Long-read RNA sequencing of archival tissues reveals novel genes and transcripts associated with clear cell renal cell carcinoma recurrence and immune evasion

Joshua Lee, Elizabeth A Snell, Joanne Brown, et al.

Genome Res. published online September 16, 2024
Access the most recent version at doi:10.1101/gr.278801.123

Long-read RNA sequencing of archival tissues reveals novel genes and transcripts associated with clear cell renal cell carcinoma recurrence and immune evasion

Joshua Lee[1, 2, 3, 4], Elizabeth A. Snell[4], Joanne Brown[5], Charlotte E. Booth[1,2], Rosamonde E. Banks[5], Daniel J. Turner[4, 6], Naveen S. Vasudev[5], and Dimitris Lagos[1,2,*]

1. Hull York Medical School, University of York, UK. 2. York Biomedical Research Institute, University of York, York, UK. 3. Department of Biology, University of York, UK. 4. Oxford Nanopore Technologies plc, Oxford, UK. 5. Leeds Institute of Medical Research at St James's, University of Leeds, St James's University Hospital, Leeds, UK. 6. Current address: ENHANC3D Genomics, Cambridge, UK.

* Corresponding author: dimitris.lagos@york.ac.uk

Running title: Long-read RNA sequencing of renal cell carcinoma

Abstract

The use of long-read direct RNA sequencing (DRS) and PCR cDNA sequencing (PCS) in clinical oncology remains limited, with no direct comparison between the two methods. We used DRS and PCS to study clear cell renal cell carcinoma (ccRCC), focussing on new transcript and gene discovery. Twelve primary ccRCC archival tumours, six from patients who went on to relapse, were analysed. Results were validated in an independent cohort of twenty patients by qRT-PCR and compared to DRS analysis of RCC4 cells. In archival clinical samples and due to long-term storage, average read length was lower (400-500nt) than that achieved through DRS of RCC4 cells (>1100nt). Still, deconvolution analysis showed a loss of immune infiltrate in primary tumours of patients who relapse as reported by others. Differentially expressed genes in patients who went on to relapse were determined with good overlap between DRS and PCS, identifying *LINC04216* and the T cell exhaustion marker *TOX* as novel candidate recurrence-associated genes. Novel transcript analysis revealed over 10,000 candidate novel transcripts detected by both methods and in ccRCC cells *in vitro*, including a novel *CD274* (*PD-L1*) transcript encoding for the soluble version of the protein with a longer 3'UTR and lower stability than the annotated transcript. Both methods identified 414 novel genes, also detected in RCC4 cells, including a novel non-coding gene over-expressed in patients who relapse. Overall, we showcase use of PCS and DRS in archival tumour samples to uncover unmapped features of cancer transcriptomes, linked to disease progression and immune evasion.

Introduction

Kidney cancer contributes approximately 2% of all newly diagnosed cancer cases worldwide (Sung et al. 2021). The most common form of kidney cancer is renal cell carcinoma (RCC) and the most frequent RCC type is clear cell RCC (ccRCC, ~75% of all RCC cases (Ricketts et al. 2018)). Inactivation of the *VHL* gene function is an almost universal hallmark of ccRCC. Secondary mutations are required in hotspot genes, including *PBRM1*, *SETD2* and *BAP1*, as well as copy number changes in Chromosomes 9p and 14q (Hsieh et al. 2018). Of

note, ccRCC tumours contain one of the highest percentages of tumour-infiltrating immune cells amongst all cancer types at approximately 30% of all cells (Aran et al. 2015; Rooney et al. 2015; Ricketts et al. 2018). Treatment of localised ccRCC typically involves removal of part or all of the kidney (radical/partial nephrectomy). Approximately one-third of patients have metastases detected at pre-operative screening and 30-50% develop metastases after the removal of the primary tumour (Rini et al. 2009). Several approaches have been proposed for assessing the risk of disease recurrence following surgery (Cotta et al. 2023). Scores based on gene expression signatures have also been proposed to refine risk prediction (Brannon et al. 2010; Rini et al. 2015; Morgan et al. 2018). However, despite a recognised need (Correa et al. 2019; Vasudev et al. 2020), so far, no set of biomarkers has reached routine clinical practice.

Aberrant co- and post-transcriptional events (e.g. alternative splicing/polyadenylation, post-transcriptional modifications, etc), drive oncogenesis but also tumour immunogenicity (Sveen et al. 2014; Smith et al. 2019). Our understanding of cancer transcriptomes is nearly exclusively based on short read sequencing platforms. Given that the average length of a mRNA is 1.5-2kb in mammals this approach requires high depth of sequencing to confidently call transcript variants and is limited with regards to reconstruction of full-length novel transcripts. Often the reliance on reference genomes/transcriptomes means that this approach misses or discards novel transcripts. Furthermore, it is extremely difficult, if not impossible, to confidently establish transcriptional co-dependencies, i.e. co-existence of distinct features (e.g. specific splice junctions and untranslated regions - UTRs) on the same transcript. Long-read direct RNA sequencing (DRS) and PCR cDNA sequencing (PCS) have emerged as transformative methodological alternatives to overcome these limitations (Nature Methods Editors 2023). Yet, in cancer, there are only a handful of reports using long read sequencing in tumour samples from patients with solid (Qu et al. 2022; Veiga et al. 2022; Mock et al. 2023) or blood cancers (Tang et al. 2020; Pratanwanich et al. 2021; Cortes-Lopez et al. 2023), with only one example of using DRS (in 3 myeloma patient

samples (Pratanwanich et al. 2021)). Currently, there are no reports directly comparing PCS to DRS in clinical samples and long-read sequencing has not been applied to kidney cancer. Here, we aimed to explore whether archival surgical fresh frozen nephrectomy tissue samples (typically stored for over 10 years) could be used for Nanopore DRS (RNA002 kit) and PCS (PCS111 kit) analyses (Garalde et al. 2018) to explore ccRCC transcriptomes. We focused on differential gene expression analysis to identify novel candidate predictors of disease relapse and discovery of novel genes and transcripts with evidence of cancer cell intrinsic expression and potential association with disease relapse.

Results

To demonstrate the feasibility of utilising long-read RNA sequencing technologies for characterising ccRCC transcriptomes in archival surgical fresh frozen specimens, we sequenced twelve snap-frozen nephrectomy samples using Nanopore PCS and DRS on ONT PromethION flow cells, using 200ng and 2µg of total RNA, respectively. These samples consisted of six specimens from patients who later developed ccRCC recurrence and six nonrecurrent controls (see Methods and **Supplemental Table S1**). For each clinical specimen, the same RNA sample was used for DRS and PCS analysis. No significant differences in RNA quality (based on RNA integrity numbers (RIN)) were observed between recurrent and nonrecurrent controls (**Supplemental Fig. S1A**). An overview of study design and data analysis pipeline is shown in **Fig. 1A**.

DRS and PCS of ccRCC nephrectomy samples

All nephrectomy specimens were successfully sequenced using both PCS and DRS. After 72 hours of sequencing, PCS generated reads ranging from 50 million to 85 million (median = 56.6 million, total = 701 million), with approximately 80% qualified as pass reads (median = 45.8 million, total = 561 million) having a minimum read Q score of 7. DRS generated between 2.4 million to 5.5 million reads (median = 4.6 million, total = 52.6 million), with approximately 70% qualified as pass reads (median = 3.2 million, total = 37.4 million).

Summary sequencing output statistics can be found in **Table 1** and **Supplemental Table S2**.

Both PCS and DRS reads were next mapped to the human reference genome. The median alignment length for PCS and DRS reads were 517 and 405 nucleotides respectively, which is lower than that typically observed (Garalde et al. 2018). A range of 21 – 37.1% (median = 25.95) of PCS- and 3.2 – 18.1% (median = 7.6%) of DRS-aligned reads represent full-length transcripts (coverage of at least 95% of the mapped reference annotation) (**Fig. 1B, C**). As RNA molecules are sequenced from 3' end, gene coverage is biased towards 3' end (**Supplemental Fig. S1B**). Overall, PCS and DRS reads achieved median accuracies of 95.5% and 90.5%. The longest aligned reads for PCS and DRS were 27,854 and 7,822 nucleotides, respectively. This was likely due to the significantly higher sequencing depth achieved by PCS. These results demonstrated the capability of Nanopore long-read RNA sequencing to produce high-depth, PCS and DRS sequencing datasets from flash-frozen historical clinical specimens. The modest average read length is likely due to long-term storage of these samples.

To evaluate the ability of long-read sequencing to capture the diversity of the transcriptome, we examined the RNA biotypes of genes identified by PCS and DRS. PCS identified 39,115 genes across the 12 samples, with a median of 26,203 mapped genes per specimen (**Supplemental Fig. S1C**). Among all PCS-mapped genes, 45.47% were classified as protein-coding genes, 29.48% as long non-coding RNAs (lncRNA) and 12.34% as processed pseudogenes (Fig. 1D). In comparison, DRS identified 26,457 genes across the specimens (median = 18,057 per sample, **Supplemental Fig. S1C**), with 32.48% classified as protein-coding genes, 28.81% as lncRNAs and 16.50% as processed pseudogenes (**Fig. 1D**). 25,692 genes were mapped by both methods (**Fig. 1E**). 13,423 genes were exclusively mapped by PCS, likely due to higher sequencing depth compared to DRS. Notably, 765 genes were exclusively mapped by DRS.

We observed that the majority of expressed genes for both PCS and DRS were protein-coding (89.4% and 91.7%, respectively), followed by mitochondrial rRNAs (mt-rRNAs)

(5.35% and 4.66%), processed pseudogenes (2.57% and 1.74%) and lncRNA (1.71% and 1.10%) (**Fig. 1D** and **Supplemental Fig. S1D**). The observed bias towards detection of protein-coding genes was likely due to both PCS and DRS using poly(A)-targeting probes for library construction, also meaning all sequenced transcripts are polyadenylated. Some read-through events were observed but we did not analyse them systematically as we were mindful of the relatively low percentage of full-length reads in our analyses of archival tissue. Despite using total RNA as input for PCS and DRS library preparation, highly abundant rRNAs were sequenced at negligible levels. Distribution of gene expression levels for each biotype by PCS and DRS are illustrated by violin plots at **Supplemental Fig. S1E**. Furthermore, amongst genes mapped by both PCS and DRS (n = 25,692), we found significant correlation in their gene expression levels (**Fig. 1F and Supplemental Fig. S2**). Overall, whilst PCS provided greater sequencing depth, our data demonstrated that both methods can capture a diverse range of transcripts from archival clinical samples, yielding highly concordant gene expression profiles.

Differential gene expression analysis reveals that ccRCC recurrence is associated with suppressed tumour immune infiltration

We then tested whether DRS and PCS can identify features associated with ccRCC recurrence. After alignment to the reference genome, PCA did not result in visually distinct gene expression clusters correlating with ccRCC recurrence status for either PCS or DRS (**Supplemental Fig. S3A**). No sample separation was observed based on RNA integrity, number of pass sequencing reads, proportion of full-length transcripts, sex, or the number of mutations on ccRCC prognostic markers (**Supplemental Fig. S3B-F**). We explored the effect of number of mutations in addition to *VHL* as we have previously shown poorer outcomes for tumours with *VHL+2* or more mutations (Scelo et al. 2014; Vasudev et al. 2023). However, differential gene expression analysis identified 159 and 68 genes with significantly differential expression $|\log_2\text{FoldChange}| \geq 2$, padj ≤ 0.1) between recurrent and nonrecurrent tumours by PCS and DRS, respectively, with substantial overlap (**Fig. 2A, B**,

Supplemental Tables S3 and S4). The directionality of gene expression amongst these differentially expressed genes (DEGs) showed strong correlation (**Fig. 2C**). We note that we did not observe any outliers with regards to the time to relapse within the recurrent disease group.

As PCS produced substantially higher number of sequencing reads, we further evaluated ccRCC gene expression patterns using randomly subsampled PCS reads (5%) compared to DRS. We selected 5% as this brings the PCS depth to similar levels of the range achieved by DRS (2-4 million passed reads). PCA did not reveal a distinct cluster correlating with ccRCC recurrence (**Supplemental Fig. S3G**). Differential gene expression analysis of 5% subsampled PCS data identified 92 DEGs, with significant overlap with DRS. The number of commonly identified DEGs between 5% PCS and DRS increased as a percentage of total DEGs identified by PCS and decreased as a percentage of total DEGs identified by DRS (**Fig. 2D-E, Supplemental Table S5**).

Within the overlapping DEGs between PCS, 5% PCS, and DRS, several key adaptive immune genes, including *CD8B*, *PDCD1*, *GZMK* and *TOX*, were significantly downregulated in recurrent samples (**Fig. 2A, B**). To evaluate variations in biological processes and pathways between recurrent and nonrecurrent ccRCC tumours, we performed Gene Ontology (GO) analysis. The top 10 most significantly enriched (by padj) GO biological processes (BP) terms from PCS data were all associated with adaptive immunity (**Supplemental Fig. S4A**). This pattern was also found using DRS data, where enrichment plot for the top 5 enriched GO BP terms (by padj) by DRS showed identified suppression of adaptive immune response related pathways (i.e. positive regulation of cell killing, T cell mediated cytotoxicity) in ccRCC recurrent samples (**Supplemental Fig. S4B**).

