Aouiche *et al. BMC Bioinformatics* 2019, **20**(Suppl 7):194
https://doi.org/10.1186/s12859-019-2740-6

BMC Bioinformatics

RESEARCH Open Access

Predicting stage-specific cancer related genes and their dynamic modules by integrating multiple datasets

Chaima Aouiche[1,2†], Bolin Chen[1,2*†] and Xuequn Shang[1,2]

From The 12th International Conference on Computational Systems Biology (ISB 2018)
Guiyang, China. 18-21 August 2018

Abstract

Background: The mechanism of many complex diseases has not been detected accurately in terms of their stage evolution. Previous studies mainly focus on the identification of associations between genes and individual diseases, but less is known about their associations with specific disease stages. Exploring biological modules through different disease stages could provide valuable knowledge to genomic and clinical research.

Results: In this study, we proposed a powerful and versatile framework to identify stage-specific cancer related genes and their dynamic modules by integrating multiple datasets. The discovered modules and their specific-signature genes were significantly enriched in many relevant known pathways. To further illustrate the dynamic evolution of these clinical-stages, a pathway network was built by taking individual pathways as vertices and the overlapping relationship between their annotated genes as edges.

Conclusions: The identified pathway network not only help us to understand the functional evolution of complex diseases, but also useful for clinical management to select the optimum treatment regimens and the appropriate drugs for patients.

Keywords: Disease genes, Clinical stages, Dynamic modules, Pathway networks, Disease evolution

Introduction

Complex diseases, such as cancers, are kinds of evolutionary diseases [1, 2], which involve successive stages from early initiation to advanced end-stages. Determining the possible biological changes associated with these stages is necessary for understanding the progression of many diseases, thereby specifying their best treatment strategy. Take the colorectal cancer for example, these stages can be classified generally into four phases based on their level of extension, lymphatic involvement and metastatic features. Specifically, stage I refers to a tumor of small size confined to the organ of origin; stage II describes the disease that has locally advanced beyond the site of origin; stage III characterizes the disease that has spread to the neighboring organs; and stage IV represents distant metastatic disease. Here, cancers at early stages (stage I or II) are usually considered curable and might only need an active surveillance compared to advanced stages (stage III or IV) which might require more radical and active treatment. Therefore, the understanding of the biological mechanism and molecular events of complex diseases through stages require the identification of stage-specific disease genes, unlike other irrelevant genes and genetic aberrations that turned out to have no functional relevance to any disease or specifically to cancer biology [3].

The availability of high dimensional datasets, and the great advancement of high-throughput technologies have enabled the identification of genes associated with

*Correspondence: blchen@nwpu.edu.cn
†Chaima Aouiche and Bolin Chen contributed equally to this work.
[1] School of Computer Science, Northwestern Polytechnical University, 710072 Xi'an, China
[2] Key Laboratory of Big Data Storage and Management, Northwestern Polytechnical University Ministry of Industry and Information Technology Xi'an, China

specific diseases, providing potential methods for precision medicine [4] and drug design [5]. Take the Cancer Genome Atlas (TCGA) project for example, it has generated multi-omics datasets over the genomic, epigenetic and transcriptome levels together with clinical data for more than 30 human tumors [6–9]. These multiple omics datasets provided many high-resolution molecular profiles, such as gene expression (microarray, RNA-seq), copy number variation (CNV or sCNA), DNA methylation, mRNA expression, somatic mutation, protein expression, as well as clinical information describing specific metrics, which including pathological stages, clinical stages, grade and age at diagnosis. They are highly variable in term of availability from disease to disease.

These datasets enabled integrative analysis focusing on the identification of cancer-related genes [10–14], unlike individual analysis with a single type of data, which represents an incomplete snapshot of a biological process and does not provide a comprehensive view of different disease states. In addition, clinical data also provided valuable insights into the genetic aberration detections, including cancer genes identification and their clinical translation [15–17].

Despite many discoveries made by these integrated genome datasets, there are only a limited number of studies that consider the associations between genomic profiles, clinical parameters, and their stage related cancer genes [18–24]. Moreover, these discoveries often neglected the fact that those identified cancer genes and functional modules are dynamically changed. Identifying the evolution of these biological modules is very important to understand the progression of many complex diseases, the key regulators of many cancer-related genes and their dysregulated pathways [25–31].

The main objective of this paper is to investigate a versatile working flow that can address the staging evolution processes of complex diseases, which including: (1) the identification of stage-specific cancer related genes, (2) the construction of their related dynamic modules, and (3) the generation of the stage related pathway networks. The rest of the paper is organized as follows. "Materials and methods" section introduces the methods and related materials. "Results and discussions" section addresses the numerical experiments and results. "Conclusion" section draws discussions and conclusions.

Materials and methods

Data sources and preprocessing

The Level 3 clinical information and genomic datasets were obtained from the FIREHOSE Broad GDAC [32]. It is one of the Genome Data Analysis Centers (GDACs) for the TCGA project that are used for prognosis and disease diagnosis. The datasets were downloaded in December 2016, which including the clinical information, gene expression and DNA methylation profiles for the same group of patients. The summarized information can be found in Table 1.

Table 1 The datasets informations for the same set of samples from Broad Firehose TCGA project

Data type	Platform	Samples
Gene expression	UNC-AgilentG4502A	219
DNA methylation	JHU-USC-HumanMethylation27	219
Clinical data	-	219

The clinical information for each patient are highly variable. Therefore, we focus on the "pathology_t_stage", which mainly describes the diagnosis stage of individual samples (t_1, t_2, t_3 and t_4). These pathological variables were converted into binary values for our following regularized regression analysis. For the sample selection, we only take those patients when the "pathology_t_stage" parameter was available. Finally, 219 samples were used to conduct our subsequent analysis.

The gene expression and DNA methylation profiles were measured for majority genes, which contain 17505 gene expressions and 26224 DNA methylations. However, we only consider the intersection of the two gene sets, which contain both gene expression and DNA methylation information in datasets. Moreover, genes with missing values, such as NA or NULL, were filtered out, and methylation CpG loci which related to multiple genes were shared equally for those genes. Eventually, a set of 12586 genes were obtained in this study.

Additionally, the HPRD PPI network (release 9) [33] were used to detect gene interactions and functional modules, which contained 9465 proteins and 37039 interactions. The workflow of the whole process is shown in Fig. 1.

Stage-specific related gene identification

The first important step to investigate the evolution progress of complex diseases is to identify signature genes for individual stages. Elastic net [34] is one of the classical feature selection algorithms, and has been widely used in biological and clinical research areas [35–42]. It employs a generalized linear regression model to handle the high-dimensional data regression issue without using any prior information. The elastic net method is based on a compromise between the Least Absolute Shrinkage and Selection Operator (LASSO) penalty (L1 norm) and the ridge penalty (L2 norm), where the LASSO penalty performs the feature selection and the coefficient estimation, while the ridge penalty shrinks those coefficients toward to zero [35].

