 684 were from DASC labels. These studies contained 56,498 total samples. We used DESeq2 to identify genes with fold changes greater than 1.5 and FDRs smaller than 0.01 as DEGs. Finally, we collected 12,082 contrasts with DEG numbers between 2 and 3,000, and 27,342 genes which have been identified as DEGs in at least 5% contrasts, as our input data. Most remaining genes are protein coding genes (Sup Fig 7).

We used a similar procedure to process data from Reount2 Brain and MAD. The only difference is that genes with fold changes greater than 1.2 and FDRs smaller than 0.01 were identified as DEGs. We collected 93 contrasts from 78 experiments from Recount2 Brain, which contains ~1,300 samples and 15,000 DEGs. We also collected 138 contrasts from 70 experiments in MAD, which contains ~1,000 samples and 10,000 DEGs.

Co-perturbation matrix generation:

First, we generated a binary DEG matrix in which the row reflects the gene, and the column reflects the experiment. If the gene is a DEG, we made the corresponding entry a 1 or a 0. Then we multiplied the DEG matrix by the transposed DEG matrix to create the gene-gene co-perturbation matrix, in which both rows and columns reflect genes, and each entry indicates how many experiments two genes co-perturbed.

Simulated co-perturbation matrix generation:

Based on the binary DEG matrix, we calculated the sum of rows and the sum of columns. Then, we used the greedy search to generate 1,000 random DEG matrices. Each simulated DEG matrix had the same sum of rows and sum of columns as the original binary matrix. We began by putting 1s in the row or column with the highest degree of freedom and then updated this data in all rows and columns. We repeated this process until the whole matrix was filled. If we failed to fill in 1s at any particular step, we would abandon the whole simulated matrix and start over. For each simulated random binary DEG matrix, we applied matrix multiplication to get the simulated co-perturbation matrix.

Beta-binomial fitting:

After generating 1,000 random co-perturbation matrices, we extracted each gene pair's co-perturbation time in 1,000 simulations. We used VGAM to fit beta-binomial distribution on every gene pair's co-perturbation distribution. The fitting was generally accurate as shown in Sup Fig 8. We used each gene pair's real co-perturbation in the true co-perturbation matrix as the threshold for the corresponding beta-binomial fitting curves. We then calculated the p-values and FDR. Gene pairs with significant FDRs were connected in the final co-perturbation network.

Determining non-linear score of co-perturbed gene pair

We defined the non-linear gene pair score as $1 - \left|1 - \frac{2N_{+/-}}{N}\right|$, where N is the number of experiments in which a gene pair co-perturbed and $N_{+/-}$ is the number of experiments in which one gene up-regulated while the other gene down-regulated. The overall idea was to make the non-linear score 1 when half of the contrasts showed concordant change and the other half showed discordant change, and to make it 0 when all contrasts showed concordant or discordant change.

Co-perturbation heatmap

We extracted contrasts that a gene pair co-perturbed; colored the contrasts that a gene upregulated red; and colored the contrasts that a gene downregulated blue. The color scale is based on the genes' log2FC.

Scatter plot of genes in pooled Recount3 data and in co-perturbed contrasts

RNA-seq data from Recount3 were log2 transformed and combined without batch correction. The overall Pearson's correlation was also calculated. We extracted contrasts in which two genes were concordantly co-perturbed, assigned samples in the same contrast the same color, and processed the discordantly co-perturbed samples the same way. We generated the color using the "rainbow" palette in R, but due to the large number of contrasts, samples from different contrasts may have similar colors.

Module detection:

We used the recursive Louvain module detection method to identify modules in the co-perturbation network. First, we ran the Louvain clustering on the whole network to get several large modules and then ran the GO term enrichment on each module. Second, if a module's genes were significantly enriched in multiple GO terms, we continued to use Louvain clustering on that module to get more sub-modules until the division would cause one sub-module to have no GO terms. Then, we returned to the previous step and kept the module whole.

Binary heatmap of gene modules

The heatmap row represented the contrasts and the color bar distinguished contrasts from different studies by using different colors. The heatmap column was genes and the color bar indicates genes from different modules. Genes in grey did not belong to a module. Genes in the same module was ordered by the DEG frequency.

GO enrichment comparison

We used the clusterProfiler package to make the plot. For each gene module, we ran the GO enrichment test and took the significant GO terms with an FDR smaller than 0.05. For each module we only show the top significant GO terms. Then, using the dot plot as a binary matrix where each row is a GO term and each column is a gene module, we clustered the modules based on their Euclidean distance.

GO term interaction density comparison:

The Go term density was defined as Density(t) = NE(t)/NP(t), where NE(t) is the number of edges between genes assigned to term t, and NP(t) is the number of gene pairs in the term, which equals to C_n^2, where n is the number of genes in t.

Validation by PPI

Protein-protein interaction data was downloaded from the STRING database. For recount data, we applied Homo sapiens PPI data, and for the MAD study we applied Mus musculus PPI data to perform the validation. We matched the gene pairs from the co-perturbation model and co-expression model with the PPI data and calculated the percentage of gene pairs validated by STRING in each model.

Correlation distribution

In the correlation figure in Recount3, we generated the background distribution with Pearson's correlation of all gene pairs in the Recount3 data. We also calculated and compared the Pearson's correlations of gene pairs that the co-perturbation model identified and gene pairs that WGCNA identified.

Since Recount3 covers gene data from different organs, we validated our results on GTExMaximum, which is the maximum Pearson's correlation of the gene pair in all GTEx tissues. We first downloaded RNA seq data from the GTEx portal and then performed batch correction for each tissue type using the limma package. We checked the distribution of gene-gene Pearson's correlations in each tissue type and removed four tissues (Bladder, Fallopian Tube, Cervix-Endocervix and Cervix-Ectocervix) with poor quality due to the small sample size. Finally, for each gene pair, we picked the maximum correlation score in all remaining tissues to represent its correlation. The background distribution contains all genes recognized by either the WGCNA or co-perturbation model. For gene modules identified by WGCNA, we

removed the largest module, which included more than 10,000 genes (which is almost half the total genes).

In the Recount2 Brain project, we validated our results by GTEx brain data, as studies from this project were all specific to the brain region. All GTEx data were batch corrected using the R package limma.

In MAD data, we matched the mouse genes with human genes first, and then used batch corrected GTEx brain data and ROSMAP data to do the validation. ROSMAP data mainly contains gene expression data from patients with Alzheimer's disease and dementia.

Sub-network examples

We chose the module examples based on linear rate, gene-gene association FDRs, and the number of genes. We chose the disease related module using an extremely high FDR cutoff and then matched these genes with the OMIM database.

