Published in partnership with the Systems Biology Institute

https://doi.org/10.1038/s41540-024-00445-2

Network modeling links kidney developmental programs and the cancer type-specificity of VHL mutations

Check for updates

Xiaobao Dong[1,4] ✉, Donglei Zhang[2,4], Xian Zhang[2], Yun Liu[3] & Yuanyuan Liu[1]

Elucidating the molecular dependencies behind the cancer-type specificity of driver mutations may reveal new therapeutic opportunities. We hypothesized that developmental programs would impact the transduction of oncogenic signaling activated by a driver mutation and shape its cancer-type specificity. Therefore, we designed a computational analysis framework by combining single-cell gene expression profiles during fetal organ development, latent factor discovery, and information theory-based differential network analysis to systematically identify transcription factors that selectively respond to driver mutations under the influence of organ-specific developmental programs. After applying this approach to VHL mutations, which are highly specific to clear cell renal cell carcinoma (ccRCC), we revealed important regulators downstream of VHL mutations in ccRCC and used their activities to cluster patients with ccRCC into three subtypes. This classification revealed a more significant difference in prognosis than the previous mRNA profile-based method and was validated in an independent cohort. Moreover, we found that EP300, a key epigenetic factor maintaining the regulatory network of the subtype with the worst prognosis, can be targeted by a small inhibitor, suggesting a potential treatment option for a subset of patients with ccRCC. This work demonstrated an intimate relationship between organ development and oncogenesis from the perspective of systems biology, and the method can be generalized to study the influence of other biological processes on cancer driver mutations.

The cancer-type specificity of driver mutations is a prevalent phenomenon observed in cancer genomic studies[1–3]. These cancer-type-specific mutations have high incidence rates in some cancers and are relatively rare in others, such as VHL mutations in ccRCC, BRAF mutations in cutaneous melanoma, and NPM1 mutations in myeloid leukemia. The presence of cancer-type specificity suggests that cells with distinct developmental origins differ in their responses to the insult of the same driver mutations; thus, the developmental programs that each cell inherits from its progenitor cells may be important in interpreting the biological effects of driver mutations[4,5].

Recently, several studies have confirmed the important impact of developmental programs in shaping the specificity of cancer driver mutations. For example, Baggiolini et al. reported that the developmental programs regulated by the chromatin-modifying gene ATAD2 and the transcription factor (TF) SOX10 are responsible for the activation of cell proliferation-associated pathways by BRAF V600E, which explains why the BRAF V600E mutation specifically transforms normal cells only in melanocyte progenitor cells[6]. Patel et al. used genetic screening techniques to identify PAX8, a transcription factor associated with kidney development, which mediates oncogenic signaling downstream of VHL in ccRCC[7]. Weiss et al. showed that transcriptional programs involved in limb development are required for the CRKL mutation-driven growth of acral melanoma[8]. These discoveries not only deepen our understanding of the mechanisms of cancer development but also provide biomarkers and therapeutic targets for cancer prevention and treatment. However, as demonstrated in these studies, to study the cancer-type specificity of driver mutations, researchers need to establish cellular or animal models harboring the same genetic variants at different developmental stages or in different types of tissues and conduct in-depth comparative studies to distinguish the effects of the same driver events in multiple cellular contexts. These complex experimental approaches tend to be expensive and time-consuming.

[1]Department of Genetics, School of Basic Medical Sciences, Tianjin Medical University, Tianjin, China. [2]Department of Hematology, Zhongnan Hospital of Wuhan University, Wuhan, Hubei, China. [3]Department of Pediatric Oncology, Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Center for Cancer, Tianjin's Clinical Research Center for Cancer, Key Laboratory of Cancer Prevention and Therapy, Tianjin, China. [4]These authors contributed equally: Xiaobao Dong, Donglei Zhang. ✉e-mail: dongxiaobao@tmu.edu.cn

Rapidly accumulating single-cell RNA sequencing data and analytical approaches in network biology provide new opportunities for systematic analysis of the relationships between driver mutations and developmental programs. For example, single-cell atlases of mouse and human organ development have been published[9,10]. The single-cell RNA sequencing (scRNA-seq) data generated by these studies provide unprecedented resolution and coverage for studying transcriptional programs during organogenesis, which involve more dynamic changes in cell differentiation compared with data from adult tissues that are dominated by differentiated cells. Moiso et al. used single-cell transcriptome data from normal mouse organogenesis to construct an artificial neural network model that can classify the origins of cancers with extreme accuracy[11]. Their work also demonstrated that transcriptional programs during organogenesis are equally conserved even in cells that have undergone malignant transformation. By statistically factorizing these single-cell transcriptomic data[12–14], it will be possible to systematically identify developmentally relevant transcriptional programs and thus study their activity in different types of cancers. This factorization-based technique has been successfully used to delineate the mutational process in cancer[15], to study the factors that influence the responses of patients with cancer to immunotherapy[16] and to analyze the metastatic process of cancer[17]. As another example, the network biology approach provides a flexible and powerful analytical tool for integrating transcriptional programs and genetic variation data by abstracting the complex functional relationships between genes into interconnected nodes[18]. This approach has been used to rank driver mutations[19], interpret genetic risk loci of cancer[20], localize biological pathways perturbed by cancer mutations[21], and predict patient response to anticancer therapy[22]. Differential network analysis techniques have also been developed to specifically study changes in gene interactions in cells under different conditions[23]. Therefore, a network biology approach may be a suitable framework for systematically unraveling important downstream regulators in carcinogenesis by associating transcriptional programs extracted from organ development data with the specificity of cancer driver mutations.

In this study, we developed a computational pipeline based on network biology and scRNA-seq data to analyze the links between developmental programs and cancer-type-specific driver mutations. We studied the impact of transcriptional programs during kidney development on the oncogenic effects of renal cancer-specific VHL mutations. We show that our approach can identify both previously known and novel important developmental regulators that influence the effects of VHL mutations. We utilized these regulators to create a more accurate prognostic risk stratification model for patients with cancer and to uncover new potential therapeutic targets for this highly aggressive subtype.

Results

Overview of the methods

Currently, studies of organ development and cancer genomics are conducted relatively independently of each other, and direct comparative analysis of data generated in the two fields presents significant challenges due to the large differences in underlying experimental assumptions and subjects. Inspired by the work of Tamayo et al.[24] and Stein-O'Brien et al.[13], we thought that it would be useful to decompose transcriptomic changes during development into variations in the activities of a much smaller number of biologically significant latent factors, which in turn would allow us to transfer the data between the two fields by analyzing only the changes in the activity of these factors in cancer cells. The main steps of our approach are shown in Fig. 1 (see "Methods" for details).

First, for a cancer-type-specific driver gene, we selected a protein-protein interaction (PPI) subnetwork centered on it to represent its functional context (Fig. 1A). Previous studies have shown that functional genes relevant to cancer biology are mainly genes that are one or two steps away from canonical cancer genes[25]. In our study we selected the driver gene and the genes that are within two steps away of it in the network as the driver gene-centered subnetwork. By selecting this subnetwork, we were able to focus our analysis on the biological pathways most relevant to the driver gene to be studied[26].

Second, we used non-negative matrix factorization (NMF)[27] from scRNA-seq data of fetal organ development to extract low-dimensional latent factors that explain changes in the gene expression profiles of subnetwork members during development (Fig. 1B). These factors, which are much smaller in number than the number of genes in the subnetwork, represent the major transcriptional programs associated with driver genes during development (also referred to as developmental programs in this work). For each driver gene and corresponding cancer type, we selected the organ of cancer origin for analysis and focused on scRNA-seq data from cells of origin[28] of this cancer type (see "Methods" for more details). We then projected the pancancer transcriptome data of TCGA into the low-dimensional space constituted by these transcriptional programs to obtain the activity profile of each factor in each cancer type. This allowed us to identify cancer-type-specific programs associated with the specifically mutated driver gene for subsequent analysis.

Finally, we updated the weights of the PPI network using the identified developmental program and analyzed the differences in communication efficiency between the cancer-type-specific driver gene and downstream TFs before and after the update. Here, we adopted the personalized PageRank algorithm[29] to update the network weights and search information[30] to measure the efficiency of communication between genes. Inferring the functional relevance of genes through network proximity has been shown to be a very effective strategy[31,32]. We postulated that genes that are neighboring in the PPI network with genes that are highly weighted in the developmental program are also functionally related with the program. The personalized PageRank algorithm used highly weighted genes in a developmental program as seeds to perform random walk on the PPI network and assign higher weights to genes neighboring these seeds. Search information (S) is an information entropy-based model that quantifies the amount of information (in bits) required for a signal to reach the target node after it is emitted from the source node when it travels along the shortest path from the source to the target node. The lower the search information between the network nodes is, the easier the signaling transmission is. The difference in search information (ΔS) between weighted and unweighted PPI networks was used to detect the impact of the developmental program. By comparing the observed ΔS with those from weighted networks updated by random seeds, we calculated the P value of the change in communication efficiency and obtained significant TFs. For convenience, we refer to these genes as developmental program-sensitive TFs (dsTFs) in this manuscript. These dsTFs are important transcriptional regulator candidates downstream of cancer-type-specific driver mutations and provide the basis for interpreting the oncogenic effects of driver mutations.

A ccRCC-specific developmental program surrounding VHL

To illustrate the validity of our approach, we applied our method to ccRCC-specific VHL mutations, which can be found in 90% of ccRCCs[33]. The inactivation of VHL, which is a key protein in the cellular sensing of oxygen, can cause the accumulation of the hypoxia-inducible factors HIF1α and HIF2α, which activate the expression of downstream cell proliferation-related genes[34]. However, though activation of the hypoxia pathway is widespread in many cancers, why VHL mutations are only specifically found in ccRCC is not yet fully understood. By studying this specific mutation type, we hope that our method can identify some previously known regulators associated with it to demonstrate the validity of our method and also aim to systematically identify more new regulators and gain new insights into the treatment of ccRCC.