Suppose there is a set of m samples (patients) and n features (gene expression or methylation profiles), the feature matrix can be denoted as a $m \times n$ dimension matrix X. Given a m dimension label vector Y (pathology stage

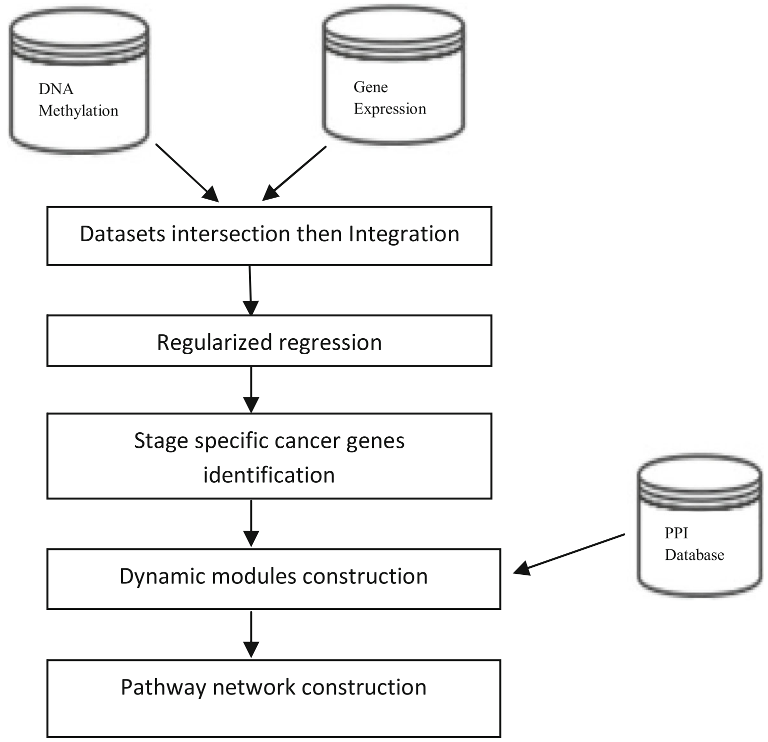
DNA
Methylation
Gene
Expression
Datasets intersection then Integration
Regularized regression
Stage specific cancer genes identification
PPI
Database
Dynamic modules construction
Pathway network construction

Fig. 1 The workflow for identifying stage-specific cancer related genes and their evolution processes through pathological_staging

labels), the problem stage-specific related gene identification is to detect a set of genes that minimize the following objective function

$$B = \|Y - X\beta\|_2 \tag{1}$$

where $\beta = (\beta_1, \beta_2, \ldots, \beta_n)^T$ is the coefficient vector for all features.

After adding a LASSO penalty and a ridge penalty, the elastic net method have a form like

$$\widehat{B} = \|Y - X\beta\|_2 + \lambda_1|\beta| + \lambda_2\|\beta\|_2, \tag{2}$$

or

$$\widehat{B} = argmin \left\{ \frac{1}{m}\sum_{i=1}^{m}\left(y_i - \sum_{j=1}^{n} x_{ij}\beta_j\right)^2 + \lambda_1\sum_{j=1}^{n}|\beta_j| + \lambda_2\sum_{j=1}^{n}{\beta_j}^2 \right\} \tag{3}$$

to be more specific, where λ_1, λ_2 are the penalty parameters related to LASSO and ridge penalty, respectively.

In this study, the gene expression profiles and the DNA methylation information were integrated to form the feature matrix X, and four binary stage-specific label vectors $Y_t, t = 1, 2, 3, 4$ were employed to identify disease related genes for individual stages, respectively (where an element in Y_t represents if that sample was recognized as the t_{th} pathology stage in the clinical dataset).

The objective function (3) was implemented in Matlab R2015a with the tuning parameter $\lambda_1 = \lambda_2 = 0.5$. The fitted least-squares regression coefficients were used for gene selection. Giving a pair of X and Y_t, the Matlab program calculated the fitted coefficients at around 50 times (automatically determined by Matlab). At each time, a set of signature genes could be selected if their coefficients were larger than a threshold. The finally stage-specific genes were determined based on the times of those genes were selected during the calculation. Table 2 summarizes the times of running, the number of selected genes across the 4 pathology stages, and the number of genes that were selected at least 20 times.

Table 2 The number of genes detected by cut off=20

Stages	Models #	Detected # of genes	# of genes at cut off=20	# of genes in the giant components
Pathology_t1	51	279	167	17
Pathology_t2	48	257	195	40
Pathology_t3	45	272	206	227
Pathology_t4	50	278	178	64

[20 implies the number of genes selected by at least 20 models]. This table illustrated the number of the models resulted at each pathology_stage, the maximum number of non-zero coefficients (genes) obtained at a specific model, the number of genes predicted by a cutoff metric and the number of genes in the giant components

Stage-specific module detection

Stage-specific modules at each pathology stage were constructed based on the giant component strategy and the human PPI network.

The selected signature genes at each stage were often isolated with each in many cellular networks, and the enrichment analyses of those genes may not get any meaningful result. To overcome this problem, we propose to use the giant component strategy to select the most functional related gene modules based on the identified genes.

To be more specific, a biological network could be employed as the basic background network, where the identified genes and their directed neighbors in the network were selected to form a subnetwork. The edges of the subnetwork were also generated based on the background network. By doing this, the obtained subnetwork often robustly linked with each other, and the initial identified genes served as seed nodes to generate the related functional modules. To further filtering out those disrelated genes, only the genes belong to the giant connected component of the subnetwork were selected as the signature genes for that stage. The rest of other genes will not consider in this study. The flowchart of this part was illustrated in Fig. 2.

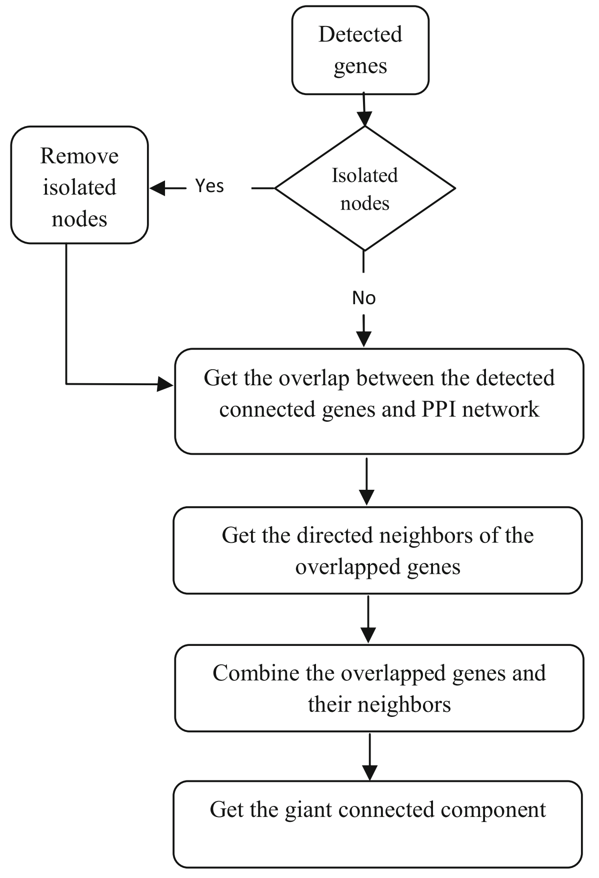
Detected genes
Isolated nodes
Yes
Remove isolated nodes
No
Get the overlap between the detected connected genes and PPI network
Get the directed neighbors of the overlapped genes
Combine the overlapped genes and their neighbors
Get the giant connected component

Fig. 2 The flow chart of the stage-specific gene identification

The HPRD PPI dataset deemed the most important interaction in this study compared to other human datasets like the human cofunction network in [43] and the InWeb_IM PPI network in [44], since it linked efficiently the genes identified at each stage.

