SCIENTIFIC REPORTS
nature research

Check for updates

OPEN

Association study based on topological constraints of protein–protein interaction networks

Hao-Bo Guo[1,2]✉ & Hong Qin[1,2,3]✉

The non-random interaction pattern of a protein–protein interaction network (PIN) is biologically informative, but its potentials have not been fully utilized in omics studies. Here, we propose a network-permutation-based association study (NetPAS) method that gauges the observed interactions between two sets of genes based on the comparison between permutation null models and the empirical networks. This enables NetPAS to evaluate relationships, constrained by network topology, between gene sets related to different phenotypes. We demonstrated the utility of NetPAS in 50 well-curated gene sets and comparison of association studies using Z-scores, modified Z′-scores, p-values and Jaccard indices. Using NetPAS, a weighted human disease network was generated from the association scores of 19 gene sets from OMIM. We also applied NetPAS in gene sets derived from gene ontology and pathway annotations and showed that NetPAS uncovered functional terms missed by DAVID and WebGestalt. Overall, we show that NetPAS can take topological constraints of molecular networks into account and offer new perspectives than existing methods.

Interactomes, particularly the protein–protein interaction networks (PINs) from model organisms and humans[1–8], have shifted our interests in molecular biology from the functions of individual genes or proteins to functional modules of PINs, including the modules associated with human diseases[9]. PINs can be treated as graphs in which vertices (nodes) are proteins and edges (links) are protein–protein interactions. PINs possess characteristics observed in other real-world graphs, such as small-world[10], scale-free[11], and error-tolerance[12,13], suggesting that the topological patterns of PINs can offer biological insights.

Tools such as Gene Set Enrichment Analysis (GSEA)[14] and pathway analysis[15] have become routine to extract shared characteristics of gene sets obtained from omics experiments. These analyses are often based on knowledge bases such as the gene ontology (GO) knowledge database[16], gene pathway databases such as KEGG[17], and the interaction networks connecting genes or gene products[18]. Gene set[19] (e.g., GSEA) and pathway[20] analysis methods typically adopt statistical methods including Fisher's, binomial, hypergeometric distribution, Chi-square, linear regression, or logistic regression[21,22] to score the associations between the gene set and GO, pathways, or other functional terms. Another work uses the cohesion coefficient to measure the association among pathways, annotations and gene sets[23,24]. An underlying assumption of these analyses is that biological events in the cell are often conducted by groups of genes via direct, physical interactions, which are collectively called the interactome[25]. In this regard, methods that are directly based on interactome, such as PIN, can take unique biological constraints into accounts and may offer more biologically relevant results than simple enrichment tests.

Multiple network-based approaches have been developed for functional predictions of gene sets. EnrichNet uses prioritization scores to expand the interested protein via random walks over the PIN[26]. WebGestalt unifies over-representation analysis (ORA), GSEA, and network topology-based analysis into a gene set analysis toolkit[27]. Other tools including NET-GE[28] and pathfindR[29] aim to identify densely connected or functionally related subnetworks from the protein–protein interactions to strengthen the enrichment analysis. The PAGER database, on the other hand, has integrated gene-set, network, and pathway analysis (GNPA) data resources into a gene-signature electronic repository[23,24].

[1]Department of Computer Science and Engineering, The University of Tennessee, Chattanooga, USA. [2]SimCenter, The University of Tennessee, Chattanooga, USA. [3]Department of Biology, Geology and Environmental Science, The University of Tennessee, Chattanooga, USA. ✉email: haobo-guo03@utc.edu; hong-qin@utc.edu

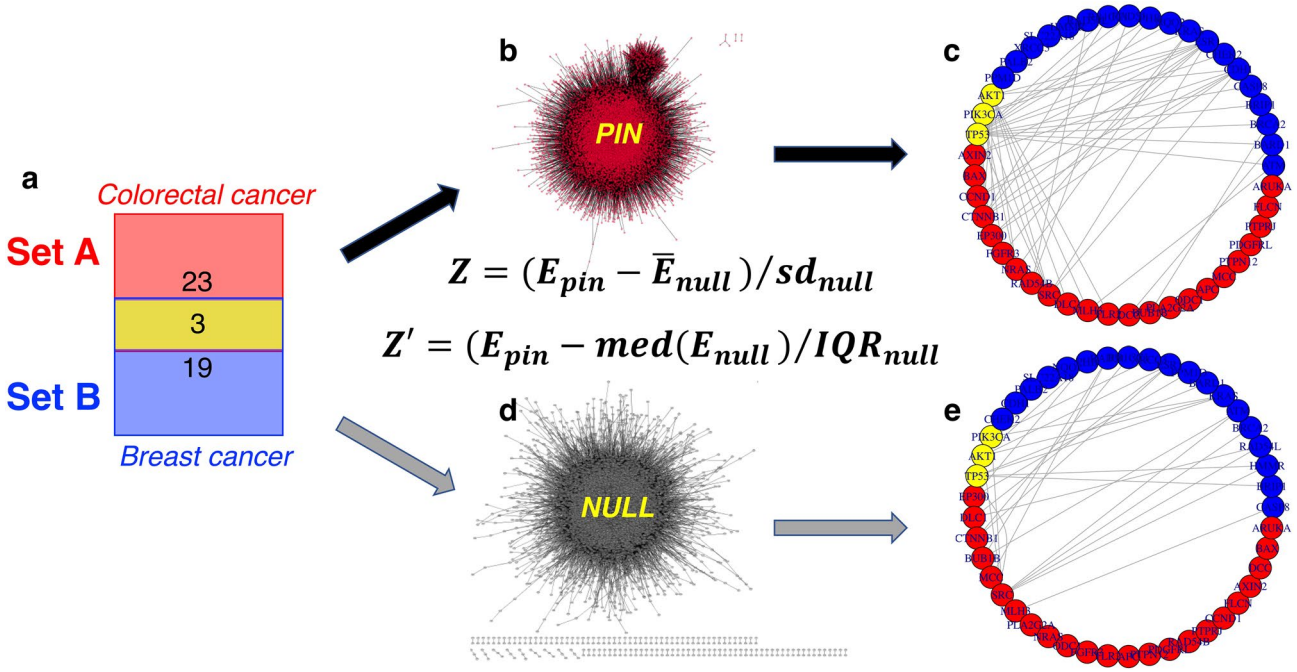

Figure 1. An example of calculating association Z and Z′ scores of two gene sets. The two gene sets are selected from OMIM for colorectal cancer (MIM entry: 114,500, set A) and breast cancer (MIM entry: 114,480, set B). (**a**) Venn diagram shows that Set A has 26 and Set B has 22 genes, respectively. There are three overlapping genes (AKT, PIK3CA, and TP53). (**b**) The human PIN. (**c**) 51 interactions between Set A and Set B observed from the PIN. (**d**) An example of the null network model. (**e**) 23 interactions between Set A and Set B observed from the null model shown in d. The interaction numbers from 10,000 null models are 25.4 ± 4.8 (mean ± sd), leading to a Z-score of 5.3; and the median (25) and IQR (6) values yield a Z′-score of 4.3. Hence there is an enriched association between both cancers.

In addition to the above network-based enrichment tools and databases, another useful and important approach comes from comparisons with random networks[13,30,31]. However, random networks such as the Erdös-Rényi (ER) random networks[32] do not have the power-law distribution of the node degrees observed in PINs and other natural networks[11]. As pointed out by Maslov and Sneppen[13], as well as Newman et al.[31], the meaningful permutations should have the node degrees preserved. Using the random (or null) networks with the preserved degree and/or additional lower-level topological parameters[33], the higher-level attributes of the original network can be abstracted from statistical comparisons. In an outstanding work using the permutation null models, the hub nodes in both PIN and regulatory networks were observed to avoid the interactions with other hub nodes, and this observation was proposed to contribute to the stability of the networks[13]. Inspired by this work we term the random (or null) networks with preserved node degrees as the MS02 null models[34].