We generated the heatmap using the same approach we used to generate the co-perturbation heatmap. The only difference is that we did not scale the color per log2FC because the number of contrasts (number of rows in the heatmap) was too large. We generated the scatter plot the same way we generated the scatter plot of genes in pooled Recount3 data.

Property of hub genes investigation

We downloaded the disease gene label from the OMIM database. We designated all genes with disease phenotypes as disease-related genes. Then, we downloaded LoF scores of all the genes from the GnomAD database and we defined the genes with LoF scores smaller than 0.35 as genes intolerant of LoF alleles. The scale-free network was generated with an FDR cutoff of 10^{-6}. We ordered the genes by their degree in the network and divided the degree into 9 groups with an increasing degree step of 2^n. Then, we calculated the percentage of OMIM genes and of genes intolerant of mutations in each group. Finally, we made the scatter plot using gene degree as the x-axis and percentage of genes as the y-axis.

Runtime estimation

In order to pinpoint the relationship between runtime, matrix size, and matrix sparsity when generating the random binary matrix under the given row sum and column sum, we made the sparsity 0.95 (the real sparsity is usually much higher), made the column dimension 50, increased the row dimension gradually from 10 to 50,000, and recorded the runtime it took to generate a single random binary matrix. Then, we made both the row and column dimensions 1,000, increased the sparsity from 0.70 to 0.99, and recorded the runtime.

Relationship between co-perturbation times and significance

We selected two gene pairs to show that higher co-perturbation times do not always have higher significance. We plotted the histogram of co-perturbation times based on 1,000 simulations and the 1-ecdf figure based on the simulation results. However, overall, higher co-perturbation times yield higher significance, which can be seen from our comparison of the histogram of co-perturbation times of all genes versus co-perturbed genes.

Soft threshold determination

We applied different FDRs as cutoffs for significantly co-perturbed gene pairs, and for each FDR, we generated the network and analyzed the relationship between the network degree and number of nodes in the network. We applied linear regression to log the transformed network degree and the transformed number of nodes and calculated the R square to represent the network's scale-free property. We chose the largest FDR as the cutoff when the next smaller FDR gave us a smaller scale free fit R square.

GTEx correlation histogram before and after batch correction

We downloaded the original GTEx expression data and calculated the Pearson's correlation as the before-batch-correction histogram. Then, we applied the limma package to remove the batch effect and performed the same calculation to get the after-batch-correction histogram.

Code availability:

Our co-perturbation model R package can be accessed though:
https://github.com/W765623192/CoPNet

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Availability of data and materials
The datasets analyzed during the current study are available:
- https://rna.recount.bio/, Recount3
- http://research.libd.org/recount-brain/, Recount2 Brain
- https://doi.org/10.7303/syn2580853, MAD

Competing interests
The authors declare no competing interests.

Funding
Research reported in this publication was supported by the Eunice Kennedy Shriver National 168Institute of Child Health & Human Development of the National Institutes of Health under 169Award Number P50HD103555for use of the Bioinformatics Core facilities. The content is 170solely the responsibility of the authors and does not necessarily represent the official views of the 171National Institutes of Health.

Acknowledgements
We thank Alexander J. Trostle, Hyun-hwan Jeong for their contributions; Sasidhar Pasupuleti for his help managing the network website.

Authors' contributions
J.W. and Z.L. led the research. J.W developed the model and pipeline. J.W., Y.-W.W. and R.A.-O. analyzed the networks. J.W., Y.-W.W. and Z.L. wrote the paper.

Reference:

[1] Sun, K., Gonçalves, J., Larminie, C., & Pržulj, N. (2014). Predicting disease associations via biological network analysis. BMC Bioinformatics, 15(1). doi: 10.1186/1471-2105-15-304

[2] Gerring, Z., Gamazon, E., & Derks, E. (2019). A gene co-expression network-based analysis of multiple brain tissues reveals novel genes and molecular pathways underlying major depression. PLOS Genetics, 15(7), e1008245. doi: 10.1371/journal.pgen.1008245

[3] Szklarczyk, D., Gable, A., Nastou, K., Lyon, D., Kirsch, R., & Pyysalo, S. et al. (2021). Correction to 'The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets'. Nucleic Acids Research, 49(18), 10800-10800. doi: 10.1093/nar/gkab835

[4] Alonso-López, D., Campos-Laborie, F., Gutiérrez, M., Lambourne, L., Calderwood, M., Vidal, M., & De Las Rivas, J. (2019). APID database: redefining protein–protein interaction experimental evidences and binary interactomes. Database, 2019. doi: 10.1093/database/baz005

[5] Slenter, D., Kutmon, M., Hanspers, K., Riutta, A., Windsor, J., & Nunes, N. et al. (2017). WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Research, 46(D1), D661-D667. doi: 10.1093/nar/gkx1064

[6] Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., & Itoh, M. et al. (2007). KEGG for linking genomes to life and the environment. Nucleic Acids Research, 36(Database), D480-D484. doi: 10.1093/nar/gkm882

[7] Obayashi, T., Kagaya, Y., Aoki, Y., Tadaka, S., & Kinoshita, K. (2018). COXPRESdb v7: a gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference. Nucleic Acids Research, 47(D1), D55-D62. doi: 10.1093/nar/gky1155

[8] Wilks, C., Zheng, S., Chen, F., Charles, R., Solomon, B., & Ling, J. et al. (2021). recount3: summaries and queries for large-scale RNA-seq expression and splicing. Genome Biology, 22(1). doi: 10.1186/s13059-021-02533-6

[9] Papatheodorou, I., Moreno, P., Manning, J., Fuentes, A., George, N., & Fexova, S. et al. (2019). Expression Atlas update: from tissues to single cells. Nucleic Acids Research. doi: 10.1093/nar/gkz947

[10] Bonett, D., & Wright, T. (2000). Sample size requirements for estimating pearson, kendall and spearman correlations. Psychometrika, 65(1), 23-28. doi: 10.1007/bf02294183

[11] Leek, J., Scharpf, R., Bravo, H., Simcha, D., Langmead, B., & Johnson, W. et al. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics, 11(10), 733-739. doi: 10.1038/nrg2825

[12] Tan J, Hammond JH, Hogan DA, Greene CS. ADAGE-Based Integration of Publicly Available Pseudomonas aeruginosa Gene Expression Data with Denoising Autoencoders Illuminates Microbe-Host Interactions. mSystems. 2016;1(1). doi:10.1128/mSystems.00025-15

[13] Tan J, Doing G, Lewis KA, et al. Unsupervised Extraction of Stable Expression Signatures from Public Compendia with an Ensemble of Neural Networks. Cell Syst. 2017;5(1):63-71.e6. doi:10.1016/j.cels.2017.06.003