Dynamic module analysis and pathway network generation

Once the signature genes were identified for each stage, the Cytoscape was used to draw the explicit graphical representations for different biological modules and subnetworks. In this study, 4 groups of dynamic giant connected functional modules were constructed, where the vertices represented the list of interested genes at each stage, and edges represented the functional relations between them obtained from the PPI network.

To further determine whether the list of signature genes identified at individual pathology stages are statistically enriched in certain biological processes or functions, functional enrichment analysis were performed using the DAVID tool [45]. A list of significant Reactome pathways has been obtained from these enrichment analysis. We then pooled these Reactome pathways altogether and get their official annotated pathway descriptions from the database. Next, a pathway evolution network was generated by pooling all those stage-specific pathways together, where the vertices in this network represent individual pathways, and connections were obtained if the two pathways have overlapped genes.

The pathway network could clearly show the dynamic evolution processes of the interested disease, since we can use the color of individual vertices to indicate their pathology stage, and the width of edges can show the overlapped score between two pathways. Here, the overlap score was calculated as follows:

$$W = \frac{k^2}{p * q}. \tag{4}$$

where k is the number of the overlapped genes between a pair of pathway P_i and pathway P_j, p and q are the total numbers of genes in P_i and P_j, respectively.

Results and discussions

The number of stage-specific related genes

In this study, we have selected those genes that were detected by at least 20 models as the seed of stage specific related genes. By using this strategy, a list of signature genes that robustly delineate early and advanced pathological stages. Table 2 summarized the number of genes selected at different stages. To be more specific, stage t1 has obtained 167 genes from 51 models; stage t2 has obtained 195 genes from 48 models; stage t3 has obtained

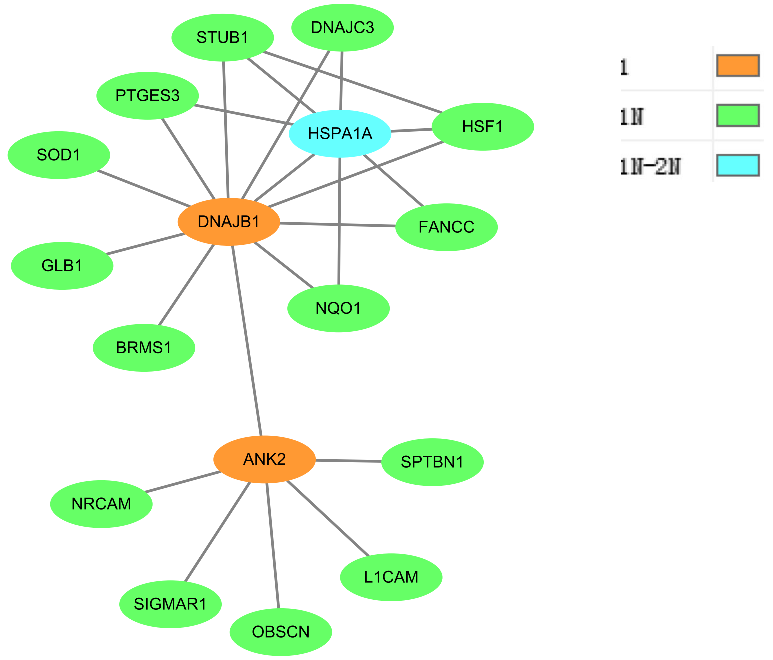

Fig. 3 Pathology_t1 stage module. This module has 17 giant component nodes (genes) interacted with 23 edges. Node colors specify: stage1 identified genes, their neighbors and also the overlapped genes from other pathology stages, where 1 indicates stage1 detected genes, 1N indicates stage1 directed neighbors and 1N-2N indicates the overlapping genes between stage1 neighbor genes and stage2 neighbor genes as shown in the code colors

206 genes from 45 models; and stage t4 has obtained 178 genes from 50 models, respectively.

All of these genes were considered as indicators or signatures to characterize the dynamics of the 4 pathological stages, due to their possible role in cancer progression.

Dynamic modules construction and visualization

The HPRD network was used to construct 4 groups of pathology stage related modules based on their identified giant components. Interactions among their identified genes were extracted to form the corresponding modules,

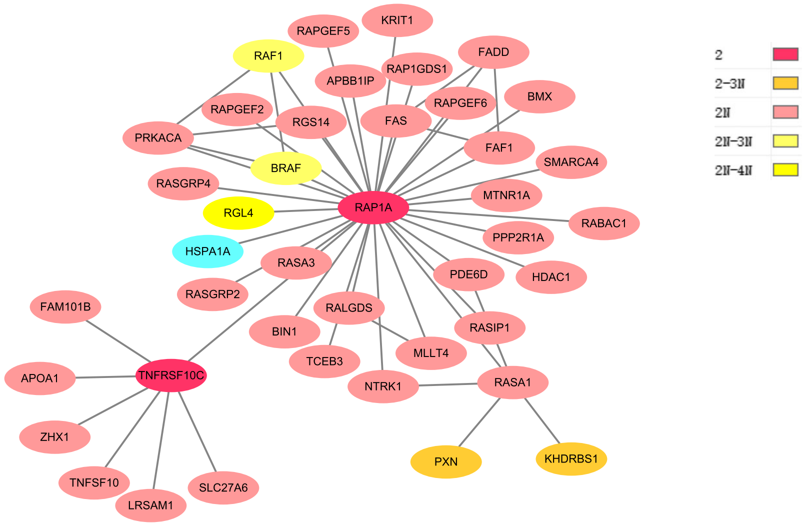

Fig. 4 Pathology_t2 stage module. This module has 42 giant component nodes interacted with 51 edges. Node colors specify: stage2 identified genes, their neighbor genes and also the overlapped genes from other pathology stages, where 2 indicates stage2 detected genes, 2N indicates stage2 directed neighbors, 2-3N indicates the overlapping genes between stage2 detected genes and stage3 neighbor genes, 2N-3N indicates overlap genes between stage2 neighbor genes and stage3 neighbor genes and 2N-4N denotes overlap genes between stage2 neighbor genes and stage4 neighbor genes which shown clearly in the code colors

which contained 17 nodes and 23 interactions for stage t1; 42 nodes and 51 interactions for stage t2; 228 nodes and 1004 interactions for stage t3; and 65 nodes and 87 interactions for stage t4.

