Molecular Genetics and Genomics (2023) 298:871–882
https://doi.org/10.1007/s00438-023-02022-4

ORIGINAL ARTICLE

Network diffusion-based approach for survival prediction and identification of biomarkers using multi-omics data of papillary renal cell carcinoma

Keerthi S. Shetty[1] · Aswin Jose[1] · Mihir Bani[1] · P. K. Vinod[1]

Received: 8 September 2022 / Accepted: 12 April 2023 / Published online: 24 April 2023

Abstract

Identification of cancer subtypes based on molecular knowledge is crucial for improving the patient diagnosis, prognosis, and treatment. In this work, we integrated copy number variations (CNVs) and transcriptomic data of Kidney Papillary Renal Cell Carcinoma (KIRP) using a network diffusion strategy to stratify cancers into clinically and biologically relevant subtypes. We constructed GeneNet, a KIRP specific gene expression network from RNA-seq data. The copy number variation data was projected onto GeneNet and propagated on the network for clustering. We identified robust subtypes that are biologically informative and significantly associated with patient survival, tumor stage and clinical subtypes of KIRP. We performed a Singular Value Decomposition (SVD) analysis of KIRP subtypes, which revealed the genes/silent players related to poor survival. A differential gene expression analysis between subtypes showed that genes related to immune, extracellular matrix organization, and genomic instability are upregulated in the poor survival group. Overall, the network-based approach revealed the molecular subtypes of KIRP and captured the relationship between gene expression and CNVs. This framework can be further expanded to integrate other omics data.

Keywords Network diffusion · Kidney papillary renal cell carcinoma · Multi-omics data · Gene interaction networks

Introduction

Pathologists traditionally classify cancers based on histological appearance and site of origin. However, this may not capture all the variations of the disease due to the different molecular aberrations comprising somatic mutations, copy number variations (CNVs) and DNA methylations (Zhao et al. 2019). Therefore, cancer is viewed as a heterogeneous disease with different subtypes. The stratification of cancer patients into clinically relevant subtypes based on different kinds of omics data (genomic, transcriptomic, and epigenomic) is crucial for precision medicine, which can improve patient diagnosis, prognosis, and treatment. Papillary renal cell carcinoma, which accounts for 15–20% of kidney cancers, is heterogeneous with histologic subtypes and variations in both disease progression and patient outcomes. KIRP has two main subtypes: Type 1, which is often multifocal, is characterized by papillae and tubular structures covered with small cells containing basophilic cytoplasm and small, uniform, oval nuclei (Delahunt and Eble 1997). Type 2, which is more heterogeneous, is characterized by papillae covered with large cells containing eosinophilic cytoplasm and large, spherical nuclei with prominent nucleoli (Delahunt and Eble 1997; Linehan et al. 2016). The Cancer Genome Atlas (TCGA) of KIRP provides molecular characterization at multiple levels, including copy number variations and transcriptomic data. Multi-omics data can be integrated to generate molecular insights, stratify patients and build predictive models. Previously, we have characterized

Keerthi S. Shetty, Aswin Jose and Mihir Bani contributed equally to this work.

✉ P. K. Vinod
vinod.pk@iiit.ac.in

Keerthi S. Shetty
keerthi.shetty@ihub-data.iiit.ac.in

Aswin Jose
aswin.jose@research.iiit.ac.in

Mihir Bani
mihir.bani@research.iiit.ac.in

[1] Center for Computational Natural Sciences and Bioinformatics, IIIT Hyderabad, Hyderabad 500032, India

the metabolic alternations of KIRP at the genome-scale level and developed predictive models based on transcriptomic and DNA methylation data to predict tumor stages of KIRP (Singh and Vinod 2020; Pandey et al. 2020; Singh et al. 2018). To identify KIRP subtypes from multi-omics data, we require an integrative method with the objective of clustering samples into disease subtypes.

The development of integrative methods for multi-omics data fusion is one of the major challenges in cancer informatics. Network Diffusion (ND) provides a powerful strategy for integrating multiple datasets by estimating the proximity between genes associated with one or more data types. Different single-omics studies have used ND approach to stratify specific cancer samples into relevant subtypes using a priori network (Hofree et al. 2013; Zhong et al. 2015; Fujimoto et al. 2016; Liu and Zhang 2015). Here, ND was applied to a binary somatic mutation matrix (genes-by-samples) that transforms it into a continuous activation profile. The resulting propagated matrix was clustered using a network-constrained non-negative matrix factorization (NMF) to find k patient groups. Gene-expression based analysis also essentially applies the ND strategy to a binary genes-by-samples matrix that represents the significantly expressed genes (Wu et al. 2015). The ND framework is also used for the integrative analysis of multi-omics data. The integration can be performed before, during, and after the ND step. It is also shown that propagating mutations with an irrelevant network may lead to erosion of pathway signals affecting the identification of subtypes. Therefore, the inference of a disease-specific network for ND may help in effective disease subtyping. Seifert and Beyer (2018) proposed a model with each gene expression changes as a linear combination of its own copy number and expression of other putative regulators for network inference followed by ND. He et al. (2017) showed improved stratification by constructing a cancer-specific co-expression network and integrating mutation data.

In this work, we propose a simple network diffusion strategy for integrative analysis of CNV and gene expression data to effectively stratify patients into clinically relevant subtypes and identify silent players in KIRP. A KIRP-specific network was constructed based on integrating a priori human protein-protein interaction (PPI) network PCNet (Huang et al. 2018a) and gene expression data of KIRP (referred to as GeneNet). The CNV data were projected onto the GeneNet and propagated on the network for clustering. Using GeneNet, we identified two robust subtypes that are biologically informative and have a strong association with clinical outcomes, such as patient survival, tumor stage and clinical subtypes of KIRP. We also compared the performance using the whole PCNet network and curated Cancer Reference Network (CRN) (Huang et al. 2018b; Forbes et al. 2017; Hanahan and Weinberg 2011; Iorio et al. 2016; Vogelstein et al. 2013). We observed that stratification using GeneNet outperformed the ones using PCNet and CRN. Further, we performed Singular Value Decomposition (SVD) analysis of cluster subtypes, which revealed CNVs associated with poor survival cluster. Interestingly, this approach also revealed the genes which are not directly affected by CNVs in KIRP (called silent players). Furthermore, the differential gene expression analysis between these two KIRP clusters showed pathways specific to poor survival and the relationship between gene expression and copy number variation data.

Methods

Data preprocessing

KIRP CNV data processed using GISTIC2 pipeline (Primary solid tumor) was downloaded from https://gdac.broadinstitute.org/. This comprises of focal amplifications and deletions with 5913 genes across 288 samples (Table S1). The copy number amplifications (gain) and deletions (loss) were treated equally as altered events. The processed data are represented as a binary matrix (0 or 1), where 1 means that the gene has been altered by genomic change. RNAseq gene expression (raw count) data of KIRP was downloaded from the GDC portal (https://portal.gdc.cancer.gov/).