We used scRNA-seq data generated from the fetal organ developmental atlas[9] to analyze changes in the expression profiles of 768 genes that are members of the VHL-centered subnetwork during fetal kidney development. These expression data originated from 28 fetuses ranging from 72 to 129 days in postconceptual age. Using the results of principal component analysis (PCA) as a guide (Fig. 2A) and after balancing the tradeoff between the stability and accuracy of the NMF results (Fig. 2B), we ultimately extracted 11

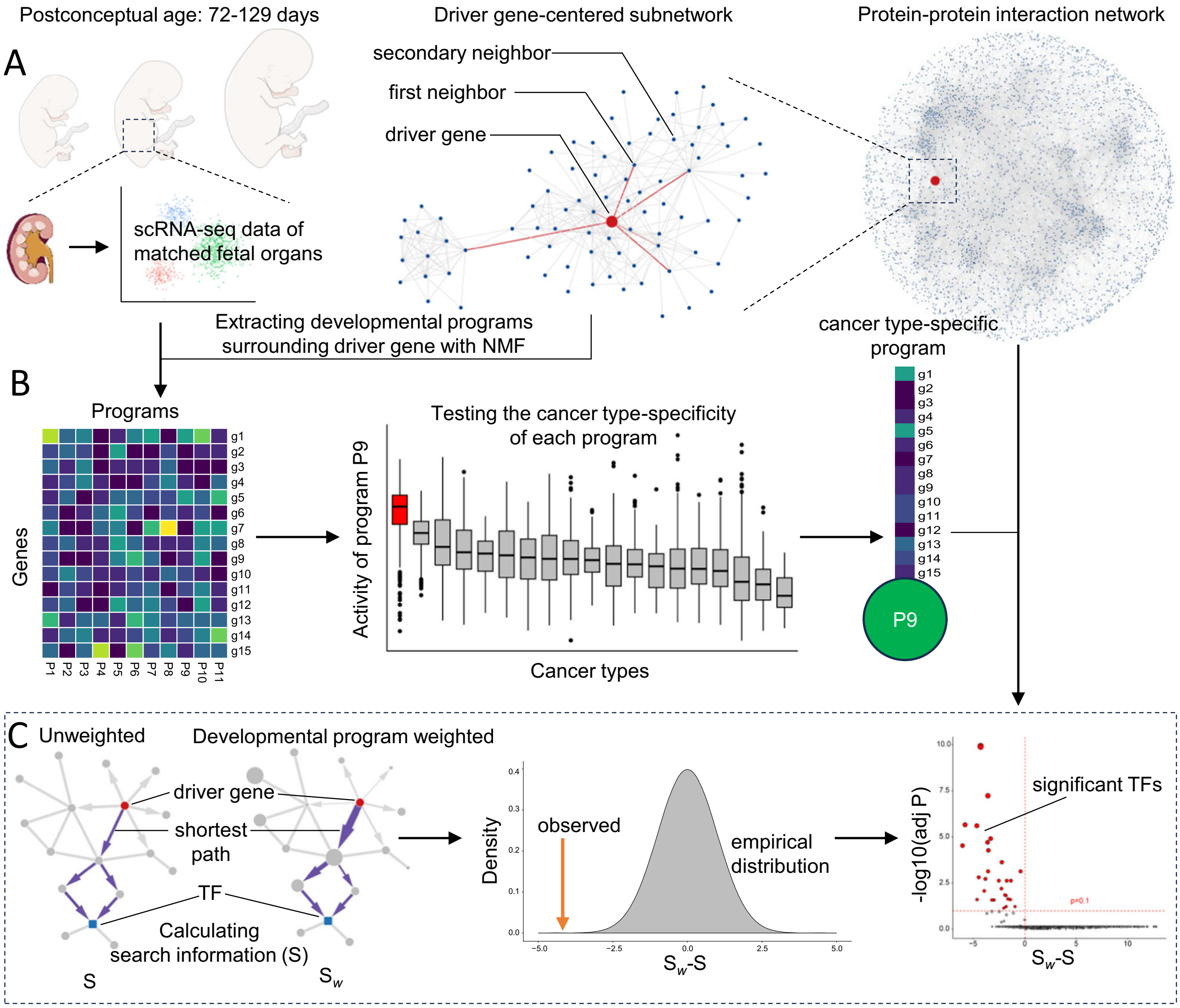
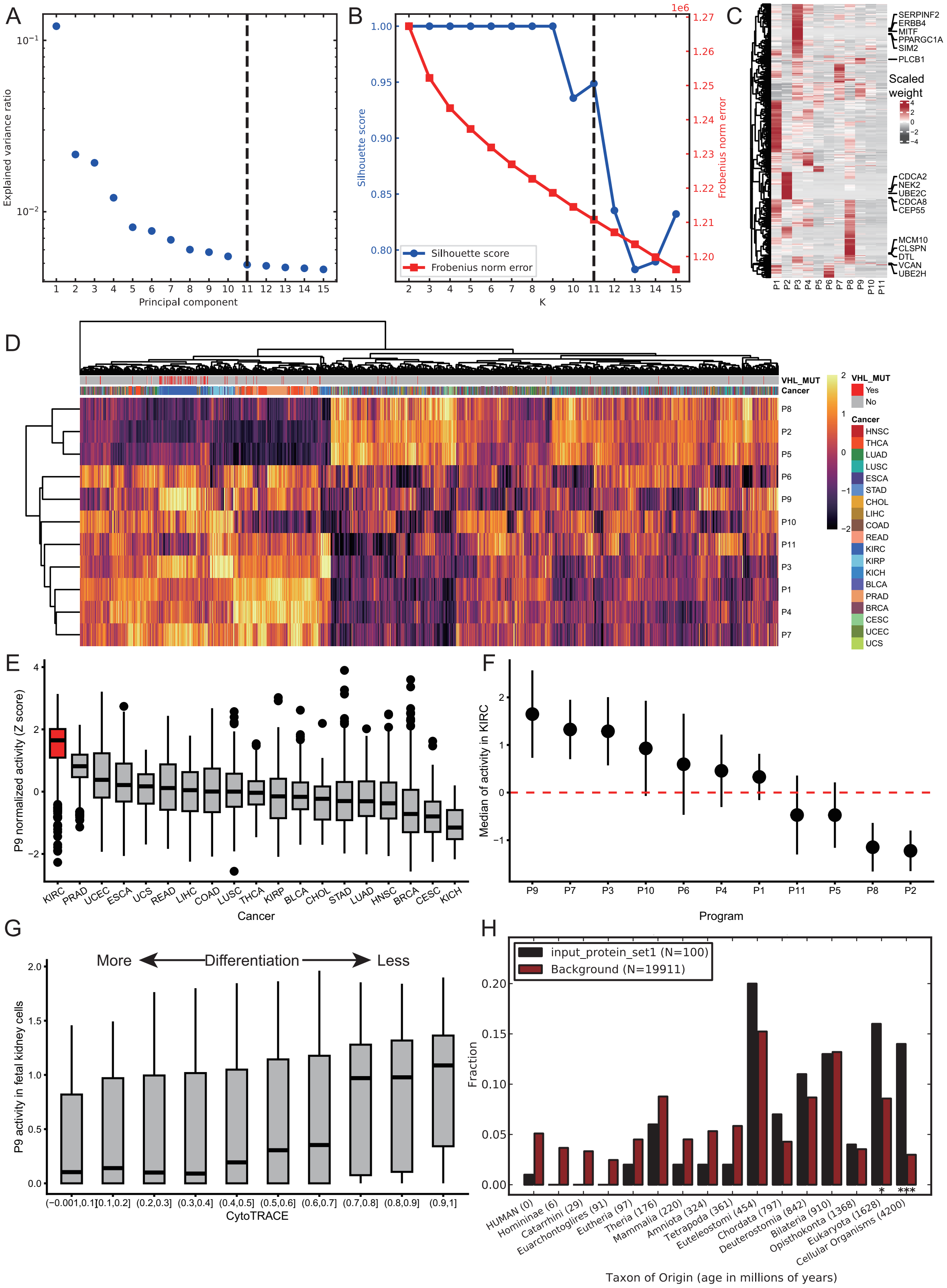

Postconceptual age: 72-129 days
Driver gene-centered subnetwork
Protein-protein interaction network
A
secondary neighbor
first neighbor
driver gene
scRNA-seq data of matched fetal organs
Extracting developmental programs surrounding driver gene with NMF
cancer type-specific program
B
Programs
Genes
g1
g2
g3
g4
g5
g6
g7
g8
g9
g10
g11
g12
g13
g14
g15
P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
Testing the cancer type-specificity of each program
Activity of program P9
Cancer types
P9
C
Unweighted
Developmental program weighted
driver gene
shortest path
TF
Calculating search information (S)
S
S_w
Density
observed
empirical distribution
S_w-S
-log10(adj P)
significant TFs
p=0.1
S_w-S

Fig. 1 | Overview of network analysis workflow. A Driver gene-centered subnetwork and fetal scRNA-seq data were integrated to extract developmental programs surrounding cancer-specific driver genes. NMF is non-negative matrix factorization. **B** The activities of developmental programs in 19 cancer types were calculated via projection analysis of TCGA RNA-seq data. A specifically activated gene (P9) was identified, and its top-weighted genes were used as seeds to update the PPI network. **C** The network communication efficiencies between driver genes and downstream TFs were measured by search information (*S*) along the shortest paths. For each driver gene–TF pair, the difference in *S* before (*S*) and (S_W) after considering developmental programs was calculated, and its statistical significance was determined by an empirical test. The observed difference between S_w and *S* were marked with orange arrow. This value is compared with the random distribution of$S_W - S$(gray), in which S_W is generated from a network with shuffled gene weight assignment. The images of the fetus and kidney in (**A**) were downloaded from Servier Medical Art (https://smart.servier.com/) that is licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/deed.en).

developmental programs (P1–P11) from ~90,000 kidney cells (see Methods for details). Some representative highly weighted genes in each developmental program are shown in Fig. 2C (Supplementary Table 1). These genes can be interpreted as having relatively high expression levels in the corresponding programs. Gene Ontology enrichment analysis (Supplementary Fig. 1) revealed that some of these genes are involved in cell differentiation and apoptosis (P6, P7, and P10), while others are involved in biological pathways that play important roles in cell proliferation (P2 and P4).

