Su et al. BMC Bioinformatics (2018) 19:215
https://doi.org/10.1186/s12859-018-2216-0

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

MGOGP: a gene module-based heuristic algorithm for cancer-related gene prioritization

Lingtao Su[1,2], Guixia Liu[1,2*], Tian Bai[1,2*], Xiangyu Meng[1,2*] and Qingshan Ma[3]

Abstract

Background: Prioritizing genes according to their associations with a cancer allows researchers to explore genes in more informed ways. By far, Gene-centric or network-centric gene prioritization methods are predominated. Genes and their protein products carry out cellular processes in the context of functional modules. Dysfunctional gene modules have been previously reported to have associations with cancer. However, gene module information has seldom been considered in cancer-related gene prioritization.

Results: In this study, we propose a novel method, MGOGP (Module and Gene Ontology-based Gene Prioritization), for cancer-related gene prioritization. Different from other methods, MGOGP ranks genes considering information of both individual genes and their affiliated modules, and utilize Gene Ontology (GO) based fuzzy measure value as well as known cancer-related genes as heuristics. The performance of the proposed method is comprehensively validated by using both breast cancer and prostate cancer datasets, and by comparison with other methods. Results show that MGOGP outperforms other methods, and successfully prioritizes more genes with literature confirmed evidence.

Conclusions: This work will aid researchers in the understanding of the genetic architecture of complex diseases, and improve the accuracy of diagnosis and the effectiveness of therapy.

Keywords: Gene prioritization, Gene module, Gene ontology, Cancer-related genes

Background

Discovering cancer-related genes has profound applications in modelling, diagnosis, therapeutic intervention, and in helping researchers get clues on which genes to explore [1–3]. Computational approaches are preferred due to their high efficiency and low cost [4, 5]. Many computational methods have been proposed, including: a) gene-based function similarity measure methods [6–9]; b) biological interaction network-based methods [10–14], and c) methods based on multiple datasets fusion [15–17]. Methods of the first kind based on the hypothesis that phenotypically similar diseases are caused by functionally related genes. Based on this hypothesis, many methods prioritize genes by computing similarity scores between the candidate genes and the known disease genes. For example, ToppGene [6] ranks genes based on similarity scores of each annotation of each candidate genes by comparing enriched terms in a given set of training genes. Endeavour [8] prioritizes candidate genes by similarity values between candidate genes and seed genes, by integrating more than six types of genomic datasets from over a dozen data sources. Methods of the second kind prioritize genes using the guilt-by-association principle, which means genes interacting with known disease genes are more likely disease-related genes. For instance, PINTA [10] prioritizes candidate genes by utilizing an underlying global protein interaction network. Other methods rank candidate genes by exploiting either local or global network information [2]. Methods of the last kind incorporate datasets such as gene expression, biomedical literature, gene ontology, and PPIs together for gene prioritization. For example, ProphNet [17] integrates information of different types of biological entities in a number of heterogeneous data networks. Taking all these methods into consideration, they are either gene-centric or network-centric.

* Correspondence: lgx1034@163.com; baitian@jlu.edu.cn; 413224445@qq.com
[1]College of Computer Science and Technology, Jilin University, Changchun 130012, China
Full list of author information is available at the end of the article

However, gene module as a basic functional unit of genes has seldom been considered.

Gene module can be defined as a protein complex, a pathway, a sub-network of protein interactions. Module detection has long been studied and many useful algorithms have been proposed, such as [18–21]. Although different methods have different module detection strategies, most of them rely on PPIs network. PPIs network suffers from drawbacks as highlighted in [22]. Firstly, the PPI network is incomplete, which only covers the interactions of well-researched proteins. For instance, of the 20,502 genes in the gene expression matrix downloaded from The Cancer Genome Atlas (TCGA), only 9078 (44.2%) and 2761 (13.4%) genes are included in Human Protein Reference Database (HPRD) [23] and Database of Interacting Proteins (DIP) [24] PPIs networks respectively. As a result, detected modules are incomplete and their accuracy are limited. Secondly, protein interactions in PPIs network suffer from high false positive and negative rates, modules discovered from such PPI data also suffer from high false rates. All these inherent limitations affect the coverage and accuracy of the inferred modules.

Nowadays, numerous public databases of protein and gene annotation information are available, such as Entrez Gene [25], Ensembl [26], PIR iProClass [27], GeneCards [28], KEGG [29], Gene Ontology Consortium [30], DAVID [31], GSEA [32] and UniProt [33]. For instance, DAVID [31] contains information on over 1.5 million genes from more than 65,000 species, with annotation types, including sequence features, protein domain information, pathway maps, enzyme substrates and reaction, protein-protein interaction data and disease associations. Gene Ontology Consortium describes the functions of specific genes, using terms known as GO (Gene Ontology). KEGG map genes to pathways while GSEA provides functional gene groups collected from BioCarta genes sets, KEGG gene sets and Reactome gene sets. With these annotation information, we can easily group genes into functional modules.

Complex diseases, especially cancer are caused by the dysfunction of groups of genes and/or gene interactions rather than the mutations of individual genes. Detecting and prioritizing cancer-related genes from the perspective of gene module is promising. Although some useful work has been conducted [34, 35], the results are still far from being satisfactory. In this study, we take the importance of not only genes but also their affiliated modules into consideration, and prioritizing genes in a heuristic way. We measure module importance by the number of differential genes within the module and the number of differential correlations between the module genes. Besides, the number of known cancer-related genes in the module is also considered. We measure the gene importance by three aspects information: a), gene's differential expression value, b), the number of differential correlations between the gene and all other module gene. c), the fuzzy measure based similarity values between the gene and all known cancer-related genes (if exist) within the module. The global rank of all genes is obtained by utilizing a rank fusion strategy.

Methods

As shown in Fig.1, MGOGP takes gene expression datasets, gene modules, known disease genes and gene ontology