The microarray data of KIRP (GEO accession number GSE2748) obtained using Affymetrix HGU133 Plus 2.0 arrays platform was used for validation. The microarray data were pre-processed using Robust Multi-array Average (RMA) method (Irizarry et al. 2003), which performs background correction, quantile normalization, and summarization. This dataset includes 22 and 12 samples in class 1 (better survival) and class 2 (poor survival), respectively.

Reconstruction of the gene expression network

KIRP specific gene expression network, GeneNet, was reconstructed based on differential expressed genes (DEGs) between tumor and tumor-matched normal samples of KIRP. We performed differential gene expression analysis by using DESeq2 (Love et al. 2014) to identify DEGs with adjusted p value (p-adj) cutoff of less than 0.05 and $|log_2(fold\ change)| > 1$. Interactions between DEGs were obtained based on the protein–protein interaction network PCNet to form KIRP-specific GeneNet, which contains 4320 genes and 166,644 interactions. We also considered PCNet (Huang et al. 2018a) (19,781 genes, 2,724,724 interactions) and curated Cancer Reference Network (CRN) (Huang et al. 2018b; Forbes et al. 2017; Hanahan and Weinberg 2011; Iorio et al. 2016; Vogelstein et al. 2013) (2291 genes, 204453 interactions) for this study. A summary of different networks and their overlap is shown in Fig. 1a. The

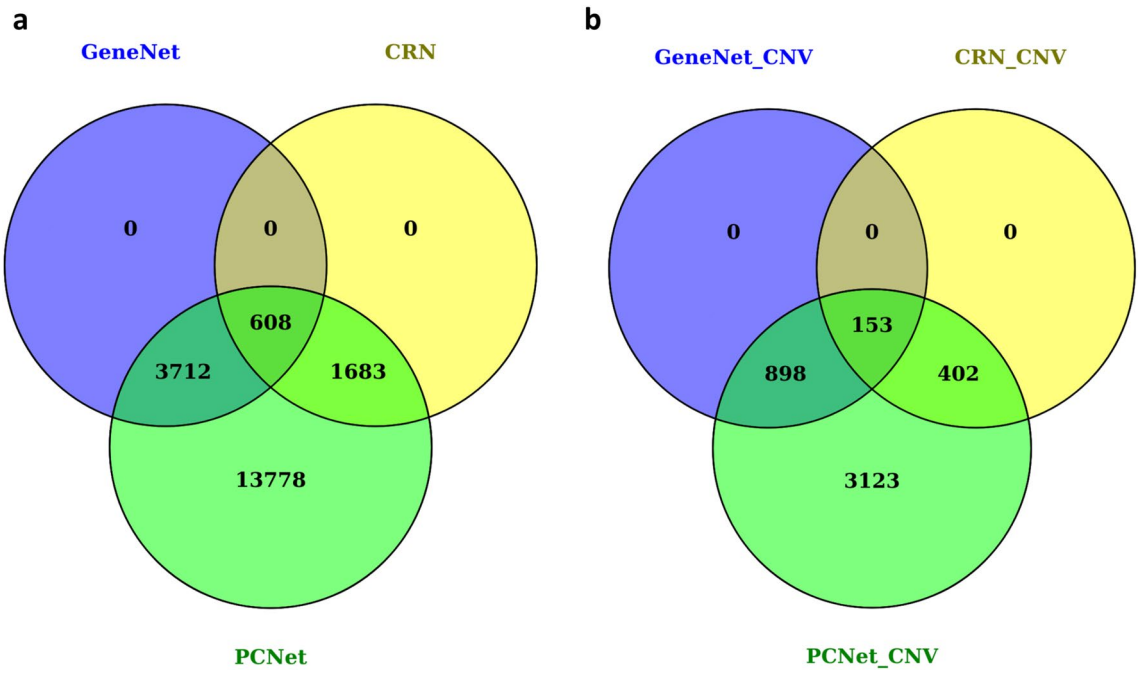

Fig. 1 **a** Overlap of different networks. **b** Overlap of networks considering only genes with CNVs

CNV binary matrix was mapped onto different networks: GeneNet, CRN, and PCNet. Figure 1b depicts the extent of overlap between network genes and CNV genes. The CNVs that are considered for the analysis vary depending on the network.

Network diffusion for integrative analysis of CNVs and gene expression

We adopted Network-Based Stratification (NBS) approach proposed by Huang et al. (2018b) for integrative analysis of CNVs and gene expression to stratify KIRP patients into clinically relevant subtypes. Figure 2 depicts the pipeline to identify subtypes of KIRP based on gene expression and CNVs. The input to the pipeline is a matrix of binary values describing tumor samples CNV data (i.e., patients × genes matrix) and the second input is the GeneNet network derived from gene expression data. A regularization graph for network-regularized non-negative matrix factorization (net-NMF) was constructed using the gene interaction network. A K-nearest neighbor (KNN) network was constructed from the gene interaction network matrix (Vandin et al. 2011), and graph Laplacian of this network was used in the non-negative matrix factorization.

Since the CNV matrix is sparse, a gene-by-gene matrix describing the influence of each gene on every other gene in the network was pre-computed by random-walk propagation. This propagation kernel was computed by independently propagating all genes in the gene interaction network. After the pre-computation of the regularization graph Laplacian and the network propagation kernel, the core steps of NBS were performed multiple times to produce multiple clusterings that were used in the consensus clustering step. The clustering was performed with the following steps:

1. Subsampling rows (samples/tumors) and columns (network genes) of the binary CNV matrix. 80% of rows and columns were subsampled.
2. Network propagation of the subsampled binary CNV matrix with coefficient (α) set to 0.7. After testing multiple values between 0.5 and 0.8, $\alpha = 0.7$ showed to produce robust results.
3. Quantile normalization of the network-smoothed mutation data to ensure that the smoothed profile for each patient follows the same distribution.
4. Non-negative Matrix Factorization (NMF) was used to decompose the matrix into k clusters. It decomposes the matrix into two lower rank non-negative matrices whose product can reasonably approximate the original matrix (Lee and Seung 1999). We applied network-regularized NMF to constrain NMF to respect the structure of the underlying gene interaction network as previously described (Cai et al. 2008). The objective is to minimize the following function:

$$min_{W,H>0} \|F - WH\|_F^2 + \lambda.trace\left(W^t LW\right) \quad (1)$$

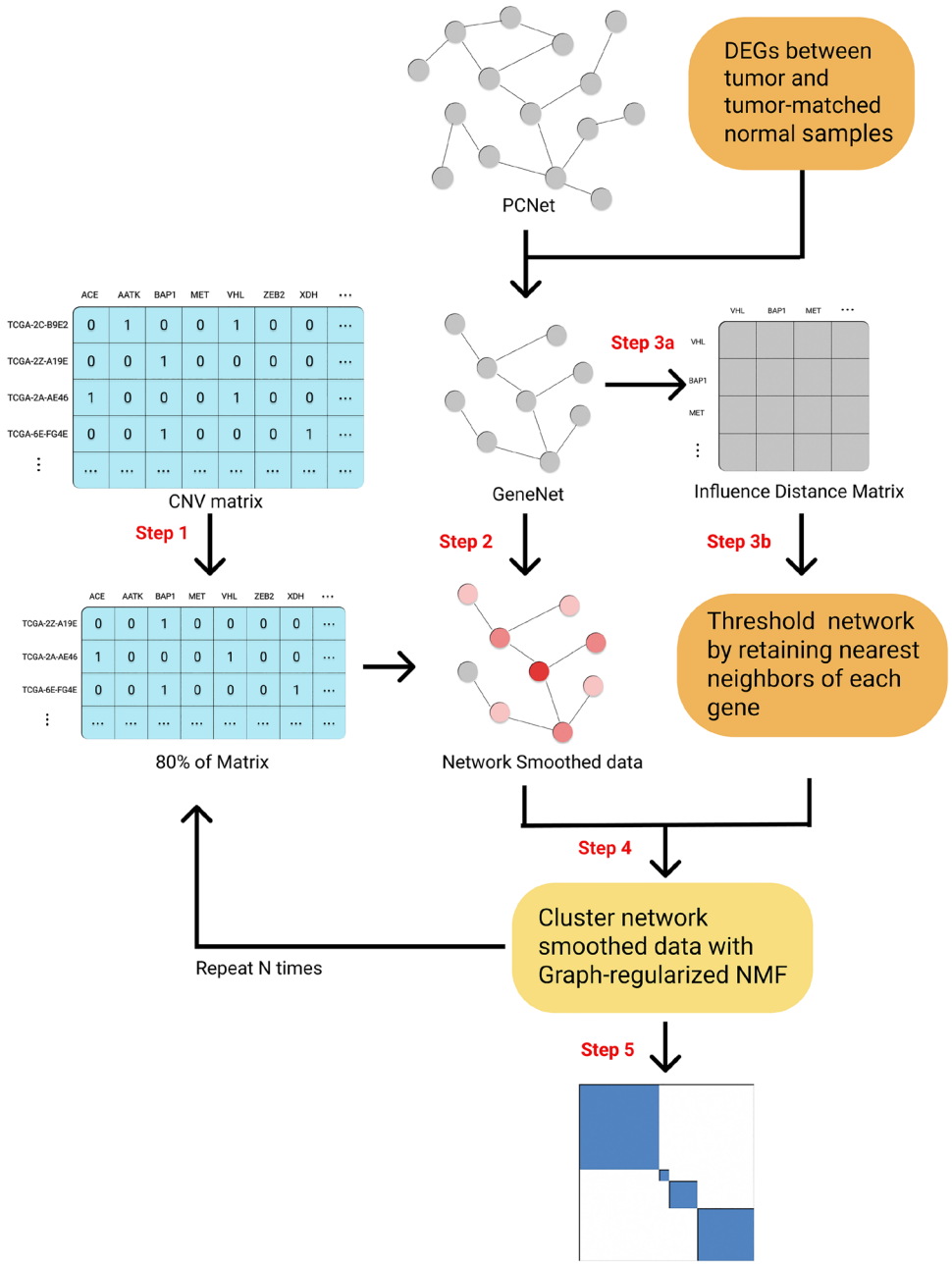

Fig. 2 Workflow of the network-based stratification method

where $\|.\|_F$ denotes the matrix Frobenius norm, W and H form a decomposition of the patients × genes matrix F, with entries in both W and H non-negative. W is a collection of basis vectors or 'metagenes', and H is the basis vector loadings. L is the graph Laplacian of a k-nearest-neighbor network. We set the number of nearest neighbours $k = 11$ as previously described by (Hofree et al. 2013). λ is the regularization parameter and the value was set to a default value of 200 (Cai et al. 2008). The iterative algorithm proposed by Cai et al. (2008) was used to find the solutions W and H. The iterations were run until the objective function converges.

Consensus clustering

Robust clustering was achieved by applying consensus clustering (Monti et al. 2003; The Cancer Genome Atlas Research Network 2011, 2012; Verhaak 2010) to produce the final subtypes. The randomly subsampled clustering was repeated 100 times for GeneNet, CRN and 1000 times for PCNet after testing for multiple values for convergence. The results of the multiple clustering make up the patient–patient similarity matrix. This matrix records the frequency of the sampling of each pair of patients and the rates at which the pairs were clustered in the same group amongst all

replicates. Hierarchical clustering with average linkage was performed using this similarity matrix.

The goodness of the cluster separation was assessed using the cophenetic correlation coefficient (ccc) value (Brunet et al. 2004). Clusters that exhibit clear patterns have ccc values over 0.99 (Zhong et al. 2015). The stability of the clusters was assessed by the Proportion of Ambiguous Clustering (PAC) (Senbabaoglu et al. 2014). If PAC is ≤ 30%, the clusters are stable. The ccc and PAC were calculated using R package 'NMF' and 'diceR' using *cophcor()* and *PAC()*, respectively. We also calculated Silhouette Width (SW) to assess whether samples are well clustered or not. Its value ranges from 1 to −1, with higher value indicating the sample is well clustered.

Characterizing the clinical, CNVs and gene expression differences in KIRP subtypes

To study the survival difference between the identified subtypes, Kaplan–Meier survival curves and Log-rank tests were performed for each identified subtype (clusters) to test the association of subtypes with survival. Fisher's exact tests were used to test the association of subtypes with tumor stage and KIRP subtypes. Further, genes that show significant CNV in each subtype were identified based on the binary CNV matrix and propagated CNV matrix. We applied Singular Value Decomposition (SVD), which helps to identify CNVs that show maximum variances. The propagated CNV matrix was used to identify genes that are not directly affected by CNVs. We also identified the DEGs between the clusters using DESeq2. The biological processes and pathway enrichment of DEGs were performed using Enrichr (Kuleshov et al. 2016).

Results

Network-based subtyping of KIRP

We studied the clustering patterns of KIRP patient samples based on CNV data by applying the network diffusion approach with KIRP specific GeneNet, CRN, and PCNet. The quality of clustering was assessed in two ways:

(i) The calculation of metrics like ccc, PAC and SW.
(ii) The significance of association between clusters and clinical information.

We tested for different cluster numbers, i.e., $K = 2$–6. Table 1 displays the results of significant associations between identified subtypes and survival in KIRP using different networks. The clustering map for each of the networks is shown in Fig. 3a and Fig. S2a and S3a. We observed similar clustering patterns with clusters $K = 2$ for three different networks with high ccc and SW values and low PAC values (Table 1). We also tested for different numbers of nearest neighbours ($k = 8$, 10, 12, 15, 25) and observed only small changes in the outcome. Similarly, varying the regularization parameter ($\lambda = 100$, 300, 500) did not impact the outcome (data not shown).