To identify ccRCC-specific programs, we analyzed program activities in different cancer types. Batch effect corrected RNA-seq data from 6142 The Cancer Genome Atlas (TCGA) patient samples across 19 cancer types[35,36], including data from ccRCC and two other kidney cancer types (papillary renal carcinoma and chromophobe renal carcinoma) that are not typically associated with VHL mutations, were obtained[33]. Notably, ccRCC, papillary renal carcinoma and chromophobe renal carcinoma were named KIRC, KIRP, and KICH, respectively, in TCGA. We projected gene expression profiles from patients with cancer into a low-dimensional space consisting of 11 developmental programs and created a pancancer landscape of program activities (Fig. 2D). For each of the three kidney cancer types, corresponding samples could be clustered together, a trend that was not observed in the other types of cancer. More importantly, samples from the three kidney cancer types were clearly separated from each other. This pattern of separation of distinct cancer subtypes with the same organ was further validated in lung cancers (Supplementary Fig. 2), suggesting that the developmental programs extracted from single-cell data sensitively reflected the subtle differences between cancers from the same organ. Many evidence have shown that the cells of origin of different cancer types within the same organ are different[28]. It may be the biological basis of this result. In addition, we also observed that the PRAD and THCA samples show a clear separation compared to the other cancer types. According to the heatmap, this separation shown by the two cancer types can be mainly attributed to the higher activity levels of the program P1 and P4 in them. These developmental programs may be shared by kidney and the organs of origin of these two cancer types.



By further examining individual developmental programs, we found that the activity of developmental program P9 was significantly greater in ccRCC than in other cancer types and that it also had the highest relative activity in ccRCC among all 11 developmental programs (Fig. 2E, F and Supplementary Fig. 3). In addition, its activation was significantly greater in VHL-mutated ccRCC samples than in non-VHL-mutated ccRCC samples (Supplementary Fig. 3). Additional analysis of gene expression profiles from patient-derived tumor xenografts[37] validated our observations (Supplementary Fig. 4). These results suggested that P9 is closely related to the selectivity of VHL mutations in ccRCC; therefore, in subsequent analyses, we focused on the relationship between P9 and VHL mutations.

Fig. 2 | Identification of the VHL mutation-associated kidney developmental program. A, **B** The optimal number (*K*) of developmental programs was determined through PCA analysis and stability analysis. The first 15 principal components extracted from gene weight matrix from NMF are shown. The dashed line marks the position corresponding to the optimal K. **C** Heatmap of gene weights in each program. Representative genes for each program are labeled. **D** Heatmap of the developmental program activities in more than 6000 cancer samples. **E** The distribution of P9 activity in 19 cancer types. **F** The distribution of activities of 11 programs in 19 patients with ccRCC (named KIRC in TCGA). **G** The correlation between P9 activity and the differential potential estimated by CytoTRACE using fetal scRNA-seq data. **H** The evolutionary age distribution of the top 100 weighted genes in P9. The significant difference between input genes and background genes in each group were marked with asterisk (Fisher's exact test). $^*P < 0.05$, $^{***}P < 0.001$. HNSC head and neck cancer, THCA thyroid cancer, LUAD lung adenocarcinoma, LUSC lung squamous cell carcinoma, ESCA esophageal carcinoma, STAD stomach adenocarcinoma, CHOL cholangiocarcinoma, LIHC liver hepatocellular carcinoma, COAD colon adenocarcinoma, READ rectum adenocarcinoma, KIRC kidney renal clear cell carcinoma, KIRP kidney renal papillary cell carcinoma, KICH kidney chromophobe, BLCA bladder urothelial carcinoma, PRAD prostate adenocarcinoma, BRCA breast invasive carcinoma, CESC cervical squamous cell carcinoma and endocervical adenocarcinoma, UCEC uterine corpus endometrial carcinoma, UCS uterine carcinosarcoma.

By utilizing the cell-type markers provided by the PanglaoDB database[38] (Supplementary Table 2), we found that several highly weighted genes in P9 are expressed in stem/progenitor cells, such as pluripotent stem cells (RFC3, CENPF, ATAD5, SALL4, CHEK2 and DNA2), embryonic stem cells (SALL4, PCGF2, KIT and LEF1) and kidney progenitor cells (WT1, MAP1B and PCGF2). Consistent with this observation, the analysis results of CytoTRACE[39], a single-cell developmental potential analysis tool, also showed that kidney cells with greater P9 activity had greater differentiation potential (Fig. 2G). KEGG-based pathway analysis[40] (Supplementary Table 3) revealed that P9-related genes are involved mainly in ribosomes ($P = 1.61 \times 10^{-14}$), pathways involved in cancer ($P = 2.65 \times 10^{-5}$) and the Wnt signaling pathway ($P = 8.96 \times 10^{-5}$). After querying the eukaryote protein origin database ProteinHistorian[41] with the top 100 highly weighted genes in P9, we found that these genes were significantly older than the average age of all the genes in the human genome (Mann–Whitney *U* test, $P = 2.19 \times 10^{-10}$), with a significant proportion having appeared at the time of eukaryotest (Fig. 2H). Although cancer-type specificity reflects a trait of multicellular organisms, our data suggest that these genes inherited from the ancient single-cell era may still play an important role in this phenomenon.

Network communication analysis identifies perturbated transcription factors downstream of VHL

To identify transcriptional regulators downstream of the VHL signaling pathway, which are affected by the P9 developmental program, we updated the weights of the PPI network with the P9 program and calculated the changes in the efficiency of communication between VHL and all TFs. Twenty-nine TFs, including the subunit (ARNT) of the VHL mutation-activated hypoxia-inducible factor and WT1, an important regulator of kidney development, exhibited significant changes (adjusted $P < 0.1$, Fig. 3A and Supplementary Table 4). In addition, all 29 of these dsTFs had negative ΔS; in other words, the P9 programs improved the communication efficiencies for all of them in our model. For the other 1011 nonsignificant TFs, both positive and negative values were observed.

However, we also noted that our approach missed multiple TFs known to be associated with VHL specificity, such as EPAS1 and PAX8. Therefore, these 29 dsTFs were further expanded using GeneMANIA[42], a network integration algorithm for predicting genes with functions similar to those of the input gene list. Finally, we identified 49 dsTFs that may be related to VHL mutations (Fig. 3B and Supplementary Table 5). We successfully retrieved PAX8, HIF1A (HIF1α), EPAS1(HIF2α) and BHLHE41 among the 20 expanded TFs (Supplementary Table 6). Their importance for effects of VHL mutations and the pathophysiology of ccRCC has been well documented. HIF1α and HIF2α are the core effectors of the VHL mutation-induced hypoxia response[33]. And PAX8 is a recently identified lineage TF that directly affects the target selection of HIFα[7]. The difference of BHLHE41 in population has been associated with variations in the risk of renal cell carcinoma, and BHLHE41 genetic variants are selected during hypoxic adaptation[43,44]. Although the links of other genes, such as PAX2, to VHL are unclear, they are closely connected to known TFs in functional networks, as are their important roles in kidney development and cancer[45], supporting their function in VHL-mutated ccRCC. To pursue the connection between these 20 expanded TFs and the program P9, we compared them with the top 100 genes with the highest weights in P9, and found that MEF2C was shared. In addition, the expanded dsTF MAML1 is homologous to MAML2, the second highest weighted gene in the P9, and both are involved in the NOTCH signaling pathway. These known TFs combined with other new TFs identified here implicate an uncharacterized regulatory network downstream of VHL in ccRCC.

Based on the above results, we analyzed the regulatory network composed of these dsTFs using gene regulatory relationships collected from the literature by NetAct[46]. This network included 672 genes and 929 interactions involving 38 dsTFs (including 22 dsTFs directly derived from network communication analysis and 16 expanded TFs, Supplementary Table 7) and their targets (Supplementary Fig. 5A). Enrichment analysis with hallmark gene sets[47] showed that genes commonly regulated by at least two dsTFs were involved in apoptosis, hypoxia and TNF-alpha signaling (Fig. 3C), which was consistent with previously reported effects of VHL mutations in ccRCC. According to the KEGG pathway analysis, many of these targets were members of pathways involved in cancer, including the renal cell carcinoma pathway. Eight genes were regulated by five or more dsTFs, including JUN, EP300, CCND1, AR, VEGFA, TP53, CREBBP, and BCL2. Among these factors, the selective expression of CCND1 in ccRCC has been suggested to be responsible for the cancer-type specificity of VHL mutations[48]. There was also mutual regulation between these dsTFs. We observed that 63% (24/38) of the dsTFs could directly regulate other dsTFs (Supplementary Fig. 5B). These genes form a core regulatory network centered on EGR, and the three main branches of the network correspond to TFs in the NOTCH signaling pathway, represented by MAML1 and RBPJ; TFs in the HIF signaling pathway, represented by HIF1A, EPAS1, and ARNT; and TFs in the renal development pathway, represented by PAX2, PAX8, and WT1.

To determine whether the states of dsTFs were able to discriminate between ccRCC and other types of cancer, we conducted a single-sample gene set analysis[49] using the expression levels of dsTF target genes as a proxy to infer the activation of dsTF activities in clinical cancer samples. Only 17 dsTFs having at least 10 tagets were considered for stability of the result. PCA analysis showed that these dsTF activity profiles were clearly different between ccRCC, normal kidney cells, and other cancers (Fig. 3E), and there was clear separation between ccRCC, normal kidney cells, and other cancers (i.e., breast and colon cancers). Moreover, this separation did not exist between other cancers, suggesting that the state of this regulatory network was specific to ccRCC. We analyzed the contribution of individual dsTFs to this discrepancy revealed by PCA analysis (Fig. 3F and Supplementary Fig. 6), and we found that although both are members of the HIF1A family, EPAS1 activity was stronger in ccRCC, and conversely, HIF1A activity was stronger in other types of cancer, supporting that HIF1A is a kidney cancer suppressor gene[50]. We also found that PAX8 activity was attenuated in ccRCC cells compared with normal kidney cells, implying that the functional state of PAX8 was not the same even though this lineage-expressed TF can promote the activation of some key downstream targets of VHL mutations and promote ccRCC according to previous work. These results provide insight for further study and understanding of the biological effects of VHL in ccRCC and illustrate the value of our method.

