Supplemental Material – Lee et al

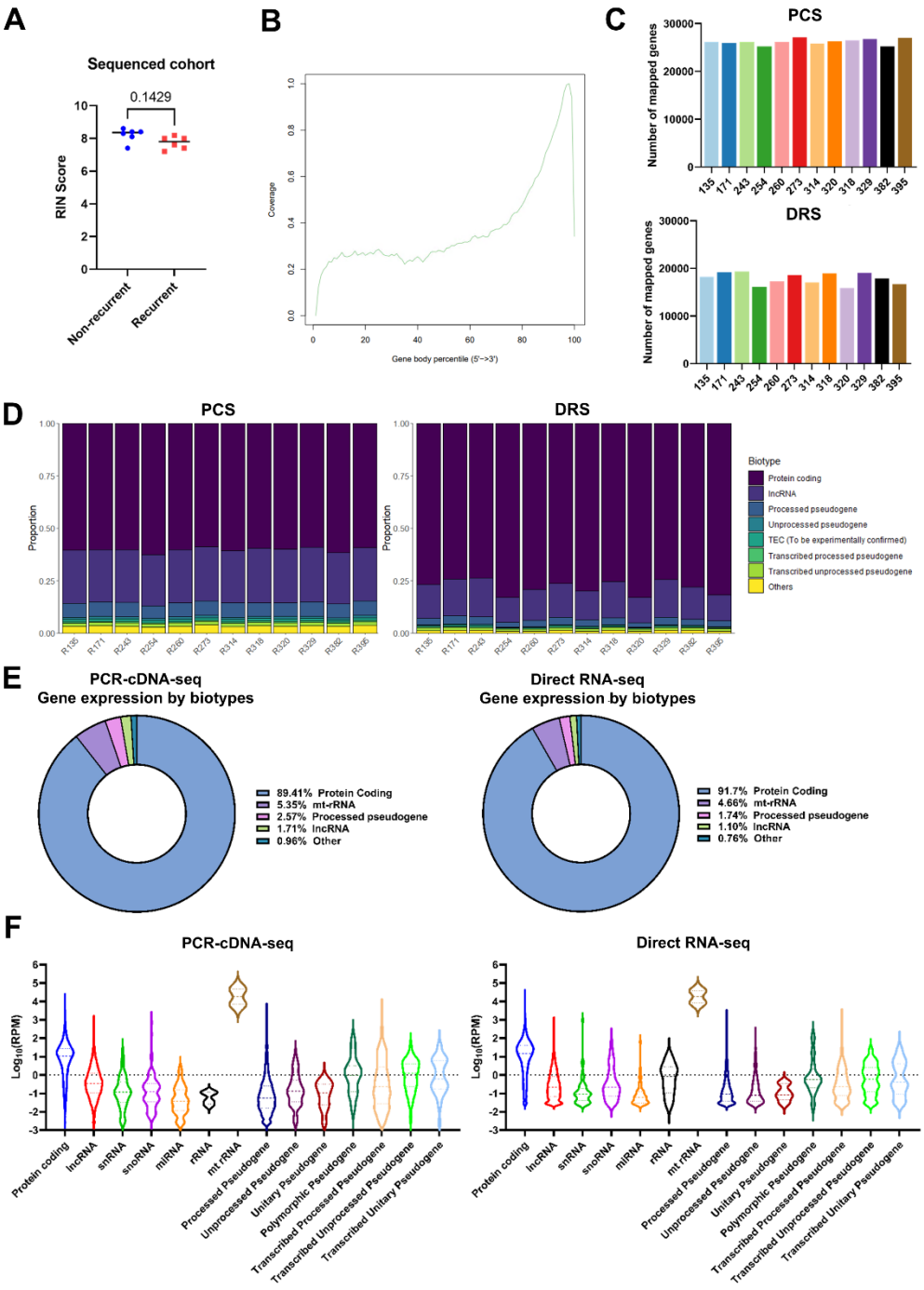
Long-read RNA sequencing of archival tissues reveals novel genes and transcripts associated with clear cell renal cell carcinoma recurrence and immune evasion

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.

