Supplementary Materials:

Discovering cooperative biomarkers for heterogeneous complex disease diagnoses

Duanchen Sun[1,2,3], Xianwen Ren[4], Eszter Ari[5], Tamas Korcsmaros[6,7], Peter Csermely[8], Ling-Yun Wu[1,2,3,*]

[1]Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China, [2]National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Sciences, Beijing 100190, China, [3]School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China, [4]Biodynamic Optical Imaging Center, Peking University, Beijing 100187, China, [5]Department of Genetics, Eötvös Loránd University, 1117 Budapest, Hungary, [6]Gut Health and Food Safety Programme, Institute of Food Research, Norwich NR4 7UA, UK, [7]Earlham Institute, Norwich Research Park, Norwich, NR4 7UH, UK, [8]Department of Medical Chemistry, Semmelweis University, H-1428 Budapest, Hungary.

*To whom correspondence should be addressed. Tel: +86-10-82541370; Email: lywu@amss.ac.cn

1. The simulation study

We compared the performance of MarkRank and NetRank in simulated datasets. The goal of the simulation study was to test whether MarkRank can discriminate the pre-set target genes (positive genes) from the remaining genes (negative genes); i.e., these target genes are expected to be on the top positions of the ranking list. The expression patterns of these pre-set target genes were well controlled and can partially mimic the potential mechanisms of disease pathogenesis in different ways. In our study, we simulated three types of expression patterns. The whole workflow of the simulation study was introduced in the main text.

For each simulation, a network G with 100 nodes and 274 edges was randomly extracted from the Human Protein Reference Database (HPRD) PPI network, which contains 9453 genes with 36874 interactions. The expression data of 50 samples were simulated, and labeled as two groups of equal size. The size of pre-set target nodes in subnetwork S_i was 10 for the first and the second scenarios. In the third scenario, each of the two disjoint subnetworks that exhibited complementary information had 5 nodes. The expression profile for a sample was drawn from a multivariate Gaussian distribution whose mean vector followed a univariate Gaussian distribution, and the covariances for adjacent gene pairs in G were relatively higher than those of non-adjacent pairs. Specifically, the mean vector was $\mu = (\mu_1, \mu_2, \cdots, \mu_{100})$ where $\mu_i \sim N(5,1)$ independently. The covariance matrix Σ was set as $\sigma_{i,j} \sim N(4,1)$ if node pair v_i, v_j was connected in G and $\sigma_{i,j} \sim N(2,1)$ otherwise. The variances of variables were gradually increased until Σ became a positive definite matrix. Finally, the sample expression profiles were drawn from the multivariate Gaussian distribution $N_{100}(\mu, \Sigma)$.

In the simulation, a parameter ρ was set to control the degree of differential expression. The fold-change of up-regulated gene expression levels followed $N(\rho, 0.1)$ in all three types of expression patterns. Each simulation test was repeated 100 times for each ρ to obtain a comparable result.

The performance was measured by the area under the curve (AUC). The pre-set target genes were the positive class, and all the remaining genes were the negative class. The AUC was computed using the prioritization score of each gene and the ground truth label (positive or negative). The averaged AUC plus/minus one-fold standard deviation as a function of differential expression degree ρ was used to compare the performance of MarkRank and NetRank.

2. Data description and pre-processing

In our study, we integrated two types of data, gene expression profiles and PPI networks to prioritize network biomarkers via MarkRank.

In our work, four microarray expression datasets were downloaded from the Gene Expression Omnibus (GEO) repository http://www.ncbi.nlm.nih.gov/geo/ (accession numbers GSE4115, GSE11223, GSE9750, GSE36895, respectively) for real dataset analysis.

The gene expression profiles for lung cancer (GSE4115) were collected from histologically normal bronchial epithelium of smokers during clinical bronchoscopy (Spira, et al., 2007) using the Affymetrix HGU133a GeneChips platform. We combined the original primary and prospective datasets, which gave a total of 97 and 90 smokers with and without lung cancer, respectively. For ulcerative colitis (GSE11223), *Noble et al.* (Noble, et al., 2008) performed transcriptional profiling of colon epithelial paired endoscopic biopsies, which were taken from 5 specific anatomical locations for RNA extraction and histology. We used the uninflamed samples in each cohort to obtain a balanced classification with 66 ulcerative colitis patients and 69 healthy control donors. The cervical cancer (GSE9750) dataset contains 24 normal cervixes versus 33 cervical cancer samples, and the detailed information can be found in Ref. (Scotto, et al., 2008). For renal cell carcinoma (RCC, GSE36895), the RNA of clear-cell renal cell carcinoma primary tumors, tumors growing in immunodeficient mice (tumor grafts), and normal kidney cortices were labeled and hybridized to Affymetrix Human Genome U133 Plus 2.0 arrays (Pena-Llopis, et al., 2012). We used the paired expression profiles of 23 clear-cell RCC patients and their related normal cortex for further analysis. For all expression datasets, we averaged the expression values of the probes mapping on the same gene. The summaries of the detailed processed datasets are shown in Table 1 in the main text.

Protein-protein interaction (PPI) data were extracted from the Human Protein Reference Database (HPRD, http://www.hprd.org). The HPRD database provides considerable resources and integrated information for the human proteome, such as post-translational modifications, interaction networks and disease associations (Keshava Prasad, et al., 2009). The original PPI network contained 9453 genes with 36874 interactions. After mapping the common genes present in both PPI data and each of the four expression profiles as described above, we further restricted our study to the largest connected component of the refined network.

Two other popular biological molecular network databases, BioGRID (Chatr-Aryamontri, et al., 2015; Stark, et al., 2006) and STRING (Szklarczyk, et al., 2015; von Mering, et al., 2003), were also used to test the robustness of the MarkRank algorithm in cross-validation. As for the PPI extracted from the BioGRID (Homo_sapiens, version 3.4.141, https://thebiogrid.org), we only retained the protein associations detected in the low-throughput experimental system. The processed interaction network contains 12791 genes with 79660 interactions. For the PPI maintained in the STRING database (http://string-db.org/), we used the R package 'STRINGdb' compiled in the Bioconductor (http://bioconductor.org/packages/STRINGdb/), which provided an R interface to the STRING protein-protein interactions database. The threshold of the interaction score was set to 900 (under version: STRING v10, Homo sapiens). The processed interaction network contains 20457 genes with 216960 interactions. Similar to the former procedure, we restricted our study to the largest connected component of the network derived from the common genes in the expression dataset and the biological molecular network dataset (red number in Table S1). The summaries of the detailed overlapped genes are shown in Table S1. (Here, we show the detailed information for the lung cancer and ulcerative colitis dataset, for we only executed the following Monte Carlo cross-validation procedure on these two datasets. The reason is explained in the next section.)

Table S1: The summaries of the detailed overlapped genes in the lung cancer and ulcerative colitis datasets.

Dataset	Original genes	HPRD	HPRD*	BioGRID	BioGRID*	STRING	STRING*
Lung Cancer	12493	7608	7244	8157	7713	12459	7963
Ulcerative Colitis	10506	5055	4244	5178	4360	9680	5295

Notes: The columns HPRD, BioGRID, and STRING show the number of common genes in the expression data and the respective network. The columns HPRD*, BioGRID* and STRING* are the number of genes in the largest connected component (LCC) of the respective refined network, which are same as the LCC genes in the main text.

For one expression dataset, the relationship of the research genes across three published biological molecular networks is shown in Fig. S1. The common genes in the three biological networks were 4294 (53.9%~59.3%) and 2247 (42.4%~52.9%) in the lung cancer and ulcerative colitis datasets, respectively, which indicated the difference in these networks.

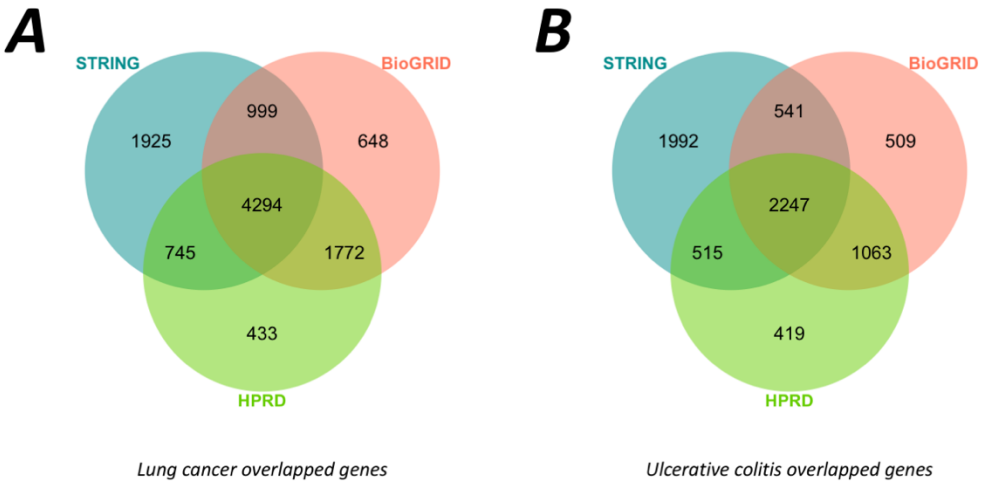

Fig. S1: The overlapped genes in three published biological molecular networks for the (A) lung cancer and (B) ulcerative colitis datasets.

3. The Monte Carlo cross-validation procedure

To compare MarkRank with other related ranking methods, we followed the multiple random validation procedure as described in Ref. (Michiels, et al., 2005). We wanted to test the classification capacity of our identified MarkRank genes to discriminate the disease and normal samples. In our study, the averaged area under the curve (AUC) in validation sets as a function of the corresponding percentage of the training set was used as an evaluation measurement. Note that the AUC derived from the Monte Carlo cross-validation here reflects the performance of classifying the disease and control samples for each ranking method, whereas the AUC measured in the simulation study reflects the performance of identifying the pre-set target genes from the remaining genes.

