Network properties of cancer prognostic genes and their gene sets in the human protein interactome

Jifeng Zhang (jifengzhang@fudan.edu.cn)
Huainan Normal University https://orcid.org/0000-0001-5716-0126
Cheng Jiang
Huainan Normal University
Zhicheng Ji
Johns Hopkins University
Chenrun Wang
Huainan Normal University

Research article

Keywords: prognostic genes, prognostic genes sets, network property, protein interactome, cancer, modules

Posted Date: September 24th, 2019

DOI: https://doi.org/10.21203/rs.2.14851/v1

Abstract

Background Identifying prognostic genes (PG) is crucial for estimating survival time and providing pinpoint treatments for patients with cancer. However, prognostic genes sets (PGS) reported in most existing research have low reproducibility and overlap ever between the same cancers or their subtypes. Their common characteristic as well as the molecular mechanism of action is still elusive. Methods Here, we obtained nine prognostic gene sets (including 1,439 prognostic genes) of different types of cancer from 23 high quality literatures, and systemically investigated eight network topological properties for PG and PGS compared with background and four other gene sets (cancer gene set CA, essential gene set ES, housekeeping gene set HK, and metastasis-angiogenesis gene set MA) based on the HPRD and String networks. Results The results showed that PG did not occupy key positions in the human protein interactome network, and were more similar to ES rather than CA. Also, PGS had significantly small intraset distance (IAD) and interset distance (IED) in comparison with random sets. Further, we also found that PGS tended to have be distributed within network modules rather than between modules, the functional intersection of the modules enriched with PGS was closely related to cancer. Conclusions Our research reveals the common properties of cancer PG and PGS in the human protein interactome network, and can help us understand and discover cancer prognostic biomarkers.

Background

Prognostic genes (PG) have many crucial clinical applications, such as accurate predictions of cancer types (or subtypes), stages and their survival time for cancer patients. In particular, precise targeted treatments and surveillance strategies could be implemented when patients have been classified into different risk groups by means of application of PG [1]. In the past 20 years, tremendous efforts have been making to investigate PG, and a large amount of prognostic biomarkers have been identified for different types of cancer [2-10], Several PG have been playing critical roles in the prognosis of certain types of cancer such as ER and HER2 for breast cancer[11].

Biological network provides an invaluable research platform to trace genetic phenomena and disease mechanisms on a system level [12, 13]. Network topology analysis helps to discover groups of nodes with special network characteristics in biological networks, as well as associations between groups (e.g., plant immunity[14] and human disease[15, 16]). Among them, study of cancer genes network has showed that they tend to have higher degree and betweenness compared with essential genes[17, 18]. Systematic studies of the PG's network properties could help identify pan-cancer PG and unveil the mechanisms of cancer. However, previous studies of PG were scattered and mostly focus on one specific cancer type or subtype. Cumulative evidence also showed that even for the same cancer type, PGS obtained by different researchers had very small overlap and questionable reproducibility[19]. Few PG studies were carried out involving multiple cancer types (e.g., [20-22]). Also, they either didn't pay attention to the network properties of PG, or were only involved in the gene co-expression network characteristics of PG in a few cancers. So, it is still not well known to their common properties of the human protein interactome.

In this study, we obtained nine PGS (including 1,439 PG) of different cancer types when we manually selected 23 high quality literatures that involve the identification of prognostic genes. Based on the two protein interactome networks (HPRD and String) and four other gene sets for comparison (cancer gene set: CA, essential gene set: ES, housekeeping gene set: HK, and metastasis-angiogenesis gene set: MA), we investigated their eight network properties. Our study showed that although PG did not possess a high degree compared with cancer genes, PGS had tighter network connections and closer inter-genesets distances than background, and the network modules they were in had many common functions that were closely related to cancer. These findings could bring insight into identification of prognostic biomarkers and understandings of their mechanisms.

Results

Overview of prognostic genes. To systematic study of prognostic genes, we obtained 25 small prognostic gene sets (ranging from 3 to 330) from 23 high quality literatures (Table 1). Similar to previous study [19], these genes had very small overlap and network connections. Only 14 genes were repeatedly mentioned 3 times in these small gene sets (see Supplementary Table 1 for details). Taking into account the number of gene sets and cancer types, we combined the gene sets with the smaller number of genes and finally got nine large prognostic gene sets (PGS), which consisted of 1,439 prognostic genes (PG) after normalizing gene names and removing duplications. For comparison, we also selected four other gene sets: cancer gene set (CA), essential gene set (ES), housekeeping gene set (HK), and metastasis-angiogenesis gene set (MA) (Supplementary Table 2). To investigate their network properties, we employed two protein-protein interaction (PPI) networks, HPRD and String, which both exhibited power-law node-degree distributions (Figure 1A and B)[23]. Figure 1C shows that cancer prognostic genes are discretely distributed in the HPRD network. Only three out of the 14 genes which appeared three times above had directly connected edges in the HPRD network.

Four network centralities of prognostic genes. For prognostic genes, we first investigated the four network centralities: Degree, Betweenness, Closeness, and Eigenvector. They are used to measure the importance of a node in a given network from different perspectives. Larger values of the four centralities indicate more importance in the network [12]. Based on the HPRD and String networks, we calculated the four centralities for all 1,439 prognostic genes, background (mean of all nodes in the network), and four other gene sets. The results were shown in Figure 2. Like ES, degree and betweenness of PG was lower than the background, while CA and MA were obviously higher than the background in two PPI networks in Figure 2A-D. However, in Figure 2E-H, Closeness of PG and four other gene sets were considerably higher than the background, while Eigenvector of PG was different from CA and MA, and its values were always lower than the background in the HPRD and String networks. Eigenvector of CA and MA, as well as degree and betweenness of HK, showed inconsistency in the both networks, probably due to the String network consists of more notes and edges[24].

Overall, the results clearly showed that: (1) CA had very similar centralities to MA and both gene sets had significantly higher centralities than other three gene sets including PG (except eigenvector in the String

network, FDR-adjusted p-values of t tests were much smaller than 0.001 in all other cases). This illustrated that PG was significantly different from cancer-related genes in terms of four network centralities; (2) Except Closeness, the other three centralities of PG were below the average of the whole network. This showed that, unlike cancer genes, prognostic genes did not occupy key positions in the network [18, 25]; (3) The four centralities of PG were not significantly different from those of ES (FDR-adjusted p-values of t test were greater than 0.1 for all four centralities).

