Computational and Structural Biotechnology Journal 20 (2022) 838–849

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Identifying individualized risk subpathways reveals pan-cancer molecular classification based on multi-omics data

Yanjun Xu [a,1], Jingwen Wang [a,1], Feng Li [a,1], Chunlong Zhang [a,1], Xuan Zheng [a], Yang Cao [b], Desi Shang [a], Congxue Hu [a], Yingqi Xu [a], Wanqi Mi [a], Xia Li [a,*], Yan Cao [b,*], Yunpeng Zhang [a,*]

[a] *College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China*
[b] *Harbin Medical University Cancer Hospital, Harbin, China*

ARTICLE INFO

Article history:
Received 30 August 2021
Received in revised form 18 January 2022
Accepted 18 January 2022
Available online 22 January 2022

Keywords:
Individualized risk subpathway
Pathway analysis
Multi-omics
Pan-cancer

ABSTRACT

Cancer is a highly heterogeneous disease with different functional disorders among individuals. The initiation and progression of cancer is usually related to dysregulation of local regions within pathways. Identification of individualized risk pathways is crucial for revealing the mechanisms of tumorigenesis and heterogeneity. However, approach that focused on mining patient-specific risk subpathway regions is still lacking. Here, we developed an individualized cancer risk subpathway identification method that was referred as InCRiS by integrating multi-omics data. Then, the method was applied to nearly 3000 samples across 9 TCGA cancer types and its robustness and reliability were evaluated. Dissecting dysregulated subpathways in these tumor samples revealed several key regions which may play oncogenic roles in cancer. The construction of risk subpathway dysregulation profile of pan-cancers revealed 11 pan-cancer molecular classification (InCRiS subtypes) with significantly different clinical outcomes. Moreover, subpathway regions that tend to be disordered in individuals of specific subtypes were examined for understanding the pathogenesis of tumor and some key genes such as CTNNB1, EP300 and PRKDC were nominated in different subtypes. In summary, the proposed method and resulting data presented useful resources to study the mechanism of tumor heterogeneity and to discovery novel therapeutic targets for precise treatment of cancer.

1. Introduction

The initiation and development of malignant tumor is a complex process involving dysregulation of multiple molecules and their interactions [1]. Thus, it is a great challenge to elucidate the pathogenesis of cancer. The emergence of large-scale cancer multidimensional omics data provides new opportunities for cancer related researches. Identification of risk pathways (pathways that significantly disordered in tumor individuals) is a useful strategy for interpreting these data and understanding the mechanism of tumorigenesis.

Currently, a series of methods and tools have been developed for identifying disease risk pathways. These methods can be classified into three categories including over-representation approach (ORA) [2], functional class scoring (FCS) [3,4] and pathway topology-based methods. ORA methods mainly evaluate whether the risk gene set (e.g. differentially expressed genes) of one disease significantly over-represented in a biological pathway by using statistical tests. For example, DAVID [5] , KOBAS [6], Enrichr [7] and g: Profiler [8,9] are popular used platforms that provide pathway enrichment analysis based on hypergeometric test. The gene set enrichment analysis (GSEA) method [10] is a representation of FCS methods, in which the degree of differential expression of genes within pathway was considered based on their ranks on the background list. In biological pathways, different genes are connected to each other, forming a complex topological network structure and different genes contribute differently to the function of pathway. For example, the P53 gene plays a central role in the P53 signaling pathway. Thus, a series of pathway topology-based methods such as SPIA [11] and TPEA [12] which consider both upstream and downstream regulatory associations between genes

[*] Corresponding authors at: College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China (Y. Zhang); Department of Urology, Harbin Medical University Cancer Hospital, Harbin, China (Y. Cao); College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China (X. Li).

E-mail addresses: lixia@hrbmu.edu.cn (X. Li), caoyan1981821@126.com (Y. Cao), zhangyp@hrbmu.edu.cn (Y. Zhang).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.csbj.2022.01.022
2001-0370/© 2022 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.

and the importance of topological positions within pathways were proposed. The occurrence and development of diseases are often caused by abnormalities of several different local regions (subpathways) within biological pathways, rather than the overall scale of pathways. Thus, identification of risk subpathway regions could capture functional dysregulation of the disease at a more precise level. Researchers have developed a series of data-driven methods and tools for subpathway identification, such as Clipper [13], Signet [14], HotNet [15] and HotNet2 [16]. However, most of these methods focused on a single omics data, transcriptome or genome. Furthermore, these methods identify disease risk pathways at the population level and ignore the individual differences.

Cancers were highly heterogeneous diseases which had great differences in drug sensitivity, survival prognosis and other aspects among individuals. Currently, many studies aimed to classify cancer subtypes for revealing the tumor heterogeneity and promoting the precision treatment. For example, Arora et al. proposed 'survClust' method and identified prognostic subtypes for 18 cancer types across multiple data [17]. Hoadley et al. performed integrative analysis using multiple omics data and reclassified human tumors into unified subtypes of pan-cancer [18,19]. There are individualized differences in the dysregulation functions among different patients. Therefore, identification of risk pathways at individual level may be of great significance for researches on the mechanism of tumor heterogeneity, and ultimately promote the realization of precision medicine. Several methods have been developed. Vaske et al. developed the "PARADIGM" method [20], which inferred cancer individual-level pathway activity based on multi-dimensional cancer omics data. The "Pathifier" method inferred pathway activity of individual samples based on gene expression [21]. "Individpath" method identified dysfunctional pathways at individualized level based on the disrupted coordination of gene expression [22]. These individualized approaches are helpful for understanding the mechanism of cancer heterogeneity, but they did not consider pathway topologies. Furthermore, these methods only focused on the whole pathway, rather than their local regions which provided more insights into the mechanism of tumorigenesis.

Here, we propose a data-driven method for identifying dysregulated subpathway regions in specific patient by integrating multi-omics data. The feasibility and reliability of method were evaluated based on cancer hallmark/oncogenic pathway data. We also identified some novel key subpathway regions in pan-cancer. The dysregulated subpathway profile was constructed and 11 subtypes with significant differences in survival outcomes were identified in pan-cancer. Dissecting subpathways that tend to be disordered in individuals of specific subtypes suggested that our method may server as a useful resource for investigating the pathogenesis of cancer.

2. Materials and methods

2.1. Data collection

2.1.1. Multi-omics data of cancer

We obtained multi-omics data for tumor individuals from the TCGA database [23] (Fig. 1A). In this study, cancer types which contained more than 10 normal samples at the expression level were selected for further analysis. Finally, multi-omics (somatic mutation, copy number data and expression profile) datasets of 2979 tumor individuals and 393 normal samples from nine cancer types (BLCA, BRCA, KIRP, LIHC, LUSC, READ, STAD, THCA and UCEC) were obtained (Supplementary Table S1). In addition, the clinical survival data (if available) for these individuals were also obtained.

Somatic mutation The mutation data were downloaded in the form of mutation annotation file (MAF). We removed the silent mutations, and then organized mutation annotation files into mutation or non-mutation matrix for cancer types of TCGA. Hypermutated samples with genes mutated more than 1000 were also removed.

Copy number data Copy number datasets of these nine cancer types from TCGA were obtained. We ran GISTIC [24] to identify genes with copy number variation, and only large segment of deletions or amplifications were taken into consideration. Additionally, both deletions and amplifications of genes were regarded as equal alterations in our method.

Expression profile In this study, we obtained expression data (level 3) of cancer types that have more than 10 normal samples. Furthermore, genes with missing expression value greater than 20% were removed.

Furthermore, to validate the robustness and reliability of method, multi-omics data of LIHC and THCA from PCAWG (https://dcc.icgc.org/pcawg) [25] were also obtained. For somatic mutation data, hypermutated samples with genes mutated more than 1000 were removed. The GISTIC result file of copy number data were downloaded and only large segment of deletions or amplifications were regarded as variation. Genes with missing expression value greater than 20% were also removed. Finally, 8 normal samples and 50 tumor samples of LIHC and 4 normal samples and 25 tumor samples for THCA in PCAWG were used to evaluate the method.

2.1.2. Biological pathway and protein–protein interaction network

Biological pathway data was obtained from our previously download KGML files which contained gene-gene interactions of KEGG database [26]. Then, the iSubpathwayMiner R package [27] was used to reconstruct these pathways into undirected gene-gene interaction graphs. Finally, 281 KEGG pathways in the form of igraph (https://igraph.org/r/) which contained gene-gene interactions were obtained.

We constructed a global protein–protein interaction (PPI) network based on pathway datasets from different resources. Firstly, the human gene-gene interactions within pathways from databases including KEGG, Biocarta (www.biocarta.com), Reactome [28], NCI/Nature Pathway Interaction Database [29], HumanCyc [30] and Panther [31] were extracted by using the graphite R package [32]. Protein-protein interactions of Human Protein Reference Database (HPRD) [33] were obtained and combined with these gene-gene interactions from pathways to construct the global protein–protein interaction network, which consist of 13,013 genes and 276,845 reliable interactions. Furthermore, we also obtained 17,939 genes and 9,972,783 interaction edges along with their combined score (used as edge weights for locating subpathways) from String database (v9.0) [34].