To further explore the relationship between ccRCC recurrence and immune infiltrate populations, we used the ESTIMATE algorithm (Yoshihara et al. 2013), which uses gene expression signatures to infer tumour purity and immune cell abundance. Using PCS data, we found that recurrent ccRCC exhibited significantly lower immune scores and higher levels of tumour purity compared with nonrecurrent controls (**Fig. 2F**, **Supplemental Fig.**

S4C). DRS data displayed a borderline non-significant trend towards decrease in immune scores in recurrent ccRCC tumours ($p = 0.0881$) and a high degree of concordance was found between DRS and PCS ESTIMATE immune scores ($R2 = 0.87$, $p < 0.0001$) (**Supplemental Fig. S4D, E**). Both PCS and DRS datasets also had significantly lower immune scores for recurrent samples according to xCell analysis (**Supplemental Fig. S4F**). Immune infiltrate profiles were further analysed using another cell-type deconvolution algorithm, CIBERSORTx (Newman et al. 2019). Both PCS and DRS exhibited a significant reduction in the fraction of CD8+ T cell within recurrent ccRCC tumours when compared to nonrecurrent controls (**Fig. 2G**, **Supplemental Fig. S4G, H**). Similarly, EPIC, another immune cell type deconvolution method (Racle and Gfeller 2020), also indicated suppression of CD8+ T cell within the immune infiltrates amongst the recurrent ccRCC tumours (**Supplemental Fig. S4I**). These findings agreed with a previously reported, qRT-PCR based, ccRCC recurrence prediction assay, which also linked lower expression levels of immune response genes with an increased likelihood of disease recurrence (Rini et al. 2015). Amongst the 11 recurrence-related gene makers examined in that study, our PCS and DRS analyses also identified that levels of *NOS3* and *CCL5* were significantly decreased in the recurrent tumours (**Supplemental Fig. S4J**).

Critically, we sought to validate our long-read sequencing results by an independent method (qRT-PCR) for *CD8B*, *PDCD1*, *GZMK* and *TOX* using samples from both the sequenced cohort but also an independent validation cohort ($n = 20$, 10 from recurrent ccRCC patients and 10 nonrecurrent controls). This analysis confirmed significant downregulation of the CD8+ T cell marker *CD8B*, the activation marker *GZMK*, and the T cell exhaustion marker *TOX* in the recurrent tumours (**Fig. 2H**; **Supplemental Fig. S5A-C**). We note that for *TOX* the effect was statistically significant both in the pooled data and when analysing the sequenced and validation cohorts separately. *PDCD1* (also known as *PD-1*) levels were not statistically different when assessed by qRT-PCR between the two groups (**Fig. 2H**; **Supplemental Fig. S5D**). To explore if the additional sequencing depth achieved by PCS could lead to identification of more candidate gene correlates of disease recurrence we used

qRT-pCR to measure levels of three DEGs (*LINC04216*, *LINC04217*, *POU4F1*) that were only significantly differentially expressed according to PCS. We found significant downregulation of *LINC04216* using the sequenced cohort and when both cohorts were pooled (**Fig. 2H**, **Supplemental Fig. S5E**). No significant changes were found for *LINC04217* and *POU4F1* by qRT-PCR (**Supplemental Fig. S5F-G**).

Collectively, these findings demonstrated that both PCS and DRS can identify differential expression signatures associated with disease relapse. PCS and DRS showed a significant suppression of immune infiltration, particularly CD8+ T cells, in tumours of patients who later experience disease recurrence, and identified the exhaustion marker *TOX* and the *LINC04216* non-coding RNA as novel candidate recurrence-associated genes.

Differential transcript usage analysis identifies candidate isoform switching events associated with ccRCC recurrence

One of the advantages of the long-read sequencing approach lies in its ability to identify and quantify transcript isoforms. By aligning sequencing reads against the reference transcriptome, isoform-level expression data can be used to detect differential transcript usage (DTU) events. We first compared reference genome- and reference transcriptome-based method in gene expression and differential gene expression analysis. For both PCS and DRS, the reference transcriptome alignment method detected similar number of genes compared to reference genome alignment, with substantial overlap (**Fig. 3A**). Gene expression levels of the PCS and DRS of nephrectomy samples also displayed strong correlation between the two alignment methods (r = 0.7685 for PCS, r = 0.7087 for DRS) (**Fig. 3B**). Differential gene expression analysis using PCS and DRS reference transcriptome alignment data identified 197 and 34 significant DEGs between recurrent and nonrecurrent controls, respectively, with good overlap with reference genome alignment method (**Fig. 3C**, **Supplemental Fig. 6A Supplemental Tables S6-7**). The directionality of gene expression amongst these DEGs showed strong correlation (**Supplemental Fig. 6B**).

DTU analysis was carried out on both PCS and DRS of ccRCC tumour samples using DRIMSeq (Nowicka and Robinson 2016) and DEXSeq (Anders et al. 2012). Analysis of the PCS data identified 31 genes that displayed isoform switching in recurrent ccRCC tumours compared to nonrecurrent controls (**Fig. 3D**, **Supplemental Table S8**). These included *CMC1* that showed statistically significant DTU by both DRIMSeq and DEXSeq and was also identified by DRS to display DTU (**Supplemental Table S9**). In PCS, most of the isoforms that are expressed in recurrent ccRCC samples are ENST000423894 and ENST00000466830 (**Fig. 3E-F**, labeled as purple and teal. respectively). In contrast to nonrecurrent counterparts, recurrent ccRCC specimens also expressed very low level of ENST00000468330 and ENST00000495428 (**Fig. 3E-F**, labeled as yellow and green, respectively). Overall, this analysis revealed a limited number of candidate disease recurrence-associated DTU events.

Long-read RNA sequencing enables the discovery of novel full-length transcripts expressed in ccRCC cells

A unique strength of long-read sequencing is the potential to discover novel transcript isoforms and genes, not currently included in the reference transcriptome. To identify novel transcript isoforms that are present in the ccRCC nephrectomy specimens, we applied StringTie2 to perform transcriptome assembly using PCS reads aligned to the reference genome. StringTie2 assembled isoforms were subsequently compared to the reference annotation (Ensembl release 105) with both SQANTI3 and gffcompare (see Methods). SQANTI3 classifies each assembled isoform into known or novel based on their splice junction matches. Known transcripts comprise FSM and ISM, whereas NIC, NNC, antisense, fusion, genic, genic intron and intergenic isoforms are classified as novel transcripts (**Fig. 4A**, **Supplemental Table S10**). Similarly, gffcompare assigns each StringTie2 assembled isoform with a transcript class code which corresponds to ‘Known’ and ‘Novel’ transcripts (**Supplemental Fig. S7**, **Supplemental Table S10**).

Both SQANTI3 and gffcompare classifications revealed that novel transcripts constitute more than 50% of the assembled transcripts from the nephrectomy specimens (**Fig. 4B**). For SQANTI3 classification, the most prominent class of assembled transcripts were FSM (36.5%, n = 19,722 out of a total 54,185, **Fig. 4C**). Within the FSM isoforms, 15.3% exhibited an alternative 3' end (n = 3,010), 11.0% contain an alternative 5' end (n = 2,170) and 4.9% of FSM transcripts display alternative 3' and 5' ends (n = 962) when compared to the reference annotation (**Supplemental Table S11**). Under a broader classification criterion, these isoforms could be considered as putative novel transcripts. Importantly, analysis by SQANTI3 also indicated that the large proportion of novel transcripts may possess coding potential (**Fig. 4D**).

Similarly, gffcompare analysis revealed that the predominant class of StringTie2 assembled transcripts was 'j' (Novel, multi-exon gene with at least 1 matched exon junction) (40.58%, n = 21,635), followed by '=' (Known, complete intron chain match) (28.32%, n = 15,099) (**Supplemental Table S12**). Applying the alternative long-read RNA-seq transcript assembler FLAIR on PCS reads, gffcompare characterisation reaffirmed that the majority of assembled transcripts are novel isoforms (**Supplemental Table S12**). Detailed characterisations of novel StringTie2 assembled transcripts by SQANTI3 and gffcompare can be found in **Supplemental Table S13**.

Next, we asked whether the novel assembled transcripts can also be detected by DRS of nephrectomy samples and, ccRCC tumour cells *in vitro*. The latter was used to avoid artefacts associated with the modest read length and full-length transcript coverage achieved in clinical samples, and to indicate the cancer cell-intrinsic origin of these transcripts. To address this, we performed DRS analysis of the *VHL*-negative, ccRCC cell line RCC4 under both untreated and IFNG and TNF treated conditions using DRS. The cytokine treatment conditions aimed to simulate in part the transcriptomic response of tumour cells to immune cells. Workflow and sequencing statistics can be found in **Supplemental Fig. S8**, and DEG analysis data can be found in **Supplemental Table S14**. The mean read length was over 1,100nt and the full-length transcript coverage over 45% for

all samples. Out of the 26,834 novel isoforms that were mapped by PCS of nephrectomy samples, 14,544 were also mapped in at least one DRS of the nephrectomy samples, and 13,336 were also detected in the DRS of RCC4 samples, whereas 10,645 novel transcript isoforms were detected in all three datasets (**Fig. 4E**). Levels of novel isoforms that were detected in all three datasets showed comparable expression levels compared to known reference isoforms (**Fig. 4F**). Furthermore, expression levels of these novel isoforms exhibited strong concordance between PCS and DRS of nephrectomy samples (R^2 = 0.6832, P = <0.0001) (**Fig. 4G**). Despite starting with a reference based on our PCS dataset, a small number of StringTie2 transcripts were mapped exclusively in DRS of nephrectomy samples (n = 80) and RCC4 (n = 332). This may be due to the assignment of multi-mapping, ambiguous sequencing reads by the sequence alignment program (minimap2). These results reveal a plethora of previously uncharacterised and unmapped transcripts within the ccRCC transcriptome. Despite differences in sequencing depth, novel transcripts from PCS could also be mapped by DRS, with a substantial proportion of these novel transcripts also detected in ccRCC cells *in vitro.*

Long-read RNA sequencing reveals the full exonic structure of ccRCC splice variants

Taking advantage of the ability of long-read sequencing to reveal whole transcript exonic structures, we next examined recently reported novel, non-reference annotated splice variants (SVs) specific to ccRCC tumours, which were supported by proteomics data and associated with clinical outcomes (Chang et al. 2022). We found sequence read evidence for all 16 reported SVs (15/16 for PCS of nephrectomies, 9/16 for DRS of nephrectomies and 13/16 for DRS of RCC4). Moreover, PCS StringTie2 assembled transcripts spanning 11 of the unannotated SVs (**Supplemental Table S15**). For example, the StringTie2 assembled transcripts *MSTRG.9279.18* and *MSTRG.9269.19* accurately replicated two ccRCC-specific SVs from *MVK* (**Fig. 4G**). This was supported by reference genome aligned reads from all three long-read RNA-seq datasets (**Supplemental Fig. S9A**). In addition, sequencing results showed that these ccRCC splice variants adopt the 3'UTR structure of *MVK-002*

(ENST00000392727) instead of the longer 3'UTR from canonical MANE transcript *MVK-001* (ENST00000228510) (**Fig. 4H**). Another example was *HPCAL1*, where two StringTie2 assembled transcripts (*MSTRG.20400.11* and *MSTRG.20400.12*) were found to span the ccRCC-specific SV (**Supplemental Fig. S9B**). The two isoforms exhibit variation in the exon 3 usage, where evidence of exon 3 retention can be found in PCS as well as RCC4 sequencing results. Overall, 3 additional SVs (*SYNPO*, *EGFR* and *FAM107B*) were found to be encompassed by 2 StringTie2 assembled transcripts (**Supplemental Table S15**). We also examined *VHL* isoform expression in PCS and DRS of nephrectomies, as well as DRS of RCC4. The 3'UTR of the *VHL* mRNA is 4kb-long. As such, even though full length *VHL* transcripts were detected, most PCS and DRS of archival samples did not span the 5' upstream exons (**Supplemental Fig. S10**). The longer sequencing read length achieved in the DRS analysis of RCC4 cells allowed us to capture mainly full-length *VHL*. The majority of expressed *VHL* transcripts in RCC4 correspond to ENST00000256474, but with a shorter 3'UTR compared to the reference gene model. Collectively, using recently reported ccRCC-related SVs (Chang et al. 2022), we demonstrated the ability of long-read sequencing to reveal transcriptomic co-dependencies, in this case the co-occurrence of specific SVs with specific UTRs, providing unparalleled insight into novel features of ccRCC.

Discovery of a novel soluble PD-L1 isoform expressed by ccRCC tumour cells

Having identified that a reduction in the immune infiltrate of ccRCC tumours was linked to disease recurrence and that the ccRCC transcriptome includes a high number of previously uncharacterised novel transcript isoforms we focused on transcripts of immune checkpoint proteins. Here, we focused on*PD-L1* (official symbol CD274). Whilst most studies on *PD-L1* have focused on the membrane-bound isoform (*mPD-L1*), recent attention has been drawn to a soluble *PD-L1* isoform (*sPD-L1*) lacking exon 5, 6, and 7. *sPD-L1* is currently unannotated in the Ensembl gene annotation, but it has been described in the NCBI GenBank database (NM_001314029). An Ensembl annotated transcript (ENST00000474218) partially overlaps with the 3'UTR of the GenBank *sPD-L1* transcript,

serving as a proxy for mapping sPD-L1. Upon closer inspection to the exon 4 and 3'UTR region of *sPD-L1* (Chr5: 5,462,800 – 5,463,400), reference genome reads coverage and StringTie2 supported two distinct isoforms with varying 3'UTR lengths (**Fig. 5A**). While the shorter *sPD-L1* represents the GenBank transcript, the alternative *sPD-L1* includes a 3'UTR more than twice the length of GenBank annotation (61nt vs 154nt) (**Fig. 5A**). StringTie2 assembly revealed this elongated 3'UTR structure, with supporting evidence stemming from reference genome aligned reads of PCS and DRS data from ccRCC tissues, as well as DRS data from RCC4 cells (**Supplemental Fig. S11A-E**). To further validate our findings, we performed short-read Illumina RNA sequencing analysis of RCC4 cells and we were able to detect reads corresponding to the novel *sPD-L1* 3'UTR. Furthermore, analysis of publicly available short-read RNA-seq data of normal human kidney and lung tissues from the GTEx project (The GTEx Consortium 2013) also validated the existence of this *PD-L1* isoform (**Supplemental Fig. S11F-G**).