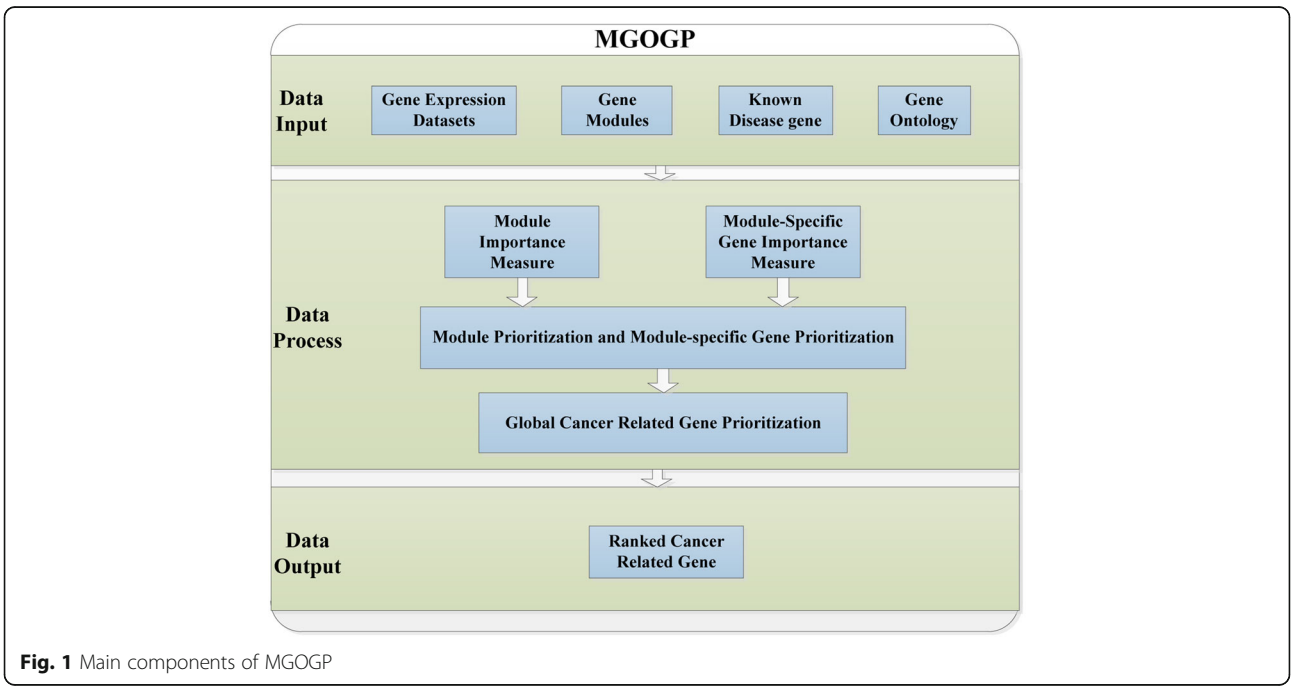

Fig. 1 Main components of MGOGP

annotation information [36] as input, and the ranked genes as output. The main parts including: module importance measure, module-specific gene importance measure, module rank and module-specific gene prioritization, and global cancer-related gene prioritization. Figure 2 schematically illustrates these steps in detail.

First, obtain functional gene modules; then get the global ranking of all modules and the local ranking of all module-specific genes based on their importance; finally, the rank fusion algorithm further gives all genes a global rank.

Input datasets

As shown in Fig. 1, MGOGP takes gene expression datasets, gene modules, known disease-related genes and gene ontology annotation information as input. In this study, all gene modules are downloaded from GSEA website (http://software.broadinstitute.org/gsea/downloads.jsp). All GO ontologies of genes are downloaded from GeneCards [37, 38]. Information of relationships between GO terms are got from Gene Ontology Consortium website.

Module importance measure

We measure the importance of a module by: the number of differentially expressed genes in the module, the number of differential correlations between module genes and the basic importance of the module itself.

We use DESeq2 for gene differential expression analysis [3, 35, 39, 40]. If genes with $padj(g_i)$ value bigger than the threshold value μ, we set $Se(g_i) = 0$. Otherwise, we set $Se(g_i) = 1$, which means the gene g_i is a candidate differential expression gene. $Se(g_i)$ is defined as follows:

$$Se(g_i) = \begin{cases} 0, \text{if } padj(g_i) > \mu \\ 1, \text{else} \end{cases} \quad (1)$$

To further improve the statistical significance of the selected candidate differential expression genes, we applied a multiple random sampling strategy. As defined in Eq. 2.

$$DEG(g_i) = \begin{cases} 0, \text{if } \frac{1}{s}\sum_{s=1}^{S} Se(g_i) < \omega \\ 1, \text{else} \end{cases} \quad (2)$$

Where S is the number of sampling; ω is a threshold value; if a gene g_i is selected as a differential expression gene we set $DEG(g_i) = 1$, Otherwise, we set $DEG(g_i) = 0$.

We define $Ncr(m_j)$ as the ratio of differential expression genes in the module m_j as shown in Eq. 3:

$$Ncr(m_j) = \frac{\sum_{i=1}^{N} DEG(g_i)}{N} \quad j \in 1, 2, 3, \ldots, M \quad (3)$$

Where, g_i is the ith gene in the module m_j; N is the total number of genes in the module m_j; $DEG(g_i)$ is defined in Eq. 2.

Next, for each pair of genes in the module m_j, two correlation values are calculated using normal and tumor samples respectively. As defined in Eqs. 4 and 5 respectively.

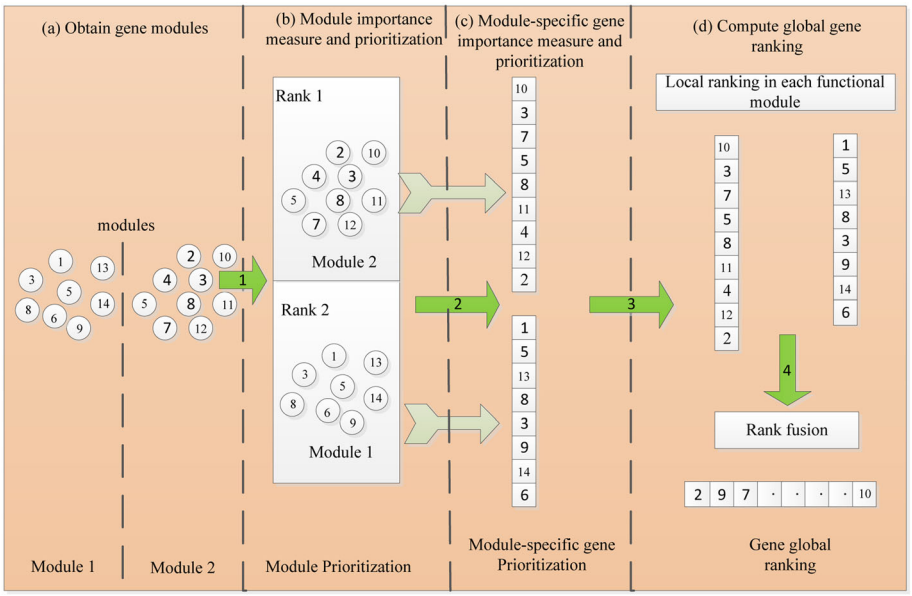

Fig. 2 MGOGP processes are illustrated. **a** Obtain gene modules, **b** Module importance measure and prioritization, **c** Module-specific gene importance measure and prioritization, **d** Compute global gene ranking

$$r_N(g_i,g_h)=\frac{\sum_{l=1}^{L}(x_l-\bar{x})(y_l-\bar{y})}{\sqrt{\sum_{l=1}^{L}(x_l-\bar{x})^2(y_l-\bar{y})^2}} \tag{4}$$

$r_N(g_i, g_h)$ is the Pearson correlation value between gene g_i and gene g_h across all normal samples. L is the normal sample number.

$$r_T(g_i,g_h)=\frac{\sum_{q=1}^{Q}(x_q-\bar{x})\left(y_q-\bar{y}\right)}{\sqrt{\sum_{q=1}^{Q}(x_q-\bar{x})^2\left(y_q-\bar{x}\right)^2}} \tag{5}$$

$r_T(g_i, g_h)$ is the Pearson correlation value between gene g_i and gene g_h across all tumor samples. Q is the tumor sample number.