Four network measures of prognostic gene sets. Most of cancer prognostic biomarkers often act as functional units in a gene set. Therefore, it was necessary to examine the network topological properties of gene sets. To this end, we first calculated clustering coefficient (CC) for nine PGS, four other gene sets and random gene sets. CC measures the tendency that the nodes in a graph cluster together. Larger CC values indicate that the nodes are more likely to form clusters in a network[26]. Figure 3A and B shows their distributions of CC. In the HPRD network, nine PGS had slightly larger CC than the random gene sets (p-value of KS test was not significant). CA and MA also had larger CC than HK and ES. For the String network with higher density, on the one hand, nine PGS had significantly smaller CC than the random gene sets (KS test, p-value < 0.05), which showed that genes within the nine PGS were more sparsely connected compared to random gene set in the network. On the other hand, HK had significantly larger CC than all the other gene sets (p-value of permutation test smaller than 0.001). This was probably due to the fact that edges were more likely to be formed between HK in the String network since HK had consistent expression patterns [27]. And it can also be clearly demonstrated by comparing the degree of HK in the two networks (Figure 2A and B).

Through the investigation of CC, we failed to obtain the significant common properties of the PGS in the network. So, we proposed three other measures, intraset distance (IAD), interset distance (IED) and genset-distribution in modules (GDM), to examine the network properties of gene sets in the network. IAD and IED were used to portray the network distance within a gene set and between two gene sets, respectively (see the methods section for more details). Their calculations were based on shortest path (SP), which can reflect the ability of network information transfer[28]. Figure 3C and D shows that IAD of nine PGS are significantly smaller than the random gene sets in both networks, indicating that there is a more compact network structure within PGS. In the four other gene sets, CA and MA had obviously smaller IAD than PGS compared to HK and ES, and considering two networks together, the ES was not the closest one to PGS in the IAD distribution.

Similarly, we found that IED between PG themselves were significantly smaller than those between PGS and the random gene sets (Figure 3E and F). The result indicates that the nine PGS are not spatially loose, but rather closely connected. Simultaneously, we also found that IED between PGS and the four other gene sets were significantly smaller than those of PGS themselves. One possible reason was that they were derived from different cancer types. Among them, we found that IED between PG and CA or MA was smaller than the other two gene sets, which may suggest that PGS is more closely related to cancer (Figure S1A and B). In addition, we can easily see that whether it was IAD and IED, the distances in the String network were smaller compared to the HPRD network.

Next, we used genset-distribution in modules (GDM) to investigate the distribution of PGS within and between modules of network. Figure 3G and H shows that GDM of nine PGS are significantly larger than random in both networks, demonstrating that they are more likely to be distributed within the module. We also found that GDM of MA were the largest of the four other gene sets, and a change in the relative position of CA and HK in the two networks. This may be due to the different modules that were derived from different networks, and the complexity of gene sets themselves.

Functional analysis of prognostic gene sets. We performed functional enrichment analysis for nine PGS based on GO terms and KEGG pathways database using Fisher test. However, more than half of the gene sets were not enriched with any significant functional terms. Genes with same or similar functions are more inclined to be in the same module of a network [29]. We then examined functions of network modules with two or more PGS. Interestingly, when we compared these functions of the modules from different networks, we found that the intersections of their functions were mostly related to cancer. Figure 4 shows the intersection of the function of module #4 of the String network and module #5 and #7 of the HPRD network. Most of functional terms could be attributed to hallmarks of cancer [30]. They included “Extracellular matrix organization”, “Leukocyte migration”, “Collagen metabolic process”, “Transforming growth factor beta receptor signaling pathway”, etc. In particular, among them, “extracellular matrix organization”was the most significant GO term. Researchers have found that its remodeling directly affects tumor growth, development, and progression[31]. And “transforming growth factor beta” (TGF-β), the main pathway of another functional terms, have been evaluated as prognostic or predictive markers for cancer patients [32](the lower half of Figure 4).

Discussion

A few nodes, the hubs, have a higher connectivity coexistence with most rarely connected nodes in a scale-free network, and they are highly influential in keeping the whole network together [33]. Hub genes were found in human cancer genes and essential genes of yeast and worms in PPI networks [17, 23]. In the study of network centralities, we also found that the degree of CA and their other three centralities were significantly larger than PG and human ES. Among them, PG and human ES were very similar (their distribution difference was not significant by t-test) in most cases, and were smaller than the background mean state, although ES was reported to have significantly more important network topology than “unnecessary genes”[34]. However, low-degree features did not affect functional genes to play an important role. For example, the metabolites with low degree were involved in essential reactions in the metabolic networks of *Escherichia coli* [35]. The importance of PG to cancer patients could be comparable to the importance of ES to the healthy group [36]. In contrast, we also found that PGS have smaller IED to MA compared to ES in the study of gene sets. This indicates that prognostic genes could be more related to MA instead of ES in terms of the causal relation from a pathology perspective [37, 38].

Although the number of genes in prognostic biomarkers had a decreasing trend as a whole [39, 40], most of the previous prognostic biomarkers were often in the form of a union set of dozens or hundreds of genes, or even a network module [6, 41-43]. This would considerably weaken the importance of individual

genes, which may be one of the reasons why the four network centralities of the single PG were not high. However, by focusing on gene sets, our study helped to make up for this deficiency. Three network properties of gene sets presented in this paper had obtained results that were significantly different from the random background. They were more conducive to further understanding of prognostic biomarkers and their mechanisms of action. Interestingly, Yang et. al. [22] also found that prognostic genes did not occupy hub positions and were more likely to appear within network modules when studying the topological properties of prognostic genes in co-expression networks based on four types of cancer. Despite using the different measurement method, Zhang and Horvath [44] also drew similar conclusions that prognostic genes for cancer survival was highly correlated with their intramodular connectivity. The network structure of PPI may be less susceptible to environmental conditions than gene co-expression network [16]. This implies that the topological properties of prognostic genes in protein interactome networks can also be shared in gene co-expression network.

Discovery and clinical applications of prognostic genes have been going on for several decades, but the secrets of prognostic genes are yet to be unveiled. Non-overlapping and non-reproducibility of research findings on prognostic genes can be due to many factors, such as types of cancer, microenvironment, patient cohorts, methodologies and technological platforms [20]. Although there are enormous difficulties to the systemic study of PG, further studies will be still essential in view of the indispensable roles of prognostic genes in cancer diagnosis and treatments as well as the exploration of cancer mechanisms. The above unfavorable situation may be improved by studies of emerging new molecular markers, such as, prognostic miRNAs, prognostic lncRNAs, prognostic circRNAs, and their combinations [45-47].

Conclusions

In summary, we systematically studied the network properties of PG and their PGS for the first time based on two protein interactome networks and eight network properties. We found that prognostic genes had significantly different network properties from CA and were similar to ES in the four network centralities. For intra-module and inter-module distances, PGS was significant smaller relative to random gene sets. And they were more easily enriched inside the modules, which were found to enrich the functions related to cancer development and progression. These characteristic will be valuable for future understanding and identification of prognostic genes. However, several disadvantages still existed in this study, including not considering emerging markers, lacking optimized combinations of gene sets, fewer gene sets and cancer types, etc. Including more datasets and developing new computational strategies could lead to more significant results in the future research.