In addition, 2212 pathways from SBmaps [35] and the PCNet [36] were obtained to validate the robustness and reliability of method. For each pathway in SBmaps, genes were map to PPI network from String database and extracted interactions with combined interaction score at least 900. Genes which involved in these extracted interactions were selected as the corresponding pathway genes. Then, pathways which contain at least 5 selected genes were retained and finally 1061 pathways in SBmaps were used for further analysis. The above filtered steps ensure that interactions between genes in each SBmaps pathway are high confidence. The PCNet contains 19,781 genes and 2,724,724 interactions.

2.1.3. Cancer hallmarks and oncogenic pathways

We obtained 35 Gene Ontology (GO) sets that belonged to 10 cancer hallmarks according to a previous study [37]. Ten oncogenic

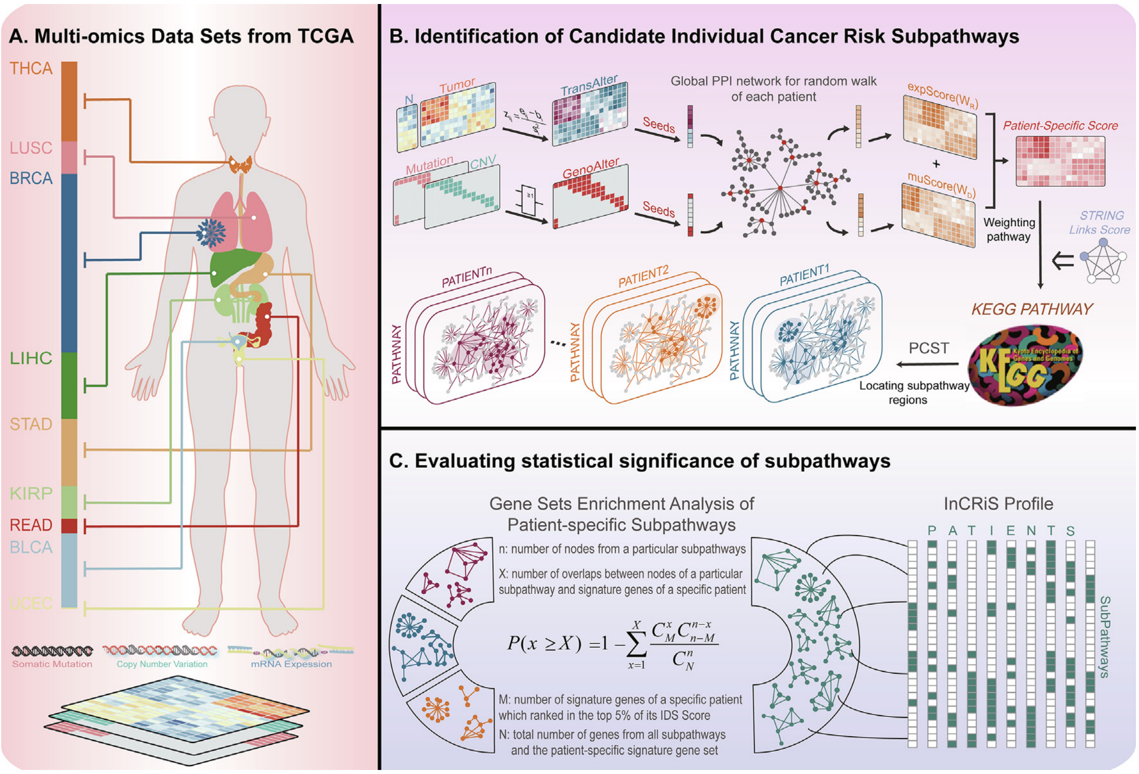
A. Multi-omics Data Sets from TCGA
THCA
LUSC
BRCA
LIHC
STAD
KIRP
READ
BLCA
UCEC
Somatic Mutation
Copy Number Variation
mRNA Expression
B. Identification of Candidate Individual Cancer Risk Subpathways
N
Tumor
$z_{ij} = \frac{e_{ij} - b_i}{s_i^2}$
TransAlter
Seeds
Global PPI network for random walk of each patient
expScore(W_R)
+
muScore(W_D)
Patient-Specific Score
Mutation
CNV
≥1
GenoAlter
Seeds
Weighting pathway
STRING Links Score
KEGG PATHWAY
PCST
Locating subpathway regions
PATIENTn
PATIENT2
PATIENT1
PATHWAY
C. Evaluating statistical significance of subpathways
Gene Sets Enrichment Analysis of Patient-specific Subpathways
n: number of nodes from a particular subpathways
X: number of overlaps between nodes of a particular subpathway and signature genes of a specific patient
$P(x \geq X) = 1 - \sum_{x=1}^{X} \frac{C_M^x C_{n-M}^{n-x}}{C_N^n}$
M: number of signature genes of a specific patient which ranked in the top 5% of its IDS Score
N: total number of genes from all subpathways and the patient-specific signature gene set
InCRiS Profile
PATIENTS
SubPathways

Fig. 1. Workflow of InCRiS method. (A) Multi-omics (mRNA expression, somatic mutation and somatic copy number) data of 9 cancer types from TCGA. (B) Locating risk subpathway regions at individual level. First, the transcriptome and genome dysregulation score of pathway genes were evaluated based on random walk algorithm. Then, we integrated patient-specific score of pathway genes and weighted pathways. Finaly, PCST method was used to locate candidate individual risk subpathway regions. (C) Evaluating the statistical significance of subpathways.

functions that highlighted by TCGA analysis working groups [38] were selected for further validation. Eight of these ten functions were matched to our used KEGG canonical pathways (RAS, Cell cycle, PI3K, p53, Notch, Wnt, Hippo, TGF-Beta).

2.2. Methods

2.2.1. Accessing the transcriptomic dysregulation of genes in individual

To access the transcriptomic dysregulation of genes in individual (Fig. 1B), we calculated patient-specific z-test statistic [39] for per tumor sample. Under formula (1),

$$z = \frac{e - b}{s^2}$$

where, z represented the z-score of gene in a certain patient; e was the particular expression value of gene in the patient; b and s, were mean and standard deviation of the gene in normal samples of the same cancer type, respectively. The patient-specific z-score was calculated to be positive if one gene overexpressed in tumor sample, and be negative opposite.

2.2.2. Construction of genomic alteration profile

We integrated somatic mutation and somatic copy number variation profile to construct genomic alteration profile (Fig. 1B). If a gene perturbed in mutation or copy number variation (CNV) level, we considered it was genomic alteration and marked 1 in the alteration profile, otherwise 0 was assigned.

2.2.3. Calculating the integrative dysregulation score of genes at individual level

We first evaluated the dysregulation score of genes within pathway at genomic and transcriptomic level, respectively. By comprehensive consideration of transcriptomic and genomic alterations, we then integrated gene scores from different omics levels as the integrative dysregulation score of genes at individual level (Fig. 1B). The detail processes were as follow.

We evaluated the transcriptomic and genomic dysregulation score of genes based on the random walk (RW) with restart algorithm [40]. RW is an important network diffusion algorithm which could effectively measure the relationships among nodes in the network. The transcriptomic (genomic) dysregulation score W_R (W_D) were calculated as follows:

$$W^{t+1} = (1 - r)MW^t + rW^0 \tag{2}$$

where M represented the adjacency matrix of the above constructed global PPI network; W correspond to W_R (W_D) in the following; W^0 was the initial weight vector of each gene on the network; when calculating the transcriptomic dysregulation score (W_R) of genes in each individual, the value of each genes in W^0 was set as their absolute value of z-score in the corresponding patient; when calculating the genomic dysregulation score (W_D) of genes in each individual, values in W^0 was set as 1 for genes with genomic alteration in the corresponding patient while the value of other genes were set as 0; W^t is the weight vector at time t and r was set as 0.7. In each individual, we evaluated the transcriptomic (genomic) dysregulation score W_R (W_D) for each gene based on the above RW with restart algorithm. Then, we calculated the integrative dysregulation score (IDS) of gene in each individual as follow:

$$IDS = W_D + W_R \tag{3}$$

where IDS was the integrative dysregulation score of a given gene in the corresponding individual.

2.2.4. Locating cancer risk subpathway regions in individual

We firstly weighted the nodes and edges for these reconstructing KEGG pathways at individual level. For each pathway, the edge weight were defined as:

$$Se_{gg'} = 1 - score/1000 \tag{4}$$

where $Se_{gg'}$ represented the edge weight between the two interacted genes *g* and *g'* within the given pathway, *score* was the combined interaction score that obtained from String database for genes *g* and *g'*. Here, *score* values in String database range from 0 to 1000, so that $Se_{gg'}$ is between 0 and 1. The nodes (genes) of each pathway were weighted at individual level. For a given pathway, the node weights in a specific patient were assigned as the *IDS* score of each gene in the corresponding patient.