To test whether the correlation coefficient between gene g_i and gene g_h is differentially correlated, we test whether $r_T(g_i, g_h)$ and $r_N(g_i, g_h)$ are significantly different. The two correlation coefficients are changed to $Z_N(g_i, g_h)$ and $Z_T(g_i, g_h)$ respectively.

$$Z_N(g_i,g_h)=\frac{1}{2}\log\frac{1+r_N(g_i,g_h)}{1-r_N(g_i,g_h)} \tag{6}$$

Similarly, $r_T(g_i, g_h)$ is changed to $Z_T(g_i, g_h)$ as Eq. (6). The differential correlation is tested based on Fisher's z-test [41]. As defined in Eq. (7):

$$Z=\frac{Z_N(g_i,g_h)-Z_T(g_i,g_h)}{\sqrt{\frac{1}{L-3}+\frac{1}{Q-3}}} \tag{7}$$

The Z value has an approximately Gaussian distribution under the null hypothesis [41]. If the *fdr* value of a gene is bigger than the threshold value v, we set $Sc(g_i, g_h) = 0$, otherwise we set $Sc(g_i, g_h) = 1$, which means the correlation coefficient is a potential differential correlation. $Sc(g_i, g_h)$ is defined as follows:

$$Sc(g_i,g_h)=\begin{cases}0, \text{if } fdr(g_i,g_h)>v\\ 1, \text{else}\end{cases} \tag{8}$$

Where $fdr(g_i, g_h)$ is the local false-discovery rate (*fdr*) derived from fdrtool package [42]; v is a threshold value.

As the way we find differential expression genes, we retain only those significantly changed correlations. As defined in Eq. 9:

$$DEE(g_i,g_h)=\begin{cases}0, \text{if } \frac{1}{s}\sum_{s=1}^{S}Sc(g_i,g_h)<\delta\\ 1, \text{else}\end{cases} \tag{9}$$

Where S is the number of sampling; δ is a threshold value; we set $DEE(g_i, g_h) = 1$ if the gene g_i and g_h are differentially correlated. Otherwise, we set $DEE(g_i, g_h) = 0$.

We define $Ecr(m_j)$ as the ratio of differential correlations among genes in the module m_j. $Ecr(m_j)$ is defined in Eq. 10:

$$\begin{aligned}&Ecr(m_j)=\frac{\sum_{k=1}^{K}DEE(g_i,g_h)}{K}\\ &K=\frac{N(N-1)}{2}\ and\ i,h\in 1,2,3,\ldots,N\end{aligned} \tag{10}$$

K and N is the edge number and the gene number of the module m_j respectively.

We measure the basic importance of a module by calculating the ratio of known disease genes in a module, as shown in Eq. 11:

$$info(m_j)=\left(num(d_j)+1\right)/N \tag{11}$$

$num(d_j)$ is the number of known disease genes in the module m_j; N is the number of genes in the module m_j.

The module importance is defined in Eq. 12.

$$\begin{aligned}&p(m_j)=\left((Ncr(m_j)+Ecr(m_j))/2\right)*info(m_j)\\ &j\in 1,2,3\ldots,M\end{aligned} \tag{12}$$

where m_j means the jth module; M is the total number of modules.

Module-specific gene importance measure

We measure the importance of a gene ($p(g_i)$) in the module by measuring: the gene's differential expression value, the number of differential correlations between the gene and all other module genes and the basic importance of the gene itself.

The number of differential correlations ($CorC(g_i)$) between the gene g_i and all other genes in the same module is calculated as in Eq. 13.

$$\begin{aligned}&CorC(g_i)=\frac{\sum_{h=1,h\neq i}^{N-1}Sc(g_i,g_h)}{N-1}\\ &i,h\in 1,2,3,\ldots,N, g_i\in m_j\\ &j\in 1,2,3,\ldots,M\end{aligned} \tag{13}$$

N is the number of genes in the module m_j; M is the total module number.

Finally, the basic importance of a gene is determined by the gene ontology-based fuzzy measure similarity values between the gene and all known disease gene (if exist) in the same module. As shown in Eq. 14.

$$info(m_j\_g_i) = \begin{cases} 0, if\ num(m_j\_d_h) = 0 \\ 1, if\ g_i\ is\ a\ known\ disease\ gene\ itself \\ \sum_{h=1}^{num(m_j\_d_h)} S_{FMS}(g_i, m_j\_d_h)/num(m_j\_d_h), else \end{cases} \quad (14)$$

$num(m_j\_d_h)$ is the number of known disease genes in the module m_j. If $num(m_j\_d_h) = 0$, which means no known disease gene in the module m_j, we set $info(m_j\_g_i) = 0$. If g_i itself is a known disease gene, we set $info(m_j\_g_i) = 1$. Otherwise, we calculate the gene importance value based on the fuzzy similarity measure between the gene and all the known disease gene in the module m_j. $S_{FMS}(\mathrm{m}_j\_g_i, \mathrm{m}_j\_d_h)$ is defined in Eq. 15, as in [43]:

$$S_{FMS}(\mathrm{m}_j\_g_i, \mathrm{m}_j\_d_h) = \frac{Sm_i(T_{\mathrm{m}_j\_g_i} \cap T_{m_j\_d_h}) + Sm_h(T_{\mathrm{m}_j\_g_i} \cap T_{m_j\_d_h})}{2} \quad (15)$$

Where Sm_i is the Sugeno measure [43] defined on GO terms of gene $\mathrm{m}_j\_g_i$ and Sm_h is the Sugeno measure defined on GO terms of module disease gene $m_j\_d_h$.