Methods

Prognostic genes and other four gene sets. To obtain reliable prognostic genes from vast amounts of existing literature, we set two criteria for a publication to be included in our study: (1) the sample size of patients with cancer in the study was larger than 100; (2) the publication was published on a high-impact biomedical journal (impact factor > 8.0) after year 2000 and the research findings are explicit and

accessible. Given these criteria, we manually selected 23 publications and each publication reported 3 to 330 prognostic genes. For convenience we also merged these genes into nine gene sets according to cancer types and the number of prognostic genes reported in each study. Each gene set consisted of 100 to 200 prognostic genes. Details of the selected publications and the corresponding prognostic genes can be found in Table 1 and Supplementary Table2.

To facilitate the comparison of network properties, we selected four other gene sets: cancer gene set (CA), essential gene set (ES), housekeeping gene set (HK) and metastasis-angiogenesis gene set (MA). Each gene set contained around 120 genes and the source of the four gene sets can be found in Supplementary Table 2. The gene names in this study were all converted to official gene symbols using HGNC database[48].

Biology networks and network modules. Two protein interactome networks were used in this study. The first network was constructed using the Human Protein Reference Database (HPRD V9.0)[49]. It consisted of 9,402 nodes and 36,746 edges after removing redundancy. The second network were constructed using human STRING Database (String v10)[24]. It consisted of 14,733 nodes and 334,463 edges after removing edges with scores less than 0.6. Network structures were visualized using Cytoscape[50]. Network modules were identified using Multi-Step Greedy (MSG) algorithm [51] and modules with at least 30 genes were reserved for further analysis.

Calculation of topological measures. Network centralities such as Degree, Betweenness, Closeness and Eigenvector were defined in previous literature [52]. The igraph package in R (http://igraph.org/r/) was used to calculate the four measures. Clustering coefficient (CC) and shortest path (SP) were calculated using previous definitions [12]. Specifically, if two nodes were not connected in the network, their SP was set to be the maximum of SP between all connected nodes in the network.

We proposed two measures to further quantify the network properties of gene sets. First, we used SP to define the distance between two nodes in the network. Then, we defined intraset distance (IAD) and interset distance (IED) based on the distance. IAD was derived from the definition of the average shortest path of complex network [53]. These two measures were used to quantify the distance (or compactness) within a gene set and the distance between two gene sets, respectively. For gene set S with N genes, the IAD was defined as follows:

$$IAD = \frac{1}{N(N-1)} \sum_{i,j \in N, i \neq j} d_{ij} \quad (1)$$

Here, d_{ij} is the shortest path between gene *i* and *j*. *IAD* is the average distance between all pairs of genes in gene set *S*.

For gene set S_p and S_q, *IED* was defined as follows:

$$D_{iS_q} = \underset{i \in N_{S_p}, j \in N_{S_q}}{\text{mean}} (d_{ij, j=1,\ldots,N.}) \quad (2)$$

$$IED = \frac{1}{N_{S_p}} \sum_{i \in N_{S_p}} D_{iS_q} + \frac{1}{N_{S_q}} \sum_{j \in N_{S_q}} D_{jS_p} \quad (3)$$

Here, D_{iS_q} is the distance between gene *i* in S_p and gene set S_q, which is defined as the average distance between gene *i* and all genes in S_q. *IED* is the average distance of all genes in one gene set to the other gene set.

In addition, we also proposed genset-distribution in modules (GDM) to investigate the distribution of a gene set in network modules. For a given gene set, *GDM* can be expressed as a proportion of edges (links between genes) within modules from all possible edges in the module or between modules. It was defined using the following formula:

$$GDM = E_{\text{intra}} / (E_{\text{intra}} + E_{\text{inter}}) \quad (4)$$

Here, E_{intra} is the total number of edges within network modules and E_{inter} is the total number of edges between network modules. Figure 5 demonstrates how *GDM* is calculated.

Functional enrichment analysis. Biological process of gene ontology (GO) and KEGG pathway enrichment analysis were performed by Fisher test. We retained only GO annotations with 30-300 genes and excluded annotations that were electronically inferred (IEA) for GO analysis. For each gene set, the background was all genes which appear in their corresponding network. Only annotations with FDR-adjusted p-values < 0.05 were considered.

Abbreviations

CA: cancer gene set; CC: Clustering coefficient; ES: essential gene set; GO: gene ontology; GDM: genset-distribution in modules; HK: housekeeping gene set; IED: interset distances; IAD: intraset distance; MA: metastasis-angiogenesis gene set;

PG: prognostic genes; PGS: prognostic genes sets; PPI: protein-protein interaction; SP: shortest path.

Declarations

Authors' contributions

This study was conceived by J.Z. and the project was led by J.Z. Collection of references and gene sets were done by C.J., Z.J., and C.W. The network topology analysis and computations were performed by J.Z., Z.J., C.J., and C.W. The manuscript was written by J.Z. and Z.J.

Acknowledgments

This work was supported by Anhui Science and Technology Major Project (no.18030701189); Huainan science and technology project (2017A0421); the Key Support Program for Outstanding Young Talents in University of Anhui Province (no. gxyqZD2016264); the Research Projects of Huainan Normal University (no. 2017hsyxkc91, 2015xj49zd, 2015hssjjd05, and 2015hsyxkc22); the Research Projects of Quality Engineering and Teaching Reform in University of Anhui Province (no. 2018mooc145, 2017sjjd026, 2015ckjh036, and 2015zdjy133); the Provincial Natural Science Research Project of Anhui Colleges (no. KJ2013B259).

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Yuan Y, Van Allen EM, Omberg L, Wagle N, Amin-Mansour A, Sokolov A, Byers LA, Xu Y, Hess KR, Diao L *et al*: **Assessing the clinical utility of cancer genomic and proteomic data across tumor types**. *Nature biotechnology* 2014, **32**(7):644-652.
2. Li J, Lenferink AE, Deng Y, Collins C, Cui Q, Purisima EO, O'Connor-McCourt MD, Wang E: **Identification of high-quality cancer prognostic markers and metastasis network modules**. *Nature communications* 2010, **1**:34.
3. Cancer Genome Atlas Research N: **Integrated genomic analyses of ovarian carcinoma**. *Nature* 2011, **474**(7353):609-615.
4. Yeoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahfouz R, Behm FG, Raimondi SC, Relling MV, Patel A *et al*: **Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling**. *Cancer cell* 2002, **1**(2):133-143.

