A novel similarity score based on gene ranks to reveal genetic relationships among diseases

Dongmei Luo[1,2]**, Chengdong Zhang**[4]**, Liwan Fu**[1]**, Yuening Zhang**[5]**, and Yue-Qing Hu**[1,3]

[1]**State Key Laboratory of Genetic Engineering, Institute of Biostatistics, School of Life Sciences, Fudan University, Shanghai, China**
[2]**Department of Information and Computing Science, School of Mathematics and Physics, Anhui University of Technology, Ma'anshan, China**
[3]**Shanghai Center for Mathematical Sciences, Fudan University, Shanghai, China**
[4]**Shanghai Public Health Clinical Center, Fudan University, Shanghai, China**
[5]**SJTU-Yale Joint Center for Biostatistics, Shanghai Jiao Tong University, Shanghai, China**

Corresponding author:
Yue-Qing Hu[1,3]

Email address: yuehu@fudan.edu.cn

ABSTRACT

Knowledge of similarities among diseases can contribute to uncovering common genetic mechanisms. Based on ranked gene lists, a couple of similarity measures were proposed in the literature. Notice that they may suffer from the determination of cutoff or heavy computational load, we propose a novel similarity score *SimSIP* among diseases based on gene ranks. Simulation studies under various scenarios demonstrate that *SimSIP* has better performance than existing rank-based similarity measures. Application of *SimSIP* in gene expression data of 18 cancer types from The Cancer Genome Atlas shows that *SimSIP* is superior in clarifying the genetic relationships among diseases and demonstrates the tendency to cluster the histologically or anatomically related cancers together, which is analogous to the pan-cancer studies. Moreover, *SimSIP* with simpler form and faster computation is more robust for higher levels of noise than existing methods and provides a basis for future studies on genetic relationships among diseases. In addition, a measure *MAG* is developed to gauge the magnitude of association of an individual gene with diseases. By using *MAG* the genes and biological processes significantly associated with colorectal cancer are detected.

INTRODUCTION

Exploring the common genetic basis of complex human diseases is often useful for understanding the disease relationships drawing on their genetic mechanisms. A similarity measure is a central component in detecting the common genetic basis among different diseases. Several common metrics have been proposed to measure similarities between diseases, such as Pearson, Spearman correlation coefficient, Euclidean distance, Manhattan distance, and Jaccard correlation coefficient (Antosh et al., 2013; Dennis et al., 2003; Serra et al., 2016; Shi et al., 2014). However, with the technological developments in molecular biology, large-scale gene expression profiling datasets produced from diverse technological platforms necessitate new and adaptive similarity measures to reveal meaningful genetic relationships across multiple platform types. Ranking genes according to their contribution to each disease can convert heterogeneous data into platform-independent rank lists, and a recent study (Serra et al., 2016) suggested that the discrimination ability of the similarity/distance measures based on ranked gene lists perform well or better than traditional measures (such as Euclidean distance and correlation coefficient), particularly for gene expression datasets produced with different biotechnologies (microarray, RNA-seq, etc.). Therefore,

46 the genetic overlaps among diseases can be detected by integrating multi-platform datasets using similarity
47 measures based on ranked gene lists, which can help us gain further insight on understanding disease
48 etiology.

49 A review of related studies in similarity measures based on ranked gene lists showed that most of
50 them depend on either a fixed cutoff position to consider overlaps between the top part of ranked gene
51 lists or variational cutoffs to select the one that generates the most significant results (Chen et al., 2015;
52 Dennis et al., 2003; Von Mering et al., 2006). To avoid the uncertainty of results owing to arbitrariness in
53 cutoff settings, global rank-based similarity measures have been developed. For example, the algorithm
54 of *GOrilla* is a tool to discover and visualize the enrichment of GO terms that ranks genes by fold-change
55 (Eden et al., 2009). The algorithm of gene set enrichment analysis (*GSEA*) allows all genes to contribute
56 to overlapping signals in proportion to their degree of differential expression and can detect the weak
57 signals that would be discarded by 'threshold' approaches (Efron and Tibshirani, 2007; Subramanian
58 et al., 2005). The algorithm of *CORaL* estimates the significant set size using the overlaps between
59 sections of the ranked gene lists by maximizing statistical likelihood (Antosh et al., 2013). The algorithm
60 of *R2KS* particularly emphasizes finding the same items near the top of the ranked gene list (Ni and
61 Vingron, 2012). Plaisier et al. (2010) presented a popular 'threshold-free' approach in neuroscience and
62 biomedicine named rank-rank hypergeometric overlap (*RRHO*), which identifies and visualizes regions of
63 significant overlap between two ranked gene lists and determines the statistical significance of enrichment
64 by hypergeometric distribution.

65 Moreover, a similarity measure *OrderedList* that focuses on evaluating whether there is significant
66 overlap between two ranked gene lists was proposed (Yang et al., 2006). The *OrderedList* is the weighted
67 sum of the number of overlapping genes with an exponentially decaying weight, where a parameter β is
68 introduced to determine how deep to go in the ranked gene lists. There are many studies on statistics-based
69 improvement from similarity measures *OrderedList*. For example, Serra et al. (2016) introduced similarity
70 measure FES_β (fraction enrichment sum) and set the exponentially decaying parameter β as 0, 0.001,
71 and 0.01. Chen et al. (2015) adopted the default series of β values in the R package *OrderedList* and
72 proposed a robust statistic minimum p value: $\min_\beta p_\beta$, where parameter β can be calculated by setting
73 a minimum weight and a default series of positions. To avoid arbitrariness and manual intervention of
74 parameter β selection, Chen et al. (2015) finally found a parameter-free similarity measure *WeiSumE**
75 with good performance to detect overlaps among ranked gene lists in simulations, which is the weighted
76 sum that normalizes the number of overlapping genes on the top genes of two ranked gene lists by its
77 expectation. Generally, existing rank-based similarity measures are mostly based on the overlaps among
78 ranked gene lists and may lead to information loss owing to fixed cutoff positions in the ranked gene
79 lists, the uncertainty of results because of the arbitrariness of the cutoff position, or heavy computational
80 burden.

81 In this study, we propose *MAG* based on the transformation of gene ranks to measure the magnitude
82 of association of the individual gene with two diseases. By exploring the summation of *MAG* over all
83 genes, we develop a novel similarity score *SimSIP* with simpler form and light computation burden to
84 gauge genetic overlap among diseases based on gene ranks instead of intersections between the top part of
85 ranked gene lists. To show the superiority of *SimSIP*, we firstly conduct a series of simulation studies to
86 demonstrate the performance of *SimSIP* compared to some existing similarity measures based on ranked
87 gene lists (*WeiSumE** (Chen et al., 2015), *OrderedList* (Yang et al., 2006), $FES_{0.001}$, $FES_{0.01}$(Serra et al.,
88 2016)) and Euclidean distance *EucD* under various scenarios. Secondly, we apply *SimSIP* to analyze
89 the gene expression data of cancers in The Cancer Genome Atlas database and find that it sheds light
90 on the genetic relationships among cancers. Thirdly, we arrange the significantly similar cancer pairs
91 among the 18 cancer types detected by *SimSIP* into a disease network in which the tendency to cluster the
92 histologically or anatomically related cancers provides basic support for pan-cancer studies. Finally, for
93 the most significantly similar cancer pair, colon adenocarcinoma (*COAD*) and rectum adenocarcinoma
94 (*READ*), found by *SimSIP*, we use *MAG* to measure the magnitude of association of each gene both with
95 *COAD* and *READ* and find the important oncogenes of colorectal cancer which are associated with *COAD*
96 and *READ* and regulated in the same pattern. Moreover, biological processes highly associated with
97 colorectal cancer are detected.

MATERIALS AND METHODS

MAG **and** *SimSIP*

Let us assume that there are n genes for disease 1 and disease 2. For gene i, $1 \leq i \leq n$, let a_i be its rank among the n genes for disease 1, and b_i is defined similarly for disease 2. Now we intend to gauge the genetic similarity or genetic relationship between these two diseases based on gene rank lists $\{a_i\}_{i=1}^{n}$ and $\{b_i\}_{i=1}^{n}$. For gene i, we have its ranks a_i and b_i for disease 1 and disease 2, respectively. The gene rank usually represents the strength of association with the disease, in the sense of small rank meaning strong association and big rank meaning weak association. For example, the rank can be assigned for each gene by the magnitude of p value reported in the study of detecting differentially expressed genes, which is the routine work in the literature. To facilitate the construction of the similarity score, we transform the ranks a_i and b_i to their reciprocal $1/a_i$ and $1/b_i$, whose values fall in the interval $(0, 1]$ and can be treated directly as the magnitude of association of gene i with diseases 1 and 2, respectively. Further, we employ the geometric mean $\sqrt{1/a_i \cdot 1/b_i}$ of $1/a_i$ and $1/b_i$ and call it *MAG*, which is a compromise between $1/a_i$ and $1/b_i$ as the magnitude of association of gene i with both diseases 1 and 2. Intuitively, a small value of *MAG* means a weak association of gene i with these two diseases, a big value means a strong association of gene i with these two diseases. Note the value of *MAG* is between n^{-1} and 1.

Now let us explore the summation $\sum_{i=1}^{n} \sqrt{1/a_i \cdot 1/b_i}$ of *MAG* over all genes, i.e. the total magnitude of association for all n genes with diseases 1 and 2. Note that this summation is actually the inner product of vectors $\{1/\sqrt{a_i}\}_{i=1}^{n}$ and $\{1/\sqrt{b_i}\}_{i=1}^{n}$ with positive components. It is easy to check that the length of each of these two vectors is $\sqrt{\sum_{i=1}^{n} 1/i}$, which depends only on n and is independent of genes' concrete ranks. So the summation is proportional to the angle between the two vectors mentioned above. Recall that the inner product in algebra or geometry theory is sometimes called the scalar product or dot product and is the projection of one vector on another in geometry space, which is a symmetric measure of closeness of two vectors. So the similarity score by inner product

$$SimSIP = \sum_{i=1}^{n} \sqrt{1/a_i \cdot 1/b_i}$$

is an appropriate candidate for measuring similarity between contributions of all n genes to diseases 1 and 2, and it can be further taken as a similarity score between two diseases derived from the corresponding gene rank lists. Considering one extreme case in which $a_i = b_i$ for every gene i, which means the rank of every gene for disease 1 is exactly equal to the rank of same gene for disease 2, these two diseases are the most similar in terms of these n genes and *SimSIP* attains its maximum $\sum_{i=1}^{n} 1/i$. Considering the other extreme case in which the gene ranking top for disease 1 would always rank bottom for disease 2, and gene ranking bottom for disease 1 would always rank top for disease 2, these two diseases are the most dissimilar in term of these n genes and *SimSIP* attains its minimum. So a big value of *SimSIP* would imply that two diseases are similar.