Table 1 Significant associations between identified subtypes (clusters) and survival in KIRP using different networks

Network	K	*p* value	SW	CCC	PAC (%)
GeneNet	2	0.002	0.95	0.98	15
CRN	2	0.006	0.95	0.98	17
PCNet	2	0.04	0.96	0.99	15

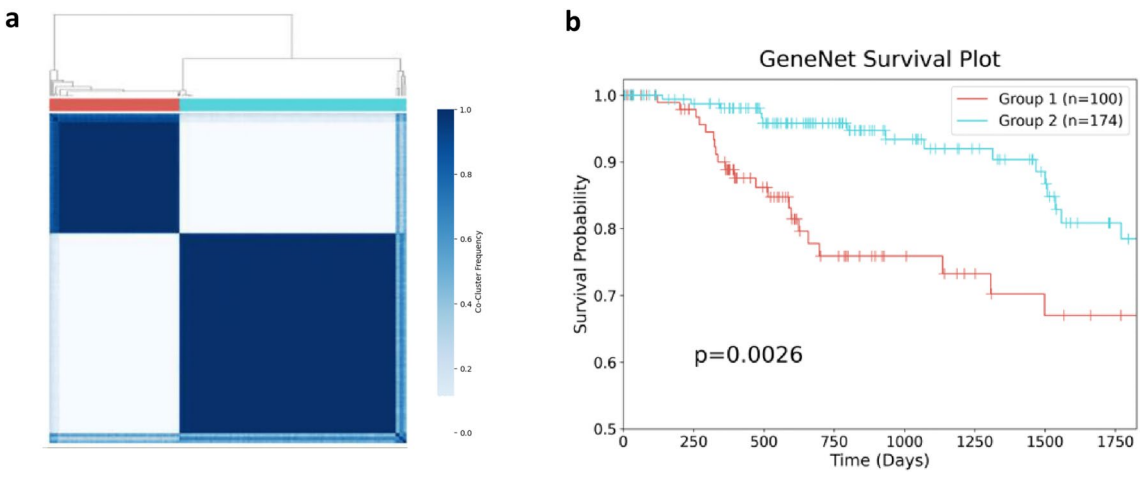

Fig. 3 **a** Clustering pattern of KIRP samples using GeneNet and CNV data. **b** KM plot showing survival difference between clusters that are obtained using GeneNet

Further, we found that clusters are not stratified properly if the clustering is performed without a network (Fig. S1).

Association of KIRP subtypes with clinical data

A significant association between two cluster subtypes and survival is observed with different networks (Fig. 3b, Fig. S2b, Fig. S3b). The stratification with GeneNet outperformed compared to CRN and PCNet. The cluster-wise distribution of KIRP subtypes and tumor stages shows that most samples in poor survival cluster (cluster C1) is associated with KIRP Type 2 subtype and consists of most number of stage 3 and 4 samples (late stages) (Fig. 4 and Fig. S4–S5). These observations demonstrate that KIRP network-based stratification reveals subtypes with strong clinical association and is consistently observed using three different networks.

Identifying subtype-specific altered genes

The CNV landscape of two clusters obtained using GeneNet shows that poor survival cluster C1 is predominantly associated with copy number loss compared to the better survival cluster C2, which has more amplification(Fig. 5). We also applied SVD to the network-smoothed matrix and binary matrix of CNV data to identify the top 100 genes in each cluster. Figure 6 shows the overlap of cluster-specific genes of KIRP obtained using these matrices. The analysis using the binary matrix resulted in unique CNVs in each cluster. The Cluster C1 was characterized by *1p36.31*, *14q24.2* deletions and C2 was characterized by *17q23.2*, *7q31.2*, *7p11.2* amplifications. Cluster C2 genes obtained using network-smoothed data showed significant overlap with candidates obtained based on the binary matrix. Genes of Cluster C2 are related to integrin-mediated cell adhesion and nucleotide metabolism. On the other hand, Cluster C1 candidate genes show only a few overlaps. We found 77 unique genes from network-smoothed data, 59 out of which map to CNV data,

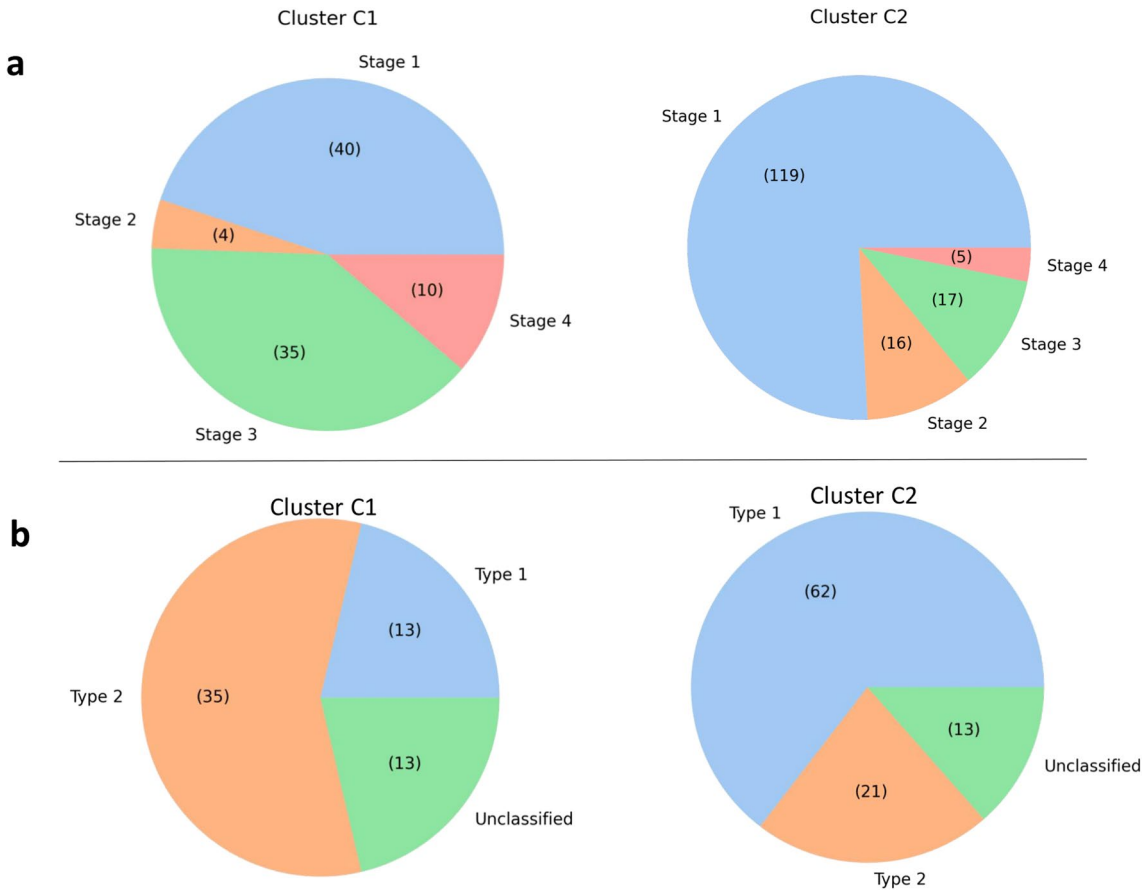

Fig. 4 **a** Association of Clusters C1 and C2 obtained using GeneNet with tumor stage. A significant difference in the distribution of stages between clusters is observed (p value=2.02e−08) by Fisher's exact test. **b** Association of Clusters C1 and C2 obtained using GeneNet with KIRP subtypes. A significant difference in the distribution of KIRP subtypes between clusters is observed (p value=2e−07) by Fisher's exact test

