Heterogeneous and Novel Transcript Expression in Single Cells of Patient-Derived ccRCC Organoids

Tülay Karakulak[1,2,3], Hella Anna Bolck [2,4], Natalia Zajac[5,3], Anna Bratus-Neuenschwander[5], Qin Zhang [5], Weihong Qi [5,3], Tamara Carrasco Oltra[5], Hubert Rehrauer[5,3], Christian von Mering[1,3], Holger Moch[2], Abdullah Kahraman[3,6,*]

1 Department of Molecular Life Sciences, University of Zurich,
2 Department of Pathology and Molecular Pathology, University Hospital Zurich
3 Swiss Institute of Bioinformatics, Lausanne
4 Centre for AI, School of Engineering, Zurich University of Applied Sciences (ZHAW), Technikumstrasse 71, 8400 Winterthur, Switzerland
5 Functional Genomics Center Zurich, ETH, Zurich,
6 School for Life Sciences, Institute for Chemistry and Bioanalytics, University of Applied Sciences Northwestern Switzerland, Muttenz
*corresponding author, abdullah.kahraman@fhnw.ch

Abstract

Splicing is often dysregulated in cancer, leading to alterations in the expression of canonical and alternative splice isoforms. This complex phenomenon can be revealed by an in-depth understanding of cellular heterogeneity at the single-cell level. Recent advances in single-cell long-read sequencing technologies enable comprehensive transcriptome sequencing at the single-cell level. In this study, we have generated single-cell long-read sequencing of Patient-Derived Organoid (PDO) cells of clear-cell Renal Cell Carcinoma (ccRCC), an aggressive and lethal form of cancer that arises in kidney tubules. We have used the Multiplexed Arrays Sequencing (MAS-ISO-Seq) protocol of PacBio to sequence full-length transcripts exceptionally deep across 2,599 single cells to obtain the most comprehensive view of the alternative landscape of ccRCC to date. On average, we uncovered 303,547 transcripts across PDOs, of which 40.5% were previously uncharacterized. In contrast to known transcripts, many of these novel isoforms appear to exhibit cell-specific expression. Nonetheless, 37.5% of these novel transcripts, expressed in more than three cells, were predicted to possess a complete protein-coding open reading frame. This finding suggests a biological role for these transcripts within kidney cells. Moreover, an analysis of the most dominant transcript switching revealed that many switching events were cell and sample-specific, underscoring the heterogeneity of alternative splicing events in ccRCC. Interestingly, one of the ccRCC organoids seemed to have a VHL-negative phenotype despite a VHL P25L mutation, underscoring the benign nature of the mutation. Overall, our research elucidates the intricate transcriptomic architecture of ccRCC, potentially exposing the mechanisms underlying its aggressive phenotype and resistance to conventional cancer therapies.

Keywords

ccRCC, single-cell sequencing, full-length sequencing, novel transcript, cell heterogeneity, PacBio

Introduction

Alternative splicing is a pivotal mechanism by which eukaryotic cells enhance their transcriptomic and proteomic diversity. By allowing a single gene to encode multiple RNA variants, alternative splicing contributes significantly to cellular complexity, tissue specificity, and organismal adaptability. In the context of human disease, notably cancer, dysregulation of alternative splicing events can lead to the expression of oncogenic isoforms, influencing tumor initiation, progression, and resistance to therapy. Despite its recognized importance, the comprehensive characterization of alternative splicing at the resolution of individual cells remains a formidable challenge, primarily due to the limitations of conventional sequencing technologies in capturing the full spectrum of splicing events.

Recent advances in single-cell RNA sequencing (scRNA-seq) have revolutionized our understanding of cellular heterogeneity in complex tissues and tumoral environments, revealing unprecedented insights into the transcriptomic variations that define cell types, states, and functions. However, most single-cell studies have relied on short-read sequencing technologies, which, despite their high throughput, fall short of accurately resolving complex splice variants due to their limited read lengths. Long-read sequencing technologies offer a promising solution to these limitations. With the ability to generate reads that span entire transcript isoforms, long-read sequencing enables the direct observation of splicing patterns and the identification of novel isoforms that would be missed or misassembled by short-read technologies (Byrne et al. 2017; Amarasinghe et al. 2020; Bolisetty et al. 2015). However, long-read sequencing was not appropriate for single-cell transcriptome measurements due to the initial lower throughput and high sequencing errors. With the recent advances in sequencing chemistries and transcript concatenation protocols, the restrictions could be overcome, allowing us to measure transcripts in the transcriptome at full-length at single-cell resolution.

Using derivates of this new technology, the research community has begun to investigate the transcriptome of various samples at single-cell resolution. For example, Shiau *et al.* identified a distinct combination of isoforms in tumor and neighboring stroma/immune cells in a kidney tumor, as well as cell-type-specific mutations like VEGFA mutations in tumor cells and HLA-A mutations in immune cells (Shiau et al. 2023). Tian *et al.* highlighted the complexity of the transcriptome in human and mouse samples by identifying thousands of novel transcripts with conserved functional modules enriched in alternative transcript usage, including ribosome biogenesis and mRNA splicing. They found drug-resistance mutations in subclones within transcriptional clusters (Tian et al. 2021). Also, Yang *et al.* observed thousands of novel transcripts in human cerebral organoids, with differentially spliced exons and retained introns (Yang et al. 2023). Cell-type-specific exons with de novo mutations were enriched in autistic patients. In another interesting study, Wan *et al.* integrated single-cell long-read sequencing with single-molecule microscopy and observed distinct but consistent bursting expression for all genes with similar nascent RNA dwell time (Wan et al. 2021; Shiau et al. 2023). The intron removal time spans minutes to hours, suggesting that the spliceosome removes introns progressively in pieces. In a recent study, Dondi *et al.* identified over 52,000 novel transcripts in five ovarian cancer samples that had not been reported previously, and similar to the studies above, discovered cell-specific transcript and polyadenylation site usages and were able to identify a gene fusion event that would have been missed using short-read sequencing (Dondi et al. 2023).

Following in the footsteps of these studies, we have applied PacBio's new Multiplexed Arrays Sequencing (MAS-ISO-Seq) protocol (Al'Khafaji et al. 2023) to probe full-length transcripts in single cells in patient-derived kidney organoids of four clear-cell renal cell carcinoma (ccRCC) patients. To our knowledge, the transcriptome and alternative splicing landscape in single-cell resolution has not been studied in ccRCC despite the heterogeneity and complexity of its tumor microenvironment (Motzer et al. 2022). However, various high-throughput studies point towards an important role of alternative splicing in ccRCC development and treatment response (Wang et al. 2022; Simmler et al. 2022; Zhang et al. 2021a). For example, recent single-cell studies have suggested VCAM1-positive renal proximal tubule cells to be the likely origin of ccRCC (Zhang et al. 2021b; Schreibing and Kramann 2022), which is consistent with the hypothesis that ccRCC is derived from the proximal tubules. Also, ccRCC tumors were found to detain many CD8+ T-cells and macrophages in immune checkpoint inhibition responsive and resistant samples, respectively (Krishna et al. 2021). The distinct response could explain the general good response of ccRCC patients to immunotherapy despite having a low mutational burden in their ccRCC tumors (Borcherding et al. 2021).

Here, for the first time, we are exploring the transcriptome landscape of ccRCC samples and one matched-normal patient-derived organoids (PDOs) in single-cell resolution using single-cell long-read sequencing technology. We detected more than 300,000 unique transcripts across samples found in at least three cells. Of those, 27% were identified as novel, unknown transcripts, of which many are specific to one sample. In addition, we evaluated the coding capability of transcripts and found a higher proportion of complete open reading frames in novel transcripts commonly expressed in many cells. Our findings elucidate the extensive heterogeneity inherent in the splicing landscape of ccRCC samples. This intricate variability underscores the resilience of ccRCC against conventional therapeutic strategies.

Results

Full-length single-cell sequencing reveals transcript diversity and the cell heterogeneity of known and novel transcripts

To discern the transcriptome diversity in ccRCC, we have applied full-length single-cell sequencing using the MAS-Seq protocol (Al'Khafaji et al. 2023) on a PacBio Sequel IIe instrument to five patient-derived organoids (PDO) samples (Fig. 1A). The resulting data included 29.4 to 58.8 million segmented reads per sample (please see Zajac N. *et al.* (accompanying submission) for the technical details). The PDOs were established from four fresh ccRCC tissue samples (Fig. 1B). We sequenced one normal PDO matching ccRCC2, which we established from adjacent normal kidney tissue. All ccRCC PDOs carried a VHL mutation, a hallmark of ccRCC (Table 1). Note that the P25L mutation in ccRCC3 has previously been described as benign (Rothberg 2001). To sequence the single-cell transcriptomes as deeply as possible, we attempted to load as many transcript molecules of as few cells as possible on the flow cell. We managed to sequence a total of 2,599 cells, ranging between 310 and 1091 cells per sample with between 76,232 and 120,658 transcripts per cell (Table 2), an unparalleled depth. All cells except one in ccRCC4 expressed more than three genes and transcripts. Calculation of the number of unique genes and transcripts expressed in at least three cells and their UMI counts per cell revealed that the ccRCC4 sample with the highest number of cells had the lowest number of transcripts, genes, and UMI per cell (Supplementary Figures 1A and 1B). The sequencing depth reached up to 25,997 reads per cell on average in the normal-PDO, followed by 21,308 in ccRCC5, 21,077 in ccRCC2, 16,246 in ccRCC3, and 6,555 in ccRCC4.

Table 1: Clinical data of patient-derived organoid (PDO) samples.