5. Bullinger L, Dohner K, Bair E, Frohling S, Schlenk RF, Tibshirani R, Dohner H, Pollack JR: **Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia**. *The New England journal of medicine* 2004, **350**(16):1605-1616.
6. Zhao H, Ljungberg B, Grankvist K, Rasmuson T, Tibshirani R, Brooks JD: **Gene expression profiling predicts survival in conventional renal cell carcinoma**. *PLoS medicine* 2006, **3**(1):e13.
7. Lau SK, Boutros PC, Pintilie M, Blackhall FH, Zhu CQ, Strumpf D, Johnston MR, Darling G, Keshavjee S, Waddell TK *et al*: **Three-gene prognostic classifier for early-stage non small-cell lung cancer**. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2007, **25**(35):5562-5569.
8. Boutros PC, Lau SK, Pintilie M, Liu N, Shepherd FA, Der SD, Tsao MS, Penn LZ, Jurisica I: **Prognostic gene signatures for non-small-cell lung cancer**. *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106**(8):2824-2828.
9. Sveen A, Agesen TH, Nesbakken A, Meling GI, Rognum TO, Liestol K, Skotheim RI, Lothe RA: **ColoGuidePro: a prognostic 7-gene expression signature for stage III colorectal cancer patients**. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2012, **18**(21):6001-6010.
10. Gerami P, Cook RW, Wilkinson J, Russell MC, Dhillon N, Amaria RN, Gonzalez R, Lyle S, Johnson CE, Oelschlager KM *et al*: **Development of a prognostic genetic signature to predict the metastatic risk associated with cutaneous melanoma**. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2015, **21**(1):175-183.
11. Weigel MT, Dowsett M: **Current and emerging biomarkers in breast cancer: prognosis and prediction**. *Endocrine-related cancer* 2010, **17**(4):R245-262.
12. Furlong LI: **Human diseases through the lens of network biology**. *Trends in genetics : TIG* 2013, **29**(3):150-159.
13. Barabasi AL, Oltvai ZN: **Network biology: understanding the cell's functional organization**. *Nature reviews Genetics* 2004, **5**(2):101-113.
14. Ahmed H, Howton TC, Sun Y, Weinberger N, Belkhadir Y, Mukhtar MS: **Network biology discovers pathogen contact points in host protein-protein interactomes**. *Nature communications* 2018, **9**(1):2312.
15. Leiserson MD, Vandin F, Wu HT, Dobson JR, Eldridge JV, Thomas JL, Papoutsaki A, Kim Y, Niu B, McLellan M *et al*: **Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes**. *Nature genetics* 2015, **47**(2):106-114.
16. Barabasi AL, Gulbahce N, Loscalzo J: **Network medicine: a network-based approach to human disease**. *Nature reviews Genetics* 2011, **12**(1):56-68.
17. Zhu K, Liu Q, Zhou Y, Tao C, Zhao Z, Sun J, Xu H: **Oncogenes and tumor suppressor genes: comparative genomics and network perspectives**. *BMC genomics* 2015, **16 Suppl 7**:S8.
18. Sun J, Zhao Z: **A comparative study of cancer proteins in the human protein-protein interaction network**. *BMC genomics* 2010, **11 Suppl 3**:S5.

19. Ein-Dor L, Zuk O, Domany E: **Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer**. *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**(15):5923-5928.
20. Gentles AJ, Newman AM, Liu CL, Bratman SV, Feng W, Kim D, Nair VS, Xu Y, Khuong A, Hoang CD *et al*: **The prognostic landscape of genes and infiltrating immune cells across human cancers**. *Nature medicine* 2015, **21**(8):938-945.
21. Martinez-Ledesma E, Verhaak RG, Trevino V: **Identification of a multi-cancer gene expression biomarker for cancer clinical outcomes using a network-based algorithm**. *Scientific reports* 2015, **5**:11966.
22. Yang Y, Han L, Yuan Y, Li J, Hei N, Liang H: **Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types**. *Nature communications* 2014, **5**:3231.
23. Jeong H, Mason SP, Barabasi AL, Oltvai ZN: **Lethality and centrality in protein networks**. *Nature* 2001, **411**(6833):41-42.
24. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP *et al*: **STRING v10: protein-protein interaction networks, integrated over the tree of life**. *Nucleic acids research* 2015, **43**(Database issue):D447-452.
25. Jonsson PF, Bates PA: **Global topological features of cancer proteins in the human interactome**. *Bioinformatics* 2006, **22**(18):2291-2297.
26. Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang D-U: **Complex networks: Structure and dynamics**. *Physics reports* 2006, **424**(4):175-308.
27. de Jonge HJ, Fehrmann RS, de Bont ES, Hofstra RM, Gerbens F, Kamps WA, de Vries EG, van der Zee AG, te Meerman GJ, ter Elst A: **Evidence based selection of housekeeping genes**. *PloS one* 2007, **2**(9):e898.
28. Latora V, Marchiori M: **Efficient behavior of small-world networks**. *Physical review letters* 2001, **87**(19):198701.
29. Sharan R, Ulitsky I, Shamir R: **Network-based prediction of protein function**. *Molecular systems biology* 2007, **3**:88.
30. Hanahan D, Weinberg RA: **Hallmarks of cancer: the next generation**. *Cell* 2011, **144**(5):646-674.
31. Sangaletti S, Chiodoni C, Tripodo C, Colombo MP: **The good and bad of targeting cancer-associated extracellular matrix**. *Current opinion in pharmacology* 2017, **35**:75-82.
32. Elliott RL, Blobe GC: **Role of transforming growth factor Beta in human cancer**. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2005, **23**(9):2078-2093.
33. Vidal M, Cusick ME, Barabasi AL: **Interactome networks and human disease**. *Cell* 2011, **144**(6):986-998.
34. Hwang YC, Lin CC, Chang JY, Mori H, Juan HF, Huang HC: **Predicting essential genes based on network and sequence analysis**. *Molecular bioSystems* 2009, **5**(12):1672-1678.