Let $T_{\mathrm{m}_j g_i}$ is the set of GO annotation terms of gene $\mathrm{m}_j\_g_i$, Sm_i, is a real value function, satisfying [44]:

1) $Sm_i(T_{\mathrm{m}_j g_i}) = 0, if\ T_{\mathrm{m}_j g_i} = \emptyset, else\ Sm_i(T_{\mathrm{m}_j g_i}) = 1.$
2) $Sm_i(T_{\mathrm{m}_j g_i}) \le Sm(T_{m_j d_h})\ if\ T_{\mathrm{m}_j g_i} \subseteq T_{m_j d_h}$
3) For all $T_A, T_B \subseteq T_{\mathrm{m}_j g_i}$ with $T_A \cap T_B = \Phi$
$Sm_i(T_A \cup T_B) = Sm_i(T_A) + Sm_i(T_B)$
$+\lambda Sm_i(T_A) Sm_i(T_B), \lambda > -1$

For a given gene annotation set $T_{\mathrm{m}_j g_i}$, the parameter λ of its Sugeno fuzzy measure can be uniquely solved as in Eq. 16:

$$(1+\lambda) = \prod_{i=1}^{n}(1+\lambda Sm_i) \quad (16)$$

This equation has a unique solution for $\lambda > -1$. Let $Sm_k = Sm(\{T_k\})$. The mapping $T_k \rightarrow Sm_k$ is called a fuzzy density function. The fuzzy density value, Sm_k, is interpreted as the importance of the single information source T_k in determining the similarity of two genes. As defined in Eq. 17:

$$Sm_k = -\ln\left(p(T_k)/\max_{T_j \in T_{g_i}}\{-\ln(p(T_j))\}\right) \quad (17)$$

Where $p(T_k)$ is defined in Eq. 18:

$$p(T_k) = \left(\frac{count(T_k + children\ of\ T_k\ in\ corpus)}{count(all\ GO\ terms\ in\ corpus)}\right) \quad 1 \le k \le |\ T_{g_i}\ | \quad (18)$$

The importance of gene ($p(g_i)$) in a module is defined in Eq. 19.

$$p(g_i) = padj(g_i) + CorC(g_i) + info(g_i) \quad i \in 1, 2, 3, \ldots, N, g_i \in m_j \quad (19)$$

N is the number of genes in the module m_j.

Global gene ranking

Most genes deploy their functions in the context of sophisticated functional modules [45, 46]. Therefore, the global rank of a gene need be decided by its own importance and the importance of its affiliated module. As in [34], a rank fusion strategy is used to fuse the local rank of genes in each module into a global rank.

The rank fusion strategy is a recursive process. It decides the rank of the *nth* gene based on all the top-ranked $n-1$ genes. We define i as the number of genes having already obtained their global ranking in the recursive process of rank fusion, $m(i,j)$ as the number of top i genes located in the module j after having determined the top i genes. $t(i,j)$ as the expectation of the number of top i genes located in the module j. $e(i,j)$ as the expectation of probability that the $i+1$ globally ranked genes come from the module j. We use the module importance value $p(m_j)$ as the probability of a disease-related gene comes from it. The relationship between i, $m(i,j)$, $t(i,j)$ and $p(m_j)$ is shown in Eq. 20:

$$t(i,j) = ip(m_j) \quad e(i,j) = t(i+1, j) - m(i,j) \quad (20)$$

Initially, the first ranked gene in the module with highest importance value is chosen as the top 1 gene in the gene's global rank, because all genes in each module have been ranked from big to small according to their importance value. Let i as the number of genes having obtained their global ranking, to decide the $i+1$ ranked gene, we need to find the module with the biggest $e(i,j)$ value, because $e(i,j)$ indicates the expectation of probability that the $i+1$ globally ranked genes from module j. So the genes ranked $m(i,j)+1$ in the module j will be chosen as the top $i+1$ ranked gene, because in the module j, top $m(i,j)$ genes has obtained the global ranking. Repeat the process until all genes get ranked. As shown in Fig. 3 (in Additional file 1).

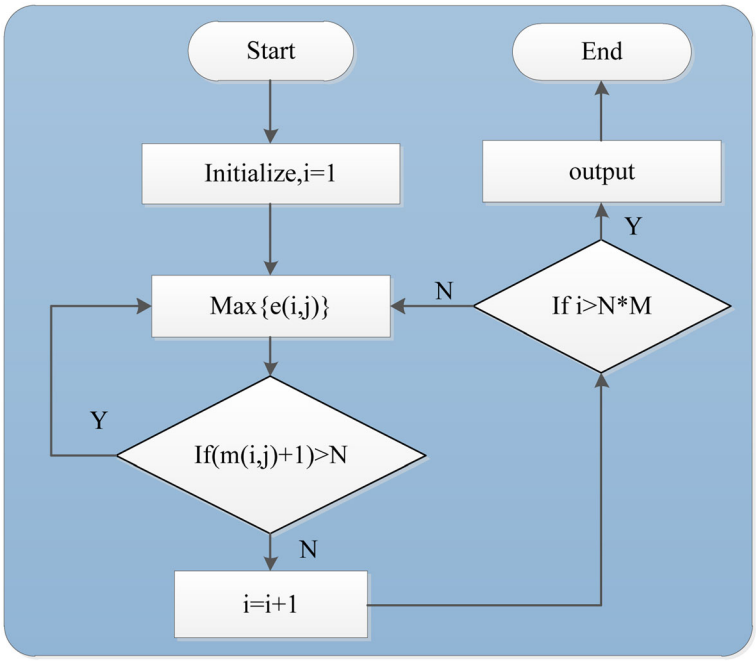

Fig. 3 Rank fusion process. *N* is the number of genes in the module *j*, *M* is the total module number

Results

Both raw count and normalized gene expression datasets are downloaded from TCGA (http://cancergenome.nih.gov/) [47], which include expression values of 20,503 genes across 102 normal samples and 779 tumor samples. Besides, gene expression datasets of Prostate adenocarcinoma containing 483 tumor samples and 51 normal samples are also downloaded from TCGA. Four thousand seven hundred twenty-six gene modules are downloaded from the website of GSEA (in Additional file 2).

Firstly, the performance of MGOGP is validated by comparing it with three module based cancer-related gene prioritization methods (MENDEAVOUR, MDK and MRWR) proposed in [34]. For comparison, the same prostate cancer network used in [34] are used, which consists of 233 genes and 1218 interactions. Modules are obtained by picking out all the GSEA modules that contain more than three genes in the prostate network after removing irrelevant module genes. Irrelevant genes are genes that are included in GSEA modules but are not included in these 233 genes. Fifteen known prostate cancer genes are obtained from OMIM (Table 1). Six genes (BRCA1, TP53, EP300, STAT3, ZFHX3, HNF1B), which are confirmed have associations with prostate cancer by Genetics Home Reference (https://ghr.nlm.nih.gov/) are used as test genes. Results are shown in Table 2.