In present work, we propose an approach to evaluate gene sets by comparing molecular networks with the MS02 null models, and we term this approach the network-permutation-based association study (NetPAS). We validated the usefulness with 50 well-curated gene sets and established consistency by using Z-scores, modified Z′-scores, p-values and Jaccard-indices. We also estimated an appropriate cutoff of Z-score to infer enriched or suppressed associations. A weighted human disease network was constructed using the NetPAS approach. Moreover, we showed that NetPAS can be applied in gene ontology and pathway enrichment analysis. We propose that the NetPAS is a useful tool to extract biological information stored in gene sets.

Results

Association Z and Z′-score of two gene sets. As illustrated in Fig. 1, we can use the Z-scores to evaluate the over- or under-representation of interactions between two gene sets A and B—where Set A is a group of genes, e.g., genes associated with colorectal cancer, and Set B is another group of genes, e.g., genes associated with breast cancer. The gene IDs for both sets are obtained from OMIM[35]. The two sets share 3 genes. NetPAS first calculates the total number of edges (interactions) between set A and set B that appear in the original network—the human InWeb_IM PIN[8] used in the present work (Fig. 1b). Then by comparing with the numbers of edges from null network models (one example is in Fig. 1d), a Z-score is calculated (see "Methods"). For interactions between both sets, 51 are observed in the PIN (Fig. 1c), compared to 25.4 ± 4.8 observed in 10,000 null network models (one example is in Fig. 1e), yielding an association Z-score of (51–25.4)/4.8 = 5.3. In Fig. 1b, very few isolated interactions can be seen. In contrast, many isolated interactions can be seen in one example of a permuted null network model in Fig. 1d. The contrast suggests that genes with single interaction tend to interact with genes with more connections. Figure 1c illustrates the importance of topological constraints in association tests. Moreover, we tested the modified Z′-scores based on the interquartile range (IQR) from the null models

(see in "Methods"). The two gene sets in Fig. 1 give $Z' = 4.3$, supporting the calculated Z-score for an enriched association between both cancers.

Application of NetPAS in hallmark gene sets. We used 50 hallmark gene sets from the molecular signature database (MSigDB)[36]. These hallmark sets can be considered "refined" benchmarks on top of > 20,000 gene sets in MSigDB (version 7), which respectively represent well-defined biological processes with coherent expressions[37]. The names and details of these hallmark sets are listed in the Table S1 of the Supporting Information (SI). The gene names can also be found in Fig. 2a, the boxplots of Z-score distributions of all hallmark sets. We calculated association Z-scores, one-tailed p-values and Jaccard-indices (see "Methods") between all pairs of gene sets (including self-interactions). Figure 2b shows the heatmap of the association Z-scores calculated from all pairwise associations among the 50 hallmark gene sets using 10,000 MS02 null models compared with the original PPI. In this heatmap, positive Z-score (red) indicate over-representation, whereas negative Z-score (blue) indicates under-representation, respectively.

The Z-score approach (Eq. 2 in "Methods") has an implicit assumption that the interaction numbers from the null models follow the normal distribution. We found that most null distributions can pass the normality test, as shown in Figure S1a in SI. We also compared the modified Z′-scores (Eq. (3) in "Methods"). For the hallmark gene sets, the estimated Z-scores and Z′-scores are highly correlated, and a comparison of the heatmaps derived from both Z- and Z′-scores is presented in Figure S1b.

We directly estimated one-tailed p-values for associations from the PPI using 10,000 MS02 null models. Heatmap of the p-values ($-\log_{10}$ scale) are plotted in Fig. 2c and the p-value distribution is highly correlated with that of the Z-scores with a Pearson's correlation coefficient (PCC) of 0.794 ($P < 2.2 \times 10^{-16}$). A comparison of the heatmaps based on the p-values and q-values ($-\log_{10}$ scale) is shown in Figure S2 of the SI.

We also calculated the Jaccard-indices (see "Methods") between the pairs of gene sets with a heatmap shown in Fig. 2d. A general agreement was also observed between association Z-scores and Jaccard indices with PCC = 0.48 ($P < 2.2 \times 10^{-16}$).

For comparison, we constructed networks to illustrate association patterns among the 50 hallmark sets using the association Z-scores, p-values, and Jaccard-indices, respectively. All networks use the gene sets as nodes and association scores as edge weights. Figure 2e–g show parts of all three networks, respectively. In the Z-score network (Fig. 2e) top 5% over-represented (red) interactions have Z larger than 11.8, and the top 5% under-represented (blue) interactions have Z smaller than − 5.8, respectively. The p-value network (Fig. 2f) shows 326 associations (for all 1,225 pairs of gene sets excluding the self-interactions) with p-value $< 1 \times 10^{-4}$ (i.e., more interactions observed in the PPI than all null models), and in this network a uniform edge-weight is applied for these interactions. For comparisons, the Z-score network in Fig. 2e has 76 positive and 61 negative interactions, the p-value network (Fig. 2f) has 326 interactions, whereas the Jaccard network (Fig. 2g) has only 40 interactions, respectively. Note that all networks would possess more interactions by using looser cutoffs. For instance, a criterion of $|Z| > 2$ for the Z-score network would lead to 509 positive and 254 negative interactions; a cutoff of p-value $< 1 \times 10^{-3}$ results in 427 interactions; and a cutoff of $J > 0$ for the Jaccard network—similar to a previous human disease network[38] in which two diseases are connected if they share at least one gene—would lead to 871 interactions, respectively.

Estimations of p-values are limited by the number of null models used. For 10,000 null models applied in present work, we cannot estimate a p-value smaller than 1×10^{-4}, which is roughly equivalent to $Z = 3.72$ for a one-tail test under normal distributions. Therefore, based on limited number of null models, it is difficult to rank the interaction strengths that have low p-values to a graph, as such, in Fig. 2f p-value $< 1 \times 10^{-4}$ interactions are visualized with a uniform weight. However, the Z-scores (Fig. 2e) spread a considerably wide range using a limited number of null models. In addition, we show that in a Z-score heatmap both enriched (red) and suppressed (blue) associations can be plotted. However, for the one-tail p-value analysis, only one of both associations can be addressed at a time, based on the null hypothesis used—such as enriched associations in Fig. 2c,f—despite both enriched and suppressed associations can be analyzed separately. Similar to using p-values, the Jaccard indices also cannot describe the under-representation information on how gene sets 'avoid' interacting with each other and is only informative on enrichment. Using Z-scores we can identify both enriched and suppressed interactions with relatively small number of null network models. In addition, the standard deviation of the Z-score are similar between using 10,000 and using 1,000 null models (see below in the discussion of random models). In this work, the empirical choice of the number of null models is set to 10,000 because it gives more accurate p-values.

Interestingly, for all the 326 hallmark-hallmark interactions with p-value $< 1 \times 10^{-4}$, their Z-scores are 9.59 ± 6.36 with a minimum of 3.78, which are equivalent to p-value $< 1 \times 10^{-4}$ under a normal distribution one-tail test. Moreover, 11 of these 326 interactions have Jaccard-index of 0: although these 11 pairs show positive interactions (p-value $< 1 \times 10^{-4}$ and $Z = 6.11 \pm 1.97$ with $Z_{min} = 4.24$), no shared genes between each pair could be found. One example is for gene sets 6 and 24 (full names in Table S1 of SI) that have $J = 0$, p-value $< 1 \times 10^{-4}$, and $Z = 10.3$, which reflects a significant over-represented interaction number (344) in the PIN compared to null models (204.5 ± 13.5), as shown in Figure S3 in the SI.

A negative Z-score calculated by NetPAS reflects under-represented interactions between two gene sets. In the box plot of Z-scores between gene sets (self-interactions are excluded, Fig. 2a), some hallmark sets appear to have a negative mean Z-score and appear to have 'avoided' interactions to most of the other hallmark sets. For example, the hallmark set 12 (full name in Table S1 of SI) has a mean Z-score of − 3.6 for interactions with all other hallmark sets. Figure S3 in the SI shows interactions between set 12 and set 25 observed from the PIN and a representative null model.