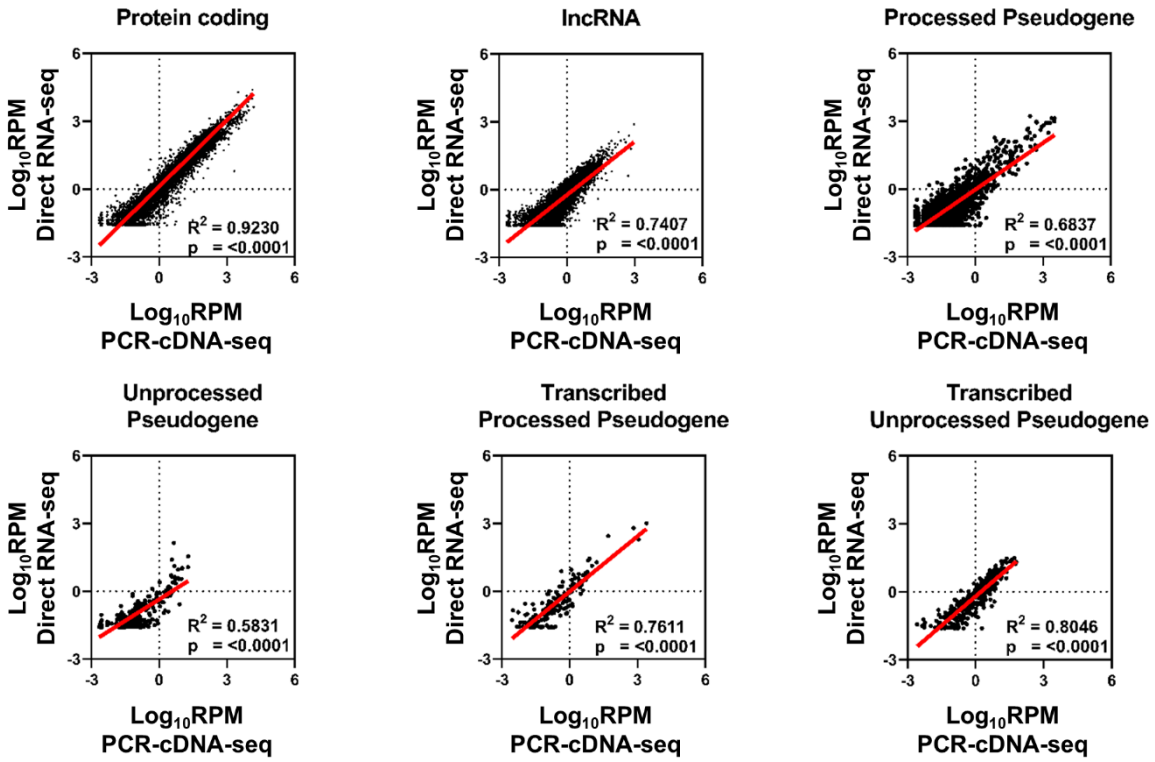
* Corresponding author: dimitris.lagos@york.ac.uk

Table of Contents



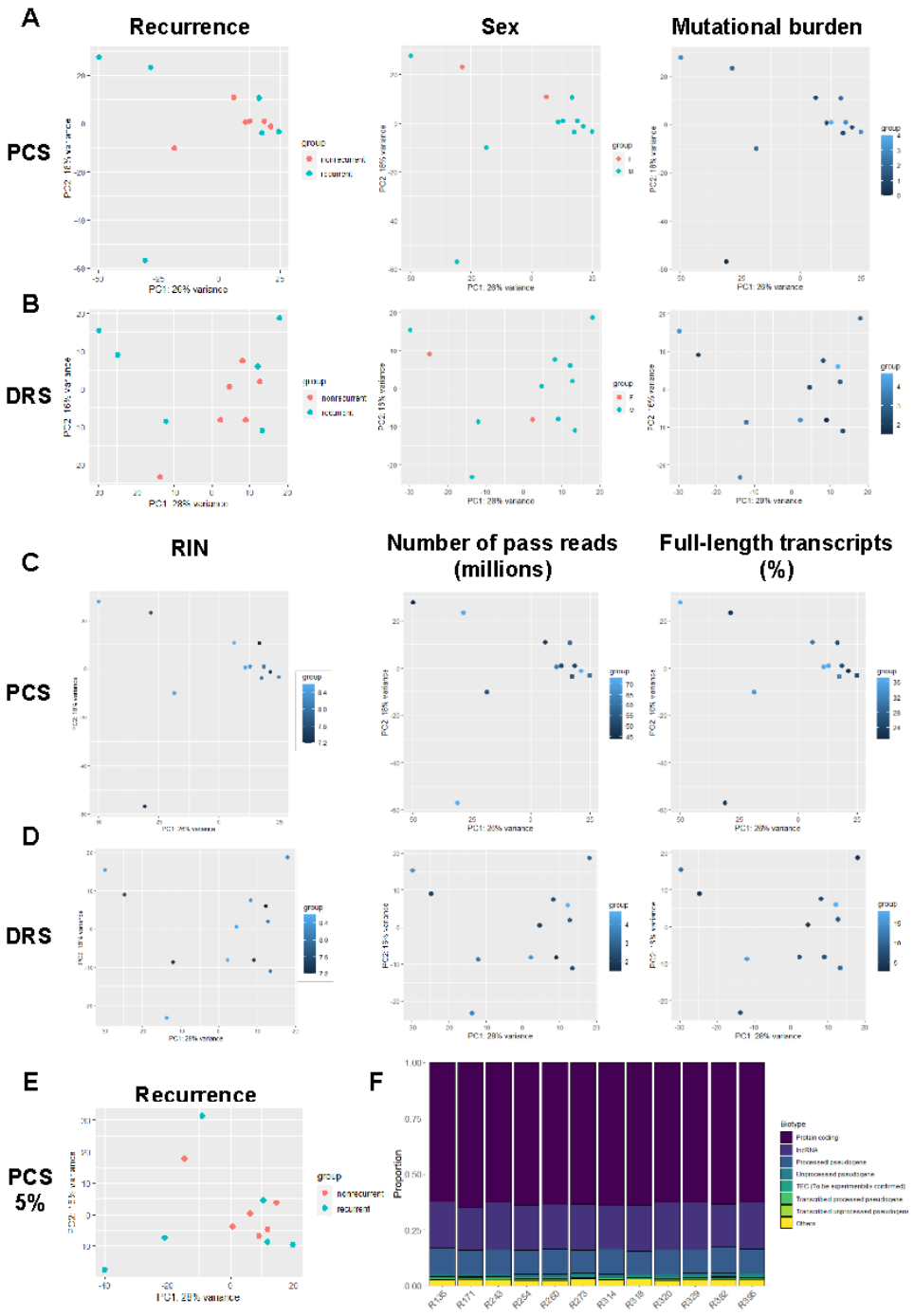
Supplemental Figure S1. DRS and PCS of ccRCC nephrectomy samples. (I) Grouped dot plot showing RIN score of RNA extracted from non-recurrent (blue) and recurrent (red) ccRCC tumours. Two-tailed Mann-Whitney U tests were used with $p \leq 0.05$ considered significant. p value of non-significant results is indicated in graph. (*B*) Metagene profile representing average 5' to 3' read coverage across gene bodies (Matched annotation from NCBI and EMBL-EBI (MANE), n = 19,316)), using 5% subsampled PCS 135 as a representative model.