As shown in Table 2, all the six genes are ranked on average within top10% of all the candidate genes, which indicates the superiority of MGOGP to other three algorithms. For further comparison, we put these 21 genes together, each time we randomly select 20 different genes as known disease genes and the remaining 1 gene

Table 1 Known prostate cancer genes retrieved from the OMIM

Gene ID	Gene Symbol	Gene name
367	AR	Androgen receptor
675	BRCA2	Breast cancer type 2 susceptibility protein
3732	CD82	CD82 antigen
11200	CHEK2	Serine/threonine-protein kinase Chk2
60528	ELAC2	Zinc phosphodiesterase ELAC protein 2
2048	EPHB2	Ephrin type-B receptor 2 precursor
3092	HIP1	Huntingtin-interacting protein 1
1316	KLF6	Krueppel-like factor 6
8379	MAD1L1	Mitotic spindle assembly checkpoint proteinMAD
4481	MSR1	Macrophage scavenger receptor types I and II
4601	MXI1	MAX-interacting protein 1
7834	PCAP	Predisposing for prostate cancer
5728	PTEN	Phosphatase and tensin homolog
6041	RNASEL	2-5A-dependent ribonuclease
5513	HPC1	Hereditary prostate cancer 1

Table 2 Ranks of six test genes in prostate cancer gene network. They are prioritized by MDK, MRWR, Endeavour and MGOGP

Gene	MDK	MRWR	Endeavour	MGOGP
BRCA1	29	6	58	63
TP53	104	132	85	24
EP300	83	70	90	11
STAT3	39	41	88	17
ZFHX3	174	174	34	19
HNF1B	44	190	109	26
Average Rank	78	102	77	26

for test. Each run we compared the ranked positions of the 1 test gene between our method and Endeavour. Results are shown in Table 3. In Table 3 some genes do not exist, because they don't exist in our GSEA gene modules or not exist in Endeavour database. According to Table 3, 11 of the 13 known prostate cancer-related genes and 4 of the 6 test genes have much higher ranks than these of the Endeavour. Moreover, the average ranking of these genes is 51 by MGOGP, which is better than 82 by Endeavour.

Table 3 Ranks of each validation gene

Gene	MGOGP	Endeavour
AR	32	30
BRCA2	29	40
CD82	169	211
CHEK2	19	35
ELAC2	64	176
EPHB2	45	165
HIP1	91	111
KLF6	88	72
MAD1L1	78	194
MSR1	60	190
MXI1	92	89
PCAP	Not Exist	Not Exist
PTEN	24	94
RNASEL	67	83
HPC1	Not Exist	Not Exist
BRCA1	46	16
TP53	5	5
EP300	11	12
STAT3	17	23
ZFHX3	59	68
HNF1B	26	12

Next, we use MGOGP for genome-wide breast cancer gene prioritization. We use 328 breast disease-related genes downloaded from SNP4Disease (http://snp4disease.mpibn.mpg.de/result.php) as seed genes (see Additional file 3). Ten well-known breast cancer-related genes (shown in Table 4, which are not contained in the 328 genes) are used to validate the effectiveness of our method. All GSEA gene modules are pre-processed by removing all the genes which do not have gene expression information (the final module list is supplied in Additional file 4). The result is shown in Fig. 4.

As shown in Fig. 4, all the 10 breast cancer-related genes are ranked within the top5% of the gene prioritization results. During the process, we set $S = 1000$, $\omega = 0.9$ and $\delta = 0.9$ (which means of the 1000 sampling results, over 90% fulfill the filter criteria). We set $\upsilon = 0.05$ and $\mu = 0.01$ as most others do [39, 41]. The performance of MGOGP under different parameter settings are supplied in Additional file 5. The top 10 ranked modules in this case study are shown in Table 5.

As can be seen from Table 5, many top-ranked modules are included in well-known breast cancer pathways, such as PI3K/AKT [48] pathway and VEGF ligand-receptor pathway. The VEGF family of ligands and receptors are intimately involved in tumor angiogenesis, lymphangiogenesis, and metastasis [49]. More importantly, of the 100 genes in the top 10 ranked modules, 20 of them are contained in the KEGG breast cancer pathway (hsa05224), which is an indication of the good performance of MGOGP for cancer gene prioritization.

Next, we validate the performance of MGOGP by comparing the gene prioritization results with results obtained by methods: Endeavour [8], GeneFriends [50], PINTA [10], TOPPGene [6] and TOPNet [13]. All the methods use the same datasets and under their default parameter

Table 4 Ten well-known breast cancer genes

Gene ID	Gene symbol	Gene name
672	BRCA1	Breast Cancer 1, Early Onset
675	BRCA2	Breast Cancer 2, Early Onset
7157	TP53	Tumor Protein P53
5728	PTEN	Phosphatase And Tensin Homolog
841	CASP8	Caspase 8, Apoptosis-Related Cysteine Peptidase
2263	FGFR2	Fibroblast Growth Factor Receptor 2
4214	MAP3K1	Mitogen-Activated Protein Kinase Kinase Kinase 1, E3 Ubiquitin Protein Ligas
11200	CHEK2	Checkpoint Kinase 2
472	ATM	ATM Serine/Threonine Kinase
83990	BRIP1	BRCA1 Interacting Protein C-Terminal Helicase 1

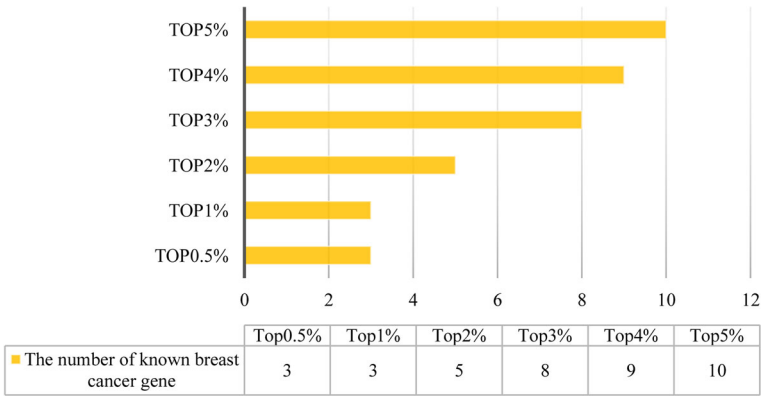

Fig. 4 Known cancer-related gene prioritization result

settings. The results are shown in Fig. 5. Brief descriptions of these methods are provided in Additional file 6. Core sourcecode of MGOGP is provided in Additional file 7. Other source codes are available from the corresponding author on reasonable request.

In Fig. 5, we count the number of breast cancer-related genes in the gene prioritization results. As is shown in Fig. 5, MGOGP outperforms other methods in detecting cancer-related genes. We use all the 328 breast disease related genes as known disease gene (Endeavour and GeneFriends used the same gene sets) and count the number of known disease genes appear in top 100–1000 prioritization results.

To do comparison more rigorously, we further compare MGOGP to Endeavour, TOPNet and TOPPGene. Each time we randomly select 100, 150 and 200 different known disease genes from the 328 breast disease-related genes for known disease genes and others are left for test (each kind of selection repeat 100 times). We count the average number of test genes appear in Top 200 gene prioritization results. Results are shown in Fig. 6.

Finally, to further validate our method, we get the top 10 ranked genes of each method in Fig. 5. The results are shown in Table 6.