The workflow of the Monte Carlo cross-validation procedure was as follows:

1) The whole mapped expression dataset was randomly split into training and testing sets according to a fixed percentage. The training set was used for gene ranking, feature selection and classifier training, whereas the testing set was used for validation without information leakage.
2) Different methods for ranking genes (see below) were performed on the training set. According to the related gene scores computed by one method, the top 10 ranked genes were selected as the biomarkers.
3) A classifier based on the identified biomarkers was trained using the training dataset.
4) A probabilistic score was assigned to each sample in the testing dataset by the trained classifier based on the identified biomarkers. Together with the ground truth information (the true sample labels on the testing dataset), the AUC was computed to evaluate different methods.

The above steps were repeated 200 times for each partition percentage. For MarkRank, the gene cooperation network G_2 and the prior information was constructed using only the training set, guaranteeing fair comparison in the Monte Carlo cross-validation procedure and preventing information leakage.

Random forest is a powerful ensemble learning method for supervised classification (Breiman, 2001) and has had many applications to biological problems in the past decades (Diaz-Uriarte and Alvarez de Andres, 2006). It corrects overfitting, an undesired property of single decision trees, by constructing a multitude of decision trees. In our study, we used the *randomForest* function provided in the R package randomForest (Version: 4.6-12 from https://cran.r-project.org/web/packages/randomForest/) to train the classifier. In our setting, the number of trees (parameter *ntree*) was set to 1000, and the default values were used for other parameters.

In this study, we used the following ranking methods to compare their performance: (i) Mutual Information (MI), the mutual information of single gene was computed using the R package mpmi as described above; (ii) The Student's t-test, the features were ranked using the p-values of t-test; (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; (vi) NetRank algorithm (Winter, et al., 2012); and (vii) MarkRank algorithm. The prior information for NetRank was the same as for MarkRank as mentioned above. Notably, NetRank itself needs an additional inner cross-validation loop to learn the model parameter α, as described in (Winter, et al., 2012), which is very time-consuming. Instead, we recorded the related performance using different values of α and λ ranging from 0 to 1 in steps of 0.1. The results of parametric sensitivity analysis on the performance of different α, λ selection can be found in Fig. S2. In addition, we also computed the related results using α equal to 0.2 and 0.8 in the NetRank algorithm to obtain a more unambiguous comparison. The random selection of genes was

also taken into consideration, which was repeated 5000 times in each percentage of the Monte Carlo cross-validation.

Our Monte Carlo cross-validation procedure was executed on the lung cancer and ulcerative colitis datasets, which have a moderate classification difficulty. Because the classes in cervical cancer and renal cell carcinoma datasets are easy to classify by all methods (see Fig. S5, S9 and S10), we did not use these two datasets to test MarkRank performance via the Monte Carlo cross-validation procedure.

4. MarkRank performance with different parameters

MarkRank has two model parameters, α and λ. In our work, we set $\alpha = 0.8$ and $\lambda = 0.2$ as default parameter settings according to the results of a sufficient quantity of pilot simulation tests. To make a reasonable comparison with NetRank, which needs an additional inner cross-validation loop to train the model parameter α, we recorded the related performance using different values of α and λ ranging from 0 to 1 in steps of 0.1 in the above Monte Carlo cross-validation procedure on the lung cancer and ulcerative colitis datasets. The results are shown in Fig. S2.

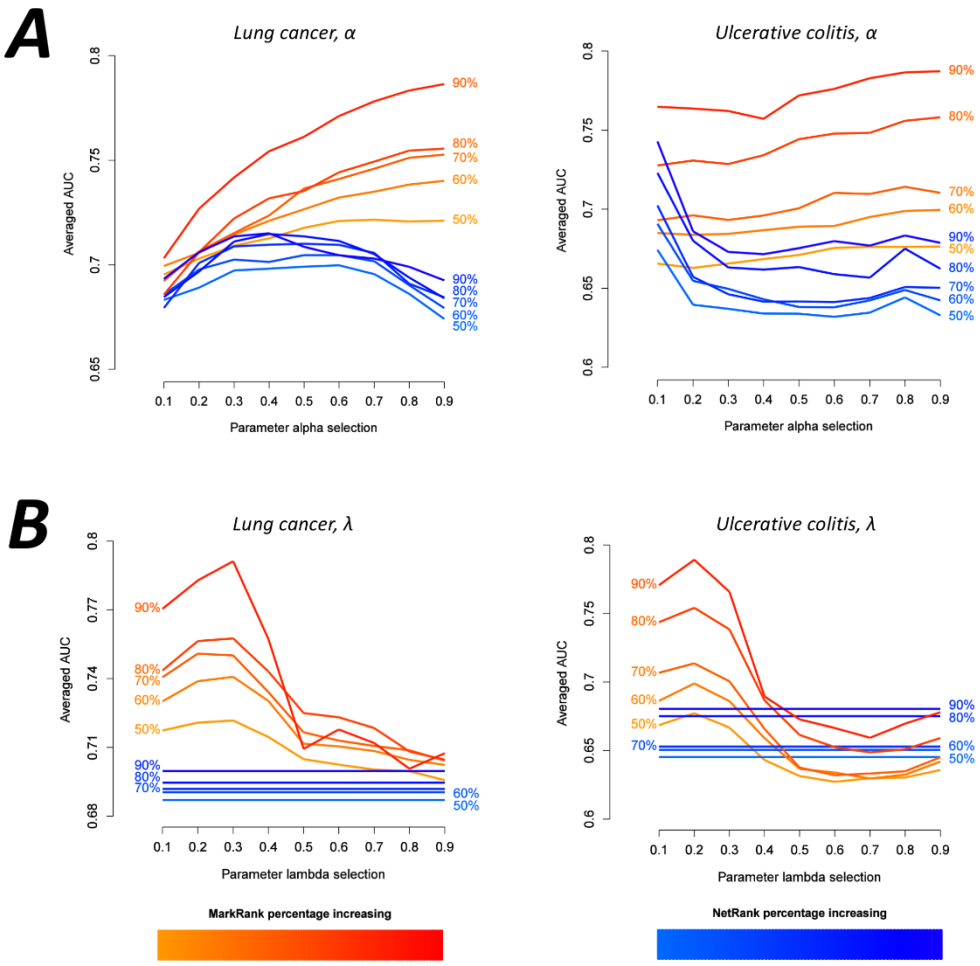

Fig. S2: The performance of different parameter selections in the Monte Carlo cross-validation procedure. (A) The performance of parameter α selection on the lung cancer (left) and the ulcerative colitis datasets (right). (B) The performance of parameter λ selection on the lung cancer (left) and the ulcerative colitis datasets (right).

In each panel, we focused on the tendency of the averaged AUC as the related parameter increasing in a fixed percentage of the training dataset partition (50% to 90%). The classifier and the top gene number were the same as in the Monte Carlo cross-validation.

In random walk-based models, parameter α balances the effect of prior information and the influence of networks. A smaller α lays more emphasis on the prior information provided by the user. In the lung cancer and ulcerative colitis datasets, the tendency of the averaged AUC computed via NetRank was different (Fig. S2A). The tendency reached a peak with $\alpha =$ 0.3 to 0.4 on the lung cancer dataset for each partition. On the ulcerative colitis dataset, the averaged AUC first decreased when $\alpha \leq 0.4$ and then started to level off, which was consistent with the fact that the performance of NetRank was inferior

to PCC in the Monte Carlo cross-validation procedure, since we used PCC as the prior information in the random walk model. In comparison to NetRank, MarkRank showed an increasing trend as a whole and had a higher averaged AUC in each partition.

λ is a specific model parameter in MarkRank that balances the relative importance of two networks. Larger λ inclines to lay more emphasis on G_1; e.g., the PPI network structure. From the results, we can clearly see that the performance of MarkRank asymptotically approached that of NetRank as λ increased, and a peak was reached when $\lambda =$ 0.2 to 0.3 for both datasets, which was in accordance with our parameter selection.

5. Performance of the Monte Carlo cross-validation trained using the top 30 ranked features

To exclude the dependency on the number of training features in the Monte Carlo cross-validation procedure, we used the top 30 ranked genes as a signature to train the classifier. Other parts of the Monte Carlo cross-validation procedure remained unchanged. The related performance is shown in Fig. S3.

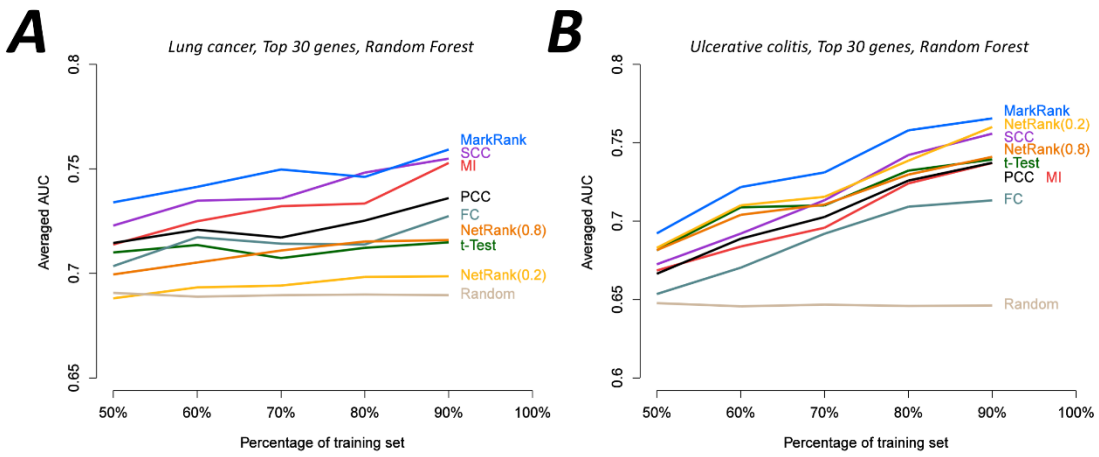

Fig. S3: The performance of the Monte Carlo cross-validation using the top 30 ranked features on the (A) lung cancer and the (B) ulcerative colitis datasets. The abbreviations for each method are the same as in Fig. 4.