Then, we used the PCST algorithm [41,42] to locate subpathway regions that were heavily disturbed and tightly connected within each entire pathway at individual level (Fig. 1B). The PCST method aims to locate subpathway(s) *G'* = (*N'*, *E'*) within entire pathway *G*= (*N*, *E*) that was (were) satisfied:

$$\min_{\substack{E' \in E, N' \in N \\ (E', N')connected}} \sum_{e \in E'} S_e - \mu \sum_{n \in N'} IDS_n \tag{5}$$

where S_e was the weight of edge *e* which was contained in the current subpathway *G'*, and IDS_n was the weight of gene *n* in the specific patient. Here, we considered edges as equally important as nodes in a pathway. Thus, the value of μ was set as 1.

2.2.5. Evaluating statistical significance of subpathways

We used hypergeometric test to assess the statistical significance of these located patient-specific subpathways (Fig. 1C). The statistical significance P-value of each subpathway in a specific patient was calculated as:

$$P(x \geqslant X) = 1 - \sum_{x=1}^{X} \frac{C_M^x C_{N-M}^{n-x}}{C_N^n}$$

in which *n* was number of genes from a particular subpathway, *M* was the number of patient-specific signature genes which were defined as genes with IDS score ranked top 5% in the corresponding patient, X was the number of overlap genes between the subpathway and patient-specific signature genes, N was total number of genes from all subpathways and the patient-specific signature gene set. Then, the P-value were corrected by using the Benjamini & Hochberg (BH) method which was implemented in the p.adjust() function of R. Finally, subpathways with BH corrected P-value < 0.01 and contained more than 5 genes (size greater than 5) were identified as individualized cancer risk subpathways.

2.2.6. Evaluating associations between subpathways and cancer hallmarks

We used hypergeometric test to evaluate association between subpathways and cancer hallmarks. First, the significance P-value of hypergeometric test between each subpathways and GO terms of cancer hallmarks was calculated to assess whether the overlap genes between them was significant. The background gene set were defined as unification of genes from all subpathways and genes from cancer hallmark processes. A subpathway was regarded to be associated with a cancer hallmark if the P-value between this subpathway and at least one GO term that belonged to the corresponding hallmark was<0.01.

2.2.7. Evaluating the specific activity of subpathway in tumor individual

We defined InteScore, which was referred the measure that used to estimate gene set expression from the study of Levine et al. [43], to evaluate the specific activity of subpathway in tumor individual. The InterScore of each subpathway was calculated as follow:

$$\text{InteScore} = \frac{mean' - mean}{S} * \sqrt{n}$$

where *mean'* was average *IDS* score of genes in a given subpathway, *mean* and *S* was mean and standard deviation IDS score of all genes and *n* was number of genes in the subpathway.

2.2.8. Survival analysis

We used R package "survival" (https://CRAN.R-project.org/package = survival) to estimate overall survival for patients and statistical significance among groups.

3. Results

3.1. Subpathways identified by InCRiS are associated with cancer hallmarks and oncogenic function

We firstly applied our individualized cancer risk subpathway identification method (InCRiS) to nine cancer types in TCGA including BLCA, BRCA, KIRP, LIHC, LUSC, READ, STAD, THCA and UCEC. Totally, 20,307 unique subpathways in 2925 individuals were identified. We further dissected the number of subpathways that dysregulated in individuals of each cancer type, and found 2781 (95%) individuals have more than 10 dysfunctional subpathway regions. The average numbers of subpathway identified for individuals were at the same level across different cancer types (Fig. 2A). However, the number of individualized dysfunctional subpathways varied within the same cancer type (Fig. 2A). We also found that most subpathways were dysregulated in few individuals; while a small subset of subpathways, denoted common subpathways, were dysregulated in many individuals (Fig. S1A). This shows that these pathway regions were widely dysfunction in cancer individuals. The above analysis suggests the heterogeneity of dysfunction in cancer individuals, which highlighted the importance of individualized subpathway identification.

Then, we evaluated the accuracy of our method by assessing the association between individualized cancer risk subpathways (InCRiSs) and cancer hallmarks/oncogenic processes (see Materials and methods). The result showed that most of InCRiSs (74%) were significantly involved in at least one cancer-related function. Mostly, InCRiSs were associated with Self Sufficiency in Growth Signals (62%) and Tissue Invasion and Metastasis (42%) (Fig. 2B), which were confirmed hallmark functions in cancer occurrence and development. Majority of patients had more than 75% of InCRiSs that related with cancer hallmarks (Fig. S1B). The high overlap between InCRiSs and cancer hallmarks implied that subpathways identified by InCRiS were functional associated with cancer hallmarks. We dissected the oncogenic pathways in which InCRiSs located for each cancer type and found that RAS, Cell Cycle and PI3K oncogenic pathways were more widely dysregulated in cancer individuals. In particular, all of READ patients had PI3K dysfunction and most of them (greater than90%) had RAS, Cell Cycle, Wnt, TGF-Beta oncogenic functions dysregulated (Fig. 2C, Fig. S2). In addition, Hippo was dysregulated in THCA (92%), yet p53 Signaling pathway was disordered in LUSC (63%) and KIRP (62%) individuals. Although p53 Signaling pathway was medium-frequently recognized in patients, the p53 and its upstream regulator p300 tend to be essential in other canonical pathways such as PI3K and Ras. In summary, our framework can efficiently locate well-

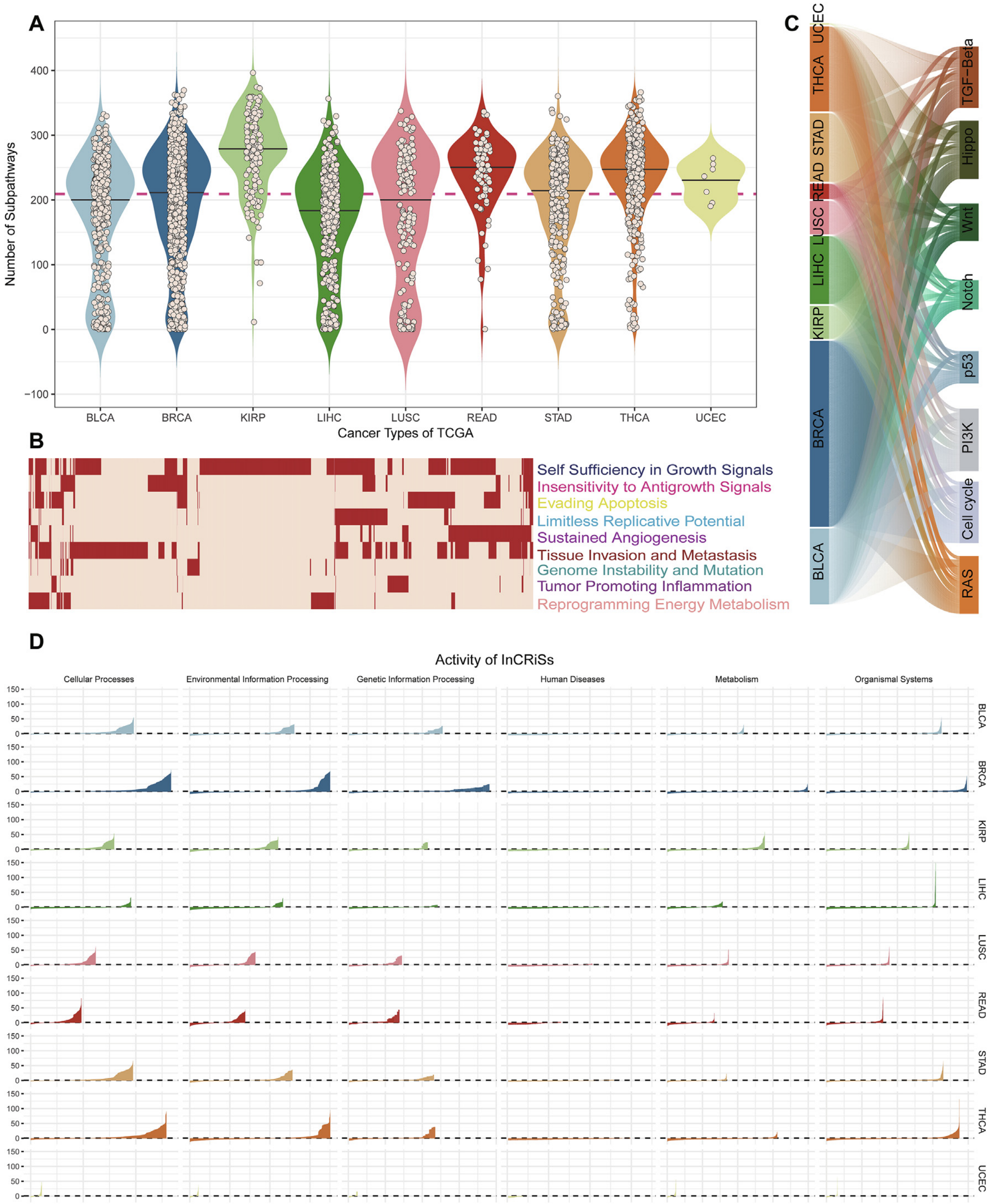

Fig. 2. Characterization of individual cancer risk subpathways. (A) Number of risk subpathways identified in each patient across TCGA cancer types. Red dashed in violin plots marks overall average value. (B) Heatmap shows which hallmarks the InCRiSs were associated (red). (C) Oncologic pathways were dysfunction in patients of 9 cancers. Lines connected tumor individuals and oncologic pathways, which represents subpathway region(s) within that pathway is (are) dysregulated in the corresponding individual. (D) Subpathway activity in TCGA cancers. Activity of InCRiSs in each TCGA cancer type was evaluated by their expression profile. Subpathways were grouped based on the systemic classes of their locating entire pathways from KEGG. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

known fundamental dysregulated functions which had various activities among cancers (Fig. 2D) and those subpathways were strongly linked to oncogenic function.