[14] Zhou W, Altman RB. Data-driven human transcriptomic modules determined by independent component analysis. BMC Bioinformatics. 2018;19(1):327. doi:10.1186/s12859-018-2338-4

[15] Taroni JN, Grayson PC, Hu Q, et al. MultiPLIER: a transfer learning framework for transcriptomics reveals systemic features of rare disease. Cell Syst. 2019;8(5):380-394

[16]Smyth, G., & Speed, T. (2003). Normalization of cDNA microarray data. Methods, 31(4), 265-273. doi: 10.1016/s1046-2023(03)00155-5
[17]Zhang, Y., Parmigiani, G., & Johnson, W. (2020). ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genomics And Bioinformatics, 2(3). doi: 10.1093/nargab/lqaa078
[18]Lee A, Park Y, Doing G, Hogan D, Greene C. Correcting for experiment-specific variability in expression compendia can remove underlying signals. 2020. doi:10.1101/2020.05.03.066597
[19]Razmara, A., Ellis, S., Sokolowski, D., Davis, S., Wilson, M., & Leek, J. et al. (2019). recount-brain: a curated repository of human brain RNA-seq datasets metadata. doi: 10.1101/618025
[20]Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics, 9(1). doi: 10.1186/1471-2105-9-559
[21]Blyth, C. (1972). On Simpson's Paradox and the Sure-Thing Principle. Journal Of The American Statistical Association, 67(338), 364-366. doi: 10.1080/01621459.1972.10482387
[22]Furlotte, N., Kang, H., Ye, C., & Eskin, E. (2011). Mixed-model coexpression: calculating gene coexpression while accounting for expression heterogeneity. Bioinformatics, 27(13), i288-i294. doi: 10.1093/bioinformatics/btr221
[23]Yi, H., Raman, A., Zhang, H., Allen, G., & Liu, Z. (2017). Detecting hidden batch factors through data-adaptive adjustment for biological effects. Bioinformatics, 34(7), 1141-1147. doi: 10.1093/bioinformatics/btx635
[24]Wang, B., Wu, P., Kwan, B., Tu, X. M., & Feng, C. (2018). Simpson's paradox: examples. *Shanghai archives of psychiatry*, *30*(2).
[25]Blondel, V., Guillaume, J., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal Of Statistical Mechanics: Theory And Experiment, 2008(10), P10008. doi: 10.1088/1742-5468/2008/10/p10008
[26]Szklarczyk, D., Gable, A., Lyon, D., Junge, A., Wyder, S., & Huerta-Cepas, J. et al. (2018). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Research, 47(D1), D607-D613. doi: 10.1093/nar/gky1131
[27]Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., & Shad, S. et al. (2013). The Genotype-Tissue Expression (GTEx) project. Nature Genetics, 45(6), 580-585. doi: 10.1038/ng.2653
[28]Bennett, D., Buchman, A., Boyle, P., Barnes, L., Wilson, R., & Schneider, J. (2018). Religious Orders Study and Rush Memory and Aging Project. Journal Of Alzheimer's Disease, 64(s1), S161-S189. doi: 10.3233/jad-179939
[29]Perron, J., & Bixby, J. (1999). Tetraspanins expressed in the embryonic chick nervous system. FEBS Letters, 461(1-2), 86-90. doi: 10.1016/s0014-5793(99)01429-5
[30]Tada, A., Islam, M., Kober, A., Fukuyama, K., Takagi, M., & Igata, M. et al. (2020). Transcriptome Modifications in the Porcine Intramuscular Adipocytes during Differentiation and Exogenous Stimulation with TNF-α and Serotonin. International Journal Of Molecular Sciences, 21(2), 638. doi: 10.3390/ijms21020638
[31]Pidugu, V., Pidugu, H., Wu, M., Liu, C., & Lee, T. (2019). Emerging Functions of Human IFIT Proteins in Cancer. Frontiers In Molecular Biosciences, 6. doi: 10.3389/fmolb.2019.00148
[32]Perng, Y., & Lenschow, D. (2018). ISG15 in antiviral immunity and beyond. Nature Reviews Microbiology, 16(7), 423-439. doi: 10.1038/s41579-018-0020-5
[33]Jang, M., Kim, E., Now, H., Nguyen, N., Kim, W., & Yoo, J. et al. (2015). Mutations in DDX58, which Encodes RIG-I, Cause Atypical Singleton-Merten Syndrome. The American Journal Of Human Genetics, 96(2), 266-274. doi: 10.1016/j.ajhg.2014.11.019
[34]Ferreira, C., Crow, Y., Gahl, W., Gardner, P., Goldbach-Mansky, R., & Hur, S. et al. (2018). DDX58 and Classic Singleton-Merten Syndrome. Journal Of Clinical Immunology, 39(1), 75-80. doi: 10.1007/s10875-018-0572-1

[35] Croft, D., Mundo, A., Haw, R., Milacic, M., Weiser, J., & Wu, G. et al. (2013). The Reactome pathway knowledgebase. Nucleic Acids Research, 42(D1), D472-D477. doi: 10.1093/nar/gkt1102
[36] Wang, T., Ong, P., Roscioli, T., Cliffe, S., & Church, J. (2012). Hepatic veno-occlusive disease with immunodeficiency (VODI): First reported case in the U.S. and identification of a unique mutation in Sp110. Clinical Immunology, 145(2), 102-107. doi: 10.1016/j.clim.2012.07.016
[37] Peng, S., Meng, X., Zhang, F., Pathak, P., Chaturvedi, J., & Coronado, J. et al. (2022). Structure and function of an effector domain in antiviral factors and tumor suppressors SAMD9 and SAMD9L. Proceedings Of The National Academy Of Sciences, 119(4). doi: 10.1073/pnas.2116550119
[38] Landrum, M., Chitipiralla, S., Brown, G., Chen, C., Gu, B., & Hart, J. et al. (2019). ClinVar: improvements to accessing data. Nucleic Acids Research, 48(D1), D835-D844. doi: 10.1093/nar/gkz972
[39] Zhu, H., Tang, Y., Zhan, G., Su, C., & Zheng, C. (2021). The Critical Role of PARPs in Regulating Innate Immune Responses. Frontiers In Immunology, 12. doi: 10.3389/fimmu.2021.712556
[40] Fehr, A. R., Singh, S. A., Kerr, C. M., Mukai, S., Higashi, H., & Aikawa, M. (2020). The impact of PARPs and ADP-ribosylation on inflammation and host-pathogen interactions. Genes & development, 34(5-6), 341–359. https://doi.org/10.1101/gad.334425.119