From the results, we can see that MarkRank, which still outperformed traditional methods on both the lung cancer and the ulcerative colitis datasets, had little dependency on the trained feature number. The gap between MarkRank and the suboptimal method was narrowed when the top 30 ranked genes were used as a signature. In addition, the averaged AUC dropped approximately 5% using MarkRank, whereas it increased to different extents in the traditional methods, which means that the key genes with strong discriminative power identified by traditional methods were lower-ranking. Generally speaking, biomarkers with unlimited feature number are inconvenient for the analysis of their biological function and the underlying etiology mechanism. An excessive number of biomarker genes is also inapplicable for clinical prediction in practice. Therefore, MarkRank being able to identify the key genes that ranked at the top of the list has a considerable practical value. The biomarkers identified via MarkRank can further contribute to the understanding of disease mechanisms, diagnosis and therapy.

6. Performance of the Monte Carlo cross-validation trained by other classifiers

To exclude the effect of the classifier in the Monte Carlo cross-validation procedure and compare MarkRank with other related ranking methods in a broader perspective, we used the following alternative classifiers to replace the random forest classifier in the Monte Carlo cross-validation procedure:

(1) Support Vector Machine (SVM).

(2) Naïve Bayes.

Other parts of the Monte Carlo cross-validation procedure remained unchanged.

In our study, we used the *svm* and *naiveBayes* functions provided in the R package e1071 (version 1.6-7, https://cran.r-project.org/web/packages/e1071/) to obtain the corresponding classifiers. For the SVM classifier, we used linear and radial basis function (RBF) kernels. We set the parameter $C = 1000$ for the linear kernel and $C = 100,\ \gamma = 0.01$ for the RBF kernel. The results of randomly selected genes were also taken into consideration, which was repeated 5000 times in each percentage of the Monte Carlo cross-validation. The related results are shown in Fig. S4.

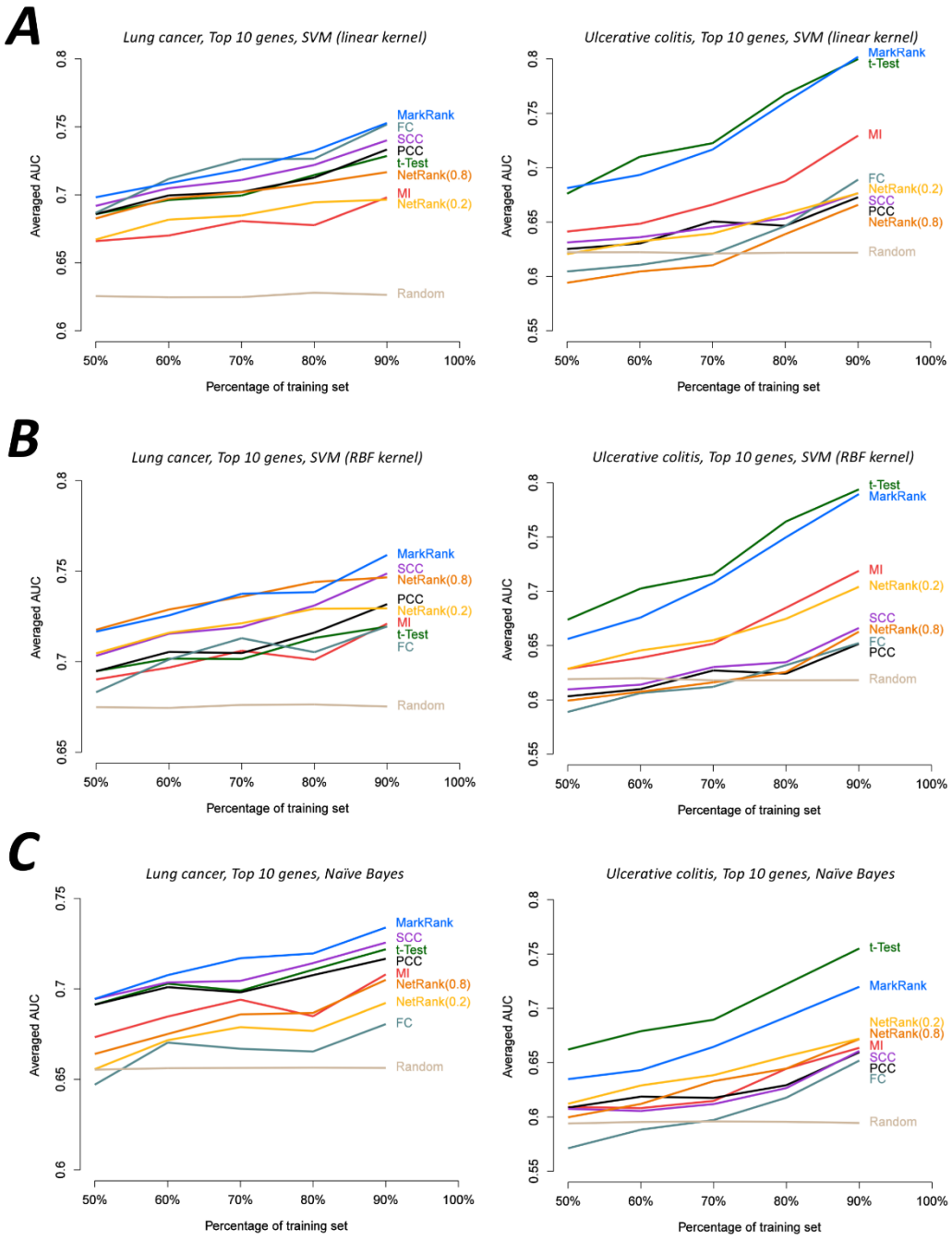

Fig. S4: The performance of other trained classifiers. (A) The performance of the SVM classifier (linear kernel). (B) The performance of the SVM classifier (RBF kernel). (C) The performance of the Naïve Bayes classifier. The corresponding Monte Carlo cross-validation was executed on the lung cancer (left) and the ulcerative colitis datasets (right). The abbreviations for each method are the same as in Fig. 4.

The results show that MarkRank, which ranked first or second place, had little dependency on the selection of classifier and model parameter.

Compared to the randomForest classifier, the performance of random feature selection decreased when trained using either the SVM or Naïve Bayes classifier on both datasets, which was related to the attributes of the classifier itself. The performance of the SVM classifier was relative to the selection of kernel functions and model parameters. For lung cancer, the highest averaged AUC of MarkRank was nearly 0.75, which was a little lower than that of the randomForest classifier. However, there was a

measurable increase for other traditional methods. For the ulcerative colitis datasets, MarkRank and Student's t-test were still the two top-ranked methods when compared to other methods whose average AUC decreased when compared to the *randomForest* classifier. The performance of the Naïve Bayes classifier was inferior to other classifiers whose average AUC decreased 5% in both datasets. In conclusion, these results suggest that the MarkRank method was little affected by the selection of classifier or model parameter and showed a consistent dominance over other traditional methods.

7. Results of the Monte Carlo cross-validation in the cervical cancer and the renal cell carcinoma datasets

In our work, we aimed at comparing our method with other related methods via the Monte Carlo cross-validation procedure on datasets that have a moderate classification difficulty. This comparison is meaningless when the related dataset is relatively easier to classify. To this end, we first examined the averaged AUC of traditional methods on each dataset to allow a preliminary filtration. The performance of the Monte Carlo cross-validation on the cervical cancer and renal cell carcinoma datasets is shown in Fig. S5.

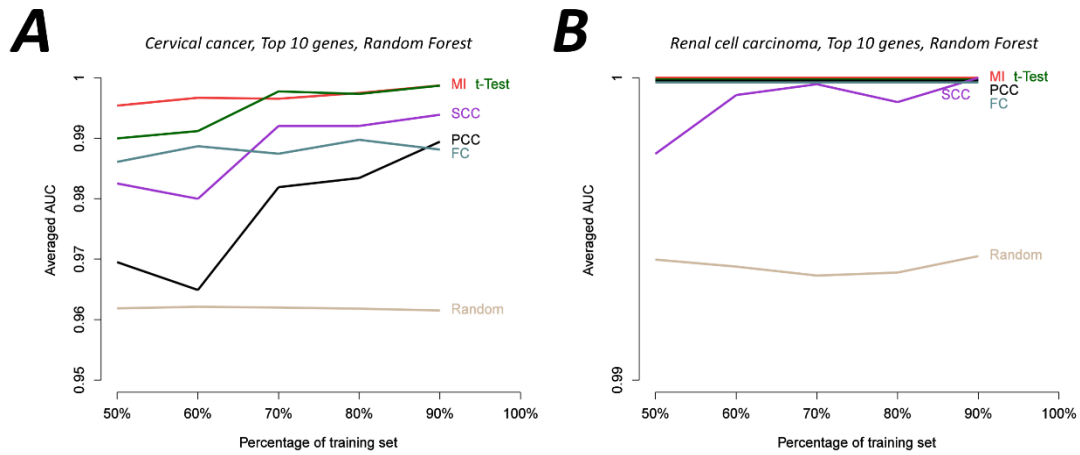

Fig. S5. The performance of five ranking methods in the Monte Carlo cross-validation on the (A) cervical cancer and (B) renal cell carcinoma datasets. The abbreviations for each method are the same as in Fig. 4.