35. Samal A, Singh S, Giri V, Krishna S, Raghuram N, Jain S: **Low degree metabolites explain essential reactions and enhance modularity in biological networks**. *BMC bioinformatics* 2006, **7**:118.
36. Liao BY, Zhang J: **Null mutations in human and mouse orthologs frequently result in different phenotypes**. *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105**(19):6987-6992.
37. Zetter BR: **Angiogenesis and tumor metastasis**. *Annual review of medicine* 1998, **49**:407-424.
38. Bergers G, Benjamin LE: **Tumorigenesis and the angiogenic switch**. *Nature reviews Cancer* 2003, **3**(6):401-410.
39. Zuo S, Dai G, Ren X: **Identification of a 6-gene signature predicting prognosis for colorectal cancer**. *Cancer cell international* 2019, **19**:6.
40. Tang J, Kong D, Cui Q, Wang K, Zhang D, Gong Y, Wu G: **Prognostic Genes of Breast Cancer Identified by Gene Co-expression Network Analysis**. *Frontiers in oncology* 2018, **8**:374.
41. Lenz G, Wright G, Dave SS, Xiao W, Powell J, Zhao H, Xu W, Tan B, Goldschmidt N, Iqbal J *et al*: **Stromal gene signatures in large-B-cell lymphomas**. *The New England journal of medicine* 2008, **359**(22):2313-2323.
42. Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, Faria D, Bull S, Pawson T, Morris Q, Wrana JL: **Dynamic modularity in protein interaction networks predicts breast cancer outcome**. *Nature biotechnology* 2009, **27**(2):199-204.
43. Wu G, Stein L: **A network module-based method for identifying cancer prognostic signatures**. *Genome biology* 2012, **13**(12):R112.
44. Zhang B, Horvath S: **A general framework for weighted gene co-expression network analysis**. *Statistical applications in genetics and molecular biology* 2005, **4**:Article17.
45. Calin GA, Croce CM: **MicroRNA signatures in human cancers**. *Nature reviews Cancer* 2006, **6**(11):857-866.
46. Yang Z, Xie L, Han L, Qu X, Yang Y, Zhang Y, He Z, Wang Y, Li J: **Circular RNAs: Regulators of Cancer-Related Signaling Pathways and Potential Diagnostic Biomarkers for Human Cancers**. *Theranostics* 2017, **7**(12):3106-3117.
47. Qi P, Du X: **The long non-coding RNAs, a new cancer diagnostic and therapeutic gold mine**. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc* 2013, **26**(2):155-165.
48. Gray KA, Daugherty LC, Gordon SM, Seal RL, Wright MW, Bruford EA: **Genenames.org: the HGNC resources in 2013**. *Nucleic acids research* 2013, **41**(Database issue):D545-552.
49. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A *et al*: **Human Protein Reference Database–2009 update**. *Nucleic acids research* 2009, **37**(Database issue):D767-772.
50. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T: **Cytoscape 2.8: new features for data integration and network visualization**. *Bioinformatics* 2011, **27**(3):431-432.

51. Schuetz P, Caflisch A: **Multistep greedy algorithm identifies community structure in real-world and computer-generated networks**. *Physical review E, Statistical, nonlinear, and soft matter physics* 2008, **78**(2 Pt 2):026112.
52. Ozgur A, Vu T, Erkan G, Radev DR: **Identifying gene-disease associations using centrality on a literature mined gene-interaction network**. *Bioinformatics* 2008, **24**(13):i277-285.
53. Zhou S, Mondragon RJ: **Accurately modeling the internet topology**. *Physical review E, Statistical, nonlinear, and soft matter physics* 2004, **70**(6 Pt 2):066108.

Tables

Table 1 Literature sources of prognostic genes and sizes of PGS in this study.

Study§	Disease	Number of prognostic genes in study	Gene set	Number of prognostic genes in gene set
Gentles et al.(Nat Med. 2015)	multiple tumor types	various	S1	120*
The Cancer Genome Atlas Research Network (Nature. 2011)	Ovarian Carcinoma	190	S2	185
Lenz et al.(N Engl J Med. 2008)	(Diffuse)Large-B-Cell Lymphomas	39,283,71	S3	330
Zhao et al.(PLoS Med. 2006)	Renal Cell Carcinoma	259	S4	222
Dave et al.(N Engl J Med. 2006)	Burkitt's Lymphoma	217	S5	200
Bullinger et al.(N Engl J Med. 2004)	Acute Myeloid Leukemia(AML)	133	S6	103
Liu et al.(J Natl Cancer Inst. 2014)	(Triple-negative) Breast cancer	11	S7	135
Wang et al.(Lancet. 2005)	(Lymph-node-negative) Breast cancer	76		
van de Vijver et al.(N Engl J Med. 2002)	Breast cancer	70		
Wistuba et al. (Clin Cancer Res. 2013)	Lung adenocarcinoma	31	S8	118
Tang et al. (Clin Cancer Res. 2013)	Non-Small Cell Lung Cancer(NSCLC)	12		
Xie et al. (Clin Cancer Res. 2011)	NSCLC	59		
Zhu et al. (J Clin Oncol. 2010)	NSCLC	15		
Boutros et al. (Proc Natl Acad Sci U S A. 2009)	NSCLC	6		
Lau et al. (J Clin Oncol. 2007)	NSCLC	3		
Gerami et al. (Clin Cancer Res. 2015)	Melanoma	28	S9	174
Wu et al. (Proc Natl Acad Sci U S A. 2013)	Prostate cancer	32		
Li et al. (J Clin Oncol. 2013)	AML	24		
Lohavanichbutr et al. (Clin Cancer Res. 2013)	Oral squamous cell carcinomas(OSCC)	13		
Sveen et al. (Clin Cancer Res. 2012)	Colorectal Cancer	7		
Smith et al. (Gastroenterology. 2010)	Colon Cancer	34		
Ramaswamy et al.Nat Genet. 2003	Solid tumors	17		
Yeoh et al. (Cancer Cell. 2002)	Acute lymphoblastic leukemia(ALL)	7--20		

§Please see supplementary table1 for details of references

*: It consists of top 60 of adversely prognostic genes and top 60 of favorably prognostic genes based on global meta-z score.