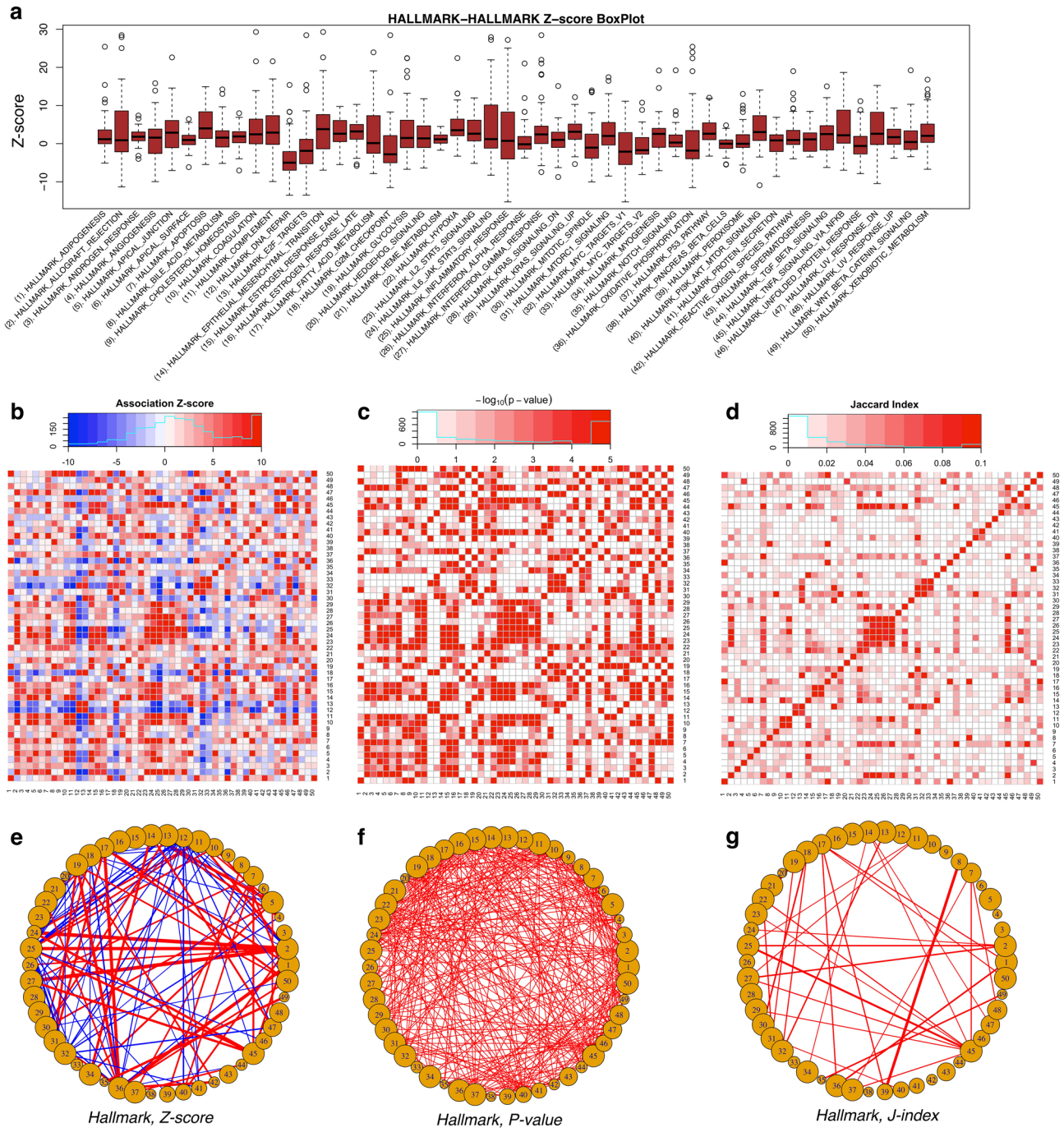

Figure 2. Association studies of 50 hallmark gene sets from MSigDB. (**a**) The boxplots of association Z-scores of the 50 gene sets with others—self-interactions are not considered in this plot. The names of the hallmark gene sets are listed along with their serial numbers. Heatmaps of (**b**) association Z-scores, (**c**) p-values, and (**d**) Jaccard-indices of the 50 gene sets as illustrated by their serial numbers. The names of the gene sets can be found in the Table S1 of the SI. Both enriched (red) and suppressed (blue) interactions can be revealed by Z-score. One-tail p-values (−log10 scale is used in the heatmap) are calculated for enriched interactions in the PIN compared to MS02 null models. When the observed interactions in the PIN are more than that in each of all 10,000 null models, we can only infer that $P < 1 \times 10^{-4}$ instead of $P = 0$. We used P = 1e−5 (or $-\log_{10}P = 5$) for these situations. Networks of the hallmark gene sets have been generated for (**e**) Z-scores, the top 5% enriched (Z > 11.8, red) and top 5% suppressed (Z < − 5.8, blue) interactions are shown in the network; (**f**) p-values, only those of $P < 1 \times 10^{-4}$ are have been shown, and (**g**) Jaccard-indices.