From the results we can clearly see that the averaged AUC can reach 0.96 and 0.99 in the cervical cancer and renal cell carcinoma datasets, respectively, when using randomly selected features. Each traditional ranking method can perfectly classify each class when the percentage of training set was large enough. Therefore, we only executed the Monte Carlo cross-validation procedure on the lung cancer and ulcerative colitis datasets.

8. The performance of MarkRank on BioGRID and STRING networks

To exclude the effect of the network information and to test the robustness of the MarkRank algorithm, we also used two published biological molecular networks (BioGRID and STRING) to compare MarkRank with other state-of-the-art methods. The detailed information about these networks can be found in the *Data description and pre-processing* section.

In this section, we used a cross-validation procedure to compare each ranking method. First, we randomly selected 90% samples as the training set and the remaining 10% samples as the testing set. Second, for each ranking method, all genes were scored only using the information of the training set. Finally, a random forest classifier was trained using the top 10 genes and then used to predict the testing samples (disease or normal), which obtained an AUC score. The above cross-validation procedure was repeated 200 times to achieve a robust result, and the averaged AUC was used to evaluate each method. Note that the index in each randomly split was identical to each method, which can receive a more comparable result.

In addition, to show the robustness of our method across different sources, we also plotted the overlap of the top 100 MarkRank genes identified from all samples (training set + testing set) using three network databases.

The results are shown in Fig. S6.

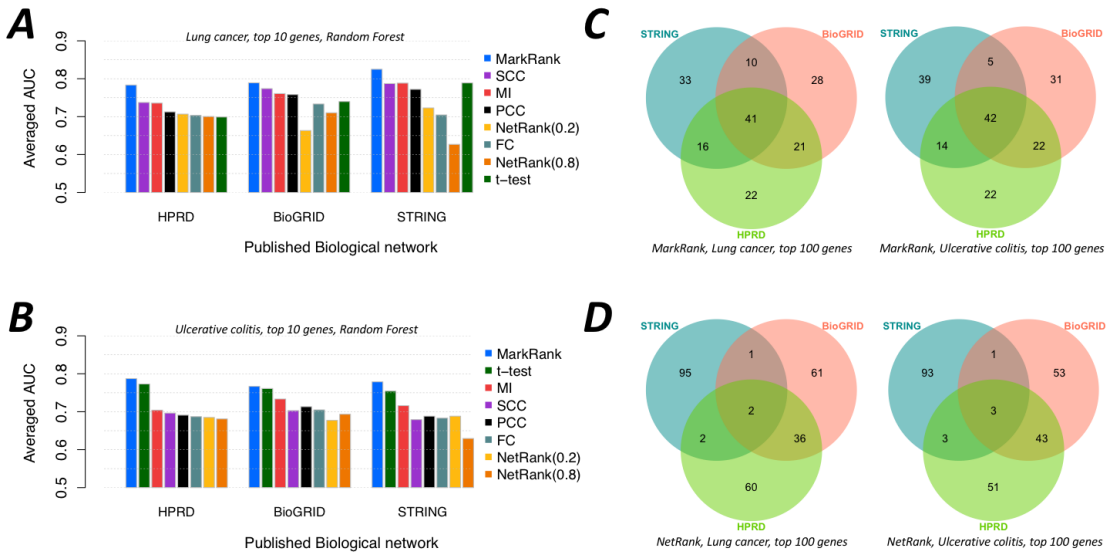

Fig. S6: The performance of MarkRank on other published network databases. The averaged AUC of MarkRank in cross-validation compared with other ranking methods in three published networks (HPRD, BioGRID, STRING) on the (A) lung cancer and (B) ulcerative colitis datasets. (C) The number of the overlapped genes in the top 100 MarkRank genes. (D) The number of the overlapped genes in the top 100 NetRank genes. The abbreviations for each method are the same as in Fig. 4.

The results show that the performance of MarkRank was consistently superior to other ranking methods in each biological network. The network has very little influence on the cross validation performance of MarkRank in both datasets. It is noteworthy that the second-best method, SCC for lung cancer and Student's t-test for ulcerative colitis, was only related to the expression dataset itself. These methods did not have a consistent result across different datasets. On the other hand, NetRank suffered a prominent influence from the selection of the network, and the number of overlapped genes identified via NetRank was only 2 and 3, respectively, which showed that NetRank had a much greater dependence on the network in contrast to MarkRank (overlapped gene: 41 and 42, respectively). In conclusion,

MarkRank was less affected by the selection of biological network, and the consistently prominent results indicated the robustness of MarkRank across either expression dataset or biological molecular network.

9. Principal component analysis of identified genes on four real datasets

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to map observations of possibly correlated variables from an original space to a new space, in which the variables are linearly uncorrelated over the dataset. In this study, we used PCA to execute the dimensionality reduction for visualizing. For each ranking method, we selected the top 10 genes as original gene signatures to reduce the dimension and kept the top two principal components to visualize the distribution of samples. Moreover, PCA on the original dataset (all genes) was also performed for comparison. The results are shown in Figs. S7-S10.

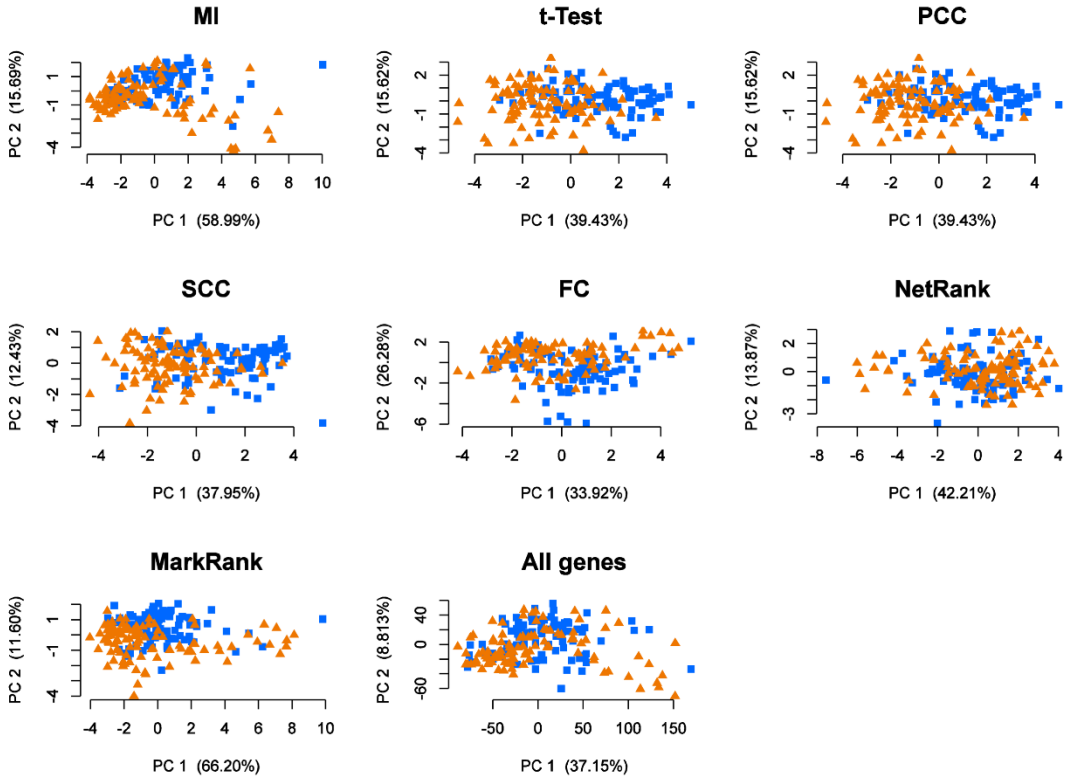

Fig. S7: Principal component analysis of each ranking method on the lung cancer dataset. The blue solid squares represent the normal samples, and the orange triangles represent the disease samples. The numbers in parentheses show the percentages of the contributions of the first and the second principal components.

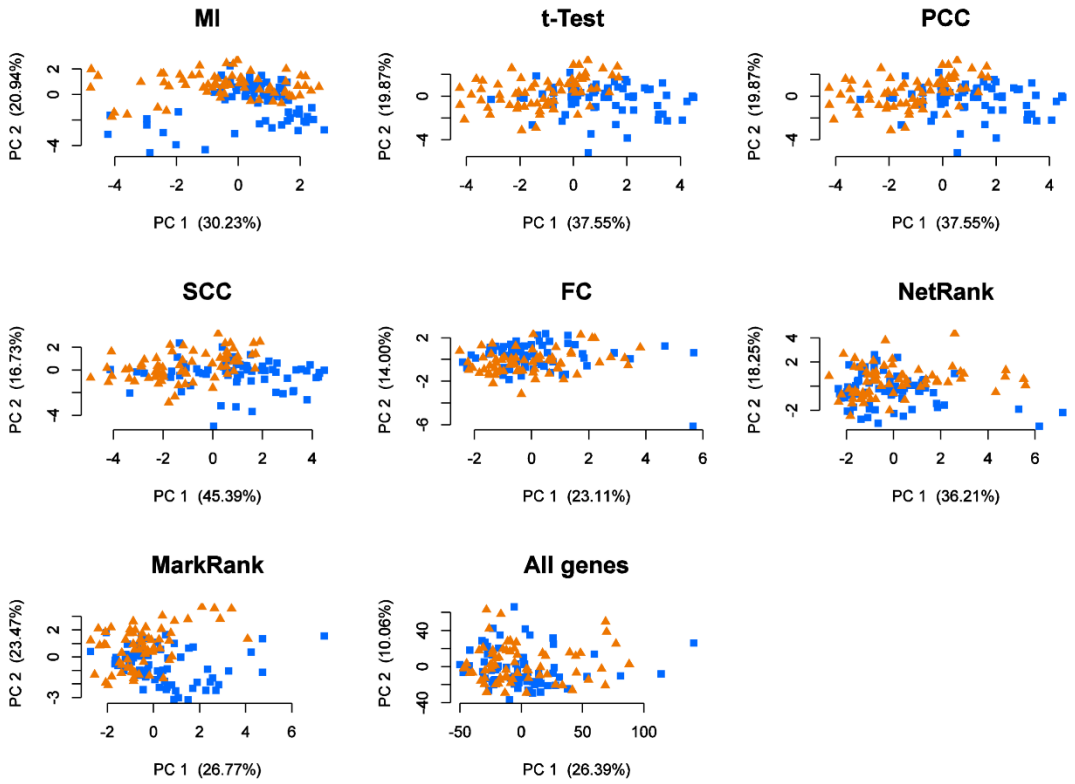

Fig. S8: Principal component analysis of each ranking method on the ulcerative colitis dataset. The annotations are the same as in Fig. S7.

