OXFORD

Briefings in Bioinformatics, 20(1), 2019, 89–101

doi: 10.1093/bib/bbx090
Advance Access Publication Date: 8 August 2017
Paper

Discovering cooperative biomarkers for heterogeneous complex disease diagnoses

Duanchen Sun, Xianwen Ren, Eszter Ari, Tamas Korcsmaros, Peter Csermely and Ling-Yun Wu

Corresponding author. Ling-Yun Wu, 55 Zhongguancun East Road, Beijing 100190, China. Tel: +86-10-82541872; E-mail: lywu@amss.ac.cn

Abstract

Biomarkers with high reproducibility and accurate prediction performance can contribute to comprehending the underlying pathogenesis of related complex diseases and further facilitate disease diagnosis and therapy. Techniques integrating gene expression profiles and biological networks for the identification of network-based disease biomarkers are receiving increasing interest. The biomarkers for heterogeneous diseases often exhibit strong cooperative effects, which implies that a set of genes may achieve more accurate outcome prediction than any single gene. In this study, we evaluated various biomarker identification methods that consider gene cooperative effects implicitly or explicitly, and proposed the gene cooperation network to explicitly model the cooperative effects of gene combinations. The gene cooperation network-enhanced method, named as MarkRank, achieves superior performance compared with traditional biomarker identification methods in both simulation studies and real data sets. The biomarkers identified by MarkRank not only have a better prediction accuracy but also have stronger topological relationships in the biological network and exhibit high specificity associated with the related diseases. Furthermore, the top genes identified by MarkRank involve crucial biological processes of related diseases and give a good prioritization for known disease genes. In conclusion, MarkRank suggests that explicit modeling of gene cooperative effects can greatly improve biomarker identification for complex diseases, especially for diseases with high heterogeneity.

Key words: heterogeneous diseases; cancer biomarker; network-based; cooperative biomarker; gene expression

Introduction

Complex diseases such as cancer, Alzheimer's disease and diabetes mellitus have received widespread public and research interest in the past decade. Cancer has become one of the most lethal diseases worldwide, and cancer-related deaths increased dramatically in recent years [1]. It is well accepted that cancer is a complex disease involving many pathways [2, 3], and the underlying pathogenesis resulting from high heterogeneity, rapid proliferation and metastasis is still not clearly known. With the rapid development of genomics technologies, the available big data allow comprehensive characterization of cancers and unbiased identification of specific biomarkers for understanding the disease mechanisms and improving the diagnosis and therapies. Many molecular biomarkers have been revealed in recent decades. For example, Botling *et al.* [4] identified that *CADM1* was significantly associated with survival of non-small lung cancer by conducting meta-analysis. Gentles *et al.* [5] built a nine-gene molecular prognostic index for

Duanchen Sun is a PhD candidate in bioinformatics at the Academy of Mathematics and Systems Science, Chinese Academy of Sciences, and School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, China.
Xianwen Ren is an associate professor at the Biodynamic Optical Imaging Center, Peking University, Beijing, China.
Eszter Ari is an assistant lecturer at Department of Genetics, Eötvös Loránd University, Budapest and a research fellow at Synthetic and Systems Biology Unit, Biological Research Centre of Hungarian Academy of Science, Szeged, Hungary.
Tamas Korcsmaros is a researcher at the Institute of Food Research and the Earlham Institute, Norwich, UK.
Peter Csermely is a professor at the Department of Medical Chemistry, Semmelweis University, Budapest, Hungary.
Ling-Yun Wu is a professor at the Academy of Mathematics and Systems Science, Chinese Academy of Sciences, and School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, China.
Submitted: 24 January 2017; **Received (in revised form):** 19 June 2017

non-small lung cancer. *VEGF* and related proteins were widely studied as predictive biomarkers for renal cell carcinoma [6].

With the appearance of systems biology as well as advanced technologies generating high-throughput omics data, the types of disease biomarkers have gradually changed from individual genes to network-based disease biomarkers [7, 8]. Individual biomarkers may be powerless to overcome the high heterogeneity of complex diseases. But network-based biomarkers can exploit the cooperativity among genes, which are expected to shed fresh light into the disease mechanisms and further improve the diagnosis and treatment of diseases. Various biological networks obtained from large-scale experimental techniques [9–12], e.g. protein–protein interaction (PPI) networks, have been used to improve the identification of network-based biomarkers for complex diseases, which show high reproducibility and provide deep insights into the molecular mechanisms of diseases [13–17].

During the past years, numerous methods that integrate the information from gene expression profiles and biological networks have been developed to identify biomarkers for various diseases [18–24]. For example, Chuang *et al.* [14] and Li *et al.* [19] defined scoring functions for subnetworks and searched discriminative subnetworks using heuristic optimization algorithms. Winter *et al.* [24] developed a network-based ranking method NetRank to discover the cancer biomarkers. Compared with the network-based biomarker identification methods that search most discriminative subnetworks, the ranking-based approaches relax the hard constraint of connected subnetwork to a soft constraint that marker genes are close to each other in the network. This relaxation dramatically reduces the computational complexity of network-based biomarker identification. On the other hand, the relaxation might also improve the effectiveness of discovering biomarkers for heterogeneous diseases, as the disease genes may involve two or more distinct (and disconnected) pathways because of the high heterogeneity of complex diseases.

However, the existing network-based gene ranking approaches [17, 24] do not explicitly consider the cooperative effects of gene combination. The biomarkers for heterogeneous diseases often exhibit strong cooperative effects that the gene combination can achieve more accurate outcome predictions than any individual gene in the combination [14, 17]. Owing to the heterogeneity, the disease samples may be classified into several subtypes, each of which has different mechanism and can be predicted by a few different marker genes. Then, the marker genes from all subtypes constitute cooperative biomarkers with better overall performance. In essence, biomarker identification is a feature selection problem in the perspective of machine learning [25]. Many hybrid and wrapper feature selection methods based on meta-heuristics, such as genetic algorithm and ant colony optimization algorithm, can be applied for identifying the cooperative biomarkers for heterogeneous diseases [26]. However, it is not easy for these optimization-based methods to explicitly and properly integrate the biological network information (e.g. the node importance calculated by random walk model) into their evaluation frameworks. On the other hand, although the biological networks might implicitly include the information of gene cooperation to some extent, the lack of explicit consideration and evaluation of gene cooperative effects may greatly hinder the effectiveness of network-based gene ranking methods.

To evaluate this issue, in this study, we proposed the gene cooperation network to model the cooperative effects of gene combinations and compared the gene cooperation network-enhanced biomarker identification method with the traditional biomarker ranking techniques. Our study suggests that explicit modeling of gene cooperative effects can greatly improve biomarker identification for complex diseases, especially for diseases with high heterogeneity. The gene cooperation network proposed in this study provides a tool to enhance the existing network-based biomarker ranking techniques.

Related work

In this study, we focused on the ranking-based approaches for identifying biomarkers. The traditional biomarker ranking methods only use the expression profile to identify the related biomarkers. In this study, the following popular methods were evaluated to compare their performance: (i) the mutual information (MI) of single-gene profile with the sample label. MI was computed using the R package mpmi as in MarkRank; (ii) the Student's *t*-test of the expression values on normal samples with disease samples. The genes were ranked using the *P*-values of *t*-test in ascending order; (iii) the Pearson correlation coefficient (PCC) of gene expression with the sample label; (iv) the Spearman correlation coefficient (SCC) of gene expression with the sample label; (v) fold change (FC), as defined by the ratio of average expression values in normal over disease samples.

Many network-based gene ranking techniques have been widely used in the field of bioinformatics, for example random walk-based algorithms [17, 24, 27, 28], topological properties-based methods [29–31], differential kernel scheme [32], Bayesian methods [33, 34], Markov random field [35] and order statistics [36]. Briefly speaking, in network-based gene ranking approaches, all candidate genes are first scored and ranked based on a scoring scheme integrated both prior information (e.g. known disease genes or most discriminative genes) and biological network. Then, the top ranked genes are identified as the most potentially 'important' genes. However, most studies in literature focused on the problem of disease gene prioritization [29–33] instead of biomarker identification [17, 24]. Although these two problems are closely related, there exist major differences between two problems in the goals and the evaluation criteria. The disease gene prioritization often used the known disease–gene association database such as OMIM [37] as a gold standard, while the biomarker identification emphasized the prediction capability of genes on the samples of related disease. NetRank is a popular network-based biomarker ranking method, which was successfully applied on many cancer data sets to predict diagnosis and prognosis outcome [17, 24].