Upon evaluating the expression of *PD-L1* in ccRCC tumours, no significant disparity in gene-level expression was found between recurrent and nonrecurrent nephrectomy samples (**Fig. 5B**). However, at the isoform level, whilst *mPD-L1* was suppressed in the recurrent samples, both *sPD-L1* isoforms (NM_001314029 and novel *sPD-L1*) showed no significant differences (**Fig. 5C**). We note that in the clinical samples *mPD-L1* is the most abundant *PD-L1* transcript, whereas expression of the novel and annotated *sPD-L1* isoforms is comparable (**Fig. 5C**). Subsequent expression validation via qRT-PCR with the sequenced and independent validation cohort displayed the same pattern of results, where *mPD-L1* displayed a borderline non-significant (p=0.09) downregulation, while *sPD-L1* isoforms remained unchanged (**Supplemental Fig. S12**).

As all *PD-L1* transcripts, including the novel *sPD-L1* isoform, were detected in RCC4 cells by DRS, we sought to further explore their regulation in cancer cells. The expression of all *PD-L1* isoforms increased in response by IFNG and TNF treatment (**Fig. 5D**). However, expression levels of *mPD-L1* were profoundly more responsive to cytokine treatment than the soluble isoforms (approximately 30-fold induction of *mPD-L1* as opposed to 3-10-fold

induction of *sPD-L1* isoforms; **Fig. 5D, E**). mRNA stability assays revealed that cytokine treatment significantly reduced the stability of *sPD-L1* but not *mPD-L1* (**Fig. 5F**). Furthermore, the novel *sPD-L1* isoform exhibited lower stability than the total *sPD-L1* isoforms. Taken together, our findings revealed the existence of an up-to-now uncharacterised *sPD-L1* isoform with a longer 3'UTR and low stability, and key differences in the regulation of membrane and soluble *PD-L1* isoforms in ccRCC tumours and in response to inflammatory cytokines *in vitro*.

Discovery of novel genes associated with ccRCC recurrence

In addition to the characterisation of novel isoforms within known genes, long-read RNA-seq also enables the discovery of novel genes that are absent from the reference gene annotation. Using the PCS StringTie2 assembly, we identified 1,350 novel genes (curated by SQANTI3) that were mapped in the PCS dataset. The majority of these genes were classified as either intergenic (59.76%) or antisense (39.86%) transcripts (**Fig. 6A**). Most of these novel genes have a single isoform, with the majority being non-coding, multi-exon isoforms featuring canonical splice sites (**Fig. 6B**, **Supplemental Fig. S13A-C**). The expression levels of these novel genes are similar to those of reference annotated genes, with the coding novel genes demonstrating higher expression levels than non-coding novel genes (**Supplemental Fig. S13D**). Importantly, of the 1,350 novel genes that were mapped by PCS, 982 (72.7%) were also detected in the DRS data of tumour nephrectomies, and 414 (30.7%) novel genes were also mapped in the DRS data of RCC4 cells (**Fig. 6C**). This suggests that a large number of novel genes might be expressed in ccRCC tumour cells.

Next, we performed DEG analysis with the PCS StringTie2 assembly and identified a set of significantly differentially expressed ($|\log_2\text{FoldChange}| \geq 2$, padj ≤ 0.1) novel genes (n = 40 for PCS, n = 4 for DRS) between recurrent and nonrecurrent samples (**Fig. 6D, Supplemental Fig. S13E, Supplemental Tables S16-19**). The directionality of gene expression for these differentially expressed novel genes demonstrated strong concordance between PCS and DRS (**Fig. 6E**). 13 differentially expressed novel genes were also

identified between untreated and IFNG and TNF treated RCC4 cells (**Supplemental Fig. S13F**).

To further validate our sequencing findings, we sought to experimentally measure the levels in recurrent and nonrecurrent ccRCC nephrectomies of two novel genes: *MSTRG.29728* and *MSTRG.38727* using qRT-PCR. *MSTRG.29728*, a StringTie2 assembled gene is located on Chromosome 5, with its nearest reference annotated genes (5': *CSNK1G3*, 3': *LINC01170*) situated more than 300 kb away (**Fig. 6F**). Notably, the presence of this novel gene was also supported by reference genome aligned reads from the DRS and our short-read RNA sequencing analysis of RCC4 cells (**Fig. 6F**). Based on coverage data from all sequencing experiments, the most highly expressed *MSTRG.29728* isoform consists of 2 exons (**Supplemental Fig. S14A, B**) and its levels of expression are intermediate (**Supplemental Tables S16-18**). Analysis of publicly available cell line data further validated the existence of this gene and suggested that it was enriched in kidney cancer cell lines (**Supplemental Fig. S14C**). *MSTRG.29728* was significantly upregulated in recurrent ccRCC tumours in both PCS and DRS datasets. This upregulation was confirmed by qRT-PCR in both sequenced and independent validation cohorts (**Fig. 6G**). The second tested novel gene, *MSTRG.38727*, is located on Chromosome X with read coverage from PCS and DRS of nephrectomy specimens, albeit absent in DRS data from RCC4 (**Supplemental Fig. S15A-C**). PCS sequencing results showed that *MSTRG.38727* expression was highly elevated in 3 of the 6 recurrent ccRCC tumours (**Supplemental Fig. S15D**). This was corroborated through qRT-PCR validation within the sequenced cohort, but was not validated in the independent validation cohort (**Supplemental Fig. S15E**).

Overall, long-read sequencing revealed a high number of candidate novel genes present in ccRCC transcriptomes. Further testing for two such genes by orthogonal methods and in independent patient cohorts provide further support for their existence and, critically, identified *MSTRG.29728* as a novel non-coding RNA gene associated with ccRCC recurrence in both study cohorts. We provisionally term *MSTRG.29728* as *RECART*, for Renal Carcinoma Recurrence Associated Transcript.

Discussion

Long-read sequencing technologies represent a new era in cancer genomics and RNA medicine (Sakamoto et al. 2020; Wang et al. 2023). We used DRS and PCS to explore transcriptomes of primary ccRCC tumours. Our study aimed to demonstrate the methodological application of long-read sequencing, both PCS and DRS, in cancer and specifically ccRCC, focusing on use of archival fresh frozen tissue samples and new gene and transcript discovery, and using disease recurrence as a proof-of-principle context. Even though we analysed a sequencing and an independent validation cohort, the relatively low total number of study participants (n=32) is a limitation of our study that should be considered. In addition, the relatively modest read length acquired for clinical samples should also be considered as a limitation. As a mitigation for this, we used the RCC4 DRS dataset as a high-quality reference, as well as further validation for selected isoforms and genes. Using this approach, we showcase how long-read RNA sequencing can lead to discovery of novel disease-associated transcripts and genes, the existence of which is supported by multiple approaches including short-read Illumina sequencing, targeted qRT-PCR, and validation in independent cohorts. We opted not to perform more detailed analyses such as estimation of poly(A) length per transcript or post-transcriptional RNA modification analyses (Krause et al. 2019). However, our work sets the foundation for follow up studies using the new Nanopore DRS platform (RNA004) comparing the ability to detect changes in such features in fresh and archival samples.

From a methodological point of view, the distinguishing features of our study are (i) the use of long-term stored tissue, (ii) the direct comparison between DRS and PCS of clinical samples, (iii) the successful sequencing of archival fresh frozen tissue samples, and (iv) the use of total RNA as starting material for DRS and PCS library preparation. In reference to the latter point, compared to the pg-ng range of total RNA input requirement for short-read RNA sequencing library preparation, previous studies using ONT DRS have typically used 50 – 500 ng of poly(A) enriched RNA, which is hugely demanding for clinical samples (Jain

et al. 2022). Here we used 2 µg and 200 ng total RNA for DRS and PCS from tissues, respectively without poly(A) enrichment. Indeed, it has been suggested that poly(A) selection can introduce a potential bias towards mRNAs with longer poly(A) tails (Viscardi and Arribere 2022). PCS achieved a higher depth and, consequently, detected a higher number of transcripts and genes in all tested samples, and a higher number of DEGs in primary tumours of patients who experienced recurrence than DRS. We note that we did not multiplex samples for PCS, but used a subsampling approach (5% PCS) that identified similar gene expression patterns both with regards to DEGs and enriched pathways. The 5% subsampling level was chosen to reduce the PCS read depth within the range achieved by DRS (2-4 million passed reads). On the other hand, DRS does not include a PCR amplification step, providing further confidence in the overlapping gene sets between the two methods. With regards to read-length, both methods produced long reads. On average, raw reads generated by PCS were longer than DRS reads likely because of the fact that raw reads from PCS have additional ligated reverse transcription, PCR amplification primer and unique molecular identifier. Once aligned to the reference genome, both methods achieved similar read lengths, although PCS achieved a higher percentage of full-length transcripts likely due to the size selection step following PCR (Bayega et al. 2022). Alignment to reference transcriptome showed good correlation with genome mapping and DTU analysis identified candidate DTU events associated with recurrence, including changes in *CMC1* transcript usage. It should be noted however that, when using historical samples and achieving relatively modest read lengths, there might be limitations in the ability to accurately measure ratios of different splice variants of the same gene. We note that as we used archival tumour samples, our DTU analysis should be interpreted with caution as it is likely to be underestimating the number of disease relapse-associated DTU events. This is why we focused on gene level comparisons and discovery of novel transcripts and genes that could be validated by other methods including DRS of RCC4 cells that achieved higher quality measures and by qRT-PCR.

The primary biological objective of our study was to use DRS and PCS to explore ccRCC recurrence-associated transcriptome features including previously uncharacterised genes and transcripts. Our differential expression and deconvolution analyses identified a loss of immune infiltrate and specifically CD8+ T cells as a key feature of primary tumours that go on to relapse after surgery. This is reported by others (Ghatalia et al. 2019; Peng et al. 2022) providing a biological validation of our findings. Despite the low numbers of samples tested in our sequencing cohort we were able to see similar expression patterns for *NOS3* and *CCL5*, two previously reported recurrence markers (Rini et al. 2015), but also the novel finding of downregulation of *TOX* and *LINC04216* linked to recurrence (note that *TOX* levels were not measured in the study that identified loss of *NOS3* and *CCL5* as recurrence markers (Rini et al. 2015)). In addition, our study also identified upregulation of a novel gene, *MSTRG.29728* or *RECART*, as a candidate marker of disease relapse. These candidate prognostic biomarkers of relapse will need to be validated in the future in independent cohorts.

A unique strength of long-read sequencing is the ability to determine preferential use of specific UTRs by specific SVs, which can suggest tissue- or disease-specific co-transcriptional processing mechanisms. Indeed, we demonstrated this for recently identified ccRCC-associated SVs (Chang et al. 2022), including *MVK* and *HPCAL1*. Focusing these analyses on immune checkpoints led to the discovery of a novel *sPD-L1* transcript with a longer 3'UTR than the currently annotated *sPD-L1*. This means that the novel *sPD-L1* is likely to be controlled by additional post-transcriptional mechanisms, including microRNA-mediated silencing or regulation by RNA-binding proteins. Of note, regulation through the 3'UTR is a major determinant of *mPD-L1* expression (Sun et al. 2018; Yamaguchi et al. 2022). Indeed, the novel *sPD-L1* transcript demonstrates lower stability than the other *PD-L1* transcripts under homeostatic or inflammatory conditions *in vitro*. We found that there is a trend for downregulation for tumour *mPD-L1* but no differences in *sPD-L1* in patients that experience recurrence. This is consistent with the observed loss of CD8+ T cells from these tumours and the enhanced responsiveness of mPD-L1 to IFNG and TNF observed *in vitro*.

Clinically, this is important as PD-1/PD-L1-targeted checkpoint inhibitors are currently being explored in the adjuvant setting (Gorin et al. 2022; Motzer et al. 2023; Choueiri et al. 2024) and expression of *sPD-L1* has been linked with ccRCC prognosis and immunotherapy treatment outcome (Larrinaga et al. 2021; Mahoney et al. 2022). Future studies will have to explore the relative contributions of the different *PD-L1* transcripts, including the novel one reported here, to tumour immune evasion and response to immunotherapy.

Overall, we demonstrate feasibility of both DRS and PCS in archival clinical samples with significant overlap between the two methods with regards to detectable transcripts, differential gene expression analysis, pathway enrichment analysis, and novel transcript and gene discovery. We also identify a common limitation in that when using historical samples that have been stored for long periods of times (years) both methods might result in relatively shorter read length. Higher depth can be achieved for PCS, which might be beneficial for initial comparative analyses. On the other hand, demonstrating the feasibility of DRS using archival clinical samples opens the way for future studies exploring questions that can only be addressed by DRS (e.g. RNA post-transcriptional modifications) avoiding biases associated with reverse transcription and PCR amplification. We provide evidence for the existence of thousands of novel transcript isoforms and hundreds of novel genes detected by both DRS and PCS in primary ccRCC tumours but also *in vitro* in ccRCC cell lines. We describe loss of *TOX* and *LINC04216* and upregulation of *RECART* as novel candidate predictors of relapse. We discover and validate through orthogonal methods a novel *sPD-L1* isoform with differential stability. These findings demonstrate that application of long-read RNA sequencing, even in long-term stored tissue samples, has the potential to lead to a radical revision of our understanding of cancer transcriptomes.

Methods

Study participants and ethics

In this observational study, we used 32 ccRCC tumour nephrectomy samples (16 nonrecurrent and 16 recurrent cases) collected between 2000 and 2012 and stored in the

Leeds multidisciplinary research tissue bank. 12 samples were used as a discovery cohort for DRS and PCS sequencing (6 nonrecurrent and 6 recurrent) and 20 (10 in each group) were used as an independent validation cohort. For the recurrence group, median time to relapse was 23 months (5 – 176). For the control group median follow-up without relapse was 11 years (7 – 18). Groups were matched for demographic, pathological, and clinical characteristics including TNM and Leibovich score (**Supplemental Table S1**, age is shown in 5-year intervals). The sample/kidney IDs were not known to anyone outside the research group. Mutation status of each sample was determined as described (Scelo et al. 2014; Vasudev et al. 2023). This study was approved regional ethics committee approval: Yorkshire & The Humber – Leeds East Research Ethics Committee, reference 15/YH/0080. The research conforms with the principles of the Declaration of Helsinki. All patients gave written informed consent for their participation in this study.