Assessment of significance

It is easily observed from the expression of *SimSIP* that the more identically ranked genes associated with diseases 1 and 2 there are, the more similar two diseases are, and the larger *SimSIP* is. Therefore, an observed *SimSIP* larger than expected under the null hypothesis that two diseases have non-overlapping genes means significant. When we say that two diseases have no overlapping genes in the genetic perspective, we mean that the two gene rank lists are two random shufflings of $1 \sim n$. Based on the expression of *SimSIP*, we describe a procedure to generate the empirical distribution of *SimSIP* under the null hypothesis, which is used to evaluate the significance of the similarity score. Without loss of generality, we fix the first gene rank list as $\{i\}_{i=1}^{n}$ and random permutate $1 \sim n$ as $\{b_i\}_{i=1}^{n}$ for S times, then we obtain the corresponding $SimSIP^s, s = 1, 2, \ldots, S$. The p value of *SimSIP* is computed as

$$p = \frac{\sum_{s=1}^{S} I(SimSIP^s \geq SimSIP)}{S}, \tag{1}$$

where $I(\cdot)$ is the indicator function assigning the value 1 or 0 relying on whether the condition within brackets is met. Let p^r denote the p value and $r = 1, 2, \ldots, R$ for R replications, and the power under the

alternative hypothesis or the type I error under the null hypothesis for a given significance level α is

$$Power = \frac{\sum_{r=1}^{R} I(p^r \leq \alpha)}{R}. \tag{2}$$

Regarding *MAG*, we can employ a similar procedure to evaluate the significance of an observed *MAG*. Under the null hypothesis that two diseases having non-overlapping genes, the two gene ranks a_i and b_i involved in are randomly drawn from $\{1, 2, \ldots, n\}$ with replacement, and it is not difficult to get the distribution of *MAG* under the null hypothesis. The significance of the observed *MAG* can then be calculated.

Results
Simulation Study
Parameter setting

We carry out simulations to evaluate the performance of *SimSIP* and the existing rank-based methods *WeiSumE** (Chen et al., 2015), *OrderedList* (Yang et al., 2006), $FES_{0.001}$, and $FES_{0.01}$ (Serra et al., 2016), which are weighted sum of the number of overlapping genes between the top part of ranked gene lists, and Euclidean distance *EucD* between the original values, which is a method with no need of ranking.

As done in Chen et al. (2015), we set the number of genes $n = 6000, 11000$, and 25000 and randomly choose two sets of d genes from the n genes as the associated genes with diseases 1 and 2, respectively. We fix the $d = 1, 5, 10$, and 20 in the simulation. The number o of overlapping associated genes in the two sets is taken as 0 for evaluating the type I error rate and positive for evaluating the power. For convenience, genes 1 to d are assumed to be associated with disease 1 and genes $d - o + 1$ to $2d - o$ are associated with disease 2, where $d \geq o$. As performed by Chen et al. (2015), two normal distributions $N(0, \sigma^2)$ and $N(1, \sigma^2)$ are employed to generate the rank of each gene. Specifically, $N(1, \sigma^2)$ is for the associated genes and $N(0, \sigma^2)$ for the remaining $n - d$ non-associated genes, in either disease 1 or 2. The gene rank is from the value of normal distribution. The variance σ^2 in the normal distribution is taken as 0.01, 0.05, 0.1, 0.5 when $o = 0$; 0.07, 0.09, 0.11, 0.13 when $o = 1$; 0.1, 0.15, 0.2, 0.25 when $o = 5$; 0.2, 0.25, 0.3, 0.35 when $o = 10$; and 0.3, 0.35, 0.4, 0.45 when $o = 20$. The variance σ^2 plays an importance role in discerning the associated genes. The number of replications R is set to 1000 in the computation of powers/type I error rates, and the nominal significance level is set to 0.05. We set $S = 1000$ in the simulation study and $S = 10$ millions in real data analysis of TCGA. As an illustration, the empirical distribution of *SimSIP* for total genes number $n = 6000, 11000, 25000$ are given in Figures S1-S3 and the empirical distribution of *MAG* for $n = 6000, 11000, 25000$ are given in Figures S4-S6.

Type I error

We firstly show the simulation results under the null model in evaluating the type I error rate of *SimSIP* and five existing methods under various scenarios. For the null setting, the empirical p values of existing rank-based methods (*WeiSumE**, *OrderedList*, $FES_{0.001}$, $FES_{0.01}$) are computed by permuting ranked gene lists from the value of the normal distribution, and the null distribution of *EucD* is obtained by permuting the value of the normal distribution generated. All empirical sizes shown in Supplementary Table S1 are around the significance 0.05 and are well controlled.

Power comparison

In the second set of simulations, the performance of *SimSIP* is compared with five other existing measures in terms of power. The results show that the powers of six measures are all monotonic decreasing for higher noise (variance σ^2) and monotonic increasing for a bigger ratio of o to d. The findings are shown in Figure 1. Firstly, for the same total gene number n, the smaller the number of associated genes d becomes, the more obvious the advantage of the *SimSIP* is. For example, in the scenarios of $d = 1$ (in subfigures A, E and I of Figure 1), the power of *SimSIP* has almost a 10% increase compared with that of *WeiSumE** (Chen et al., 2015) and *OrderedList* (Yang et al., 2006), is more than 2 times the powers of $FES_{0.01}$ (Serra et al., 2016), and almost more than 8 times the powers of $FES_{0.001}$ (Serra et al., 2016) and *EucD*. The scenario of $d = 5$ is analogy to the scenario of $d = 1$. However, in the scenarios of increasing d up to 10 and 20, the advantage of *SimSIP* is no longer obvious. Secondly, for the same d, the larger the number of genes (n) becomes, the more obvious the advantages of *SimSIP* is. For example, in the
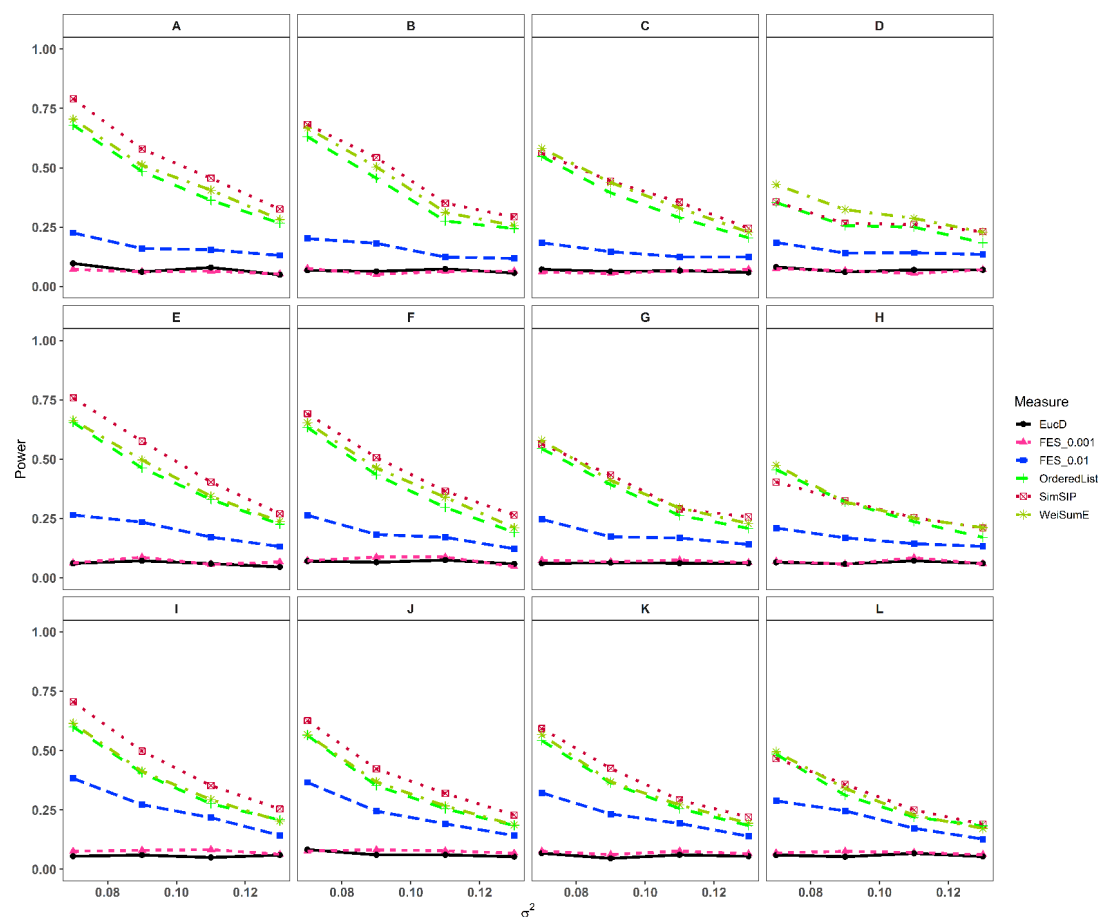
Figure 1. Powers of *EucD*, $FES_{0.001}$, $FES_{0.01}$, *OrderedList*, *SimSIP* and *WeiSumE** when $o = 1$ with 12 scenarios. (A) n=6000, d=1; (B) n=6000, d=5; (C) n=6000, d=10; (D) n=6000, d=20; (E) n=11000, d=1; (F) n=11000, d=5; (G) n=11000, d=10; (H) n=11000, d=20; (I) n=25000, d=1; (J) n=25000, d=5; (K) n=25000, d=10; (L) n= 25000, d=20. The arrangement of variance σ^2 on *x* axis is a series (0.07, 0.09, 0.11, 0.13) on which the power of six measures can be ranged from 0.1 to 0.95.

scenarios of $n = 6000$ and $n = 11000$, the similarity measure *WeiSumE** is the most powerful when the number of associated genes $d = 20$; however, *SimSIP* becomes superior when compared with the other five measures when n increases to 25000. Thirdly, when the number of associated genes d is relatively large, *SimSIP* gradually manifests certain advantages with the increase in noise (variance σ^2), which can be observed by six subfigures C, D, G, H, K and L in Figure 1 (in the scenarios of $d = 10$ and $d = 20$). When increasing the number of overlapping associated genes o up to 5, *SimSIP* always maintains the superiority compared with the other five methods, and its power has obvious advantages for the larger ratio of o to d with a fixed n and for the larger n with a fixed d (see Figure 2). When the number of overlapping associated genes o increases from 5 to 10 or 20, the advantages of *SimSIP* are much more significant (see Figures S7-S8 in Supplementary Information).