In order to further know how the four pathology stages involved and interacted to each other, the overlapping cancer genes between them were identified from the combined set, and the connections of these genes along with their neighbors at individual stage compared to other stages were shown in Figs. 3, 4, 5 and 6, respectively. These figures show originally detected genes, neighbor genes and their overlapped genes of individual pathology stages, which are highlighted by different colors.

Annotated functions and pathway enrichment analysis

Pathway analysis has become the first choice for gaining insight into the underlying biology of differentially expressed and methylated genes, as it reduces complexity and has increased explanatory power. Thus, to validate our results, a DAVID functional annotation tool with a Reactome pathway annotation [46] was carried out for the identified pathological significant genes to identify their potential pathways and thereafter construct the corresponding pathway network.

In this study, a considerable number of stage-specific genes have been successfully identified across 4 pathological stages. These identified genes robustly associated with each other to produce meaningful biological modules, and significantly enriched in some key biological pathways. Those pathways, in turn, also interacted at a higher level based on the overlapping between their annotated genes. The rich set of interactions between these pathways results in a valuable pathway network, capturing highly evolved pathways through the 4 pathological stages.

The network provides novel insights into the cancer disease evolution. As can be seen in Fig. 7, the evolution histories or communities could be clearly classified into 6 groups: (a), (b), (c), (d), (e) and (f). Specifically, in group (a), the cancer evolved from stage 1 (red nodes) to stage 2 (light blue nodes), and continued to evolve through stage 3 (dark blue nodes) to the end of stage 4 (green nodes). Similar to group (b) which involved an evolution start from

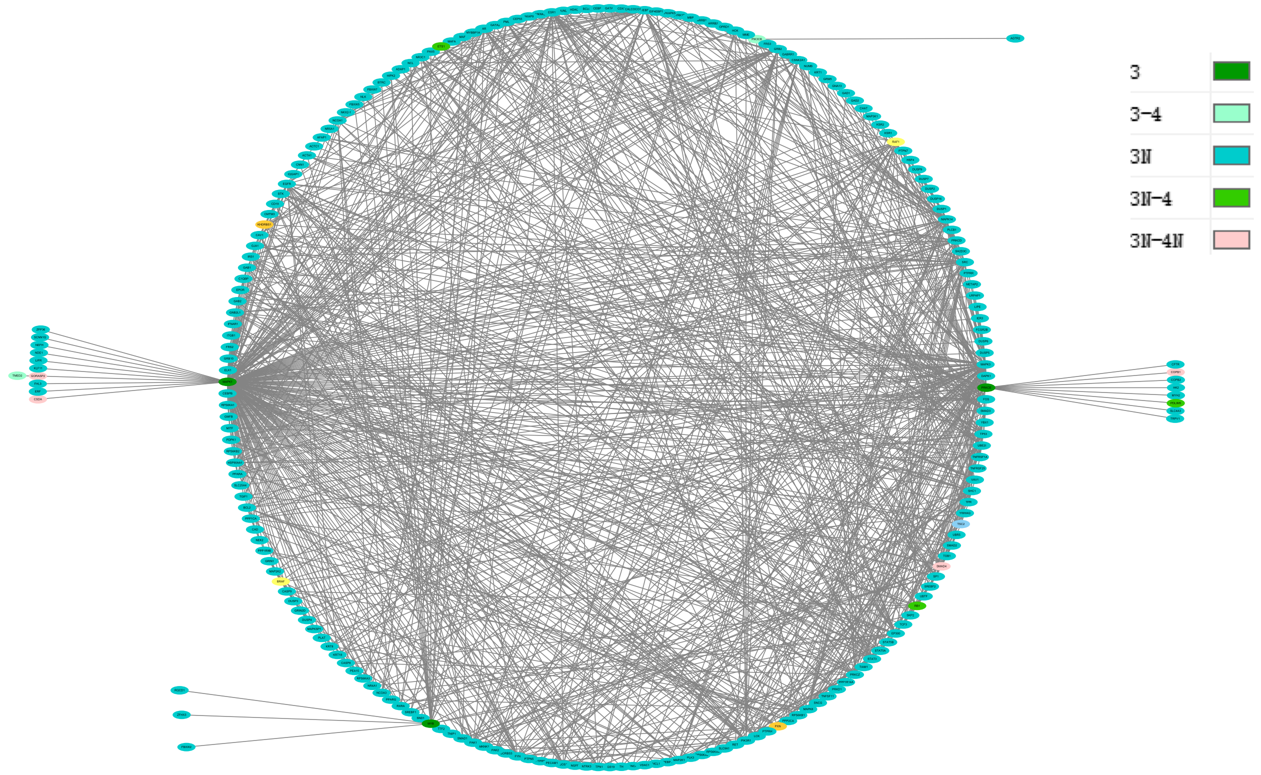

Fig. 5 Pathology_t3 stage module. This module has 228 giant component nodes interacted with 1004 edges. Node colors specify: stage3 identified genes, their neighbor genes and also the overlapped genes from other pathology stages, where 3 indicates stage3 detected genes, 3N indicates stage3 directed neighbors, 3-4 indicates the overlapping genes between stage3 detected genes and stage4 detected genes, 3N-4 indicates overlap genes between stage3 neighbor genes and stage4 detected genes and 3N-4N denotes overlap genes between stage3 neighbor genes and stage4 neighbor genes which shown clearly in the code colors