(*C*) Bar charts showing the number of genes mapped by PCS and DRS from each ccRCC tumour sample. (*D*) Stacked bar charts depicting the proportions of gene biotypes of all mapped genes from reference genome (Ensembl release 105, GRCh38) by PCS and DRS of sequenced tumour samples. (*E*) Donut charts depicting the average proportions of RNA biotypes of mapped genes by expression levels from PCS and DRS data. (*F*) Violin plots depicting the distribution of gene expression levels (Log_{10}RPM) of mapped genes by biotypes from PCS and DRS, with first and third quartiles and median shown as horizontal line within each plot.



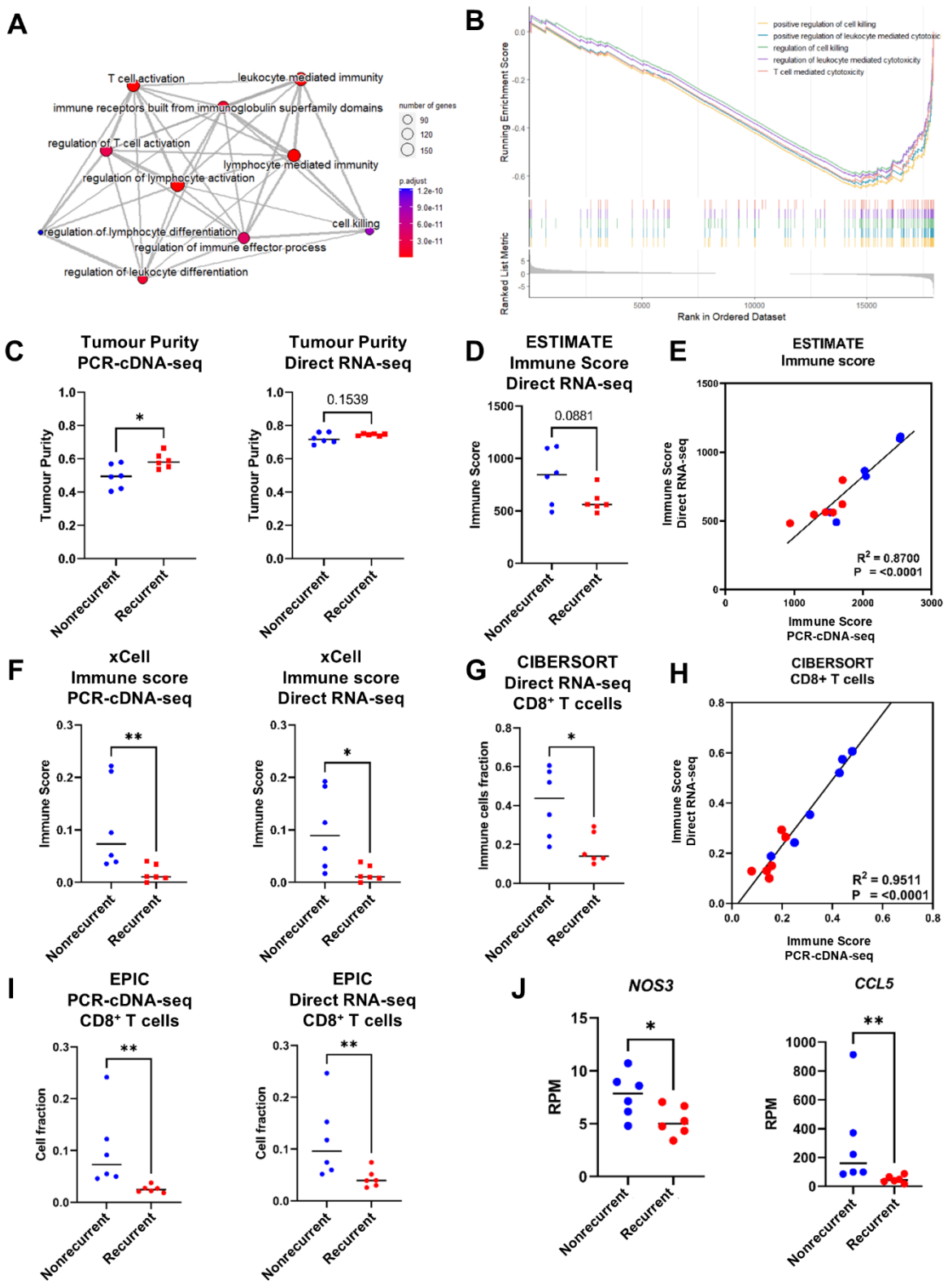
Supplemental Figure S2. Correlation between PCS and DRS data. Correlation between gene expression levels (Log_{10}RPM) of Protein coding, lncRNA, Processed Pseudogene, Unprocessed Pseudogene, Transcribed Processed Pseudogenes and Transcribed Unprocessed Pseudogenes mapped by PCS and DRS of ccRCC tumour samples. Throughout, diagonal lines represent the line of best fit. R^2 values were computed to measure

goodness-of-fit and p values were generated from F-test, with $p<0.05$ considered statistically significant.