To investigate the merit of explicitly modeling gene cooperative effects, we developed an enhanced biomarker ranking method, named as MarkRank, by integrating a gene cooperation network into the NetRank model. The workflow of MarkRank is shown in Figure 1. MarkRank first constructs a gene cooperation network from given gene expression profiles to explicitly model the gene cooperative effects. An edge in the gene cooperation network indicates a possible cooperative gene pair that will improve the prediction power of biomarkers. The gene cooperation network integrates the discriminative power of single genes and the cooperative effects of gene combinations. MarkRank then uses a modified random walk algorithm on two networks to rank candidate genes. Sorting the genes by MarkRank scores, the top ranked genes can be used for many downstream analyses such as diagnosis prediction, survival time prediction, disease gene prediction and drug target prediction.

MarkRank has been implemented in the R package Corbi, publicly available at the CRAN Web site (http://cran.r-project.org/web/packages/Corbi/), which can be readily installed and

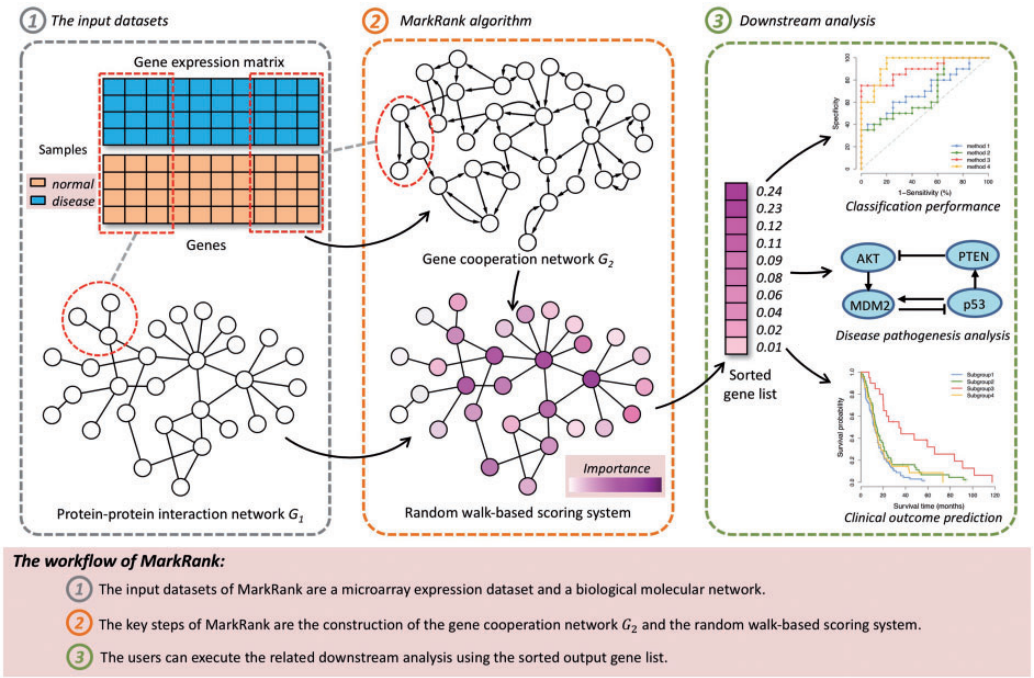

Figure 1. The workflow of the gene cooperation network-enhanced biomarker identification method MarkRank. Data sets in the left dashed box, a microarray expression matrix and a mapped biological network, are the model inputs of MarkRank. The key steps for MarkRank are the construction of a novel gene cooperation network G_2 and a random walk-based method for scoring potential biomarkers (middle dashed box). The users could execute the related downstream analyses, such as the classification of unknown samples or the clinical outcome predictions, using the sorted gene list (right dashed box).

used in R. Additional Materials (including the open-source codes, original data sets and functional enrichment results) are available at http://doc.aporc.org/wiki/MarkRank.

In the framework of random walk-based ranking such as NetRank and MarkRank, it is expected that one node's score is high if its neighbors in the network have high scores. This intuitive interpretation consists of the guilt by association assumption of gene interactions during disease progression, which could benefit the biomarker identification. The special innovation of MarkRank is that we integrated the biological network and the gene cooperation network to discover the most cooperative and discriminative genes, which makes MarkRank distinct from the traditional random walk-based ranking methods [17, 24, 27, 28].

Results

Explicit consideration of gene cooperative effects facilitates effective identification of biomarkers for both homogeneous and heterogeneous diseases in simulation

We first compared the performance of MarkRank with NetRank [17, 24], in a series of well-designed simulated data sets. The simulation study was designed to test whether MarkRank can prioritize the preset target genes on the top of the ranking list. The workflow of simulation studies is shown in Figure 2A. In detail, we first extracted a global network G from a real PPI network, after which one (to simulate homogeneous diseases and simple heterogeneous diseases) or two (to simulated complicated heterogeneous diseases) fixed size subnetworks S_i from G were randomly extracted, and the nodes in S_i were regarded as the target marker genes. Subsequently, we generated a simulated gene expression data set, where the expression values of the above target marker genes had the desired biological characteristics, which will be described below. Finally, we computed the MarkRank and NetRank scores and compared their ranking results for preset target genes (positive genes) and the remaining genes (negative genes) through sufficient replicates.

We simulated three different expression patterns. In the first scenario (Figure 2B), we simply simulated a situation, where the overall expression values of the oracle biomarker genes in disease samples were upregulated. All the disease samples are homogeneous and no subgroup existed. In the second scenario (Figure 2C), we simulated the disease heterogeneity via allowing that each target gene was not significantly differentially expressed. However, when they were integrated together with other target genes and functioned as a module, their overall degree of differential expression was noteworthy. We upregulated the expression values of one cluster of target genes only in a subset of disease samples. Different gene clusters shared little sample overlap with each other. In the third scenario, we assumed that a sole subnetwork or pathway may not have significant discriminative power, but integrating several subnetworks together may reach a prominent improvement in a holistic perspective. Therefore, we simulated two disjoint subnetworks that exhibited complementary differential expression patterns. These three types of simulation studies, taking both homogeneity and heterogeneity into consideration, provided a series of benchmark data sets to evaluate the performance of biomarker identification methods from different perspectives. More details about the simulation studies can be found in the Supplementary Materials.

Results on the simulated data sets are shown in Figure 3. The performance of MarkRank was excellent for identifying biomarkers for homogenous diseases. As measured by the averaged area under the receiver operating characteristic curve (AUC), MarkRank demonstrated superior performance over NetRank and had a smaller variance (Figure 3A). In the second

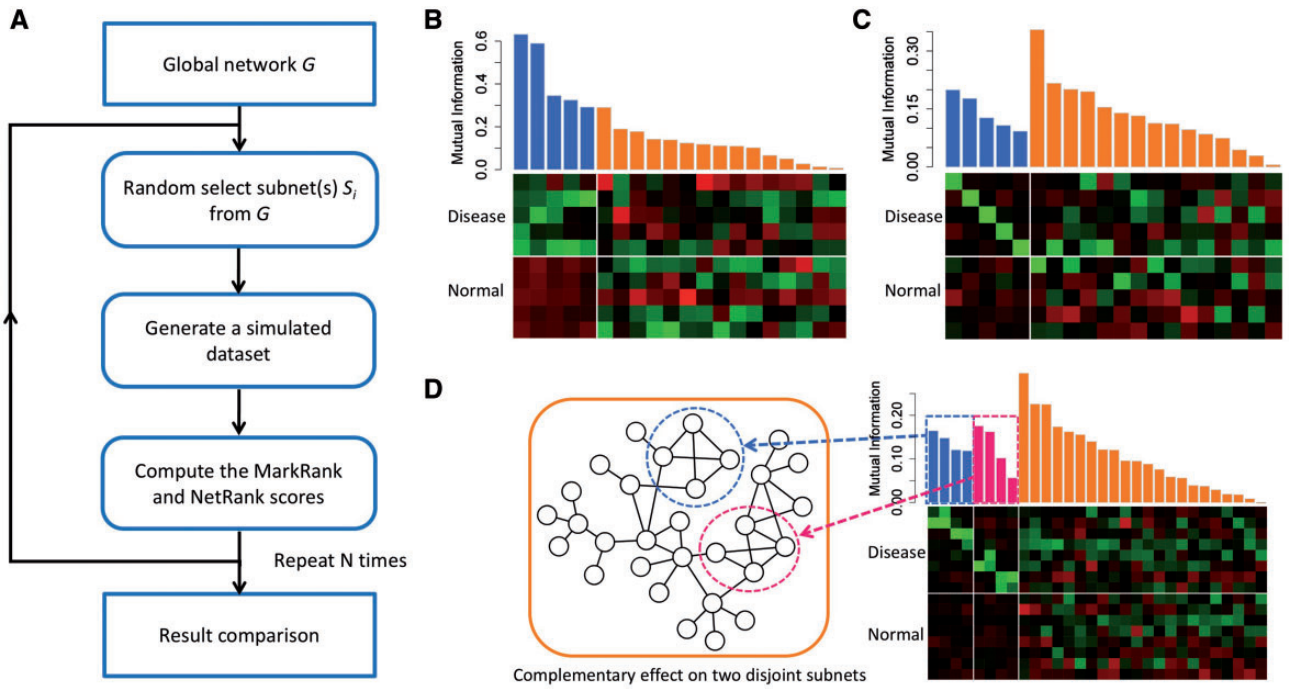

Figure 2. The workflow of simulation comparison and sketch maps of data sets for three simulation scenarios. (**A**) The workflow of simulation comparison. (**B–D**) The sketch maps of simulated data sets in three scenarios. The heat maps were plotted using column normalized ($\mu = 0,\ \sigma = 1$) expression data sets, and the bar plots represent the MI of corresponding genes. These simulation patterns had biological sense and focused on different perspectives.