Sample Names in the Manuscript	FGCZ Sample No	VHL Status	Grade
Normal	030669/1	WT	-
ccRCC2	030669/2	c.286C>T	3
ccRCC3	030669/3	c.74C>T	4
ccRCC4	030669/4	c.227T>C	4
ccRCC5	030669/5	c.230insT	4

The Iso-seq pipeline classified transcripts into four categories using SQANTI3 in SMRT-Link. Based on the alignment profile of exon coordinates of transcripts to the reference transcriptome, SQANTI3 (Pardo-Palacios et al. 2023) categorized the transcripts as full-splice match (FSM), incomplete-splice match (ISM), novel in catalog (NIC), and novel not in catalog (NNC) (Fig. 1C). FSM transcripts perfectly align with reference transcripts at their junctions; ISM transcripts have fewer exons at the 5' or 3' ends, while the rest of the internal junctions align with the reference transcript junctions. The novel transcript categories NIC or NNC are

made of new combinations of known splice junctions or have at least one new donor or acceptor site, respectively. We grouped the remaining SQANTI3 transcripts, namely antisense, genic intron, genic genomic, and intergenic, into a single category called 'Other'. On average, we identified 291,459 transcripts across all samples (Table 2). 37.2% of the transcripts were identified as ISM, followed by 36.9% novel transcripts, of which 14.6% and 22.3% were identified as NIC and NNC, respectively. 21.9% of transcripts were annotated as FSM across samples (Fig. 1D, Table 2).

Table 2: Statistics on the number of cells, genes, and transcripts sequenced in this project. *Number of cells expressing 100 genes or transcripts. **Number of genes or transcripts found in at least three cells. FSM: Full splice match, ISM: Incomplete splice match, NIC: Novel In Catalog, NNC: Novel Not In Catalog. Other: Genic, antisense, intergenic, fusion, more Junctions.

	Normal	ccRCC2	ccRCC3	ccRCC4	ccRCC5
Cells	437	373	310	1091	388
Cells*	437	373	310	1090	388
Unique Genes	29,138	26,260	26,657	29,074	30,065
Unique Genes**	18,028	16,759	17,366	17,741	18,074
Unique Transcripts	346,107	303,547	216,926	289,556	301,160
Unique Transcripts**	120,658	97,445	76,232	99,541	99,996
FSM (%)	71,205 (20.5%)	68,352 (22.5%)	50,983 (23.5%)	62,582 (21.6%)	64,378 (21.4%)
ISM (%)	126,715 (36.6%)	101,665 (33.5%)	96,380 (44.4%)	109,263 (37.7%)	102,536 (34%)
NIC (%)	52,740 (15.2%)	49,142 (16.2%)	23,054 (10.6%)	42,698 (14.7%)	49,049 (16.3%)
NNC (%)	83,308 (24.1%)	73,820 (24.3%)	37,997 (17.5%)	62,710 (21.7%)	72,372 (24.0%)
Other (%)	12,139 (3.5%)	10,568 (3.5%)	8,512 (3.9%)	12,303 (4.2%)	12,825 (4.3%)
FSM** (%)	39,078 (32.4%)	35,784 (36.7%)	26,510 (34.8%)	33,379 (33.5%)	34,527 (34.5%)
ISM** (%)	52,173 (43.2%)	37,000 (38.0%)	36,959 (48.5%)	43,085 (43.3%)	39,060 (39.1%)
NIC** (%)	15,093 (12.5%)	12,808 (13.1%)	6,433 (8.4%)	12,117 (12.2%)	13,756 (13.8%)
NNC** (%)	13,599 (11.3%)	11,170 (11.5%)	5,925 (7.8%)	10,176 (10.2%)	11,887 (11.8%)
Other** (%)	715 (0.006%)	683 (0.007%)	405 (0.005%)	784 (0.008%)	766 (0.008%)

FSM transcripts mainly consist of transcripts having alternative 3' ends, while ISM transcripts have different 5' prime ends (Supplementary Figure 2). Calculation of the number of transcripts detected per gene across all samples showed that 28% of genes in the normal sample expressed more than ten transcripts. ccRCC3 had the smallest number of genes per cell expressing more than ten transcripts, with only one for 27% of genes. (Fig. 1E, Supplementary Fig. 1E for the transcripts found at least in 3 cells). Most gene transcripts were found to be expressed in only one cell across all samples (Fig. 1F). FSM and NIC transcripts tended to be the longest and to have similar lengths on average (t-test p-value=0.82) (Fig. 1G), while the ISM and NNC transcripts showed shorter lengths compared to FSM and NIC (t-test, p-value < 2.2e-16). The 'Other' category showed the shortest transcripts (t-test, p-value < 2.2e-16). ~50% of transcripts found in a single cell were novel, while the transcripts found in more than 150 cells were mostly FSM (Fig. 1H). To investigate whether the exon number affects the formation of new novel transcripts, we calculated the correlation between the number of exons and novel transcript numbers detected per gene. Interestingly, we observed a higher

correlation for ISM than for novel transcripts (Supplementary Figure 3B, R-value between 0.44 and 0.47, p<2.2e-16).

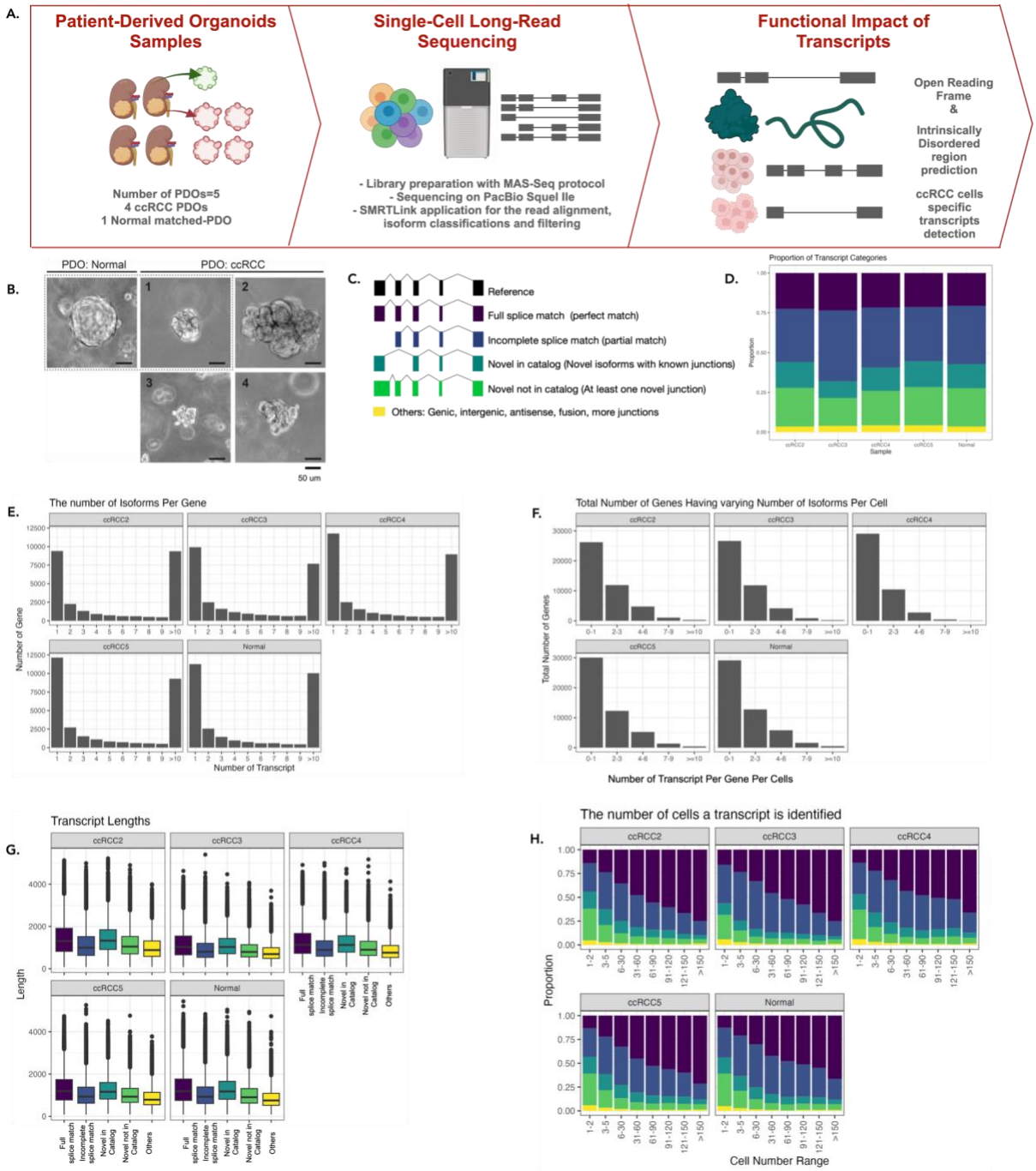

Fig 1. Transcript landscape and cell heterogeneity in normal and ccRCC-PDOs: (A) Schematic design of the project showing how patient-derived organoid (PDO) samples are established, sequenced using single-cell long-read sequencing, and functionally characterized (illustrations were created by Biorender). **(B)** The brightfield representative images of our organoids. The dotted line marks the matched pair. The scale bar is 50um. (C). SQANTI3 transcript categories. **(D).** The proportion of transcript categories found across four ccRCC-

PDO and one normal PDO (please see the (C) for the color code). **(E).** The number of identified transcripts per gene in each PDO. The x-axis denotes the number of transcripts per gene, categorized into bins (1, 2, 3, 4, 5, 6, 7, 8, 9, and >10), while the y-axis represents the number of genes. The height of each bar reflects the count of genes that express the corresponding number of transcripts. **(F).** The number of identified transcripts per gene per cell in each PDO. The x-axis shows the number of transcripts detected per gene per cell, categorized into different bins, while the y-axis denotes the total number of genes, with the height of each bar reflecting the count of genes that express the corresponding number of transcripts per cell. **(G).** Distribution of transcript lengths for each structural category across samples. **(H).** Proportional distribution of identified transcripts' structural categories across cell number ranges.