Tissue sample preparation

Following surgical removal, tissue samples were washed in phosphate-buffered saline (PBS), blotted on a tissue before being enveloped in aluminium foil and snap frozen in liquid nitrogen. Once thawed, samples were immediately used for RNA extraction without further freeze-thawed cycles. All cases underwent pathology review of a parallel formalin fixed paraffin embedded (FFPE) block to confirm ccRCC histology and tumour cell viability, as part of a separate study (Scelo et al. 2014).

Cell culture and cytokine treatment

RCC4 cells were maintained at 37^{o}C in a humidified atmosphere of 5% CO_2 and grown in complete Dulbecco's Modified Eagle's Medium (DMEM, Gibco 21969-05), supplemented with 10% foetal bovine serum (FCS) (Gibco A5256701), 1% 200 mM L-Glutamine (Gibco 25030) and 1% penicillin/streptomycin (Gibco 15140). For RCC4 Direct RNA sequencing experiment, 1 x 106 RCC4 cells were seeded in 15 mL of complete DMEM in T75 flasks. 24 hours after seeding, media were changed into complete DMEM, with or without the addition

of IFNG (1000U/mL, Peprotech 300-02) and TNF (25 ng/mL, Peprotech 300-01). Cells were harvested 24 hours later for RNA extraction. Three flasks of T75s were used for each replicate for the sequencing experiment.

RNA extraction

Total RNA was extracted from nephrectomy specimens or cultured cells using QIAzol (Qiagen 79306) and RNeasy kits (Qiagen 74004) with on-column DNase I digestion step, according to manufacturer's instruction. Nephrectomy specimens were homogenized in QIAzol using a TissueLyser LT (Qiagen 85600) with stainless steel beads (Qiagen 69997). RNA integrity number (RIN) was determined using the 2100 Bioanalyzer with RNA Nano kit (Agilent 5067) and quantified using Qubit RNA HS assay kit (Invitrogen, Q32852). Total RNA from RCC4 for Direct RNA sequencing was enriched for poly(A)$^+$ RNA molecules using the Dynabeads Oligo(dT)25 mRNA isolation kit (Invitrogen 61002).

Library preparation and RNA sequencing

Sequencing libraries used for PCR-cDNA-seq and Direct RNA-seq were generated using the SQK-PCS111 and SQK-RNA002 kit (Oxford Nanopore Technologies, ONT), respectively. For the nephrectomy specimens, 200 ng and 2 µg of extract total RNA were used as input for each sequencing library for PCR-cDNA-seq and Direct RNA-seq, respectively. 500 ng of poly(A)$^+$ RNA was used for each sequencing library for Direct RNA-seq of RCC4 cells. For PCR-cDNA-seq, cDNA libraries were prepared with the SQK-PCS111 kit according to manufacturer's instruction with 14 cycles of PCR cycles. For Direct RNA-seq, libraries were prepared with the SQK-RNA002 kit according to manufacturer's instruction including the optional reverse transcriptase step. Sequencing libraries for each experiment were prepared together to mitigate batch effects. All sequencing libraries were sequenced on ONT PromethION sequencer with R9.4.1 PromethION flow cells (ONT) for 72 hours. Basecalling and FASTQ files generation were performed with Guppy (v5.1.12, ONT).

602 **Quality control and reads alignment**

603 Sequencing reads generated from Direct RNA-seq, and PCR-cDNA-seq with a minimum
604 read quality score (Q score) of 7 were used for mapping and downstream analysis. FASTQ
605 files generated from sequencing runs were concatenated using catfishq (v1.4.0,
606 https://github.com/philres/catfishq). PCR-cDNA-seq reads were orientated by pychopper
607 (v2.5.0, https://github.com/nanoporetech/pychopper), filtered for the presence of 5' and 3'
608 sequencing adaptors and trimmed by cutadapt (v4.1) (Martin 2011). Direct RNA-seq and
609 processed PCR-cDNA-seq reads were aligned to either human genome, transcriptome
610 (GRCh38, Ensembl release 105) or StringTie2 assembly using minimap2 (v2.24), with
611 recommended parameters (Genome alignment: -ax splice –uf –k14; Transcriptome
612 alignment: -ax map-ont –p 0 -N 10) (Li 2018). Aligned reads were sorted, merged and
613 indexed to BAM files with SAMtools (v1.13) (Li et al. 2009). For subsampling, reference
614 genome aligned PCS reads were randomly selected using the SAMtools view command with
615 '-s 0.05'. The workflows for reads alignment are available at **Supplemental Code** and
616 https://github.com/joshuacylee/DRS and https://github.com/joshuacylee/PCR-cDNAseq.
617 Mapping data quality and statistics of sequencing data were analysed by Nanoplot and
618 bamslam (De Coster et al. 2018), https://github.com/josiegleeson/BamSlam). Illumina
619 sequencing reads were processed with FastQC, trimmed using cutadapt (version 1.18) to
620 remove sequence adaptors, followed by reference genome alignment with HISAT2 (Kim et
621 al. 2019).
622
623 **Differential gene expression**
624 Gene-level expression quantification was performed using featureCounts with long-read
625 counting mode (-L) (subread v2.0.0) (Liao et al. 2014). Transcript isoform quantification was
626 performed using Salmon (v1.7.0) with Oxford Nanopore long-reads mode (--ont) (Patro et al.
627 2017). Normalisation and identification of differentially expressed genes (padj ≤ 0.1 and
628 $\log_2$FoldChange ≥ 2) were performed using the R package (R Core Team 2022) DESeq2
629 (v1.40.2) (Love et al. 2014). PCA plots were generated by DESeq2 and volcano plots were

generated with the R package EnhancedVolcano (v1.18.0). Workflow for differential gene expression identification is available at **Supplemental Code** and (https://github.com/joshuacylee/DESeq2).

Gene set enrichment analysis and tumour-infiltrating immune cell analysis

Gene set enrichment analysis was performed using clusterProfiler (v4.4.4) (Wu et al. 2021). Gene Ontology biological process and molecular function databases were used for functional enrichment analysis. Parameters used for Gene Ontology enrichment were as follows: Permutations (nPerm):10000, minimum gene set size (minGSSize): 5, Maximum gene set size (maxGSSize): 500, Minimum p-value (pvalueCutoff) = 0.05, Organism (Orgdb) = org.Hs.eg.db, pAdjustMethod = Benjamini-Hochberg (BH). Tumour purity and tumour-infiltrating immune cell population abundance was estimated using two gene signature-based algorithms: ESTIMATE(v1.0.13) and xCell (v1.1.0) (Yoshihara et al. 2013; Aran et al. 2015). Tumour-infiltrating immune cell type deconvolution was performed using CIBERSORTx and EPIC (Newman et al. 2019; Racle and Gfeller 2020).

Differential transcript usage

Differential transcript usage analysis of DRS and PCS data was performed with RNAseqDTU (version 3.14) workflow, which employs both DRIMSeq and DEXSeq, followed by stageR statistical post-processing. Isoform quatification was scaled and normalized (dtuScaledTPM) before analysis. Analysis was performed on transcripts which had a minimum expression levels of 5 (normalized TPM) across all 12 tumour samples, with 5% of total gene expression in at least half of the samples in at least half of the samples. Genes with padj values below 0.1 were considered significant.

Transcriptome assembly and novel gene/isoform discovery

Using reference genome aligned PCR-cDNA-seq BAM files, transcript assembly was performed with StringTie2 (v2.2.1) and FLAIR (Kovaka et al. 2019). StringTie2 assembly

was performed with long-reads processing mode (-L), guided by reference gene annotation (Ensembl release 105). StringTie2 transcriptome assemblies from all sequenced nephrectomy specimens were then merged using the –merge option to generate transcript annotation file (GTF file). StringTie2 annotation used for novel gene mapping was performed by merging all nephrectomy assemblies with reference gene annotation (Ensembl release 105). FLAIR assembly was generated using the 'flair correct' and 'flair collapse' commands, with the long-read optimised option selected (--trust_ends). Generated transcript annotation files from StringTie2 and FLAIR were compared to Ensembl reference gene annotation (with –r option) where each assembled transcript was classified with a classcode using GffCompare (v0.12.6). In accordance with (Gleeson et al. 2022), transcripts were categorised into 3 main categories: 'Known' ('=' : Complete intron chain match, 'c': Partial intron chain match), 'Novel' ('j' : Multi-exon with at least 1 matched junction', 'k' : Containing reference, 'm' : Retained intron(s) – all covered, 'n' : Retained intron(s) – not all covered, 'i' : Contained within intron, 'o' : Overlapped exon, 'x' : Overlapped antisense, 'y' : Containing reference within intron, 'u' : None of above/Unknown), and 'Potential Artefacts' ('p' : No over-lap, 'e' : Single exon partially covering an intron, 's' :Intron matched on opposite strand, 'r' : Repeat).

StringTie2 assembled transcripts were also characterised by SQANTI3 (v5.1.2), which classifies genes as 'annotated' or 'novel', and isoforms as full splice match (FSM), incomplete splice match (ISM), novel in catalog (NIC), novel not in catalog (NNC), antisense, genic intron, genic genomic and intergenic (Pardo-Palacios et al. 2024). FSM represent isoforms with the exact same splice junctions and number of exons with the reference annotation. ISM represent isoforms with fewer exons from the 5' end but with the remaining internal splice junction sites matching with the reference annotation. NIC isoforms contain novel combinations of known splice junctions/exons compared with the reference annotation. NNC represents isoforms with at least one novel, unannotated splice site. In the SQANTI3 model, FSM and ISM represent the 'Known' transcripts, whereas NIC, NNC, antisense, genic intron, genic genomic and intergenic isoforms represent the 'Novel'

transcripts. SQANTI3 also predicts coding potential of transcripts using the GeneMarkS-T model (Tang et al. 2015). Integrative Genomics Viewer (IGV) tracks and reference genome mapped reads aligned to the region of novel genes and isoforms were visualized using IGV viewer (Robinson et al. 2011). Evidence of novel transcript expression was derived from analysis of the GTEx (Genotype-Tissue Expression) project RNA-seq data of normal tissues (The GTEx Consortium 2013) and LocExpress RNA-seq of cancer cell lines (Hou et al. 2016).

cDNA synthesis and qRTPCR

RNA molecules were reverse transcribed to cDNA molecules using oligo(dT) primer (Novagen 69896) and SuperScript II reverse transcriptase (Invitrogen 18064022). qPCR assays were performed using Fast SYBR Green master mix (Applied Biosystem 4385612) and pre-validated primers (Eurofins) on a StepOnePlus real-Time PCR system (Applied Biosystem) for 40 amplification cycles. Relative transcript levels were determined using the ΔΔCt (cycle threshold) method with *GAPDH* and *ACTB* used as loading controls. Details on the primers used can be found in **Supplemental Table S20**.

RNA stability assay

4 x 10^4 RCC4 cells were seeded in 12 well plates. 24 hours after seeding, media were changed into complete DMEM, with or without the addition of IFNG (1000U/mL) and TNF (25 ng/mL). 24 hours later, Actinomycin D (2 μg/mL, Generon) was added and cells were harvested after 0, 2, 4, and 8 hours of incubation for RNA extraction and qPCR. 3 wells were used as technical replicates for each biological replicate (n = 3) for each time point.

Statistical analysis

Statistical analysis was performed using GraphPad Prism 9. two-tailed Mann-Whitney *U* tests were used to compare non-parametric analysis of gene or transcript isoform expression levels, tumour purity estimations, immune scores and relative immune cell

populations between experimental groups, with $p \leq 0.05$ considered statistically significant. For comparison of more than two groups, Kruskal-Wallis test was used with $p \leq 0.05$ considered significant. For correlative analysis, R^2 (coefficient of determination) was used to calculate the goodness of fit between datasets, and P values were generated from F-test, with $p \leq 0.05$ considered statistically significant. Differential gene expression analysis by DESeq2 implements the Wald test, followed by false discovery rate correction by the Benjamini-Hochberg Method. Genes with padj < 0.05 and $|\log_2\text{FoldChange}| \geq 2$ are considered to be significantly differentially expressed. All p values of non-significant results are indicated in graphs.

Data access

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession numbers: GSE242204 (PCS), GSE241932 (DRS), GSE242084 (RCC4 DRS), GSE246408 (RCC4 short-read Illumina sequencing). The workflows for read alignment are available at GitHub (https://github.com/joshuacylee/DRS and https://github.com/joshuacylee/PCR-cDNAseq) and in Supplemental Code. Workflow for differential gene expression identification is available at GitHub (https://github.com/joshuacylee/DESeq2) and as Supplemental Code.

Competing interest statement

E.A.S. and D.J.T. are employees of and stock option holders in Oxford Nanopore Technologies. As of November 2023, J.L. is also an employee of Oxford Nanopore technologies.

N.S.V. has received grants, speaker honoraria and/or advisory fees from Bristol Myers Squibb, Ipsen, EUSA pharma, Eisai and Pfizer, all outside the submitted work. The remaining authors declare no competing interests.

Acknowledgements

This work was funded the Biotechnology and Biological Sciences Research Council (BBSRC) White Rose doctoral training partnership (BB/J014443/1) through an Industrial Cooperative Awards in Science and Engineering (iCASE) studentship supported by Oxford Nanopore Technologies. Additional support was provided by the Hull York Medical School and Oxford Nanopore Technologies. We are indebted to the study participants and their families for contributing to medical research. We thank Aino Järvelin for support with bioinformatics analyses of long-read sequencing data. We thank staff at the Genomics Lab in the University of York Bioscience Technology Facility for technical assistance with short-read Illumina sequencing.