In Fig. 7, the number of Known Disease Gene is the number of genes supplied for training each method that fall within the top 10. For example, in Table 6, PTEN, VEGFB, and MCM2 are three genes fall within the top 10 of the gene ranking result, so the number of Known Disease Gene of MGOGP in Fig. 7 is 3. For each gene within the top 10 gene ranking results of each method, we search the number of articles in PubMed mention the association between the gene and breast cancer. We count the number of genes has more than 10 PubMed article reference. As shown in Fig. 7, genes detected by MGOGP have more article supports than other methods.

Table 5 Top 10 ranked modules

Rank	Module name	Gene number	Importance value
1	zerbini_response_to_sulindac_dn	6	0.542
2	reichert_g1s_regulators_as_pi3k_targets	8	0.523
3	sa_g2_and_m_phases	8	0.492
4	reactome_vegf_ligand_receptor_interactions	10	0.478
5	biocarta_srcrptp_pathway	11	0.461
6	honrado_breast_cancer_brca1_vs_brca2	18	0.447
7	tcga_glioblastoma_mutated	8	0.445
8	pid_vegf_vegfr_pathway	10	0.444
9	liang_silenced_by_methylation_dn	11	0.411
10	agarwal_akt_pathway_targets	10	0.410

Discussion and conclusion

Results of omics experiments commonly consist of a large set of genes, while researchers usually only care about the behaviour of several genes. In this paper, a heuristic algorithm is proposed for prioritizing disease-associated genes by utilizing gene ontology annotation information and known disease-related genes as heuristic information. Different from existing methods, we propose to rank genes considering the importance of both individual genes and their affiliated modules, and utilize Gene Ontology (GO) based fuzzy measure value as well as known disease genes as heuristics, and use rank fusion strategy to obtain the global gene prioritization. Results show that MGOGP

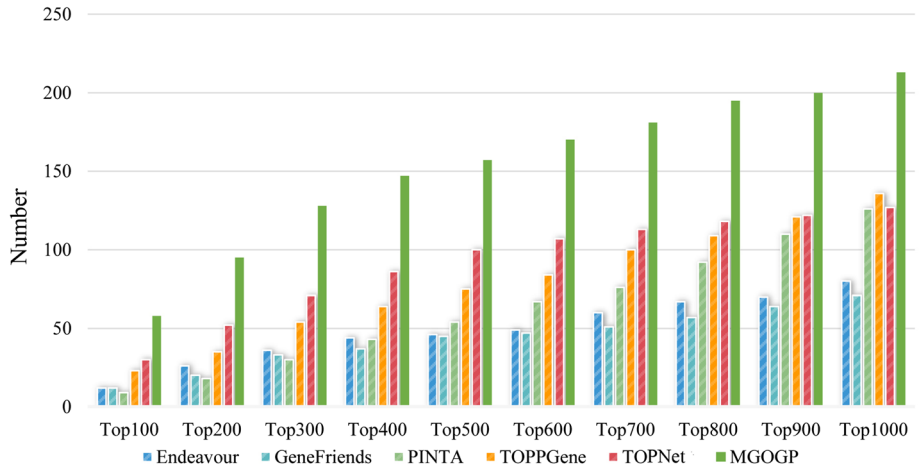

Fig. 5 Comparison results between 6 methods. Endeavour, GeneFriends, PINTA, TOPPGene, TOppNet, and MGOGP

outperforms many other methods in cancer-related gene prioritization.

Different from other module-based gene prioritization methods, where modules are detected by partitioning the network using the network clustering methods, we obtain modules through gene function annotation, that is, functionally related genes are grouped into the same modules. Because gene interaction networks often suffer from the problems of high rates of false positive/negative interactions, and modules detected by network clustering algorithms often have limited accuracy, so our method is more advanced. One important difference between modules used in this study and modules detected through network partition is that no edges in our module. Instead, we use statistical methods detecting differential correlations between genes within a module, which could help avoid the preference of genes or modules that are well-researched (because currently obtained network is far from complete, the number of interactions among well-researched genes may be much more than that of newly discovered genes).

Different from module-based methods in [34], MGOGP ranks modules considering three aspects of information: module-specific gene importance, differential correlations, and importance of the module itself. In [34], the author considers the importance of a module by considering only the number of disease genes and the size of the module, which may bias toward big modules. Furthermore, gene as the major component of the module whose importance is not considered when measuring the importance of a module in [34]. While in our method, when measuring the importance of a module, we consider: the importance of the module itself, the importance of module contained genes as well as differential correlations within the module, which are the main improvements of our method.

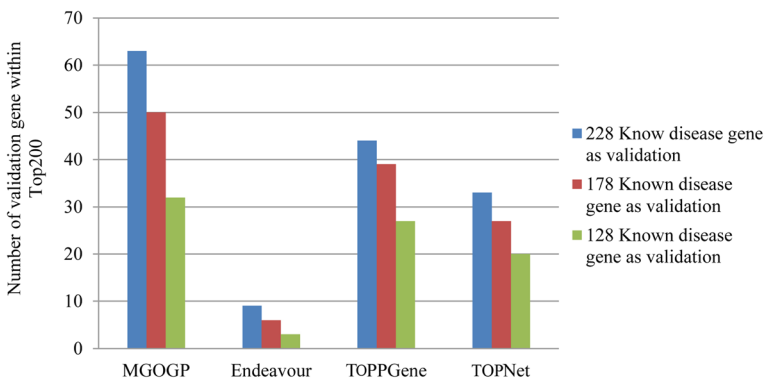

Fig. 6 Comparison results between MGOGP, Endeavour, TOPPGene, and TOPNet with different number of known disease genes as input

Table 6 Top 10 ranked genes of each method

	MGOGP	Endeavor	GeneFriends	PINTA	ToppGene	ToppNet
Top 10 gene	CCNB1IP1	SNRPF	LURAP1L	MGP	RAD51	APP
	CCNE2	BUB3	PVRL2	EEF1A1	APEX1	ELAVL1
	NEK1	MSH2	CYFIP1	TPT1	SIRT2	NTRK1
	NRP1	SSBP1	FAM120A	RPS6	NOC2L	RPA1
	CDC25C	RFC4	IL13RA1	RPL3	NEDD1	XPO1
	VIM	EZH2	MYO1B	RPS27	TERT	EED
	PTEN	CENPF	BCL9L	ACTB	EPN3	CUL3
	VEGFB	BLMH	NQO1	SCGB2A2	PPARGC1A	BARD1
	MCM2	KIF20B	RIN2	RPL11	NBN	HSP90AA1
	PTGS2	BAZ1A	SDC4	PIP	ATR	NXF1
Known disease genes fall in the top 10 gene	PTEN	MSH2	NQO1	SCGB2A2	RAD5	BARD1
	VEGFB	EZH2		PIP	TERT	
	MCM2				NBN	
					ATR	

In Table 6, each method is run with default parameter settings and use same training genes. Top 10 gene means the top 10 genes prioritized by each method and Known disease genes fall in the top 10 gene means genes supplied for training each method falls in the top 10 genes. Detail statistic results are shown in Fig. 7

Compared with other non-module-based prioritization methods, our algorithm also has obvious advantages. First, it is easier to find the potential pathogenic genes that cause the disease from the point of view of gene modules. Second, it takes cross-validation strategy which could guarantee the stability of the recognition results. And our method works with heuristic information which could effectively avoid the blindness of the search.