Figures

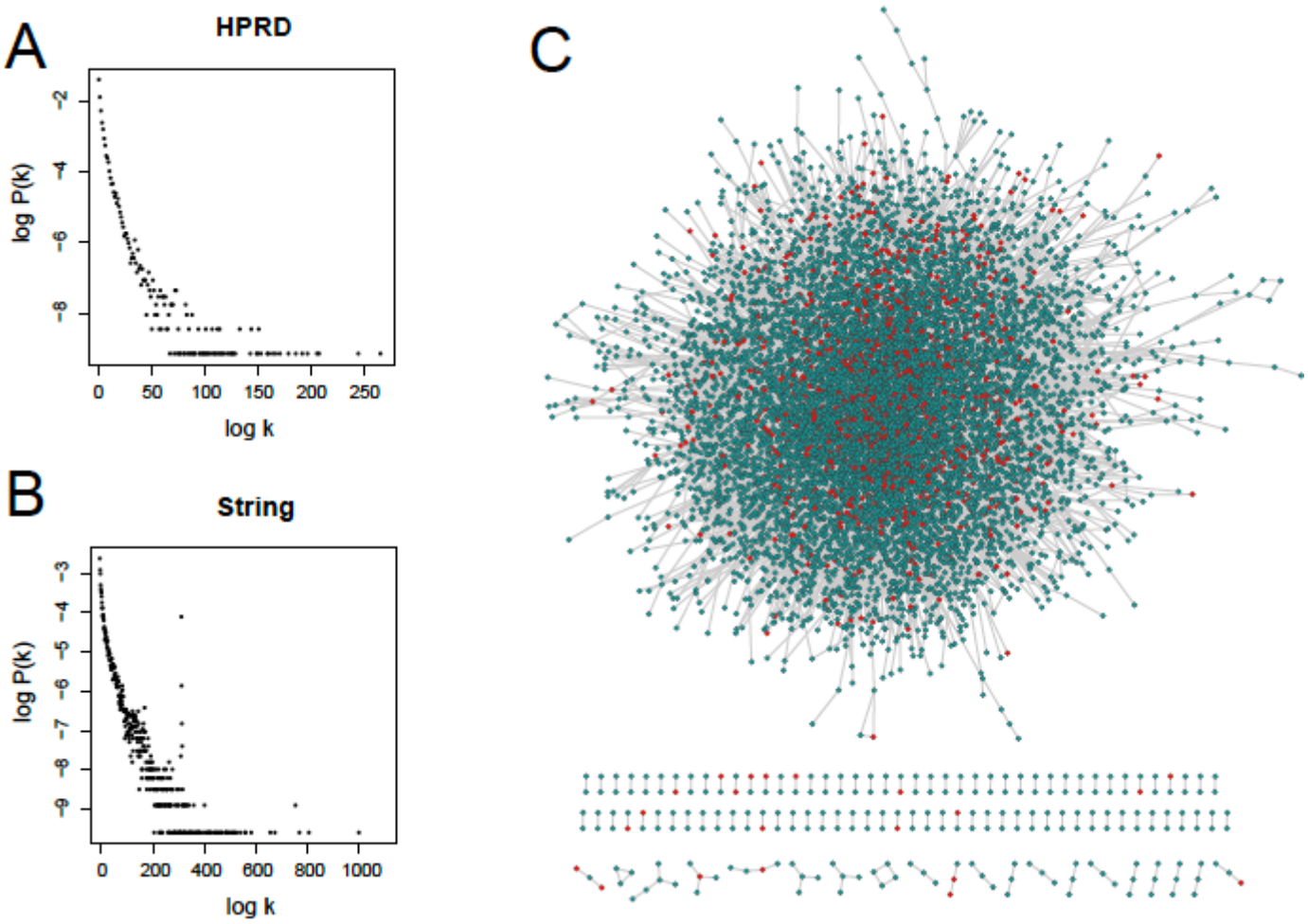

Figure 1

Human protein-protein interaction networks and their node degree distributions. (A) and (B) represent their power-law degree distributions of the HPRD network and the String networks respectively; (C) the HPRD network consisting of 9,402 notes and 36,746 edges (V9.0) and the scattered red nodes represent prognostic genes.

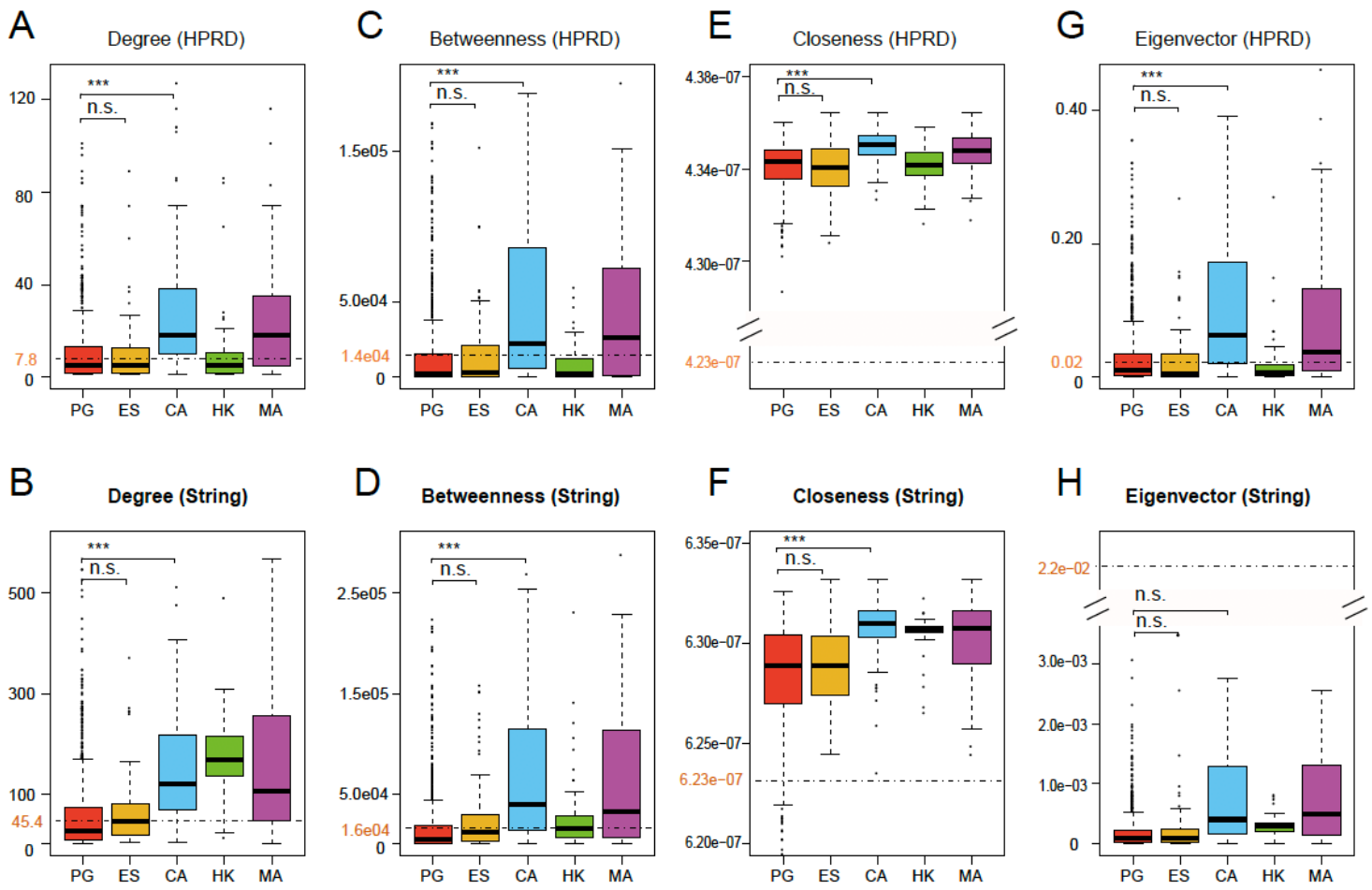

Figure 2

Boxplot of Degree (A and B), Betweenness (C and D), Closeness (E and F) and Eigenvector (G and H) of 1,439 prognostic genes and four other gene sets for comparison based on the HPRD and String networks. One tailed t-test was used to test whether the four network centrality measures had significantly different averages between the union set of all prognostic genes (PG) and ES, CA (triple asterisks, p-value< 0.001; n.s., not significant). The black dashed lines and the numbers in red display the average levels of respective centrality measures for the whole network. The figure shows the four network properties of PG are significantly different from CA and MA but are close to ES.