As shown in Figure 1-2 and S7-S8, the similarity measure *WeiSumE** performs better than other measures when there are less but significant overlap between two diseases with lower noise (variance σ^2), and our method *SimSIP* performs better in all the other scenarios. Especially for $n = 25000$, which approximates the number of genes in the full human genome, the performance of *SimSIP* is always the best. Thus, compared with other measures, *SimSIP* is more appropriate for detecting genetic similarity between longer gene lists and works well when more overlapping genes occur among diseases. Most importantly, our method *SimSIP* is more robust for higher noise than the other five methods. Therefore, *SimSIP* is more suitable for the study of human diseases than the existing methods, especially for the study of cancers in which there are more genetic overlaps.
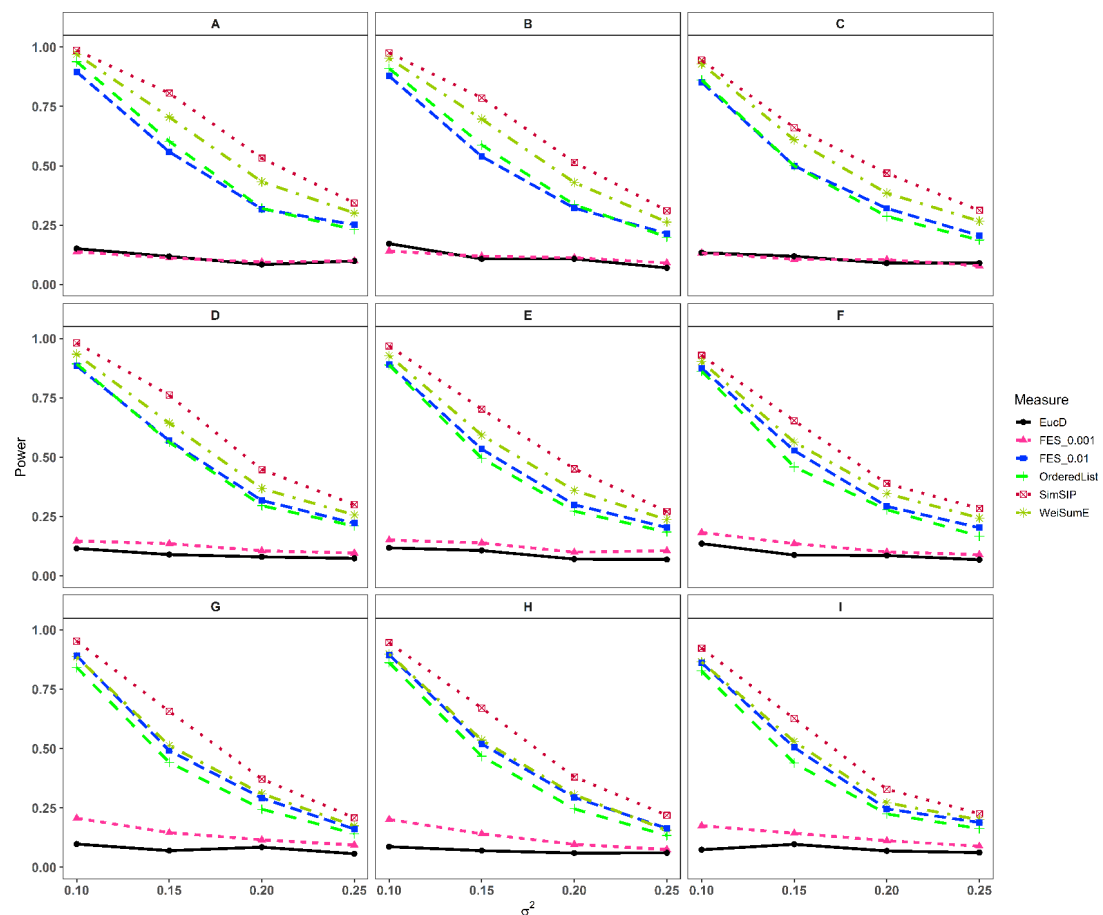
Figure 2. Powers of *EucD*, $FES_{0.001}$, $FES_{0.01}$, *OrderedList*, *SimSIP* and *WeiSumE** when $o = 5$ with 9 scenarios. (A) n=6000, d=5; (B) n=6000, d=10; (C) n=6000, d=20; (D) n=11000, d=5; (E) n=11000, d=10; (F) n=11000, d=20; (G) n=25000, d=5; (H) n=25000, d=10; (I) n= 25000, d=20. The arrangement of variance σ^2 on *x* axis is a series (0.1, 0.15, 0.2, 0.25) on which the power of six measures can be ranged from 0.1 to 0.95.

191   Furthermore, the similarity measure *OrderedList* has a higher power with fewer overlaps (for $o = 1$)
192   and $FES_{0.01}$ has a higher power with more overlaps (for $o > 1$). Their performances are superior to
193   those of similarity measure $FES_{0.001}$ and Euclidean distance *EucD*, which always have low power in all
194   scenarios. We also find that the rank-based measures always perform better than Euclidean distance *EucD*,
195   which uses the value of the normal distribution rather than gene ranks. These results demonstrate that the
196   rank of a gene can provide additional information on its contribution to each disease upon converting real
197   data to ranks.

198   In general, compared with existing similarity measures, *SimSIP* performs better in almost every
199   simulation, which indicates that it is feasible to construct similarity measure by replacing overlaps
200   between top ranked gene sets with gene ranks.

201   **TCGA Data Analysis**
202   TCGA is a combined effort by multiple research institutes, in which tumor and normal samples from more
203   than 11000 patients are publicly available, comprising 37 types of (epi)genetic and clinical data for 33
204   types of cancer. We download gene expression datasets (whose gene expression profiles were determined
205   experimentally using the Illumina HiSeq 2000 RNA Sequencing platform) of 33 types of cancer using
206   the UCSC Xena functional genome browser (*https* : *//xenabrowser.net/datapages/*). Expression data
207   of 20530 genes are available for each cancer, and 18 types of cancer are selected (see details in Table 1)
208   based on the criteria that the sample size of the control group is not smaller than 5. We rank all genes

by their *p* values derived from the R package LIMMA for differential expression analysis and apply our proposed method to gauge the genetic similarity among the 18 types of cancer in TCGA. In this section, we choose *WeiSumE** (Chen et al., 2015), which performs only second to *SimSIP* under all simulation scenarios, to compare with *SimSIP*.

For the $(18 \times 17)/2 = 153$ cancer pairs among 18 cancers, we compute their similarity scores by using *SimSIP* and *WeiSumE** respectively and standardize them by normalization method $\frac{T-T_{min}}{T_{max}-T_{min}}$. The normalized similarity scores about *SimSIP* are in [0.11, 0.77] with mean 0.24 and variance 0.015, and that about *WeiSumE** are in [0.064, 0.53] with mean 0.13 and variance 0.005. Obviously, the normalized similarity scores about *SimSIP* are generally bigger and more spread than those about *WeiSumE**. So *SimSIP* is more discriminative than *WeiSumE** in quantifying the relationships among diseases.

Table 1. Sample sizes of control group and case group per cancer type in TCGA

Abbreviation	Cancer type^a	n_0^{b}	n_1
BLCA	Bladder urothelial carcinoma	19	407
BRCA	Breast invasive carcinoma	114	1097
CHOL	Cholangiocarcinoma	9	36
COAD	Colon adenocarcinoma	41	286
ESCA	Esophageal carcinoma	11	184
GBM	Glioblastoma multiforme	5	154
HNSC	Head and neck squamous cell carcinoma	44	520
KICH	Kidney chromophobe	25	66
KIRC	Kidney renal clear cell carcinoma	72	533
KIRP	Kidney renal papillary cell carcinom	32	290
LIHC	Liver hepatocellular carcinoma	50	371
LUAD	Lung adenocarcinoma	59	515
LUSC	Lung squamous cell carcinoma	51	502
PRAD	Prostate adenocarcinoma	52	497
READ	Rectum adenocarcinoma	10	94
STAD	Stomach adenocarcinoma	35	415
THCA	Thyroid carcinoma	59	505
UCEC	Uterine corpus endometrial carcinoma	24	176

^{a} From 33 types of cancer in TCGA, 18 types of cancer with control group sample size being 5 or more are selected.

^{b} n_0 is sample size of control group and n_1 is sample size of case group for a given cancer

Significant cancer pairs in TCGA

To illustrate the application of *SimSIP* and *WeiSumE** in exploring significant relationships among 18 cancers, we compute the empirical *p* values of *SimSIP* and *WeiSumE** based on the null distribution (shown in Figures S9-S10) for the 153 cancer pairs among 18 cancers.