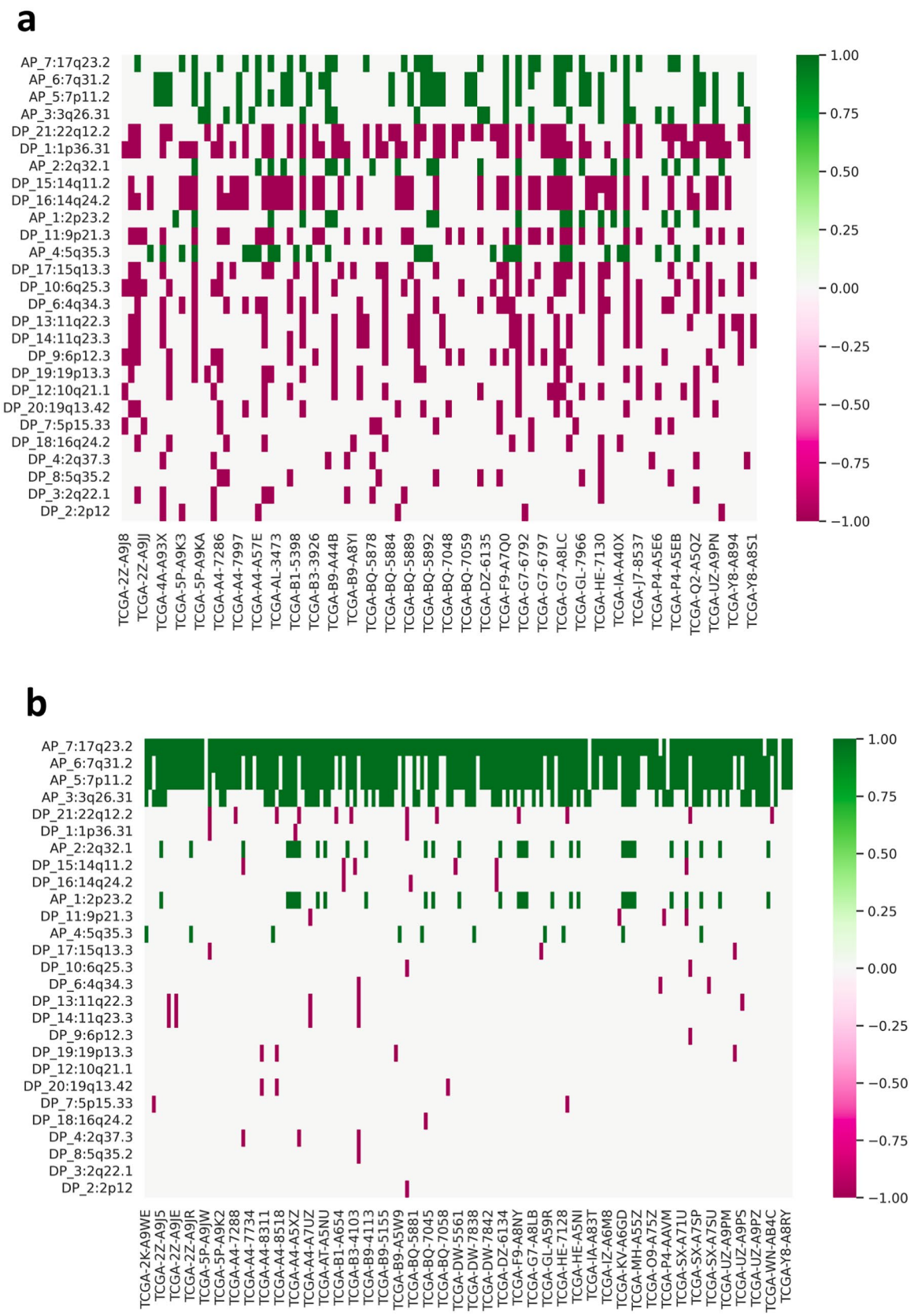

Fig. 5 The landscape of CNVs specific to **a** Cluster C1 and **b** Cluster C2

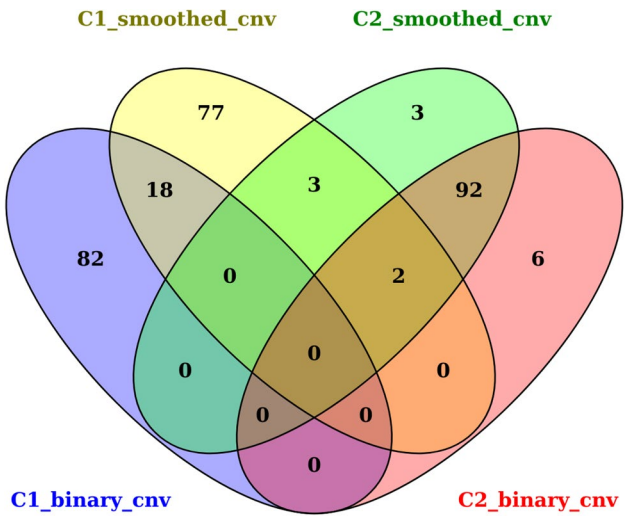

Fig. 6 Overlap of genes specific to the KIRP clusters C1 and C2 obtained from binary and smoothed CNV matrices using SVD approach. Cluster C1 specific CNVs: C1_smoothed_cnv and C1_binary_cnv. Cluster C2 specific CNVs: C2_smoothed_cnv and C2_binary_cnv

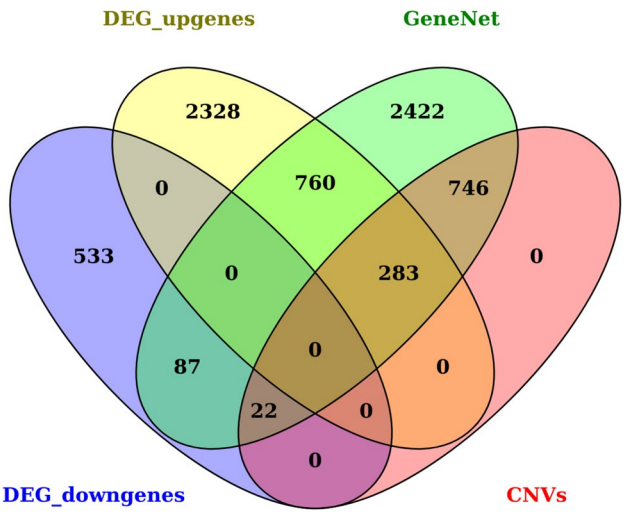

Fig. 7 Overlap of differentially expressed genes with GeneNet and CNVs

while the rest, 18, are not directly affected by CNVs. These 18 candidate genes emerge as a result of the network propagation, which helps identify silent players. The identified 18 genes are: INS, ALB, TCF4, CAMK2B, NTRK1, EGF, F2, NEUROD2, IL6, CXCL8, JAK3, PRL, ACTA2, CYP2C9, PDGFRA, E2F1, FGFR2, ITGA2. We observed that these 18 genes are significantly enriched for Pathways in Cancer, PI3K-Akt signaling pathway, Prostate Cancer, Regulation of actin cytoskeleton, JAK-STAT signaling pathway, Ras signaling pathway, Calcium signaling pathway, and HIF-1 signaling pathway.