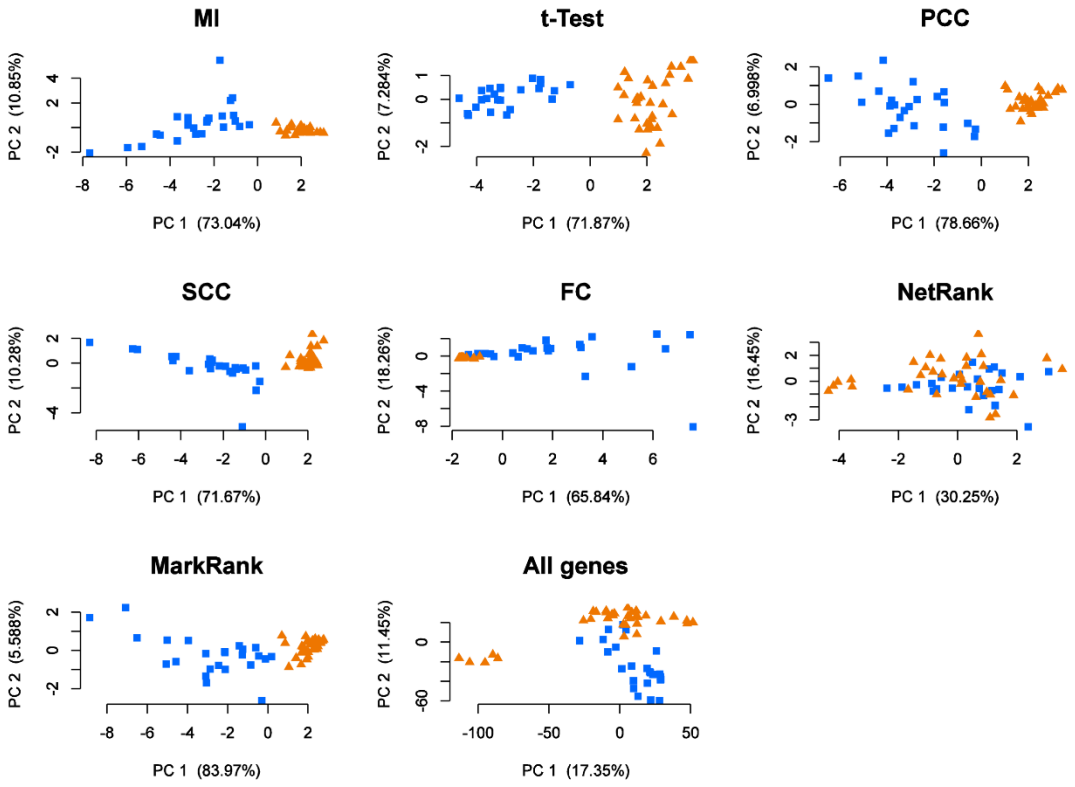

Fig. S9: Principal component analysis of each ranking method on the cervical cancer dataset. The annotations are the same as in Fig. S7.

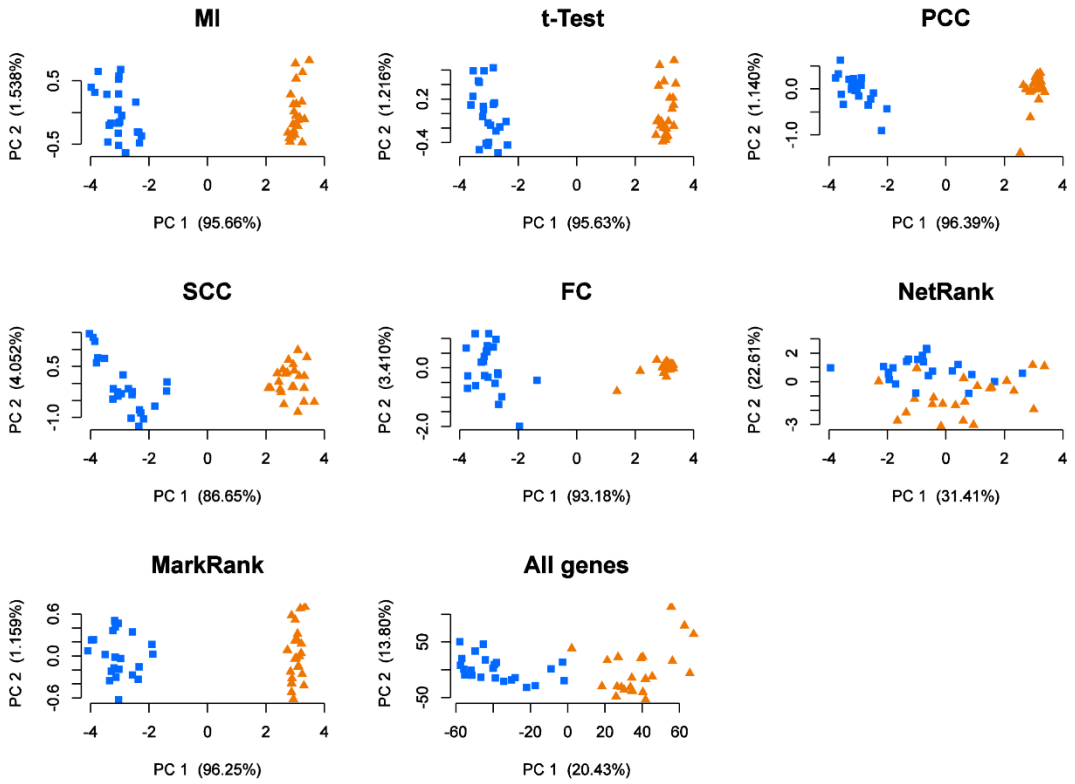

Fig. S10: Principal component analysis of each ranking method on the renal cell carcinoma dataset. The annotations are the same as in Fig. S7.

The boundary between the normal and disease samples using MarkRank genes as the signature was displayed more clearly than that of the original dataset for lung cancer and ulcerative colitis, which both exhibited a chaotic distribution. On the contrary, a well-defined boundary can be found for the cervical cancer and renal cell carcinoma datasets even when using all genes to execute the PCA dimensionality reduction, which was consistent with their high classification accuracy (Fig. S5) and can validate the classification capacity of each method from another perspective.

10. Statistical significance of gene connectivity

To plot the network view of a selected node set S, we first obtained the distance matrix of selected nodes in the PPI network, after which a minimum spanning tree (MST) was computed to construct a connected graph. Then, we linked the node pairs whose distance was less than or equal to 2. In this way, edges were grouped into three categories as described in Fig. 5.

Since MarkRank takes both the network structure and the discriminative power of cooperative gene combinations into consideration, we hope that the genes identified via MarkRank have relatively tighter connection structures when compared with the genes identified via the traditional methods. To test the statistical significance of the identified genes connectivity, we defined the gene connectivity statistics as

$$L(k) = \sum_{1 \le i < j \le n} I\big(d(v_i, v_j) \le k\big).$$

That is, the number of node pairs whose distance was shorter than a preset threshold k. Here, $n = |S|$ was the size of selected node set (in our work, we used $n = 30$, the same number of nodes in the network topological graphs (Fig. 5B-E)). $d(v_i, v_j)$ was the shortest distance between nodes v_i and v_j in the PPI network and the following condition was valid: $v_i, v_j \in S$. To compare the connectivity significance of our identified gene set with the gene sets identified by other ranking methods, random sampling was performed to obtain a null distribution. For each k=1,2,3,4,5, we random sampled node sets S' with the same node size 100,000 times and computed the related statistics. Finally, the p-value was reported as the statistical significance of network derived from S. The null distribution of $L(k)$ for each k and the summaries of statistical p-values for each ranking method are shown in Fig. S11 and Tables S2- S5.