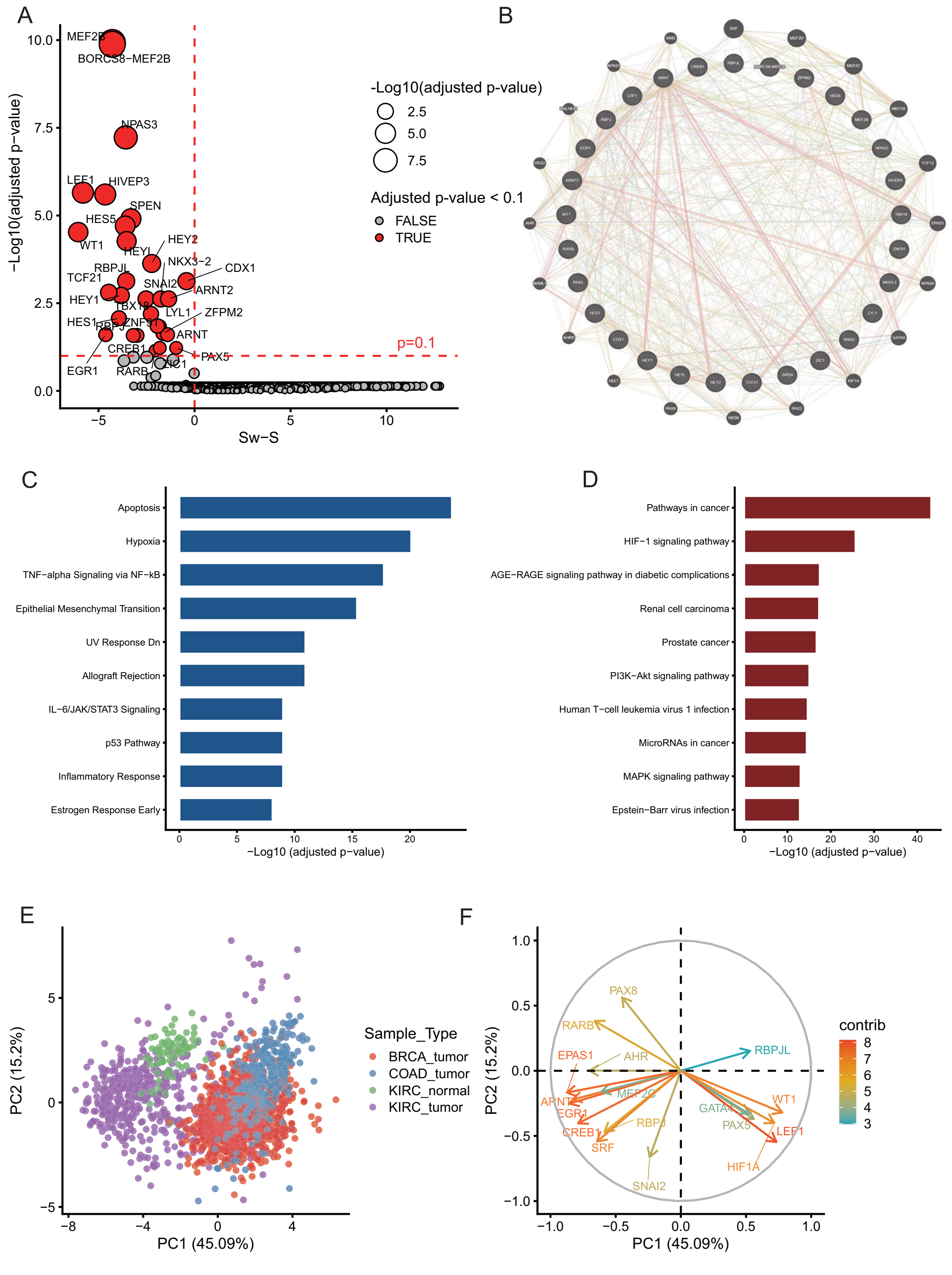


Fig. 3 | Developmental program-sensitive TFs (dsTFs). **A** The dsTFs identified by network communication analysis. The distribution of search information for weighted and unweighted PPI network can be found in Supplementary Fig. 12. **B** Functional connections between all the dsTFs. All functional connections are added by GeneMAINA webserver with default settings when using 29 dsTFs as input seed genes. The colors of the edges in the figure indicate different types of gene interactions, and the width of the edges indicate the confidence of the connections. The detailed information of each interaction can be found in Supplementary Table 5. The genes in the inner circle are significant dsTFs derived directly from network communication analysis, and the genes in the outer circle are expanded dsTFs predicted by GeneMAINA based on the dsTFs in the inner circle. **C** Enriched functional hallmarks of target genes regulated by at least two dsTFs. **D** Enriched KEGG pathways for target genes regulated by at least two dsTFs. **E** PCA of dsTF activity in cancer samples. **F** Contribution of each dsTF to the first and second principal components in the left figure.

Regulator activity clustering reveals three ccRCC subtypes with different prognoses

We further investigated the clinical significance of the dsTFs by analyzing their activity and prognosis in patients with ccRCC in TCGA (KIRC). Using the activities of the 17 dsTFs as features, a sample–sample similarity network was constructed for ccRCC. After applying the network clustering algorithm[51], we obtained three highly stable clusters (named C0, C1, and C2; Fig. 4A and Supplementary Fig. 7). A heatmap of regulatory activities showed that genes were enriched in each cluster (Fig. 4B). The samples in C0 had relatively higher activity of the dsTFs RBPJL, GATA4, HIF1A, WT1, PAX5, and LEF1. The activated dsTFs in C1 included SNAI2, MEF2C, EGR1, RBPJ, ARNT, CREB1, SRF, and EPAS1. Moreover, the activities of three dsTFs, PAX8, RARB, and AHR, were higher in C2 than in others. Survival analysis showed that patients in these three clusters had significantly different prognoses (Fig. 4C, $P < 2 \times 10^{-8}$, log-rank test). Specifically, patients in the C0 group had the worst prognosis, those in the C1 group had the best prognosis, and those in the C2 group had a prognosis between those of the other two groups. These differences remained significant even after accounting for patient sex, age, and cancer stage ($P < 0.01$, Cox proportional hazard test, Fig. 4D). It is worth noting that our classification results were different from those of the TCGA ccRCC subtypes established using the whole mRNA transcriptome and more significantly different for patient survival ($P < 3 \times 10^{-6}$ in TCGA). The relatively even distribution of VHL mutations in the three clusters suggested that the states of these dsTFs were also influenced by other factors.

Genetically, C0 exhibited an increased frequency of BAP1 mutations (Fig. 4E and Supplementary Fig. 8), which has been linked to poor prognosis in patients with ccRCC. However, these mutations were only observed in 17% of patients in the C0 group. For most of these patients, the prognosis could not be attributed to a single genetic mutation. To verify the validity of our classification on other data, we trained a support vector machine (SVM)-based multiclass classifier using dsTF activity as features and three clusters as labels. This model achieved 97% accuracy in fivefold cross-validation species (Fig. 4E). We applied the SVM model to an independent dataset of 106 patients with ccRCC from Asia (Tokyo-ccRCC). The results showed that there was still a significant difference in prognosis among the three predictive groups of patients, and the trends were fully consistent with our TCGA data (Fig. 4G).

We analyzed the contributions of different dsTFs for classification using the SHapley Additive exPlanations (SHAP) algorithm[52]. The SHAP algorithm provides an additive explanation for the contributions of individual features to a machine learning prediction result via a game theory approach. For samples predicted to be in the C0 group, the three most important features were PAX5, EGR1, and RBPJL activity (Supplementary Fig. 9A). For samples predicted to be in the C1 group, the most important features were MEF2C, RBPJL and EGR1 activity (Supplementary Fig. 9B). For samples in the C2 group, the most important features were the activities of MEF2C, PAX8 and PAX5 (Supplementary Fig. 9C). An example of the prediction results from the Tokyo-ccRCC data is shown in Fig. 4C. The predicted outcome group for this patient (ccRCC-16) was C1. The lower activity of PAX8 and RBPJ, as well as the higher activity of ARNT, positively contributed to the prediction results, even though the lower activity levels of other dsTFs, such as SRF, led the SVM model to consider this sample somewhat different from the C0 samples in the training set. Therefore, the type of each sample is the result of the combinatorial regulation of multiple dsTFs rather than a single gene. This finding also illustrates the importance of systematically identifying downstream regulated genes to accurately understand the role of VHL in ccRCC.

Epigenetic regulator EP300 as a therapeutic target for the ccRCC subtype with the worst prognosis

The core regulatory network composed of key transcription factors is not only involved in maintaining aberrant tumor cell proliferation but may also be a potential therapeutic target. According to our findings, the C0 group had the worst prognosis; therefore, we focused our analysis and explored strategies that could disable the C0 group-associated regulatory network. Transcription factors generally bind poorly to small-molecule compounds, and there are fewer drugs that directly target them. However, they are often dependent on specific epigenetic factors for their function[53]. Therefore, we analyzed the epigenetic factors that functionally interacted with the six dsTFs (i.e., RBPJL, GATA4, HIF1A, WT1, PAX5, and LEF1) that exhibited relatively high activity in the C0 group. Network analysis revealed that the histone acetyltransferase EP300 had the highest degree in the PPI subnetwork comprising these dsTFs (Fig. 5A) and that five of these six dsTFs interacted with it directly, suggesting that EP300 may be critical for maintaining the C0-associated core regulatory network.

To test our hypothesis from a functional point of view, we analyzed the genetic screening data of 16 ccRCC cell lines in the DepMap database[54]. Of these 16 cell lines, 15 were predicted to be part of the C0 group by our SVM model. Taking KMRC-20 as an example, we can see that PAX5, GATA4, WT1, and HIF1A among the 6 dsTFs associated with C0 all contributed significantly to the prediction results (Fig. 5B). In line with our expectation, we observed that 13 of these 15 C0 cell lines showed some degree of growth inhibition after EP300 knockdown, and three of them showed strong inhibition, with Chronos scores[55] of less than −0.5 (Fig. 5C). Furthermore, we analyzed the inhibitory effect of A-485, a small-molecule inhibitor of EP300, on ccRCC cell lines. A-485 is a potent, highly selective, and drug-like inhibitor that can bind to the catalytic active site of p300 (the protein produced by EP300)[56]. Of the 13 cell lines with drug responses and data, 12 were predicted to be in the C0 group (Fig. 5D). The growth of eight of these cell lines was inhibited by the addition of A-485, five of which (VMRC-RCZ, KMRC-20, A498, and UO31) exhibited strong inhibition (log2(FC) < − 1). Taken together, our analysis of the dsTF regulatory network suggested that the epigenetic regulator EP300 may be a potential therapeutic target for ccRCC.

Discussion

The cancer-type specificity exhibited by driver mutations has been found to be one of the most obvious patterns in accumulated cancer genomics data. To address this problem, researchers need to compare the functions of mutations among many different cancer types, improving the time and cost of experimental studies. As a result, the underlying molecular mechanisms are still unclear in most cases. Harnessing the big data in developmental biology offered by single-cell sequencing technology, we designed a computational biology approach based on network analysis to overcome this challenge and study the impact of kidney developmental programs on ccRCC-specific VHL mutations. We successfully recapitulated many important transcriptional regulators downstream of VHL mutations and discovered a new therapeutic target based on the core regulatory network we identified, demonstrating the usefulness of our approach.