3.2. Method comparison and evaluation

In this section, we aim to evaluate the feasibility of our InCRiS method. First, InCRiS was compared with other methods including GSEA [10], Clipper [13], HotNet2 [16] and LEANR [44]. For convenience of method comparison, subpathways identified by InCRiS and Clipper were mapped to entire pathways. These methods were compared at entire pathway level. Firstly, we obtained 41 experimentally validated cancer associated pathways from our previously developed CPAD database [45] which were used as gold standard. We found that InCRiS recalled the most number of experimentally validated cancer associated pathways in 5 (BLCA, KIRP, LIHC, LUSC and THCA) out of these 8 cancer types (Fig. 3). InCRiS also recalled relative high number of these validated pathways in

BRCA (11) and STAD (7). This suggests that identifying risk pathways at the individual level can capture more comprehensive cancer pathways. Furthermore, we found that clipper, a method that identified risk subpathway regions as InCRiS, also identified relative high number of experimentally validated pathways across these eight cancer types (Fig. 3). This indicates that subpathway strategy could identify cancer risk pathways more accurately. Next, we dissected pathways uniquely identified by InCRiS in each cancer type and found that many of these pathways have been reported to be associated with the initiation and progression of tumor (Fig. S3). InCRiS thus not only provide risk pathway regions at individual level, but also display considerable potential to complement currently pathway identification methods.

Then, we further evaluate the robustness and reliability of InCRiS method. We applied our approach to updated network (PCNet), pathways (SBmaps) and independent multi-omics datasets (LIHC and THCA) (see data collection) respectively to evaluate effects of the choice of these conditions on the results. First, we updated the original integrated PPI network with PCNet and assessed the consistency of identified subpathways with the original results (subpathways identified based on the integrated PPI) in individuals of LIHC and THCA from TCGA (see supplementary method). We evaluated the overlap significance of entire pathways which correspond to subpathways identified based on two different networks in each individual. We obtained one significance P-value of hypergeometric test for each individual and counted the number of individuals that with significant consistency (overlap) (hypergeometric test $p < 0.05$) of results. In 98% (770 of 786) individuals, significant consistency of results at entire pathway level were observed. We next calculated the consistency and recall ratios of identified subpathways in each individual (see supplementary method). The higher consistency ratio represents the greater proportion of subpathways identified in individual *S* consistent with the original results. The higher recall ratio represents the greater proportion of original KEGG subpathways recalled in individual *S*. Higher consistency and recall ratios indicate higher robustness of InCRiS method with respect to the influence of factors such as network and thus InCRiS method was reliable for identifying individualized risk subpathways. Evaluation results shown that consistency and recall ratios were greater than 0.7 in 74% and 70% individuals respectively (Fig. S4A). This indicated that results before and after network switching showed relatively high consistency at the sub-pathway level. Then, we identified SBmaps pathways in individuals of LIHC and THCA from TCGA. Results shown that SBmaps pathways identified in individuals exhibit high consistency and recall rates in more than 90% individuals (see supplementary method, Fig. S4B). The above results suggest network and pathway data have limit effect on the results of InCRiS method. Finally, we evaluated InCRiS method using independent datasets with small sample size of LIHC (50 tumor samples) and THCA (25 tumor samples) from PCAWG database. We assessed the consistency of entire pathways and subpathways identified in different proportions of samples (≤1%, 1%∼5%, 5%∼15%, 15%∼25%, 25%∼50%,50%∼75%, 75%∼100%) of LIHC and THCA cancer types from two different data sources (see supplementary method). It shown that the results of independent datasets had high consistency at both entire pathway and subpathway levels with the original results (Fig. S4C). In summary, all the above results demonstrate that InCRiS method is robust with respect to the influence of factors including network, pathway, dataset and also sample size and thus reliable for identifying InCRiSs.

3.3. Dissecting InCRiSs reveals key oncogenic subpathway regions

Although dysregulation of subpathway regions has individual difference, some subpathways which may be core function for the initiation and progression of cancer were widely dysregulated across tumor individuals. Thus, we next focused on subpathways that widely disturbed in individuals of different cancer type. Firstly, the InteScore [43] for subpathways which represents the specific activity of subpathway in tumor individuals were calculated (see Materials and Methods). Then, we defined InCRiSs whose InteScore ranked in top 5% and dysregulated in more than 30% individuals as common subpathways. Here, we excluded UCEC because the small sample numbers (only 7). In total, 440 common subpathways were defined across the remaining 8 cancer types, and 200 of these subpathways were related with the corresponding cancer types. All these cancer-related common subpathways were participated in 93 entire pathways (functions). In all these cancers, we found over 60% of common functions (65% in BLCA, 85% in BRCA, 60% in KIRP, 81% in LIHC, 63% in LUSC, 66% in READ, 66% in STAD and 60% in THCA) have been reported that have influence the initiation and progression processes of the corresponding cancer type (Table S2-S3). Especially, 44 of 52 (85%) and 39 of 48 (81%) common functions were supported by literatures to actively participate in BRCA and LIHC respectively. For example, activation of Toll-like Signaling Pathway inhibits cell proliferation and tumor growth in breast cancer [46]. Via MAPK signaling pathway, telekin showed anti-proliferation effects against human hepatocellular carcinoma cells [47]. These results demonstrated that InCRiS could efficiently capture oncogenic subpathway regions which may server as key resource for researches of oncogenesis mechanism.

Next, we further identified key subpathway regions that widely dysregulated in pan-cancer by combing clinical survival data of cancer individuals. Candidate subpathways were firstly screened from these 200 cancer-related common subpathways by univariate cox regression analysis (P-value < 0.2). Considering heterogeneity between cancer types, we further used stratified cox regression to select subpathways from the above candidates. We selected subpathways which had significant influence on survival with P-value < 0.2 of univariate cox regression analysis and with P-value < 0.05 of stratified cox regression from all common subpathways as pan-cancer oncogenic subpathways. As a result, 9 subpathways were identified (Table S4) and the topological structure of 7 subpathways within KEGG were shown in Fig. 4. A classical oncogenic subpath (Ras (HRAS, KRAS)->Raf (RAF1)->MEK (MAP2K1, MAP2K2)->ERK (MAPK1, MAPK3)) within PI3K-AKT pathway was identified (Fig. 4). Subpathway regions within some famous signaling pathways such as PI3K-AKT, ErbB and Apoptosis contained well-known cancer genes (KRAS, JUN, AKT etc.). Some cancer genes such as PAK1 not only had high degree of dysregulation at transcription level, but also has a relatively high frequency of genomic (copy number) variation across tumor individuals (Fig. 4). Furthermore, some cancer genes such as KRAS and RAF1 may be altered at only one omics level. This further indicated that integrating multiple omics data can more effectively capture subpathways associated with tumorigenesis.

3.4. Individualized subpathways applied for molecular classification of pan-cancer

The division of cancer molecular subtypes is a key step towards achieving precision medicine. The distinct patterns of subpathway dysregulation in individuals may lead to tumor heterogeneity. Therefore, we next investigated whether these identified individualized risk subpathways can be applied to split tumor samples from different cancer types into unified molecular classification (subtype). We performed molecular classification by the following steps: 1) select InCRiSs that dysregulated in more than 5% tumor samples; 2) remove samples that have<11 dysregulated subpathways; 3) construct a binary matrix that represented InCRiS profile of pan-cancer, in which the element 1(0) represented the InCRiS