Figure 1

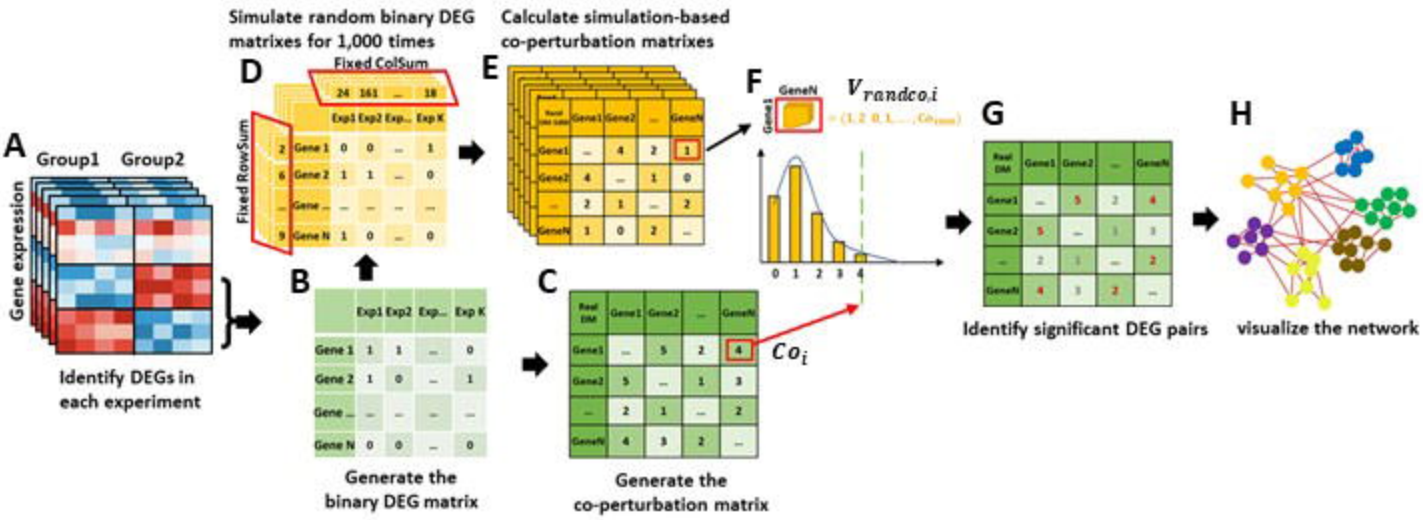

Figure 1. Co-perturbation network framework. **A)** Identify co-perturbed DEG genes in all contrasts. Each contrast contains two groups with a given label or a DASC label. **B)** Generate the binary DEG matrix with rows of genes and columns of contrasts. DEG in the corresponding contrast is marked as 1 otherwise 0. **C)** Generate the co-perturbation matrix with both rows and columns of genes. The value in the matrix represents the co-perturbation time of each gene pair. **D)** 1,000 random binary DEG matrices will be simulated with the same row and column sums as the true binary DEG matrix. **E)** 1,000 simulation-based co-perturbation matrices will be calculated based on each of the random binary DEG matrices. **F)** Generate the random co-perturbation distribution for each gene pair. Apply beta-binomial distribution to fit it and take the true co-perturbation times as the cutoff to identify significant DEG pairs. **G)** Calculate the FDR to find the most significant DEG pairs. H) Visualized the network based on the adjacency matrix.