Given the nominal significance level of 0.05, the Bonferroni adjustment of empirical *p* values 0.05/153 is employed to detect significant similar cancer pairs. 91 significant similar cancer pairs are detected by *SimSIP*, 82 pairs by *WeiSumE**, with 81 pairs in their intersection (see Table S2 and Figure S11). For the 81 cancer pairs detected both with *SimSIP* and *WeiSumE** (Chen et al., 2015), Figure 3 displays the empirical *p* values of the two methods for the same cancer pair. Clearly, the empirical *p* values of *SimSIP* are generally smaller than that of *WeiSumE** (except for cancer pair *KIRP* and *LUAD* with index 63 in Figure 3), which is consistent with the results given in the simulation.

As shown in Figure S11, *SimSIP* finds more possibly significant cancer pairs which include almost all the significant cancer pairs found with *WeiSumE**. Among the 10 significant cancer pairs (shown in red in Table S2) found with *SimSIP*, not *WeiSumE**, five cancer pairs can be explained in pan-cancer studies: For cancer pair *COAD* and *UCEC*, the diversity of high antigen-specific TCR repertoires correlates with the improved prognostic progression-free interval in *COAD* and *UCEC* (Thorsson et al., 2018); the expression of TMEM173 in tumor tissues is significantly upregulated and hypomethylated in cancer pair *COAD* and *THCA* but significantly downregulated and hypermethylated in cancer pair *LUSC* and *PRAD* (An

237 et al., 2019). Furthermore, there are high mutations rates for TBK1 in cancer pair *COAD* and *UCEC*, and
238 the expression of TMEM173 is positively associated with the infiltration of immune cells in cancer pair
239 *BRCA* and *THCA* (An et al., 2019). A high IRF3 expression yields a poor prediction of prognosis in
240 patients with *KIRC* and *PRAD* (An et al., 2019). In addition, cancer pair *COAD* and *UCEC* depends on
241 components of the EGFR pathway at similar frequencies (Wormald et al., 2013).
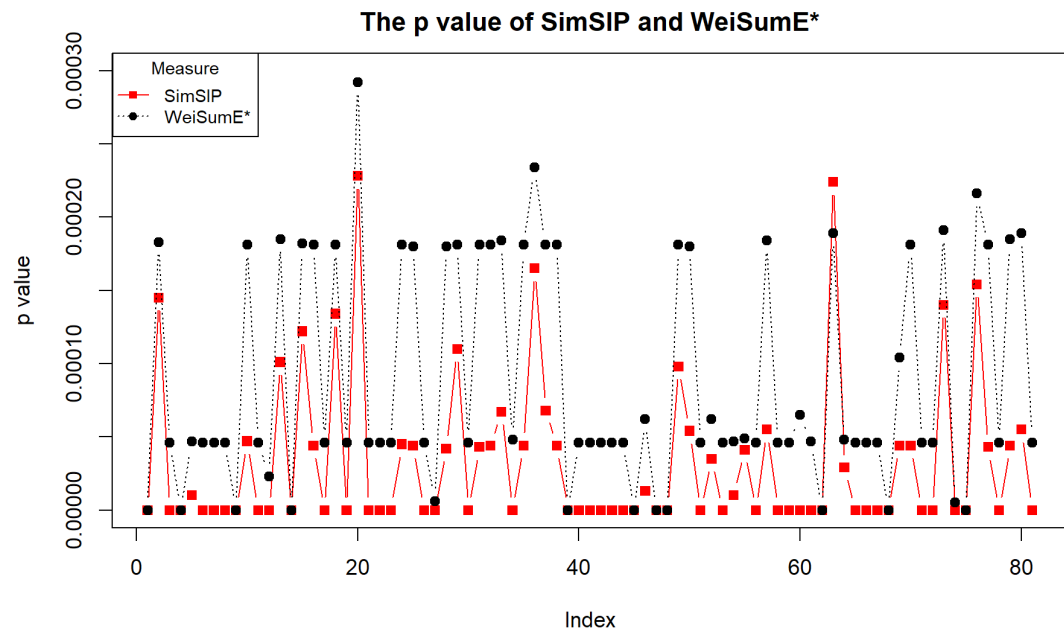
Figure 3. The empirical *p* values of *SimSIP* (black dot) and *WeiSumE** (red square) for 81 cancer pairs found with both *SimSIP* and *WeiSumE**.

242 ### *Genetic similarity among 18 types of cancer*
243 Depicting the 91 significant similar cancer pairs found with *SimSIP* in a disease network by using software
244 Cytoscape (*https*://*cytoscape.org*/) as in Figure 4, the following observations could be addressed:

245 1. *The vertex HNSC has the largest degree, 15, which shows that the HNSC has significant similarity*
246   *with the other 15 cancers, except KICH and KIRC. The degrees of vertices such as BLCA and*
247   *UCEC reach 14; READ, ESCA, LUSC, and HNSC reach 13; and LUAD and LIHC reach 12. There*
248   *are close intrinsic genetic relationships among the 18 cancers. The seven types of cancers ESCA,*
249   *LUSC, HNSC, BLCA, BRCA, STAD, and UCEC (their vertices with deeper color and bigger size)*
250   *with a higher degree are closer to each other and tend to form a pivotal hub of the disease network*
251   *of 18 cancer types.*
252 2. *Cancers originating from the same organ or tissue tend to co-cluster, such as cancer pair READ and*
253   *COAD or KIRP and KIPC with the most similar relationship. In addition, cancers with proximity*
254   *also tend to group together, such as LIHC and CHOL with a highly significant relationship. These*
255   *provide evidence that tumors with closer physical distance in human organs have similar sources*
256   *of endoderm development or exposure to a common cancer-causing environmental factor (Sell*
257   *and Dunsford, 1989). Further, for the three types of kidney tumors, KICH, KIRP, and KIPC, the*
258   *similarity between KIRP and KIPC is more significant than KICH and KIPC or KIRP and KIPC,*
259   *which may be explained by the fact that KIRC and KIRP are cancers of the proximal tubule*
260   *segments, whereas KICH is a cancer of the distal tubule segments (Chen et al., 2016; Davis et al.,*
261   *2014; Lee et al., 2015). Compared to other cancer pairs from the same tissue, the degree of*
262   *similarity between the two types of lung tumor LUSC and LUAD is not so significant, which may be*
263   *due to the derivation of cell types of the two types of lung tumors: LUSC originates from squamous*
264   *epithelial cells in the respiratory tract and alveoli, whereas LUAD originates from a large number*
265   *of glandular or alveolar cells (Li et al., 2015; Mainardi et al., 2014; Sutherland et al., 2014).*
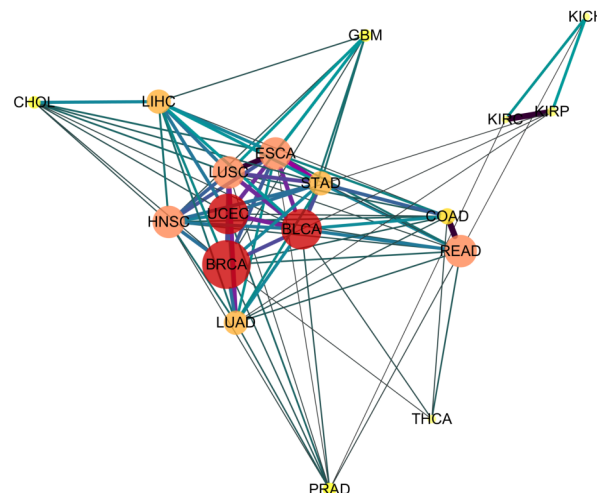
Figure 4. The diseases network of 18 cancer types. Taking Bonferroni adjustment of empirical *p* values given nominal significance 0.05 as a threshold, we select 91 significantly similar cancer pairs found with *SimSIP* to demonstrate the genetic relationships among cancers by software Cytoscape (*https : //cytoscape.org/*). The width, length and color of the edge in the network are determined by the magnitude of *SimSIP*, the size and color of the vertex in the network is determined by the degree of vertex (the number of cancers which is significantly related to this cancer).

3. *There are significant similarities among gastrointestinal tumors (READ, COAD, STAD, and ESCA), which is consistent with the results of integrative clustering across data types in the miRNA, mRNA, and RPPA platforms (Hoadley et al., 2018). In the disease network (Figure 4), squamous cell carcinomas (BLCA, ESCA, HNSC, and LUSC) are also co-clustered, and the similarity score of the cancer pair ESCA and LUSC is the top three and has more significant similarity. Hoadley et al. (2018) reported a similar discovery based on a multi-platform dataset (including miRNA, mRNA, RPPA, aneuploidy, and DNA methylation data) in TCGA, and Abrams et al. (2018) also suggested that regardless of the tissue types of squamous cell carcinoma, potential similarities were detected among the transcription factor expression profiles of BLCA, ESCA, HNSC, and LUSC. In addition, the gynecologic tumors UCEC and BRCA are also close to each other in the network, which is consistent with the results of previous studies (Hoadley et al., 2018).*

4. *Finally, similar to previous studies (Abrams et al., 2018; Hoadley et al., 2018), three types of cancer, PRAD, THCA, and GBM, are relatively independent in the network, the relationships among them and other cancers are relatively weak. In addition, instead of squamous cancers being clustered together, adenocarcinomas (PRAD, COAD, LUAD, STAD, and READ) appear to be scattered around the edge of the network.*

From the above observations of the disease network (Figure 4) obtained with *SimSIP*, we find them to be analogous to those of previous pan-cancer studies (Abrams et al., 2018; Hoadley et al., 2018): histologically or anatomically related cancers tend to cluster together, which provide basic support for analyses of pan-gastrointestinal, pan-squamous, pan-kidney, and pan-gynecological cancers.