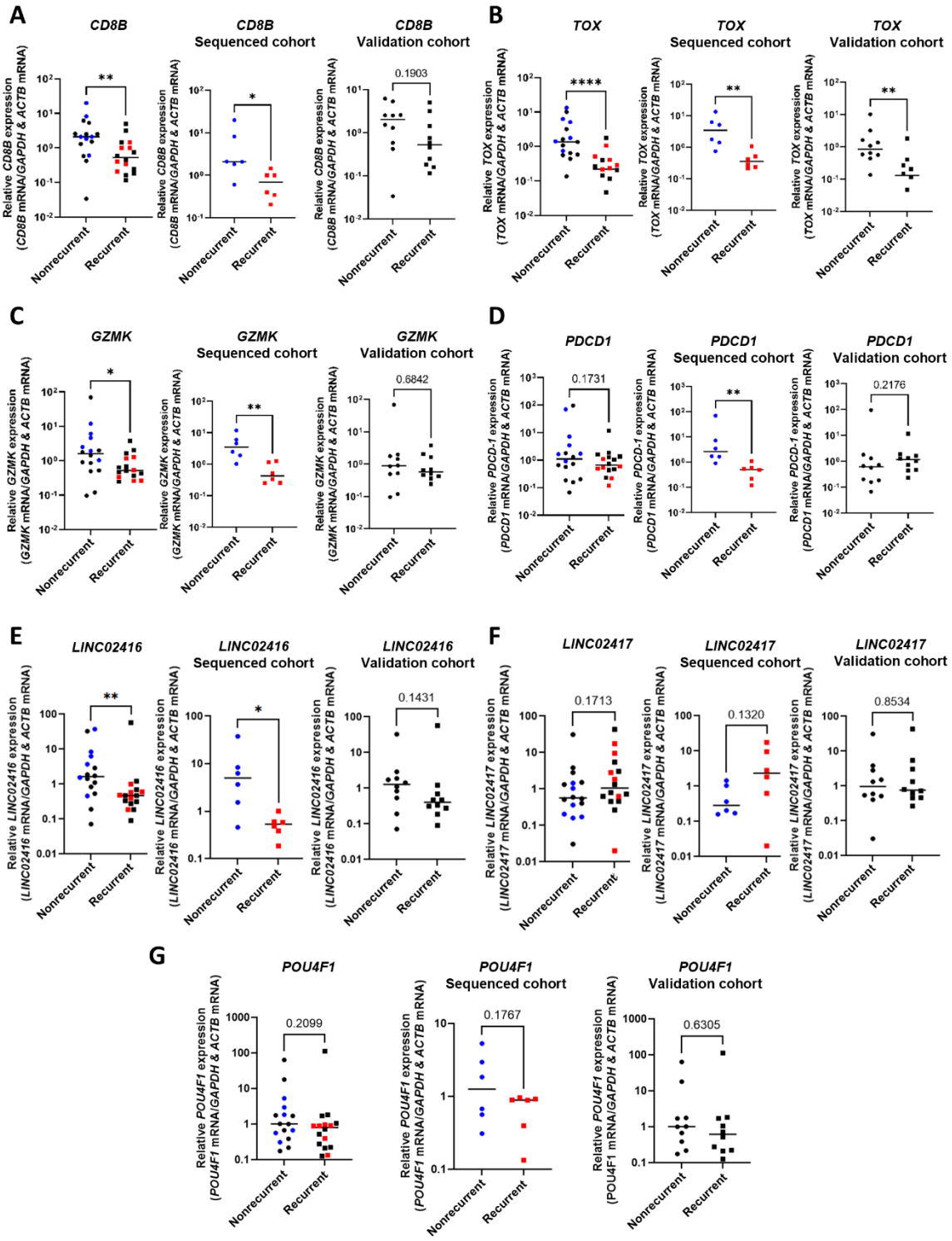
Supplemental Figure S3. Principal Component Analysis (PCA) plots on ccRCC tumours gene expression data. (*A*) DESeq2 generated PCA plots using PCS expression data showing

PCA of recurrent vs non-recurrent ccRCC groups, sex, and degree of mutational burden. (*B*) DESeq2 generated PCA plots using DRS expression data showing PCA of recurrent vs non-recurrent ccRCC groups, sex, and degree of mutational burden. (*C*) PCA plots using PCS expression data showing PCA of samples with respective RIN numbers, number of pass reads, and percentage of full-length transcripts. (*D*) PCA plots using DRS expression data showing PCA of samples with respective RIN numbers, number of pass reads, and percentage of full-length transcripts. (*E*) DESeq2 generated PCA plots using subsampled (5%) PCS expression data showing PCA of recurrent vs non-recurrent ccRCC groups. (*F*) Stacked bar charts depicting the proportions of gene biotypes of all mapped genes from reference genome (Ensembl release 105, GRCh38) by subsampled (5%) PCS sequenced tumour samples.



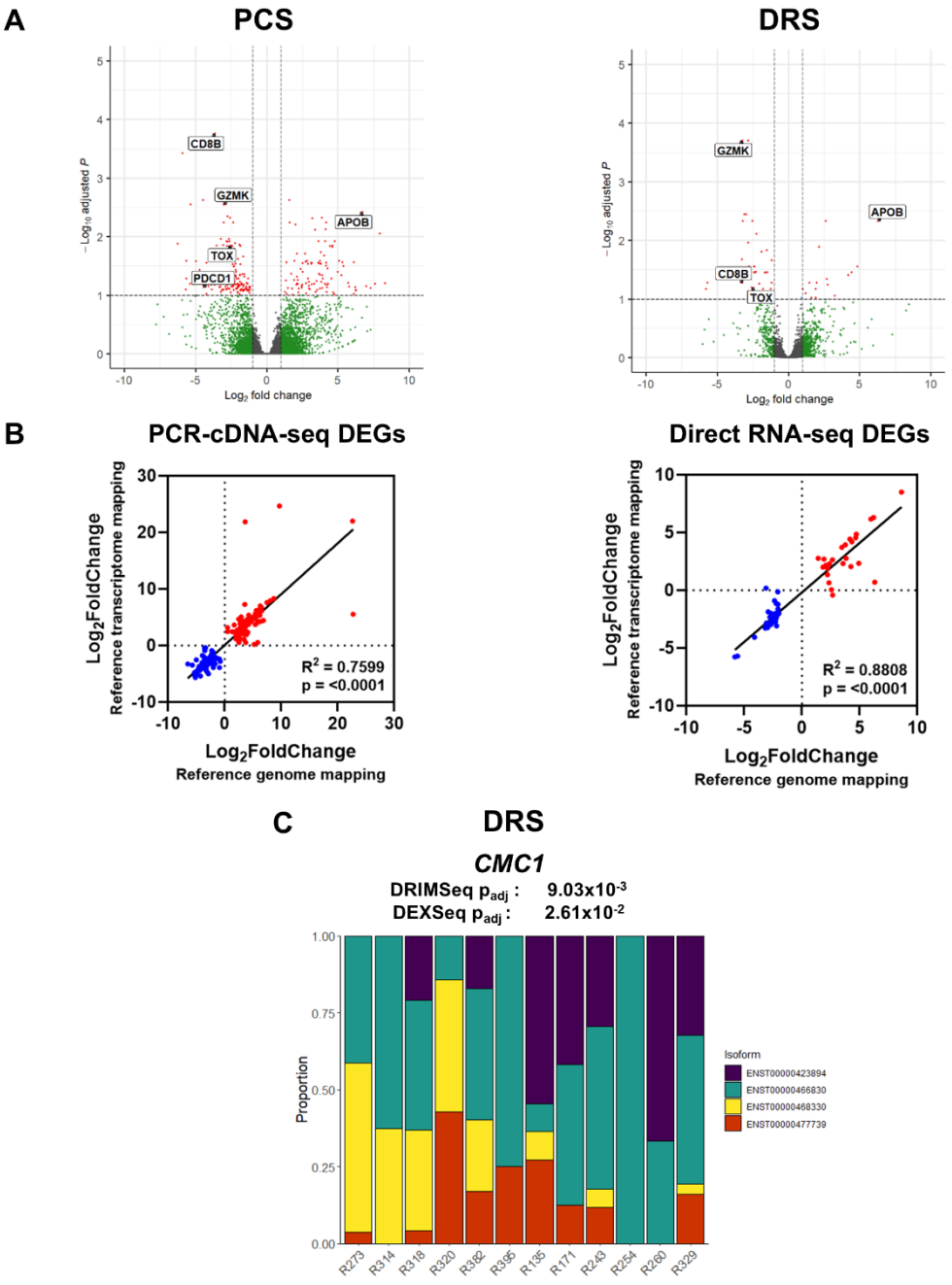