Identifying differentially expressed genes between subtypes

We performed differential gene expression analysis to identify DEGs between the two clusters obtained using GeneNet. We found that the 642 genes are downregulated and 3371 genes are upregulated (adjusted p value < 0.05, $|log_2(fold\ change)| > 1$) in the poor survival cluster C1. This contrasts with CNV data which showed more deletions than amplifications in the poor survival group. The upregulated genes are associated with cytokine–cytokine receptor interaction, complement cascade, immunoglobulin receptor binding, signaling by VEGF, Extracellular matrix organization, Kinesins and cell cycle (adjusted p value< 0.05). The correlation between upregulated genes and survival in KIRP was also analyzed. The higher expression of genes in these pathways affects patient survival in KIRP. The upregulation of kinesins (CENPE, KIF18A, KIFC1, KIF4A, KIF2C, KIF11, KIF3C, KIF20A, KIF15, CENPA, CENPF, TOP2A, TPX2) and cell cycle (CDC20, CCNB2, CCNB1, BUB1B, CDC25C, NDC80, FOXM1, RRM2, MYBL2) genes in poor survival group may indicate an increase in genomic instability in KIRP. The expression of FLT1 and VEGFA in the poor survival group is high, suggesting angiogenesis may play a role in cancer progression. The expression of IL15RA and IL20RB in JAK-STAT signalling and TWIST1 involved in epithelial–mesenchymal transition (EMT) is also associated with poor survival (KIRP cluster C1).

The DEGs obtained between the clusters was further verified using an independent microarray data. We performed DEG analysis using GEO2R (Barrett et al. 2013) between two groups: class 1 and class 2, which show survival differences. We observed most candidate genes belonging to kinesins family and cell cycle are upregulated in the poor survival group, consistent with our clustering results.

Relationship between gene expression and CNV data in KIRP

Figure 7 represents the overlap of differentially expressed genes between two identified clusters with GeneNet and CNVs. 283 up-regulated genes in the poor survival group overlap with CNVs (Table S2). We observed that 76 genes are amplified, and their expression is up-regulated in the poor survival group (cluster C1). Table 2 displays the relationship between the up-regulated genes and CNVs. The 22 down-regulated genes in the poor survival group also overlap with CNVs (Table S3). We found that 17 genes are deleted and their expression down-regulated in the poor survival group. Table 3 displays the relationship between the down-regulated genes and CNVs. Further, it can be observed that most DEGs between clusters show less overlap with

Table 2 List of up-regulated genes in poor survival group overlapping with GeneNet and CNVs

Genes	Cytoband	log_2FC	Genes	Cytoband	log_2FC
CPLX2	5q35.3	5.514898	SST	3q26.31	3.585653
GPR87	3q26.31	3.140913	SUCNR1	3q26.31	2.743622
SHOX2	3q26.31	2.737492	TBX4	17q23.2	2.686825
TLX3	5q35.3	2.678660	TMEM207	3q26.31	2.509615
GIP	17q23.2	2.505385	RTP2	3q26.31	2.476133
C5orf46	5q35.3	2.436666	CLDN11	3q26.31	2.165281
TEX19	17q23.2	2.139518	CA4	17q23.2	2.121938
HRG	3q26.31	2.056223	FAT2	5q35.3	1.966245
FOXI1	5q35.3	1.939237	EPHB3	3q26.31	1.885303
HTR3C	3q26.31	1.856657	SNCB	5q35.3	1.843131
KNG1	3q26.31	1.822792	CACNG5	17q23.2	1.788486
SPINK13	5q35.3	1.785771	C3orf80	3q26.31	1.703210
CAMK2N2	3q26.31	1.694831	ADRA1B	5q35.3	1.674463
CDC25C	5q35.3	1.661267	NXPH3	17q23.2	1.640833
FABP6	5q35.3	1.616459	MECOM	3q26.31	1.608033
PPP2R2B	5q35.3	1.604318	GABRB2	5q35.3	1.596282
SLC16A5	17q23.2	1.550386	CAMK2A	5q35.3	1.547912
EBF1	5q35.3	1.539271	GABRP	5q35.3	1.526010
GCGR	17q23.2	1.522458	LRRC31	3q26.31	1.519712
NOTUM	17q23.2	1.489252	SCN4A	17q23.2	1.463961
NPTX1	17q23.2	1.461581	CACNA1G	17q23.2	1.458183
GFRA3	5q35.3	1.453373	PCDH1	5q35.3	1.447664
AADAC	3q26.31	1.444615	HOXB13	17q23.2	1.420125
KIF4B	5q35.3	1.406033	KIF20A	5q35.3	1.393249
SLC34A1	5q35.3	1.381899	SLC36A2	5q35.3	1.373802
GPX3	5q35.3	1.366102	IL12B	5q35.3	1.349399
SYNPO	5q35.3	1.331220	FLT4	5q35.3	1.294444
TP63	3q26.31	1.276950	KCNIP1	5q35.3	1.273421
PTTG1	5q35.3	1.269612	BIRC5	17q23.2	1.241669
RTP1	3q26.31	1.237080	NEURL1B	5q35.3	1.232511
WNT9B	17q23.2	1.222900	HMMR	5q35.3	1.175546
PRR11	17q23.2	1.156051	SSTR2	17q23.2	1.153727
NMUR2	5q35.3	1.144714	PYCR1	17q23.2	1.123311
CA10	17q23.2	1.116561	ZNF750	17q23.2	1.086681
BTNL9	5q35.3	1.074589	MSX2	5q35.3	1.072187
BRIP1	17q23.2	1.047996	SERPINI1	3q26.31	1.042734
SLC7A14	3q26.31	1.023263	KLHL3	5q35.3	1.019897
PFN3	5q35.3	1.004210	DRD1	5q35.3	1.004194

GeneNet. This suggests that our approach also captures the intra-group variation in cancer samples compared to the variation between normal and cancer samples.

Discussion

Accumulation of omics data by next-generation sequencing technologies provides the scope to identify cancer subtypes. In this work, we adopted a network-based approach to identify KIRP subtypes by integrating CNVs and gene expression data, which may help to cluster patients based on alterations in similar network regions. A KIRP-specific network (GeneNet) was constructed from gene expression data to integrate the copy number variation data. A network diffusion of mutated genes and clustering revealed two consensus clusters associated with clinical information: patient survival, tumor stages, and histological subtypes (Figs. 3 and 4). Identified clusters showed differences in the pattern of gene expression and CNVs, which can help to distinguish clusters.