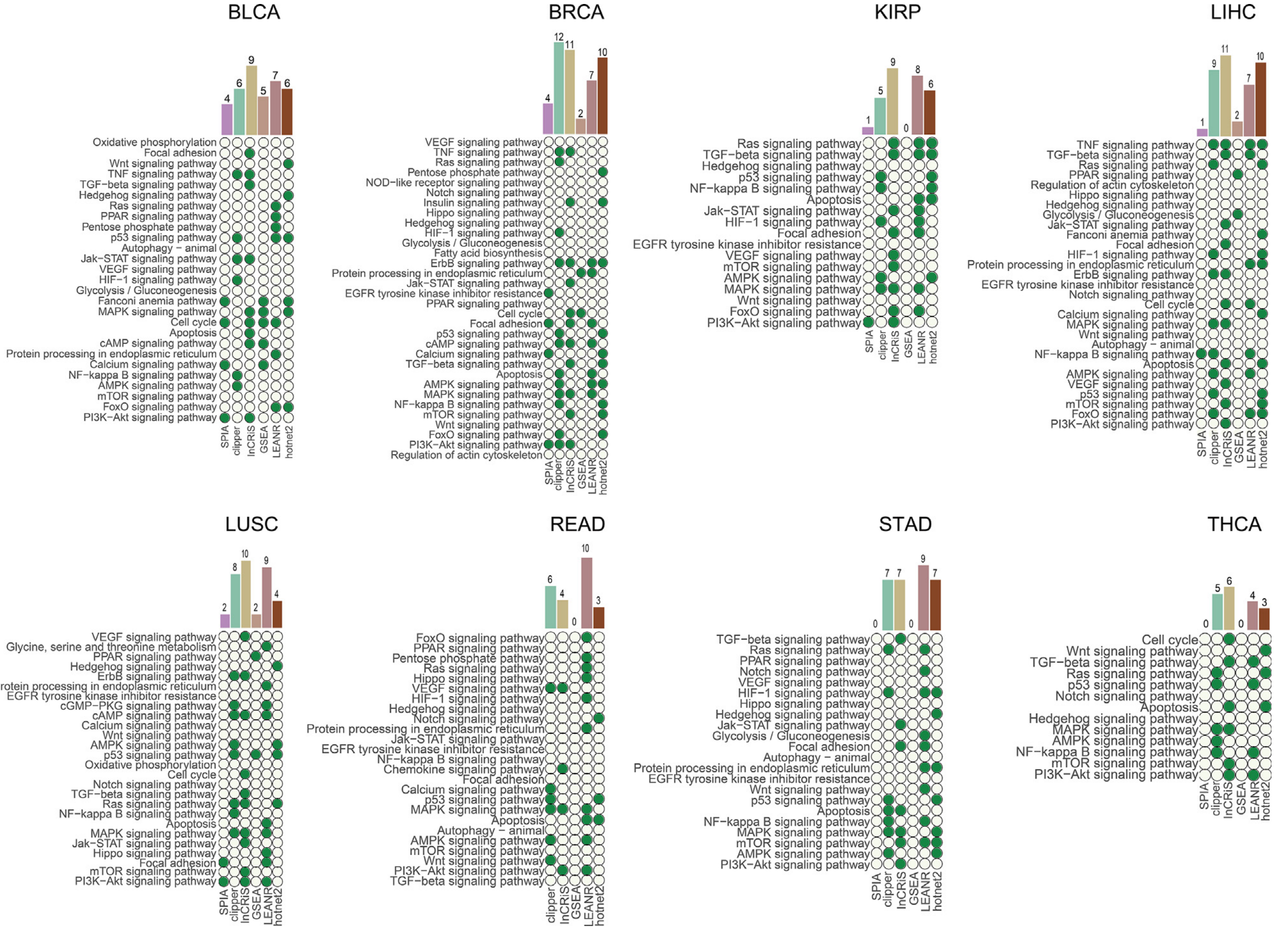

Fig. 3. Comparison of InCRiS method with LEANR, GSEA, Hotnet2 and clipper for identifying cancer related pathways. CPAD onco-pathways are used as gold standard and green dots marked onco-pathways which were identified by each algorithm. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

have (not) dysregulated in the corresponding tumor individual; 4) classify subtypes based on the hierarchical clustering for InCRiS profile (see supplementary method). As a result, 11 clusters of samples (InCRiS subtypes) across nine cancer types were classified (Fig. 5A). We first validated the reliability of the 11 molecular classifications of pan-cancer (i.e. 11 InCRiS subtypes). The validation processes were as follows: 1) the cancer samples were randomly separated into two groups (one group contains 50% samples) as training set and test set respectively; 2) performed molecular classification on the two datasets respectively (Fig. S5); 3) identifying featured subpathways for each cluster (InCRiS subtype). A subpathway is considered to be a featured subpathway of the cluster if the subpathway dysregulated in more than 70% individuals of the cluster; 4) the consistency of featured subpathways and sample distribution for 11 molecular subtypes in different datasets were evaluated. The results shown that 7 of the 11 clusters (C1, C2, C3, C6, C7, C9 and C10) from training and test sets exhibit relative high consistency with respect to the featured subpathways and sample distribution (Table S5 and Figs. S5-S6).

Dissection of the sample distribution found that almost one third samples were derived from BRCA and they distributed in all 11 InCRiS subtypes, especially in subtype C1, C2 and C9 (Fig. 5B). This is consistent with previous research findings that BRCA has a high degree of heterogeneity [48–51]. Then, we dissected tumor samples within each subtype and found that C5, C7 and C8 clusters mainly contained samples from three different cancer types (BRCA, STAD and THCA) which all belonged to adenocarcinoma type. C4 subtype (BRCA-free subtype) contains only one BRCA sample and samples from KIRP, LUSC, READ and THCA (Fig. 5B). In particular, samples from KIRP and READ are mainly distributed in the C4 subtype, which may imply an associated molecular mechanism of these individuals from these two cancer types.

To evaluate the clinical relevance of subtypes that stratified based on InCRiSs, we performed overall survival analysis based on the survival time data of tumor samples. Result showed that the overall survival rate of these 11 subtypes were significantly difference (log-rank: p = 1.69e-05, Fig. 5C). We calculated the mean and median survival time of individuals in each cluster (Table S6). A mixed cluster (C7), which comprised a small number of BLCA samples, a few BRCA samples and some STAD samples, had worse prognosis (the shortest median survival time: 16.7 months). Cluster C6, which was a cluster nearby C7, was also a mixed cluster subjected by BRCA and THCA, had much better prognosis (median survival time: 23.18 months). In addition, cluster C8 which was dominated by samples from BRCA and THCA exhibited better prognosis than patients in other clusters (median survival time: 31.7 months) (Fig. 5B-C & Table S6). In summary, the above analysis suggest these InCRiS subtypes could provide the similarities and differences of pan-cancer samples at histopathology and clinical outcome.

Breast cancer has high heterogeneity which greatly impacts the clinical outcomes and drug response of individuals [52]. In our study, breast cancer samples were contained in most of pan-cancer subtypes and we found that patients contained in different

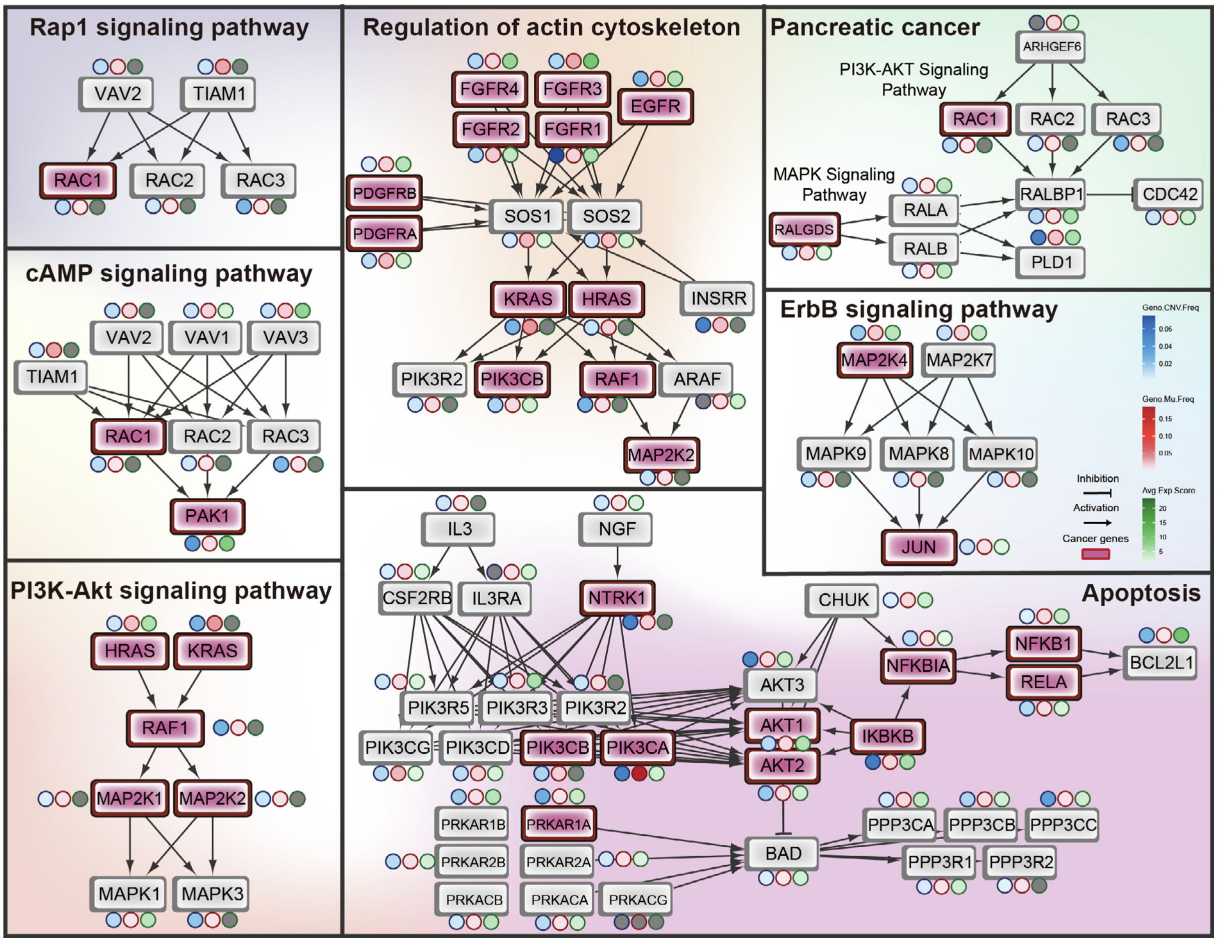

Fig. 4. Subpathway regions that commonly dysregulated in tumor individuals. Seven common subpathways are showed regarding KEGG pathway topology. Cancer drivers are marked in red squares. The points sticking on nodes represents CNV frequency, somatic mutation frequency and the average expression differential score (absolute value of z-score) among individuals respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