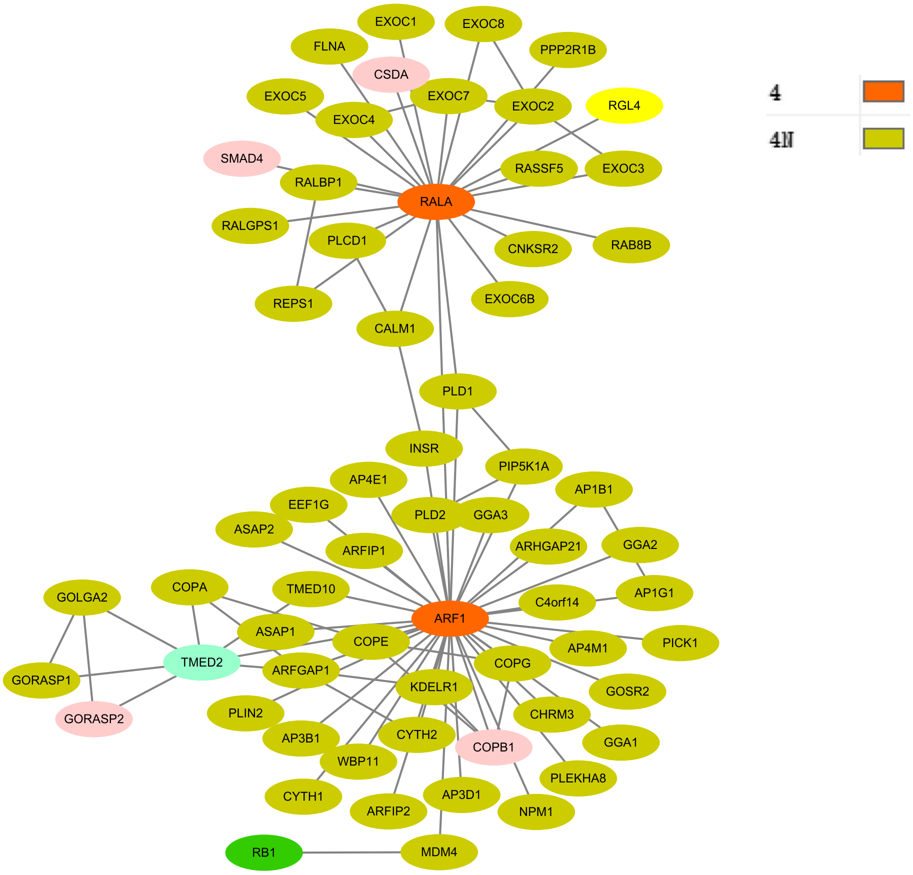

Fig. 6 Pathology_t4 stage module. This module has 65 giant component nodes interacted with 87 edges. Node colors specify: stage4 identified genes, their neighbor genes and also the overlapped genes from other pathology stages, where 4 indicates stage4 detected genes, 4N indicates stage4 directed neighbors. In addition to the overlapped genes from stage 2 and 3 as shown in previous code colors

stage 2 (blue node) to stage 2 and 3 by their common pathways (light blue and dark blue). Then, from stage 3 (dark blue) to stage 4 (which shown in green nodes). For group (c), an evolution also happened from stage 2 to stage 3 pathways. In group (d) the evolution happened through stage 1, stage 2 to stage 3, and in group (e) the evolution involved in stage 2 and stage 3. The final group (f), which includes a large set of stage 3 pathways strongly related to each other, suggesting the metastasis growth of cancer disease.

Moreover, to illuminate the biological significance role of the extracted pathway network and their evolved pathways, we also defined their annotated functions, which are shown in Fig. 8. For stage 1, the annotated functions mainly belong to cellular responses to external stimuli group. For stage 2, the annotated functions carry (1) immune system, (2) signal transduction tumor and (3) programmed cell death. For stage 3, the annotated functions evolve to (1) neuronal system, (2) immune system, (3) signal transduction and (4) developmental biology. The last stage includes annotated functions such as (1) metabolism of proteins, (2) vesicle mediated transport, (3) disease and (4) cell cycle.

To be more specific, for example, in the group (a), the evolved pathways were highly related to functions which starts from (1) cellular responses to (2) external stimuli, then go through (3) neuronal system signal transduction, (4) developmental biology, and ends up with (5) metabolism of proteins. In group (b), the evolution starts from (1) hemostasis to (2) signal transduction and ended with (3) cell cycle. Whereas in group (c), a set of communities enriched in common functions including (1) cancer, (2) disease and cancer, (3) disease and tumor, (4) signal transduction and (5) immune system. For group (f), the most pathways successfully enriched in programmed cell death function. However, in the group (d), different functions have been defined including (1) circadian clock and (2) gene expression (transcription). Overall, the description details of these