Figure 2

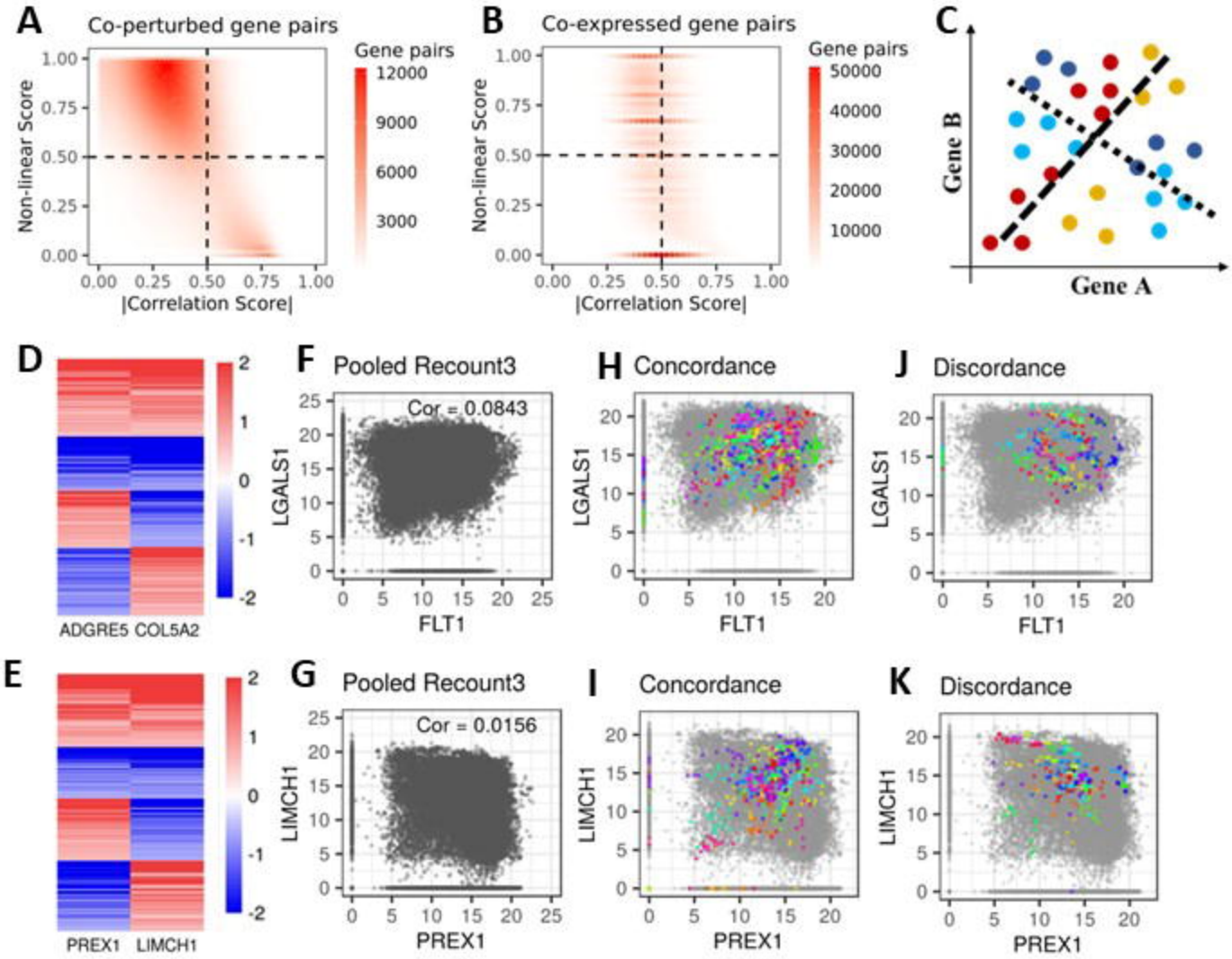

Figure 2. Co-perturbation model can identify non-linear correlation in the integrated data. **A)** Co-perturbed gene pairs with low correlation scores tend to have higher non-linear scores in the integrated data. **B)** Co-expressed gene pairs tend to have a higher mean correlation score regardless of non-linear score. **C)** Schematic diagram of non-linear correlation. Samples in warm colors are concordantly correlated while samples in cold colors are discordantly correlated. **D,E)** The log2FC heatmap of non-linear correlated gene pairs. **F,G)** The gene pair's Pearson's correlation score. **H,I)** Samples from contrasts that are concordantly co-perturbed. **J,K)** Samples from contrasts that are discordantly co-perturbed.

Figure 3

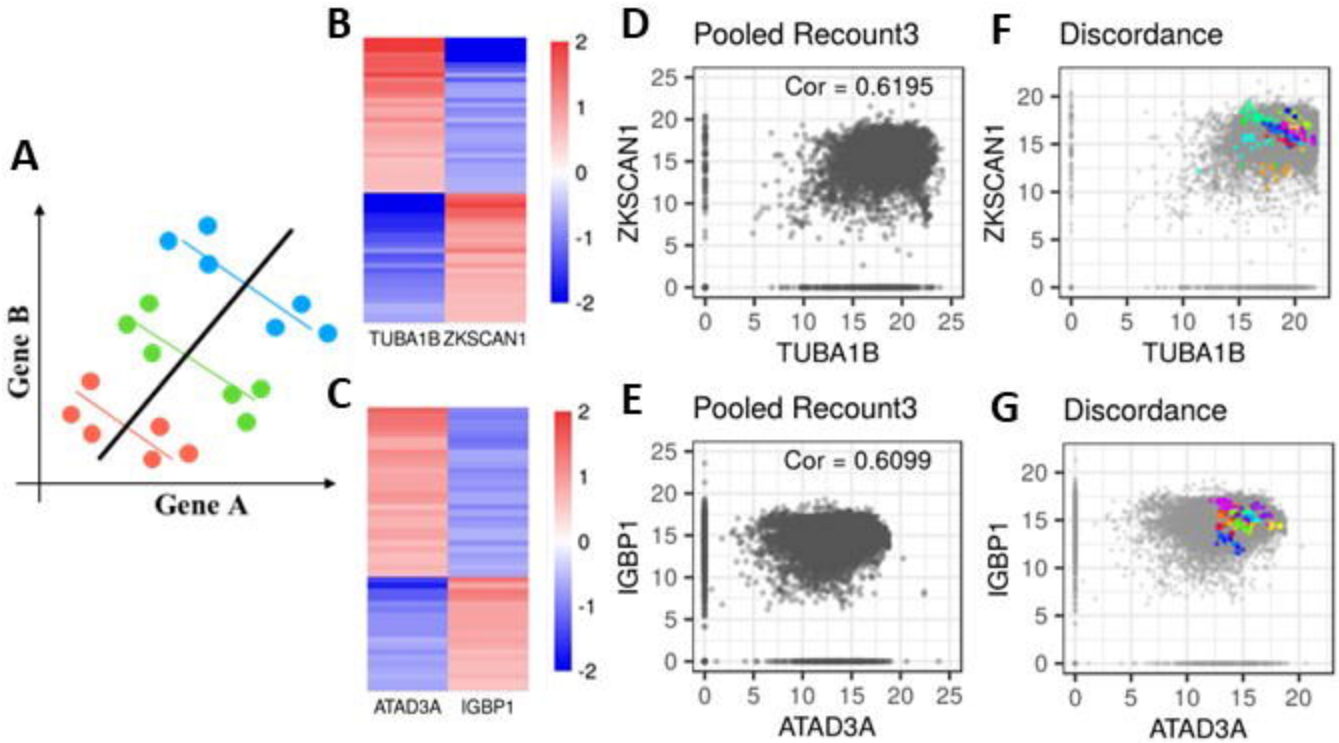

Figure 3. Simpson's paradox: A) Schematic diagram of Simpson's paradox. B,C) The log2FC heatmap of negatively correlated gene pairs. D,E) The gene pair's Pearson's correlation score is positive. F,G) Samples are only discordantly co-perturbed in each contrast.