Associated genes with colorectal cancer

For the most significantly similar cancer pair *COAD* and *READ* found with *SimSIP*, we try to explore the common underlying genetic mechanisms by sorting *MAG* values of genes in descending order. The significance of the difference of expression level of the top 5 genes in *COAD* and *READ* are shown in Table S3, which are obtained from the TCGA data mining website UALCAN (http://ualcan.path.uab.edu/). Table S3 provides evidence that the top 5 genes are associated with *COAD* and *READ* and they are regulated in the same pattern (either up-regulated or down-regulated). It is worth mentioning that, for the top 500 genes associated with *COAD* and *READ* and the top 500 genes associated with *KICH* and *STAD*, which are the most similar and the least similar cancer pairs found with *SimSIP* respectively, the proportion of genes regulated in the same pattern is 1 and 0.648 respectively. Furthermore, we compute

²⁹⁶ the correlation coefficient between the logarithmic transformation of *SimSIP* and the proportion of genes
²⁹⁷ regulated in the same pattern in the top N ($N = 500, 1000, 1500, 2000$) genes associated with each of 153
²⁹⁸ cancer pairs (shown in Figure S12). Obviously, there is a significant relationship between the degree
²⁹⁹ of overlap between diseases and the proportion of genes regulated in the same pattern. These findings
³⁰⁰ suggest that the associated genes of highly overlapping cancers may be regulated in the same pattern.

³⁰¹ Through gene annotation by using Metascape (http://metascape.org/) and literature review, among
³⁰² the top 5 genes, one gene (CDH3) is associated with multiple cancers (shown in Figure S13), and two
³⁰³ genes (ETV4, KRT24) are associated with colorectal cancer; the remaining genes can be considered as
³⁰⁴ candidate susceptibility genes of colorectal cancer (Table S4). CDH3, the top 1 gene detected by *MAG*,
³⁰⁵ is located in a region on the long arm of chromosome 16. Paredes et al. (2012) advised that genetic
³⁰⁶ or epigenetic changes in this gene or changes in its protein expression often lead to tissue disorders,
³⁰⁷ cellular dedifferentiation, and enhanced invasiveness of tumor cells. This gene is associated with intestinal
³⁰⁸ infections (Van Marck et al., 2011) and colon cancer (Van Marck et al., 2011; Sun et al., 2011). In addition,
³⁰⁹ its over-expression is also associated with tumor progression and low survival in non-small-cell lung
³¹⁰ cancer (Imai et al., 2018) and in the loss of heterozygous events for breast and prostate cancer (Royo
³¹¹ et al., 2016; Sousa et al., 2014; Vieira et al., 2017). Moreover, CDH3 is significantly associated with
³¹² liver cancer, gastric cancer, bladder cancer, and cervical adenocarcinoma (Bauer et al., 2014; Paredes
³¹³ et al., 2012; Sun et al., 2011; Van Marck et al., 2011). Searching for CDH3 in the CancerMine database
³¹⁴ (https://www.mycancergenome.org/) reveals that 12 cancers are associated with the over-expression of
³¹⁵ CDH3. This gene is therefore very important in the genetic mechanism of cancer. The top 3 gene ETV4
³¹⁶ is strongly linked to metastasis of colorectal cancer (Dumortier et al., 2018) and enriched in pathway
³¹⁷ transcriptional misregulation in cancer (Table S4). The top 5 gene KRT24 is over-expressed in patients
³¹⁸ with colorectal cancer and is a susceptibility gene for early onset of colorectal cancer (Hong et al., 2007).

³¹⁹ ***Associated signaling pathway with colorectal cancer***
³²⁰ Calculating empirical p values of *MAG* for each gene by its null distribution (described in Figure S14),
³²¹ 1838 genes significantly associated with colorectal cancer are obtained. By DAVID (https://david.ncifcrf.gov
³²² /tools.jsp), these 1838 genes associated with colorectal cancer are clustered in diverse functional cate-
³²³ gories.

Table 2. The result of pathway enrichment analysis of colorectal
cancer

ID	KEGG pathway	Benjamini
hsa04713	Circadian entrainment	2.30E-04
hsa04024	cAMP signaling pathway	2.97E-04
hsa03008	Ribosome biogenesis in eukaryotes	3.68E-04
hsa04080	Neuroactive ligand-receptor interaction	7.92E-04
hsa04020	Calcium signaling pathway	0.003328211
hsa04723	Retrograde endocannabinoid signaling	0.006342379
hsa04724	Glutamatergic synapse	0.006620968
hsa04911	Insulin secretion	0.010800326
hsa04728	Dopaminergic synapse	0.011092804
hsa04725	Cholinergic synapse	0.017964832
hsa04022	cGMP-PKG signaling pathway	0.021342644
hsa04970	Salivary secretion	0.023476401
hsa04261	Adrenergic signaling in cardiomyocytes	0.024338123
hsa00071	Fatty acid degradation	0.029045641
hsa05032	Morphine addiction	0.036794268
hsa04978	Mineral absorption	0.038043761
hsa04971	Gastric acid secretion	0.045576
hsa04726	Serotonergic synapse	0.045974199
hsa00910	Nitrogen metabolism	0.047420936
hsa05033	Nicotine addiction	0.047658725
hsa05031	Amphetamine addiction	0.048006097

As shown in Table 2, of 21 functional categories, 3 are associated with cancer (cAMP signaling pathway (Akgul, 2009; Myklebust et al., 1999), fatty acid degradation (Currie et al., 2013), and nitrogen metabolism (Sanchez-Vega et al., 2018)) and 11 appear in colorectal cancer studies, including calcium signaling pathway (Dallol et al., 2003), circadian entrainment (Lévi et al., 2010), ribosome biogenesis in eukaryotes (Lafontaine, 2015), cGMP-PKG signaling pathway (Li et al., 2013), dopaminergic synapse (Xu et al., 2010), retrograde endocannabinoid signaling (Proto et al., 2012), cholinergic synapse (Frucht et al., 1999), insulin secretion (Giovannucci, 2001), gastric acid secretion (Morton et al., 2011), morphine addiction (Jin et al., 2012), and nicotine addiction (Shureiqi et al., 2003; Yang and Frucht, 2001). The remaining signaling pathways can be considered as candidates for studies on biological processes associated with colorectal cancer.

In detail, the circadian entrainment pathway closely interacts with the cell division cycle and pharma-cological pathways in the treatment of metastatic colorectal cancer and accelerates or slows down cancer growth through modifications of host and tumor circadian clocks, which drives 24h changes in drug metabolism, cellular proliferation and apoptosis, cell cycle events, DNA repair, and angiogenesis (Lévi et al., 2007, 2010). For the insulin secretion pathway, because insulin and insulin-like growth factor axes are major determinants of cell proliferation and apoptosis, an increase in their circulating concentrations is associated with a high risk of colonic neoplasia. However, the dietary pattern with high saturated fatty acid intake can stimulate insulin resistance or secretion (Giovannucci, 2001), and cellular proliferation requires fatty acids to synthesize cell membranes and signaling molecules (Currie et al., 2013). In addition, the growth of colon tumor cells is selectively inhibited by nonsteroidal anti-inflammatory drugs that activate the cGMP/PKG pathway to suppress Wnt/β-catenin signaling (Li et al., 2013). Up to 15% of colorectal cancers are distinguished by DNA microsatellite instability and manifested by the presence of DNA replication errors (Jass et al., 1998).

DISCUSSION

SimSIP is a novel similarity score that measures the genetic relationships among diseases by (1) introduc-ing a suitable transformation of gene ranks that converts gene ranks into the magnitude of association of gene with the disease; and (2) comparing the similarity between gene rank lists in geometric space instead of comparing the overlaps between the top part of ranked gene lists. In this study, three additional tasks are also fulfilled: constructing a disease network of 18 cancers and offering some support for pan-cancer studies; developing a measure *MAG* to gauge the magnitude of association of an individual gene with two diseases (note that *MAG* can be generalized to multiple diseases); and finding some important genes and signaling pathways associated with colorectal cancer.

Extensive simulations show that *SimSIP* has better performance than existing methods in scenarios with larger numbers of overlapping associated genes (o) and larger number of genes (n), whereas in scenarios with smaller o and smaller n, such as $o = 1$, $n = 6000$ or 11000, *WeiSumE** performs better. It is desirable that *SimSIP* can clearly identify the differences among diseases with more overlapping associated genes. Thus, in real data application, we use *SimSIP* to measure genetic similarities among cancers based on the differential expression analyses of multiple datasets in TCGA with the R package LIMMA. The results show that *SimSIP* can find more significant cancer pairs than *WeiSumE**, such as cancer pair *COAD* and *UCEC*, *COAD* and *THCA*, *LUSC* and *PRAD*, *BRCA* and *THCA*, or *KIRC* and *PRAD* demonstrating the usefulness of *SimSIP*. Furthermore, for the 81 cancer pairs found with both *SimSIP* and *WeiSumE**, the p value of *SimSIP* is smaller than *WeiSumE**, therefore, the *SimSIP* may be more powerful than *WeiSumE** in detecting the significant cancer pairs.

Overall, *SimSIP* has a simpler form and faster computation (time consumption for six measures is shown in Supplementary Table S5), is more robust for higher levels of noise, and is more suitable for the study of human diseases than existing methods, especially for the study of cancers in which there are more genetic overlaps. In order to make this conclusion considerably stronger, we extend the range of the simulations to cover bigger number of overlapping associated genes, such as $o = 50, 100$ and 200. As shown in Supplementary Figure S15-S17, *SimSIP* still performs well or better in the scenarios with bigger ratio of o to d when $o = 50, 100$ and 200, which is analogous to previous simulations.

CONCLUSIONS

This article proposes a similarity score based on lists of gene ranks to measure the genetic relationships among diseases from gene expression data. Our method *SimSIP* gives a new perspective to detect the genetic relationships among diseases that does not depend on a threshold as fraction enrichment (Serra et al., 2016) nor on the weighted sum of overlaps between ranked gene lists as *OrderedList* (Yang et al., 2006).

Similar to other rank-based measures, our method relies on the correctness and scientific quality of the gene ranking. If gene ranking does not reflect the contribution of the gene to each disease, the rank-based measure is not necessarily superior to the commonly used measures. There are many common methods for ranking genes in practice, such as by the magnitude of *p* values of t-test, of marginal regression analysis and of some methods for detecting differentially expressed genes (say LIMMA, edgeR and DESeq2 et al.). Although *p* values derived from the R package LIMMA for differential expression analysis were used in this paper to rank the gene, different researchers can choose different ranking methods based on their specific needs. In summary, in contrast to existing measures that are all based on the number of overlapping genes in top ranked gene lists among diseases, we creatively describe the genetic relationships among diseases from the spatial similarity between the transformation of gene ranks, which provides a new research direction for studies of similarity measures to reveal genetic relationships among diseases.