subtypes had significantly different survival times (Fig. S7). Currently, breast cancer were classified into several intrinsic molecular subtypes (Luminal A, Luminal B, HER2-enriched and Basal-like, etc.) based on gene signatures such as ER,PR and HER2, which help the treatment options and have significantly improved therapeutic effects of breast cancer. However, there are still some patients within the same molecular subtype that failed to respond to the selected therapeutic strategy. For example, approximately 20 percent of the ER + patients were insensitive to endocrine therapy or developed acquired drug resistance [52]. Thus, we further aimed to classify breast cancer patients based on these identified InCRiSs and hierarchical clustering. As a result, six subtypes of breast cancer were identified (Fig. S8A-B). In HER2 and Luminal A, survival analysis found that patients contained in different subtypes which were classified based on InCRiSs had significantly different ($P < 0.05$) survival outcomes (Fig. S8C-D). This indicated that these InCRiSs-based subtypes may provide novel insights into precise discrimination of breast cancer subtype.

3.5. Functional characterization of pan-cancer InCRiS subtypes

Changes of genes at different omics levels in the subpathway region will lead to the corresponding functional disorders, and then perturb cancer hallmark processes and finally result in the occurrence and development of cancer. The heterogeneity of dysfunction in cancer individuals may contribute to the difference of clinical outcomes and drug response of samples in different subtypes. Then, we looked to see what functional alterations are shared within these pan-cancer subtypes. As a result, individuals of subtype C1, C2, C4, C5 and C11 have distinct dysregulated subpathway regions (Fig. 5). To explore how these functionally dysregulated subpathways contribute to the pathogenesis of cancer, several subpathway regions (Wnt, Proteoglycans, Rap1, Cell cycle and MAPK subpathways) were further examined (Fig. 6A). Among them, a beta-catenin centered subpathway region that locates within Wnt signaling pathway tend to dysregulate in individuals of C5 subtype. Functional analysis of this subpathway region found it associated with cancer progression related hallmark processes including ‘tissue invasion and metastasis’ and ‘genome instability and mutation’ (Table S7), which was consistent with the relatively poor clinical outcomes of this subtype. Further exploring this subpathway found that genomic variation of genes such as CTNNB1, EP300 and CREBBP were dominated by mutation (i.e. mutation frequency of these genes rank in top 1% in samples of the corresponding subtype), while that were dominated by copy number variation (CNV) for genes such as CCND1, MYC, VANGL2 and NKD2 in subtype C5 (Fig. 6B). The same situation was also found for genes including MAP3K13, MECOM (upstream) and TP53 (downstream) within MAPK subpathway region that tend to dysregulate in individuals of C4 subtype. These genes may be potential biomarkers and targets for precise diagnosis and treatment of cancer. We also identified one cell cycle subpathway region that mainly dysregulated in individuals of C2 and C11 subtypes (Fig. 6A). It is worth to note that dysregulation of PRKDC gene was mainly at genomic

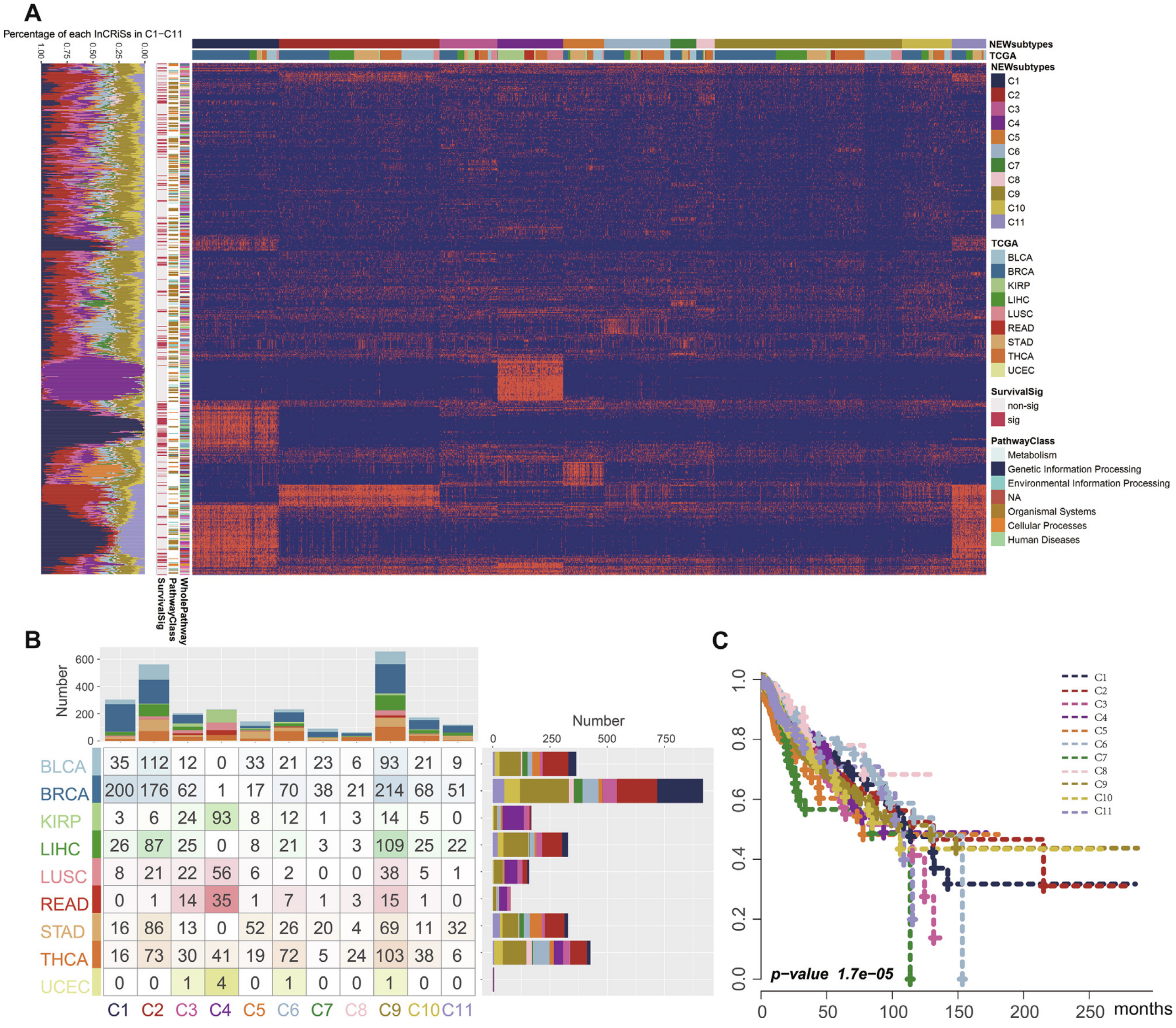

Fig. 5. Molecular stratification in pan-cancer. (A) Heatmap presents subpathways (row) identified in tumor patients (column), if a subpathway were identified in individual, the corresponding cell is colored orange, else is blue. Eleven distinct molecular subtypes were annotated in upper column bar and TCGA cancer types under. As for the row annotation, we tagged: which pathway the subpathway belonged (right); which pathway class of entire pathway that contains the corresponding subpathway region (middle); if the subpathways associate with survival (left); we assume that one subpathway associate with survival if there is significant difference (p value < 0.05) on overall survival between it dysregulated samples and others. Bar plot presents the distribution of these lnCRiS in C1-C11. (B) Distribution of lnCRiS subtype samples across different TCGA cancer types. (C) Overall survival of eleven lnCRiS subtypes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

level, but that were dominated by different variation types in these two subtypes (mutation for C2 and CNV for C11 respectively). Except for genomic variation, some genes such as CDK1, CDC25C and PKMYT1 in this subpathway region are mainly transcriptional dysregulation in the corresponding subtypes (C2 and C11) (Fig. 6B). The above analysis suggests changes of genes at different omics levels may synergistically lead to functional dysregulation of their locating subpathway regions. Thus, integrating multi-omics data may identify dysfunctional subpathways more comprehensively. Furthermore, a proteoglycans subpathway region, which associated with 'Insensitivity to Antigrowth Signals', 'Limitless Replicative Potential' and 'Tissue Invasion and Metastasis' hallmark processes, was mainly dysregulated in individuals of C1 subtype (Fig. 6A and Table S7). We found that the Ras->Raf->MEK->ERK path within this subpathway which is a crosstalk region between RAP1 and proteoglycans pathway was also mainly dysregulated in individuals of C11 subtype. This suggest that identification and analysis of subpathways at individualized level can provide more detailed information for the pathogenesis of cancer.