Figure 4

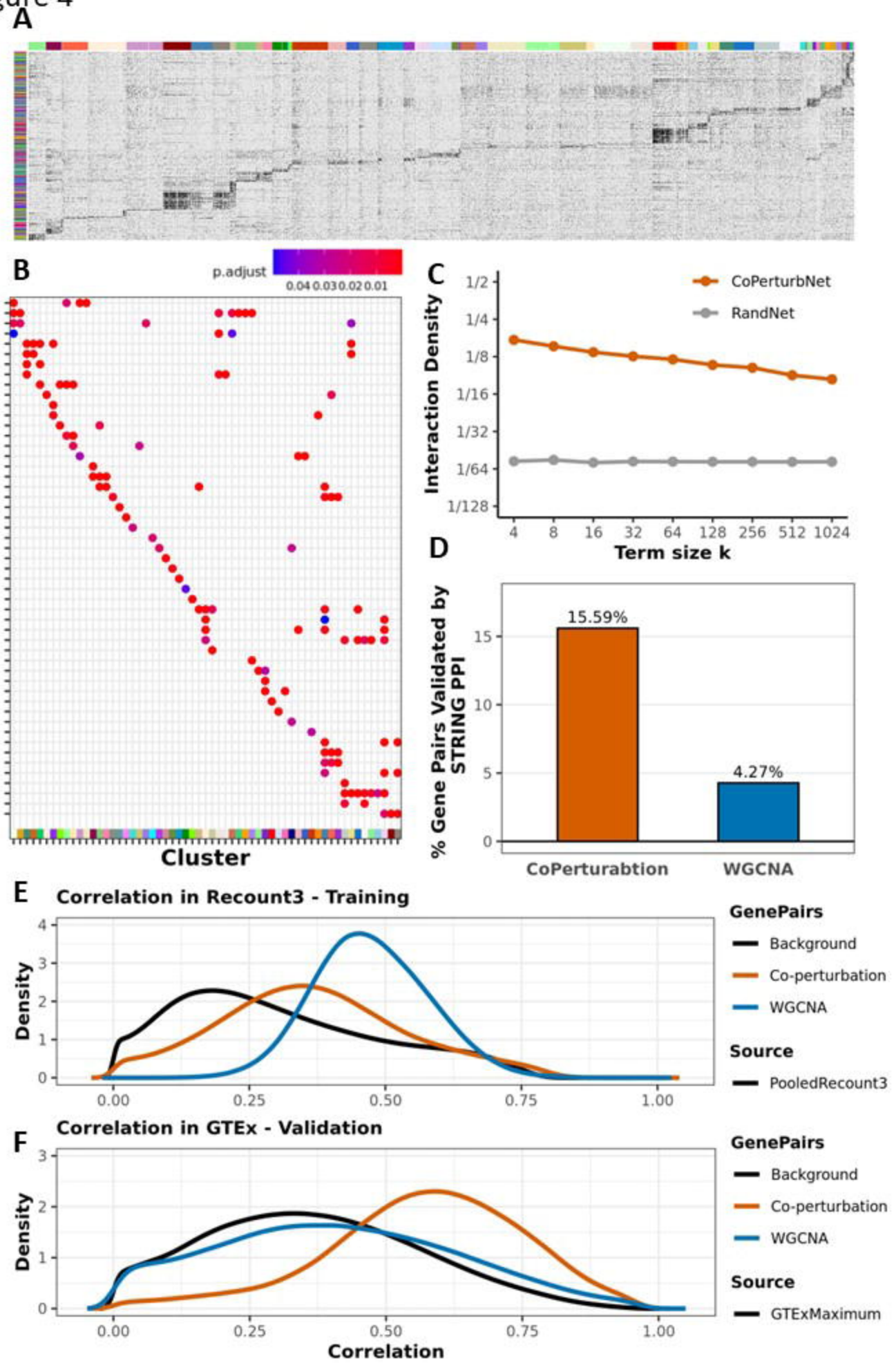
A
B
p.adjust
0.04 0.03 0.02 0.01
Cluster
C
CoPerturbNet
RandNet
Interaction Density
1/2
1/4
1/8
1/16
1/32
1/64
1/128
4 8 16 32 64 128 256 512 1024
Term size k
D
% Gene Pairs Validated by STRING PPI
15.59%
4.27%
0
5
10
15
CoPerturabtion
WGCNA
E
Correlation in Recount3 - Training
Density
0.00 0.25 0.50 0.75 1.00
GenePairs
Background
Co-perturbation
WGCNA
Source
PooledRecount3
F
Correlation in GTEx - Validation
Density
0.00 0.25 0.50 0.75 1.00
Correlation
GenePairs
Background
Co-perturbation
WGCNA
Source
GTExMaximum

Figure 4. The co-perturbation network on Recount3. **A)** Binary co-perturbation heatmap with rows of contrasts and columns of genes. The left color bar represents different contrasts, and the top color bar represents genes in different modules. **B)** GO enrichment comparison between different gene modules. **C)** GO term interaction density between Recount3 co-perturbation network and random network. **D)** PPI validation between co-perturbed gene pairs and co-expressed gene pairs. **E)** Comparison of Pearson's correlation between co-expressed gene pairs and co-perturbed gene pairs in Recount3 data. **F)** Validation by comparing Pearson's correlation of co-expressed and co-perturbed gene pairs from Recount3 in the independent GTEx data.