The innovativeness of our approach can be summarized in the following aspects. First, we linked normal developmental programs to abnormal tumor states by latent variable analysis. While single-cell data offer unprecedented resolution and sample size compared to bulk data, they are also noisier and more unstable. By compressing the transcriptional profiles of single cells into low-dimensional latent variables, our approach improves the stability and interpretability of results obtained when analyzing transcriptional patterns across platforms. Second, we designed an information theory-based network modeling approach to study the impact of developmental programs on the communication of cancer driver genes. Because only the developmental program and the topology of the molecular network need to be considered in each analysis, our model is simpler and avoids the interference of the presence of other driver mutations in our analysis. Finally, this approach provides a framework for analyzing the interactions between cellular context and driver mutations, and with some minor modifications of the data types, the method could also be used to study other factors (such as wound healing, chronic inflammation or aging) that may influence the effects of driver mutations only if suitable scRNA-seq data are available.

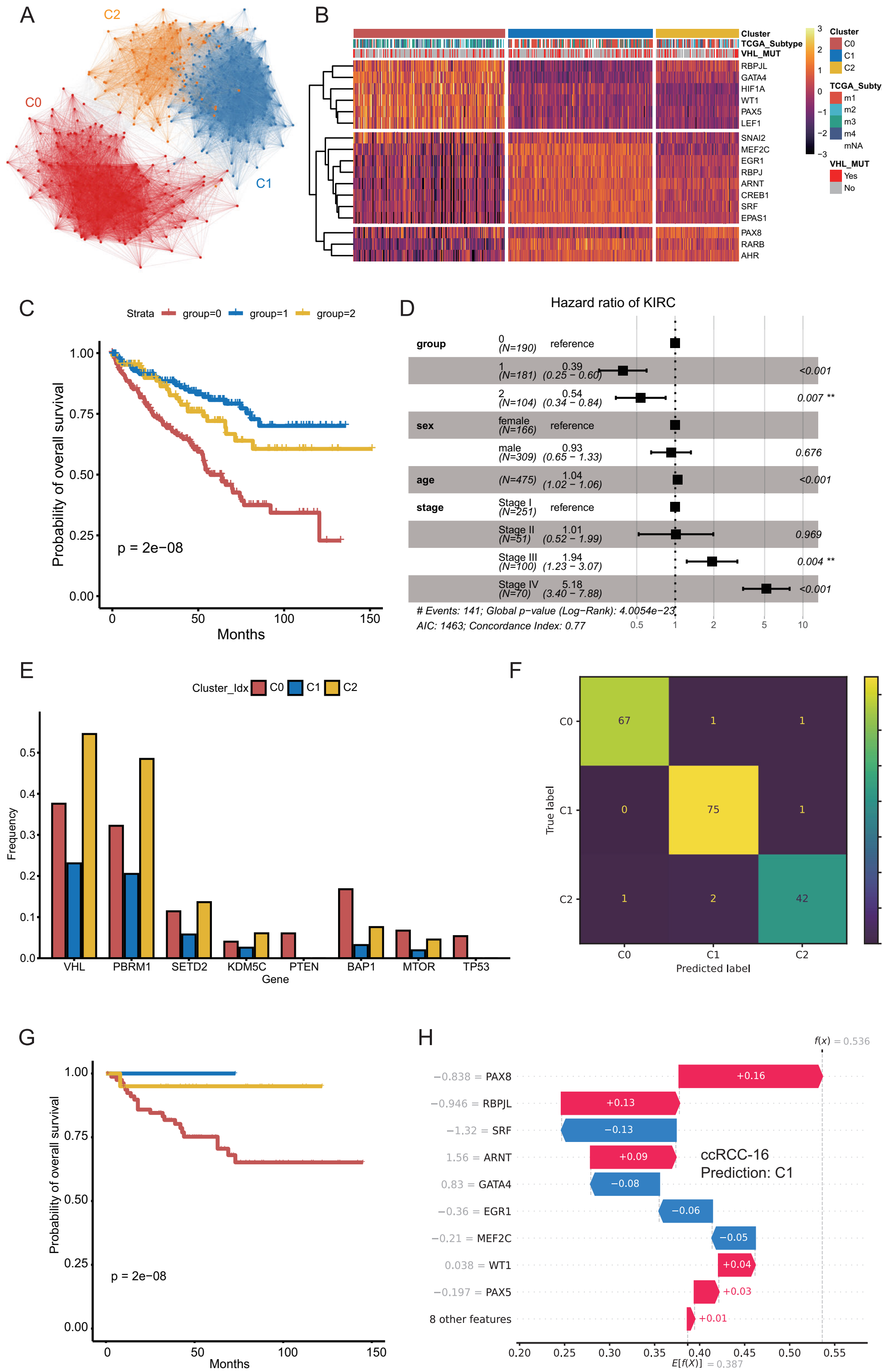


Fig. 4 | Cluster analysis of ccRCC based on dsTF activity. A Sample–sample similarity network and network clustering results. Each dot represents a patient sample of ccRCC, and the edges indicate similarities in their dsTF activities (PCC ≥ 0.4). **B** Distribution of dsTF activity in three groups of ccRCCs. **C** Kaplan-Meier plot of patients with ccRCC (TCGA). **D** Cox proportional hazards model of patients with ccRCC (TCGA). **E** Distribution of driver mutations in the three groups of patients with ccRCC. **F** Performance of the SVM-based multiclass classification model. **G** Kaplan-Meier plot of an Asian cohort of patients with ccRCC (Tokyo-ccRCC). Subtypes were predicted by the SVM model trained on TCGA data. **H** The importance of dsTFs according to SHAP analysis for subtype prediction of one patient in the Tokyo-ccRCC cohort (ccRCC-16). The horizontal axis indicates the probability predicted by the model and the vertical axis shows each dsTF and its corresponding activity value. The red and blue arrows indicate the positive and negative contributions of each dsTF value to the final predicted probability. dsTF activities move the prediction value $E[f(x)]$ from the expected model output and the final prediction result $f(x)$ is interpreted as the sum of the contributions of each dsTF.

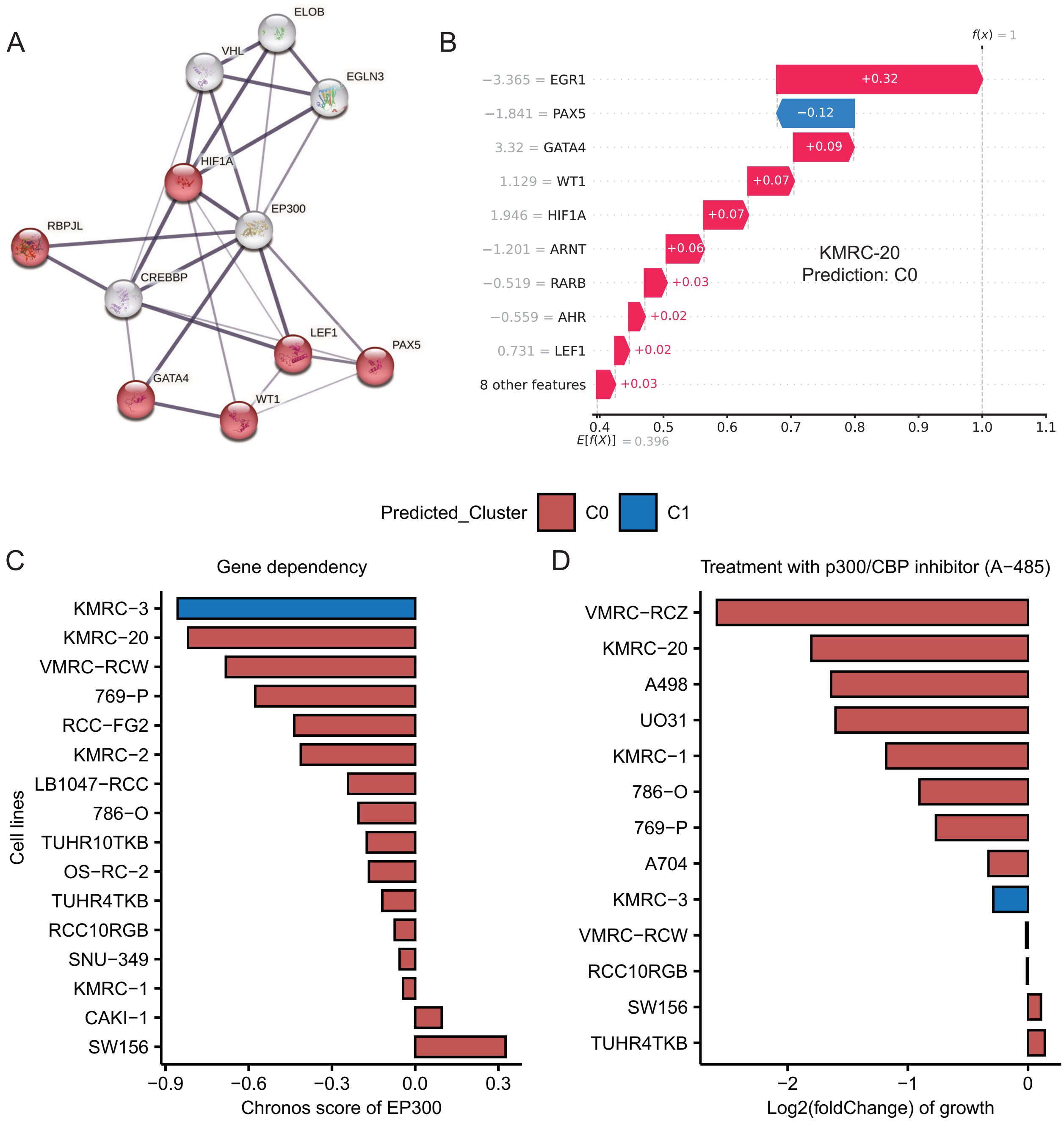


Fig. 5 | Prediction of drug targets for patients with ccRCC with poor prognosis. **A** Genes (white) that are functionally connected with dsTFs (red). The edges indicate the confidence level of the interactions in STRING network. **B** The importance of dsTFs according to SHAP analysis for the subtype prediction of a ccRCC cell line (KMRC-20). The meanings of the individual elements of the diagram are consistent with those in Fig. 4H. **C** Gene dependency score of EP300 in ccRCC cell lines. **D** Changes in the growth of ccRCC cell lines after treatment with the p300/CBP inhibitor A-485.