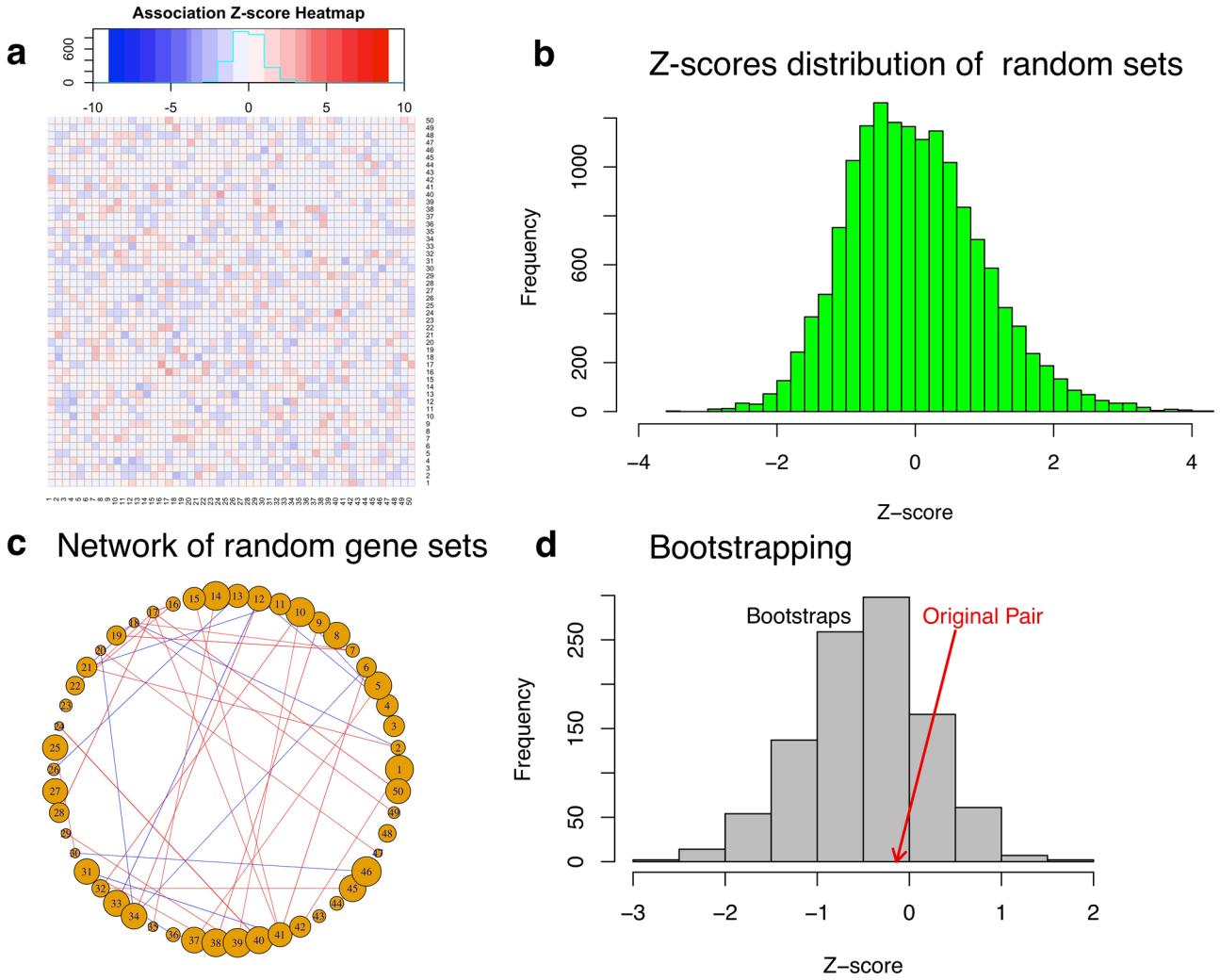

Figure 3. Association Z-scores of random gene sets and statistical validation of cutoffs. (**a**) Heat map of association Z-scores among 50 randomly constructed gene sets with gene numbers in the range of [15,200]. (**b**) Histograms of association Z-scores calculated from five sets of random networks (see Methods and Figure S2 in the SI). Among all 15,000 Z-scores between random gene sets, 451 have $Z > 2$ ($P = 0.030$) and 160 have $Z < -2$ ($P = 0.011$), respectively, indicating that using $|Z| > 2$ as the cutoff would be appropriate. (**c**) Network of the random gene sets weighted by the association Z-scores. A cutoff of $|Z| > 2$ is used for both enhanced (red) and suppressed (blue) interactions. (**d**) We then bootstrapped the affiliations of all genes from the two randomly constructed gene sets, each contains 100 genes. The association Z-score of the original pair is −0.14 (red arrow) whereas the Z-score of the 1,000 bootstraps are -0.47 ± 0.66.

Recommended cutoffs for application in practice based on background Z-scores. To find out recommended Z-score cutoffs for application of NetPAS in practice, we constructed random gene sets with comparable sizes to MSigDB in Fig. 3a (Figure S2 of SI). The association Z-scores among these random sets are narrowly centered around zero (color bar in Fig. 3a). In contrast, association Z-scores of the 50 hallmark sets have a long-tailed distribution with a skewed-peak at the positive upbound (color bar in Fig. 2a). The association Z-scores between the random gene sets (Fig. 3a) are much less and looser than the hallmark gene sets. Moreover, as randomly constructed networks reflect the genetic background, distributions of the Z-scores among these random gene sets can be used to validate the cutoffs for quantifying associations of gene sets. For all 15,000 association Z-scores between random gene sets, 451 have $Z > 2$ and 160 have $Z < -2$, corresponding to p-value = 0.030 and p-value = 0.011, respectively (Fig. 3b). Self-associations are excluded in Fig. 3b although they do not show noticeable differences to non-self-associations for the random sets. For the random gene sets, using $|Z| > 2$ as a cutoff we observed a limited number of enriched (red, 30) and suppressed (blue, 12) associations (Fig. 3c), which are much less than the hallmark gene sets. Similar trends are found in random networks of different sizes (Figure S4 of SI).

The association Z-score between two gene sets—say, set A and B—reflects how likely the genes in set A favor ($Z > 0$) or avoid ($Z < 0$) the interactions with those in set B, and vice versa. Note that for normal distributions $|Z| > 2$ is roughly equivalent to $p < 0.023$ from a one-tailed t-test; however, different cutoff in the Z-scores may

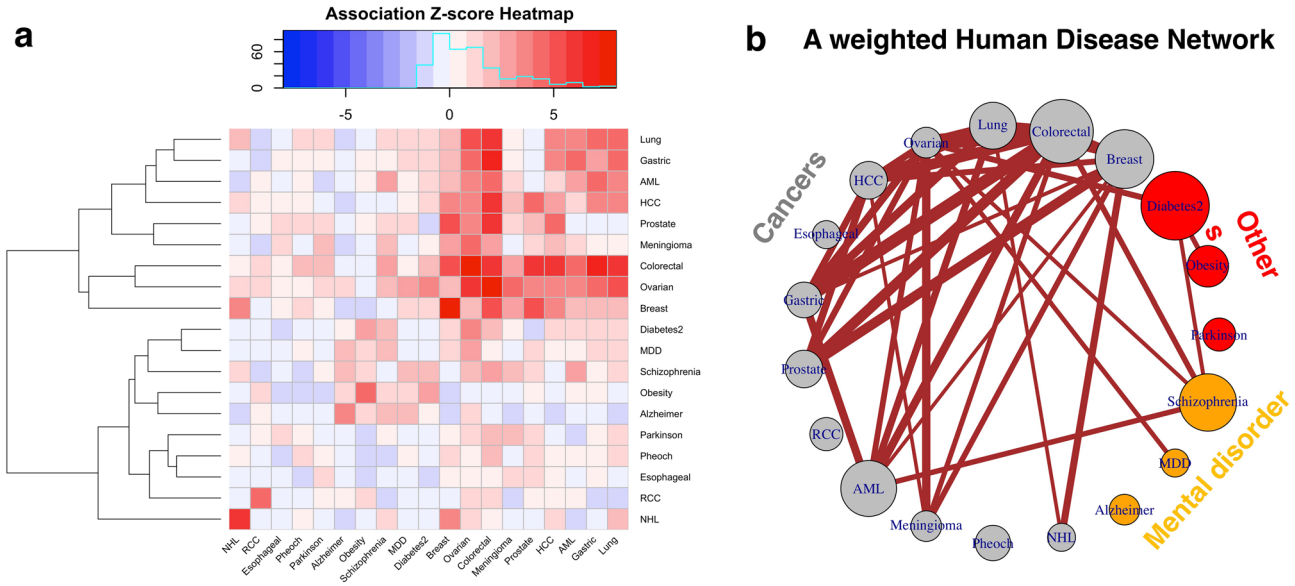

Figure 4. A weighted human disease network (HDN) generated by NetPAS. (**a**) The association Z-score heatmap of 19 diseases, which include 13 cancers (grey), 3 mental disorders (orange) and three other diseases (red). (**b**) A weighted human disorder network constructed from the Z-score matrix indicates that most cancers highly interacted with each other. Some mental disorders including Schizophrenia and Major depression disorder (MDD) are associated with certain cancers, and Type II Diabetes (Diabetes2) is associated with ovarian cancer. Note that in this wHDN all interactions are positive, and no negative or suppressed interactions among these diseases (i.e., $Z < -2$) have been observed.

lead to different interpretations. Nevertheless, we observed that compared to randomly constructed gene sets a cutoff of $|Z| > 2$ is appropriate for a single enrichment test (Fig. 3b).

To further understand how to interpret association Z-scores, we selected two randomly constructed gene sets, each comprised of 100 genes, that have no apparent association with $Z = -0.14$. In this example the number of the bootstrap combinations is $\binom{200}{100} = 910^{58}$. We did not sample all bootstraps. Instead, using 1,000 bootstraps, the Z-scores are distributed in -0.47 ± 0.66, as shown in Fig. 3d. The ratio of $|Z| > 2$ is 0.016 for all bootstraps (i.e., $p = 0.016$ for a two-tail test). Therefore, we suggest that in practice, a cutoff of $|Z| > 2$ is appropriate, in line with the discussions of the random constructed gene sets as shown in Fig. 3b.

The above analysis indicates that the background association of gene sets has relatively small Z-scores and it is an appropriate practice to use a cutoff such as $|Z| > 2$ to infer an association between two gene sets. For multiple comparisons, we would recommend the use of false-discovery rates to control multiple statistical tests.

Constructing a weighted human disease network. A previous work[38] analyzed more than a thousand of human disorders with associated genes maintained by OMIM[35]. This work produced the "human disease network" (HDN), assuming that two disorders are connected if they share at least one gene, i.e., the Jaccard-index > 0. It was shown that the genes associated with the same disorder have a tenfold increase of likelihood to interact with each other than those that are not associated[38].