Filtering out transcripts identified in only one or two cells decreased the number of total sequenced transcripts (from 291,459 to 97,412 transcripts on average). Of the remaining transcripts, only 22.7% were novel compared to 37% prior to filtering (Supplementary Figure 1C). The highest UMI counts were found with FSM transcripts, while novel transcripts tend to have the lowest UMI (Supplementary Figure 1E, 1F, 1G). Altogether, these results show that the majority of novel transcripts are lowly expressed in one or two cells.
However, there are some exceptions, for example, the novel transcripts of the genes Glyceraldehyde 3-phosphate dehydrogenase (*GAPDH)*, Pyruvate kinase (*PKM)*, Aldolase A (*ALDOA)*, Angiopoietin-like 4 (*ANGPTL4),* Vimentin (*VIM)* (see Supplementary Table 1 for the full list) are expressed in at least 50% of ccRCC2, ccRCC4 or ccRCC5. Among those, *GAPDH*, *PKM*, and *ALDOA* are involved in glycolysis, and it is reported that enzymes having a role in glycolysis are upregulated in the occurrence of VHL-deficient ccRCC due to the upregulation in hypoxia-inducible factor 1alpha (HIF-1a) (Miranda-Poma et al. 2023). *ANGPTL4* is another hypoxia-inducible gene, and its expression has been shown as a potential diagnostic marker for ccRCC (Verine et al. 2010). *VIM* is a mesenchymal marker overexpressed in epithelial-to-mesenchymal transition (Landolt et al. 2017; Xu et al. 2020). Our findings suggest that those novel transcripts expressed more broadly across cells might play an important role in the pathogenesis of ccRCC.

Transcripts common to many cells have more translation capability

To assess whether the novel transcripts are protein coding, we predicted the Open Reading Frame (ORF) using TransDecoder (Haas BJ.). Based on the occurrence of start and stop codons and coding regions, TransDecoder assigned transcripts into varying sub-ORF categories, including 3' partial (transcripts with missing stop codons), 5' partial (transcripts with missing start codons), internal (transcripts that miss both start and stop codons), and complete transcripts (including all necessary parts to code a protein). For about 77.5% of the novel transcripts, we were able to predict an ORF (Fig. 2A). ISM transcripts had the lowest proportion of complete ORFs, which is a consequence of the lack of their terminal exons which might lead to truncated or completely missing ORFs (Fig. 2B). We then investigated the prevalence of sub-ORF categories across varying cell number ranges. Transcripts commonly expressed in a sample tend to show more complete ORFs as compared to cell-specific transcripts (Fig. 2C), independent of the transcript class (Pearson's chi-square test: $p<2.2e-16$). To understand whether the predicted protein isoforms form a stable protein structure that could hint towards a biological function, we predicted intrinsically disordered regions for all isoforms with

complete ORFs using iupred2 (Mészáros et al. 2018). The calculations demonstrated that ISM transcripts had the highest proportion of disordered residues (pairwise Wilcoxon rank sum test, (ISM-FSM & ISM-NIC, p-value: <2e-16, ISM-NNC, p-value=6.7e-11) (Fig. 2D) while novel transcripts with intron retention showed a higher disordered score than those with a new splice site or a new combination of a splice site/junction (Fig. 1E). For example, we identified six novel transcripts of Nicotinamide-N-methyltransferase (*NNMT*), each comprising three to four exons, and each with a complete ORF. The protein sequences encoded by these transcripts are characterized by more than 88% of their residues being ordered in the ccRCC2 PDO. NNMT was previously characterized as a promising drug target for ccRCC (Reustle et al. 2022). These transcripts were found to be expressed in a range of 3 to 250 cells. On the other hand, protein sequences of ADP Ribosylation Factor Like GTPase 6 Interacting Protein 4 (ARL6IP4) exhibited more than 90% of their residues as disordered. Our observations suggest that most novel transcripts are similarly disordered as their canonical counterparts.

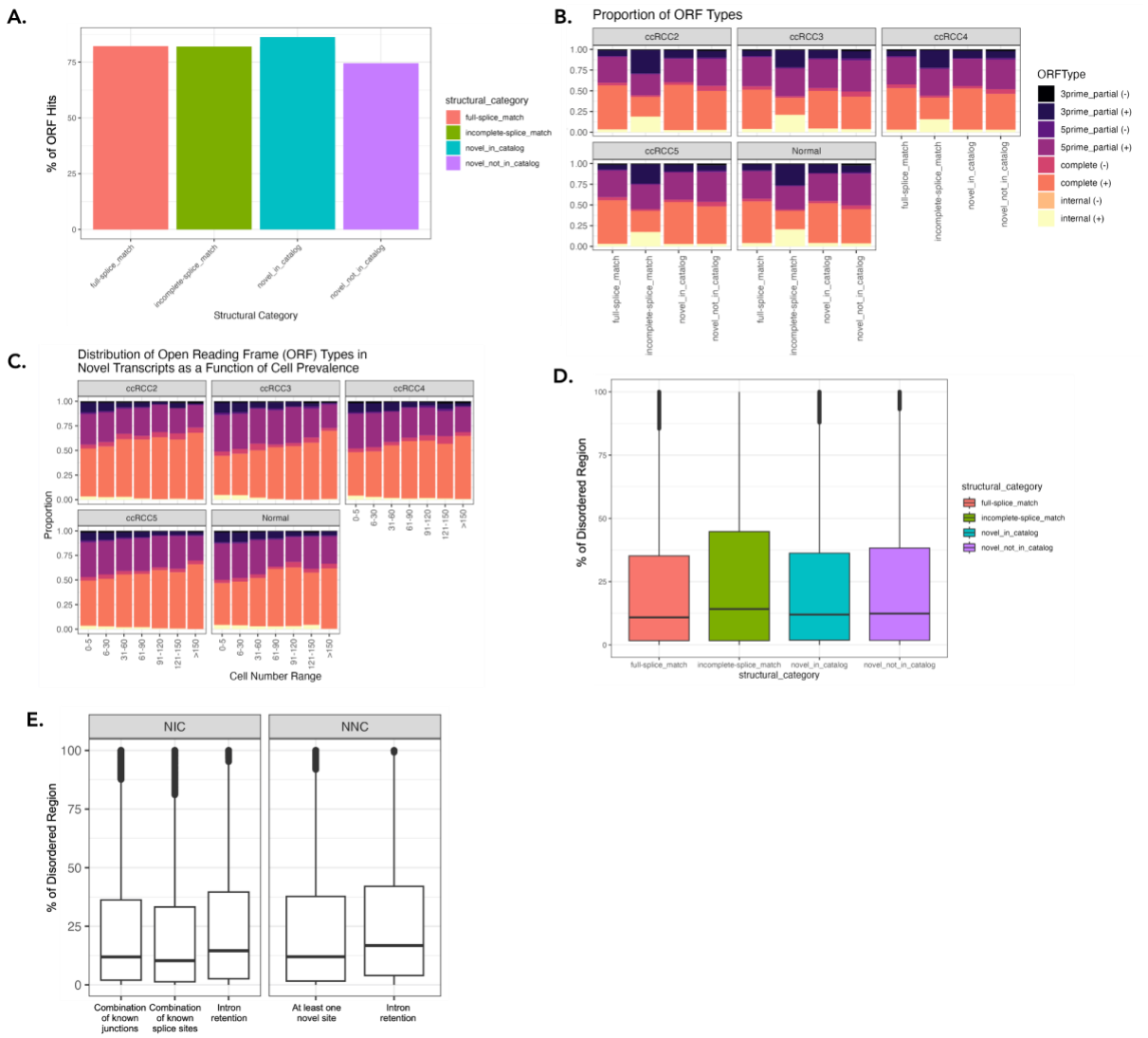

Fig 2: Distribution of open reading frame (ORF) categories and intrinsically disordered protein predictions: (A). Percentage of ORF hits across different structural categories in all datasets. **(B).** The fraction of different ORF types across datasets in each structural category. Each color represents various ORF types. **(C).** Distribution of ORF types in novel transcripts as a function of cell number range across datasets. The x-axis categorizes the cell number

range, while the y-axis shows the proportion of each ORF type (see (B) for the legends). **(D).** Comparison of disordered scores for the protein sequences of complete-ORF transcripts across structural categories. **(E).** Comparison of disordered scores for the protein sequence of NIC and NNC showing complete ORFs across different sub-structural categories.

ccRCC Cell-Specific Transcripts have a Higher Novel Transcript Proportion

Using the Seurat clustering algorithm, we identified six different cell populations among 2598 cells from one normal and four ccRCC-derived PDO (Fig. 3A). To evaluate the cell types in our samples, we examined the expression of ccRCC-specific markers (CA9, ANGPTL4) and kidney markers (EPCAM and PAX8). ccRCC markers were predominantly expressed in PDO cells from samples ccRCC2, ccRCC4, and ccRCC5 (Supplementary Figure 4C). Interestingly, kidney markers were expressed predominantly in the normal sample and in the PDO cells of ccRCC3 (Supplementary Figure 4B). The transcript expression profile of ccRCC3 stood out compared to the other ccRCC organoids. Looking closely at the *VHL* mutations (Table 1), ccRCC3 had a P25L variant. This variant was previously described as a polymorphic likely benign mutation, which could explain the VHL-negative-like expression profile of ccRCC3 (Rothberg 2001; Nickerson et al. 2008) lacking overexpression of the VHL-HIF pathway. We did not observe significant expression of endothelial markers (CDH5 and FLT1) in our PDO cells, indicating that organoids contained little or no stromal cells. As ccRCC originates from the proximal tubule (PTC), we also found that nearly all cells express PTC markers (GGT1, RIDA) (Supplementary Figure 4D).