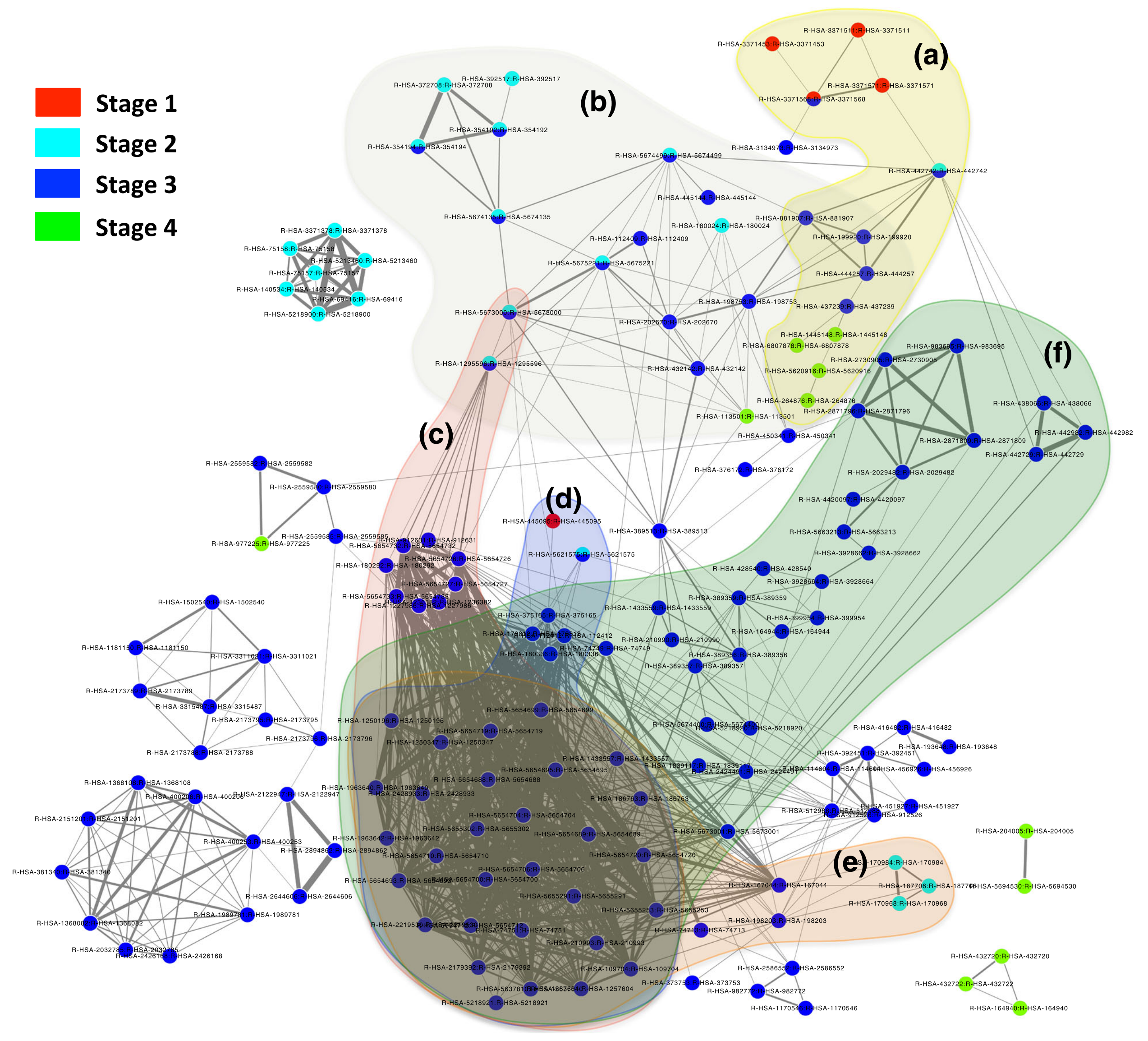
Stage 1
Stage 2
Stage 3
Stage 4
(a)
(b)
(c)
(d)
(e)
(f)
R-HSA-3371511:R-HSA-3371511
R-HSA-3371453:R-HSA-3371453
R-HSA-3371571:R-HSA-3371571
R-HSA-3371568:R-HSA-3371568
R-HSA-3134973:R-HSA-3134973
R-HSA-442742:R-HSA-442742
R-HSA-392517:R-HSA-392517
R-HSA-372708:R-HSA-372708
R-HSA-354192:R-HSA-354192
R-HSA-354194:R-HSA-354194
R-HSA-5674499:R-HSA-5674499
R-HSA-445144:R-HSA-445144
R-HSA-5674135:R-HSA-5674135
R-HSA-881907:R-HSA-881907
R-HSA-180024:R-HSA-180024
R-HSA-112409:R-HSA-112409
R-HSA-199920:R-HSA-199920
R-HSA-5675221:R-HSA-5675221
R-HSA-444257:R-HSA-444257
R-HSA-198753:R-HSA-198753
R-HSA-437239:R-HSA-437239
R-HSA-5673000:R-HSA-5673000
R-HSA-202670:R-HSA-202670
R-HSA-1445148:R-HSA-1445148
R-HSA-6807878:R-HSA-6807878
R-HSA-983695:R-HSA-983695
R-HSA-2730905:R-HSA-2730905
R-HSA-1295596:R-HSA-1295596
R-HSA-432142:R-HSA-432142
R-HSA-5620916:R-HSA-5620916
R-HSA-438066:R-HSA-438066
R-HSA-264876:R-HSA-264876
R-HSA-2871796:R-HSA-2871796
R-HSA-113501:R-HSA-113501
R-HSA-450341:R-HSA-450341
R-HSA-442982:R-HSA-442982
R-HSA-2871809:R-HSA-2871809
R-HSA-442729:R-HSA-442729
R-HSA-376172:R-HSA-376172
R-HSA-2029482:R-HSA-2029482
R-HSA-4420097:R-HSA-4420097
R-HSA-3371378:R-HSA-3371378
R-HSA-75158:R-HSA-75158
R-HSA-5213460:R-HSA-5213460
R-HSA-75157:R-HSA-75157
R-HSA-140534:R-HSA-140534
R-HSA-69416:R-HSA-69416
R-HSA-5218900:R-HSA-5218900
R-HSA-2559582:R-HSA-2559582
R-HSA-2559580:R-HSA-2559580
R-HSA-2559585:R-HSA-2559585
R-HSA-977225:R-HSA-977225
R-HSA-445095:R-HSA-445095
R-HSA-5621575:R-HSA-5621575
R-HSA-389513:R-HSA-389513
R-HSA-5663213:R-HSA-5663213
R-HSA-3928662:R-HSA-3928662
R-HSA-428540:R-HSA-428540
R-HSA-3928664:R-HSA-3928664
R-HSA-389359:R-HSA-389359
R-HSA-1433559:R-HSA-1433559
R-HSA-399954:R-HSA-399954
R-HSA-375165:R-HSA-375165
R-HSA-164944:R-HSA-164944
R-HSA-210990:R-HSA-210990
R-HSA-112412
R-HSA-74749
R-HSA-389356:R-HSA-389356
R-HSA-389357:R-HSA-389357
R-HSA-1502540:R-HSA-1502540
R-HSA-1181150:R-HSA-1181150
R-HSA-3311021:R-HSA-3311021
R-HSA-2173789:R-HSA-2173789
R-HSA-3315487:R-HSA-3315487
R-HSA-2173795:R-HSA-2173795
R-HSA-2173796:R-HSA-2173796
R-HSA-2173788:R-HSA-2173788
R-HSA-1250196:R-HSA-1250196
R-HSA-5654699:R-HSA-5654699
R-HSA-5654719:R-HSA-5654719
R-HSA-1250347:R-HSA-1250347
R-HSA-5218920
R-HSA-416482:R-HSA-416482
R-HSA-193648:R-HSA-193648
R-HSA-392451
R-HSA-456926:R-HSA-456926
R-HSA-1368108:R-HSA-1368108
R-HSA-400206
R-HSA-2122947:R-HSA-2122947
R-HSA-2151201:R-HSA-2151201
R-HSA-5654695
R-HSA-5654688:R-HSA-5654688
R-HSA-186763:R-HSA-186763
R-HSA-451927:R-HSA-451927
R-HSA-912526
R-HSA-5654704
R-HSA-5655302:R-HSA-5655302
R-HSA-5654689
R-HSA-5673001:R-HSA-5673001
R-HSA-204005:R-HSA-204005
R-HSA-400253
R-HSA-2894862
R-HSA-5654710:R-HSA-5654710
R-HSA-5654720:R-HSA-5654720
R-HSA-381340
R-HSA-170984:R-HSA-170984
R-HSA-5654706:R-HSA-5654706
R-HSA-167044:R-HSA-167044
R-HSA-187706
R-HSA-5694530:R-HSA-5694530
R-HSA-5654700
R-HSA-2644606
R-HSA-5655291:R-HSA-5655291
R-HSA-170968:R-HSA-170968
R-HSA-1989781
R-HSA-5655253
R-HSA-1368082
R-HSA-198203:R-HSA-198203
R-HSA-74751
R-HSA-74713
R-HSA-210993
R-HSA-2426168
R-HSA-432720:R-HSA-432720
R-HSA-2586552
R-HSA-2179392:R-HSA-2179392
R-HSA-109704:R-HSA-109704
R-HSA-373753
R-HSA-432722:R-HSA-432722
R-HSA-982772
R-HSA-1257604
R-HSA-5218921:R-HSA-5218921
R-HSA-1170546
R-HSA-164940:R-HSA-164940

Fig. 7 The dynamic structure of the pathway interaction network. The nodes in this network referred to pathways and edges between nodes indicated interaction of pathways determined through overlapped genes. In particular, nodes in red indicate stage1 pathways; nodes in blue indicate stage2 pathways; nodes in dark blue indicate stage3 pathways ;nodes in green indicate stage4 pathways, whereas nodes in 2 colors refer to the common pathways of 2 stages. The color of nodes represent the cancer stages

functions suggesting the important biological role of this study.

Conclusion

In this paper, we introduced a working flow which mainly adressed 3 important biological aspects such as stage-specific cancer genes identification, multi-omics data integration, biological modules and cellular pathway network construction.