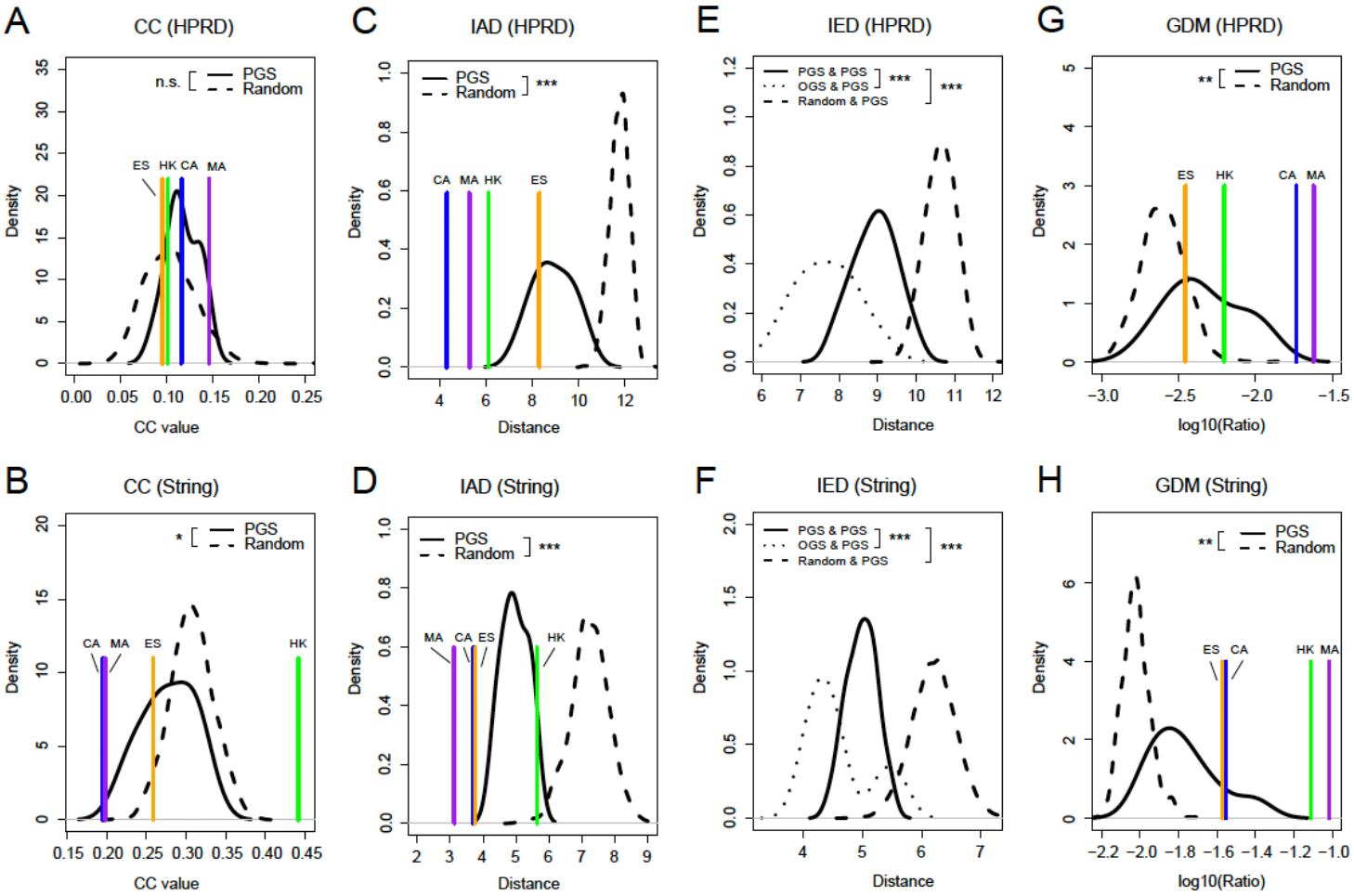

Figure 3

Distribution of clustering coefficient (CC) (A and B), intraset distance (IAD) (C and D), interset distances (IED) (E and F), and genset-distribution in modules (GDM) (G and H) of nine prognostic gene sets (PGS), random sets, and four other gene sets for comparison based on the HPRD and String networks. The random sets were sampled from the whole HGNC gene database 1000 times with each sample containing 120 genes. Differences in the distribution of four network properties between PGS and the random gene sets (or other gene sets) were estimated using one-tailed KS test (triple asterisks, p-value< 0.001; double asterisks, p-value< 0.01; single asterisks, p-value< 0.05; n.s., not significant). In general, four network properties were significantly different between PGS and the random gene sets. Random indicates the random gene sets, OGS indicates other gene sets, namely, CA, MA, ES, and HK. Here, PGS were considered as separate individuals, and PGS & PGS indicates IED between two pairs in PGS.