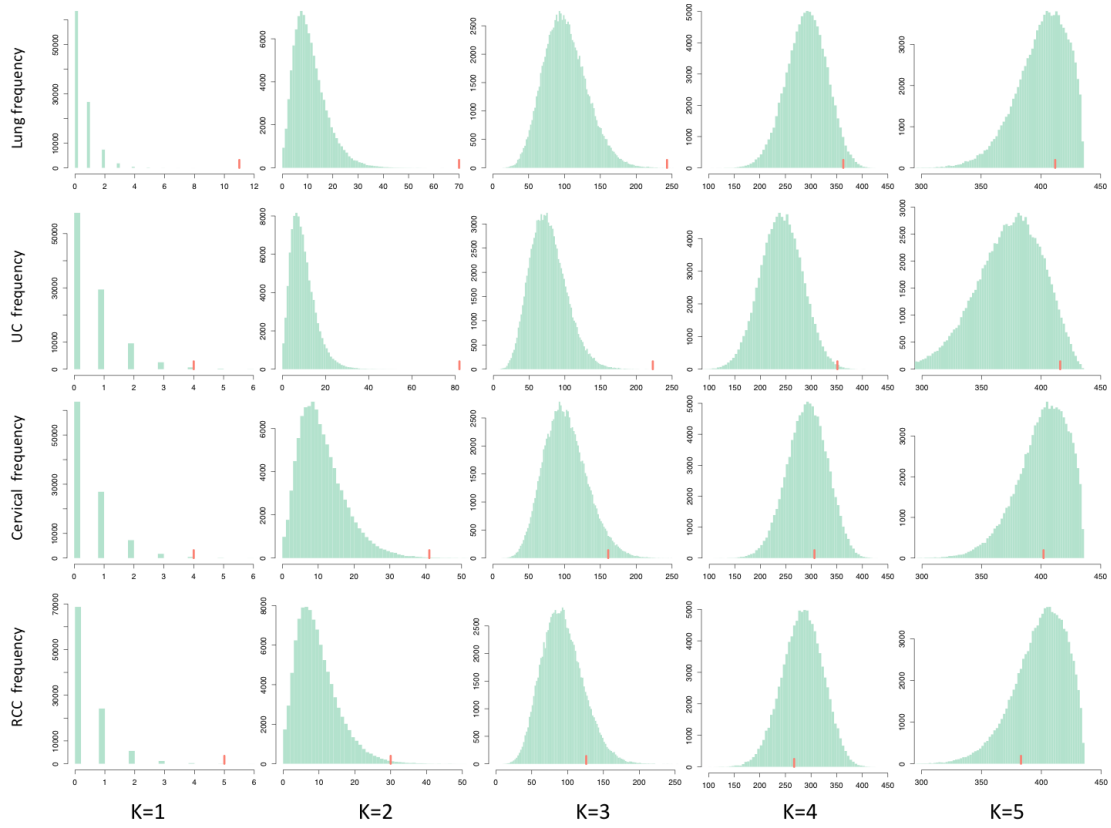

Fig. S11: The null distribution of $L(k)$ for each k on each dataset. For each selection, random sampling was performed 100,000 times to simulate the related distribution. The red bar in each histogram is corresponding $L(k)$ to the top 30 MarkRank genes. UC and RCC indicate ulcerative colitis and renal cell carcinoma, respectively.

Table S2: The statistical p-values of gene connectivity significance for each ranking method on the lung cancer dataset. The number in brackets after the statistical p-values are the $L(k)$ for corresponding methods. The red elements show the significant p-values. The abbreviations for each method are the same as in Fig. 4.

Methods	K=1	K=2	K=3	K=4	K=5
MI	0.09544 (2)	0.16766 (18)	0.42935 (106)	0.42418 (301)	0.10109 (427)
t-Test	1.00000 (0)	0.97248 (3)	0.94826 (57)	0.66657 (276)	0.64714 (396)
PCC	1.00000 (0)	0.97248 (3)	0.93296 (60)	0.55409 (288)	0.63150 (397)
SCC	0.36320 (1)	0.89588 (5)	0.77040 (79)	0.78087 (262)	0.93776 (365)
FC	0.36320 (1)	0.19955 (17)	0.63960 (90)	0.89891 (241)	0.94211 (364)
NetRank	0.00000 (100)	0.00000 (418)	0.00000 (435)	0.00000 (435)	0.00760 (435)
MarkRank	0.00000 (11)	0.00000 (70)	0.00006 (243)	0.03130 (363)	0.36164 (412)

Table S3: The statistical p-values of gene connectivity significance for each ranking method on the ulcerative colitis dataset. The number in brackets after the statistical p-values are the $L(k)$ for corresponding methods. The red elements show the significant p-values. The abbreviations for each method are the same as in Fig. 4.

Methods	K=1	K=2	K=3	K=4	K=5
MI	1.00000 (0)	0.33635 (12)	0.27164 (92)	0.35666 (258)	0.47900 (378)
t-Test	1.00000 (0)	0.12880 (17)	0.14451 (105)	0.11164 (293)	0.12379 (406)
PCC	1.00000 (0)	0.15701 (16)	0.18635 (100)	0.19620 (278)	0.15305 (403)
SCC	1.00000 (0)	0.33635 (12)	0.40542 (82)	0.40271 (253)	0.31217 (390)
FC	0.42468 (1)	0.46647 (10)	0.12271 (108)	0.13089 (289)	0.20816 (398)
NetRank	0.00000 (87)	0.00000 (381)	0.00000 (435)	0.00000 (435)	0.00032 (435)
MarkRank	0.00894 (4)	0.00000 (82)	0.00002 (223)	0.00331 (351)	0.05009 (416)

Table S4: The statistical p-values of gene connectivity significance for each ranking method on the cervical cancer dataset. The number in brackets after the statistical p-values are the $L(k)$ for corresponding methods. The red elements show the significant p-values. The abbreviations for each method are the same as in Fig. 4.

Methods	K=1	K=2	K=3	K=4	K=5
MI	0.09530 (2)	0.08031 (22)	0.36559 (111)	0.63059 (280)	0.50905 (404)
t-Test	0.00485 (4)	0.01674 (30)	0.09715 (143)	0.13545 (336)	0.59809 (399)
PCC	1.00000 (0)	0.11583 (20)	0.22916 (124)	0.24793 (320)	0.36039 (412)
SCC	0.09530 (2)	0.02492 (28)	0.14278 (135)	0.29163 (315)	0.32292 (414)
FC	1.00000 (0)	0.23163 (16)	0.66470 (88)	0.64021 (279)	0.58095 (400)
NetRank	0.00000 (99)	0.00000 (417)	0.00000 (435)	0.00000 (435)	0.00733 (435)
MarkRank	0.00485 (4)	0.00162 (41)	0.03673 (161)	0.37496 (306)	0.54512 (402)

Table S5: The statistical p-values of gene connectivity significance for each ranking method on the renal cell carcinoma dataset. The number in brackets after the statistical p-values are the $L(k)$ for corresponding methods. The red elements show the significant p-values. The abbreviations for each method are the same as in Fig. 4.

Methods	K=1	K=2	K=3	K=4	K=5
MI	0.07045 (2)	0.39170 (11)	0.21404 (116)	0.39854 (295)	0.64238 (394)
t-Test	0.31123 (1)	0.61245 (8)	0.62213 (83)	0.76380 (255)	0.78769 (383)

PCC	0.31123 (1)	0.69227 (7)	0.49667 (92)	0.80002 (250)	0.77659 (384)
SCC	0.07045 (2)	0.01971 (26)	0.04113 (148)	0.07944 (339)	0.54407 (400)
FC	1.00000 (0)	0.95401 (3)	0.95696 (49)	0.92194 (226)	0.71363 (389)
NetRank	0.00000 (104)	0.00000 (420)	0.00000 (435)	0.00000 (435)	0.00480 (435)
MarkRank	0.00062 (5)	0.00791 (30)	0.13589 (126)	0.66756 (267)	0.78769 (383)

The results showed that the MarkRank genes had significant gene connectivity in the lung cancer, ulcerative colitis and cervical cancer datasets when compared with random sampling (largest p-value 0.03673) for the preset threshold distance of node pair parameter $k \leq 3$. A similar result was obtained for the renal cell carcinoma dataset for $k \leq 2$ (largest p-value 0.00791). However, the genes identified by traditional methods did not show a consistent significant performance in all datasets, whereas the genes identified via NetRank were prone to gather together on the network. These methods lay emphasis on either the discriminative power of genes or the network structure; hence, the results are consistent with our expectations. Contrasted with these methods, the genes identified via MarkRank, taking both aspects into consideration using an effective method, not only had a strong topological relationship in the PPI network but also had a superior classification accuracy than the other methods.

11. Topological properties of the genes identified by each method

MarkRank balances the network structure and the discriminative power of cooperative gene combinations. To compare MarkRank with other ranking methods from another perspective, we analyzed the topological properties of the corresponding identified genes. In this section, two main categories, node importance indexes (node degree, betweenness centrality) and module importance indexes (clustering coefficient, the number of connected components), were selected as the measurements of each identified gene set. All these indexes were computed in the original connected PPI network. The results are shown in Fig. S12-S15.

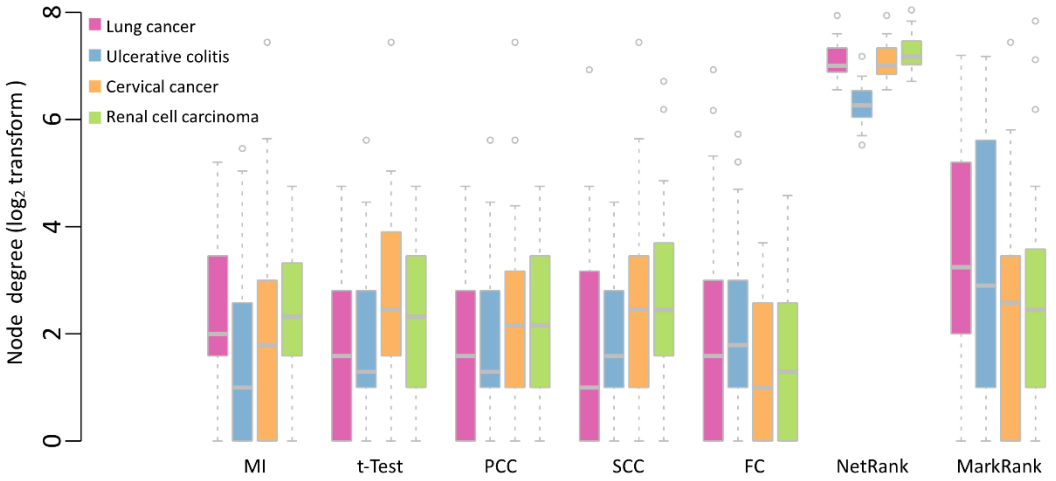

Fig. S12: The boxplot of node degrees for each gene on four datasets. The log2 transform was performed for clarity. The abbreviations for each method are the same as in Fig. 4.