Here, we use NetPAS to estimate the association Z-scores of 19 descriptive entries from OMIM. These entries are associated with different disorders and contain at least 5 associated genes for each entry. These entries include 13 cancers, 3 mental disorders and 3 other disorders (Table S2 of the SI). Although there is no association between certain diseases, such as Alzheimer's and colorectal cancer shown in Fig. 3D, the associations between some diseases are significant. The Z-score heatmap and the resulting weighted human disease network (wHDN) are shown in Fig. 4. This wHDN has several isolated nodes, including esophageal cancer, renal cell carcinoma (RCC, a type of kidney cancer), pheochromocytoma (Pheoch, rare cancer related to the adrenal gland), Alzheimer's and Parkinson's diseases. Each isolated node contains 5–8 genes. However, some nodes with similar sizes are strongly associated to other diseases, such as ovarian cancer (6 genes), non-Hodgkin Lymphoma (NHL, 5 genes) and meningioma (6 genes). Therefore, the strength of associations between gene sets (disorders in this example) is not determined by the number of genes. Instead, the direct interactions between genes associated with the gene sets (disorders), and with comparisons to those observed in null network models, have contributed to determining the association strength of two gene sets.

The wHDN shown in Fig. 4b indicates that 10 out of 13 cancers (except three isolated cancers mentioned above) have strong associations with each other. The mental disorders Schizophrenia and Major Depression Disorder (MDD) are highly associated with certain cancers, whereas Alzheimer's is not. The Type-II Diabetes is also associated with cancer as well as Schizophrenia. Obesity is not directly associated with cancer but is associated with Type-II Diabetes. This result may be useful to the understanding of disease-disease relationships. In

summary, NetPAS is useful to evaluate the associations among diseases based on physical interactions, instead of overrepresentations of genes.

GO and pathway enrichment analyses using NetPAS. A GO term or a pathway functional term can be regarded as a gene set affiliated to this term. Because NetPAS can be used to estimate the association strength between any two gene sets, it is straightforward to extend the NetPAS approach to the GO and pathway enrichment analysis. For a given target gene set, its association Z-scores with all gene sets related to the GO/pathway functional terms can be separately calculated and ranked, from which the enriched or suppressed functional terms can be inferred. To demonstrate this utility, we performed the GO[16] term and KEGG[17] pathway enrichment analysis of the 50 hallmark gene sets (see above), and compared the results with those obtained by a traditional enrichment method DAVID[39]. In this analysis, the association Z-scores are calculated between the target gene set and the 18,033 gene sets derived from 17,715 GO terms and 318 KEGG pathways (see "Methods"). All GO terms and KEGG pathways are then ranked to infer both enriched and suppressed functional terms.

The top 10 enriched terms by both NetPAS and DAVID for one example, *HALLMARK_OXIDATIVE_PHOSPHORYLATION*, are shown in Fig. 5a. Consistency between the two methods can be seen: 9 out of 10 BP terms, 10 out of 10 CC terms, 8 out of 10 MF terms, and 8 out of 10 KEGG terms predicted by NetPAS are also predicted by DAVID. However, some functional terms detected by NetPAS are missed by DAVID and other enrichment tools, such as the BP term GO:0015990 ("electron transport coupled proton transport"). In this example, the target hallmark gene set has 94 interactions with genes that carry the term GO:0015990, observed in the PIN. In contrast, there are only 4.7 ± 2.1 interactions from the 10,000 null models, leading to a large Z-score of 42.9 for this GO term. For all 50 hallmark sets and the top-10 enriched GO terms by NetPAS, 73.4% BP, 70.2% CC and 55.0% MF terms were verified by DAVID. For all functional terms suggested by NetPAS but missed by DAVID, the enrichment signals come from the fact that more interactions between the target set and the function annotation term have been observed in the PIN than random null models. Figure 5b shows the subnetworks for interactions between the hallmark set exemplified in Fig. 5a and the gene sets affiliated with the top-10 BP terms by NetPAS.

As a network-permutation-based approach, NetPAS is sensitive to the subnetwork configuration within the gene sets, including its global cluster coefficient, maximal cluster size, and maximal clique degree, summarized in Table S3 of SI. Consequently, NetPAS can yield substantial differences when the subnetwork under study is weakly connected. To illustrate the sensitivity of NetPAS to subnetwork topology, four synthetic gene sets *SynGS-1a*, *SynGS-1b*, *SynGS-2a* and *SynGS-2b* have been constructed based on the hallmark sets 20 and 28 (Figure S5 in the SI). Both *SynGS-1a* and *SynGS-2a* contain genes that are highly connected, whereas *SynGS-1b* and *SynGS-2b* contain genes that do not interact with other genes in these gene sets (Figure S5). For both *SynGS-1a* (Fig. S5c) and *SynGS-2a* (Fig. S5g), NetPAS showed more enriched BP terms than the original hallmark sets 20 (Fig. S5b) and 28 (Fig. S5f), respectively. In contrast, the number of enriched terms obtained by DAVID decreased. These contrasting changes support that NetPAS is more sensitive to the highly connected cliques than DAVID, and also suggest that DAVID is more sensitive to the sizes of input gene sets than NetPAS. Less shared GO terms are found for both *SynGS-1b* and *SynGS-2b*, which may be attributed to that these two synthetic sets contain less-connected nodes than *SynGS-1a* and *SynGS-2a*. Interestingly, for *SynGS-2a*, both network-based tools NetPAS and WebGestalt uncovered more enriched terms (668:504 for NetPAS and 23:7 for WebGestalt, respectively) and there were more shared terms (10:1) among all three methods. For the hallmark set 28, NetPAS, DAVID and WebGestalt showed one shared BP term GO:0008015 ("blood circulation"). For *SynGS-2a*, this BP term was not found by either DAVID or WebGestalt. However, it was scored $Z = 14.049$ by NetPAS. The difference between NetPAS and WebGestalt can be illustrated by SynGS-1b and 2b (Fig. S5d and S5h). Because NetPAS looks for interactions and WebGestalt looks for traversal paths, NetPAS can detected enriched BP terms through associated genes, but WebGestalt gave zero enriched terms.

Overall, these results show that NetPAS can serve as a useful complementary tool to DAVID and WebGestalt.

Discussion

No gene or protein functions alone[40]. The cellular functions can be regarded as being conducted by functional modules or communities[40] of genes/proteins in the interactome[25]. The concept of disease module[9,41] has also been proposed based on the fact that the genes associated with the same disease are more likely to interact with each other (the "local" and "disease module" hypotheses in[9]). This principle can also be applied to other curated gene sets such as those in MSigDB[36,42]. Indeed, for the 50 hallmark sets (Fig. 2), the mean association Z-score excluding self-interactions (Fig. 2d) is 2.0. However, the self-interactions for all gene sets have a significantly higher mean Z-score of 17.8. This trend holds for the 19 diseases shown in Fig. 5. For the random sets shown in Fig. 3, however, the mean Z-score is -0.04 and the mean self-association Z-score is -0.20. Therefore, our results indicate over-presented interactions for genes in the curated data sets, such as those from MSigDB or related to diseases, in contradict with random chances.

A biological network such as PIN is scale-free with the degrees of all nodes following the power law. Because in the null models of present work all node degrees have been preserved, they have the same power-law distribution as those in the original PIN. In biological networks, low-degree nodes tend to connect to high-degree nodes, or hubs[13]. For example, there are 1,004 nodes in the PIN with the degree $k = 1$, i.e., each of them only has a single interaction. In the PIN, only two interacting pairs (*CLEC2A*:*KLRF2* and *REC114*:*MEI4*) are formed by such nodes, constituting two isolated interacting pairs. However, for 10,000 null models, there are 113.3 ± 47.6 isolated pairs with a minimum value of 13. Figure 6 in the Methods shows the histograms of the number of isolated pairs in all 10,000 null models, and an example is shown in Fig. 1C.