Moreover, to explore the gene and transcript diversity between typical ccRCC and non-ccRCC cells, we categorized cells based on their CA9 expression. CA9 expression is a result of HIF up-regulation due to VHL inactivation. ccRCC2, ccRCC4, and ccRCC5 samples contained 361, 42, and 217 ccRCC cells, respectively (Figure 3C). Differential gene expression analysis between ccRCC and non-ccRCC cells revealed upregulation of several ccRCC-related genes in ccRCC cells, including NADH dehydrogenase 1 alpha subcomplex, 4-like 2 (*NDUFA4L2*), Lysyl oxidase (*LOX*), Vascular Endothelial Growth Factor A (*VEGFA*), ANGPTL4, Egl-9 Family Hypoxia Inducible Factor 3 (*EGLN3*) (Fig. 3D). Each of these genes is known to have a role in the progression of ccRCC through various mechanisms. NDUFA4L2 and EGLN3 are critical for the adaptation of ccRCC cells to hypoxic conditions (Wang et al. 2017a; Tamukong et al. 2022), VEGFA is a key factor for new blood vessel formations, essential for tumor metastasis, and LOX contributes to ccRCC progression by increasing the stiffness of the collagen matrix, which in turn, facilitates the cellular migration (Di Stefano et al. 2016).

Next, we explored the splicing diversity between ccRCC and non-ccRCC cells and found that ccRCC cell-specific transcripts show a high proportion of novel transcripts (Fig. 3E). 840 genes commonly found expressing ccRCC-cell specific novel transcripts in ccRCC2 and ccRCC5-PDOs (see Supplementary Figure 5 and Supplementary Table 2), and they were mostly associated with ccRCC-relevant pathways, including glycolysis and oxidative stress response (Fig. 3F). For example, more than 20 novel transcripts of the *NDUFA4L2* gene were found to be expressed in CA9+ cells in the ccRCC5 and ccRCC2 samples. The upregulation of novel

transcripts of these genes in ccRCC is yet another demonstration of the heterogeneity of the splicing landscape in ccRCC malignancies.

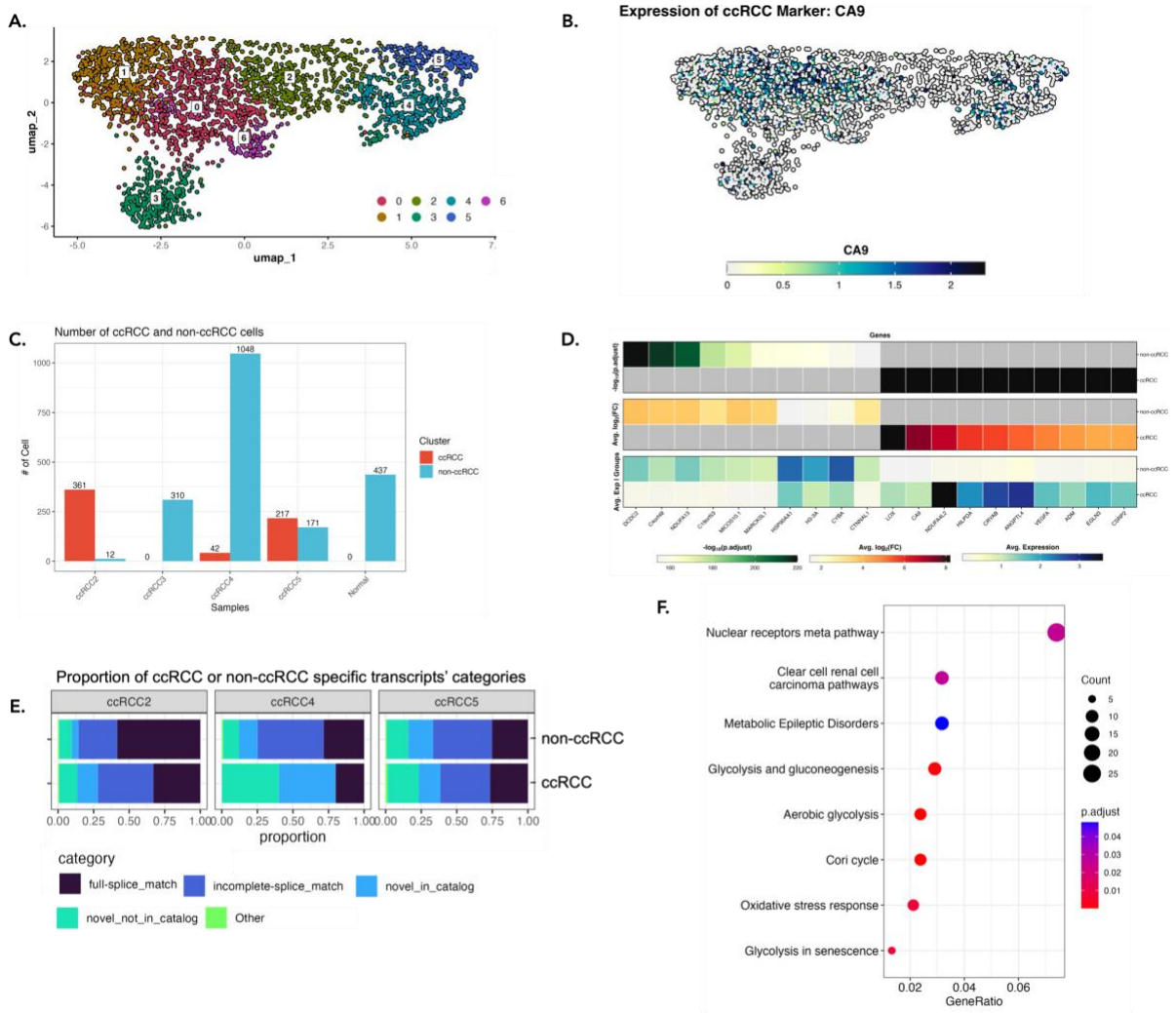

Fig 3: Categorizing cells as ccRCC and non-ccRCC in PDOs and the fraction of transcript categories. (A). UMAP plot of batch-corrected analysis of all datasets displaying each cluster by different colors. **(B).** UMAP plot of the ccRCC marker CA9 expression across cells, with darker colors indicating higher expression levels. **(C).** The table shows the number of ccRCC and non-ccRCC cells in each PDO categorized based on their CA9 expression. **(D).** The heatmap shows the differential gene expression between ccRCC and non-ccRCC cells. **(E).** The proportion of explicitly expressed transcripts' structural categories across ccRCC and non-ccRCC cells (bottom). **(F).** Over-representation analysis of genes expressing novel transcripts explicitly in ccRCC cells.

Matched Transcripts are widespread across cells

Next, we explored the number of overlapping transcripts to understand the intertumor heterogeneity of alternative splicing between patients. As the PacBio Iso-Seq pipeline assigns transcript IDs randomly, we matched the transcripts based on their exon-boundaries as described before by Healey et al. (Healey et al. 2022). Using the Tama tool, we could detect 15,939 common transcripts from 11,518 genes expressed in at least three cells of each

sample. 2,244 transcripts from 1,799 genes were found only in ccRCC2, ccRCC4, and ccRCC5 PDOs, not in the Normal and ccRCC3. (Fig 4A). Interestingly, we could see that common transcripts in all samples were expressed in more cells than transcripts unique to each sample (Figure 4B, Wilcoxon test, p-value < 2.2e-16). A comparison of the number of matched transcripts based on the Jaccard similarity index revealed the highest similarities between Normal:ccRCC5 and ccRCC2:ccRCC5 PDOs. In total, we identified 35-41% matched transcripts between ccRCC5:Normal (ccRCC5: 41,984/101,597, normal: 41,984/121,865) and ccRCC2:ccRCC5 (ccRCC2: 37,882/98,275, ccRCC5: 37,882/101,597) which was followed by ccRCC2:Normal (ccRCC2: 40,920/98,275, normal: 40,920/121,865). However, we observed the highest similarity of ISM transcripts between ccRCC2:Normal PDOs (Supplementary Figure 7). In addition, little similarity between ccRCC3 and the other ccRCC samples was detected, providing evidence that ccRCC3 is not a typical ccRCC sample and most likely resembles a VHL- phenotype.

The comparison of transcripts found explicitly in CA9+ or CA9- cells in each sample revealed 593 transcripts commonly detected only in ccRCC cells of ccRCC2 and ccRCC5 PDOs (Supplementary Fig 6). Of those, 251 are annotated as novel transcripts coded by 166 genes. Among those 251 novel transcripts, one of the frequently found novel transcripts belongs to Transmembrane protein 176A (*TMEM176A*), found in 186 and 13 ccRCC cells of ccRCC2 and ccRCC5 PDOs, respectively (ccRCC2 ID: PB.72551.11, ccRCC5 ID: PB.99360.3). This transcript is categorized as NIC and formed by a combination of known junctions comprising seven exons and has a complete ORF. Compared to the FSM transcripts, they differ in their 5' terminal region (Fig. 4C). TMEM176 has previously been shown to have a role as a tumor suppressor in esophageal squamous cell carcinoma in colorectal cancer, and its methylation has been defined as a prognostic biomarker (Gao et al. 2017; Wang et al. 2017b). In addition to this common novel transcript, the most frequently expressed transcript across cells are two distinct NNC transcripts (ccRCC2 ID: PB.72551.1, ccRCC5 ID: PB.99360.1) in both samples, and both transcripts have an additional exon at their 5' ends (Fig. 4C). ORF prediction of transcripts showed that both NIC and NNC transcripts express CD20-like family domain (PF04103). The occurrence of the NIC novel transcript, specifically in CA9+ cells, makes it a potential diagnostic biomarker candidate. In contrast, the NNC transcript is found across both CA9+ and CA9- cells, suggesting a potential contribution of transcripts to the total protein expression.