The PPI network we used in this study is an integrated functional network that encompasses as much knowledge as possible about interactions between genes without considering their biological context. Nevertheless, we actually filtered the genes in the subnetwork for functional relevance in subsequent analyses, including filtering out genes with low variability using organ-specific scRNA-seq data, assigning weights to genes by NMF analysis, and excluding components that are weakly related to the target cancer type by comparing the activities of the developmental program across 19 cancer types. Therefore contextual information about different organs and cancer types is captured in our results.

We showed that the kidney developmental program P9 was selectively activated in ccRCC. Unexpectedly, a significant fraction of the genes with high weights in P9 originated early in eukaryotic evolution. Enrichment analysis of gene functions revealed that the most abundant genes in P9 were those encoding ribosomal proteins. Traditionally, the composition of ribosomes, which act as protein translation machines, has been thought to be highly similar in cells of all tissues. However, recent studies have shown that the situation is much more complicated. The specific proteins that make up the ribosome may differ considerably in different tissues and developmental stages, which is also known as ribosome heterogeneity[57,58]. The preferential translation of mRNAs by ribosomes with different compositions[59] may lead to different translation efficiencies of the same mRNAs in different cells, resulting in cell-type-specific posttranslational regulatory mechanisms. In this way, these ancient genes may also make important contributions to the cancer specificity of mutations. Although our data do not provide any direct evidence for this possibility, they do provide a potential explanation for the presence of multiple ribosomal genes in P9 and warrant further analysis in future work.

By interacting with DNA, transcription factors directly control the expression of target genes in the nucleus and are therefore key factors in bridging upstream cellular signaling and downstream cellular responses[60]. With this in mind, we focused on the identification of developmental programs impacting TFs downstream of VHL mutations. The 49 dsTFs we identified included multiple well-established VHL-dependent factors, such as HIF1A, EPAS1 and PAX8. Inevitably, some of these findings will be determined to be false-positive predictions, but many genes (including PAX2, WT1, RARB, BHLHE41, and others) regulate kidney development or participate in the HIF pathway, which supports a role in VHL mutation-related ccRCC. In addition, some previously known VHL-dependent factors, such as EZH1, ZHX2, and ZNF395, were still missed by our methods[61–63]. EZH1 is not a TF and has not been considered in network communication analysis. We did not observe significant changes in the ΔS of ZHX2 or ZNF395.

Based on the identified dsTFs and the resulting gene regulatory network, we developed a prognostic classification model for patients with ccRCC and suggested that EP300 may be critical for maintaining dsTF activity in patients with a poor prognosis. The EP300-encoded protein p300 generally forms a complex with CREB-binding protein (CBP) that activates gene transcription by working in concert with other transcription factors[64]. p300/CBP can interact with HIFα to promote its activation of downstream target genes, and some pioneering works have reported the effectiveness of p300/CBP inhibitors (including CSC646, HBS1, and CPTH2) for the treatment of ccRCC[65]. However, the poor potency and selectivity of these p300/CBP inhibitors have prevented their clinical application. In fact, we found that CSC464 did not have a strong inhibitory effect on ccRCC cells in the C0 group. However, when we analyzed the experimental data of A-485, which has good selectivity and potency, we observed growth inhibition in several ccRCC cell lines. Combining dsTF activity analysis and next-generation p300/CBP inhibitors[56,66] may help to further develop more suitable therapeutic regimens for patients with ccRCC.

Our approach rests on several basic assumptions; although this makes our analysis more efficient, some key factors have been ignored that may affect cancer specificity. First, we only considered the possible effects of normal developmental programs on driver mutations. However, as noted in some studies, the activated developmental programs in tumor cells and the programs during normal development may be very different[67]. Second, we ignored interactions between mutations. When a driver gene mutated, there may already have existed one or more driver mutations on the genome of tumor cells, and these early mutations may play important roles in shaping the fitness landscape of later mutations. Fortunately, at least for VHL, it is generally considered to be the first mutated driver gene in ccRCC[68]; therefore, the impact of this issue on our conclusions may not be obvious. Third, search information-based analysis of network communication assumes that oncogene signals are propagated along the shortest paths in the network, an assumption that may oversimplify the signaling process in the cell.

Finally, the pipeline presented in the manuscript can be further improved. For example, more powerful latent variable analysis models than NMF have been proposed[69,70] that can improve the accuracy and interpretability of the results. Gene interaction network reconstruction algorithms based on single-cell multiomics data are also developing rapidly[71,72]. The gene networks derived from these methods have better cell specificity and can be used as an alternative to the PPI network used here for more accurate analysis.

In conclusion, our approach provides a flexible solution for analyzing interactions between mutations and cellular contexts. With the accumulation of single-cell, molecular interaction, and cancer genomic data, we hope that these findings will play more roles in resolving the biological effects of cancer driver mutations and can complement genetic screening approaches.

Methods

Single-cell RNA-seq datasets and preprocessing

Human fetal kidney scRNA-seq data (raw gene count matrix) and associated cell annotations were downloaded from https://descartes.brotmanbaty.org/[9]. The cells of origin of ccRCC have been identified as from proximal tubule[73,74] that is developed from fetal metanephric cells[75]. For these reasons, only expression data from cells annotated with “Kidney-Metanephric cells” and protein-coding genes were retained. We used the Bioconductor package Seurat[76] for data quality control and normalization. Specifically, we filtered cells that had feature counts greater than 7500 or less than 200. There were no cells with mitochondrial counts >5%. Gene expression data were normalized using “LogNormalize” method, and the scale factor was set to 10,000. The top 5000 highly variable genes were identified through the Seurat function FindVariableFeatures (selection.method = “vst”). Finally, we obtained a scRNA-seq gene expression matrix of 5000 genes and 89,714 metanephric cells for further analysis.

Protein–protein interaction network

The functional protein–protein interaction (PPI) network was downloaded from STRING[77] database v11.5 (https://string-db.org/). We considered only interactions with high confidence (interaction scores greater than 750). Then, the largest connected component of the PPI network was retained for analysis, resulting in a network with 15,193 genes and 210,014 interactions between them. To determine the functional context surrounding VHL, we further selected VHL and its first and secondary network neighbors in the PPI network according to Qing et al.’s work[25] showing that functional genes biologically relevant to cancer are predominantly distributed within this range.

We also tested the inclusion of more distant genes in the VHL subnetwork, but this resulted in an over-inflated subnetwork that was not specific to VHL. For example, if genes three steps away from VHL were included in the subnetwork, the size of the subnetwork would include more than 10,000 genes. The distribution of distances between genes and VHL is also shown in Supplementary Fig. 10. We named the selected genes and their interactions the VHL-centered subnetwork, which included 2785 genes.

Extraction of developmental programs

We used the expression data of 768 genes that were shared between the scRNA-seq gene expression matrix and VHL-centered network to extract developmental programs. To do this, the single-cell gene expression profiles of the 768 genes were decomposed with non-negative matrix factorization (NMF) algorithm[27] provided by Scikit-learn[78] package. To minimize the influence of batch effects and differences in gene expression profiling platforms, we transformed the expression levels of genes into ranks in a cell using the SciPy[79] function scipy.stats.rankdata (method = ‘min’) before NMF decomposition. For the gene expression profile matrix $X \in R^{n\times p}$, where n is the number of cells and p is the number of genes, the NMF algorithm decomposes the gene into two non-negative matrices $W \in R^{n\times k}$ and $H \in R^{k\times p}$:

$$X \sim WH \tag{1}$$

Suppose $n > k$, $p > k$. W is the transformed gene expression data. And H is the component matrix, representing the weight of each gene in each component. The k components are also called developmental programs in this manuscript. We determined the value of k using the method of Kotliar et al[14]. Briefly, we first estimated an appropriate range of k according to the proportion of variance explained by components via principal component analysis (PCA). We observed that when $k \geq 11$, the ability of more components to explain variation no longer changes significantly (Fig. 2A), suggesting that gene expression matrix can be effectively represented in a space consisting of about 11 independent components. Second, we measured the quality of the matrix decomposition using Frobenius norm error ($\|X - WH\|$). The lower the Frobenius norm error, the smaller the loss after matrix decomposition. We observed that Frobenius norm error is decreasing as k increases. However, larger k may also lead to overfitting of the results. Hence, we further calculated the silhouette score to test the stability of the solution. For each k, we repeated the NMF analysis 10 times with different random seeds. The component matrices (H) obtained from these 10 matrix decompositions were then clustered with K-means clustering ($K = k$), and the quality of the clustering was measured with a silhouette score. The closer the silhouette score is to 1 the more reliable and stable the NMF results are. As illustrated in Fig. 2B, the silhouette score decreases rapidly when k is greater than 11. According to these results, we chose $k = 11$.

Functional annotation of developmental programs

For each program, the top 50 genes with the highest weights were submitted to Enrichr[80] webserver for functional enrichment analysis. We used gene sets of KEGG pathways, GO terms and PanglaoDB. A false discovery rate (FDR)-adjusted $P < 0.05$ was considered to indicate significant enrichment.

Calculation of the activities of developmental programs in TCGA samples

We downloaded batch effect-corrected RNA-seq data[36]. The normalized TCGA tumor gene expression data in FPKM format were used for analysis. The data included 6142 gene expression profiles across 19 cancer types. After picking genes that overlapped with the 768 genes used in the developmental programs, we transformed the expression levels of genes into ranks in a tumor sample. The TCGA gene expression data were projected into the space of developmental programs through the Moore–Penrose pseudoinverse[24]:

$$\hat{W} = YH^{-1} \tag{2}$$

where Y is the rank-normalized TCGA tumor gene expression matrix, H^{-1} is the pseudoinverse of H, which was extracted from scRNA-seq data, and $\hat{W}$ is the matrix, in which the values represent the activities of the developmental programs in TCGA samples.

We also downloaded the VHL mutation states of these TCGA samples from cBioPortal[81]. A sample was deemed VHL mutated when at least one putative driver mutation (according to the annotations provided by cBioPortal) of VHL was identified.