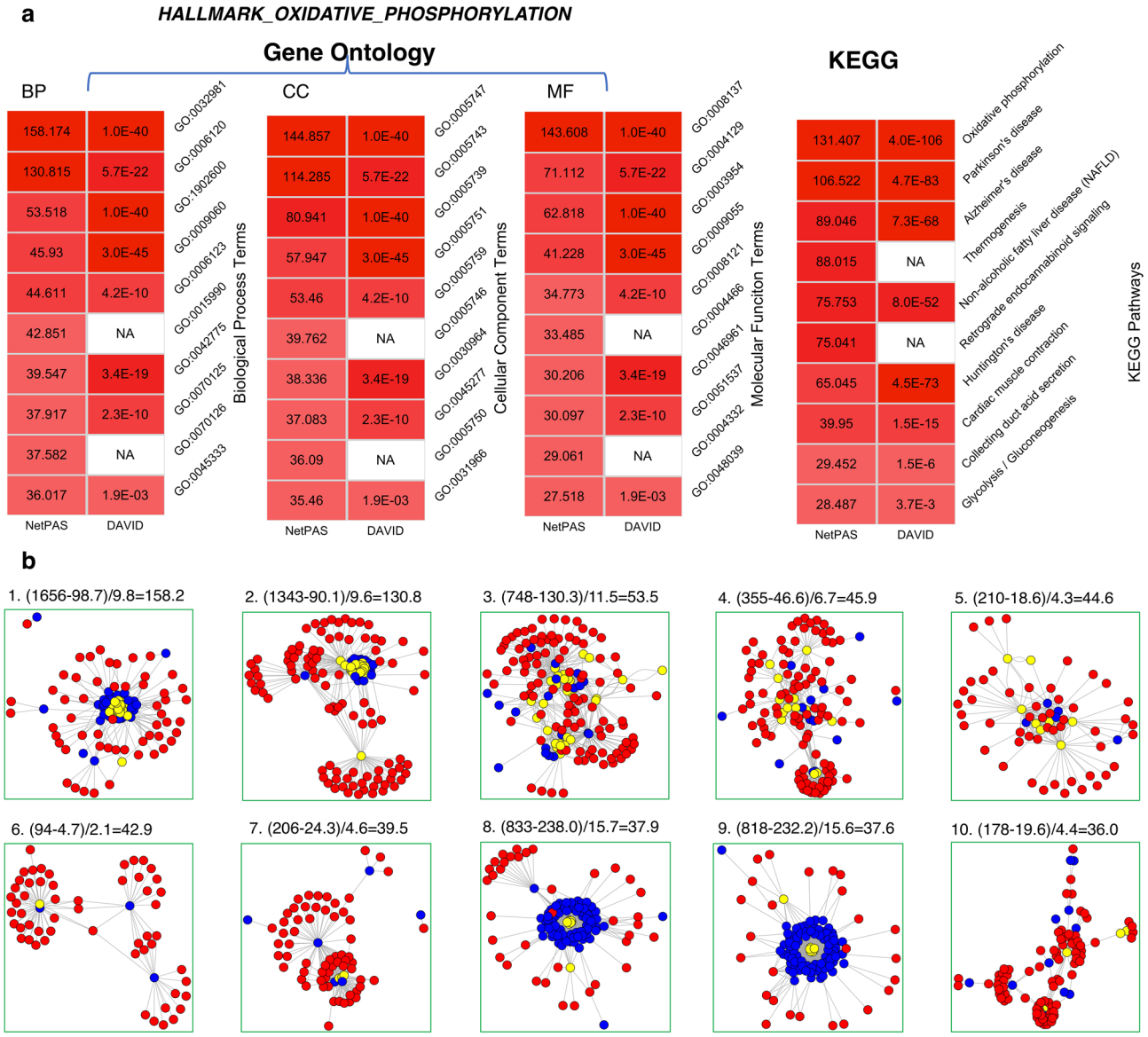

Figure 5. An example of GO and pathway enrichment performed by NetPAS using the hallmark set HALLMARK_OXIDATIVE_PHOSPHORYLATION. (**a**) The Top ten enriched gene ontology terms, including biological process (BP), cellular component (CC) and molecular function (MF) (left), and enriched KEGG pathway terms (right) using NetPAS (left) and DAVID (right). The magnitude of enrichments is scaled by the colors from white ($Z = 0$) to red ($Z = Z_{max}$). The p-values estimated by DAVID were converted to Z-scores using a two-tailed normal distribution for coloring. (**b**) Interaction sub-network between the target gene (red nodes) and the genes affiliated with the top 10 biology process (BP) GO terms (blue nodes); the genes that are both affiliated with the functional term and belong to the target gene set are shown in yellow nodes. Formulas used for calculating the Z-scores for each BP term are written on top of each subnetwork.

A modified Z′-score approach using the interquartile ranges (IQRs) may also be useful to the association study (Figure S1 and Methods). We found that for the well-connected gene sets such as the hallmark sets shown in Fig. 2, both Z- and Z′-scores yielded quantitatively consistent results (Figure S1). We noticed that the Z′-score approach may encounter a numerical challenge, i.e., the IQR may return to zero when few interactions could be observed in the null models, which leads to an infinite Z′-score—even though the standard deviation is non-zero. For example, this kind of numerical errors often happens when the gene set related to a GO term only contains a handful of genes.

Several limitations of NetPAS need to be emphasized, however. The first limitation is the incompleteness of the resources including the interactomes, the coverage of genes in different gene sets, and gene annotations in GO or pathway knowledge databases. In addition, protein–protein interactions are dynamic[43], and may vary significantly among different tissues or cell types[44,45]. These limitations may be addressed in future studies by the integration of tissue-specific or cell-type-specific interactomes to further our understanding of the biological significances of different gene sets.

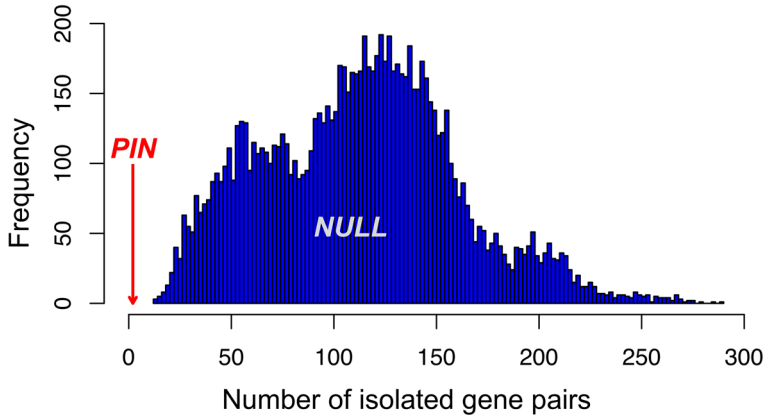

Figure 6. Topological differences between the original PIN and null models exemplified by the number of isolated pairs (interactions between k = 1 nodes) in the PIN with observation (red arrow) of 2 isolated pairs. However, the number of isolated pairs in 10,000 null models is distributed from 13 to 289 with a median value of 115 (blue histograms). For comparison also see Fig. 1b,c. The number of isolated pairs in the null models is significantly larger than that in the PIN ($p < 1 \times 10^{-4}$, for 10,000 null models).

In summary, we show that NetPAS can quantify the association between two different gene sets by taking network constraints into account. We demonstrate the utility of using Z-scores in NetPAS compared to using p-values or Jaccard-scores. NetPAS is useful in classifications of gene sets, including those associated with different diseases. We also show that NetPAS can be applied in GO and pathway enrichment analysis, in which every single GO or pathway functional term is regarded as an affiliated gene set. The NetPAS approach can be applied to extrapolate the biological association between different gene sets such as potential relationships between various gene sets behind different phenotypes and diseases. NetPAS can also be applied in other types of networks to estimate the association strengths between network subsets.

Methods

MS02 null permutation of the PPI network. The permutation-based network null model is based on a work of Maslov and Sneppen in 2002[13] (hence named MS02 null model in present work). The human PIN used in the present work contains 592,685 edges spreading on 16,641 nodes. This PIN is considered as simple graphs, i.e., it is undirected and does not contain self-interactions (self-loops) or multi-interactions.