scenario with disease heterogeneity, MarkRank showed a prominent improvement when the degree of differential expression increased, whereas NetRank did not exhibit an obvious rising trend and became stabilized at AUC values of 0.7–0.75 (Figure 3B). This result is consistent with our expectation, as MarkRank not only exploits the structure information of the PPI network but also considers the cooperation potential of gene combinations, which helps to identify the heterogeneous biomarker genes. Further, in the third scenario, MarkRank still outperformed NetRank (Figure 3C).

We also compared the averaged MI score between the combined subnetwork and the individual connected subnetworks. With the complementary effect of two subnetworks, the score of the combined subnetwork gradually increased and offered distinct advantages over individual ones (Figure 3D). For a better illustration, we calculated and ranked the MI scores of all individually connected subnetworks within a threshold size using an enumerative algorithm. The results showed that two target subnetworks were ranked from several thousands to millions (Figure 3D). These individual subnetworks generally cannot be identified through conventional algorithms because of their lower MI scores and ranks. However, the combined subnetwork did contain discriminative potential in a holistic perspective as shown by its high MI scores. These results showed that MarkRank not only has excellent performance for identifying biomarkers of homogeneous diseases but also can identify biomarkers for the diseases with heterogeneity.

The gene cooperation network-enhanced biomarker identification method outperforms traditional gene ranking methods on various real data sets

To test the efficiency and effectiveness of MarkRank in realistic situations, we used six microarray data sets of different complex diseases to execute real data set analysis, including lung cancer, ulcerative colitis, cervical cancer, renal cell carcinoma, bladder cancer and gastric cancer. We executed the Monte Carlo cross-validation procedure on the lung cancer and the ulcerative colitis data sets to compare MarkRank with other existing biomarker identification methods (see the 'Methods' section and Supplementary Materials). Based on the observation that classes in cervical cancer and renal cell carcinoma data sets are much easier to classify (Supplementary Figures S5, S10 and S11), we did not test the classification performance of MarkRank on these two data sets via the Monte Carlo cross-validation procedure. The bladder cancer and gastric cancer data sets were used to further validate the classification performance of MarkRank (Supplementary Figure S7).

After mapping the common genes present in both PPI network and each of the six expression profiles, we further restricted our study to the largest connected component of the refined network. The summary of the used data sets is shown in Table 1, and more details about the data sets and preprocessing can be found in the 'Methods' section and Supplementary Materials.

The following biomarker ranking methods were tested in the Monte Carlo cross-validation: (i) MI, (ii) Student's *t*-test, (iii) PCCs, (iv) SCCs, (v) FC, (vi) NetRank and (vii) MarkRank. In addition, random gene selection was also taken into consideration as a control method. For each method, a random forest classifier was trained using the top 10 ranked genes as signatures, and the performance was evaluated using the averaged AUC computed from sufficient quantity of repeats.

MarkRank, which had a superior averaged AUC, outperformed the traditional methods on both lung cancer and ulcerative colitis data sets (Figure 4). The superior performance of MarkRank was independent on the classifier algorithms and the number of selected genes. The performance of MarkRank with the top 30 ranked genes as signatures and other classifiers (Support Vector Machine and Naïve Bayes) can be found in Supplementary Figures S3 and S4. Notably, our method was

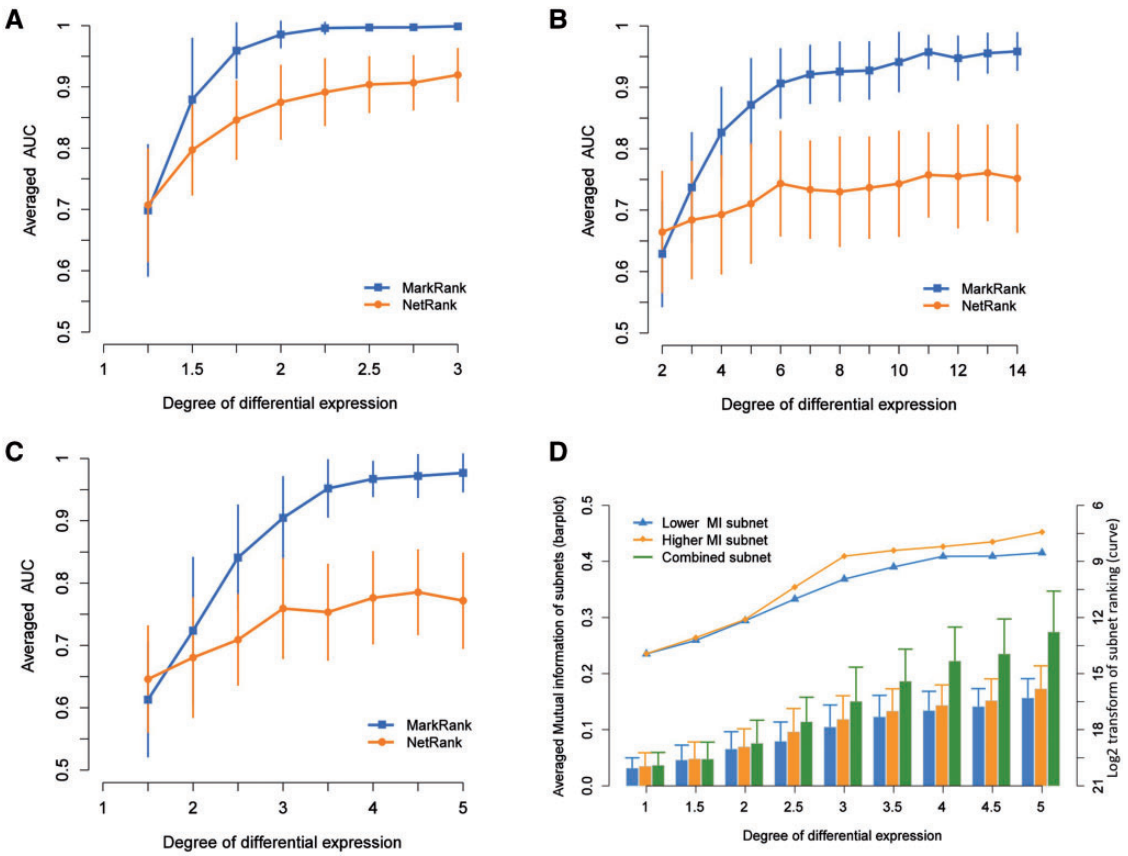

Figure 3. The results of simulation studies. The averaged AUC plus/minus 1-fold SD of MarkRank and NetRank at different degrees of gene expression upregulation in (**A**) first, (**B**) second and (**C**) third simulation scenarios is illustrated. The results of MarkRank are drawn in the curve with squares, and those of NetRank are shown in the curve with dots. (**D**) The bar plot represents the averaged MI of three subnets with 1-fold SDs. The line chart represents the median of ranking position after a log2 transformation. For clarity, we recorded the subnets with lower and higher MI scores in each iteration and computed the related statistical properties.

Table 1. The description of the expression data sets used in this study

Expression data set	Samples	Original genes	Common genes	LCC genes
Lung cancer	187	12 493	7608	7244
Ulcerative colitis	135	10 506	5055	4244
Cervical cancer	57	12 494	7608	7244
Renal cell carcinoma	46	20 108	8859	8539
Bladder cancer	92	20 514	8729	8381
Gastric cancer	94	20 514	8729	8381

Notes: Common genes are the overlapped genes between the expression data set and PPI network. LCC genes are common genes in the largest connected component of the PPI network.

prone to identify the key genes with strong discriminative power when the selected gene number was limited.

It is noteworthy that we also used the PPI networks extracted from two other biological molecular network databases, BioGRID [38, 39] and STRING [40, 41], to test the influence of the network on the performance of MarkRank. The details about these two networks can be found in the Supplementary Materials. The performance of MarkRank was consistently superior to other ranking methods on each biological network and was less affected by the selection of biological networks (Supplementary Figure S6). Principal component analysis using the identified signatures (Supplementary Figures S8 and S9) showed that the MarkRank algorithm was an effective method for ranking genes and had a more robust performance than traditional approaches.