Supplemental Figure S4. Recurrence of ccRCC is associated with lower tumour immune infiltration. (*A*) Gene Ontology Biological Process GSEA map showing top 10 enriched terms associated with ccRCC recurrence-associated differential gene expression by PCS. (*B*) GSEA

enrichment plot for the top 5 enriched Gene Ontology Biological Process terms associated with ccRCC recurrence differential gene expression from DRS. The x-axis shows genes represented in each pathway and the y-axis shows enrichment scores. (*C*) Grouped dot plot showing estimated tumour purity of non-recurrent (blue) and recurrent (red) ccRCC tumours by the ESTIMATE algorithm using gene expression data from PCS and DRS. (*D*) Grouped dot plot showing estimated immune score of non-recurrent (blue) and recurrent (red) ccRCC tumour by the ESTIMATE algorithm, using DRS gene expression data. (*E*) Correlation between ESTIMATE immune scores of non-recurrent (blue) and recurrent (red) ccRCC tumours, generated by PCS and DRS gene expression data. (*F*) Grouped dot plot showing estimated tumour purity of non-recurrent (blue) and recurrent (red) ccRCC tumours by xCell using gene expression data from PCS and DRS. (*G*) Grouped dot plot showing relative population of $CD8^+$ T cells within immune infiltrates of non-recurrent (blue) and recurrent (red) ccRCC tumours estimated by CIBERSORTx using DRS gene expression data. (*H*) Correlation between CIBERSORTx estimated $CD8^+$ T cells fraction amongst immune infiltrates in non-recurrent (blue) and recurrent (red) ccRCC tumours, generated by PCS and DRS gene expression data. (*I*) Grouped dot plot showing relative cell fraction of $CD8^+$ T cells in non-recurrent (blue) and recurrent (red) ccRCC tumours estimated by EPIC using PCS and DRS gene expression data. (*J*) Grouped dot plot showing *NOS3* and *CCL5* gene expression (Reads per million (RPM)) in non-recurrent (blue) and recurrent (red) ccRCC PCS gene expression data. For (*E*) and (*H*), R^2 values were computed to measure goodness-of-fit and p values were generated from F-test, with $p \leq 0.05$ considered statistically significant. For (C), (D), (F), (G), (I) and (J), two-tailed Mann-Whitney U tests were used with $p \leq 0.05$ considered significant. * = $p < 0.05$, ** = $p < 0.01$, **** = $p < 0.0001$. p value of non-significant results is indicated in graph. Centre line represents median for each group.