The main objective of the study was to gain biological and clinical insights into the progression of cancer diseases through pathological staging mechanism. Since complex diseases, include but not limited to cancers, are evolutionary diseases that don't directly end up with a mortal situation. They evolve cross multiple stages that can be determined not only through the lens of biological modules but more importantly through pathway networks, which contain rich biological information and provide more detailed molecular mechanisms.

Therefore, we constructed individual pathological modules based on the overlap between identified giant specific genes associated through a PPI network. We also performed pathway analysis on these genes and built a valuable pathway network based on the enriched pathways

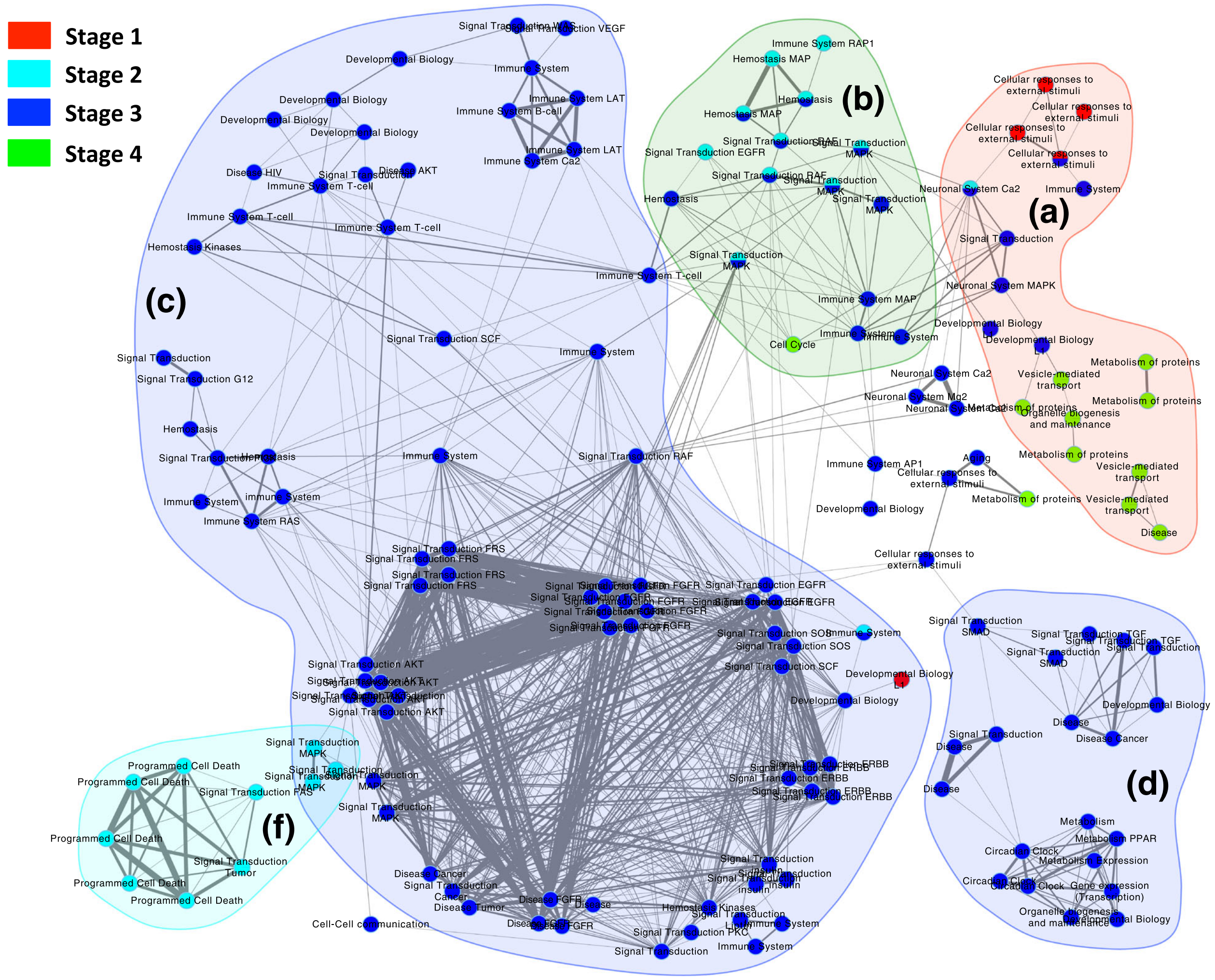

Fig. 8 The annotated functions of the extracted pathway network. This figure illustrated the different pathways functions enriched at every pathology stage and the highlighted (**a**), (**b**), (**c**), (**d**) and (**f**) cases determine the main histories functions detected in this pathway network. The color of nodes represent the cancer stages

and the overlap between their annotated genes, which captured highly evolved pathways that involved different successive stages determining the real evolution of cancer diseases.

This process has furthered this understanding by identifying significant differences between different diseases stages and determining their evolution through pathways perspective, which have important implications not only for the classification of diseases/phenotypes, but also with clinical management by helping to select the most appropriate treatment modality for patients, holding promise for finding potential drugs.

Our understanding of cancer biology through the lens of the pathway and network analyses is promising. Especially when a disease reaches a metastasis status, which is the pivotal cause of patient deaths. The metastatic status is an advanced status that can be deeply defined by the TNM (Primary tumor (T), Lymph nodes (N) and Distant metastasis (M)) criteria, which are major parameters in the staging technique. Thus, we see ample opportunities to address this issue in future work. Furthermore, integrating more datasets at various levels (e.g gene expression, DNA methylation, and somatic CNV) might further facilitate the discovery of more robust staging modules and pathways, that easily determine the evolutionary process of many diseases revealing more comprehensive information of disease states. However, substantial additional experiments will be required to validate the predicted findings.

Abbreviations

CNV: Copy number variation; DAVID: The database for annotation, visualization and integrated discovery; GDACs: Genome data analysis centers; HPRD: Human protein reference database; InWeb_IM: InWeb_InBioMap; LASSO: Least absolute shrinkage and selection operator; PPI: Protein-protein interaction;

sCNA: Somatic copy number alteration; TCGA: The cancer genome atlas; TNM: Primary tumor, lymph node and distant metastasis

Acknowledgements

Not applicable.

Funding

Publication costs are funded by the National Natural Science Foundation of China under [Grant No. 61602386, 61772426 and 61332014]; the Natural Science Foundation of Shaanxi Province under [Grant No. 2017JQ6008]; and the Top International University Visiting Program for Outstanding Young scholars of Northwestern Polytechnical University.

Availability of data and materials

The data and models analyzed in the current study are available in this article and databases.

About this supplement

This article has been published as part of *BMC Bioinformatics Volume 20 Supplement 7, 2019: Selected papers from the 12th International Conference on Computational Systems Biology (ISB 2018)*. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-20-supplement-7.