Biomarkers identified via MarkRank are more disease-specific and highly connected to each other on the PPI network

In the following exploration of the MarkRank method, we first ranked the genes of four real disease data sets described above. The top gene set selected via MarkRank contained nodes that were either hub nodes in the PPI network (high-degree nodes in

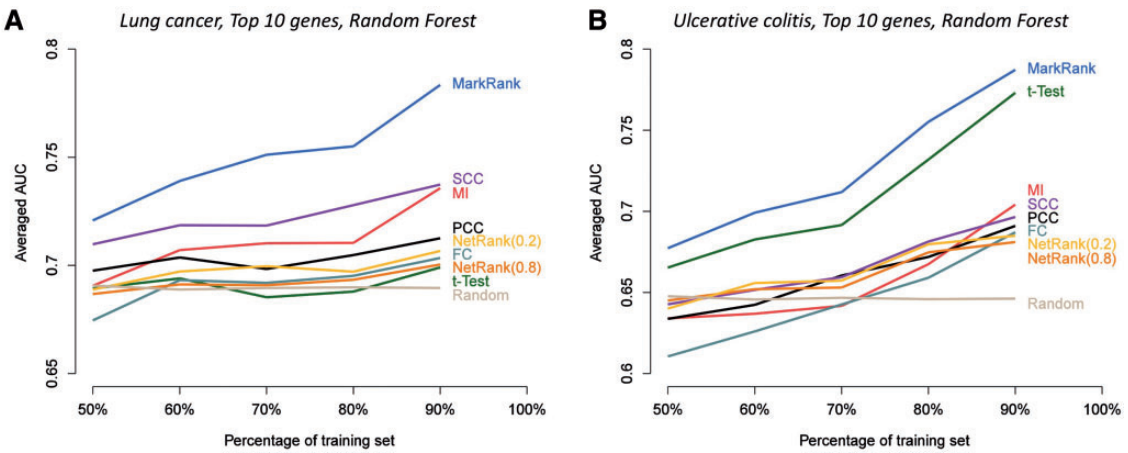

Figure 4. Comparative results of various biomarker identification methods for the (**A**) lung cancer and (**B**) ulcerative colitis data sets in terms of classification power. All gene signatures selected by related approaches were subjected to the Monte Carlo cross-validation procedure. The averaged AUC was computed using a random forest classifier trained by the top 10 genes. The number in brackets after the NetRank method legend stands for different selections of parameter α. The results showed that the MarkRank method had a more robust classification performance and was superior to all other ranking methods.
Note: MI, mutual information; FC, fold change; PCC and SCC, Pearson and Spearman correlation coefficients, respectively.

G_1) or nodes prone to make an improvement to its neighbors in the gene cooperation network (high in-degree nodes in G_2). The detailed information about the full gene list can be found in the Additional Materials (http://doc.aporc.org/wiki/MarkRank).

We plotted the Venn diagram of the top 100 genes from four data sets (Figure 5A). Genes selected via MarkRank showed a strong disease specificity: there were no common genes in four diseases, and there were few common genes between any two diseases. On the contrary, 62 common genes in four diseases were identified by NetRank, with few disease-specific genes, which will bring difficulty to the interpretation of gene ranking results, as shown by the functional enrichment analysis results in Additional Materials (http://doc.aporc.org/wiki/MarkRank). From this perspective, the expression dataset was not fully used by NetRank, as the PPI network dominated the result of random walk algorithm, which was verified by the results on the PPI networks from different databases (Supplementary Figure S6).

The top 10 genes identified by MarkRank in each type of disease contain many known biomarkers. For example, *UBQLN4*, *P4HA2*, *RABAC1*, *ARCN1* and *P4HB* have close relationship with lung cancer in literature. Dong *et al.* [42] performed a GWAS analysis and identified *P4HA2* as a novel susceptibility locus for lung cancer. *RABAC1* is one of the invasion-associated four-gene signature identified by Hsu *et al.* [43] in non-small cell lung cancer. *ARCN1* is on a list of the top 25 predictive genes selected for the weighted-voting outcome classifier for lung cancer [44]. For the ulcerative colitis, *MAPKAPK2* is a candidate gene of inflammatory bowel disease (IBD) SNP (rs3024505) [45]. *EWSR1* was proved that it can fused to the N-terminally truncated and C-terminally intact active domain of *NFATc2*, which are related with the development and metastasis of cancer [46]. For the cervical cancer, *KRT13* has been reported as a marker gene in the diagnosis of human cervix carcinomas [47]. The chromosomal position 11q22.3-q23.1, at which *CRYAB* located, was previously reported to be frequently lost in cervical cancers [48]. *SPRR3*, *HSPB8* and *SPINK5* were identified as three novel differentially expressed genes by >2-FC for cervical cancer [49]. *CALB1* was reported as a differentially expressed gene for renal cell carcinoma [50]. *SPAG4* was identified as a novel *HIF-1* target and was reported as a novel biomarker gene for renal cell carcinoma [51].

To explore whether the identified genes are located closely in the PPI network, we visualized their network topological relationship (Figure 5B–E). Based on the distance matrix of selected nodes and the derived minimum spanning tree (MST), edges were grouped into three categories (see Supplementary Materials). To obtain an explicit and succinct view, we used the top 30 genes to illustrate their relationship. The network views of the identified genes clearly showed that several connected components exist and most genes are close to the connected components. These locally tight connection structures are crucial to identify novel disease-related pathways.

We further quantitatively compared these locally tight connection structures of MarkRank with that of the traditional methods. We tested the statistical significance of gene connectivity for each ranking method by random sampling (see Supplementary Materials). The number of gene pairs with shortest path distance $k \leq 3$ was significantly larger than expected by chance for the top genes ranked by MarkRank for the lung cancer, ulcerative colitis and cervical cancer data sets (largest *P*-value 0.03673, Supplementary Figure S12, Supplementary Tables S2–S5). A similar result was obtained for renal cell carcinoma for $k \leq 2$ (largest *P*-value 0.00791). However, the top genes ranked by traditional methods did not show a consistent statistical significance in any data set except that the genes identified via NetRank were always prone to gather together in the PPI network.

We also analyzed the topological properties of the genes identified via MarkRank from the viewpoint of complex networks. Two main categories, node importance indexes (degree, betweenness centrality) and module importance indexes (clustering coefficient, the number of connected components), were selected as the measurements of each identified gene set. The complex network indexes showed that the network-based methods, compared with the traditional gene-based methods, can identify the key genes in the PPI network, which have a relatively higher importance in network structure (Supplementary Figures S13–S16).

In summary, MarkRank appropriately balances the discriminative power of gene combinations and the information of biological networks, and the identified genes not only had superior classification accuracy and strong disease specificity but also had significant network connectivity and topological importance in the PPI network.