Search information

The search information[30] (S) is a measurement that quantifies the information one needs to go from source node s to target node t along possible shortest paths in a network. In an unweighted network can be expressed as:

$$S(s \to t) = -\log_2\left(\sum_i^N \frac{1}{k_s}\prod_j \frac{1}{k_j - 1}\right) \tag{3}$$

where N is the number of shortest paths from s to t, i is an index for a specific shortest path, k_s is the degree of s, j is the index of each node (other than s and t) along path i, and k_j is the degree of node j. Search information is an entropy-based method for measuring the uncertainty for s to t. The unit is bits. A high S means that more information is needed if one wants to efficiently send a specific signal from s to t. In a weighted network, the possibility of a signal transmitting from one node j to another specific neighbor node $j + 1$ can be replaced by the proportion of the edge weight in the sum of the weights of all possible edges in the next step from j, indicating that an edge with a higher weight is more likely to be chosen as an exit link. One can obtain weighted network-based search information S_w.

To model the influence of a developmental program on oncogenic signal transduction, we tested the difference in search information from driver genes (such as VHL) to downstream transcription factors (TFs) with or without imposing the developmental program on the network. Specifically, we first computed the search information S using the driver gene as the source node and a specific TF as the target node in the unweighted PPI network; then, we computed a S_w on the developmental program-weighted PPI network. Finally, we compared the difference between S_w and S:

$$\Delta S = S_W - S \tag{4}$$

If $\Delta S < 0$, means developmental program improves the communication efficiency between the driver gene and the specific downstream TF. Note that the PPI network used here includes all edges with high confidence.

To weight PPI network with a developmental program, we performed personalized PageRank[29] analysis with the top 100 genes and their weights in the developmental program as the "personalization vector". As a result, all genes were assigned scores reflecting their relative activity under the influence of the developmental program. We also tested personalized PageRank analysis using top 50, 150, or 200 genes. The results were very similar with that of using top 100 genes (Supplementary Fig. 11). We assumed that a more activated gene more readily received signals from other genes. Thus, we used the PageRank score of node $j + 1$ as the edge weight in computing S_w. Personalized PageRank analysis was performed with pagerank function in NetworkX[82].

To control the false-positive rate, we also randomly selected a subset of genes of the same size as in the above "personalization vector" to assign them gene weights in the developmental program and computed ΔS_r. This procedure was repeated 1000 times. Then, an empirical P value of ΔS was calculated through fitting a normal distribution of ΔS_r with the R package fitdistrplus.

The names of 1639 TFs were downloaded from http://humantfs.ccbr.utoronto.ca/[83]. A total of 1040 TFs that could be found in the PPI network were used for analysis. For a specific developmental program and a driver gene, TFs were analyzed one by one. Each TF was set as the target node in neach time. The P values of all the TFs were adjusted using the Benjamini–Hochberg method provided in p.adjust function in R. For convenience, we refer to TFs with significant ΔS (adjusted $P < 0.1$) values as developmental program-sensitive TFs (dsTFs) in this manuscript.

Expanded dsTF list with GeneMANIA

The dsTFs were submitted to the GeneMANIA webserver[42] to explore additional TFs that have functions similar to these but were missed by the search information analysis. All types of protein-protein interaction data were used. The 20 most similar genes recommended by GeneMANIA with default settings were added to the input TFs to obtain an expanded dsTF list.

Regulatory network analysis

TF–target information was retrieved from the human gene regulatory network (hDB.rdata) in the NetAct[46] package. These data included 875 TFs and 16,364 high-quality literature-based TF–target relationships complied from multiple gene regulation databases. We filtered out TF–target relationships showing only weak gene expression correlations (defined as an absolute Pearson's correlation coefficient (PCC) < 0.03, suggested by SCENIC's protocol[84]) in the TCGA-KIRC gene expression profiles, which are unlikely to be functional in this disease. We defined retained TF–target relationships as activation when the PCC was >0 or as inhibition when the PCC was ≤0. All target genes of one TF combined with TCGA-KIRC expression profiles were subjected to single-sample gene set scoring to compute the regulatory activity of the TF in every TCGA sample. A dsTF was excluded if it was not included in the NetAct human regulatory network or if it had fewer than 10 target genes. Single-sample gene set scoring was also applied to the gene expression profiles of the TCGA-BRCA and TCGA-COAD cohorts. DAVID[85] and Enrichr were used to perform functional enrichment analysis on the targets coregulated by at least two dsTFs. A false discovery rate (FDR)-adjusted P value < 0.05 indicated a significant enrichment. PCA of the regulatory activities of the dsTFs in the cancer samples was conducted using the R function prcomp and the R package factoextra.

Clustering of ccRCC samples

The TF regulatory activity matrix $X \in R^{n \times p}$ was used to cluster TCGA-KIRC samples into three clusters, where n is the number of samples and p is the number of TFs. To do this, we first conducted data standardization and computed the PCCs between samples. Two samples were connected if their PCC≥cutoff to construct a sample–sample similarity network. Next, we used the Louvain community detection algorithm[51] to cluster network nodes into modules (i.e., sample clusters). We tested different cutoffs and evaluated the clustering results with the silhouette coefficient, the Calinski–Harabasz index and the coverage of clusters and identified 0.4 as the best cutoff value, corresponding to three clusters covering all samples.

To apply the clustering results to new samples, we trained a multiclass support vector machine (SVM) model using TF regulatory activities in TCGA-KIRC as features and cluster indices obtained in the above step as labels. Its accuracy was evaluated using fivefold cross-validation. We used the SHAP algorithm[52] to analyze the contribution of each feature to the prediction.

We used the louvain_communities function from the NetworkX package to perform sample–sample network clustering and the svm.SVC function from Scikit-learn to construct the SVM model. SHAP analyses were conducted with the Python package shap.

Survival analysis

Clinical data from the TCGA-KIRC cohort were downloaded using the Bioconductor package TCGAbiolinks. Clinical data and preprocessed RNA microarray data for patients in the Tokyo-ccRCC cohort were obtained from the supplementary data of the corresponding article[86]. Survival analyses, including Kaplan–Meier analysis, log-rank test and Cox proportional hazards model construction, were performed using the R packages survival and survminer.

Therapeutic target analysis

Six Cluster C0-related TFs (RBPJL, GATA4, HIF1A, WT1, PAX5 and LEF1) were submitted to STRING webserver, and an expanded subnetwork was identified using default settings. Gene dependency data[54] for EP300 and drug sensitivity data[87] for an EP300 inhibitor (A-485) were obtained by querying the DepMap database (https://depmap.org/portal/). Gene expression data for the cell lines were downloaded from Cell Model Passports[88] (https://cellmodelpassports.sanger.ac.uk/).

Statistical analysis and data visualization

All the statistical analyses that were not specified were conducted in R 4.2.3 or 4.3.2. The figures were generated with the R packages ggplot2, ggpubr, pheatmap, ComplexHeatmap, viridis, maftools, cowplot, survminer and the Python packages Scikit-learn and shap. The networks were visualized with Cytoscape[89].

Data availability

The scRNA-seq data for fetal development are available at https://descartes.brotmanbaty.org/. Human PPI data are available at the STRING website (https://cn.string-db.org/). Batch effect-corrected TCGA RNA-seq data are provided on figshare at https://figshare.com/articles/dataset/Data_record_3/5330593. The human TF list is available at http://humantfs.ccbr.utoronto.ca/. The human gene regulatory relationships reported in the literature can be found in the hDB.rdata file of NetAct at https://github.com/lusystemsbio/NetAct. There are no new experimental data generated in this work.

Code availability

The final model scripts, files, and information are available at https://github.com/NeoDong/OncoNicheDev.

Received: 14 June 2024; Accepted: 21 September 2024;
Published online: 03 October 2024

References

1. Haigis, K. M., Cichowski, K. & Elledge, S. J. Tissue-specificity in cancer: the rule, not the exception. *Science* **363**, 1150–1151 (2019).
2. Sondka, Z. et al. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).
3. Martínez-Jiménez, F. et al. A compendium of mutational cancer driver genes. *Nat. Rev. Cancer* **20**, 555–572 (2020).
4. Garraway, L. A. & Sellers, W. R. Lineage dependency and lineage-survival oncogenes in human cancer. *Nat. Rev. Cancer* **6**, 593–602 (2006).
5. Schneider, G., Schmidt-Supprian, M., Rad, R. & Saur, D. Tissue-specific tumorigenesis: context matters. *Nat. Rev. Cancer* **17**, 239 (2017).
6. Baggiolini, A. et al. Developmental chromatin programs determine oncogenic competence in melanoma. *Science* **373**, eabc1048 (2021).
7. Patel, S. A. et al. The renal lineage factor PAX8 controls oncogenic signalling in kidney cancer. *Nature* **606**, 999–1006 (2022).
8. Weiss, J. M. et al. Anatomic position determines oncogenic specificity in melanoma. *Nature* **604**, 354–361 (2022).
9. Cao, J. et al. A human cell atlas of fetal gene expression. *Science* **370**, eaba7721 (2020).
10. Cao, J. et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
11. Moiso, E. et al. Developmental deconvolution for classification of cancer origin. *Cancer Discov.* **12**, 2566–2585 (2022).
12. Stein-O'Brien, G. L. et al. Enter the matrix: factorization uncovers knowledge from Omics. *Trends Genet.* **34**, 790–805 (2018).
13. Stein-O'Brien, G. L. et al. Decomposing cell identity for transfer learning across cellular measurements, platforms, tissues, and species. *Cell Syst.* **12**, 203 (2021).
14. Kotliar, D. et al. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *eLife* **8**, e43803 (2019).
15. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* **3**, 246–259 (2013).
16. Davis-Marcisak, E. F. et al. Transfer learning between preclinical models and human tumors identifies a conserved NK cell activation signature in anti-CTLA-4 responsive tumors. *Genome Med.* **13**, 129 (2021).
17. Puram, S. V. et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* **171**, 1611–1624.e24 (2017).
18. Parikshak, N. N., Gandal, M. J. & Geschwind, D. H. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nat. Rev. Genet.* **16**, 441–458 (2015).
19. Chen, J. C. et al. Identification of causal genetic drivers of human disease through systems-level analysis of regulatory networks. *Cell* **159**, 402–414 (2014).
20. Castro, M. A. A. et al. Regulators of genetic risk of breast cancer identified by integrative network analysis. *Nat. Genet.* **48**, 12–21 (2016).
21. Yang, L., Chen, R., Goodison, S. & Sun, Y. An efficient and effective method to identify significantly perturbed subnetworks in cancer. *Nat. Comput. Sci.* **1**, 79–88 (2021).
22. Kong, J. et al. Network-based machine learning in colorectal and bladder organoid models predicts anti-cancer drug efficacy in patients. *Nat. Commun.* **11**, 5485 (2020).
23. Ideker, T. & Krogan, N. J. Differential network biology. *Mol. Syst. Biol.* **8**, 565 (2012).
24. Tamayo, P. et al. Metagene projection for cross-platform, cross-species characterization of global transcriptional states. *Proc. Natl. Acad. Sci. USA* **104**, 5959–5964 (2007).
25. Qing, T. et al. Cancer relevance of human genes. *JNCI: J. Natl Cancer Inst.* **114**, 988–995 (2022).
26. Chang, J. T. et al. A genomic strategy to elucidate modules of oncogenic pathway signaling networks. *Mol. Cell* **34**, 104–114 (2009).