Authors contributions: J.L. contributed to experimental design, data generation, data analysis, figure design, and manuscript writing. E.A.S. contributed to experimental design; data generation. C.E.B. contributed to data generation. J.B. and R.E.B. managed clinical sample and data collection and maintenance. D.J.T. assisted with study conception and design. N.S.V. contributed to clinical sample and data collection, study design, and data interpretation. D.L. assisted with study conception, study design, data interpretation, study supervision and manuscript writing. All authors read and approved the final manuscript.

References

Anders S, Reyes A, Huber W. 2012. Detecting differential usage of exons from RNA-seq data. *Genome Res* **22**: 2008-2017.

Aran D, Sirota M, Butte AJ. 2015. Systematic pan-cancer analysis of tumour purity. *Nat Commun* **6**: 8971.

Bayega A, Oikonomopoulos S, Wang YC, Ragoussis J. 2022. Improved Nanopore full-length cDNA sequencing by PCR-suppression. *Front Genet* **13**: 1031355.

Brannon AR, Reddy A, Seiler M, Arreola A, Moore DT, Pruthi RS, Wallen EM, Nielsen ME, Liu H, Nathanson KL et al. 2010. Molecular Stratification of Clear Cell Renal Cell Carcinoma

769 by Consensus Clustering Reveals Distinct Subtypes and Survival Patterns. *Genes Cancer* **1**:
770 152-163.
771 Chang A, Chakiryan NH, Du D, Stewart PA, Zhang Y, Tian Y, Soupir AC, Bowers K, Fang B,
772 Morganti A et al. 2022. Proteogenomic, Epigenetic, and Clinical Implications of Recurrent
773 Aberrant Splice Variants in Clear Cell Renal Cell Carcinoma. *Eur Urol* **82**: 354-362.
774 Choueiri TK, Tomczak P, Park SH, Venugopal B, Ferguson T, Symeonides SN, Hajek J,
775 Chang YH, Lee JL, Sarwar N et al. 2024. Overall Survival with Adjuvant Pembrolizumab in
776 Renal-Cell Carcinoma. *N Engl J Med* **390**: 1359-1371.
777 Correa AF, Jegede O, Haas NB, Flaherty KT, Pins MR, Messing EM, Manola J, Wood CG,
778 Kane CJ, Jewett MAS et al. 2019. Predicting Renal Cancer Recurrence: Defining Limitations
779 of Existing Prognostic Models With Prospective Trial-Based Validation. *J Clin Oncol* **37**:
780 2062-2071.
781 Cortes-Lopez M, Chamely P, Hawkins AG, Stanley RF, Swett AD, Ganesan S, Mouhieddine
782 TH, Dai X, Kluegel L, Chen C et al. 2023. Single-cell multi-omics defines the cell-type-
783 specific impact of splicing aberrations in human hematopoietic clonal outgrowths. *Cell Stem*
784 *Cell* doi:10.1016/j.stem.2023.07.012.
785 Cotta BH, Choueiri TK, Cieslik M, Ghatalia P, Mehra R, Morgan TM, Palapattu GS, Shuch B,
786 Vaishampayan U, Van Allen E et al. 2023. Current Landscape of Genomic Biomarkers in
787 Clear Cell Renal Cell Carcinoma. *Eur Urol* **84**: 166-175.
788 Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, Pantic N, Admassu T,
789 James P, Warland A et al. 2018. Highly parallel direct RNA sequencing on an array of
790 nanopores. *Nat Methods* **15**: 201-206.
791 Ghatalia P, Gordetsky J, Kuo F, Dulaimi E, Cai KQ, Devarajan K, Bae S, Naik G, Chan TA,
792 Uzzo R et al. 2019. Prognostic impact of immune gene expression signature and tumor
793 infiltrating immune cells in localized clear cell renal cell carcinoma. *J Immunother Cancer* **7**:
794 139.

Gorin MA, Patel HD, Rowe SP, Hahn NM, Hammers HJ, Pons A, Trock BJ, Pierorazio PM, Nirschl TR, Salles DC et al. 2022. Neoadjuvant Nivolumab in Patients with High-risk Nonmetastatic Renal Cell Carcinoma. *Eur Urol Oncol* **5**: 113-117.

Hou M, Tian F, Jiang S, Kong L, Yang D, Gao G. 2016. LocExpress: a web server for efficiently estimating expression of novel transcripts. *BMC Genomics* **17**: 1023.

Hsieh JJ, Le VH, Oyama T, Ricketts CJ, Ho TH, Cheng EH. 2018. Chromosome 3p Loss-Orchestrated VHL, HIF, and Epigenetic Deregulation in Clear Cell Renal Cell Carcinoma. *J Clin Oncol* **36**: JCO2018792549.

Jain M, Abu-Shumays R, Olsen HE, Akeson M. 2022. Advances in nanopore direct RNA sequencing. *Nat Methods* **19**: 1160-1164.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**: 907-915.

Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. 2019. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol* **20**: 278.

Krause M, Niazi AM, Labun K, Torres Cleuren YN, Muller FS, Valen E. 2019. tailfindr: alignment-free poly(A) length measurement for Oxford Nanopore RNA and DNA sequencing. *RNA* **25**: 1229-1241.

Larrinaga G, Solano-Iturri JD, Errarte P, Unda M, Loizaga-Iriarte A, Perez-Fernandez A, Echevarria E, Asumendi A, Manini C, Angulo JC et al. 2021. Soluble PD-L1 Is an Independent Prognostic Factor in Clear Cell Renal Cell Carcinoma. *Cancers (Basel)* **13**.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.

Mahoney KM, Ross-Macdonald P, Yuan L, Song L, Veras E, Wind-Rotolo M, McDermott DF, Stephen Hodi F, Choueiri TK, Freeman GJ. 2022. Soluble PD-L1 as an early marker of progressive disease on nivolumab. *J Immunother Cancer* **10**.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet* **17**: 10-12.

Mock A, Braun M, Scholl C, Frohling S, Erkut C. 2023. Transcriptome profiling for precision cancer medicine using shallow nanopore cDNA sequencing. *Sci Rep* **13**: 2378.

Morgan TM, Mehra R, Tiemeny P, Wolf JS, Wu S, Sangale Z, Brawer M, Stone S, Wu CL, Feldman AS. 2018. A Multigene Signature Based on Cell Cycle Proliferation Improves Prediction of Mortality Within 5 Yr of Radical Nephrectomy for Renal Cell Carcinoma. *Eur Urol* **73**: 763-769.

Motzer RJ, Russo P, Grunwald V, Tomita Y, Zurawski B, Parikh O, Buti S, Barthelemy P, Goh JC, Ye D et al. 2023. Adjuvant nivolumab plus ipilimumab versus placebo for localised renal cell carcinoma after nephrectomy (CheckMate 914): a double-blind, randomised, phase 3 trial. *Lancet* **401**: 821-832.

Nature Methods Editors. 2023. Method of the Year 2022: long-read sequencing. *Nat Methods* **20**: 1.

Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, Khodadoust MS, Esfahani MS, Luca BA, Steiner D et al. 2019. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol* **37**: 773-782.

Nowicka M, Robinson MD. 2016. DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. *F1000Res* **5**: 1356.

Pardo-Palacios FJ, Arzalluz-Luque A, Kondratova L, Salguero P, Mestre-Tomas J, Amorin R, Estevan-Morio E, Liu T, Nanni A, McIntyre L et al. 2024. SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. *Nat Methods* **21**: 793-797.

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**: 417-419.

Peng YL, Xiong LB, Zhou ZH, Ning K, Li Z, Wu ZS, Deng MH, Wei WS, Wang N, Zou XP et al. 2022. Single-cell transcriptomics reveals a low CD8(+) T cell infiltrating state mediated by fibroblasts in recurrent renal cell carcinoma. *J Immunother Cancer* **10**.

Pratanwanich PN, Yao F, Chen Y, Koh CWQ, Wan YK, Hendra C, Poon P, Goh YT, Yap PML, Chooi JY et al. 2021. Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore. *Nat Biotechnol* **39**: 1394-1402.

Qu H, Wang Z, Zhang Y, Zhao B, Jing S, Zhang J, Ye C, Xue Y, Yang L. 2022. Long-Read Nanopore Sequencing Identifies Mismatch Repair-Deficient Related Genes with Alternative Splicing in Colorectal Cancer. *Dis Markers* **2022**: 4433270.

R Core Team. 2022. R: A Language and Environment for Statistical Computing. *https://wwwR-projectorg*.

Racle J, Gfeller D. 2020. EPIC: A Tool to Estimate the Proportions of Different Cell Types from Bulk Gene Expression Data. *Methods Mol Biol* **2120**: 233-248.

Ricketts CJ, De Cubas AA, Fan H, Smith CC, Lang M, Reznik E, Bowlby R, Gibb EA, Akbani R, Beroukhim R et al. 2018. The Cancer Genome Atlas Comprehensive Molecular Characterization of Renal Cell Carcinoma. *Cell Rep* **23**: 3698.

Rini B, Goddard A, Knezevic D, Maddala T, Zhou M, Aydin H, Campbell S, Elson P, Koscielny S, Lopatin M et al. 2015. A 16-gene assay to predict recurrence after surgery in localised renal cell carcinoma: development and validation studies. *Lancet Oncol* **16**: 676-685.

Rini BI, Campbell SC, Escudier B. 2009. Renal cell carcinoma. *Lancet* **373**: 1119-1132.

Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29**: 24-26.

Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N. 2015. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* **160**: 48-61.

Sakamoto Y, Sereewattanawoot S, Suzuki A. 2020. A new era of long-read sequencing for cancer genomics. *J Hum Genet* **65**: 3-10.

Scelo G, Riazalhosseini Y, Greger L, Letourneau L, Gonzalez-Porta M, Wozniak MB, Bourgey M, Harnden P, Egevad L, Jackson SM et al. 2014. Variation in genomic landscape of clear cell renal cell carcinoma across Europe. *Nat Commun* **5**: 5135.

878 Smith CC, Selitsky SR, Chai S, Armistead PM, Vincent BG, Serody JS. 2019. Alternative
879 tumour-specific antigens. *Nat Rev Cancer* **19**: 465-478.
880 Sun C, Mezzadra R, Schumacher TN. 2018. Regulation and Function of the PD-L1
881 Checkpoint. *Immunity* **48**: 434-452.
882 Sveen A, Johannessen B, Teixeira MR, Lothe RA, Skotheim RI. 2014. Transcriptome
883 instability as a molecular pan-cancer characteristic of carcinomas. *BMC Genomics* **15**: 672.
884 Tang AD, Soulette CM, van Baren MJ, Hart K, Hrabeta-Robinson E, Wu CJ, Brooks AN.
885 2020. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic
886 leukemia reveals downregulation of retained introns. *Nat Commun* **11**: 1438.
887 Tang S, Lomsadze A, Borodovsky M. 2015. Identification of protein coding regions in RNA
888 transcripts. *Nucleic Acids Res* **43**: e78.
889 The GTEx Consortium. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*
890 **45**: 580-585.
891 Vasudev NS, Hutchinson M, Trainor S, Ferguson R, Bhattarai S, Adeyoju A, Cartledge J,
892 Kimuli M, Datta S, Hanbury D et al. 2020. UK Multicenter Prospective Evaluation of the
893 Leibovich Score in Localized Renal Cell Carcinoma: Performance has Altered Over Time.
894 *Urology* **136**: 162-168.
895 Vasudev NS, Scelo G, Glennon KI, Wilson M, Letourneau L, Eveleigh R, Nourbehesht N,
896 Arseneault M, Paccard A, Egevad L et al. 2023. Application of Genomic Sequencing to
897 Refine Patient Stratification for Adjuvant Therapy in Renal Cell Carcinoma. *Clin Cancer Res*
898 **29**: 1220-1231.
899 Veiga DFT, Nesta A, Zhao Y, Deslattes Mays A, Huynh R, Rossi R, Wu TC, Palucka K,
900 Anczukow O, Beck CR et al. 2022. A comprehensive long-read isoform analysis platform
901 and sequencing resource for breast cancer. *Sci Adv* **8**: eabg6711.
902 Viscardi MJ, Arribere JA. 2022. Poly(a) selection introduces bias and undue noise in direct
903 RNA-sequencing. *BMC Genomics* **23**: 530.
904 Wang D, Liu B, Zhang Z. 2023. Accelerating the understanding of cancer biology through
905 the lens of genomics. *Cell* **186**: 1755-1771.

Yamaguchi H, Hsu JM, Yang WH, Hung MC. 2022. Mechanisms regulating PD-L1 expression in cancers and associated opportunities for novel small-molecule therapeutics. *Nat Rev Clin Oncol* **19**: 287-305.

Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, Torres-Garcia W, Trevino V, Shen H, Laird PW, Levine DA et al. 2013. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun* **4**: 2612.

Figure and Table Legends

Figure 1. DRS and PCS of ccRCC nephrectomy samples. (*A*) Summary of study design and data analysis workflow – figure made with Biorender. (*B*) Violin plot showing Log_{10} transformed raw read lengths of passed reads generated by PCS. (*C*) as in (B), but for DRS. (*D*) Pie chart depicting the proportions of gene biotypes of all mapped genes from reference genome (Ensembl release 105, GRCh38) mapped PCS and DRS reads of sequenced tumour samples. (*E*) Venn diagram showing the overlap between PCS and DRS mapped genes. (*F*) Correlation between gene expression levels (Log_{10} Reads per million (RPM)) of all genes mapped by both PCS and DRS (n = 25,692). Diagonal line represents the line of best fit. R^2 value was computed to measure goodness-of-fit and p value was generated from F-test, with $p<0.05$ considered statistically significant. Lowest expression values shown correspond to the minimum normalised abundance derived for genes detected only at 1 read in the sample with the highest total number of reads.