A network is regarded as a graph $G = (V, E)$ with order of $|V| = N$, the vertices (or nodes) are $V = \{v_1, v_2, \ldots v_N\}$. An MS02 null model, $G^{null} = (V^{null}, E^{null})$, has.

$$V^{null} = V, k(v_i) = k(v_i^{null}), \tag{1}$$

where $k(v_i)$ is the degree or edge numbers—also known as connectivity—associated with v_i, i.e., it uses the same vertex set and degrees for all vertices are preserved as the original network.

It is worth noting that all MS02 null network models follow the power-law because the node degrees are the same as the observed PIN. However, there are significant topological differences between the null models and the PIN. For example, in the PIN there are only two isolated pairs (*CLEC2A*:*KLRF2* and *REC114*:*MEI4*) connected by the nodes with degree k = 1. Here we define the two k = 1 nodes that interact with each other as an "isolated pair", because they are not connected to any other nodes in the network. However, for 10,000 null models, there are 113.3 ± 47.6 isolated pairs with a minimum value of 13 connected by the k = 1 nodes (Fig. 6). The abundance of isolated interactions in MS02 null models indicates that the power-law distribution of node degrees does not originate that the low-degree vertices tend to connect to the hub vertices, as suggested previously[11]. Instead, the low-degree vertices tend to interact with the high-degree vertices may be a unique feature of the natural networks such as the PINs, compared to MS02 null network models. We observed that the number of isolated pairs of the null models may not follow a normal distribution. Nevertheless, we did observe the normality of the interaction numbers between hallmark gene sets or between the random gene sets (see Figure S1a in the SI for an example).

Z-score, modified Z′-score, p-value, q-value and Jaccard indices. The Z-score calculation follows the original analysis based on MS02 models[13]:

$$Z = \left(E_{pin} - \overline{E}_{null}\right)/sd_{null} \tag{2}$$

E_{pin} is the edge number between two gene sets based on the PIN, and $\overline{E}_{null}$ and sd_{null} are the mean and standard deviation of the edge numbers based on the MS02 null models (10,000 models are used in present work). A modified Z′-score is also used:

$$Z' = \left(E_{pin} - \mathrm{med}(E_{null})\right)/IQR_{null}, \quad (3)$$

in which $\mathrm{med}(E_{null})$ is the median edge number and $\mathrm{IQR}_{\mathrm{null}}$ is the interquartile range of the edge numbers from the null models, respectively. When scarce interactions existed between the interested gene sets, the Z′-scores may be useful to exclude errors from potential outliers of the null models. For extensively interacting gene sets such as the MSigDB hallmark sets both Z-scores and Z′-scores are highly consistent (Figure S1 in the SI).

If the interaction numbers between two gene sets A and B is E_{AB}, for the null hypothesis that A and B have more interactions in null models than the PIN, the p-value is $p = n/10{,}000$, where n is the number of null models from which A and B satisfy the null model. The p-value is less than 1×10^{-4} for $n = 0$. The q-value R package[46] had been used to convert the p-values for the hallmark-hallmark interactions to q-values, and a comparison of the heatmaps based on p-values and q-values was shown in Figure S2 in the SI.

The standard dissimilarity measure of Jaccard index J_{AB} is calculated as $J_{AB} = |A \cap B|/|A \cup B|$ between two gene sets A and B. J_{AB} is between 0 (0% overlap) and 1 (100% overlap).

Enrichment analysis by DAVID and WebGestalt. We compared NetPAS enrichment results with those obtained from DAVID[39] and WebGestalt[27]. The medium clustering stringency and other default options in DAVID (version 6.7) was used in the functional classification of gene sets. For WebGestalt, the BIOGRID[47] PIN for *Homo sapiens* was used, the false discovery rates (FDR) were used as the significant scores, and the network expansion method was used for the network constructions. The calculated Z-scores from NetPAS and the p-values estimated by DAVID and WebGestalt have been used for cross comparisons using Venn diagrams (see below).

Venn-diagram. The criteria used are $Z > 5$ for NetPAS, p-value < 0.01 for DAVID and FDR < 0.01 for WebGestalt, and the R package VennDiagram[48] had been employed to plot the Venn diagrams of the shared enrichments by these three methods (Figure S5 in the SI).

Protein–protein interaction network (PIN). Human protein symbols of both the PIN and protein sets (see below) adopt the HUGO gene nomenclature[49]. The InWeb_IM network (v 2016_09_12)[8] was taken for the human PPI network. This network does not allow self-loops (as the MS02 models) and comprises 592,685 edges (downloaded in August 2018). InWeb_IM is one of the largest protein physical interaction networks; e.g., $1.8\times$ of a recent release of the BioGIRD[47] human PIN (v3.5.168, 326,529 physical-interaction edges). Importantly, this network showed excellent performance in representing the gene–gene relationships across hundreds of human pathways[8,50], as well as in assisting the discovery of genes associated with diseases such as cancers[51], which is particularly suitable to the goal of present work.

Gene ontology (GO) annotations and pathways. All GO annotations were downloaded from the Gene Ontology Consortium website (http://www.geneontology.org)[16,52] updated February 2019. The GO terms are grouped in three basic ontologies: biological process (BP), molecular function (MF) and cellular component (CC). The GO annotations for human genome include 11,883 BP terms for 17,697 genes, 4,128 MF terms for 17,498 genes, and 1,704 CC terms for 18,697 genes, respectively. The KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways have been obtained from the ConcensusPathDB[53]. Altogether 318 KEGG pathways for 7,344 genes have been used.

For all 50 hallmark gene sets from MSigDB, we used the conventional enrichment tool DAVID[39] for the GO and pathway enrichment analysis. We also used the network-topology-based tool WebGestalt[27] for the predictions, see Fig. S5 in the SI.

Gene sets. Fifty sets of human proteins from the MSigDB hallmark gene set (category H) collection[36,42] were used to evaluate the interactions between gene sets. 19 gene sets associated with different human disorders were obtained from OMIM[35], including 13 cancers, 3 mental disorders, and 3 other diseases, as listed in the Table S2 of the SI.

We also randomly constructed six gene-set groups, each of which comprises 50 gene sets. Gene sets in groups I, II, III, and IV have the same number of 15, 50, 100 and 200 genes for each set, respectively. For each gene set in group V, the gene number is randomly set in the range of [15,200], and for each gene set in group VI, the gene number is randomly set in the range of [1,100]. Each group has 50 self-associations and $\binom{50}{2} = 1225$ unique non-self-associations. The distribution of Z-scores of these associations are shown in Fig. 3.

All calculations are performed using R. The resources and simulation codes are deposited in the GitHub repository at https://github.com/QinLab/NetPAS.

Received: 20 March 2020; Accepted: 15 June 2020
Published online: 01 July 2020

References

1. Li, S. *et al.* A map of the interactome network of the metazoan C. elegans. *Science* **303**, 540–543. https://doi.org/10.1126/science.1091403 (2004).
2. Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **98**, 4569–4574. https://doi.org/10.1073/pnas.061034498 (2001).