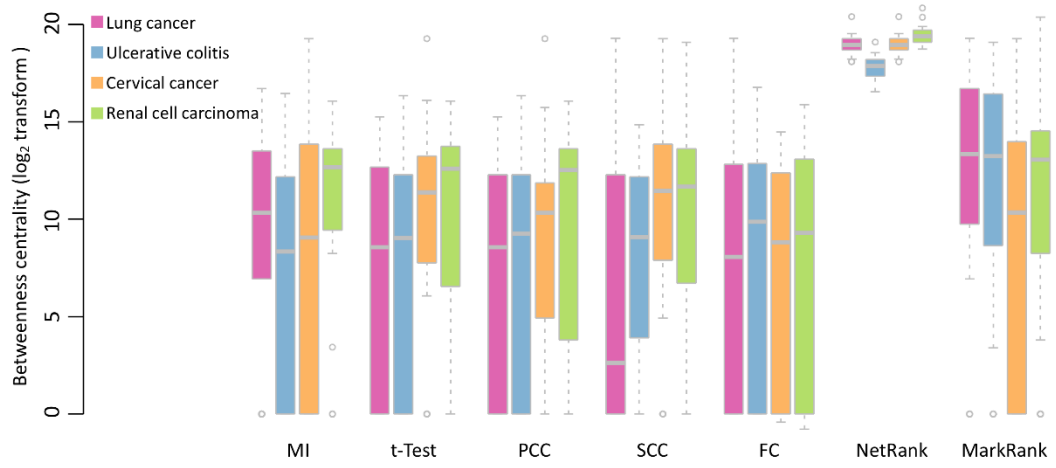

Fig. S13: The boxplot of betweenness centralities for each gene on four datasets. The log2 transform was performed for clarity. The abbreviations for each method are the same as in Fig. 4.

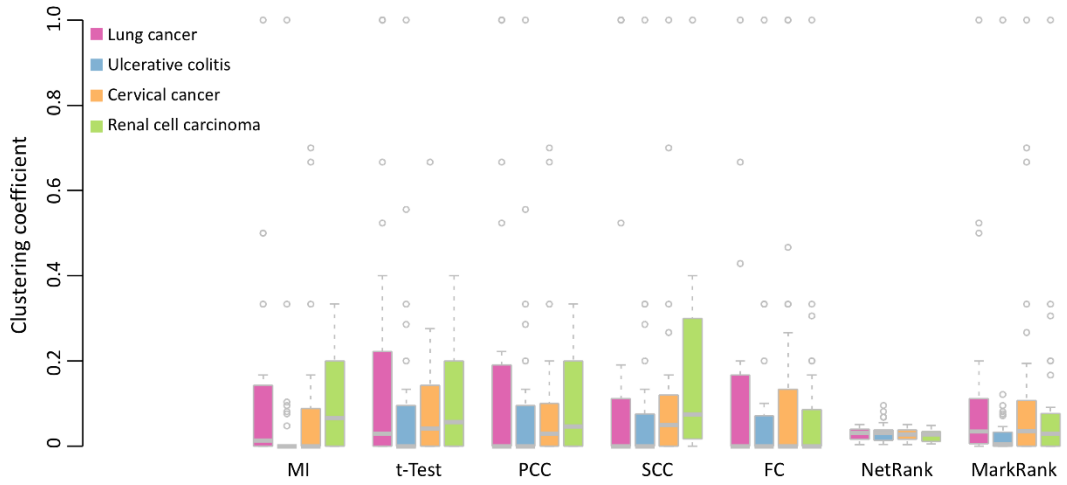

Fig. S14: The boxplot of clustering coefficients for each gene on four datasets. The abbreviations for each method are the same as in Fig. 4.

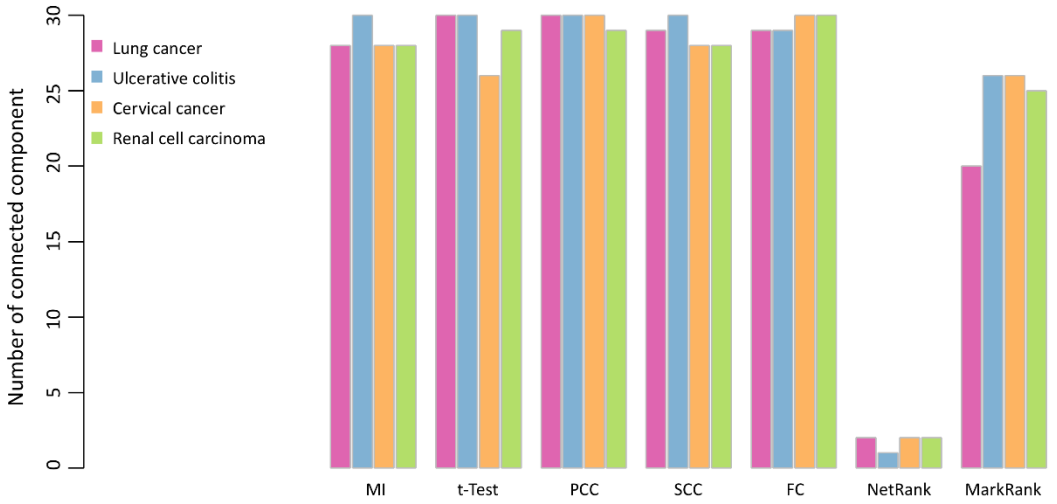

Fig. S15: The bar plot of the number of connected components formed by the identified gene set on four datasets. The abbreviations for each method are the same as in Fig. 4.

The results showed that the network-based methods had relatively higher node degree and betweenness centrality for the node importance indexes. Specifically, the performance of NetRank was significantly better than other methods, since it was prone to identify the hub nodes in PPI network. As for the module importance indexes, MarkRank, balancing the network structure and the discriminative power of cooperative gene combinations, had a moderate number of connected components, which is consistent with our initial motivation. There is no significant difference in clustering coefficients except that NetRank had a much smaller variance, which may also be because there were more hub nodes in the genes identified by NetRank since it is well-known that the hub nodes in PPI network have low clustering coefficients.

12. Enrichment analysis of gene set identified via MarkRank

In this study, we used two different enrichment approaches to evaluate the genes identified by MarkRank: one based on the hypergeometric test and another based on the Kolmogorov-Smirnov test.

The hypergeometric test - equivalent to a one-tailed Fisher's exact test (Fisher, 1922) - is an over-representation analysis using a 2×2 table. With the test, we would like to answer if the observed overlap between the genes of interest (here, the top 100 MarkRank genes were used) and if the degree that genes are related to a Gene Ontology category is any better than that obtained by chance alone. In our work, we used the Cytoscape plugin BiNGO (Maere, et al., 2005) to perform the enrichment analysis of Gene Ontology categories, which interactively uses molecular interaction networks visualized in Cytoscape. The Benjamini & Hochberg False Discovery Rate (FDR) correction was performed in multiple testing corrections, where we set the significance level at 0.05.

The hypergeometric test requires a strict cut-off in the list of MarkRank genes; therefore, the results are dependent on the chosen threshold. On the other hand, the Kolmogorov-Smirnov test (Massey, 1951), which is also widely used for gene set enrichment analyses, does not need the arbitrary threshold. The K-S test is a functional class scoring approach that first scores genes according to their order (e.g., MarkRank) and then transforms gene level scores into database entry level scores. The K-S test is very suitable for testing whether a given gene set (e.g., disease pathway) is significantly prioritized in a ranked full gene list. We applied the GSEAPreranked tool of the GSEA software (Subramanian, et al., 2005) with a correlation-weighted K-S test.

Notably, by applying the K-S test, we used specific gene sets that are the most relevant to related diseases to validate the capability of gene ranking methods for prioritizing known important disease-specific genes. Four of these gene sets were downloaded from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000) pathway database and one from the Molecular Signatures Database of Broad Institute (MSigDB) curated gene sets (c2, CGP: chemical and genetic perturbations) (Liberzon, et al., 2011). Specifically, enrichments were tested in the following gene sets: (1) lung cancer was tested in KEGG hsa05223 (non-small cell lung cancer) and KEGG hsa05222 (small cell lung cancer); (2) ulcerative colitis was tested in KEGG hsa05321 (inflammatory bowel disease, IBD); (3) cervical cancer was tested in MSigDB c2 (cervical cancer proliferation cluster); and (4) renal cell carcinoma was tested in KEGG hsa05211 (renal cell carcinoma). Here we used MSigDB for cervical cancer as KEGG does not contain a gene set specific for this disease. As for lung cancer, we executed the enrichment test on both non-small and small cell lung cancer gene sets since we do not know the exact proportion of subtypes in the lung cancer dataset GSE4115. The results of major subtype (non-small cell lung cancer), which accounts for approximately 85% of lung cancer, are shown in Fig. 6, while the related results of minor subtype (small cell lung cancer), which accounts for approximately 15% of lung cancer, can be found in Fig. S16. Detailed descriptions of these gene sets can be found in the Additional Materials.

13. The enriched biological processes of MarkRank genes using BiNGO

The top 100 genes identified by MarkRank for each dataset were employed to perform the enrichment analysis of Gene Ontology categories on the biological processes (BP) domain using the Cytoscape plugin BiNGO. The top 10 enriched GO terms for each dataset are shown in Tables S6-S9 (the full list can be found in Additional Materials).