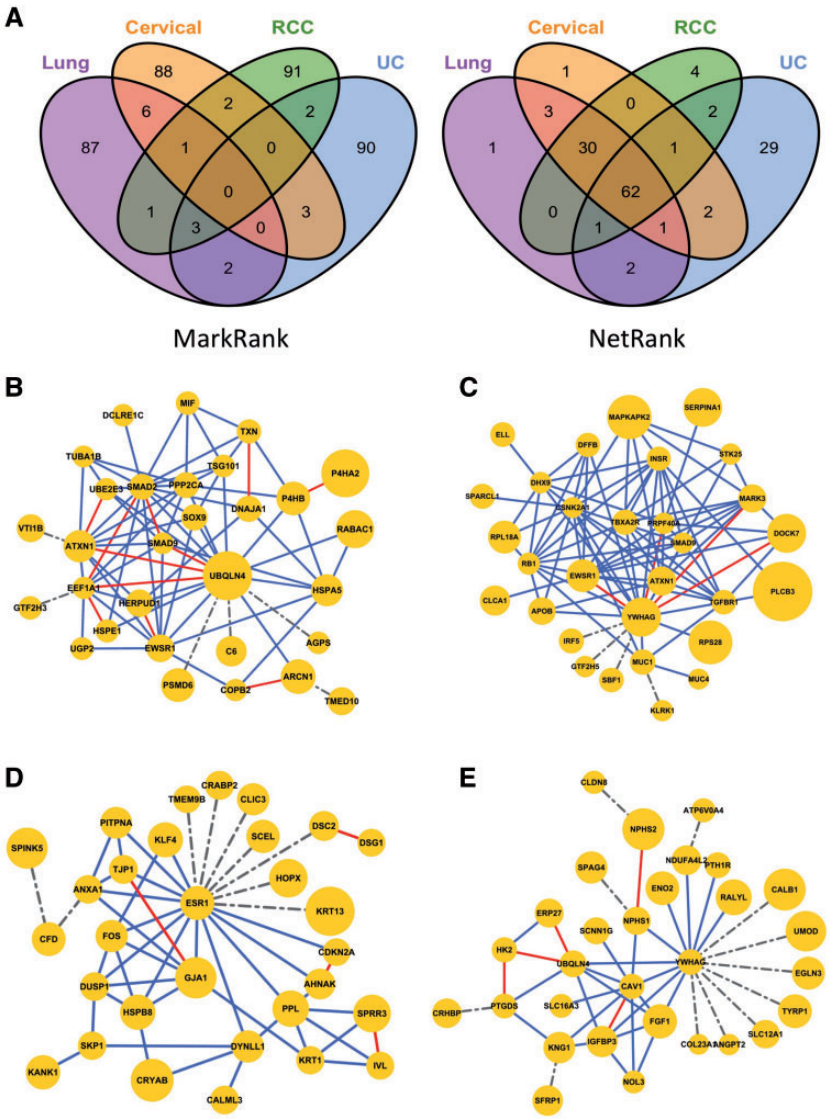

Figure 5. Comparative results for four real disease data sets in terms of network properties. (**A**) The Venn diagrams of the top 100 genes selected via MarkRank and NetRank in four data sets. UC and RCC indicate ulcerative colitis and renal cell carcinoma, respectively. (**B–E**) Network views of the top 30 identified genes in lung cancer, ulcerative colitis, cervical cancer and renal cell carcinoma. Edges were grouped into three categories based on the derived MST. The red line represents directly connected nodes. The blue line and the gray dash dot line represent the distance of node pairs in PPI network 'equal to 2' and 'larger than 2', respectively. The node size is proportional to the MarkRank score.

Functional enrichment analyses of MarkRank biomarkers reveal crucial biological processes of related diseases

The top 10 enriched gene ontology (GO) terms in the MarkRank biomarkers of each data set were shown in Supplementary Tables S6–S9. For lung cancer, the enriched GO terms of MarkRank genes were mainly for cell death, apoptosis and several key protein metabolic processes, which had been proven to have close relationships with oncogenesis [52, 53]. As for ulcerative colitis, the top functional terms were the regulation of phosphate, phosphorylation and response to stimulus. Interleukin-13 (*IL-13*) has been shown to be a key Th2 cytokine in ulcerative colitis, inducing the phosphorylation of *STAT-6*, a key transcription factor, in colonic epithelial cells and having further effects on epithelial tight junctions, apoptosis and cell restitution [54, 55]. In addition, *IL-13* caused rapid phosphorylation of the three of four members of Janus family of kinases (JAKs) and phosphorylated insulin response substrate-1, *IL-4R* p140, *JAK1* and *Tyk2* in human colon carcinoma cell lines [56]. The MarkRank identified gene set from the cervical cancer data set was enriched in 'epithelial cell' and 'epidermal cell differentiation' functions, which is consistent with the histological characteristic of human papillomavirus and cervical carcinoma [57–59]. MarkRank genes of renal cell carcinoma were enriched in 'excretion' and 'chemical homeostasis' functions, which are the main functions of renal tissue [60]. The results showed that the top genes identified via MarkRank involved crucial biological processes of the related diseases.

We further evaluated whether the known important disease-specific pathways or genes are significantly prioritized for each data set by performing the Kolmogorov–Smirnov test (K-S test). In this study, five specific gene sets [four from the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and one from the Molecular Signatures Database of Broad Institute

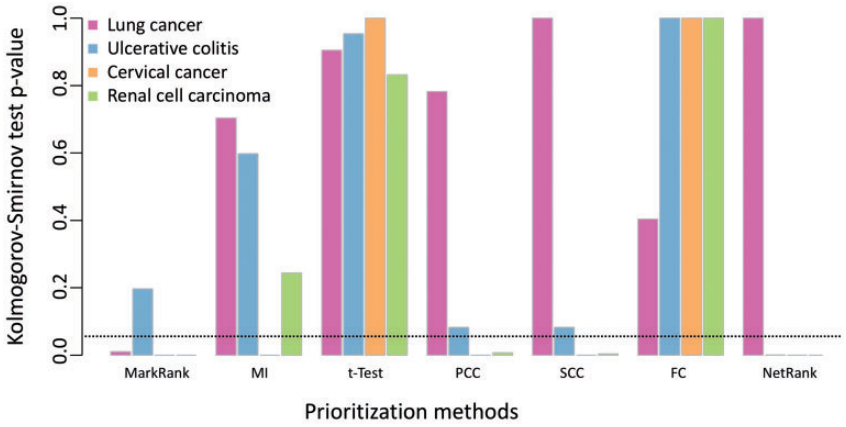

Figure 6. Pathways enriched in the top ranked list of each biomarker identification method for four real data sets. The black dotted line represents the $P = 0.05$ significance level. The results showed that the MarkRank method had significant *P*-values in the K-S test and gave a good prioritization for known important disease-specific genes when compared with all other ranking methods. The abbreviations for each method are the same as in Figure 4.

(MSigDB)-c2] are the most relevant to the corresponding diseases. The descriptions of the five gene sets and the full ordered gene lists can be found in the Additional Materials (http://doc.aporc.org/wiki/MarkRank). *P*-values for each data set are shown in Figure 6 and Supplementary Figure S17.

The K-S test *P*-values of the MarkRank method exhibited excellent overall enrichment performance, while genes sorted by Student's *t*-test and FC had the worst performance. Although all methods in this study have high classification accuracy on cervical cancer and renal cell carcinoma data sets (Supplementary Figures S5, S10 and S11), several traditional gene-based ranking methods, such as MI, Student's *t*-test and FC, failed to significantly prioritize the known important disease genes. Compared with traditional gene-based ranking methods, MarkRank and NetRank, which integrate the biological network information, showed a preferable performance and good balance for biological explanation and classification accuracy. Both MarkRank and NetRank had significant enrichment in three of four data sets. Notably, only the MarkRank genes were significantly enriched in the lung cancer data set ($P = 0.0108$), while the lowest *P*-value of the other methods was $P = 0.4041$. MarkRank did not produce significant enrichment in the ulcerative colitis data set ($P = 0.1966$), noted that ulcerative colitis is only a specific type of IBD, and other subtypes of this disease family (such as the abundant Crohn's disease) may introduce a bias to the KEGG pathway. In general, the results of pathway enrichment tests revealed that MarkRank gave a good prioritization for known disease-related genes.

Discussion and conclusion

In this article, we introduced and compared the ranking-based biomarker identification approaches, including traditional ranking methods and network-based ranking methods. Traditional ranking methods only use the gene expression data, and their results may be not reliable because of the high noise and heterogeneity in gene expression data. The network-based biomarker ranking methods such as NetRank and MarkRank improve the reliability and reproducibility by integrating the complicated relationships between genes involved in the development of complex diseases.

The main difference between network-based biomarker ranking methods and other network-based approaches that identify the subnetworks or modules as biomarkers is that ranking-based methods relaxed the constraint that one single connected gene set accounts for one target biomarker. On one hand, the inevitable shortcomings of the biological technical capabilities, such as measurement errors and noise, resulted in the incompleteness of current PPI networks. With limited knowledge of the proteome, PPI networks are generally incomplete not only in their edges but also in their nodes. On the other hand, the major assumption that key disease genes gather together in a small connected subnetwork may conflict with disease heterogeneity because key genes may involve two or more distinct (and separate) pathways because of the high heterogeneity of complex diseases. The efficacy of identifying individual connected subnetworks may be restricted, as they rarely overlap with multiple pathways. Instead, network-based ranking methods permitted the existence of several connected components scattered on the PPI network, which is important in the identification of novel disease-related pathways and is not affected by the incompleteness of the PPI network or other specificities in disease mechanisms.

In this study, we further enhanced the network-based biomarker ranking method by explicitly considering the cooperative effects of gene combination. The advantages of MarkRank can be briefly summarized as follows. First, MarkRank has a higher classification capability than existing methods, which is independent from the selection of classifier, the training feature number and the biological molecular network. Second, MarkRank can discover disease biomarkers for complex diseases with or without heterogeneity. Third, the biomarkers identified by MarkRank exhibit both strong disease specificity and significant network connectivity. Finally, the biomarkers identified by MarkRank involve crucial biological processes of related diseases, and MarkRank gives a good prioritization for known disease genes.

Compared with other network-based ranking approaches such as NetRank, explicit modeling the cooperative effects significantly improves the classification accuracy and the disease specificity. Compared with the traditional ranking approaches, subtle integration of the biological network and the gene expression data to take both source of information into consideration makes the identified biomarkers much more reliable and biologically meaningful as shown by the statistical tests. Last but not least, the modified random walk model offers a flexible framework to integrate different source of information.