By applying MGOGP on different datasets, we demonstrate that MGOGP performs better than previous gene or network-centric methods in terms of potential disease-related genes prediction. Firstly, the performance of MGOGP is validated by comparing it with three module based cancer-related gene prioritization methods. Results show that all test genes are ranked on average within top10% of all the candidate genes. According to our results, many top-ranked modules are included in well-known cancer pathways, and top-ranked genes have more supporting PubMed articles. All of the results show that our methods perform better than the state of the art methods.

Prioritization methods are useful for assisting scientists at early research stages, and to formulate novel hypotheses of interest. In the future, one of our main goals is to see how our method behaves in other prioritization problems when using different entities and sources of data sets not covered in this study. Furthermore, we plan to study in more detail the quality of the datasets and their influence on algorithm performance, and design new methods to try to improve the results. As we all know that the methods become more mature the results will become increasingly accurate and more biologically meaningful.

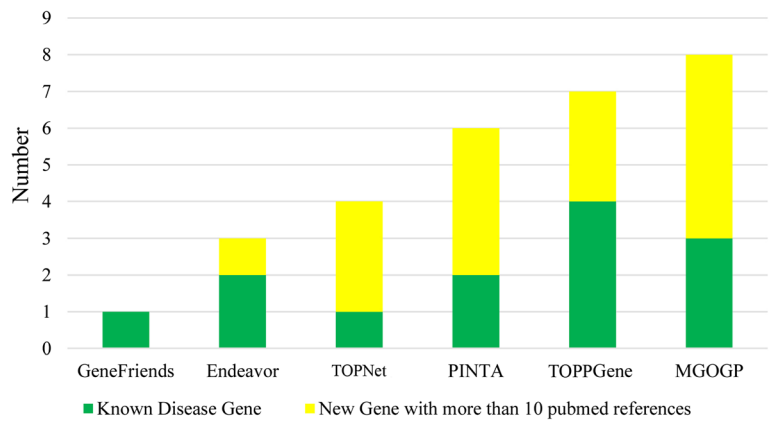

Fig. 7 Detail statistic results of results in Table 6

Additional files

Additional file 1: A step by step example of Rank Fusion process. This file provides an example of how to get the final gene rank. (DOCX 275 kb)

Additional file 2: GSEA gene module. This file is all the gene modules downloaded from GSEA website. (TXT 2837 kb)

Additional file 3: Breast-Cancer-Gene. This is the known breast cancer-related genes downloaded from SNP4Disease. (TXT 2 kb)

Additional file 4: Final module list. This is the refined module list after removing irrelevant genes. (TXT 2736 kb)

Additional file 5: Parameters discussion. This file discusses the performance of MGOGP under different parameter settings. (DOCX 65 kb)

Additional file 6: Brief description of gene prioritization methods. This file provides the short description of comparison methods, including their input datasets, limitations, and type. (DOCX 17 kb)

Additional file 7: Sourcecode. Some core code of our method. (TXT 5 kb)

Acknowledgments
The results here are in whole or part based upon datasets generated by the TCGA Research Network: http://cancergenome.nih.gov/.

Funding
This work is supported by Graduate Innovation Fund of Jilin University (No. 2016031); The National Nature Science Foundation of China (No. 61373051, No. 61502343, No. 61772226 and No. 61702214); Science and Technology Development Program of Jilin Province (No. 20140204004GX); The Science Research Funds for the Guangxi Universities (No. KY2015ZD122); The Science Research Funds for the Wuzhou University (2014A002); Project of Science and Technology Innovation Platform of Computing and Software Science (985 Engineering); The Key Laboratory for Symbol Computation and Knowledge Engineering of the National Education Ministry of China; The Fundamental Research Funds for the Central. China Postdoctoral Science Foundation (No. 2014M561293); Development Project of Jilin Province of China (No. 20150520064JH).

Availability of data and materials
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Authors' contributions
LS made contributions to method design and data analysis, and a major contributor in writing the manuscript. GL involved in drafting the manuscript and revision. TB analyzed the results and made contributions to method implementation. XM performed comparative analysis. QM made contributions to results interpretation and also involved in data acquisition and manuscript writing. All authors read and approved the final manuscript.

Ethics approval and consent to participate
In this study, all gene expression datasets were downloaded from TCGA database (https://tcga-data.nci.nih.gov/tcga/). There are no restrictions on the use of TCGA data for research and data analysis purposes. All datasets can be downloaded and used freely, and not require an ethics statement.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]College of Computer Science and Technology, Jilin University, Changchun 130012, China. [2]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China. [3]The First Clinical Hospital of Jilin University, Changchun 130021, China.