Figure 2. ccRCC recurrence is associated with suppressed tumour immune infiltration. (*A*) Volcano plots showing differentially expressed genes (red) between recurrent and nonrecurrent ccRCC tumours from PCS and DRS data using Ensembl genome reference (Ensembl release 105). (*B*) Venn diagram showing overlaps of DEGs identified by both PCS and DRS. (*C*) Correlation between log_2FoldChange of DEGs identified by either or both PCS and DRS (recurrent vs nonrecurrent ccRCC tumours). Diagonal line represents the line of best fit. R^2 value was computed to measure goodness-of-fit and p value was generated from

F-test, with p ≤ 0.05 considered statistically significant. (*D*) Volcano plots showing differentially expressed genes (red) between recurrent and nonrecurrent ccRCC tumours from 5% subsampled PCS data using Ensembl genome reference (Ensembl release 105). (*E*) Venn diagram showing overlaps of DEGs identified by both 5% subsampled PCS and DRS. (*F*) Grouped dot plot showing estimated immune score of nonrecurrent (blue) and recurrent (red) ccRCC tumour by the ESTIMATE algorithm, using PCS gene expression data. (*G*) Grouped dot plot showing relative population of CD8+ T cells within immune infiltrates of nonrecurrent (blue) and recurrent (red) ccRCC tumours estimated by CIBERSORTx using PCS gene expression data. (H) *CD8B*, *TOX*, *PD-1*, *GZMK* and *LINC02416* mRNA levels measured by qRT-PCR in recurrent and nonrecurrent tumours from sequenced cohort (blue and red, n = 12) and validation cohort (black, n = 20), relative to average mRNA levels in nonrecurrent tumours. mRNA levels were normalised to *GAPDH* and *ACTB*. For (*A*) and (*D*), Blue and red dots represent significantly down- and upregulated genes by either or both PCS and DRS. Dotted lines indicate significance threshold ($|\log_2 \text{FoldChange}| \geq 2$, padj ≤ 0.1). Names of genes that were validated by qRTPCR with validation cohort are shown. For (*F*) - (*H*), two-tailed Mann-Whitney *U* tests were used with p ≤ 0.05 considered significant. * = p < 0.05, ** = p < 0.01, **** = p < 0.0001. Line represents median for each group.

Figure 3. Differential transcript usage events associated with ccRCC recurrence. (*A*) Venn diagram showing overlaps between reference genome- and reference transcriptome-alignment method mapped genes in PCS and DRS of nephrectomy samples (*B*) Correlation between gene expression levels (Log_{10} Reads per million (RPM)) of all genes mapped by both reference genome alignment method and reference transcriptome alignment method in PCS (n = 35,797) and DRS (n = 22,038). Diagonal line represents the line of best fit. r value denote Pearson's correlation coefficient and p value was generated from F-test, with p<0.05 considered statistically significant. (*C*) Venn diagram showing the overlaps of DEGs identified by both between reference genome- and reference transcriptome- alignment

962 method in PCS and DRS of nephrectomy samples (*D*) Venn diagram showing the overlaps
963 of genes that displayed significant DTU by DRIMSeq and DEXSeq in PCS nephrectomy
964 samples (*E*) Stack bar graphs representing proportions of *CMC1* isoforms in ccRCC tumours
965 using PCS data. DRIMSeq and DEXSeq padj values for DTU of *CMC1* are indicated in
966 graph. (*F*) Graphical representation of CMC1 isoforms Ensembl reference annotations in
967 IGV, with black boxes representing exons. Sashimi plots of *CMC1* from PCS and DRS
968 recurrent (135) and nonrecurrent (273) ccRCC samples. Junction lines are shown for
969 junction coverages with at least 5% of total *CMC1* reads.
970
971 **Figure 4.** Long-read RNA sequencing enables the discovery of full-length novel transcripts.
972 (*A*) Graphical representation of the major SQANTI3 isoform categories (antisense, genic
973 intron, genic genomic and intergenic not shown here). (*B*) Bar chart showing the proportion
974 of Novel and known transcripts in StringTie2 assembly as curated by SQANTI3 and
975 gffcompare. (*C*) Pie chart depicting the distribution of SQANTI3 isoform categories amongst
976 StringTie2 assembled transcripts (n = 54,185). (*D*) Bar chart showing the proportion of
977 coding and non-coding StringTie2 assembled transcripts by SQANTI3 isoform categories.
978 (*E*) Venn diagram showing the number of overlapping mapped StringTie2 novel transcripts
979 between PCS and DRS of ccRCC tumour samples, and DRS of ccRCC cell line RCC4. (*F*)
980 Violin plot showing the expression levels (Log_2RPM) of known and novel transcripts in PCS
981 and DRS of ccRCC tumour samples, and DRS of ccRCC cell line RCC4. The width of the
982 violin plots represents the density of transcripts at different expression levels. Black dots
983 represent mean expression levels. The top and bottom of box plots represent upper and
984 lower quartiles, respectively. (*G*) Correlation between transcripts expression levels (Log_{10}
985 Reads per million (RPM)) of all StringTie2 novel transcripts mapped by both PCS and DRS
986 (n = 14,544). Diagonal line represents the line of best fit. R^2 value was computed to measure
987 goodness-of-fit and p value was generated from F-test, with $p<0.05$ considered statistically
988 significant. Lowest expression values shown correspond to the minimum normalised
989 abundance. (*H*) IGV visualisation of *MVK* reference annotations (blue), ccRCC specific *MVK*

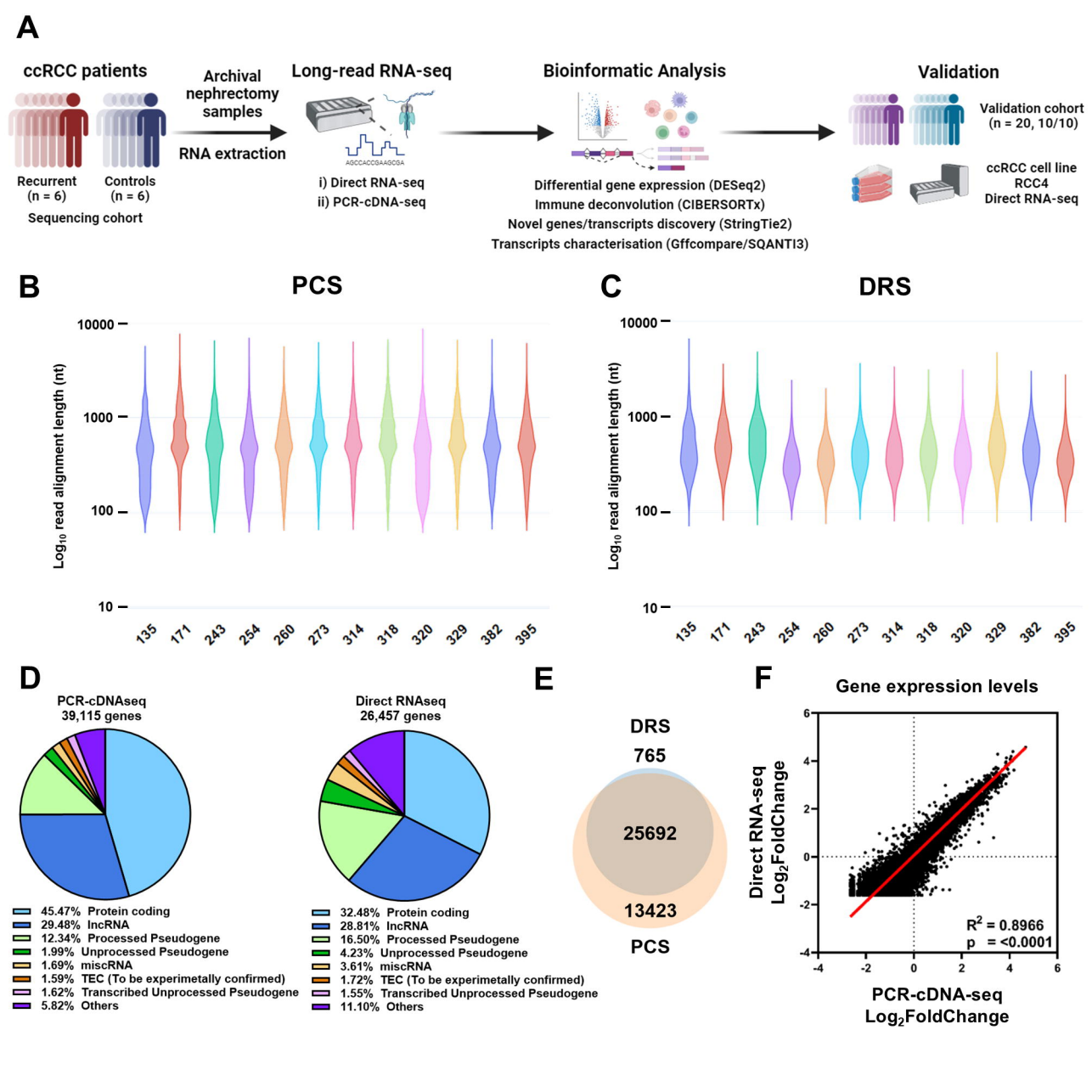
splice junctions (black), StringTie2 assembled novel transcripts (green), PCS coverage track (grey) illustrating the depth of sequence coverage across the region of interest (red bar, hg38 Chr12:109,594,200 – 109,598,600) and PCS sequencing reads aligned to the reference genome in the region of interest.

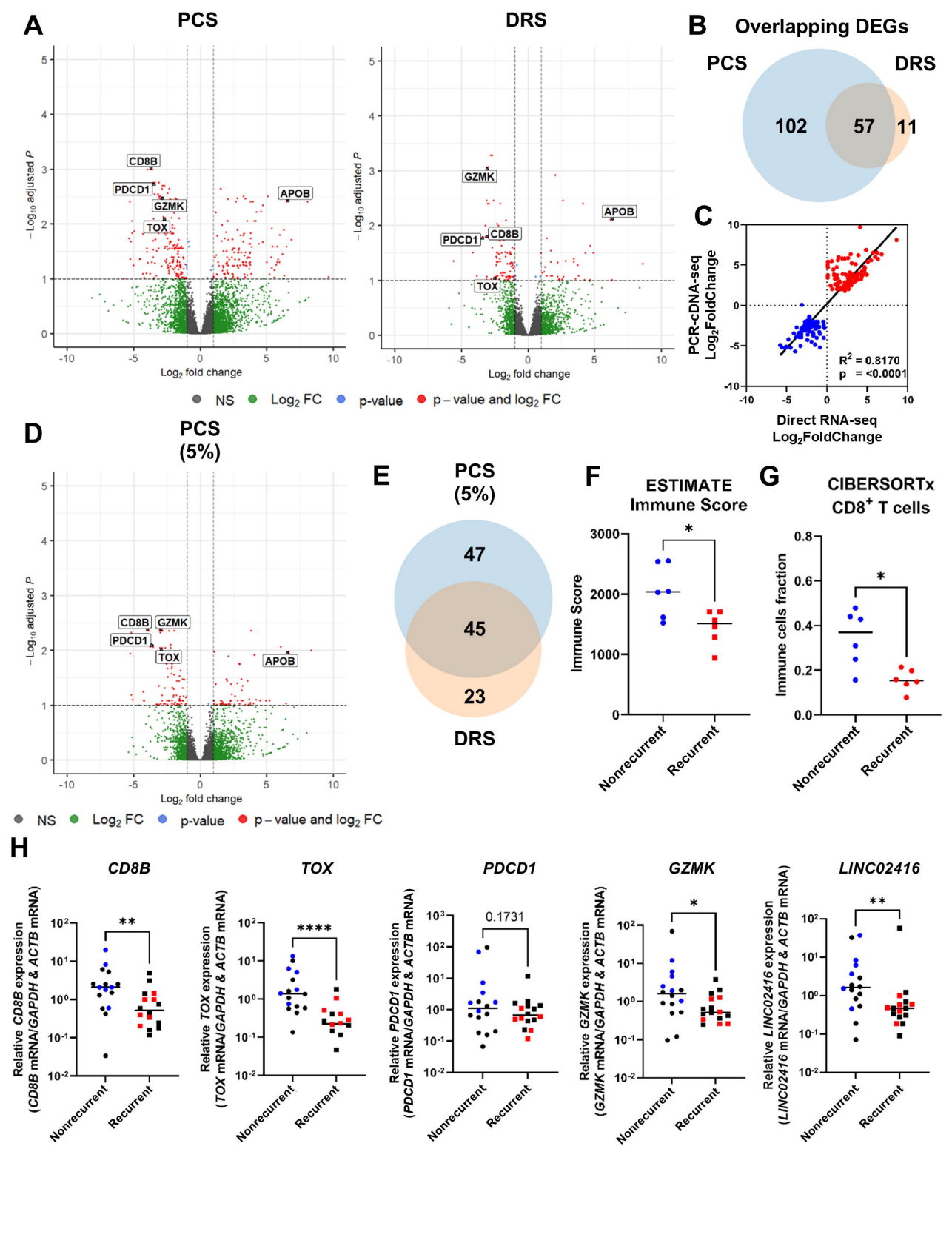
Figure 5. Discovery of a novel *sPD-L1* isoform expressed by ccRCC tumour cells. (*A*) IGV visualisation of reference annotation of *mPD-L1* isoform (black, ENST00000381577), *sPD-L1* (black, NM_001314029) and StringTie2 reference annotation (green) (Top tracks); Graphical representation of membrane, soluble and novel soluble *PD-L1* exon 4; Ensembl (black) and StringTie2 reference annotations (green) and IGV coverage tracks for PCS of ccRCC tumours (red) and DRS of RCC4 (green). (*B*) Grouped dot plot showing reference DESeq2 normalised *PD-L1* expression in nonrecurrent (blue) and recurrent (red) tumours' PCS data. DESeq2 padj value is shown in graph. Centre line represents median for each group. (*C*) Grouped dot plots showing normalised *mPD-L1*, *sPD-L1* and novel *sPD-L1* expression ($\log_2$(RPM+1)) in nonrecurrent (blue) and recurrent (red) tumours' PCS data. (*D*) Grouped dot plots showing mPD-L1, sPD-L1 (all isoforms) and novel sPD-L1 mRNA levels measured by qRT-PCR in recurrent and nonrecurrent tumours from sequenced cohort (blue and red, n = 12) and validation cohort (black, n = 20) relative to average mRNA levels in nonrecurrent tumours. (*E*) Stacked bar graphs representing proportions of *mPD-L1*, *sPD-L1* and novel *sPD-L1* isoforms in RCC4 cells based on DRS data. For (*C*) - (*E*), two-tailed Mann-Whitney *U* tests were used with $p \leq 0.05$ considered significant. * = $p < 0.05$. Centre line represents median for each group. (F) mRNA decay curves for *mPD-L1*, *sPD-L1* and novel *sPD-L1* in unstimulated (blue) and IFNG + TNF treated (red) RCC4 cells. Half-lives of isoforms are indicated in graph (blue for unstimulated, red for IFNG+TNF treated RCC4). Comparisons were made using unpaired student's *t*-test with $p \leq 0.05$ considered significant. n.s. = not significant, * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$.