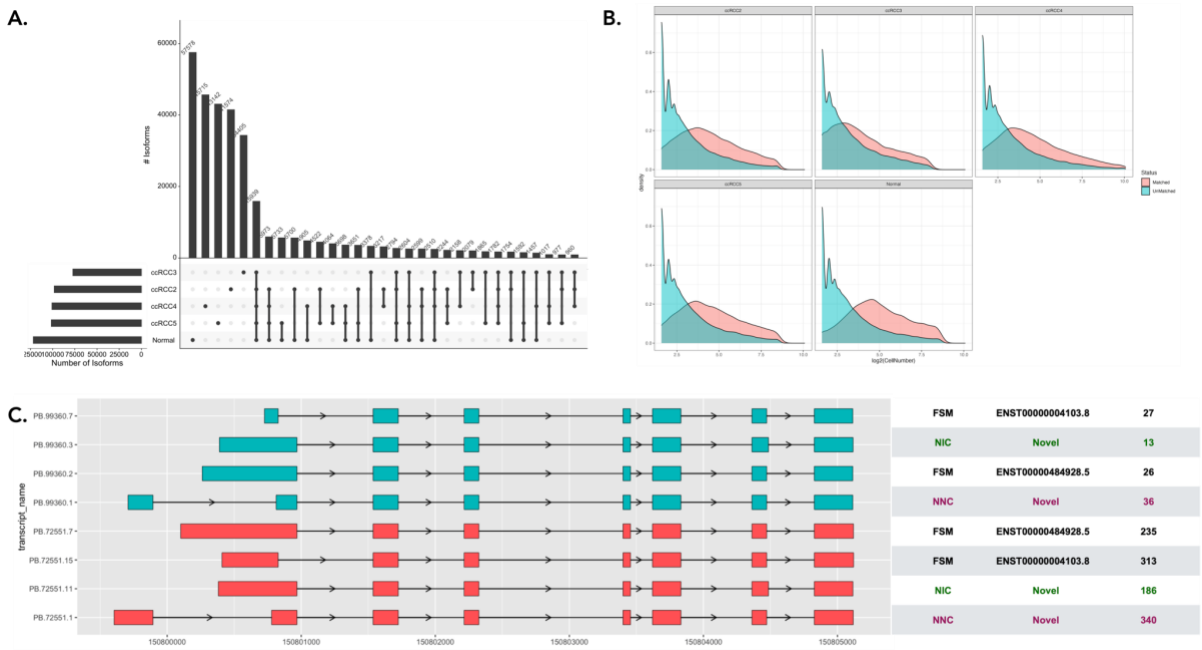

Fig 4: Overlapping transcripts among ccRCC and non-ccRCC cells and exon structures of TMEM176A transcripts: (A). The number of overlapping transcripts across samples. **(B)** Density plots show the distribution of the number of cell barcodes corresponding to overlapping (matched) and non-overlapping (unmatched) transcripts. The x-axis categorizes the transcripts into 'Matched' and 'Unmatched' groups, while the y-axis indicates the number of cell barcodes in which these transcripts are detected (Wilcoxon test, $p < 2.2e-16$). **(C).** Exon structure of four TMEM176A transcripts. Transcripts from ccRCC2 and ccRCC5 are depicted in red and blue, respectively. The table next to the transcript structures lists the SQANTI3 category of transcripts, aligned reference transcripts, and the number of cells in which the transcripts were identified. The most frequently observed transcripts in samples are highlighted in purple, and the common novel transcript is marked in green.

Most Dominant Transcripts Switching Events in ccRCC Cells

As alternatively spliced transcripts can have different exons, they may result in different protein domains, disrupt protein interactions, or form interaction with new protein partners. Previous research has shown that most protein-coding genes have one most dominant transcript (MDT) expressed at a significantly higher level than any other transcript of the same gene. These dominant transcripts can be tissue-specific (Ezkurdia et al. 2015; Gonzàlez-Porta et al. 2013; Tung et al. 2022). We previously demonstrated that these MDTs switch during malignant transition in cancer, including in ccRCC (Kahraman et al. 2020). To explore variations in MDT profiles between ccRCC and non-ccRCC cells, we analyzed MDT switches between ccRCC and non-ccRCC cells across three PDOs, ccRCC2, ccRCC4, and ccRCC5. In total, we identified 1,450 unique cancer-specific MDTs in 547 single cells (Supplementary Table 3), ranging between 1 and 26 switches per cell (Fig. 5A). 549 of the cancer-specific MDTs were found to have a complete Open Reading Frame (ORF), of which 96 were found in only one cell as a cancer-specific MDT. Among MDTs with complete ORF, 47 genes with MDT switch were found in both ccRCC2 and ccRCC5 PDOs (Fig. 5B). Over-representation analysis of the genes revealed functional roles in mRNA-splicing pathways, respiratory electron transport, and G2/M transitions. One of the cancer-specific MDT that was expressed in 135 cells of

ccRCC2 was MYL6 (Fig. 5D). The *MYL6* gene encodes the myosin light chain 6 protein, which is a component of the Myosin ATPase cellular motor protein complex. Both transcripts of MYL6 (ccRCC MDT: PB.105063.81, non-ccRCC MDT: PB.105063.80) have been annotated as full-splice matches having alternative 5' end with SQANTI3. While the ccRCC MDT aligns to ENST00000550697.6, non-ccRCC MDT maps to ENST00000547649.5. Both transcripts translate to 151 amino acids long proteins but with distinct C-terminal sequences. Interestingly, Ensembl (v.111) annotates the ccRCC MDT as canonical), while the non-canonical non-ccRCC MDT is the top-ranked transcript in the Kidney Corext and Medulla tissue on GTEx (https://www.gtexportal.org). The difference between transcript exon structures is that the ccRCC MDT transcript has one additional exon (exon 6) (Fig. 5E), making it the gene's longest transcript. Exon inclusion and exclusion of MYL6 have been previously shown to be cell type-specific, with exon inclusion observed more in the muscle tissue than non-muscle tissue (Olivieri et al. 2021). In contrast to our finding in ccRCC CA9+ cells, exon skipping has been shown to increase in metastatic pancreatic cancer samples (Jbara et al. 2023).

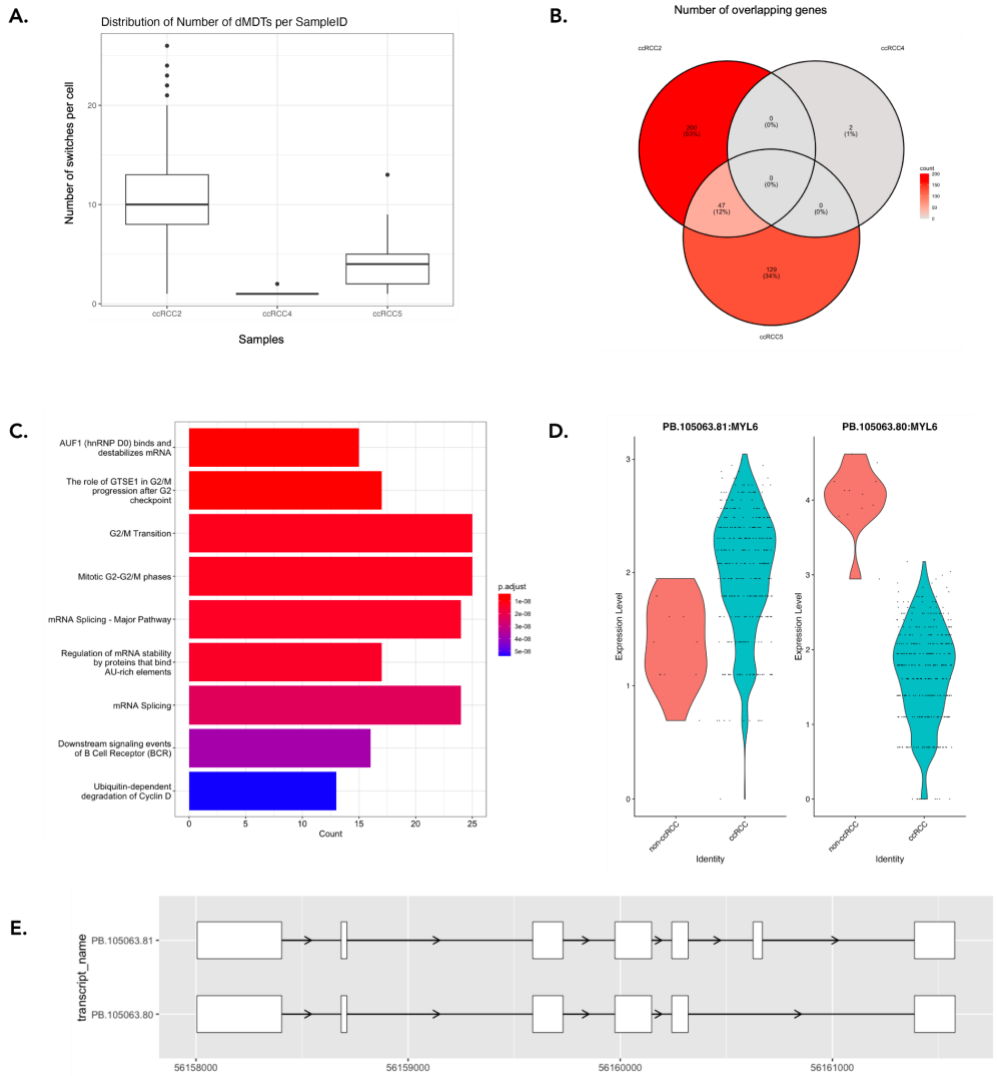

Fig 5: MDT Switches between ccRCC and non-ccRCC cells: (A). Distribution of number of switches in ccRCC2, ccRCC4, and ccRCC5 PDOs. **(B)** The number of overlapping genes showing transcript switching events across three datasets. **(C).** Reactome pathway over-representation analysis of genes commonly showing transcript switching events in ccRCC2

and ccRCC5 PDO datasets. **(D).** Normalized expression profile of two MYL6 transcripts showing different MDTs in ccRCC and non-ccRCC cells. PB.105063.81 is ccRCC MDT, while PB.105063.80 represents the non-ccRCC MDT. **(E).** Exon structures of ccRCC (PB.105063.81) and non-ccRCC (PB.105063.80) MDTs.