4. Discussion and conclusion

Dysregulation of genes in cancer cells at different levels of omics affect the corresponding dysfunction of pathways in which they located, thus leading to the occurrence and development of malignant tumors. However, the disordered pathways vary in different cancer individuals. Therefore, it is of great significance to integrate multidimensional omics data to identify dysregulated pathways at the individual level for the study of pathogenesis and heterogeneity of cancer.

We firstly provided an individualized subpathway identification method by considering changes of genes in the pathway at different omics levels and the strength of connections between genes.

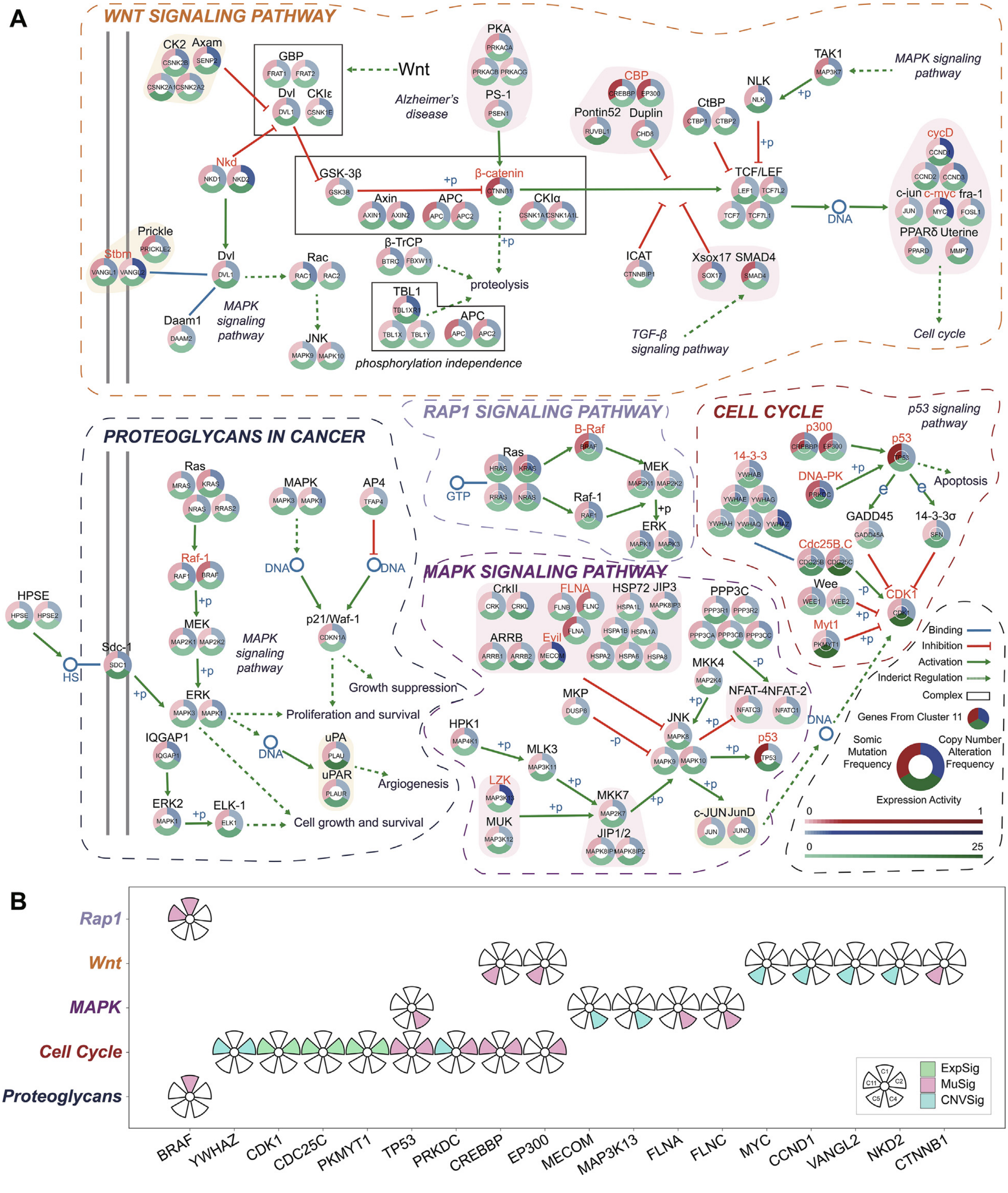

Fig. 6. Five subpathway regions that tend to dysregulate in individuals of C1 (Proteoglycans in cancer), C2 (cell cycle), C4 (MAPK signaling pathway), C5 (WNT signaling pathway) and C11 (Rap1 signaling pathway, cell cycle). (A) Pies are genes involved in subpathways that tend to dysregulate in individuals of C11 subtype and cyclic annular are genes involved in subpathways that tend to dysregulate in individuals of other subtypes. The colors in pies and circles represents somatic mutation frequency (red), CNV frequency (blue) and gene expression activity (green) of the gene in the corresponding subtype, respectively. (B) Significantly dysregulated genes at different omics level among individuals of different subtypes. If the dysregulated (mutation, CNV or differently expressed) frequency of one gene (involved in subtype specifically dysregulated subpathway) rank top 1% across the subtype individuals, it is defined as significantly dysregulated gene. Genes with z-test Pvalue < 0.01 are regarded as differently expressed in individual. Different markers are painted in prasinous (differently expressed), pink (mutation) and lightblue (CNV). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Then, we applied the method to 2925 individuals across nine cancer types of TCGA and validated the reliability of method by evaluating associations between identified subpathways and cancer hallmark processes/oncogenic pathways. In the pipeline of InCRiS method, three types of data were used including pathway data, PPI network and multi-omics data of tumor samples. Databases such as Biocarta, Reactome, HumanCyc were integrated to construct a comprehensive PPI network that was used in the InCRiS pipeline. And KEGG is a widespread used pathway data source. When the InCRiS method was used in other studies, only multiple omics data of individuals need to be changed. Furthermore, we also evaluated and demonstrated the robustness and reliability of InCRiS method with respect to PPI network, pathway, independent datasets and sample size. Thus, the InCRiS framework may have the potential to be widely used in other studies. On the other hand, we removed genes with missing expression value greater than 20% in the expression profile to ensure that the gene used for analysis expressed in as many samples as possible. However, this might remove many important mutated genes. Thus, a lower filtration threshold can be considered when applied to other studies.

InCRiS results can provide information about disordered subpathway regions at individual level. In addition, based on the dysregulated subpathways in individuals identified by InCRiS, it is also can study common dysregulated subpathways in cancer or cancer-specific subpathways. We identified core subpathway regions which were widely dysregulated in pan-cancer samples based on the InCRiS results in 2925 individuals. These subpathway regions may help researchers to uncover the pathogenesis of cancer and key genes in these pathway regions may have the potential as targets for developing novel therapeutic drugs. Based on these individualized risk subpathway regions, we constructed the pan-cancer dysregulation subpathway profile and identified 11 pan-cancer molecular subtypes with significant differences in clinical outcomes.

Further analyzing subpathway regions that tend to dysregulate in individuals of specific subtypes provided some biological insights for the pathogenesis of cancer. First, dysregulation of different genes within risk subpathway may be dominated by different omics levels. This suggests that disorder of these risk subpathways may be the result of the synergistic effects for genes in it at different omics levels. Thus, it is essential to integrate multi-omics data for identifying cancer risk pathways. Second, we discovered that some key genes were significantly changed at different omics levels within subpathways such as (Fig. 6B). These genes may have the potential to act as novel therapeutic targets and biomarkers for molecular subtype classification of cancer. Third, dissecting several subpathway regions found that ‘Tissue Invasion and Metastasis’ cancer hallmark process tend to disturb in individuals of multiple subtypes such as C1, C4, C5 and C11 (Table S7), but the dysregulated subpathway regions were different among individuals of these subtypes. This further demonstrates that the identification of subpathway could provide more precise information.

Recently, there are many popular used methods such as GSEA [10]), SPIA [11], clipper [13], HotNet [15], HotNet2 [16] and PARADIGM [20] for identification of risk pathways in human diseases. Among these methods, clipper and HotNet/HotNet2 identified risk subpathways (subnetworks) based on only single omics data (transcriptome or genomic data respectively) at cohort level, while PARADIGM inferred patient-specific pathway activities by integrating multi-omics data. Compared with these methods, our method was characterized by focusing on identification of subpathway regions at the individual level and had several unique advantages. First, individualized identification of risk subpathways can comprehensively capture oncogenic pathways (Fig. 3). Second, identification of subpathway can help researchers to more precisely target the individual carcinogenic pathway region, and provide more refined guidance for the precise targeted therapy and pathogenic mechnisms of cancer. Thus, our method has considerable potential to complement these currently existed methods.