Table S6. The top 10 enriched GO categories (Biological Process domain) for the lung cancer dataset using Cytoscape plugin BiNGO

GO ID	Description	p-value	FDR
0051246	regulation of protein metabolic process	4.10E-10	6.66E-7
0032268	regulation of cellular protein metabolic process	2.11E-09	1.71E-6
0051789	response to protein stimulus	5.94E-08	1.83E-5
0042981	regulation of apoptosis	7.53E-08	1.83E-5
0031399	regulation of protein modification process	8.37E-08	1.83E-5
0043067	regulation of programmed cell death	8.82E-08	1.83E-5
0009892	negative regulation of metabolic process	9.54E-08	1.83E-5
0031324	negative regulation of cellular metabolic process	9.77E-08	1.83E-5
0010941	regulation of cell death	1.01E-07	1.83E-5
0052547	regulation of peptidase activity	3.12E-07	5.06E-5

Table S7. The top 10 enriched GO categories (Biological Process domain) for the ulcerative colitis dataset using Cytoscape plugin BiNGO

GO ID	Description	p-value	FDR
0050793	regulation of developmental process	4.43E-7	6.25E-4
0001934	positive regulation of protein amino acid phosphorylation	9.82E-7	6.25E-4
0016310	phosphorylation	1.80E-6	6.25E-4
0042327	positive regulation of phosphorylation	1.90E-6	6.25E-4
0045937	positive regulation of phosphate metabolic process	2.25E-6	6.25E-4
0010562	positive regulation of phosphorus metabolic process	2.25E-6	6.25E-4
0042221	response to chemical stimulus	2.97E-6	6.25E-4
0043434	response to peptide hormone stimulus	3.00E-6	6.25E-4
0006468	protein amino acid phosphorylation	3.35E-6	6.25E-4
0009725	response to hormone stimulus	4.10E-6	6.67E-4

Table S8. The top 10 enriched GO categories (Biological Process domain) for the cervical cancer dataset using Cytoscape plugin BiNGO

GO ID	Description	p-value	FDR
0030855	epithelial cell differentiation	7.75E-11	1.17E-7
0008544	epidermis development	2.77E-10	2.10E-7
0009913	epidermal cell differentiation	6.13E-10	2.91E-7
0007398	ectoderm development	7.69E-10	2.91E-7
0009888	tissue development	2.75E-9	8.32E-7
0030216	keratinocyte differentiation	8.09E-9	2.04E-6
0018149	peptide cross-linking	1.52E-8	3.29E-6
0060429	epithelium development	3.38E-7	6.40E-5
0009628	response to abiotic stimulus	2.58E-6	4.01E-4
0007005	mitochondrion organization	2.65E-6	4.01E-4

Table S9. The top 10 enriched GO categories (Biological Process domain) for the renal cell carcinoma dataset using Cytoscape plugin BiNGO

GO ID	Description	p-value	FDR
0007588	excretion	1.29E-9	2.14E-6
0048878	chemical homeostasis	3.81E-8	2.40E-5
0050801	ion homeostasis	4.35E-8	2.40E-5
0006950	response to stress	8.63E-8	3.57E-5
0032501	multicellular organismal process	2.05E-7	5.85E-5
0042221	response to chemical stimulus	2.84E-7	5.85E-5
0048731	system development	3.17E-7	5.85E-5
0042127	regulation of cell proliferation	3.27E-7	5.85E-5
0048856	anatomical structure development	3.32E-7	5.85E-5
0042592	homeostatic process	3.77E-7	5.85E-5

14. Kolmogorov-Smirnov test on the small cell lung cancer gene set

Lung cancer is categorized into two main subtypes: non-small cell lung cancer (NSCLC), which accounts for approximately 85% of lung cancer, and small cell lung cancer (SCLC). Except for the Kolmogorov-Smirnov test (K-S test) executed on Kyoto Encyclopedia of Genes and Genomes (KEGG) hsa05223 (non-small cell lung cancer) we also performed the enrichment test on KEGG hsa05222 (small cell lung cancer) gene set, since there was no evidence supporting an exact subtype proportion of lung cancer for GSE4115. The related K-S test results are shown in Fig. S16.

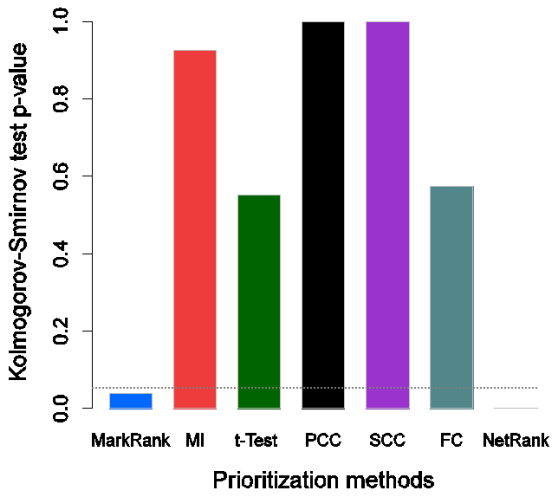

Fig. S16: The p-values of gene set enrichment analyses of KEGG pathway hsa05222 for each ranking method on GSE4115. The gray dotted line represents the p=0.05 significance level. The abbreviations for each method are the same as in Fig. 4.

The result showed that MarkRank exhibited a preferable enrichment performance (K-S test p-value < 0.05) with KEGG pathway hsa05222 when compared with other traditional ranking methods. Combined with the K-S test result in Fig. 6, genes identified via MarkRank give a good prioritization of the lung cancer genes for both NSCLC and SCLC subtypes.

15. Alternative methods for constructing the gene cooperation network

The key component of the MarkRank method is the construction of the gene cooperation network G_2. In our MarkRank model, we followed the subnetwork scoring function, f, as in previous studies and used the mutual information increment as the weight on related edge in G_2.

In the MarkRank algorithm, constructing G_2 needs the computation of mutual information for all pairs of possible genes. The time complexity for computing all mutual information is polynomial on the number of genes and is thus acceptable for large datasets. In practice, it may be time consuming when compared with other computation steps such as random walk iteration. Alternatively, we designed another two forms of G_2 construction to reduce the computation time. Here we used the same notations as introduced in Materials and Methods.

First, as the start of our inquiry, we tested that whether a source gene i pointing to a target gene j is approximately equivalent to the fact that $MI(e(j), y) \geq MI(e(i), y)$. If so, we can simplify the calculation in the G_2 construction by just using a fast ordering of single gene mutual information. Therefore, we used a related version of weight calculation as follows:

$$\widetilde{w}_{i,j} = \max\left\{0, f\left(x_j = 1, others = 0\right) - f(x_i = 1, others = 0)\right\}$$

In simulation studies, we found that for the same dataset, the overlap of non-zero terms in an adjacent matrix derived by $w_{i,j}$ and $\widetilde{w}_{i,j}$ was approximately 70% on average, whereas the AUC performance computed using $\widetilde{w}_{i,j}$ was far below the corresponding performance derived by $w_{i,j}$ (data not shown). Since this modeling method neglects the gene combination effect diverging from our original motivation, no further analysis using this method was performed.

Second, we added parameter d to reduce the computation time. Precisely, only the gene pairs whose shortest distances in the PPI network are less than d participate in G_2 construction. A smaller d restricts our search on the gene pairs with smaller PPI distance and $d = \infty$ equivalent to our original computation method. The user can set an appropriate d to balance the calculation depth with computation time in our implemented *markrank* function in the R package Corbi.

In conclusion, the former method neglects the gene combination effect diverging from our original motivation, and the latter one simplifies the calculation at the cost of losing complementary information of genes with long distance in the PPI network. Therefore, to fully use the information of expression dataset, we recommend the computation method introduced in the Materials and Methods section.

References

Breiman, L. Random forests. *Machine Learning* 2001;45(1):5-32.
Chatr-Aryamontri, A.*, et al.* The BioGRID interaction database: 2015 update. *Nucleic Acids Res* 2015;43(Database issue):D470-478.
Diaz-Uriarte, R. and Alvarez de Andres, S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* 2006;7:3.
Fisher, R.A. On the interpretation of x(2) from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society* 1922;85:87-94.
Kanehisa, M. and Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;28(1):27-30.
Keshava Prasad, T.S.*, et al.* Human Protein Reference Database--2009 update. *Nucleic Acids Res* 2009;37(Database issue):D767-772.
Liberzon, A.*, et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 2011;27(12):1739-1740.
Maere, S., Heymans, K. and Kuiper, M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 2005;21(16):3448-3449.
Massey, F.J. The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association* 1951;46(253):68-78.
Michiels, S., Koscielny, S. and Hill, C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* 2005;365(9458):488-492.
Noble, C.L.*, et al.* Regional variation in gene expression in the healthy colon is dysregulated in ulcerative colitis. *Gut* 2008;57(10):1398-1405.
Pena-Llopis, S.*, et al.* BAP1 loss defines a new class of renal cell carcinoma. *Nat Genet* 2012;44(7):751-759.
Scotto, L.*, et al.* Identification of copy number gain and overexpressed genes on chromosome arm 20q by an integrative genomic approach in cervical cancer: potential role in progression. *Genes, chromosomes & cancer* 2008;47(9):755-765.
Spira, A.*, et al.* Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nat Med* 2007;13(3):361-366.
Stark, C.*, et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 2006;34(Database issue):D535-539.
Subramanian, A.*, et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* 2005;102(43):15545-15550.
Szklarczyk, D.*, et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* 2015;43(Database issue):D447-452.
von Mering, C.*, et al.* STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res* 2003;31(1):258-261.
Winter, C.*, et al.* Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLoS Comput Biol* 2012;8(5):e1002511.