Discussion

The recent advent of single-cell long-read sequencing technologies provides a unique opportunity to gain insight into intra- and inter-tumor heterogeneity of tumors and to discover potential novel predictive biomarkers. To reveal the heterogeneity in ccRCC, we utilised the MAS-Seq single-cell long-read sequencing protocol of PacBio. We generated a comprehensive catalogue of known and novel transcripts for one normal and four ccRCC Patient-Derived Organoids (PDOs) without employing short-read single-cell sequencing data. PDOs with the highest number of sequenced cells, had, as expected, the least number of detected transcripts per gene per cell. However, sequencing a low number of cells might also cause a loss the essential cell diversity in the samples.

Here, for the first time, we uncovered over 300,000 unique isoforms across samples, of which 27% are novel transcripts with new combinations of known exons or new junctions. To interpret the biological impact of transcripts sequenced, we investigated the prevalence across cells together with their protein-coding capability. Our analysis revealed that, on average, 66% of identified novel transcripts were found only in a few cells, suggesting they might be artifacts with limited biological relevance. In contrast, widely expressed transcripts found corresponded mainly to known transcripts. In addition, frequently identified known and novel transcripts had more complete open reading frames, underscoring their protein-coding capability.

Even though the highest proportion of transcripts in our data was found to be ISM transcripts, they showed the least fraction of complete ORFs and the highest disordered score for complete ORFs. This finding suggests that proteins encoded by these transcripts may exhibit enhanced functional diversity or regulatory capacity due to their lack of stable protein structure. To understand the splicing diversity between ccRCC and non-ccRCC cells, we investigated explicitly expressed transcripts in each category, revealing that the ccRCC cells have more unique novel isoforms having a role in ccRCC-related pathways, including oxidative stress, and glycolysis, proposing functional contribution of those novel transcripts to cancer progression in ccRCC. Our most dominant switch analysis between ccRCC and non-ccRCC cells showed that many switching events are cell-specific. Nevertheless, genes showing switching events have a main role in mRNA-splicing pathways, highlighting that these transcripts are pivotal in regulating cell function through alternative splicing. The detected transcripts in long-read sequencing data must be validated using proteomics since the transcript expression is known to be buffered before protein expression. However, as Miller et al. show, long-read data can also be exploited during the validation process to detect new protein isoforms. The authors constructed a protein reference database based on the full-length transcript sequences. They used the reference database to search for matched mass-spectrometry-based proteomics data. They were able to confirm novel peptide sequences in the proteomics data as well as translated intronic sequences. The total number of these identifications was low but highlighted the possibility of transcript translations commonly ignored or overseen in classical proteomics experiments.

Some of our results might be due to artefacts in PCR amplification which is an essential part of the MAS-ISO-seq protocol. However, a recent study by Lee *et al,* who performed long-read sequencing on PCR amplified cDNA molecules and direct RNA sequencing using Oxford Nanopore sequencing, demonstrated a good overlap between differentially expressed genes using both complementary methods (Lee et al. 2023). Additional problems can arise when

equivalent transcripts are merged under a single transcript ID. Iso-Seq uses exon-boundaries and sequence similarity to determine the equivalency of transcripts. However, Iso-Seq is meant to work only on single samples, why the identification of equivalent transcripts over multiple samples is challenging. The merging over multiple samples is however important for studying the intertumor heterogeneity of tumor samples, where equivalent transcripts must be compared over multiple samples. We have used Tama's merge function for our study, however, more in depth analysis and benchmarking are required for this crucial step.

In conclusion, single-cell long-read sequencing of patient-derived organoids offers an unprecedented and detailed view of the transcriptome landscape of individual cancer patients. It reveals hundreds of thousands of novel transcripts, of which only the minority are commonly expressed in single and multiple patients, highlighting the intra- and intertumor heterogeneity of ccRCC. The discovery of frequently found novel transcripts provides insights into cancer progression and a new avenue for discovering potential novel biomarkers or therapeutic targets. The functional role of the commonly expressed novel transcripts remains to be further explored and validated.

Methods

Generation and Characterization of ccRCC Patient-Derived Organoid Samples

Patient tissue samples were provided by the Department of Pathology and Molecular Pathology at University Hospital Zürich. They were collected and biobanked according to previously described procedures (Bolck et al. 2019). The study was approved by the local Ethics Committee (BASEC# 201 9-01 959) and in agreement with Swiss law (Swiss Human Research Act). All patients gave written consent. Organoids were established as previously described (Bolck et al. 2021). Surgically resected renal tissue was reviewed by a pathologist with specialisation in uropathology (Holger Moch) and suitable specimens were stored at 4 °C in transport media (RPMI (Gibco) with 10 % fetal calf serum (FCS, Gibco) and Antibiotic-Antimycotic® (Gibco)). For organoid derivation, tissue specimens were further processed within 24 hours by rinsing them once with PBS, finely cutting and digesting them in 0.025 mg/ml Liberase (Roche) for 15 min at 37 °C. The slurry was then passed through a 100 µm cell strainer and centrifuged at 1000 rpm for 5 min. Cells were washed once with PBS and erythrocytes were lysed in ACK buffer (150 mM NH_4Cl, 10 mM $KHCO_3$, 100 mM EDTA) for 2 min at room temperature. After a final wash with PBS, appropriate amounts of cell suspension were resuspended in CK3D medium (Advanced DMEM/F12 (Gibco) with 1X Glutamax (Gibco), 10 mM HEPES (Sigma-Aldrich), 1.5X B27 supplement (Gibco), Antibiotic-Antimycotic (Gibco), 1 mM N-Acetylcysteine (Sigma-Aldrich), 50 ng/mL Human Recombinant EGF (Sigma-Aldrich), 100 ng/mL Human Recombinant FGF-10 (Peprotech), 1 mM A-83-01 (Sigma-Aldrich), 10 mM Nicotinamide (Sigma-Aldrich), 100 nM Hydrocortisone (HC, Sigma-Aldrich), 0.5 mg/ml epinephrine (Sigma-Aldrich), 4 pg/mL Triiodo-L-thyronine (T3, Promocell), R-Spondin (conditioned media, self-made) and mixed with a two volumes of growth factor reduced Matrigel (Corning). Drops of cell suspension/Matrigel were distributed in a 6-well low attachment cell culture plate (Sarstedt) and allowed to solidify for 30 min at 37 °C, upon which

CK3D media was added to cover the drops. To evaluate the growth of PDOs, bright-field images were captured using a microscope. Organoids at approximately 100-500 um were passaged, and at least 10,000 cells were collected for cell model validation using targeted DNA sequencing of the *VHL* gene. To achieve this, DNA was isolated using the Maxwell® 16 DNA Purification Kit (Promega) and corresponding Maxwell instrument. PCR and sequencing of *VHL* were performed as previously described (Rechsteiner et al. 2011).

Full-length single-cell isoform sequencing and data processing of PDO cells via MAS-ISO-Seq

To obtain single cell suspension, cell culture media was removed and PDOs from one well of a ULA 6-well plate were collected in ice-cold Cell Recovery Solution (Corning) and incubated for 1 hour at 4°C to resolve the Matrigel. Subsequently, PDOs were dissociated with TrypLE by incubation on a thermal shaker set to 37 °C, 300 rpm. Every 2 min, the samples were picked up and mechanically dissociated by pipetting up and down and the progress of dissociation was evaluated under a microscope using a small fraction of the cells and tryphan blue. After dissociation, PBS supplemented with 20 % FBS, was added to stop the reaction. Samples were centrifuged at 1000 *g* for 5 min and the supernatant was aspirated. The pellet was washed once in 1X PBS with 0.04 % BSA and filtered through a 70 µm strainer. Finally, cells were counted and diluted to the target cell concentration using PBS with 0.04 % BSA.

Generation of full-length cDNA with 10x Genomics platform and PacBio MAS-Seq library preparation and sequencing

10x Genomics Chromium platform was used to analyze the dissociated organoid cells (Zheng et al. 2017). Library preparation was conducted following the 10x Genomics Single Cell 3' Reagent Kits v3.1 (Dual Index) User Guide. We targeted to recover 700 cells per library preparation to have a greater sequencing depth using the PacBio platform. The single-cell full-length cDNAs were directed for single-cell MAS-Seq (Multiplexed Arrays Sequencing) library preparation using the MAS-Seq 10x Single Cell 3' kit (Pacific Bioscience, CA, USA). Each single-cell MAS-seq library was used to prepare the sequencing DNA-Polymerase complex and further sequenced on a single 8M SMRT cell (Pacific Bioscience), on Sequel IIe sequencer (Pacific Bioscience) yielding in ~ 2 M HiFi reads and~ 30M segmented reads per sample. The method is described in detail in Zajac N. et al., (accompanying submission).

SMRTLink Iso-Seq pipeline

In our study, we utilized the "Read Segmentation and Iso-Seq workflow" from SMRTLink version 11.1 to process our long-read sequencing data. For two specific samples, Normal and ccRCC2, we combined the data from three SMRTcells to enhance coverage. HiFi reads were converted into segmented reads using the skera tool, followed by the IsoSeq protocol in SMRTLink (Zajac N. et al., (accompanying submission)). We aligned the reads to the human genome (GRCh38.p13) using pbmm2 and analyzed them with the GenomicAlignments R package, ensuring compatibility with selected Illumina cell barcodes (Lawrence et al. 2013). Piegon was used to identify and filter the unique isoforms to remove various artifacts. The gene and isoform count matrices were generated with the make-seurat function.