ACKNOWLEDGMENTS

We thank two anonymous reviewers for their constructive comments and suggestions that improve the presentation of the manuscript greatly. we thank our lab partners Liming Li and Yuquan Wang for code guidance, and our colleagues Zhongzhi Wang, Yongjin Zhang, and Wenxi Li for career help.

REFERENCES

Abrams, Z. B., Zucker, M., Wang, M., Taheri, A. A., Abruzzo, L. V., and Coombes, K. R. (2018). Thirty biologically interpretable clusters of transcription factors distinguish cancer type. *BMC Genomics*, 19(1):738.

Akgul, C. (2009). Mcl-1 is a potential therapeutic target in multiple types of cancer. *Cellular and Molecular Life Sciences*, 66(8):1326–1336.

An, X., Zhu, Y., Zheng, T., Wang, G., Zhang, M., Li, J., Ji, H., Li, S., Yang, S., Xu, D., Li, Z., Wang, T., He, Y., Zhang, L., Yang, W., Zhao, R., Hao, D., and Li, X. (2019). An analysis of the expression and association with immune cell infiltration of the cGAS/STING pathway in pan-cancer. *Molecular Therapy-Nucleic Acids*, 14:80–89.

Antosh, M., Fox, D., Cooper, L. N., and Neretti, N. (2013). CORaL: comparison of ranked lists for analysis of gene expression data. *Journal of Computational Biology*, 20(6):433–443.

Bauer, R., Valletta, D., Bauer, K., Thasler, W. E., Hartmann, A., Müller, M., Reichert, T. E., and Hellerbrand, C. (2014). Downregulation of P-cadherin expression in hepatocellular carcinoma induces tumorigenicity. *International Journal of Clinical and Experimental Pathology*, 7(9):6125.

Chen, F., Zhang, Y., Şenbabaoğlu, Y., Ciriello, G., Yang, L., Reznik, E., Shuch, B., Micevic, G., De Velasco, G., Shinbrot, E., Noble, M. S., Lu, Y., Covington, K. R., Xi, L., Drummond, J. A., Muzny, D., Kang, H., Lee, J., Tamboli, P., Reuter, V., Shelley, C. S., Kaipparettu, B. A., Bottaro, D. P., Godwin, A. K., Gibbs, R. A., Getz, G., Kucherlapati, R., Park, P. J., Sander, C., Henske, E. P., Zhou, J. H., Kwiatkowski, D. J., Ho, T. H., Choueiri, T. K., Hsieh, J. J., Akbani, R., Mills, G. B., Hakimi, A. A., Wheeler, D. A., and Creighton, C. J. (2016). Multilevel genomics-based taxonomy of renal cell carcinoma. *Cell Reports*, 14(10):2476–2489.

Chen, Q., Zhou, X. J., and Sun, F. (2015). Finding genetic overlaps among diseases based on ranked gene lists. *Journal of Computational Biology*, 22(2):111–123.

Currie, E., Schulze, A., Zechner, R., Walther, T. C., and Farese Jr, R. V. (2013). Cellular fatty acid metabolism and cancer. *Cell Metabolism*, 18(2):153–161.

Dallol, A., Morton, D., Maher, E. R., and Latif, F. (2003). SLIT2 axon guidance molecule is frequently inactivated in colorectal cancer and suppresses growth of colorectal carcinoma cells. *Cancer Research*, 63(5):1054–1058.

Davis, C. F., Ricketts, C. J., Wang, M., Yang, L., Cherniack, A. D., Shen, H., Buhay, C., Kang, H., Kim, S. C., Fahey, C. C., Hacker, K. E., Bhanot, G., Gordenin, D. A., Chu, A., Gunaratne, P. H., Biehl, M.,

Seth, S., Kaipparettu, B. A., Bristow, C. A., Donehower, L. A., Wallen, E. M., Smith, A. B., Tickoo, S. K., Tamboli, P., Reuter, V., Schmidt, L. S., Hsieh, J. J., Choueiri, T. K., Hakimi, A. A., The Cancer Genome Atlas Research Network, Chin, L., Meyerson, M., Kucherlapati, R., Park, W.-Y., Robertson, A. G., Laird, P. W., Henske, E. P., Kwiatkowski, D. J., Park, P. J., Morgan, M., Shuch, B., Muzny, D., Wheeler, D. A., Linehan, W. M., Gibbs, R. A., Rathmell, W. K., and Creighton, C. J. (2014). The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell*, 26(3):319–330.

Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., and Lempicki, R. A. (2003). DAVID: database for annotation, visualization, and integrated discovery. *Genome Biology*, 4(9):R60.

Dumortier, M., Ladam, F., Damour, I., Vacher, S., Bièche, I., Marchand, N., de Launoit, Y., Tulasne, D., and Chotteau-Lelièvre, A. (2018). ETV4 transcription factor and MMP13 metalloprotease are interplaying actors of breast tumorigenesis. *Breast Cancer Research*, 20(1):73.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, 10(1):48.

Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes. *The Annals of Applied Statistics*, 1(1):107–129.

Frucht, H., Jensen, R. T., Dexter, D., Yang, W.-L., and Xiao, Y. (1999). Human colon cancer cell proliferation mediated by the M3 muscarinic cholinergic receptor. *Clinical Cancer Research*, 5(9):2532–2539.

Giovannucci, E. (2001). Insulin, insulin-like growth factors and colon cancer: a review of the evidence. *The Journal of Nutrition*, 131(11):3109S–3120S.

Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., Shen, R., Taylor, A. M., Cherniack, A. D., Thorsson, V., Akbani, R., Bowlby, R., Wong, C. K., Wiznerowicz, M., Sanchez-Vega, F., Robertson, A. G., Schneider, B. G., Lawrence, M. S., Noushmehr, H., Malta, T. M., The Cancer Genome Atlas Network, S., M, J., Benz, C. C., and Laird, P. W. (2018). Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, 173(2):291–304.

Hong, Y., Ho, K. S., Eu, K. W., and Cheah, P. Y. (2007). A susceptibility gene set for early onset colorectal cancer that integrates diverse signaling pathways: implication for tumorigenesis. *Clinical Cancer Research*, 13(4):1107–1114.

Imai, S., Kobayashi, M., Takasaki, C., Ishibashi, H., and Okubo, K. (2018). High expression of P-cadherin is significantly associated with poor prognosis in patients with non-small-cell lung cancer. *Lung Cancer*, 118:13–19.

Jass, J., Do, K.-A., Simms, L., Iino, H., Wynter, C., Pillay, S., Searle, J., Radford-Smith, G., Young, J., and Leggett, B. (1998). Morphology of sporadic colorectal cancer with DNA replication errors. *Gut*, 42(5):673–679.

Jin, S., Wang, S., Gao, X., and Xie, G. (2012). Effects of postoperative analgesia with dezocine and morphine on T-lymphocyte subsets and NK cell in colorectal cancer patients. *Acta Academiae Medicinae Qingdao Universitatis*, 48(4):364–370.

Lafontaine, D. L. (2015). Noncoding RNAs in eukaryotic ribosome biogenesis and function. *Nature Structural & Molecular Biology*, 22(1):11.

Lee, J. W., Chou, C.-L., and Knepper, M. A. (2015). Deep sequencing in microdissected renal tubules identifies nephron segment–specific transcriptomes. *Journal of the American Society of Nephrology*, 26(11):2669–2677.

Lévi, F., Focan, C., Karaboué, A., de la Valette, V., Focan-Henrard, D., Baron, B., Kreutz, F., and Giacchetti, S. (2007). Implications of circadian clocks for the rhythmic delivery of cancer therapeutics. *Advanced Drug Delivery Reviews*, 59(9-10):1015–1035.

Lévi, F., Okyar, A., Dulong, S., Innominato, P. F., and Clairambault, J. (2010). Circadian timing in cancer treatments. *Annual Review of Pharmacology and Toxicology*, 50:377–421.

Li, F., He, J., Wei, J., Cho, W. C., and Liu, X. (2015). Diversity of epithelial stem cell types in adult lung. *Stem Cells International*, 2015:1–11.

Li, N., Xi, Y., Tinsley, H. N., Gurpinar, E., Gary, B. D., Zhu, B., Li, Y., Chen, X., Keeton, A. B., Abadi, A. H., Moyer, M. P., Grizzle, W. E., Chang, W.-C., Clapper, M. L., and Piazza, G. A. (2013). Sulindac selectively inhibits colon tumor cell growth by activating the cGMP/PKG pathway to suppress wnt/β-catenin signaling. *Molecular Cancer Therapeutics*, 12(9):1848–1859.

Mainardi, S., Mijimolle, N., Francoz, S., Vicente-Dueñas, C., Sánchez-García, I., and Barbacid, M. (2014). Identification of cancer initiating cells in K-Ras driven lung adenocarcinoma. *Proceedings of the*

National Academy of Sciences, 111(1):255–260.

Morton, M., Prendergast, C., and Barrett, T. D. (2011). Targeting gastrin for the treatment of gastric acid related disorders and pancreatic cancer. *Trends in Pharmacological Sciences*, 32(4):201–205.

Myklebust, J. H., Josefsen, D., Blomhoff, H. K., Levy, F. O., Naderi, S., Reed, J. C., and Smeland, E. B. (1999). Activation of the CAMP signaling pathway increases apoptosis in human B-precursor cells and is associated with downregulation of Mcl-1 expression. *Journal of Cellular Physiology*, 180(1):71–80.

Ni, S. and Vingron, M. (2012). R2KS: a novel measure for comparing gene expression based on ranked gene lists. *Journal of Computational Biology*, 19(6):766–775.