Table 3 List of down-regulated genes in poor survival group overlapping with GeneNet and CNVs

Genes	Cytoband	log_2FC
RAD51AP2	2p12	−2.152967
CMA1	14q24.2	−2.017084
LRFN5	14q24.2	−1.951584
SCN2B	11q23.3	−1.946861
LHB	19q13.42	−1.891491
COX6B2	19q13.42	−1.739315
TTLL2	6q25.3	−1.645787
CTSG	14q24.2	−1.552605
ZDHHC22	14q24.2	−1.503521
CLCNKB	1p36.31	−1.494294
PRKCG	19q13.42	−1.433049
PLCH2	1p36.31	−1.202188
NLRP9	19q13.42	−1.181815
CPNE7	16q24.2	−1.142501
VSIG2	11q23.3	−1.015379
DNAAF1	16q24.2	−1.014154
ASB18	2q37.3	−1.006950

We observed that stratification using GeneNet outperformed the one using the whole protein–protein network (PCNet) or generic cancer network (CRN) (Table 1). Cluster C1 is predominantly stages 3 and 4 samples (Type 2 papillary subtype), and cluster C2 is stage 1 (Type 1 papillary subtype). Cluster C2 was characterized by gain/amplification, while cluster C1 was characterized by deletions resulting in poor survival. It can be noted that the binary matrix used for stratification does not distinguish between amplification or deletion. This suggests that the deletions (or amplification) in KIRP patients are mapped to similar regions in the network, thereby clustering them together. The gain/amplification in cluster C2 maps to chromosomes 7 and 17, consistent with changes in Type 1 papillary subtype (Linehan et al. 2016). The known MET1 amplification is associated with this cluster. On the other hand, the C1 cluster is majorly associated with the loss of *1p36* and *14q24* compared to the loss of *9p21* and *3p* reported in a few Type 2 papillary subtype samples. The candidate genes include tumor suppressor ERRFI1, which inhibits EGFR. It promotes apoptosis and positively correlates with survival in different cancers (Cui et al. 2021). CASZ1 is another tumor suppressor gene mapping to *1p36*, which is deleted in different cancers (Bhaskaran et al. 2018). Another tumor suppressor gene that is associated with cluster C1 is PRDM16, which controls HIF-targeted gene expression in kidney cancer and recruitment of immune cells in different cancers (Kundu et al. 2020; Li et al. 2022). TNFRSF14 is also deleted in cluster C1 and is known to control T-cell activation, and tumor-infiltrating leukocytes recruitment (Aubert et al. 2021).

Network propagation of mutations helped identify candidates that show maximum variance due to the presence of other CNVs in the neighborhood. The candidate genes are enriched for cancer pathways. This approach also helped to identify genes such as NTRK1, ALB, EGF, IL6, and CXCL8 that are not directly affected by CNVs but are in similar network regions of CNVs. NTRK1 is commonly mutated in different cancers AACR Project GENIE Consortium 2017, but in KIRP, it emerges due to the effects of other CNVs. ALB and EGF are reported as hub genes based on the gene expression pattern of KIRP patients (Xu et al. 2021). IL6 is involved in all aspects of tumorigenesis by regulating proliferation, apoptosis, metabolism, survival, angiogenesis, and metastasis (Kumari et al. 2016; Masjedi et al. 2018). The chemokines family gene CXCL8 plays a major role in cancer prognosis (Kohli et al. 2022).

Further, the differential gene expression analysis between these two KIRP clusters showed pathways specific to poor survival. We found that kinesins and cell cycle genes are differentially expressed between clusters, suggesting that these gene signatures can be utilized to differentiate them. This may prove beneficial in clinical settings. The expression profile of some of these genes is known to predict patient outcome in multiple cancers (Carter et al. 2006). We also found pathways related to immune signalling and complement cascade to be upregulated in cluster C1 along with genes associated with genomic instability. The upregulation of the immune signature suggests immune dysregulation in the poor survival group. Immune to stromal scores of Type 2 KIRP is significantly higher than Type 1 KIRP, and increased immune risk correlates with advanced stage of KIRP (Luo et al. 2021; Wang et al. 2019). Immune cells play a role in cancer-associated inflammation, which can be tumour promoting by controlling angiogenesis, proliferation, and invasiveness (Hanahan and Weinberg 2011; Gonzalez et al. 2018). This view is different from the canonical picture that the immune system helps eradicate tumors. The complement cascade is shown to have both tumor suppressor and promoter roles in cancers (Revel et al. 2020). A high expression of components of the classical complement cascade is shown in ccRCC. This can be the inflammatory mechanism activated by the cooperation between tumor cells and tumor-associated macrophages (Roumenina et al. 2019).

A relationship between gene expression and copy number variation was observed in the poor survival cluster. BIRC5, SERPINI1, WNT9A, C5orf46, and SPINK13 are amplified in the poor survival cluster. BIRC5 is a member of the apoptosis family, and it can promote cell proliferation (Frazzi 2021). SERPINI1 has been found to be expressed in different cancers and associated with EMT and the overall survival of patients (Matsuda et al. 2016). WNT9A plays a role in balancing the

progenitor cell expansion and differentiation during kidney development (Karner et al. 2011). C5orf46 is linked to renal cancer cell proliferation and migration and controls the immune microenvironment of renal cancer (Ma et al. 2022). MECOM (PRDM3) controls the process of Histone lysine methylation, which is involved in epigenetic control. This is in accordance with observations that renal cancer harbors frequent mutations in HMTs (Yan et al. 2019). PRDM3 and PRDM16 are the most mutated genes of the PRDM family in multiple cancers and are also linked to immune cell recruitment (Li et al. 2022). SPINK13 expression is associated with poor survival of KIRC patients (Liao et al. 2022).

In conclusion, the cancer-specific network was reconstructed based on gene expression data that incorporates the relevant biological knowledge for effective subtyping and better predictive performance of survival of KIRP patients. The sparse nature of CNVs data and heterogeneity in mutations can be overcome by integrating with the cancer-specific network. This approach also revealed the interplay of different biological processes and captured the relationship between gene expression and CNVs. This can be further expanded to integrate other omics data, such as DNA methylation. Deep learning-based network representation learning and clustering may improve the clinically relevant subtyping performance.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00438-023-02022-4.

Acknowledgements This work was supported by iHUB-Data, International Institute of Information Technology, Hyderabad, India.

Author contributions Conceptualization: PKV; methodology: KSS, AJ, MB; formal analysis and investigation: KSS, AJ, MB; writing—original draft preparation: KSS; writing—review and editing: KSS, AJ, MB, PKV; funding acquisition: PKV; supervision: PKV.

Funding Not applicable

Availability of supporting data All data generated or analysed during this study are included in this published article. Code is available in: https://github.com/Cancer-Research-Project/NBS-KIRP-CNV.

Declarations

Conflict of interest The authors have declared that no competing interests exist.

Ethical approval and consent to participate Not applicable.

Human and animal ethics Not applicable.

Consent for publication Not applicable.