Full-length single-cell data analysis

Transcript Types and Their Prevalence Across Cells

The transcripts were categorized into structural categories by SQANTI3 in SMRT-Link. We calculated the percentage of each transcript structural category and their length in each sample using scisoseq_classification.filtered_lite_classification.txt files. We then checked transcript prevalence across varying cell number ranges, and the number of transcripts per gene and cell.

Functional Annotation of Long-read Sequencing Transcripts

Open Reading Frames (ORFs) were identified on long-read transcript sequences listed in fasta files from the Iso-Seq collapse function using Transdecoder v5.7.1 (Haas BJ.). Transdecoder.LongOrfs function was used to predict all possible ORFs with a minimum ORF length of 100 nucleotides. To calculate protein sequences from the predicted ORFs, an extensive human reference database containing 226,259 canonical and alternatively spliced isoform protein sequences was generated using Uniprot (release date: 2023-11). The predicted ORFs were aligned to this database via blastp, setting the e-value to 1e-5. In addition, hmmscan v3.4 was applied to predict potential Pfam domains using the Pfam database (release date: 2023-09-12) with a maximum e-value of 1e-10. The results from both hmmscan and blastp were used to predict the final ORFs using Transdecoder.Predict function. We then selected one ORF for each transcript based on the highest score assigned by TransDecoder. We applied iupred2a on the transcripts having complete ORFs to predict their intrinsically disordered regions (IDRs). A residue was annotated as ordered or disordered, if its iupred2a score was below or above 0.5, respectively. We calculated the percentage of disordered residues for each transcript and assigned a percentage disordered score for each transcript.

Correlation between Number of Exons, Gene Length, and Number of Transcripts

To calculate the correlation between the number of exons, gene length, and number of FSM, ISM, and novel transcript per gene identified in each sample, we downloaded information on genes, transcripts, gene lengths, and exons from Biomart, Ensembl v110 by selecting Gene stable ID, Transcript stable ID, Gene name, and Exon stable ID fields. The correlation of number of exons and transcripts were calculated in R.

Cell Type Annotations

Seurat (Hao et al. 2024) was used for quality control and integration of the samples using the output files of the Iso-Seq make-seurat function. For gene-level analysis, each sample was normalized by the SCTransform function. 3000 features were selected using SelectIntegrationFeatures, and anchors for integration were identified with FindIntegrationAnchors. The samples were integrated with the IntegrateData function using the SCT normalization. Subsequently, the Seurat object was scaled with the ScaleData function and PCA, and UMAP analyses were performed using the RunPCA and RunUMAP functions, respectively. Markers for each cluster were defined with the PrepSCTFindMarkers and FindAllMarkers functions. To categorize the cells in each PDO, we analyzed the samples separately. SCT normalized gene expression matrices were scaled, and the cells were categorized into two categories using the scGate R package (Andreatta et al. 2022) by defining the CA9 as a ccRCC positive marker. The other cells were assigned as non-ccRCC. The genes expressing distinct transcripts in ccRCC cells of ccRCC2 and ccRCC5 were analyzed

with ClusterProfiler's enrichWP function for overrepresentation analysis. We used SCpubr R package to visualize marker expressions and clusters (Blanco-Carmona 2022).

Transcript Matching among Samples

Due to Iso-seq assigning transcript IDs randomly, we first converted all *sqanti_classification.filtered_lite.gff* files to the bed format using bedparse gtf2bed function (Healey et al. 2022). The columns were modified to include gene ID separated by transcript id by semicolon. Tama's tama_merge.py function was used to combine all transcript ids among samples using their exon and junction coordinates, allowing for 50 and 100 nucleotide flexibilities at the 5' and 3' ends, and 5 nucleotides flexibility at the exon junctions. The similarities of the samples were calculated in R using the Jaccard similarity matrix, i.e. the number of overlapping transcript IDs divided by the total number of transcripts found in two samples. The heatmaps were visualized using the pheatmap function in R, and the number of overlapping transcripts was plotted by UpsetR's upset function (Conway et al. 2017).

Most Dominant Transcripts Switches between ccRCC and non-ccRCC cells

To assess the Most dominant Transcripts (MDTs), we have used transcript UMI counts in each sample. Each MDT was required to have at least two times higher UMI counts than the second most abundant transcript (Kahraman et al. 2020). The MDT identifications were compared between ccRCC and non-ccRCC cells in the ccRCC2, ccRCC4, and ccRCC4 PDO samples based on the following strict rules:

- MDT in a ccRCC cell is not found as MDT in any non-ccRCC cell.
- At For at least 50% of non-ccRCC cells, a distinct MDT should be identified
- UMI counts of MDTs in ccRCC cells should be higher than the mean of the MDTs UMI count in non-ccRCC cells.

An MDT switch event was called, only when an MDT fulfilled all criteria. ClusterProfiler's enrichPathway function was used for the overrepresentation analysis of genes showing MDT switches between ccRCC2 and ccRCC5 PDOs. ggVennDiagram R package was used to generate a Venn diagram of overlapping cancer-specific MDTs among samples (Gao et al. 2021), and exons structures of the transcripts were generated with ggtranscript R package (Gustavsson et al. 2022).

Data Access

The raw single-cell long-read RNA sequencing data will be available at ENA under the accession number PRJEB73513.
The files and the codes used in the manuscript can be found at: https://github.com/KarakulakTulay/ccRCC_scLongRead

Funding

This work was funded by Krebsliga Zurich.

Competing Interest Statement

None declared.

Acknowledgments

We would like to thank Mark Robinson from UZH for valuable discussions about single-cell data analysis. We also would like to thank Harini Lakshminarayanan and Adriana Von Teichmann from the Department of Pathology and Molecular Pathology of USZ for their technical assistance during the preparation of PDO samples for the single-cell analysis.

References

Al'Khafaji AM, Smith JT, Garimella KV, Babadi M, Popic V, Sade-Feldman M, Gatzen M, Sarkizova S, Schwartz MA, Blaum EM, et al. 2023. High-throughput RNA isoform sequencing using programmed cDNA concatenation. *Nat Biotechnol*. https://www.nature.com/articles/s41587-023-01815-7 (Accessed October 24, 2023).

Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. 2020. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* **21**: 30.

Andreatta M, Berenstein AJ, Carmona SJ. 2022. scGate: marker-based purification of cell types from heterogeneous single-cell RNA-seq datasets ed. A. Mathelier. *Bioinformatics* **38**: 2642–2644.

Blanco-Carmona E. 2022. *Generating publication ready visualizations for Single Cell transcriptomics using SCpubr*. Bioinformatics http://biorxiv.org/lookup/doi/10.1101/2022.02.28.482303 (Accessed February 27, 2024).

Bolck HA, Corrò C, Kahraman A, Von Teichman A, Toussaint NC, Kuipers J, Chiovaro F, Koelzer VH, Pauli C, Moritz W, et al. 2021. Tracing Clonal Dynamics Reveals that Two- and Three-dimensional Patient-derived Cell Models Capture Tumor Heterogeneity of Clear Cell Renal Cell Carcinoma. *European Urology Focus* **7**: 152–162.

Bolck HA, Pauli C, Göbel E, Mühlbauer K, Dettwiler S, Moch H, Schraml P. 2019. Cancer Sample Biobanking at the Next Level: Combining Tissue With Living Cell Repositories to Promote Precision Medicine. *Front Cell Dev Biol* **7**: 246.

Bolisetty MT, Rajadinakaran G, Graveley BR. 2015. Determining exon connectivity in complex mRNAs by nanopore sequencing. *Genome Biol* **16**: 204.

Borcherding N, Vishwakarma A, Voigt AP, Bellizzi A, Kaplan J, Nepple K, Salem AK, Jenkins RW, Zakharia Y, Zhang W. 2021. Mapping the immune environment in clear cell renal carcinoma by single-cell genomics. *Commun Biol* **4**: 122.

Byrne A, Beaudin AE, Olsen HE, Jain M, Cole C, Palmer T, DuBois RM, Forsberg EC, Akeson M, Vollmers C. 2017. Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat Commun* **8**: 16027.

Conway JR, Lex A, Gehlenborg N. 2017. UpSetR: an R package for the visualization of intersecting sets and their properties ed. J. Hancock. *Bioinformatics* **33**: 2938–2940.

Di Stefano V, Torsello B, Bianchi C, Cifola I, Mangano E, Bovo G, Cassina V, De Marco S, Corti R, Meregalli C, et al. 2016. Major Action of Endogenous Lysyl Oxidase in Clear Cell Renal Cell Carcinoma Progression and Collagen Stiffness Revealed by Primary Cell Cultures. *The American Journal of Pathology* **186**: 2473–2485.

Dondi A, Lischetti U, Jacob F, Singer F, Borgsmüller N, Coelho R, Tumor Profiler Consortium, Aebersold R, Ak M, Al-Quaddoomi FS, et al. 2023. Detection of isoforms and genomic alterations by high-throughput full-length single-cell RNA sequencing in ovarian cancer. *Nat Commun* **14**: 7780.

Ezkurdia I, Rodriguez JM, Carrillo-de Santa Pau E, Vázquez J, Valencia A, Tress ML. 2015. Most Highly Expressed Protein-Coding Genes Have a Single Dominant Isoform. *J Proteome Res* **14**: 1880–1887.

Gao C-H, Yu G, Cai P. 2021. ggVennDiagram: An Intuitive, Easy-to-Use, and Highly Customizable R Package to Generate Venn Diagram. *Front Genet* **12**: 706907.