The key component of MarkRank is the gene cooperation network that reflects the cooperative effects of gene pairs. We followed the subnetwork scoring function as in previous studies of network-based biomarker identification and used the MI increment as the weight of related edge in the gene cooperation network. This scoring function is relatively simple and could not identify the cooperation between a pair of negative correlated genes, as their effects will cancel each other. An appropriate scoring function that could discover both positive and negative correlated gene pairs may further improve the performance of MarkRank. Using the bi-clustering approaches in the construction of gene cooperation network is also an alternative solution. The popular bi-clustering methods such as Plaid [61], QUBIC [62] and FABIA [63], can not only handle the positive and negative correlations but also identify the strong–weak correlations between genes. We can first use the bi-clustering tool to identify gene modules, which embed the information of gene correlations. The genes belong to the same module might share similar biological functions, and these modules can serve as several meta-genes. Then, we can replace the genes by meta-genes in the construction of gene cooperation network, which might simplify the networks, improve the performance and more importantly reduce the computation time. Another possible improvement of scoring function is directly modeling the changes of classification accuracy instead of estimation using MI. However, the classification accuracy depends on the selected classifier, and unbiased estimation needs complicated and time-consuming cross-validation.

The execution time and memory usage of MarkRank on real data sets are shown in Supplementary Tables S10–S11. As expected, the construction of the gene cooperation network takes long time, as it needs to compute MI for all pairs of possible genes. However, the time complexity for computing all MI is polynomial (quadratic) on the number of genes and thus acceptable for large data sets. In practice, it may be time-consuming when compared with other computation steps such as random walk iteration. Alternatively, we designed another two alternative approaches to reduce the computation time. The one simplified the calculation in the gene cooperation network construction by using a fast ordering of single-gene MI and another one added an additional parameter d to obtain a similar effect (see Supplementary Materials). However, the former method neglects the combined effects of cooperative genes diverging from our original motivation, and the latter one simplifies the calculation at the cost of missing cooperative information of genes with long distance in the PPI network. Therefore, to fully use the information from the expression data set, we recommended the version introduced in the 'Methods' section.

The gene cooperation network in this study is mainly based on two gene combinations. Although the gene cooperation network contains higher-order combination effects to some extent, there may be limited effectiveness on overall performance. It is insufficient to identify a crucial combination of three or more genes by using just the gene cooperation network based on two-gene combinations. In fact, the scoring function itself can evaluate any number of genes. But modeling an appropriate higher-order combination effect may need more complex calculation and suffer the curse of dimensionality. It may require a new modeling perspective to exquisitely generalize MarkRank to high-order cooperation, which is one of the goals of our future research.

The computational framework of MarkRank itself is more flexible than traditional random walk-based methods for integrating multiple sources of biological data. In this work, only two types of data, PPI networks and gene expression profiles, were used to discover disease biomarkers. Instead of using the gene expression data in both gene cooperation network and prior information, the performance of MarkRank might be improved by incorporating other information from multiple sources into the model. On the other hand, the construction of a gene cooperation network and prior information in this framework can be adapted flexibly in terms of the data available to achieve the goal of the study. The prior information, for instance, can be determined by other independent sources such as known disease genes [64] extracted from the literature and databases (e.g. OMIM [37]) or the driver genes inferred from somatic mutation data [65] to settle the corresponding biological problems. The gene cooperation network can also be refined by exploiting better evaluation function and more information. For example, the exclusivity between mutated genes is often considered as the main source of cancer heterogeneity [66, 67], which might be used to discover and validate informative links in the gene cooperation network.

In conclusion, the random walk-based MarkRank algorithm that we proposed in this article to discover disease biomarkers integrates multisource information including biological networks, prior information about related diseases and the discriminative power of cooperative gene combinations, and achieved a superior performance compared with existing methods in both predesigned simulation studies and real data sets. Moreover, MarkRank exhibited high specificity associated with the related diseases. Our studies suggest that MarkRank is a promising, ease-to-use R package tool to explore the underlying pathogenesis of complex diseases from a new perspective.

Methods

Microarray data sets

In this study, six microarray expression data sets were downloaded from the Gene Expression Omnibus (GEO) repository http://www.ncbi.nlm.nih.gov/geo/ (accession numbers GSE4115, GSE11223, GSE9750, GSE36895, GSE31189 and GSE64951) for real data set analysis. The related human complex diseases of these datasets are lung cancer, ulcerative colitis, cervical cancer, renal cell carcinoma, bladder cancer and gastric cancer, respectively. These data sets are all based on Affymetrix Human Genome U133 Plus 2.0 arrays. We averaged the expression values of the probes mapping on the same gene. The first four data sets were used in all analyses except that the third and fourth data sets were not used in Monte Carlo cross-validation. The last two data sets were used to further validate the classification performance of evaluated methods in a broader perspective. More information including the preprocessing procedure can be found in Supplementary Materials.

Biomolecular networks

In our work, three PPI networks were extracted from the HPRD [68], BioGRID [39], STRING [41] databases, respectively. The HPRD network was used to perform the main analyses, while the BioGRID and STRING networks were used to test the influence of PPI networks and the robustness of the network-based methods. The preprocessing procedure of PPI networks can be found in Supplementary Materials.

MarkRank

The whole workflow of MarkRank is shown in Figure 1. Briefly, the model inputs of MarkRank are a PPI network G_1 and a

mapped microarray mRNA expression data set, with expression values of n genes across m samples. The procedure to identify the cooperative biomarker using MarkRank mainly constitutes two steps, which were described in next sections: the construction of the gene cooperation network G_2 and the random walk-based scoring system to rank all candidate genes.

Gene cooperation network

Based on the gene expression data set, MarkRank constructs a directed weighted gene cooperation network G_2 following an information-theoretic scheme proposed by Chuang *et al.* [14]. Denote $e(i)$ is the expression profile of gene i across all samples, and y is the binary (0-1) vector indicating the phenotype of each sample. $y(i) = 1$ means that the i-th sample is diagnosed with disease, and $y(i) = 0$ marks the opposite case. Chuang *et al.* defined an evaluation function of discriminative power for a selected gene set S by computing the MI of the aggregated activity of genes in S and the label y. Precisely, denote $x_i \in \{0,\ 1\}$ is the indicator variable to reflect whether the i-th gene is selected in S, which defined as $x_i = 1$ if $i \in S$ and $x_i = 0$ if $i \notin S$. Using the above symbols, the evaluation function is computed as:

$$f(x_1, x_2, \ldots, x_n) = MI\ \left(\ y,\ \sum_{i \in S} \frac{e(i)}{\sqrt{|S|}}\right),$$

where $|S|$ means the size of selected gene set S. Specially, $f(x_i = 1, others = 0)$ denotes the MI of i-th gene $MI(e(i), y)$. In this study, the MI was computed using the R package mpmi (version 0.41 from https://cran.r-project.org/web/packages/mpmi/).

The gene cooperation network G_2 has the same node set as the PPI network G_1. The weight of directed edge $(i,\ j)$ is defined as the potential improvement of discriminative power when gene j is added:

$$w_{i,j} = \max\{0, f(x_i = 1, x_j = 1, others = 0) - f(x_i = 1, others = 0)\}.$$

That is, only if the integration of gene j can make an improvement to the MI of gene i, the directed edge (i, j) from node i to j and the weight of this edge $w_{i,j}$ is retained. In this way, G_2 approximately evaluates the combined effects of gene combination $\{i, j\}$ over a single gene $\{i\}$ using the MI-based evaluation function, and the edges in this network would capture the cooperation potential of related genes.

Modified random walk model

Random walk is one of the most fundamental types of stochastic processes and plays an important role in network science [69]. In the past few decades, many modified versions of random walks were produced, with extensive applications to many problems, such as network searching [70], nodes or edges centralities ranking [71, 72] and community structure detection [73, 74].

MarkRank takes both the PPI network and the gene cooperation network into consideration for ranking genes. Denote r_i as the biomarker score of gene i, which indicates the possibility that gene i is selected as a biomarker. It is expected that r_i is high if the neighbors of gene i in the PPI network have high scores. If gene i is selected as a biomarker (the score r_i is high) and there is a directed edge from gene i to j in the gene cooperation network, r_j is assumed to be high, as adding gene j into the gene combination may improve the classification capability.

Therefore, the biomarker scores are required to satisfy the following equation, which is essentially the random walk model with restart in two networks:

$$r_i = (1-\alpha)e_i + \alpha\left[\lambda \sum_{u \in N^1(i)} \frac{1}{\deg^1(u)} r_u + (1-\lambda) \sum_{u \in N^2_{in}(i)} \frac{w_{ui}}{\sum_{j \in N^2_{out}(u)} w_{uj}} r_u\right].$$

Here, r_i denotes the score of gene i, and e_i is the prior information (i.e. distribution of seed nodes). In this study, the absolute value of PCC between a gene's expression profile and the disease/control label y in the training data set is used as prior information.