References

AACR Project GENIE Consortium (2017) Aacr project genie: powering precision medicine through an international consortium. Cancer Discov 7:818–831

Aubert N, Brunel S, Olive D et al (2021) Blockade of hvem for prostate cancer immunotherapy in humanized mice. Cancers 13:3009

Barrett T, Wilhite SE, Ledoux P et al (2013) NCBI GEO: archive for functional genomics data sets-update. Nucleic Acids Res 41:D991–D995

Bhaskaran N, Liu Z, Saravanamuthu SS et al (2018) Identification of casz1 as a regulatory protein controlling t helper cell differentiation, inflammation, and immunity. Front Immunol 9:184

Brunet JP, Tamayo P, Golub TR et al (2004) Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci USA 101:4164–4169

Cai D, He X, Wu X, et al (2008) Non-negative matrix factorization on manifold. In: 8th IEEE Int Conf Data Mining, p 63–72

Carter S, Eklund A, Kohane I et al (2006) A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. Nat Genet 38:1043–1048

Cui M, Liu D, Xiong W et al (2021) Errfi1 induces apoptosis of hepatocellular carcinoma cells in response to tryptophan deficiency. Cell Death Discov 7:274

Delahunt B, Eble JN (1997) Papillary renal cell carcinoma: a clinicopathologic and immunohistochemical study of 105 tumors. Modern Pathol Off J U. S. Can Acad Pathol Inc 10:537–544

Forbes SA, Beare D, Boutselakis H et al (2017) Cosmic: somatic cancer genetics at high-resolution. Nucleic Acids Res 45:D777–D783

Frazzi R (2021) Birc3 and birc5: multi-faceted inhibitors in cancer. Cell Biosci 11(1):1–14

Fujimoto A, Furuta M, Totoki Y et al (2016) Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. Nat Genet 48:500–509

Gonzalez H, Hagerling C, Werb Z (2018) Roles of the immune system in cancer: from tumor initiation to metastatic progression. Genes Dev 32:1267–1284

Hanahan D, Weinberg RA (2011) Hallmarks of cancer: the next generation. Cell 144:646–674

He Z, Zhang J, Yuan X et al (2017) Network based stratification of major cancers by integrating somatic mutation and gene expression data. PLoS One 12(e0177):662

Hofree M, Shen JP, Carter H et al (2013) Network-based stratification of tumor mutations. Nat Methods 10:1108–1115

Huang JK, Carlin DE, Yu MK et al (2018a) Systematic evaluation of molecular networks for discovery of disease genes. Cell Syst 6:484–495

Huang JK, Jia T, Carlin DE et al (2018b) pynbs: a python implementation for network-based stratification of tumor mutations. Bioinformatics 34:2859–2861

Iorio F, Knijnenburg TA, Vis DJ et al (2016) A landscape of pharmacogenomic interactions in cancer. Cell 166:740–754

Irizarry RA, Hobbs B, Collin F et al (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4:249–264

Karner CM, Das A, Ma Z et al (2011) Canonical wnt9b signaling balances progenitor cell expansion and differentiation during kidney development. Development 138(7):1247–1257

Kohli K, Pillarisetty VG, Kim TS (2022) Key chemokines direct migration of immune cells in solid tumors. Cancer Gene Ther 29:10–21

Kuleshov MV, Jones MR, Rouillard AD et al (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 44:W90–W97

Kumari N, Dwarakanath BS, Das A et al (2016) Role of interleukin-6 in cancer progression and therapeutic resistance. Tumour biology J Int Soc Onco Dev Biol Med 37:11,553-11,572
Kundu A, Nam H, Shelar S et al (2020) Prdm16 suppresses hif-targeted gene expression in kidney cancer. J Exp Med 217(e20191):005
Lee D, Seung H (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401:788–791
Li M, Ren H, Zhang Y et al (2022) Mecom/prdm3 and prdm16 serve as prognostic-related biomarkers and are correlated with immune cell infiltration in lung adenocarcinoma. Front Oncol 12:772686
Liao C, Wang Q, An J et al (2022) Spinks in tumors: potential therapeutic targets. Front Oncol 12(833):741
Linehan WM, Spellman PT, Ricketts CJ et al (2016) Comprehensive molecular characterization of papillary renal-cell carcinoma. N Engl J Med 374:135–145
Liu Z, Zhang S (2015) Tumor characterization and stratification by integrated molecular profiles reveals essential pan-cancer features. BMC Genom 16:503
Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome Biol 15:550
Luo L, Zhou H, Su H (2021) Identification of 4-genes model in papillary renal cell tumor microenvironment based on comprehensive analysis. BMC Cancer 21:553
Ma M, Zhang Z, Liu Y et al (2022) Preliminary study on the role of the c5orf46 gene in renal cancer. Transl Oncol 21(101):442
Masjedi A, Hashemi V, Hojjat-Farsangi M et al (2018) The significant role of interleukin-6 and its signaling pathway in the immunopathogenesis and treatment of breast cancer. Biomed Pharmacother 108:1415–1424
Matsuda Y, Miura K, Yamane J et al (2016) Serpini1 regulates epithelial-mesenchymal transition in an orthotopic implantation model of colorectal cancer. Cancer Sci 107(5):619–628
Monti S, Tamayo P, Mesirov J et al (2003) Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Mach Learn 52:91–118
Pandey N, Lanke V, Vinod PK (2020) Network-based metabolic characterization of renal cell carcinoma. Sci Rep 10:5955
Revel M, Daugan MV, Sautés-Fridman C et al (2020) Complement system: promoter or suppressor of cancer progression? Antibodies (Basel) 9:57
Roumenina LT, Daugan MV, Noé R et al (2019) Tumor cells hijack macrophage-produced complement c1q to promote tumor growth. Cancer Immunol Res 7:1091–1105
Seifert M, Beyer A (2018) regnet: an r package for network-based propagation of gene expression alterations. Bioinformatics 34:308–311
Senbabaoglu Y, Michailidis G, Li J (2014) Critical limitations of consensus clustering in class discovery. Sci Rep 4:6207
Singh NP, Vinod PK (2020) Integrative analysis of dna methylation and gene expression in papillary renal cell carcinoma. Mol Genet Genom 295:807–824
Singh NP, Bapi RS, Vinod PK (2018) Machine learning models to predict the progression from early to late stages of papillary renal cell carcinoma. Comput Biol Med 100:92–99
The Cancer Genome Atlas Research Network (2011) Integrated genomic analyses of ovarian carcinoma. Nature 474:609–615
The Cancer Genome Atlas Research Network (2012) Comprehensive molecular portraits of human breast tumours. Nature 490:61–70
Vandin F, Upfal E, Raphael BJ (2011) Algorithms for detecting significantly mutated pathways in cancer. J Comput Biol J Comput Mol Cell Biol 18:507–522
Verhaak RG et al (2010) Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. Cancer Cell 17:98–110
Vogelstein B, Papadopoulos N, Velculescu VE et al (2013) Cancer genome landscapes. Science (New York, NY) 339:1546–1558
Wang Z, Song Q, Yang Z et al (2019) Construction of immune-related risk signature for renal papillary cell carcinoma. Cancer Med 8:289–304
Wu L, Liu Z, Xu J et al (2015) Netbags: a network-based clustering approach with gene signatures for cancer subtyping analysis. Biomark Med 9:1053–1065
Xu Y, Kong D, Li Z et al (2021) Screening and identification of key biomarkers of papillary renal cell carcinoma by bioinformatic analysis. PLoS One 16(e0254):868
Yan L, Zhang Y, Ding B et al (2019) Genetic alteration of histone lysine methyltransferases and their significance in renal cell carcinoma. PeerJ 7:e6396
Zhao L, Lee VHF, Ng MK et al (2019) Molecular subtyping of cancer: current status and moving toward clinical applications. Brief Bioinform 20:572–584
Zhong X, Yang H, Zhao S et al (2015) Network-based stratification analysis of 13 major cancer types using mutations in panels of cancer genes. BMC Genomics 16:S7