Authors' contributions

BC initialized this study. CA and BC discussed many times to finalized the work plan. CA conducted the majority of numerical experiments. XS gave suggestions many time to modify this study. CA drafted the manuscript. Everyone read the manuscript and revised it, and agreed with the final version.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Published: 1 May 2019

References

1. Horne S, Chowdhury S, Heng H. Stress, genomic adaptation, and the evolutionary trade-off. Front Genet. 2014;5:92.
2. Horne S, Pollick S, Heng H. Evolutionary mechanism unifies the hallmarks of cancer. Int J Cancer. 2015;136:2012–21.
3. Vogelstein B, Kinzler K. Cancer genes and the pathways they control. Nat Med. 2004;10:789–99.
4. Jorgensen J. A challenging drug development process in the era of personalized medicine. Drug Discov Today. 2011;16:891–7.
5. Wang X, Gulbahce N, Yu H. Network-based methods for human disease gene prediction. Brief Funct Genomics. 2011;10:280–93.
6. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008;455:1061–1068.
7. Shen R, Olshen A, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009;25:2906–12.
8. Vucic E, Thu K, Robison K, Rybaczyk L, Chari R, Alvarez C, Lam W. Translating cancer "omics" to improved outcomes. Genome Res. 2012;22:188–95.
9. Chin L, Hahn W, Getz G, Meyerson M. Making sense of cancer genomic data. Genes Dev. 2011;25:534–55.
10. Greenawalt D, Sieberts S, Cornelis M, Girman C, Zhong H, et al. Integrating genetic association, genetics of gene expression, and single nucleotide polymorphism set analysis to identify susceptibility loci for type 2 diabetes mellitus. Am J Epidemiol. 2012;176:423–30.
11. Li Q, Seo J, Stranger B, McKenna A, Pe'er I, et al. Integrative eqtl-based analyses reveal the biology of breast cancer risk loci. Cell. 2013;152:633–41.
12. Serizawa R, Ralfkiaer U, Steven K, Lam G, Schmiedel S, et al. Integrated genetic and epigenetic analysis of bladder cancer reveals an additive diagnostic value of fgfr3 mutations and hypermethylation events. Int J Cancer. 2011;129:78–87.
13. Lee J, Zhao X, Yoon I, Lee J, Kwon N, Wang Y, et al. Integrative analysis of mutational and transcriptional profiles reveals driver mutations of metastatic breast cancers. Cell Discov. 2016;2:16025.
14. Guanghui Z, Hui Y, Xiao C, Jun W, Yong Z, Xing-Ming Z. Cstea: a webserver for the cell state transition expression atlas. Nucleic Acids Res. 2017;45:103–8.
15. van Vliet M, Horlings H, van de Vijver M, Reinders M, Wessels L. Integration of clinical and gene expression data has a synergetic effect on predicting breast cancer outcome. PLoS ONE. 2012;7:40358.
16. Xiong Q, Ancona N, Hauser E, Mukherjee S, Furey T. Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets. Genome Res. 2012;22:386–97.
17. Seoane J, Day I, Gaunt T, Campbell C. A pathway-based data integration framework for prediction of disease progression. Bioinformatics. 2013;30:838–45.
18. Hsu F, Serpedin E, Hsiao T, Bishop A, Dougherty E, Chen Y. Reducing confounding and suppression effects in tcga data: an integrated analysis of chemotherapy response in ovarian cancer. BMC Genomics. 2012;13:13.
19. Parker J, Mullins M, Cheang M, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27:1160–7.
20. Curtis C, Shah S, Chin S, Turashvili G, Rueda O, Dunning M, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486:346–52.
21. Kittaneh M, Montero A, Gluck S. Molecular profiling for breast cancer: a comprehensive review. Biomark Cancer. 2013;5:61–70.
22. Li A, Walling J, Ahn S, Kotliarov Y, Su Q, Quezado M, et al. Unsupervised analysis of transcriptomic profiles reveals six glioma subtypes. Cancer Res. 2009;69:2091–9.
23. Shen L, Toyota M, Kondo Y, Lin E, Zhang L, Guo Y, et al. Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer. Proc Natl Acad Sci U S A. 2007;104:18654–9.
24. van't Veer L, Dai H, van de Vijver M, He Y, Hart A, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415:530–6.
25. Akavia U, et al. An integrated approach to uncover drivers of cancer. Cell. 2010;143:1005–17.
26. Danussi C, et al. Rhpn2 drives mesenchymal transformation in malignant glioma by triggering rhoa activation. Cancer Res. 2013;73:5140–50.
27. Sonabend A, et al. The transcriptional regulatory network of proneural glioma determines the genetic alterations selected during tumor progression. Cancer Res. 2014;74:1440–51.
28. Carro M, et al. The transcriptional network for mesenchymal transformation of brain tumours. Nature. 2010;463:318–25.
29. Hoadley K, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell. 2014;158:929–44.
30. Liu K, Liu Z, Hao J, et al. Identifying the dysregulated pathways in cancers from pathway interaction networks. BMC Bioinformatics. 2012;13:126.
31. Wu G, Feng X, Stein L. A human functional protein interaction network and its application to cancer data analysis. Genome Biol. 2010;11:53.
32. FIREHOSE Broad GDAC. http://gdac.broadinstitute.org/. Accessed Sept 2014.
33. Human Protein Reference Database. http://www.hprd.org/. Accessed Sept 2014.
34. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970;12:55–67.
35. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer; 2009.
36. Cho S, Kim H, Oh S, Kim K, Park T. Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis. BMC Proc. 2009;3:25.
37. Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau K, Greninger P, Thompson IR, Luo X, Soares J, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature. 2012;483:570–5.

38. Horvath S. Dna methylation age of human tissues and cell types. Genome Biol. 2013;14:3156.
39. Zemmour C, Bertucci F, Finetti P, Chetrit B, Birnbaum D, Filleron T, Boher J. Prediction of early breast cancer metastasis from dna microarray data using high-dimensional cox regression models. Cancer Informat. 2015;14:129–38.
40. Whelan R, Watts R, Orr CA, Althoff RR, Artiges E, Banaschewski T, Barker GJ, Bokde ALW, Büchel C, Carvalho FM, et al. Neuropsychosocial profiles of current and future adolescent alcohol misusers. Nature. 2014;512:185–9.
41. Lee H, Flaherty P, Ji H. Systematic genomic identification of colorectal cancer genes delineating advanced from early clinical stage and metastasis. BMC Med Genomics. 2013;6:54.
42. Lee H, Palm J, Grimes S, Ji H. The cancer genome atlas clinical explorer: a web and mobile interface for identifying clinical–genomic driver associations. Genome Med. 2015;7:112.
43. Lee I, Blom U, Wang P, Shin J, Marcotte E. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res. 2011;21:1109–21.
44. Li T, Wernerse R, Hansen R, Horn H, Mercer J, et al. A scored human protein-protein interaction network to catalyze genomic interpretation. Nat Methods. 2017;14:61–4.
45. DAVID Bioinformatics Resources 6.8. http://david.abcc.ncifcrf.gov/. Accessed Oct 2016.
46. Reactome. http://reactome.org/. Accessed Jan 2017.