Parameter $\alpha \in [0,\ 1]$ is the restart probability of random walk and balances the effect of prior information e_i and the influence of networks, whereas parameter $\lambda \in [0,\ 1]$ is the retain probability in the PPI network and balances the importance of the two networks. Larger λ inclines to lay more emphasis on G_1, i.e. the PPI network. The results of the performance of parameter settings can be found in Supplementary Figure S2. In this study, we set $\alpha = 0.8$ and $\lambda = 0.2$ as default parameter setting.

Denote $E = [e_1, \ldots, e_n]^T$. A_1 is the symmetric adjacent matrix of G_1 and A_2 is the weighted adjacent matrix of G_2. D_1 is a diagonal matrix, where the elements are the degrees of corresponding nodes in G_1. D_2 is another diagonal matrix, where the elements are the sum of weights of outgoing edges of corresponding nodes in G_2. The MarkRank score can be computed iteratively using the matrix format:

$$R^{(k)} = (1-\alpha)E + \alpha\left[\lambda A_1^T D_1^{-1} + (1-\lambda)A_2^T D_2^{-1}\right] R^{(k-1)},$$

where $(\cdot)^T$ means matrix transposition, and $R^{(k)} = [r_1^{(k)}, \ldots, r_n^{(k)}]^T$ is the MarkRank scores in k-iteration.

NetRank

NetRank [24] is a representative of network-based biomarker ranking method based on the Google's PageRank algorithm [71] and propagates the gene values to their neighbors through the biological network. The rank $r_j^{(k)}$ of gene j in the k-th iteration is updated using the following formula:

$$r_j^{(k)} = (1-\alpha)e_j + \alpha \sum_{i=1}^{N} \frac{m_{ij}}{\deg(i)} r_i^{(k-1)},$$

where parameter α is the damping factor and selected using an additional inner cross-validation loop. m_{ij} is the related element in $N \times N$ adjacent matrix M. The prior information e_j for NetRank was the same as for MarkRank as mentioned above.

Evaluation of classification performance

In this work, we used several different evaluation procedures to compare the classification performance of biomarker identification methods. In the simulation studies with preselected truth biomarkers, each method was directly evaluated using the AUC. The goal of the simulation studies was to test whether each method can discriminate the preset biomarker genes from the remaining genes.

For the real data sets where the truth biomarkers are not known, a Monte Carlo cross-validation procedure [24] was adopted to test the classification power of the top ranked biomarker genes as the effectiveness measurement of the corresponding biomarker identification method. In this work, the

random forest, support vector machine and Naïve Bayes classifiers were trained to classify the samples. Note that both the gene cooperation network and the prior information were constructed only based on the training data in each iteration of the cross-validation to avoid the evaluation bias because of information leakage. See Supplementary Materials for more details.

Functional enrichment analyses

In this study, we used two different enrichment approaches to evaluate the marker genes identified by MarkRank:

1. The hypergeometric test-based functional enrichment analyses. We used the Cytoscape plugin BiNGO [75] to perform the enrichment analysis of GO categories. We would like to assess if the degree that a gene list (here the top 100 MarkRank genes) related to a GO category is any better than that observed by chance alone.
2. The K-S test-based gene set enrichment analyses. The K-S test is used for testing whether a given gene set (e.g. disease pathway) is significantly prioritized in a ranked full gene list. The K-S test does not require a strict cutoff. We adopted the GSEAPreranked tool of the GSEA software [76] with a correlation-weighted K-S test. Notably, we selected four specific gene sets from the KEGG [77] pathway database and one from the MSigDB curated gene sets (c2, CGP: chemical and genetic perturbations) [78], which are the most relevant to related diseases. See Supplementary Materials for more details.

Key Points

- The biomarkers for heterogeneous diseases often exhibit strong cooperative effects. A set of genes may achieve more accurate outcome prediction than any single gene.
- We proposed the gene cooperation network to evaluate the gene cooperative effects on biomarker identification.
- The biomarkers identified by the gene cooperation network-enhanced biomarker identification tool MarkRank not only have a better prediction accuracy but also have stronger topological and functional relationships in the biological network and disease settings, providing important guidance and tools for practical usage.
- This study suggests that explicit modeling of gene cooperative effects can greatly improve biomarker identification for complex diseases, especially for diseases with high heterogeneity.

Supplementary data

Supplementary data are available at *BIB* online.

Acknowledgements

The authors are grateful to Dezső Módos and Johanne Brooks for their expert help in the disease-specific gene set analysis.

Funding

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (grant number XDB13040600) and the National Natural Science Foundation of China (grant numbers 11131009, 11631014, 91330114 and 11661141019). TK's work was supported by a fellowship in computational biology at the Earlham Institute (Norwich, UK) in partnership with the Institute of Food Research (Norwich, UK) and strategically supported by the Biotechnological and Biosciences Research Council, UK. Work in PC's laboratory was supported by the Hungarian National Research, Development and Innovation Office (grant number K115378).

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. *CA Cancer J Clin* 2015;**65**:5–29.
2. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. *Nat Med* 2004;**10**:789–99.
3. Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell* 2000;**100**:57–70.
4. Botling J, Edlund K, Lohr M, *et al*. Biomarker discovery in non-small cell lung cancer: integrating gene expression profiling, meta-analysis, and tissue microarray validation. *Clin Cancer Res* 2013;**19**:194–204.
5. Gentles AJ, Bratman SV, Lee LJ, *et al*. Integrating tumor and stromal gene expression signatures with clinical indices for survival stratification of early-stage non-small cell lung cancer. *J Natl Cancer Inst* 2015;**107**(10):djv211.
6. Maroto P, Rini B. Molecular biomarkers in advanced renal cell carcinoma. *Clin Cancer Res* 2014;**20**:2060–71.
7. Srinivas PR, Kramer BS, Srivastava S. Trends in biomarker research for cancer detection. *Lancet Oncol* 2001;**2**:698–704.
8. Liu ZP, Wang Y, Zhang XS, *et al*. Network-based analysis of complex diseases. *IET Syst Biol* 2012;**6**:22–33.
9. Schwikowski B, Uetz P, Fields S. A network of protein-protein interactions in yeast. *Nat Biotechnol* 2000;**18**:1257–61.
10. Tong AH, Lesage G, Bader GD, *et al*. Global mapping of the yeast genetic interaction network. *Science* 2004;**303**:808–13.
11. Lee TI, Rinaldi NJ, Robert F, *et al*. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* 2002;**298**:799–804.
12. Breitkreutz A, Choi H, Sharom JR, *et al*. A global protein kinase and phosphatase interaction network in yeast. *Science* 2010;**328**:1043–6.
13. Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004;**5**:101–13.
14. Chuang HY, Lee E, Liu YT, *et al*. Network-based classification of breast cancer metastasis. *Mol Syst Biol* 2007;**3**:140.
15. Dehmer M, Mueller LA, Emmert-Streib F. Quantitative network measures as biomarkers for classifying prostate cancer disease states: a systems approach to diagnostic biomarkers. *PLoS One* 2013;**8**:e77602.
16. Patel VN, Gokulrangan G, Chowdhury SA, *et al*. Network signatures of survival in glioblastoma multiforme. *PLoS Comput Biol* 2013;**9**:e1003237.
17. Roy J, Winter C, Isik Z, *et al*. Network information improves cancer outcome prediction. *Brief Bioinform* 2014;**15**:612–25.
18. He D, Liu ZP, Chen L. Identification of dysfunctional modules and disease genes in congenital heart disease by a network-based approach. *BMC Genomics* 2011;**12**:592.
19. Li J, Roebuck P, Grunewald S, *et al*. SurvNet: a web server for identifying network-based biomarkers that most correlate with patient survival data. *Nucleic Acids Res* 2012;**40**:W123–6.
20. Nibbe RK, Koyuturk M, Chance MR. An integrative -omics approach to identify functional sub-networks in human colorectal cancer. *PLoS Comput Biol* 2010;**6**:e1000639.