Gao D, Han Y, Yang Y, Herman JG, Linghu E, Zhan Q, Fuks F, Lu ZJ, Guo M. 2017. Methylation of *TMEM176A* is an independent prognostic marker and is involved in human colorectal cancer development. *Epigenetics* **12**: 575–583.

Gonzàlez-Porta M, Frankish A, Rung J, Harrow J, Brazma A. 2013. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biol* **14**: R70.

Gustavsson EK, Zhang D, Reynolds RH, Garcia-Ruiz S, Ryten M. 2022. *ggtranscript* : an R package for the visualization and interpretation of transcript isoforms using *ggplot2* ed. A. Mathelier. *Bioinformatics* **38**: 3844–3846.

Haas BJ. Haas, BJ. Transdecoder v5.7.1. https://github.com/TransDecoder/TransDecoder.

Hao Y, Stuart T, Kowalski MH, Choudhary S, Hoffman P, Hartman A, Srivastava A, Molla G, Madad S, Fernandez-Granda C, et al. 2024. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* **42**: 293–304.

Healey HM, Bassham S, Cresko WA. 2022. Single-cell Iso-Sequencing enables rapid genome annotation for scRNAseq analysis ed. A. Sanchez Alvarado. *Genetics* **220**: iyac017.

Jbara A, Lin K-T, Stossel C, Siegfried Z, Shqerat H, Amar-Schwartz A, Elyada E, Mogilevsky M, Raitses-Gurevich M, Johnson JL, et al. 2023. RBFOX2 modulates a metastatic signature of alternative splicing in pancreatic cancer. *Nature* **617**: 147–153.

Kahraman A, Karakulak T, Szklarczyk D, Von Mering C. 2020. Pathogenic impact of transcript isoform switching in 1,209 cancer samples covering 27 cancer types using an isoform-specific interaction network. *Sci Rep* **10**: 14453.

Krishna C, DiNatale RG, Kuo F, Srivastava RM, Vuong L, Chowell D, Gupta S, Vanderbilt C, Purohit TA, Liu M, et al. 2021. Single-cell sequencing links multiregional immune landscapes and tissue-resident T cells in ccRCC to tumor topology and therapy efficacy. *Cancer Cell* **39**: 662-677.e6.

Landolt L, Eikrem Ø, Strauss P, Scherer A, Lovett DH, Beisland C, Finne K, Osman T, Ibrahim MM, Gausdal G, et al. 2017. Clear Cell Renal Cell Carcinoma is linked to Epithelial-to-Mesenchymal Transition and to Fibrosis. *Physiol Rep* **5**: e13305.

Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for Computing and Annotating Genomic Ranges ed. A. Prlic. *PLoS Comput Biol* **9**: e1003118.

Lee J, Snell EA, Brown J, Banks RE, Turner DJ, Vasudev NS, Lagos D. 2023. *Long-read RNA sequencing redefines the clear cell renal cell carcinoma transcriptome and reveals novel genes and transcripts associated with disease recurrence and immune evasion*. Oncology http://medrxiv.org/lookup/doi/10.1101/2023.09.08.23295204 (Accessed March 14, 2024).

Mészáros B, Erdős G, Dosztányi Z. 2018. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Research* **46**: W329–W337.

Miranda-Poma J, Trilla-Fuertes L, López-Vacas R, López-Camacho E, García-Fernández E, Pertejo A, Lumbreras-Herrera MI, Zapater-Moros A, Díaz-Almirón M, Dittmann A, et al. 2023. Proteomics Characterization of Clear Cell Renal Cell Carcinoma. *JCM* **12**: 384.

Motzer RJ, Jonasch E, Agarwal N, Alva A, Baine M, Beckermann K, Carlo MI, Choueiri TK, Costello BA, Derweesh IH, et al. 2022. Kidney Cancer, Version 3.2022, NCCN Clinical Practice Guidelines in Oncology. *Journal of the National Comprehensive Cancer Network* **20**: 71–90.

Nickerson ML, Jaeger E, Shi Y, Durocher JA, Mahurkar S, Zaridze D, Matveev V, Janout V, Kollarova H, Bencko V, et al. 2008. Improved Identification of von Hippel-Lindau Gene Alterations in Clear Cell Renal Tumors. *Clinical Cancer Research* **14**: 4726–4734.

Olivieri JE, Dehghannasiri R, Wang PL, Jang S, De Morree A, Tan SY, Ming J, Ruohao Wu A, Tabula Sapiens Consortium, Quake SR, et al. 2021. RNA splicing programs define tissue compartments and cell types at single-cell resolution. *eLife* **10**: e70692.

Pardo-Palacios FJ, Arzalluz-Luque A, Kondratova L, Salguero P, Mestre-Tomás J, Amorín R, Estevan-Morió E, Liu T, Nanni A, McIntyre L, et al. 2023. *SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms*. Bioinformatics http://biorxiv.org/lookup/doi/10.1101/2023.05.17.541248 (Accessed October 24, 2023).

Rechsteiner MP, Von Teichman A, Nowicka A, Sulser T, Schraml P, Moch H. 2011. *VHL* Gene Mutations and Their Effects on Hypoxia Inducible Factor HIFα: Identification of Potential Driver and Passenger Mutations. *Cancer Research* **71**: 5500–5511.

Reustle A, Menig L, Leuthold P, Hofmann U, Stühler V, Schmees C, Becker M, Haag M, Klumpp V, Winter S, et al. 2022. Nicotinamide-N-methyltransferase is a promising metabolic drug target for primary and metastatic clear cell renal cell carcinoma. *Clinical & Translational Med* **12**: e883.

Rothberg P. 2001. Is the P25L a “Real” VHL mutation? *Molecular Diagnosis* **6**: 49–54.

Schreibing F, Kramann R. 2022. Mapping the human kidney using single-cell genomics. *Nat Rev Nephrol* **18**: 347–360.

Shiau C-K, Lu L, Kieser R, Fukumura K, Pan T, Lin H-Y, Yang J, Tong EL, Lee G, Yan Y, et al. 2023. High throughput single cell long-read sequencing analyses of same-cell genotypes and phenotypes in human tumors. *Nat Commun* **14**: 4124.

Simmler P, Cortijo C, Koch LM, Galliker P, Angori S, Bolck HA, Mueller C, Vukolic A, Mirtschink P, Christinat Y, et al. 2022. SF3B1 facilitates HIF1-signaling and promotes malignancy in pancreatic cancer. *Cell Reports* **40**: 111266.

Tamukong PK, Kuhlmann P, You S, Su S, Wang Y, Yoon S, Gong J, Figlin RA, Janes JL, Freedland SJ, et al. 2022. Hypoxia-inducible factor pathway genes predict survival in metastatic clear cell renal cell carcinoma. *Urologic Oncology: Seminars and Original Investigations* **40**: 495.e1-495.e10.

Tian L, Jabbari JS, Thijssen R, Gouil Q, Amarasinghe SL, Voogd O, Kariyawasam H, Du MRM, Schuster J, Wang C, et al. 2021. Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing. *Genome Biol* **22**: 310.

Tung K-F, Pan C-Y, Lin W. 2022. Dominant transcript expression profiles of human protein-coding genes interrogated with GTEx dataset. *Sci Rep* **12**: 6969.

Verine J, Lehmann-Che J, Soliman H, Feugeas J-P, Vidal J-S, Mongiat-Artus P, Belhadj S, Philippe J, Lesage M, Wittmer E, et al. 2010. Determination of Angptl4 mRNA as a Diagnostic Marker of Primary and Metastatic Clear Cell Renal-Cell Carcinoma ed. E.W. Steyerberg. *PLoS ONE* **5**: e10421.

Wan Y, Anastasakis DG, Rodriguez J, Palangat M, Gudla P, Zaki G, Tandon M, Pegoraro G, Chow CC, Hafner M, et al. 2021. Dynamic imaging of nascent RNA reveals general principles of transcription dynamics and stochastic splice site selection. *Cell* **184**: 2878-2895.e20.

Wang L, Peng Z, Wang K, Qi Y, Yang Y, Zhang Y, An X, Luo S, Zheng J. 2017a. NDUFA4L2 is associated with clear cell renal cell carcinoma malignancy and is regulated by ELK1. *PeerJ* **5**: e4065.

Wang Y, Zhang Y, Herman JG, Linghu E, Guo M. 2017b. Epigenetic silencing of TMEM176A promotes esophageal squamous cell cancer development. *Oncotarget* **8**: 70035–70048.

Wang Z, Zhu L, Li K, Sun Y, Giamas G, Stebbing J, Peng L, Yu Z. 2022. Alternative splicing events in tumor immune infiltration in renal clear cell carcinomas. *Cancer Gene Ther* **29**: 1418–1428.

Xu H, Xu W-H, Ren F, Wang J, Wang H-K, Cao D-L, Shi G-H, Qu Y-Y, Zhang H-L, Ye D-W. 2020. Prognostic value of epithelial-mesenchymal transition markers in clear cell renal cell carcinoma. *Aging* **12**: 866–883.

Yang Y, Yang R, Kang B, Qian S, He X, Zhang X. 2023. Single-cell long-read sequencing in human cerebral organoids uncovers cell-type-specific and autism-associated exons. *Cell Reports* **42**: 113335.

Zhang D, Zhang W, Sun R, Huang Z. 2021a. Novel insights into clear cell renal cell carcinoma prognosis by comprehensive characterization of aberrant alternative splicing signature: a study based on large-scale sequencing data. *Bioengineered* **12**: 1091–1110.

Zhang Y, Narayanan SP, Mannan R, Raskind G, Wang X, Vats P, Su F, Hosseini N, Cao X, Kumar-Sinha C, et al. 2021b. 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.

Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, et al. 2017. Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**: 14049.