Figure 5

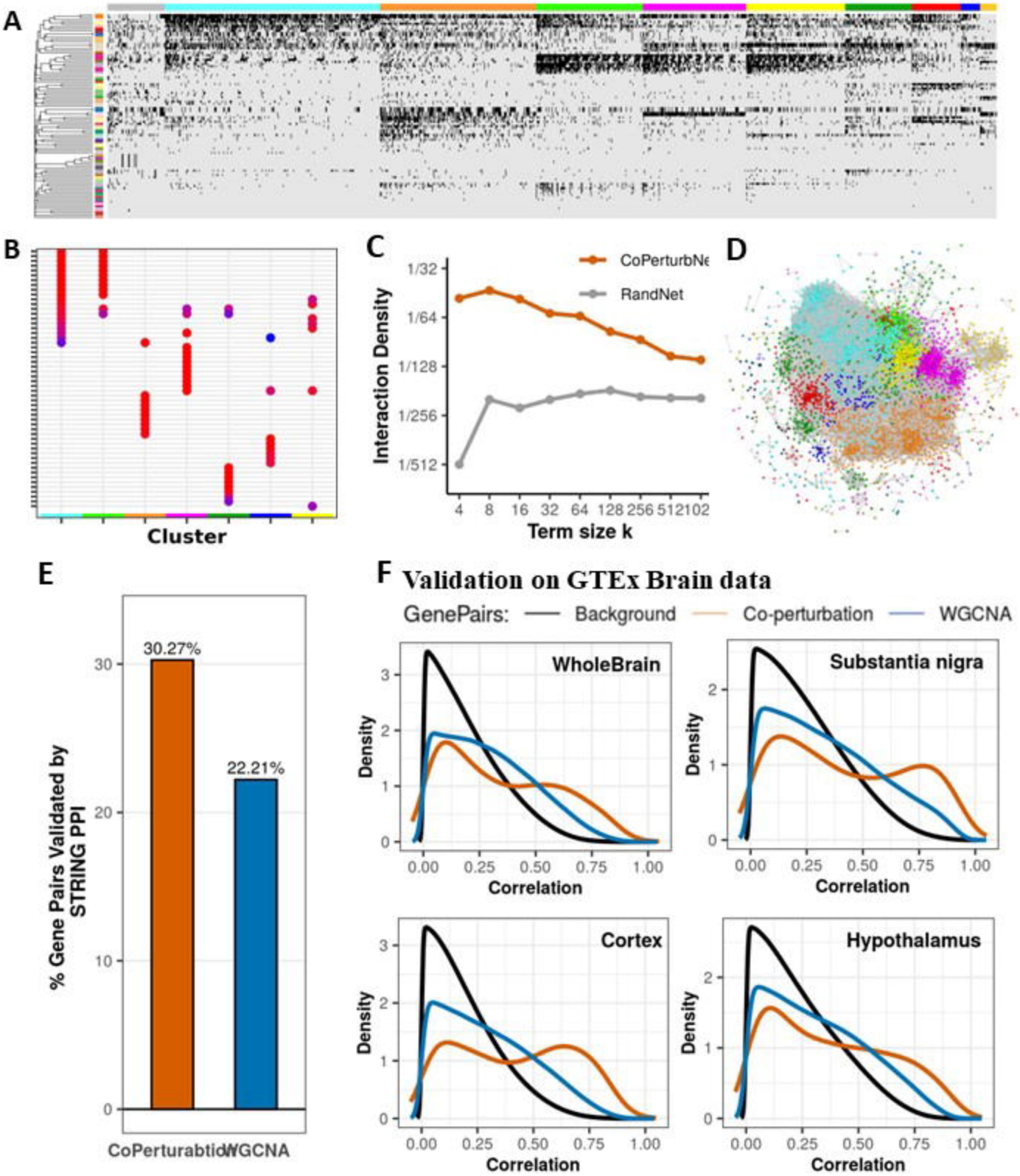

Figure 4. The co-perturbation network on Recount2 Brain. **A)** Binary co-perturbation heatmap with rows of contrasts and columns of genes. The left color bar represents different contrasts, and the top color bar represents genes in different modules. **B)** GO enrichment comparison between different gene modules. **C)** GO term interaction density between Recount2 Brain co-perturbation network and random network. **D)** Visualized Recount2 Brain co-perturbation network. **E)** PPI validation between co-perturbed gene pairs and co-expressed gene pairs. **F)** Validation by comparing Pearson's correlation of co-expressed and co-perturbed gene pairs from Recount2 Brain in the independent GTEx data.

Figure 6

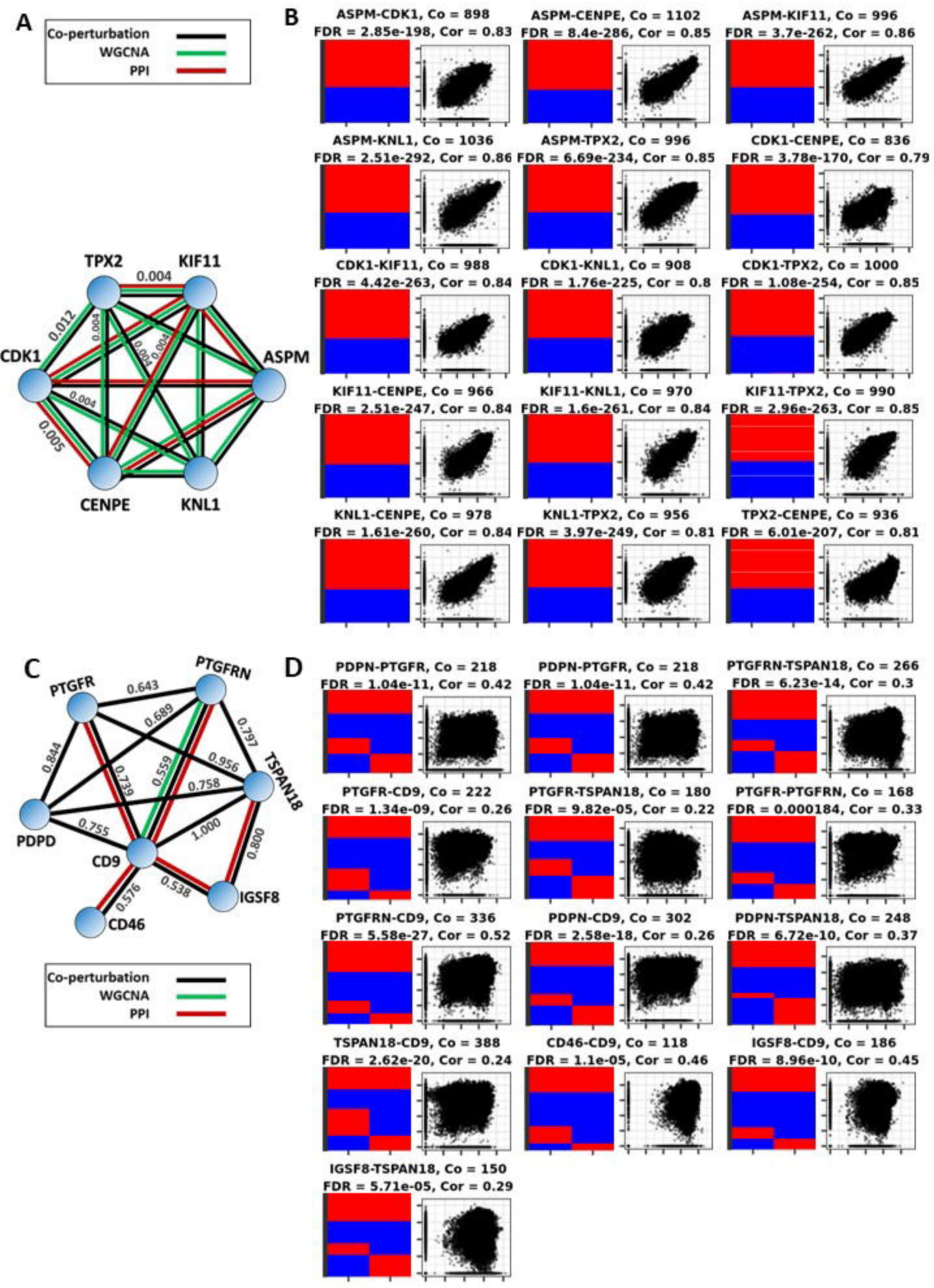
A
Co-perturbation
WGCNA
PPI
TPX2
KIF11
CDK1
ASPM
CENPE
KNL1
0.004
0.012
0.005
B
ASPM-CDK1, Co = 898
FDR = 2.85e-198, Cor = 0.83
ASPM-CENPE, Co = 1102
FDR = 8.4e-286, Cor = 0.85
ASPM-KIF11, Co = 996
FDR = 3.7e-262, Cor = 0.86
ASPM-KNL1, Co = 1036
FDR = 2.51e-292, Cor = 0.86
ASPM-TPX2, Co = 996
FDR = 6.69e-234, Cor = 0.85
CDK1-CENPE, Co = 836
FDR = 3.78e-170, Cor = 0.79
CDK1-KIF11, Co = 988
FDR = 4.42e-263, Cor = 0.84
CDK1-KNL1, Co = 908
FDR = 1.76e-225, Cor = 0.8
CDK1-TPX2, Co = 1000
FDR = 1.08e-254, Cor = 0.85
KIF11-CENPE, Co = 966
FDR = 2.51e-247, Cor = 0.84
KIF11-KNL1, Co = 970
FDR = 1.6e-261, Cor = 0.84
KIF11-TPX2, Co = 990
FDR = 2.96e-263, Cor = 0.85
KNL1-CENPE, Co = 978
FDR = 1.61e-260, Cor = 0.84
KNL1-TPX2, Co = 956
FDR = 3.97e-249, Cor = 0.81
TPX2-CENPE, Co = 936
FDR = 6.01e-207, Cor = 0.81
C
PTGFR
PTGFRN
TSPAN18
PDPD
CD9
IGSF8
CD46
0.643
0.689
0.797
0.844
0.739
0.559
0.956
0.758
0.755
1.000
0.800
0.576
0.538
Co-perturbation
WGCNA
PPI
D
PDPN-PTGFR, Co = 218
FDR = 1.04e-11, Cor = 0.42
PDPN-PTGFR, Co = 218
FDR = 1.04e-11, Cor = 0.42
PTGFRN-TSPAN18, Co = 266
FDR = 6.23e-14, Cor = 0.3
PTGFR-CD9, Co = 222
FDR = 1.34e-09, Cor = 0.26
PTGFR-TSPAN18, Co = 180
FDR = 9.82e-05, Cor = 0.22
PTGFR-PTGFRN, Co = 168
FDR = 0.000184, Cor = 0.33
PTGFRN-CD9, Co = 336
FDR = 5.58e-27, Cor = 0.52
PDPN-CD9, Co = 302
FDR = 2.58e-18, Cor = 0.26
PDPN-TSPAN18, Co = 248
FDR = 6.72e-10, Cor = 0.37
TSPAN18-CD9, Co = 388
FDR = 2.62e-20, Cor = 0.24
CD46-CD9, Co = 118
FDR = 1.1e-05, Cor = 0.46
IGSF8-CD9, Co = 186
FDR = 8.96e-10, Cor = 0.45
IGSF8-TSPAN18, Co = 150
FDR = 5.71e-05, Cor = 0.29