21. Tao H, Guo S, Ge T, *et al*. Depression uncouples brain hate circuit. *Mol Psychiatry* 2013;**18**:101–11.
22. Wang YC, Chen BS. A network-based biomarker approach for molecular investigation and diagnosis of lung cancer. *BMC Med Genomics* 2011;**4**:2.
23. Li C, Li H. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* 2008;**24**: 1175–82.
24. Winter C, Kristiansen G, Kersting S, *et al*. Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLoS Comput Biol* 2012;**8**:e1002511.
25. Hilario M, Kalousis A. Approaches to dimensionality reduction in proteomic biomarker studies. *Brief Bioinform* 2008;**9**: 102–18.
26. Liu JJ, Cutler G, Li WX, *et al*. Multiclass cancer classification and biomarker discovery using GA-based algorithms. *Bioinformatics* 2005;**21**:2691–7.
27. Morrison JL, Breitling R, Higham DJ, *et al*. GeneRank: using search engine technology for the analysis of microarray experiments. *BMC Bioinformatics* 2005;**6**:233.
28. Chen J, Aronow BJ, Jegga AG. Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinformatics* 2009;**10**:73.
29. Krauthammer M, Kaufmann CA, Gilliam TC, *et al*. Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. *Proc Natl Acad Sci USA* 2004;**101**:15148–53.
30. Wu X, Jiang R, Zhang MQ, *et al*. Network-based global inference of human disease genes. *Mol Syst Biol* 2008;**4**:189.
31. George RA, Liu JY, Feng LL, *et al*. Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Res* 2006;**34**:e130.
32. Kohler S, Bauer S, Horn D, *et al*. Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet* 2008;**82**:949–58.
33. Franke L, van Bakel H, Fokkens L, *et al*. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet* 2006;**78**: 1011–25.
34. Lage K, Karlberg EO, Storling ZM, *et al*. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol* 2007;**25**:309–16.
35. Ma X, Lee H, Wang L, *et al*. CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data. *Bioinformatics* 2007;**23**:215–21.
36. Aerts S, Lambrechts D, Maity S, *et al*. Gene prioritization through genomic data fusion. *Nat Biotechnol* 2006;**24**:537–44.
37. Hamosh A, Scott AF, Amberger JS, *et al*. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 2005;**33**: D514–17.
38. Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, *et al*. The BioGRID interaction database: 2015 update. *Nucleic Acids Res* 2015;**43**:D470–8.
39. Stark C, Breitkreutz BJ, Reguly T, *et al*. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 2006;**34**: D535–9.
40. Szklarczyk D, Franceschini A, Wyder S, *et al*. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* 2015;**43**:D447–52.
41. von Mering C, Huynen M, Jaeggi D, *et al*. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res* 2003;**31**:258–61.
42. Dong J, Hu Z, Wu C, *et al*. Association analyses identify multiple new lung cancer susceptibility loci and their interactions with smoking in the Chinese population. *Nat Genet* 2012;**44**:895–9.
43. Hsu YC, Yuan S, Chen HY, *et al*. A four-gene signature from NCI-60 cell line for survival prediction in non-small cell lung cancer. *Clin Cancer Res* 2009;**15**:7309–15.
44. Tomida S, Koshikawa K, Yatabe Y, *et al*. Gene expression-based, individualized outcome prediction for surgically treated lung cancer patients. *Oncogene* 2004;**23**:5360–70.
45. Jostins L, Ripke S, Weersma RK, *et al*. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 2012;**491**:119–24.
46. Szuhai K, Ijszenga M, de Jong D, *et al*. The NFATc2 gene is involved in a novel cloned translocation in a Ewing sarcoma variant that couples its function in immunology to oncology. *Clin Cancer Res* 2009;**15**:2259–68.
47. Carrilho C, Alberto M, Buane L, *et al*. Keratins 8, 10, 13, and 17 are useful markers in the diagnosis of human cervix carcinomas. *Hum Pathol* 2004;**35**:546–51.
48. Umayahara K, Numa F, Suehiro Y, *et al*. Comparative genomic hybridization detects genetic alterations during early stages of cervical cancer progression. *Genes Chromosomes Cancer* 2002;**33**:98–102.
49. Wong YF, Cheung TH, Tsao GS, *et al*. Genome-wide gene expression profiling of cervical cancer in Hong Kong women by oligonucleotide microarray. *Int J Cancer* 2006;**118**:2461–9.
50. Diegmann J, Junker K, Gerstmayer B, *et al*. Identification of CD70 as a diagnostic biomarker for clear cell renal cell carcinoma by gene expression profiling, real-time RT-PCR and immunohistochemistry. *Eur J Cancer* 2005;**41**:1794–801.
51. Shoji K, Murayama T, Mimura I, *et al*. Sperm-associated antigen 4, a novel hypoxia-inducible factor 1 target, regulates cytokinesis, and its expression correlates with the prognosis of renal cell carcinoma. *Am J Pathol* 2013;**182**:2191–203.
52. Kerr JF, Winterford CM, Harmon BV. Apoptosis. Its significance in cancer and cancer therapy. *Cancer* 1994;**73**:2013–26.
53. Williams GT. Programmed cell death: apoptosis and oncogenesis. *Cell* 1991;**65**:1097–8.
54. Heller F, Florian P, Bojarski C, *et al*. Interleukin-13 is the key effector Th2 cytokine in ulcerative colitis that affects epithelial tight junctions, apoptosis, and cell restitution. *Gastroenterology* 2005;**129**:550–64.
55. Sartor RB. Current concepts of the etiology and pathogenesis of ulcerative colitis and Crohn's disease. *Gastroenterol Clin North Am* 1995;**24**:475–507.
56. Murata T, Noguchi PD, Puri RK. IL-13 induces phosphorylation and activation of JAK2 Janus kinase in human colon carcinoma cell lines: similarities between IL-4 and IL-13 signaling. *J Immunol* 1996;**156**:2972–8.
57. McCance DJ, Kopan R, Fuchs E, *et al*. Human papillomavirus type 16 alters human epithelial cell differentiation in vitro. *Proc Natl Acad Sci USA* 1988;**85**:7169–73.
58. Bosch FX, Lorincz A, Munoz N, *et al*. The causal relation between human papillomavirus and cervical cancer. *J Clin Pathol* 2002;**55**:244–65.
59. Lee MY, Chou CY, Tang MJ, *et al*. Epithelial-mesenchymal transition in cervical cancer: correlation with tumor progression, epidermal growth factor receptor overexpression, and snail up-regulation. *Clin Cancer Res* 2008;**14**:4743–50.
60. Motzer RJ, Bander NH, Nanus DM. Renal-cell carcinoma. *N Engl J Med* 1996;**335**:865–75.
61. Lazzeroni L, Owen A. Plaid models for gene expression data. *Stat Sin* 2002;61–86.

62. Li G, Ma Q, Tang H, *et al*. QUBIC: a qualitative biclustering algorithm for analyses of gene expression data. *Nucleic Acids Res* 2009;**37**:e101.
63. Hochreiter S, Bodenhofer U, Heusel M, *et al*. FABIA: factor analysis for bicluster acquisition. *Bioinformatics* 2010;**26**:1520–7.
64. Vanunu O, Magger O, Ruppin E, *et al*. Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol* 2010;**6**:e1000641.
65. Hofree M, Shen JP, Carter H, *et al*. Network-based stratification of tumor mutations. *Nat Methods* 2013;**10**:1108–15.
66. Vandin F, Upfal E, Raphael BJ. De novo discovery of mutated driver pathways in cancer. *Genome Res* 2012;**22**:375–85.
67. Zhao J, Zhang S, Wu LY, *et al*. Efficient methods for identifying mutated driver pathways in cancer. *Bioinformatics* 2012;**28**: 2940–7.
68. Keshava Prasad TS, Goel R, Kandasamy K, *et al*. Human protein reference database–2009 update. *Nucleic Acids Res* 2009; **37**:D767–72.
69. Masuda N, Porter MA, Lambiotte R. Random walks and diffusion on networks. ArXiv e-prints 2016; arXiv:1612.03281v2 [physics.soc-ph].
70. Lv Q, Cao P, Cohen E, *et al*. Search and replication in unstructured peer-to-peer networks. In: *Proceedings of the 16th International Conference on Supercomputing*. ACM, New York, NY, USA 2002, 84–95.
71. Page L, Brin S, Motwani R, *et al*. The PageRank citation ranking: bringing order to the web. Technical Report. Stanford InfoLab. Stanford, CA, USA. 1999.
72. Masuda N, Kawamura Y, Kori H. Impact of hierarchical modular structure on ranking of individual nodes in directed networks. *N J Phys* 2009;**11**:113002.
73. Delvenne JC, Yaliraki SN, Barahona M. Stability of graph communities across time scales. *Proc Natl Acad Sci USA* 2010;**107**: 12755–60.
74. Lambiotte R, Delvenne JC, Barahona M. Random walks, Markov processes and the multiscale modular organization of complex networks. *IEEE Trans Netw Sci Eng* 2014;**1**:76–90.
75. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 2005;**21**: 3448–9.
76. Subramanian A, Tamayo P, Mootha VK, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005; **102**:15545–50.
77. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 2000;**28**:27–30.
78. Liberzon A, Subramanian A, Pinchback R, *et al*. Molecular Signatures Database (MSigDB) 3.0. *Bioinformatics* 2011;**27**: 1739–40.