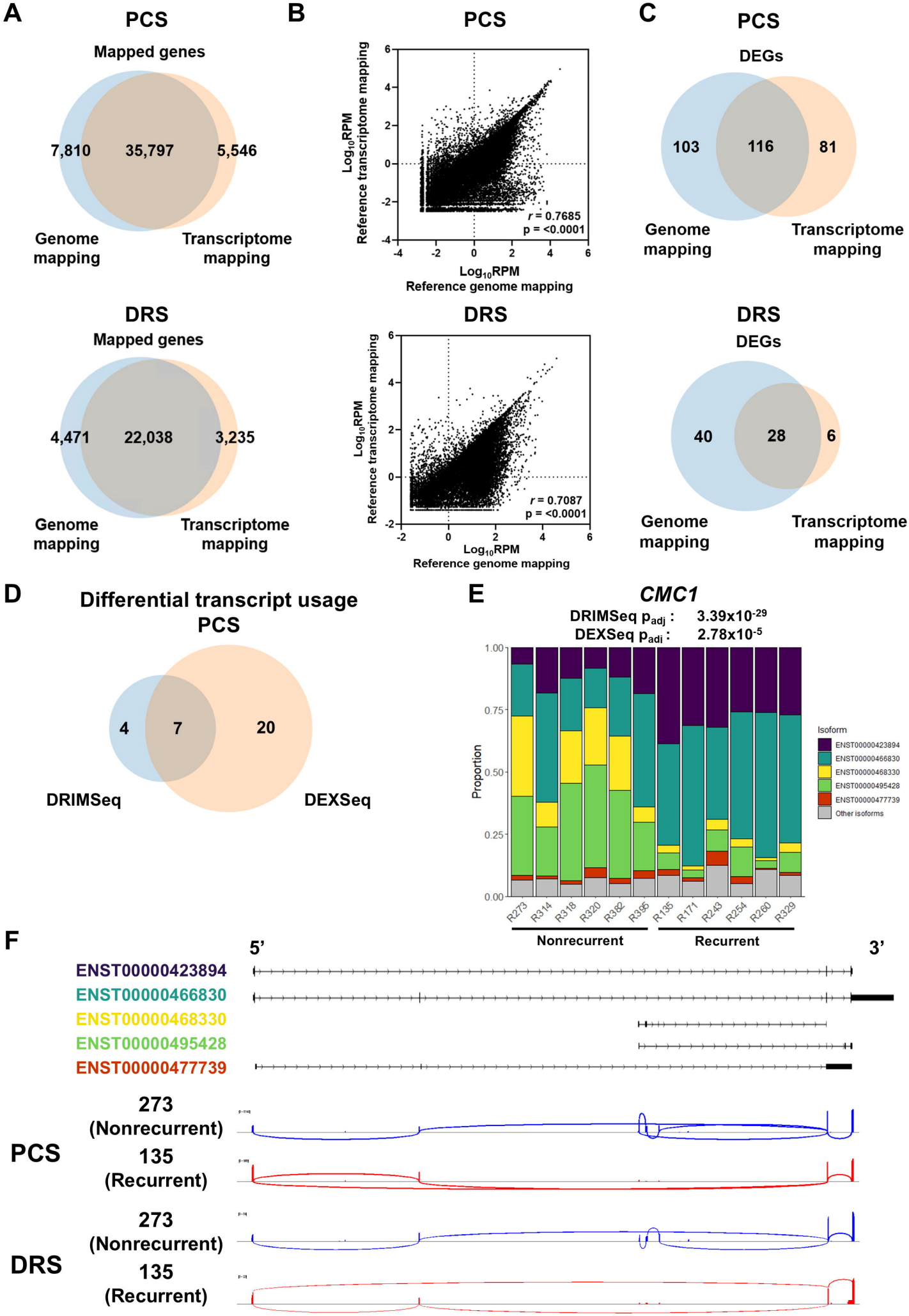
Figure 6. Discovery of ccRCC recurrence associated novel genes by long-read RNA-seq. (*A*) Bar chart showing the isoform classifications of StringTie2 assembled transcripts from novel genes as classified by SQANTI3. (*B*) Pie chart illustrating the proportion of coding and non-coding StringTie2 assembled transcripts from novel genes as classified by SQANTI3. (*C*) Venn diagram showing the number of overlapping mapped novel genes between PCS and DRS of ccRCC tumour samples, and DRS of RCC4. (*D*) Volcano plots showing differentially expressed genes (red) between recurrent and nonrecurrent tumours from PCS and DRS data using StringTie2 assembled reference. Number of differentially expressed novel and known genes are shown in table below plots. Names of novel genes that were validated by qPCR with validation cohort are shown on plots. (*E*) Correlation between $\log_2$FoldChange of differentially expressed novel genes identified by either or both PCS and DRS between recurrent vs nonrecurrent tumours (n = 40). (*F*) IGV visualisation of *MSTRG.29728* isoforms StringTie2 reference annotation (green) and the closest neighbouring genes (*LINC01170* and *CSNK1G3*) in the Ensembl reference annotation (Ensembl release 105) at Chr5:123,500,000-124,300,000 (Top track); Sashimi plot showing abundance of reference genome aligned reads and splicing patterns along *MSTRG.29728* (Chr5:123,859,000-123,868,000) for PCS (red) and DRS (blue) of ccRCC tumour samples, and DRS (green) and short-read Illumina sequencing (orange) of RCC4; Representative PCS sequencing reads (grey) aligned to the reference genome in the region of interest. (*G*) *MSTRG.29728* mRNA levels measured by qRT-PCR in recurrent and nonrecurrent tumours from sequenced cohort (blue and red, n = 12) and validation cohort (black, n = 20), relative to average mRNA levels in nonrecurrent tumours. mRNA levels were normalised to *GAPDH* and *ACTB*. Two-tailed Mann-Whitney *U* test was used with $p \leq 0.05$ considered significant. * = $p < 0.05$. Centre line represents median for each group.

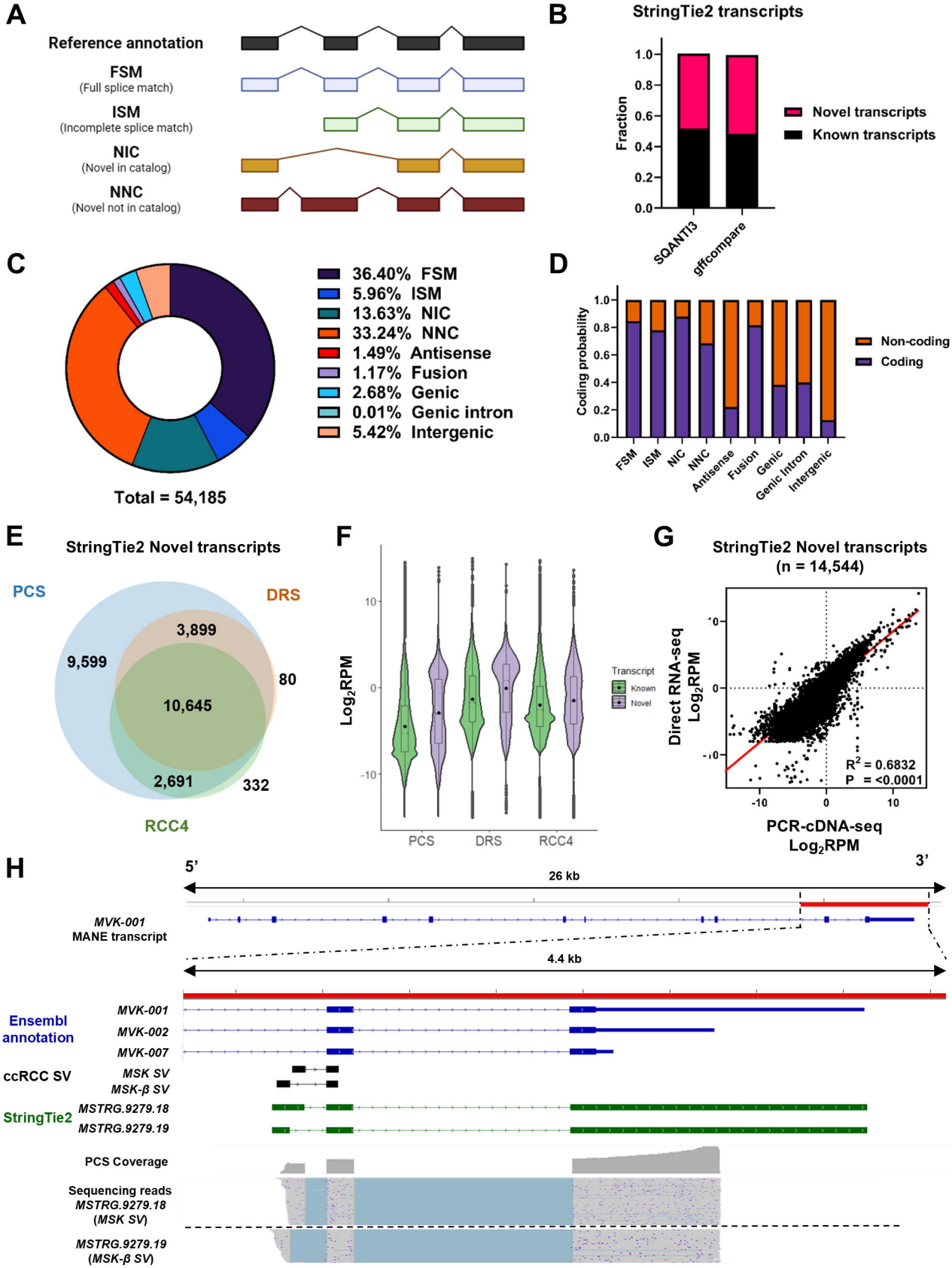
Table 1. Sequencing statistics of PCS and DRS of archival ccRCC tumour samples. Tables showing the number of passed reads (Q >7), median reference genome (Ensembl release 105, GRCh38) alignment length (nt), median read accuracy (%), and percentage of reads

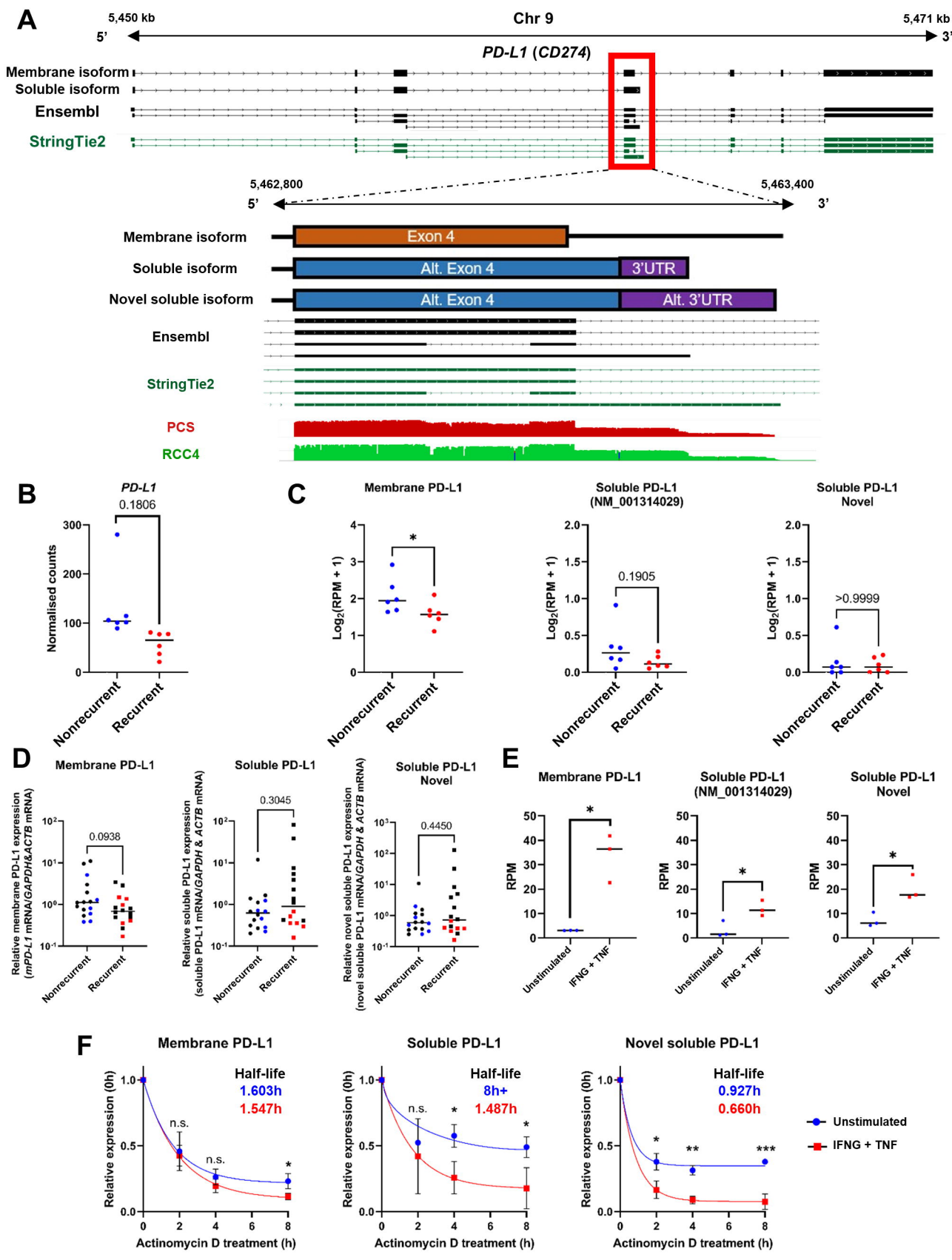
1045 representing full-length transcripts (95%+ coverage of reference transcript isoform) of
1046 sequenced archival ccRCC tumour samples.
1047