Paredes, J., Figueiredo, J., Albergaria, A., Oliveira, P., Carvalho, J., Ribeiro, A. S., Caldeira, J., Costa, Â. M., Simões-Correia, J., Oliveira, M. J., Pinheiro, H., Pinho, S. S., Mateus, R., Reis, C. A., Leite, M., Fernandes, M. S. F., Schmitt, F., Carneiro, F., Figueiredo, C., Oliveira, C., and Seruca, R. (2012). Epithelial E-and P-cadherins: role and clinical significance in cancer. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1826(2):297–311.

Plaisier, S. B., Taschereau, R., Wong, J. A., and Graeber, T. G. (2010). Rank–rank hypergeometric overlap: identification of statistically significant overlap between gene-expression signatures. *Nucleic Acids Research*, 38(17):e169–e169.

Proto, M. C., Gazzerro, P., Di Croce, L., Santoro, A., Malfitano, A. M., Pisanti, S., Laezza, C., and Bifulco, M. (2012). Interaction of endocannabinoid system and steroid hormones in the control of colon cancer cell growth. *Journal of Cellular Physiology*, 227(1):250–258.

Royo, F., Zuñiga-Garcia, P., Torrano, V., Loizaga, A., Sanchez-Mosquera, P., Ugalde-Olano, A., González, E., Cortazar, A. R., Palomo, L., Fernández-Ruiz, S., Lacasa-Viscasillas, I., Berdasco, M., Sutherland, J. D., Barrio, R., Zabala-Letona, A., Martín-Martín, N., Arruabarrena-Aristorena, A., Valcarcel-Jimenez, L., Caro-Maldonado, A., Gonzalez-Tampan, J., Cachi-Fuentes, G., Esteller, M., Aransay, A. M., Unda, M., Falc¨®n-Pérez, J. M., and Carracedo, A. (2016). Transcriptomic profiling of urine extracellular vesicles reveals alterations of CDH3 in prostate cancer. *Oncotarget*, 7(6):6835.

Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W. K., Luna, A., La, K. C., Dimitriadoy, S., Liu, D. L., Kantheti, H. S., Saghafinia, S., Chakravarty, D., Daian, F., Gao, Q., Bailey, M. H., Liang, W.-W., Foltz, S. M., Shmulevich, I., Ding, L., Heins, Z., Ochoa, A., Gross, B., Gao, J., Zhang, H., Kundra, R., Kandoth, C., Bahceci, I., Dervishi, L., Dogrusoz, U., Zhou, W., Shen, H., Laird, P. W., Way, G. P., Greene, C. S., Liang, H., Xiao, Y., Wang, C., Iavarone, A., Berger, A. H., Bivona, T. G., Lazar, A. J., Hammer, G. D., Giordano, T., Kwong, L. N., McArthur, G., Huang, C., Tward, A. D., Frederick, M. J., McCormick, F., Meyerson, M., Cancer Genome Atlas Research Network; Allen, E. M. V., Cherniack, A. D., Ciriello, G., Sander, C., and Schultz, N. (2018). Oncogenic signaling pathways in the cancer genome atlas. *Cell*, 173(2):321–337.

Sell, S. and Dunsford, H. (1989). Evidence for the stem cell origin of hepatocellular carcinoma and cholangiocarcinoma. *The American Journal of Pathology*, 134(6):1347.

Serra, F., Romualdi, C., and Fogolari, F. (2016). Similarity measures based on the overlap of ranked genes are effective for comparison and classification of microarray data. *Journal of Computational Biology*, 23(7):603–614.

Shi, X., Yi, H., and Ma, S. (2014). Measures for the degree of overlap of gene signatures and applications to TCGA. *Briefings in Bioinformatics*, 16(5):735–744.

Shureiqi, I., Jiang, W., Zuo, X., Wu, Y., Stimmel, J. B., Leesnitzer, L. M., Morris, J. S., Fan, H.-Z., Fischer, S. M., and Lippman, S. M. (2003). The 15-lipoxygenase-1 product 13-S-hydroxyoctadecadienoic acid down-regulates PPAR-δ to induce apoptosis in colorectal cancer cells. *Proceedings of the National Academy of Sciences*, 100(17):9968–9973.

Sousa, B., Ribeiro, A. S., Nobre, A. R., Lopes, N., Martins, D., Pinheiro, C., Vieira, A. F., Albergaria, A., Gerhard, R., Schmitt, F., Baltazar, F., and Paredes, J. (2014). The basal epithelial marker P-cadherin associates with breast cancer cell populations harboring a glycolytic and acid-resistant phenotype. *BMC Cancer*, 14(1):734.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.

Sun, L., Hu, H., Peng, L., Zhou, Z., Zhao, X., Pan, J., Sun, L., Yang, Z., and Ran, Y. (2011). P-cadherin promotes liver metastasis and is associated with poor prognosis in colon cancer. *The American Journal*

536   *of Pathology*, 179(1):380–390.

537   Sutherland, K. D., Song, J.-Y., Kwon, M. C., Proost, N., Zevenhoven, J., and Berns, A. (2014). Multiple
538   cells-of-origin of mutant K-Ras–induced mouse lung adenocarcinoma. *Proceedings of the National*
539   *Academy of Sciences*, 111(13):4952–4957.

540   Thorsson, V., Gibbs, D. L., Brown, S. D., Wolf, D., Bortone, D. S., Yang, T.-H. O., Porta-Pardo, E., Gao,
541   G. F., Plaisier, C. L., Eddy, J. A., Ziv, E., Culhane, A. C., Paull, E. O., Sivakumar, I. K. A., Gentles,
542   A. J., Malhotra, R., Farshidfar, F., Colaprico, A., Parker, J. S., Mose, L. E., Vo, N. S., Liu, J., Liu, Y.,
543   Rader, J., Dhankani, V., Reynolds, S. M., Bowlby, R., Califano, A., Cherniack, A. D., Anastassiou, D.,
544   Bedognetti, D., Mokrab, Y., Newman, A. M., Rao, A., Chen, K., Krasnitz, A., Hu, H., Malta, T. M.,
545   Noushmehr, H., Pedamallu, C. S., Bullman, S., Ojesina, A. I., Lamb, A., Zhou, W., Shen, H., Choueiri,
546   T. K., Weinstein, J. N., Guinney, J., Saltz, J., Holt, R. A., Rabkin, C. S., Cancer Genome Atlas Research
547   Network; Lazar, A. J., Serody, J. S., Demicco, E. G., Disis, M. L., Vincent, B. G., and Shmulevich, I.
548   (2018). The immune landscape of cancer. *Immunity*, 48(4):812–830.

549   Van Marck, V., Stove, C., Jacobs, K., Van den Eynden, G., and Bracke, M. (2011). P-cadherin in
550   adhesion and invasion: Opposite roles in colon and bladder carcinoma. *International Journal of Cancer*,
551   128(5):1031–1044.

552   Vieira, A. F., Dionísio, M. R., Gomes, M., Cameselle-Teijeiro, J. F., Lacerda, M., Amendoeira, I., Schmitt,
553   F., and Paredes, J. (2017). P-cadherin: a useful biomarker for axillary-based breast cancer decisions in
554   the clinical practice. *Modern Pathology*, 30(5):698.

555   Von Mering, C., Jensen, L. J., Kuhn, M., Chaffron, S., Doerks, T., Krüger, B., Snel, B., and Bork, P.
556   (2006). STRING 7-ecent developments in the integration and prediction of protein interactions. *Nucleic*
557   *Acids Research*, 35(suppl_1):D358–D362.

558   Wormald, S., Milla, L., and O'Connor, L. (2013). Association of candidate single nucleotide poly-
559   morphisms with somatic mutation of the epidermal growth factor receptor pathway. *BMC Medical*
560   *Genomics*, 6(1):43.

561   Xu, B., Goldman, J., Rymar, V., Forget, C., Lo, P., Bull, S., Vereker, E., Barker, P., Trudeau, L., Sadikot,
562   A., and Kennedy, T. E. (2010). Critical roles for the netrin receptor deleted in colorectal cancer in
563   dopaminergic neuronal precursor migration, axon guidance, and axon arborization. *Neuroscience*,
564   169(2):932–949.

565   Yang, W.-L. and Frucht, H. (2001). Activation of the PPAR pathway induces apoptosis and COX-2
566   inhibition in HT-29 human colon cancer cells. *Carcinogenesis*, 22(9):1379–1383.

567   Yang, X., Bentink, S., Scheid, S., and Spang, R. (2006). Similarities of ordered gene lists. *Journal of*
568   *Bioinformatics and Computational Biology*, 4(03):693–708.