Figure 6. Examples of Recount3 sub-networks. **A)** A linearly correlated sub-network. Scores on the edges are the non-linear scores between two genes. Edges without scores are completely linear. **B)** Gene pair details in figure A including the up-down-regulation heatmap (Each column is a gene, and each row is a co-perturbed contrast, red color means the gene is upregulated in that contrast and blue color means the gene is downregulated). "Co" means the number of contrasts the pair co-perturbed. FDR is the significance of co-perturbation, and "Cor" means the Pearson's correlation score. **C)** A non-linearly correlated sub-network. **D)** Gene pair details in figure C including the up-down-regulation heatmap, co-perturbation times, FDR, and Pearson's correlation score. (Same explanation as Figure B)

Figure 7

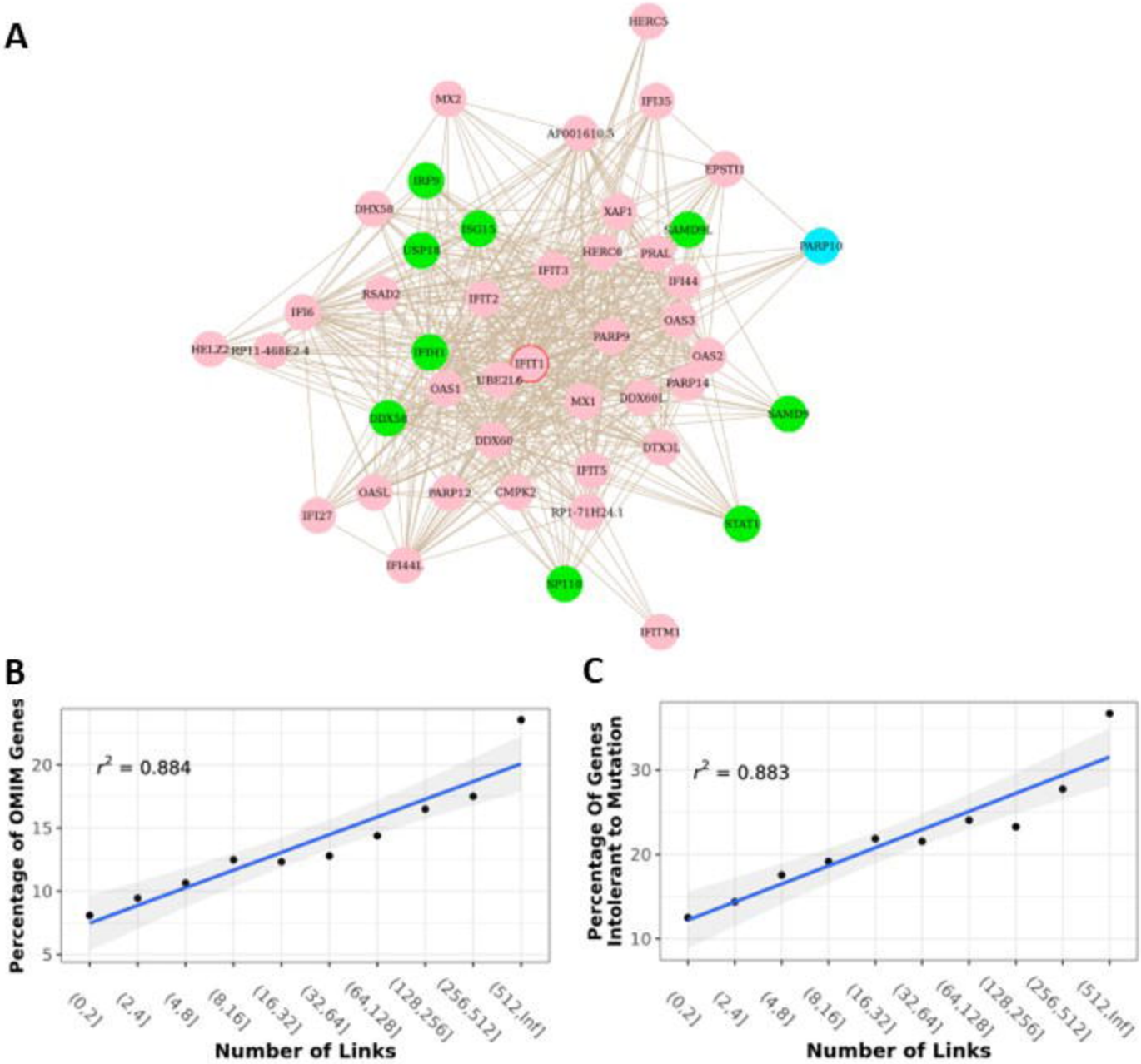

Figure 7. Disease-related gene function prediction. Based on the assumption of guilt-by-association, we demonstrated a subnetwork for gene function prediction. Besides, we observed a striking correlation between the hubness of our co-perturbation network and the disease gene frequency. **A)** The IFIT family subnetwork. The green genes have a known OMIM phenotype. The blue genes lack known OMIM phenotypes but are recorded as pathogenic in ClinVar. The pink genes lack known OMIM phenotypes and are recorded as non-pathogenic in ClinVar. **B)** The node degree is positively related with the percentage of OMIM genes. **C)** The node degree is positively related with the percentage of lethal genes.