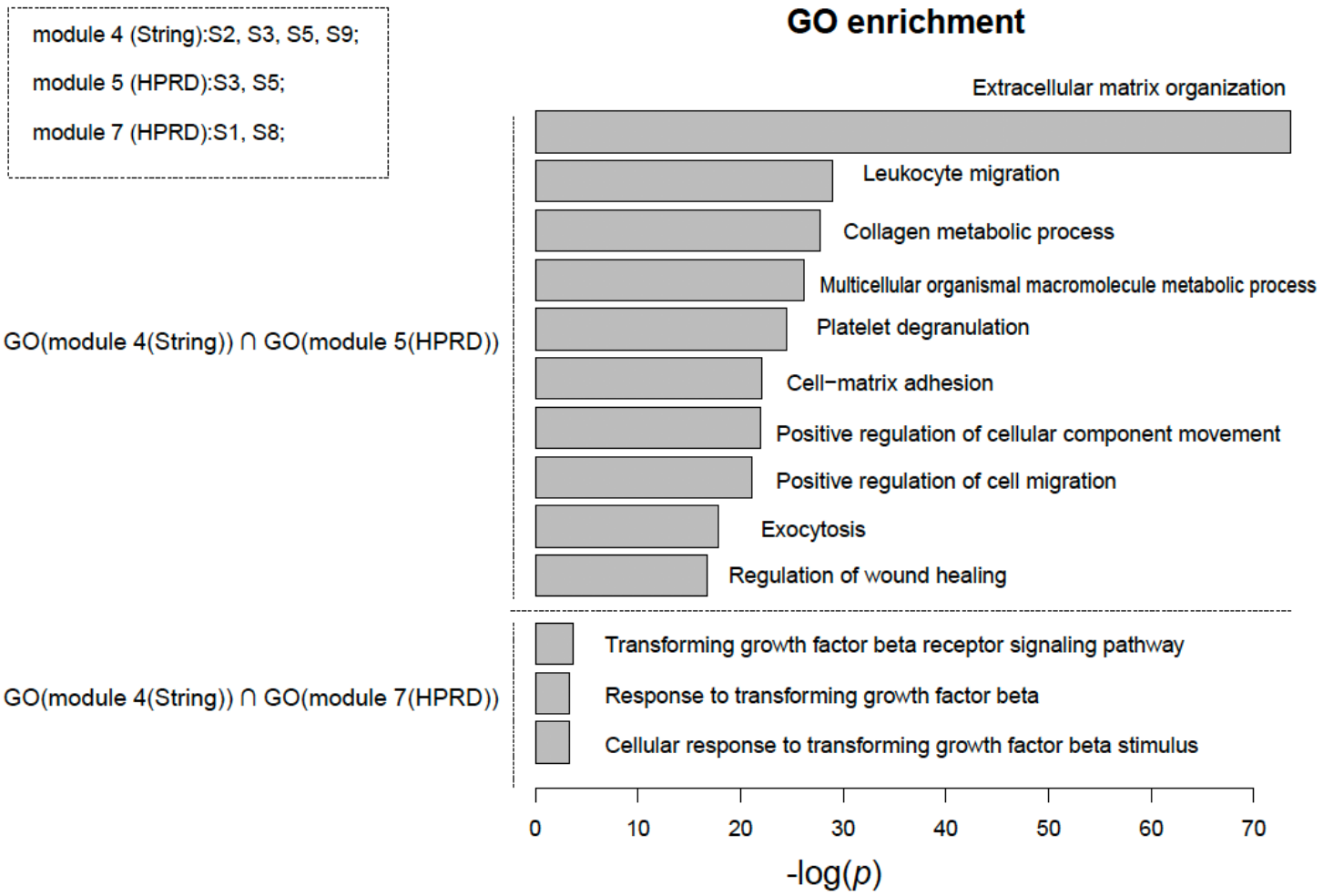

Figure 4

Intersections of enriched GO terms of network modules containing at least two PGS by functional enrichment analysis. The top ten and only three in total of GO terms (BP) were shown separately in the upper and lower parts of the figure. They were sorted in ascending order of p-value, which were estimated by Fisher's test and adjusted by FDR, and the final p-value was the larger of the two with common GO term. The top left of the figure also shows which PGS are included in these modules.

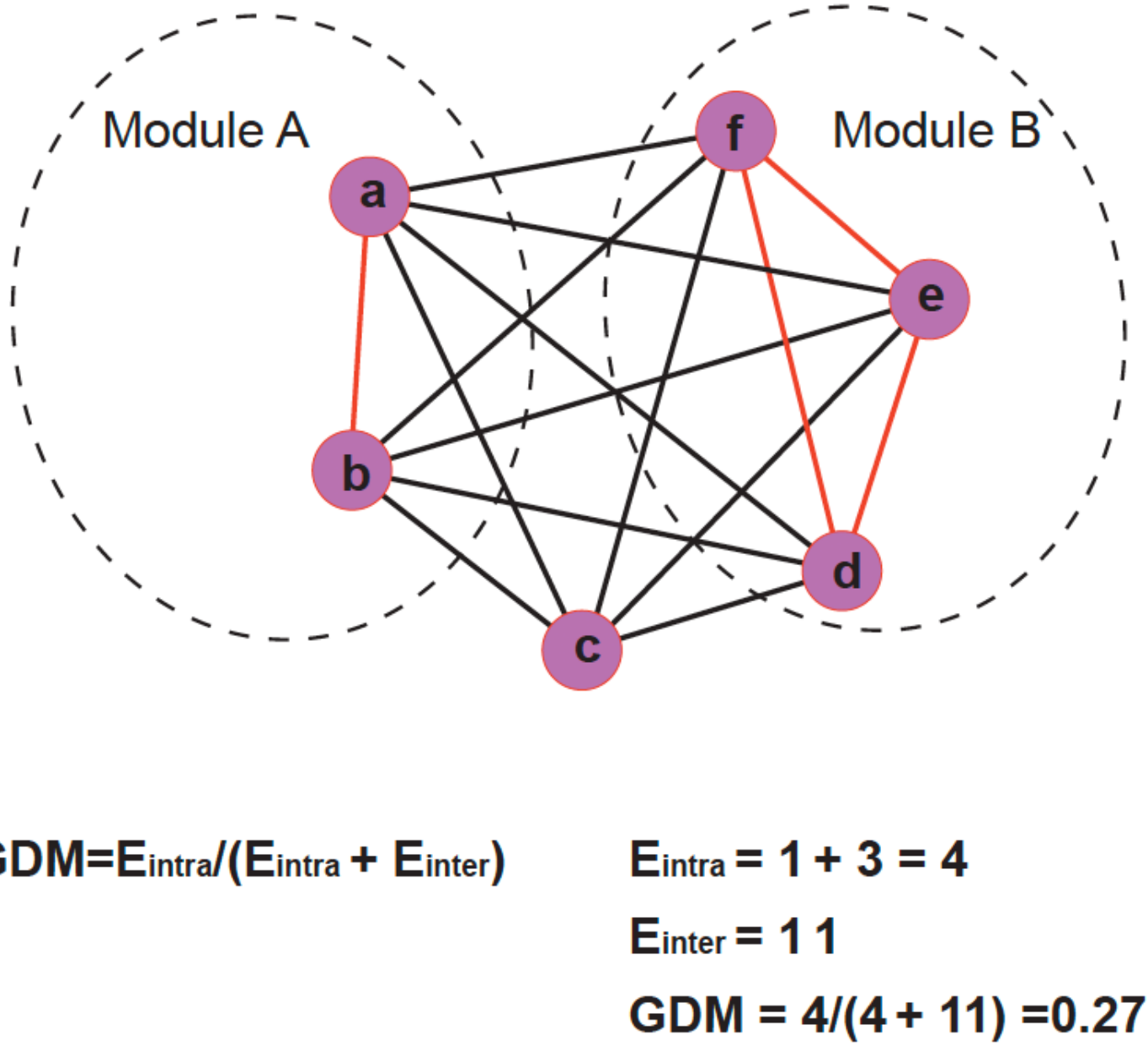

Figure 5

Schematic diagram of calculating GDM of gene sets in the network. The formula for GDM and its calculation process were provided for the given example in the chart below.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS21.xlsx
- TableS22.xlsx
- FigureS1.pdf

- TableS23.xlsx
- TableS1.xlsx