3. Stelzl, U. *et al.* A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957–968. https://doi.org/10.1016/j.cell.2005.08.029 (2005).
4. Rual, J. F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178. https://doi.org/10.1038/nature04209 (2005).
5. Hein, M. Y. *et al.* A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell* **163**, 712–723. https://doi.org/10.1016/j.cell.2015.09.053 (2015).
6. Huttlin, E. L. *et al.* Architecture of the human interactome defines protein communities and disease networks. *Nature* **545**, 505–509. https://doi.org/10.1038/nature22366 (2017).
7. Huttlin, E. L. *et al.* The BioPlex network: a systematic exploration of the human interactome. *Cell* **162**, 425–440. https://doi.org/10.1016/j.cell.2015.06.043 (2015).
8. Li, T. *et al.* A scored human protein–protein interaction network to catalyze genomic interpretation. *Nat. Methods* **14**, 61–64. https://doi.org/10.1038/nmeth.4083 (2017).
9. Barabasi, A. L., Gulbahce, N. & Loscalzo, J. Network medicine: A network-based approach to human disease. *Nat. Rev. Genet.* **12**, 56–68. https://doi.org/10.1038/nrg2918 (2011).
10. Watts, D. J. & Strogatz, S. H. Collective dynamics of "small-world" networks. *Nature* **393**, 440–442. https://doi.org/10.1038/30918 (1998).
11. Barabasi, A. L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
12. Albert, R., Jeong, H. & Barabasi, A. L. Error and attack tolerance of complex networks. *Nature* **406**, 378–382. https://doi.org/10.1038/35019019 (2000).
13. Maslov, S. & Sneppen, K. Specificity and stability in topology of protein networks. *Science* **296**, 910–913 (2002).
14. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550. https://doi.org/10.1073/pnas.0506580102 (2005).
15. Wang, K., Li, M. Y. & Bucan, M. Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Genet.* **81**, 1278–1283. https://doi.org/10.1086/522374 (2007).
16. The Gene Ontology C. The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, 330–338. https://doi.org/10.1093/nar/gky1055 (2019).
17. Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K. & Tanabe, M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res.* **47**, D590–D595. https://doi.org/10.1093/nar/gky962 (2019).
18. Barabasi, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101 (2004).
19. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13. https://doi.org/10.1093/nar/gkn923 (2009).
20. Khatri, P., Sirota, M. & Ten Butte, A. J. years of pathway analysis: Current approaches and outstanding challenges. *PLoS Comput. Biol.* **8**, e1002375. https://doi.org/10.1371/journal.pcbi.1002375 (2012).
21. Rivals, I., Personnaz, L., Taing, L. & Potier, M. C. Enrichment or depletion of a GO category within a class of genes: Which test?. *Bioinformatics* **23**, 401–407. https://doi.org/10.1093/bioinformatics/btl633 (2007).
22. de Leeuw, C. A., Neale, B. M., Heskes, T. & Posthuma, D. The statistical properties of gene-set analysis. *Nat. Rev. Genet.* **17**, 353–364. https://doi.org/10.1038/nrg.2016.29 (2016).
23. Yue, Z. L. *et al.* PAGER 2.0: an update to the pathway, annotated-list and gene-signature electronic repository for Human Network Biology. *Nucleic Acids Res.* **46**, D668–D676. https://doi.org/10.1093/nar/gkx1040 (2018).
24. Yue, Z. L. *et al.* PAGER: constructing PAGs and new PAG-PAG relationships for network biology. *Bioinformatics* **31**, 250–257. https://doi.org/10.1093/bioinformatics/btv265 (2015).
25. Ghadie, M. A., Coulombe-Huntington, J. & Xia, Y. Interactome evolution: Insights from genome-wide analyses of protein–protein interactions. *Curr. Opin. Struct. Biol.* **50**, 42–48. https://doi.org/10.1016/j.sbi.2017.10.012 (2018).
26. Glaab, E., Baudot, A., Krasnogor, N., Schneider, R. & Valencia, A. EnrichNet: Network-based gene set enrichment analysis. *Bioinformatics* **28**, i451–i457. https://doi.org/10.1093/bioinformatics/bts389 (2012).
27. Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z. & Zhang, B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.* **47**, W199–W205. https://doi.org/10.1093/nar/gkz401 (2019).
28. Di Lena, P., Martelli, P. L., Fariselli, P. & Casadio, R. NET-GE: a novel NETwork-based Gene Enrichment for detecting biological processes associated to Mendelian diseases. *BMC Genom.* **16**(Suppl 8), S6. https://doi.org/10.1186/1471-2164-16-S8-S6 (2015).
29. Ulgen, E., Ozisik, O. & Sezerman, O. U. pathfindR: An R Package for Pathway Enrichment Analysis Utilizing Active Subnetworks. https://doi.org/10.1101/272450 (2018).
30. Maslov, S., Sneppen, M. & Zaliznyak, A. Detection of topological patterns in complex networks: correlation profile of the internet. *Phys. A* **333**, 529–540 (2004).
31. Newman, M. E. J., Strogatz, S. H. & Watts, D. J. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E* **64**, 1 (2001).
32. Erdös, P. & Rényi, A. On random graphs I. *Publ. Math. Debrecen* **6**, 18 (1959).
33. Orsini, C. *et al.* Quantifying randomness in real networks. *Nat. Commun.* **6**, 1 (2015).
34. Qin, H., Lu, H. H., Wu, W. B. & Li, W.-H. Evolution of the yeast protein interaction network. *Proc. Natl. Acad. Sci.* **100**, 12820–12824 (2003).
35. Amberger, J. S., Bocchini, C. A., Scott, A. F. & Hamosh, A. OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res.* **47**, D1038–D1043. https://doi.org/10.1093/nar/gky1151 (2019).
36. Liberzon, A. *et al.* The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425. https://doi.org/10.1016/j.cels.2015.12.004 (2015).
37. Agarwal, S., Deane, C. M., Porter, M. A. & Jones, N. S. Revisiting date and party hubs: novel approaches to role assignment in protein interaction networks. *PLoS Comput. Biol.* https://doi.org/10.1371/journal.pcbi.1000817 (2010).
38. Goh, K. I. *et al.* The human disease network. *Proc. Natl. Acad. Sci. USA* **104**, 8685–8690. https://doi.org/10.1073/pnas.0701361104 (2007).
39. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57. https://doi.org/10.1038/nprot.2008.211 (2009).
40. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402**, C47–C52. https://doi.org/10.1038/35011540 (1999).
41. Menche, J. *et al.* Uncovering disease-disease relationships through the incomplete interactome. *Science* **347**, 1257601. https://doi.org/10.1126/science.1257601 (2015).
42. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740. https://doi.org/10.1093/bioinformatics/btr260 (2011).
43. Han, J. D. J. *et al.* Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* **430**, 88–93. https://doi.org/10.1038/nature02555 (2004).
44. Ellis, J. D. *et al.* Tissue-specific alternative splicing remodels protein–protein interaction networks. *Mol Cell* **46**, 884–892. https://doi.org/10.1016/j.molcel.2012.05.037 (2012).
45. Yao, V. *et al.* An integrative tissue-network approach to identify and test human disease genes. *Nat. Biotechnol.* https://doi.org/10.1038/nbt.4246 (2018).

46. Dabney, A., Storey, J. D. & Warnes, G. qvalue: Q-value estimation for false discovery rate control. *R package version* **1** (2010).
47. Oughtred, R. *et al.* The BioGRID interaction database: 2019 update. *Nucleic Acids Res.* **47**, D529–D541. https://doi.org/10.1093/nar/gky1079 (2019).
48. Chen, H. & Boutros, P. C. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinform.* **12**, 35 (2011).
49. Braschi, B. *et al.* Genenamesorg: the HGNC and VGNC resources in 2019. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gky930 (2019).
50. Li, T. B. *et al.* GeNets: a unified web platform for network-based genomic analyses. *Nat. Methods* **15**, 543–546. https://doi.org/10.1038/s41592-018-0039-6 (2018).
51. Horn, H. *et al.* NetSig: network-based discovery from cancer genomes. *Nat. Methods* **15**, 61–66. https://doi.org/10.1038/Nmeth.4514 (2018).
52. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29. https://doi.org/10.1038/75556 (2000).
53. Herwig, R., Hardt, C., Lienhard, M. & Kamburov, A. Analyzing and interpreting genome data at the network level with ConsensusPathDB. *Nat. Protoc.* **11**, 1889–1907. https://doi.org/10.1038/nprot.2016.117 (2016).

Acknowledgements

This work is supported by NSF Career award #1453078 (transferred to #1720215), BD Spoke #1761839, and internal funding of the University of Tennessee at Chattanooga. All simulations are performed using the SimCenter computing resources of the University of Tennessee at Chattanooga.

Author contributions

H.B.G. and H.Q. conceived the study, designed the workflow, performed data analysis, prepared all the figures and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-67875-w.

Correspondence and requests for materials should be addressed to H.-B.G. or H.Q.

Reprints and permissions information is available at www.nature.com/reprints.

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