In summary, we provided a computational framework to identify individualized risk subpathway regions in cancer and then applied it to pan-cancer individuals. Our method and findings presented here will be useful for researches on the pathogenesis of cancer and its precision treatment.

Author contributions

Conceptualization: YPZ, YC and XL; Data curation: XZ and CXH; Formal analysis: YJX,JWW,FL and CLZ; Funding acquisition: YPZ and XL; Investigation: YJX,JWW,FL and CLZ; Methodology: YJX and FL; Project administration: YPZ, YC and XL; Supervision: YPZ and XL; Validation: XZ,YC,DSS and YQX; Visualization: XZ,YC and WQM; Writing—original draft: YJX,JWW and FL; Writing—review & editing: YJX,JWW and CLZ. All authors read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (Grant Nos. 61873075, 31801107 and 62101164), the National Key R&D Program of China (2018YFC2000100), the Heilongjiang Touyan Innovation Team Program, the China Postdoctoral Science Special Foundation (Grant No. 2020T130162), the China Postdoctoral Science Foundation (Grant No. 2019M661295), the Heilongjiang Postdoctoral Foundation (Grant No. LBH-Z19083), the Doctor Green Seedlings Breaking Ground Project of Harbin Medical University (Grant No. QMPT-2010) and University Nursing Program for Young Scholars with Creative Talents in Heilongjiang Province Youth Science and technology innovation personnel training project (UNPYSCT-2017059).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.01.022.

References

[1] Xie C, Li SY, Fang JH, et al. Functional long non-coding RNAs in hepatocellular carcinoma. Cancer Lett 2021;500:281–91.

[2] Khatri P, Draghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics 2005;21:3587–95.

[3] Goeman JJ, van de Geer SA, de Kort F, et al. A global test for groups of genes: testing association with a clinical outcome. Bioinformatics 2004;20:93–9.

[4] Pavlidis P, Qin J, Arango V, et al. Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. Neurochem Res 2004;29:1213–22.

[5] Sherman BT, Huang da W, Tan Q, et al. DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. BMC Bioinf 2007;8:426.

[6] Xie C, Mao X, Huang J, et al. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res 2011;39:W316–22.

[7] Xie Z, Bailey A, Kuleshov MV, et al. Gene Set Knowledge Discovery with Enrichr. Curr Protoc 2021;1:e90.

[8] Raudvere U, Kolberg L, Kuzmin I, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res 2019;47:W191–8.

[9] Kolberg L, Raudvere U, Kuzmin I, Vilo J, Peterson H. gprofiler2 – an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. F1000Res 2020;9:709.

[10] Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 2005;102:15545–50.

[11] Tarca AL, Draghici S, Khatri P, et al. A novel signaling pathway impact analysis. Bioinformatics 2009;25:75–82.

[12] Yang Q, Wang S, Dai E, et al. Pathway enrichment analysis approach based on topological structure and updated annotation of pathway. Brief Bioinform 2019;20:168–77.

[13] Martini P, Sales G, Massa MS, et al. Along signal paths: an empirical gene set approach exploiting pathway topology. Nucleic Acids Res 2013;41:e19.

[14] Gouy A, Daub JT, Excoffier L. Detecting gene subnetworks under selection in biological pathways. Nucleic Acids Res 2017;45:e149.
[15] Vandin F, Upfal E, Raphael BJ. Algorithms for detecting significantly mutated pathways in cancer. J Comput Biol 2011;18:507–22.
[16] Leiserson MD, Vandin F, Wu HT, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat Genet 2015;47:106–14.
[17] Arora A, Olshen AB, Seshan VE, et al. Pan-cancer identification of clinically relevant genomic subtypes using outcome-weighted integrative clustering. Genome Med 2020;12:110.
[18] Hoadley KA, Yau C, Wolf DM, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell 2014;158:929–44.
[19] Hoadley KA, Yau C, Hinoue T, et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell 2018;173(291–304):e296.
[20] Vaske CJ, Benz SC, Sanborn JZ, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics 2010;26:i237–45.
[21] Drier Y, Sheffer M, Domany E. Pathway-based personalized analysis of cancer. Proc Natl Acad Sci USA 2013;110:6388–93.
[22] Wang H, Cai H, Ao L, et al. Individualized identification of disease-associated pathways with disrupted coordination of gene expression. Brief Bioinform 2016;17:78–87.
[23] Weinstein JN, Collisson EA, Mills GB, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 2013;45:1113–20.
[24] Mermel CH, Schumacher SE, Hill B, et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol 2011;12:R41.
[25] Pan-cancer analysis of whole genomes. Nature 2020;578:82–93.
[26] Kanehisa M, Goto S, Sato Y, et al. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 2012;40:D109–14.
[27] Li C, Han J, Yao Q, et al. Subpathway-GM: identification of metabolic subpathways via joint power of interesting genes and metabolites and their topologies within pathways. Nucleic Acids Res 2013;41:e101.
[28] Matthews L, Gopinath G, Gillespie M, et al. Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res 2009;37:D619–22.
[29] Schaefer CF, Anthony K, Krupa S, et al. PID: the Pathway Interaction Database. Nucleic Acids Res 2009;37:D674–9.
[30] Caspi R, Billington R, Ferrer L, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 2016;44:D471–80.
[31] Mi H, Muruganujan A, Thomas PD. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res 2013;41:D377–86.
[32] Sales G, Calura E, Cavalieri D, et al. graphite - a Bioconductor package to convert pathway topology to gene network. BMC Bioinf 2012;13:20.
[33] Keshava Prasad TS, Goel R, Kandasamy K, et al. Human Protein Reference Database–2009 update. Nucleic Acids Res 2009;37:D767–72.
[34] Szklarczyk D, Morris JH, Cook H, et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res 2017;45:D362–8.
[35] Kuenzi BM, Ideker T. A census of pathway maps in cancer systems biology. Nat Rev Cancer 2020;20:233–46.
[36] Huang JK, Carlin DE, Yu MK, et al. Systematic Evaluation of Molecular Networks for Discovery of Disease Genes. Cell Syst 2018;6(484–495):e485.
[37] Plaisier CL, Pan M, Baliga NS. A miRNA-regulatory network explains how dysregulated miRNAs perturb oncogenic processes across diverse cancers. Genome Res 2012;22:2302–14.
[38] Sanchez-Vega F, Mina M, Armenia J, et al. Oncogenic Signaling Pathways in The Cancer Genome Atlas. Cell 2018;173(321–337):e310.
[39] Kim YA, Wuchty S, Przytycka TM. Identifying causal genes and dysregulated pathways in complex diseases. PLoS Comput Biol 2011;7:e1001095.
[40] Kohler S, Bauer S, Horn D, et al. Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet 2008;82:949–58.
[41] Bailly-Bechet M, Borgs C, Braunstein A, et al. Finding undetected protein associations in cell signaling by belief propagation. Proc Natl Acad Sci USA 2011;108:882–7.
[42] Balbin OA, Prensner JR, Sahu A, et al. Reconstructing targetable pathways in lung cancer by integrating diverse omics data. Nat Commun 2013;4:2617.
[43] Levine DM, Haynor DR, Castle JC, et al. Pathway and gene-set activation measurement from mRNA expression data: the tissue distribution of human pathways. Genome Biol 2006;7:R93.
[44] Gwinner F, Boulday G, Vandiedonck C, et al. Network-based analysis of omics data: the LEAN method. Bioinformatics 2017;33:701–9.
[45] Li F, Wu T, Xu Y, et al. A comprehensive overview of oncogenic pathways in human cancer. Brief Bioinform. 2020;21:957–69.
[46] Cai Z, Sanchez A, Shi Z, et al. Activation of Toll-like receptor 5 on breast cancer cells by flagellin suppresses cell proliferation and tumor growth. Cancer Res 2011;71:2466–75.
[47] Li L, Zheng BB, Ma LS, et al. Telekin suppresses human hepatocellular carcinoma cells in vitro by inducing G2/M phase arrest via the p38 MAPK signaling pathway. Acta Pharmacol Sin 2014;35:1311–22.
[48] Alom MZ, Yakopcic C, Nasrin MS, et al. Breast Cancer Classification from Histopathological Images with Inception Recurrent Residual Convolutional Neural Network. J Digit Imaging 2019;32:605–17.
[49] Lan C, Peng H, McGowan EM, et al. An isomiR expression panel based novel breast cancer classification approach using improved mutual information. BMC Med Genomics 2018;11:118.
[50] Banu AB, Thirumalaikolundusubramanian P. Comparison of Bayes Classifiers for Breast Cancer Classification. Asian Pac J Cancer Prev 2018;19:2917–20.
[51] Yanovich G, Agmon H, Harel M, et al. Clinical Proteomics of Breast Cancer Reveals a Novel Layer of Breast Cancer Classification. Cancer Res 2018;78:6001–10.
[52] Chung W, Eum HH, Lee HO, et al. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Nat Commun 2017;8:15081.