Figure 1

Figure 1

Powers of EucD, FES_0.001, FES_0.01, OrderedList, SimSIP and WeiSumE* when o=1 with 12 scenarios.
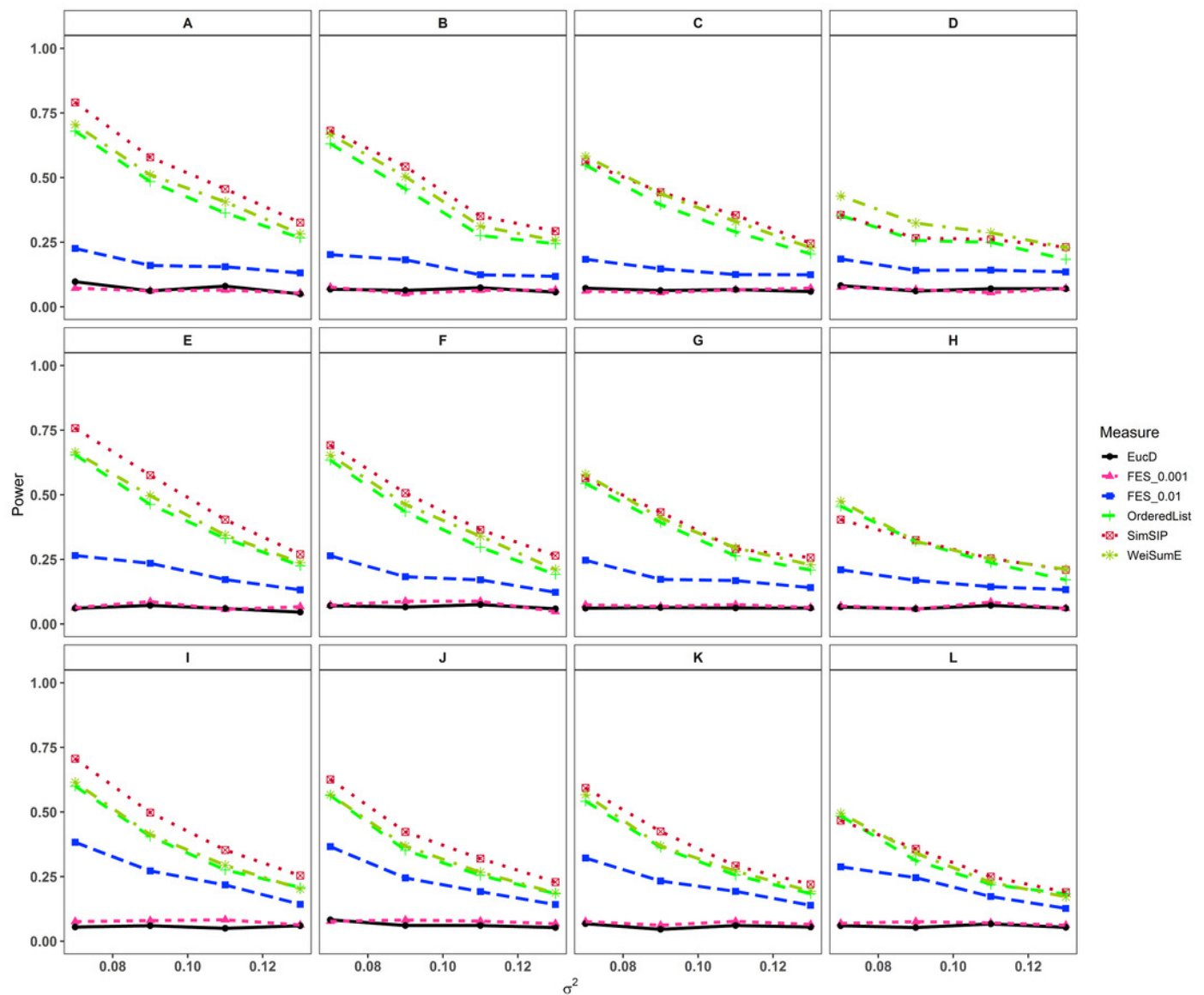
Figure 2

Figure 2

Powers of EucD, FES_0.001, FES_0.01, OrderedList, SimSIP and WeiSumE* when o=5 with 9 scenarios.
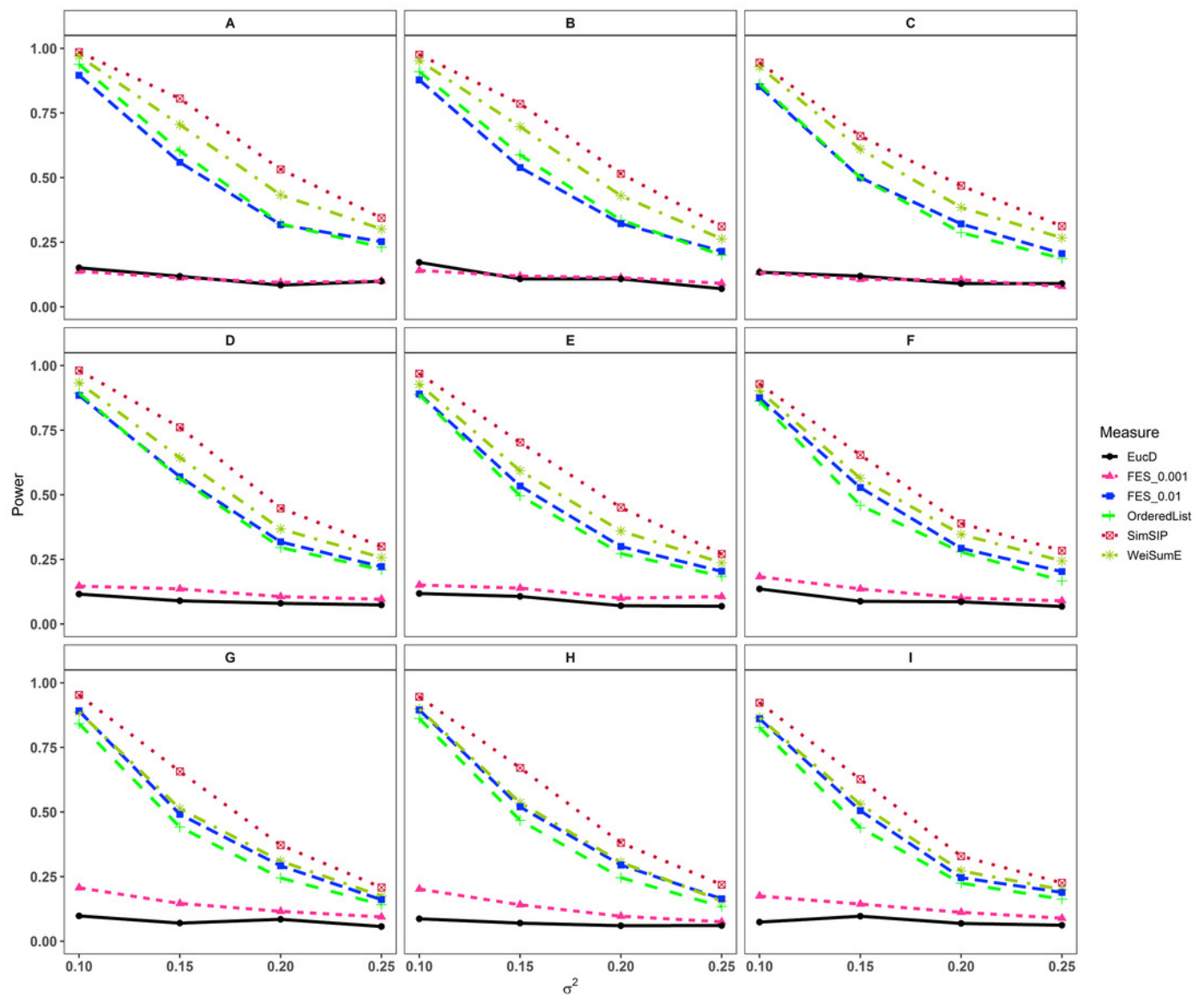
Figure 3

Figure 3

The empirical p values of SimSIP (black dot) and WeiSumE* (red square) for 81 cancer pairs found with both SimSIP and WeiSumE*
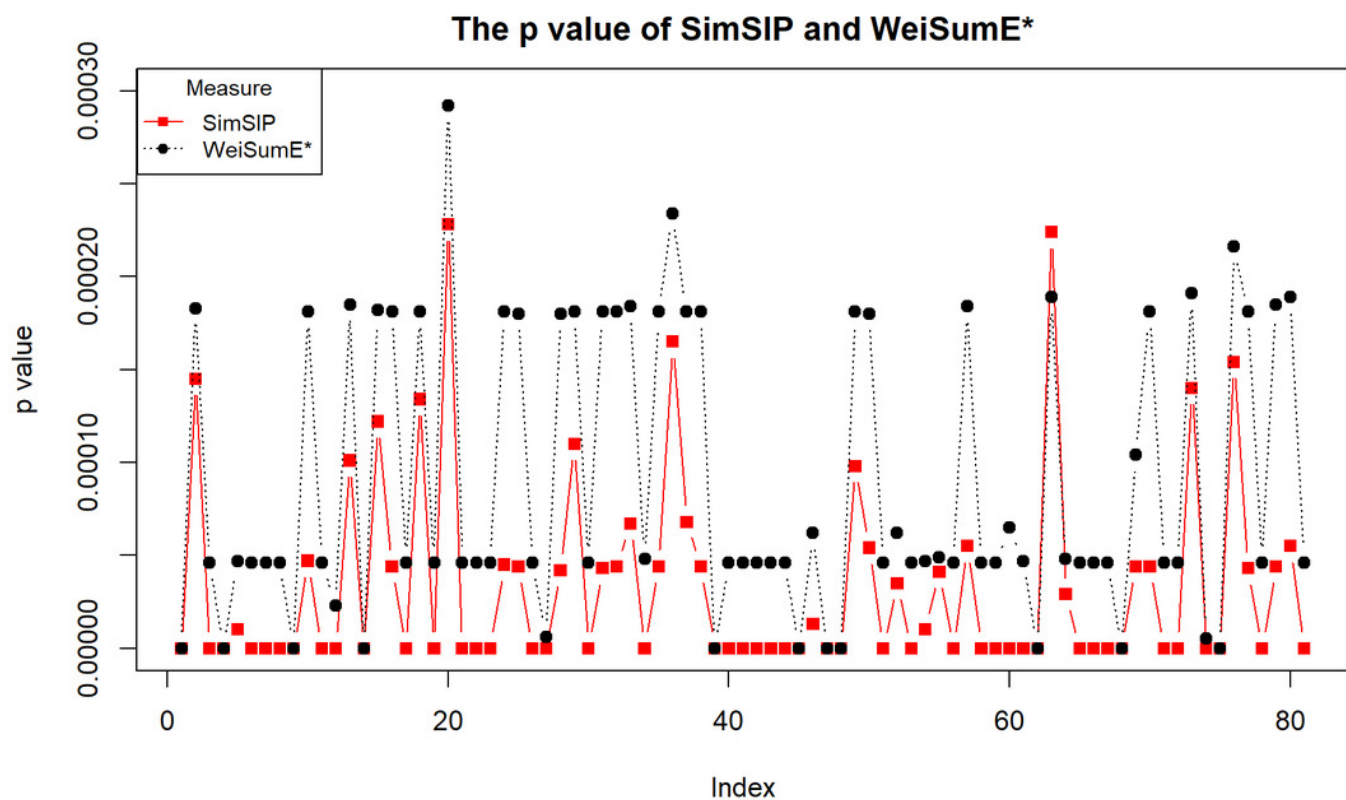
Figure 4

Figure 4

The diseases network of 18 cancer types.