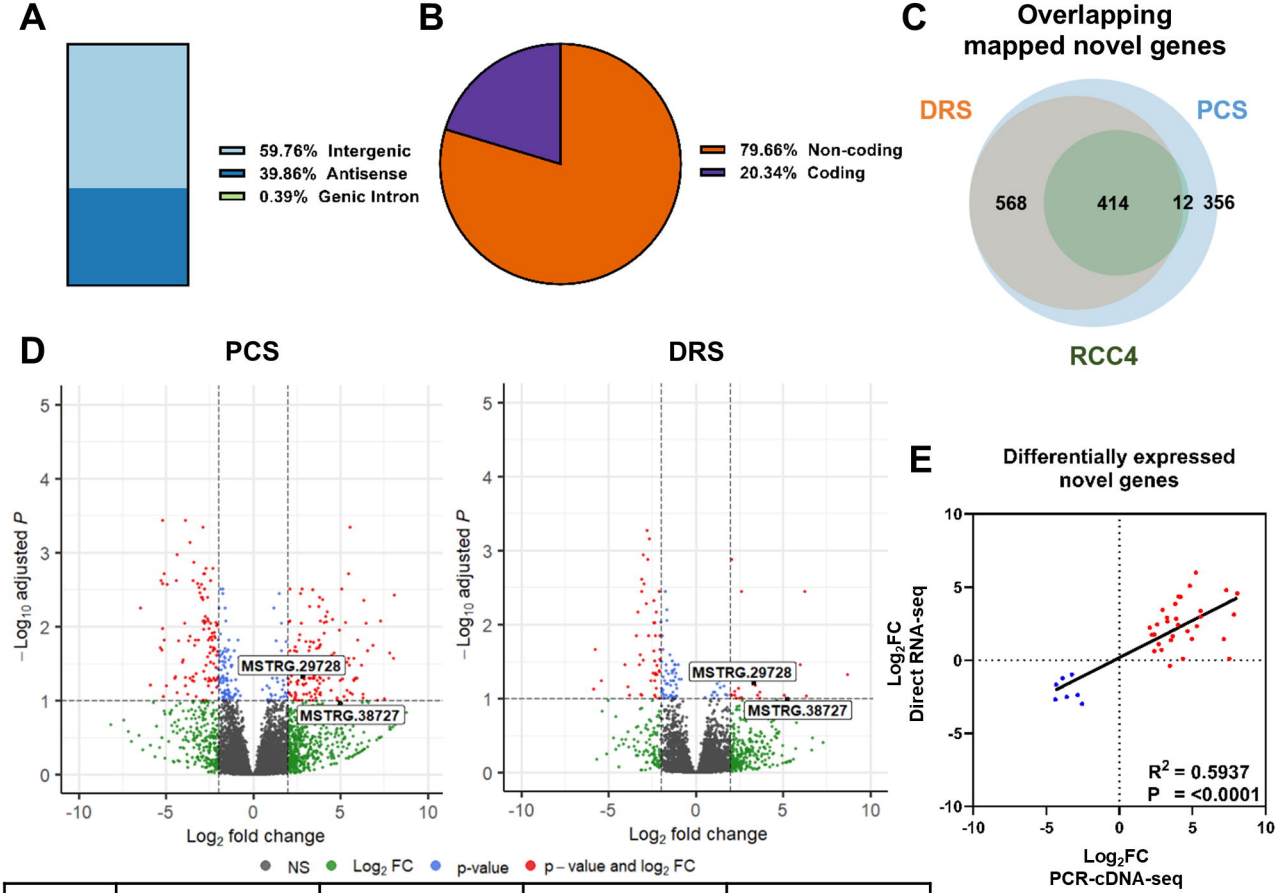

A
ccRCC patients
Recurrent (n = 6)
Controls (n = 6)
Sequencing cohort
Archival nephrectomy samples
RNA extraction
Long-read RNA-seq
i) Direct RNA-seq
ii) PCR-cDNA-seq
Bioinformatic Analysis
Differential gene expression (DESeq2)
Immune deconvolution (CIBERSORTx)
Novel genes/transcripts discovery (StringTie2)
Transcripts characterisation (Gffcompare/SQANTI3)
Validation
Validation cohort (n = 20, 10/10)
ccRCC cell line RCC4 Direct RNA-seq
B
PCS
C
DRS
Log10 read alignment length (nt)
10000
1000
100
10
135
171
243
254
260
273
314
318
320
329
382
395
D
PCR-cDNAseq 39,115 genes
45.47% Protein coding
29.48% lncRNA
12.34% Processed Pseudogene
1.99% Unprocessed Pseudogene
1.69% miscRNA
1.59% TEC (To be experimetally confirmed)
1.62% Transcribed Unprocessed Pseudogene
5.82% Others
Direct RNAseq 26,457 genes
32.48% Protein coding
28.81% lncRNA
16.50% Processed Pseudogene
4.23% Unprocessed Pseudogene
3.61% miscRNA
1.72% TEC (To be experimetally confirmed)
1.55% Transcribed Unprocessed Pseudogene
11.10% Others
E
DRS
765
25692
13423
PCS
F
Gene expression levels
Direct RNA-seq Log2FoldChange
PCR-cDNA-seq Log2FoldChange
R2 = 0.8966
p = <0.0001


A
PCS
DRS
−Log10 adjusted P
Log2 fold change
CD8B
PDCD1
GZMK
TOX
APOB
NS
Log2 FC
p-value
p – value and log2 FC
B
Overlapping DEGs
PCS
DRS
102
57
11
C
PCR-cDNA-seq
Log2FoldChange
Direct RNA-seq
Log2FoldChange
R2 = 0.8170
p = <0.0001
D
PCS
(5%)
E
PCS
(5%)
47
45
23
DRS
F
ESTIMATE
Immune Score
Immune Score
Nonrecurrent
Recurrent
G
CIBERSORTx
CD8+ T cells
Immune cells fraction
H
CD8B
TOX
PDCD1
GZMK
LINC02416
Relative CD8B expression
(CD8B mRNA/GAPDH & ACTB mRNA)
Relative TOX expression
(TOX mRNA/GAPDH & ACTB mRNA)
Relative PDCD1 expression
(PDCD1 mRNA/GAPDH & ACTB mRNA)
Relative GZMK expression
(GZMK mRNA/GAPDH & ACTB mRNA)
Relative LINC02416 expression
(LINC02416 mRNA/GAPDH & ACTB mRNA)
0.1731
Nonrecurrent
Recurrent

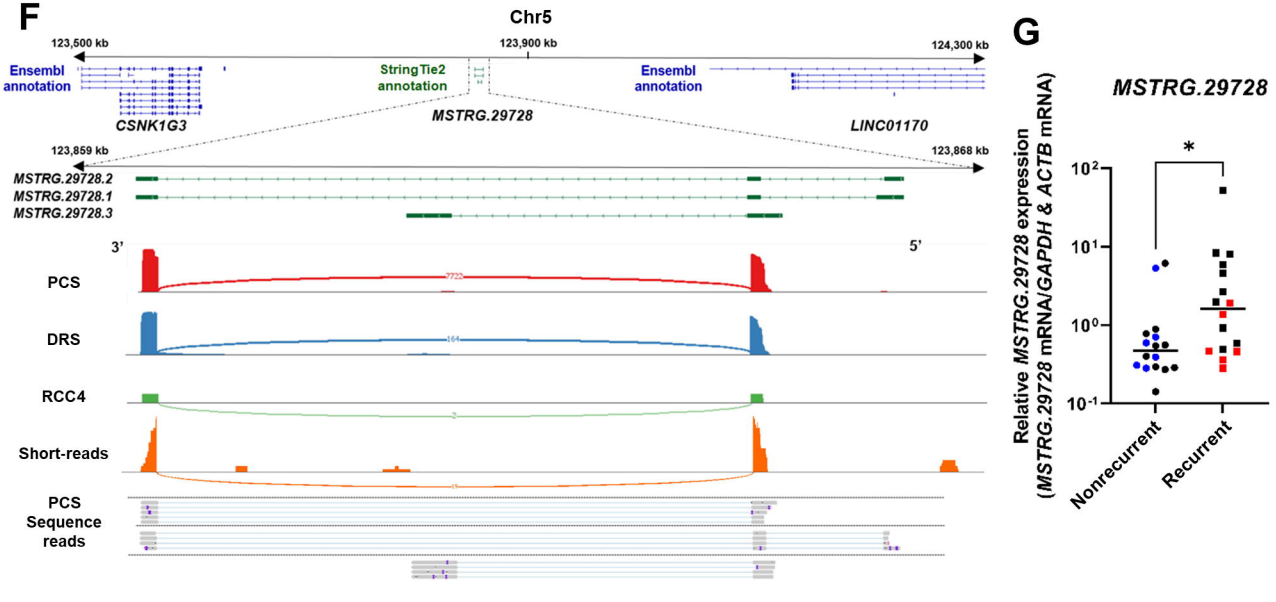



A
Reference annotation
FSM
(Full splice match)
ISM
(Incomplete splice match)
NIC
(Novel in catalog)
NNC
(Novel not in catalog)
B
StringTie2 transcripts
Fraction
1.0
0.8
0.6
0.4
0.2
0.0
SQANTI3
gffcompare
Novel transcripts
Known transcripts
C
36.40% FSM
5.96% ISM
13.63% NIC
33.24% NNC
1.49% Antisense
1.17% Fusion
2.68% Genic
0.01% Genic intron
5.42% Intergenic
Total = 54,185
D
Coding probability
1.0
0.8
0.6
0.4
0.2
0.0
FSM
ISM
NIC
NNC
Antisense
Fusion
Genic
Genic Intron
Intergenic
Non-coding
Coding
E
StringTie2 Novel transcripts
PCS
DRS
RCC4
3,899
9,599
80
10,645
2,691
332
F
Log2RPM
10
0
-10
PCS
DRS
RCC4
Transcript
Known
Novel
G
StringTie2 Novel transcripts
(n = 14,544)
Direct RNA-seq
Log2RPM
PCR-cDNA-seq
Log2RPM
R2 = 0.6832
P = <0.0001
H
5'
3'
26 kb
MVK-001
MANE transcript
4.4 kb
Ensembl
annotation
MVK-001
MVK-002
MVK-007
ccRCC SV
MSK SV
MSK-β SV
StringTie2
MSTRG.9279.18
MSTRG.9279.19
PCS Coverage
Sequencing reads
MSTRG.9279.18
(MSK SV)
MSTRG.9279.19
(MSK-β SV)


A
5,450 kb
Chr 9
5,471 kb
5'
3'
PD-L1 (CD274)
Membrane isoform
Soluble isoform
Ensembl
StringTie2
5,462,800
5,463,400
Exon 4
Alt. Exon 4
3'UTR
Alt. 3'UTR
Novel soluble isoform
PCS
RCC4
B
PD-L1
0.1806
Normalised counts
Nonrecurrent
Recurrent
C
Membrane PD-L1
Soluble PD-L1 (NM_001314029)
Soluble PD-L1 Novel
Log2(RPM + 1)
0.1905
>0.9999
D
Relative membrane PD-L1 expression (mPD-L1 mRNA/GAPDH&ACTB mRNA)
Relative soluble PD-L1 expression (soluble PD-L1 mRNA/GAPDH & ACTB mRNA)
Relative novel soluble PD-L1 expression (novel soluble PD-L1 mRNA/GAPDH & ACTB mRNA)
0.0938
0.3045
0.4450
E
RPM
Unstimulated
IFNG + TNF
F
Membrane PD-L1
Soluble PD-L1
Novel soluble PD-L1
Half-life
1.603h
1.547h
8h+
1.487h
0.927h
0.660h
Relative expression (0h)
Actinomycin D treatment (h)
n.s.
Unstimulated
IFNG + TNF



	Downregulated	Upregulated	Downregulated	Upregulated
Known	100	111	46	24
Novel	8	32	1	3



PCR-cDNA-seq												
Tumour samples	135	171	243	254	260	273	314	318	320	329	382	395
Passed reads (Q > 7, 10^6)	72.3	43.9	59.5	71.7	60.7	53.7	51.0	27.6	72.8	53.6	51.3	63.2
Median alignment length (nt)	461	554	519	447	515	552	616	539	446	552	490	510
Median accuracy (%)	95.3	95.6	95.2	95.9	95.9	96.0	95.5	95.0	95.2	95.5	92.0	95.4
Full-length transcripts (%)	22.7	37.1	25.2	21.8	25.1	34.2	36.4	30.3	21.0	31.7	25.4	26.4

Direct-RNA-seq												
Tumour samples	135	171	243	254	260	273	314	318	320	329	382	395
Passed reads (Q > 7, 10^6)	4.44	5.10	6.01	4.05	4.71	4.95	3.43	5.44	2.41	5.06	3.62	3.39
Median alignment length (nt)	426	483	507	301	342	396	384	413	362	481	419	345
Median accuracy (%)	90.1	91.0	89.8	90.6	90.8	91.0	90.7	90.5	90.0	90.6	90.5	90.3
Full-length transcripts (%)	15.8	11.3	18.1	2.76	3.20	7.20	4.90	7.40	7.80	10.3	8.14	4.20