27. Brunet, J.-P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. USA* **101**, 4164–4169 (2004).
28. Visvader, J. E. Cells of origin in cancer. *Nature* **469**, 314–322 (2011).
29. Cowen, L., Ideker, T., Raphael, B. J. & Sharan, R. Network propagation: a universal amplifier of genetic associations. *Nat. Rev. Genet.* **18**, 551 (2017).
30. Trusina, A., Rosvall, M. & Sneppen, K. Communication boundaries in networks. *Phys. Rev. Lett.* **94**, 238701 (2005).
31. Hofree, M., Shen, J. P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nat. Methods* **10**, 1108–1115 (2013).
32. Kim, S. Y., Choe, E. K., Shivakumar, M., Kim, D. & Sohn, K.-A. Multi-layered network-based pathway activity inference using directed random walks: application to predicting clinical outcomes in urologic cancer. *Bioinformatics* **37**, 2405–2413 (2021).
33. Linehan, W. M. & Ricketts, C. J. The Cancer Genome Atlas of renal cell carcinoma: findings and clinical implications. *Nat. Rev. Urol.* **16**, 539–552 (2019).
34. Kim, W. Y. & Kaelin, W. G. Role of VHL gene mutation in human cancer. *J. Clin. Oncol.* **22**, 4991–5004 (2004).
35. Ding, L. et al. Perspective on oncogenic processes at the end of the beginning of cancer genomics. *Cell* **173**, 305–320.e10 (2018).
36. Wang, Q. et al. Unifying cancer and normal RNA sequencing data from different sources. *Sci. Data* **5**, 180061 (2018).
37. Sun, H. et al. Comprehensive characterization of 536 patient-derived xenograft models prioritizes candidates for targeted treatment. *Nat. Commun.* **12**, 5086 (2021).
38. Franzén, O., Gan, L.-M. & Björkegren, J. L. M. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database* **2019**, baz046 (2019).
39. Gulati, G. S. et al. Single-cell transcriptional diversity is a hallmark of developmental potential. *Science* **367**, 405–411 (2020).
40. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* **51**, D587–D592 (2023).
41. Capra, J. A., Williams, A. G. & Pollard, K. S. ProteinHistorian: tools for the comparative analysis of eukaryote protein origin. *PLoS Comput. Biol.* **8**, e1002567 (2012).
42. Warde-Farley, D. et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* **38**, W214–W220 (2010).
43. Bigot, P. et al. Functional characterization of the 12p12.1 renal cancer-susceptibility locus implicates BHLHE41. *Nat. Commun.* **7**, 12098 (2016).
44. Huerta-Sánchez, E. et al. Genetic signatures reveal high-altitude adaptation in a set of Ethiopian populations. *Mol. Biol. Evol.* **30**, 1877–1888 (2013).
45. Robson, E. J. D., He, S.-J. & Eccles, M. R. A PANorama of PAX genes in cancer and development. *Nat. Rev. Cancer* **6**, 52–62 (2006).
46. Su, K. et al. NetAct: a computational platform to construct core transcription factor regulatory networks using gene activity. *Genome Biol.* **23**, 270 (2022).
47. Liberzon, A. et al. The molecular signatures database hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
48. Wykoff, C. C. et al. Gene array of VHL mutation and hypoxia shows novel hypoxia-induced genes and that cyclin D1 is a VHL target gene. *Br. J. Cancer* **90**, 1235–1243 (2004).
49. Foroutan, M. et al. Single sample scoring of molecular phenotypes. *BMC Bioinforma.* **19**, 404 (2018).
50. Shen, C. et al. Genetic and functional studies implicate HIF1α as a 14q kidney cancer suppressor gene. *Cancer Discov.* **1**, 222–235 (2011).
51. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.: Theory Exp.* **2008**, P10008 (2008).
52. Lundberg, S.M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* 4768–4777 (Curran Associates Inc., Long Beach, California, USA, 2017).
53. Bradner, J. E., Hnisz, D. & Young, R. A. Transcriptional addiction in cancer. *Cell* **168**, 629–643 (2017).
54. Pacini, C. et al. Integrated cross-study datasets of genetic dependencies in cancer. *Nat. Commun.* **12**, 1661 (2021).
55. Dempster, J. M. et al. Chronos: a cell population dynamics model of CRISPR experiments that improves inference of gene fitness effects. *Genome Biol.* **22**, 343 (2021).
56. Lasko, L. M. et al. Discovery of a selective catalytic p300/CBP inhibitor that targets lineage-specific tumours. *Nature* **550**, 128–132 (2017).
57. Genuth, N. R. & Barna, M. The discovery of ribosome heterogeneity and its implications for gene regulation and organismal life. *Mol. Cell* **71**, 364–374 (2018).
58. Norris, K., Hopes, T. & Aspden, J. L. Ribosome heterogeneity and specialization in development. *WIREs RNA* **12**, e1644 (2021).
59. Shi, Z. et al. Heterogeneous ribosomes preferentially translate distinct subpools of mRNAs genome-wide. *Mol. Cell* **67**, 71–83.e7 (2017).
60. Weidemüller, P., Kholmatov, M., Petsalaki, E. & Zaugg, J. B. Transcription factors: bridge between cell signaling and gene regulation. *PROTEOMICS* **21**, 2000034 (2021).
61. Chakraborty, A. A. et al. HIF activation causes synthetic lethality between the VHL tumor suppressor and the EZH1 histone methyltransferase. *Sci. Transl. Med.* **9**, eaal5272 (2017).
62. Zhang, J. et al. VHL substrate transcription factor ZHX2 as an oncogenic driver in clear cell renal cell carcinoma. *Science* **361**, 290–295 (2018).
63. Yao, X. et al. VHL deficiency drives enhancer activation of oncogenes in clear cell renal cell carcinoma. *Cancer Discov.* **7**, 1284–1305 (2017).
64. Zhu, Y. et al. The role of CREBBP/EP300 and its therapeutic implications in hematological malignancies. *Cancers* **15**, 1219 (2023).
65. Wen, Q. et al. Essential role of bromodomain proteins in renal cell carcinoma. *Mol. Med. Rep.* **28**, 139 (2023).
66. Nicosia, L. et al. Therapeutic targeting of EP300/CBP by bromodomain inhibition in hematologic malignancies. *Cancer Cell* **41**, 2136–2153.e13 (2023).
67. Fazilaty, H. & Basler, K. Reactivation of embryonic genetic programs in tissue regeneration and disease. *Nat. Genet.* **55**, 1792–1806 (2023).
68. Turajlic, S. et al. Deterministic evolutionary trajectories influence primary tumor growth: TRACERx renal. *Cell* **173**, 595–610.e11 (2018).
69. Kunes, R. Z., Walle, T., Land, M., Nawy, T. & Pe'er, D. Supervised discovery of interpretable gene programs from single-cell data. *Nat. Biotechnol.* **42**, 1084–1095 (2023).
70. Rahimikollu, J. et al. SLIDE: significant latent factor interaction discovery and exploration across biological domains. *Nat. Methods* **21**, 835–845 (2024).
71. Erbe, R., Gore, J., Gemmill, K., Gaykalova, D. A. & Fertig, E. J. The use of machine learning to discover regulatory networks controlling biological systems. *Mol. Cell* **82**, 260–273 (2022).
72. Wang, L. et al. Dictys: dynamic gene regulatory network dissects developmental continuum with single-cell multiomics. *Nat. Methods* **20**, 1368–1378 (2023).
73. Young, M. D. et al. Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. *Science* **361**, 594–599 (2018).
74. Zhang, Y. et al. Single-cell analyses of renal cell cancers reveal insights into tumor microenvironment, cell of origin, and therapy response. *Proc. Natl. Acad. Sci. USA* **118**, e2103240118 (2021).
75. Pietilä, I. & Vainio, S. J. Kidney development: an overview. *Nephron Exp. Nephrol.* **126**, 40–44 (2014).
76. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019).
77. Szklarczyk, D. et al. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.* **45**, D362–D368 (2017).
78. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

79. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
80. Kuleshov, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).
81. Cerami, E. et al. The cBio Cancer Genomics Portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
82. Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)* (eds. Varoquaux, G. et al.) (Pasadena, 2008).
83. Lambert, S. A. et al. The human transcription factors. *Cell* **172**, 650–665 (2018).
84. Van de Sande, B. et al. A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nat. Protoc.* **15**, 2247–2276 (2020).
85. Sherman, B. T. et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res.* **50**, W216–W221 (2022).
86. Sato, Y. et al. Integrated molecular analysis of clear-cell renal cell carcinoma. *Nat. Genet.* **45**, 860–867 (2013).
87. Corsello, S. M. et al. Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. *Nat. Cancer* **1**, 235–248 (2020).
88. Hall, C. et al. Cell Model Passports—a hub for clinical, genetic and functional datasets of preclinical cancer models. *Nucleic Acids Res.* **47**, D923–D929 (2018).
89. Saito, R. et al. A travel guide to Cytoscape plugins. *Nat. Methods* **9**, 1069–1076 (2012).

Acknowledgements

The authors thank members of our laboratories for discussions and for providing comments on the manuscript. This work was supported by National Natural Science Foundation of China (nos. 31801122 (X.D.) and 82000127 (X.Z.)).

Author contributions

X.D. and D.Z. designed and performed algorithm development and computational analysis. D.Z. and X.Z. performed the survival analysis. Y.L. and Y.L.L. helped with the interpretation of the gene dependency and drug response data. X.D. conceived this work and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41540-024-00445-2.

Correspondence and requests for materials should be addressed to Xiaobao Dong.

Reprints and permissions information is available at http://www.nature.com/reprints

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