Received: 13 March 2017 Accepted: 23 May 2018
Published online: 05 June 2018

References
1. Gill N, Singh S, Aseri TC. Computational disease gene prioritization: an appraisal. J Comput Biol. 2014;21(6):456–65.
2. Moreau Y, Tranchevent LC. Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nat Rev Genet. 2012;13(8):523–36.
3. Cruz-Monteagudo M, Borges F, Paz YMC, Cordeiro MN, Rebelo I, Perez-Castillo Y, Helguera AM, Sanchez-Rodriguez A, Tejera E. Efficient and biologically relevant consensus strategy for Parkinson's disease gene prioritization. BMC Med Genet. 2016;9:12.
4. Bromberg Y. Chapter 15: disease gene prioritization. PLoS Comput Biol. 2013;9(4):e1002902. https://doi.org/10.1371/journal.pcbi.1002902.
5. Doncheva NT, Kacprowski T, Albrecht M. Recent approaches to the prioritization of candidate disease genes. Wiley Interdiscip Rev Syst Biol Med. 2012;4(5):429–42.
6. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. 2009;37:W305–11.
7. Schlicker A, Lengauer T, Albrecht M. Improving disease gene prioritization using the semantic similarity of Gene Ontology terms. Bioinformatics. 2010; 26(18):i561–7.
8. Tranchevent LC, Barriot R, Yu S, Van Vooren S, Van Loo P, Coessens B, De Moor B, Aerts S, Moreau Y. ENDEAVOUR update: a web resource for gene prioritization in multiple species. Nucleic Acids Res. 2008;36:W377–84.
9. Yu W, Wulf A, Liu T, Khoury MJ, Gwinn M. Gene Prospector: an evidence gateway for evaluating potential susceptibility genes and interacting risk factors for human diseases. BMC Bioinformatics. 2008;9:528.
10. Nitsch D, Tranchevent LC, Goncalves JP, Vogt JK, Madeira SC, Moreau Y. PINTA: a web server for network-based gene prioritization from expression data. Nucleic Acids Res. 2011;39(Web Server issue):W334–8.
11. Xie B, Agam G, Balasubramanian S, Xu J, Gilliam TC, Maltsev N, Bornigen D. Disease gene prioritization using network and feature. J Comput Biol. 2015; 22(4):313–23.
12. Navlakha S, Kingsford C. The power of protein interaction networks for associating genes with diseases. Bioinformatics. 2010;26(8):1057–63.
13. Chen J, Aronow BJ, Jegga AG. Disease candidate gene identification and prioritization using protein interaction networks. BMC Bioinformatics. 2009;10:73.
14. Erten S, Bebek G, Ewing RM, Koyuturk M. DADA: degree-aware algorithms for network-based disease gene prioritization. BioData mining. 2011;4(19). https://doi.org/10.1186/1756-0381-4-19.
15. Wu C, Zhu J, Zhang X. Integrating gene expression and protein-protein interaction network to prioritize cancer-associated genes. BMC Bioinformatics. 2012;13:182.
16. Simoes SN, Martins DC Jr, Pereira CA, Hashimoto RF, Brentani H. NERI: network-medicine based integrative approach for disease gene prioritization by relative importance. BMC Bioinformatics. 2015;16(Suppl 19):S9.
17. Martínez V, Cano C, Blanco A. ProphNet: a generic prioritization method through propagation of information. BMC Bioinformatics. 2014;15(Suppl 1): S5. doi:https://doi.org/10.1186/1471-2105-15-S1-S5.
18. Zhang Y, Lin H, Yang Z, Wang J. Integrating experimental and literature protein-protein interaction data for protein complex prediction. BMC Genomics. 2015;16(Suppl 2):S4.
19. Srihari S, Yong CH, Patil A, Wong L. Methods for protein complex prediction and their contributions towards understanding the organisation, function and dynamics of complexes. FEBS Lett. 2015;589(19 Pt A):2590–602.
20. Su L, Liu G, Wang H, Tian Y, Zhou Z, Han L, Yan L. GECluster: a novel protein complex prediction method. Biotechnol Biotechnol Equip. 2014;28(4):753–61.
21. Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics. 2003;4:2.
22. Ramaprasad A, Pain A, Ravasi T. Defining the protein interaction network of human malaria parasite plasmodium falciparum. Genomics. 2012;99(2):69–75.
23. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, et al. Human protein reference database–2009 update. Nucleic Acids Res. 2009; 37(Database):D767–72.
24. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res. 2002;30(1):303–5.

25. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez gene: gene-centered information at NCBI. Nucleic Acids Res. 2011;39:D52–7.
26. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, et al. Ensembl 2011. Nucleic Acids Res. 2011; 39(Database):D800–6.
27. Wu CH, Huang H, Nikolskaya A, Hu Z, Barker WC. The iProClass integrated database for protein functional analysis. Comput Biol Chem. 2004;28(1):87–96.
28. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: integrating information about genes, proteins and diseases. Trends Genet. 1997;13(4):163.
29. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44(D1):D457–62.
30. Gene Ontology C. Gene ontology consortium: going forward. Nucleic Acids Res. 2015;43(Database issue):D1049–56.
31. Huang DW, Sherman BT, Tan Q, Kir J, Liu D, Bryant D, Guo Y, Stephens R, Baseler MW, Lane HC, et al. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res. 2007;35(Web Server issue):W169–75.
32. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–50.
33. UniProt C. UniProt: a hub for protein information. Nucleic Acids Res. 2015; 43(Database issue):D204–12.
34. Chen X, Yan GY, Liao XP. A novel candidate disease genes prioritization method based on module partition and rank fusion. OMICS. 2010;14(4):337–56.
35. Liu X, Liu ZP, Zhao XM, Chen L. Identifying disease genes and module biomarkers by differential interactions. J Am Med Inform Assoc. 2012; 19(2):241–8.
36. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25–9.
37. Belinky F, Nativ N, Stelzer G, et al. PathCards: multi-source consolidation of human biological pathways. Database: J Biol Databases and Curation. 2015; 2015:bav006. doi:https://doi.org/10.1093/database/bav006.
38. Rappaport N, Twik M, Nativ N, Stelzer G, Bahir I, Stein TI, Safran M, Lancet D. MalaCards: a comprehensive automatically-mined database of human diseases. Curr Protoc Bioinformatics/editoral board, Andreas D Baxevanis [et al]. 2014;47:1.24.21–19.
39. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12). https://doi.org/10.1101/002832.
40. Wen Z, Liu ZP, Liu Z, Zhang Y, Chen L. An integrated approach to identify causal network modules of complex diseases with application to colorectal cancer. J Am Med Inform Assoc. 2013;20(4):659–67.
41. Fukushima A. DiffCorr: an R package to analyze and visualize differential correlations in biological networks. Gene. 2013;518(1):209–14.
42. Strimmer K. fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. Bioinformatics. 2008;24(12):1461–2.
43. Popescu M, Keller JM, Mitchell JA. Fuzzy measures on the Gene Ontology for gene product similarity. IEEE/ACM Trans Comput Biol Bioinform. 2006;3(3):263–74.
44. Chen J, Xu H, Aronow BJ, Jegga AG. Improved human disease candidate gene prioritization using mouse phenotype. BMC Bioinformatics. 2007;8:392.
45. Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R. Associating genes and protein complexes with disease via network propagation. PLoS Comput Biol. 2010;6(1):e1000641.
46. Wang L, Sun FZ, Chen T. Prioritizing functional modules mediating genetic perturbations and their phenotypic effects: a global strategy. Genome Biol. 2008;9(12):R174. doi:https://doi.org/10.1186/gb-2008-9-12-r174.
47. Zhu Y, Qiu P, Ji Y. TCGA-assembler: open-source software for retrieving and processing TCGA data. Nat Methods. 2014;11(6):599–600.
48. Mukohara T. PI3K mutations in breast cancer: prognostic and therapeutic implications. Breast Cancer (Dove Med Press). 2015;7:111–23.
49. Eppenberger M, Zlobec I, Baumhoer D, Terracciano L, Lugli A. Role of the VEGF ligand to receptor ratio in the progression of mismatch repair-proficient colorectal cancer. BMC Cancer. 2010;10:93.
50. van Dam S, Craig T, de Magalhaes JP. GeneFriends: a human RNA-seq-based gene and transcript co-expression database. Nucleic Acids Res. 2015; 43(Database issue):D1124–32.