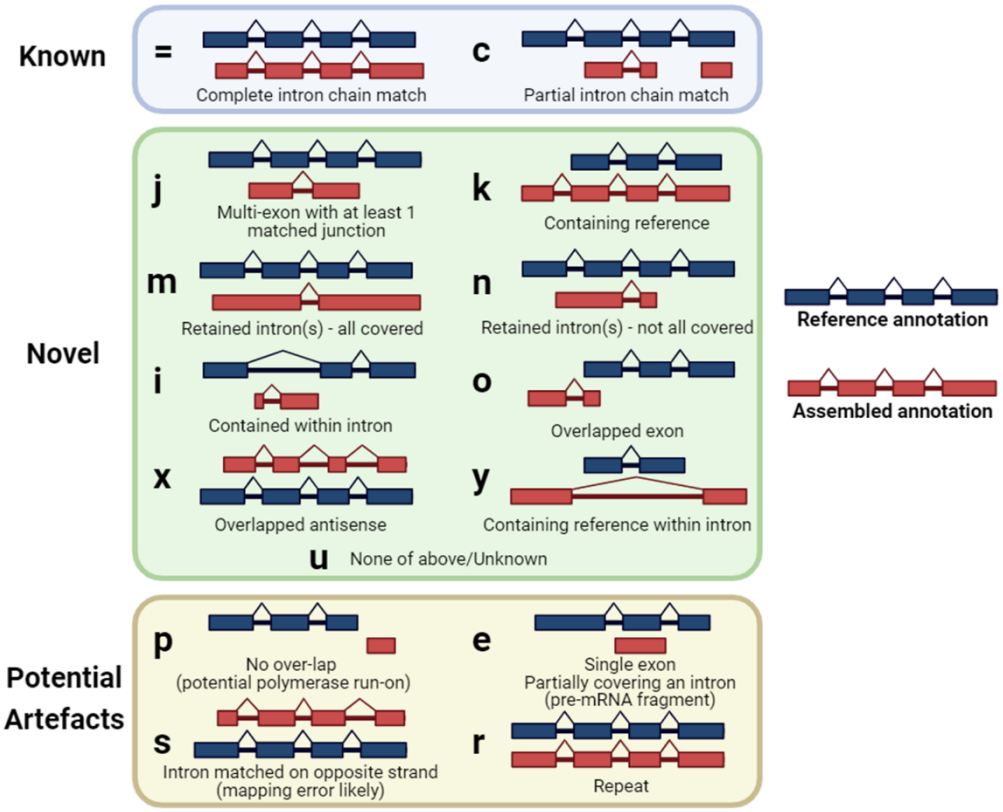
Supplemental Figure S5. Validation of sequencing results via qRT-PCR. (*A*) *CD8B*, (*B*) *TOX*, (*C*) *GZMK* ,(*D*) *PD-1* , (*E*) *LINC02416* ,(*F*) *LINC02417* and (*G*) *POU4F1* mRNA levels measured by qRT-PCR in recurrent and non-recurrent tumours from sequenced cohort (blue and red, middle, n = 12) and validation cohort (black, right, n = 20) relative to average mRNA levels in non-recurrent tumours. mRNA levels were normalised to *GAPDH* and *ACTB.* Plots

showing data from both sequenced cohort and validation cohort (left) replicate content of Fig. 2 from the main body of the paper to provide a clearer visual representation of data. two-tailed Mann-Whitney U tests were used with $p \leq 0.05$ considered significant. * = $p < 0.05$, ** = $p < 0.01$, **** = $p < 0.0001$. p value of non-significant results is indicated in graph. Centre line represents median for each group.

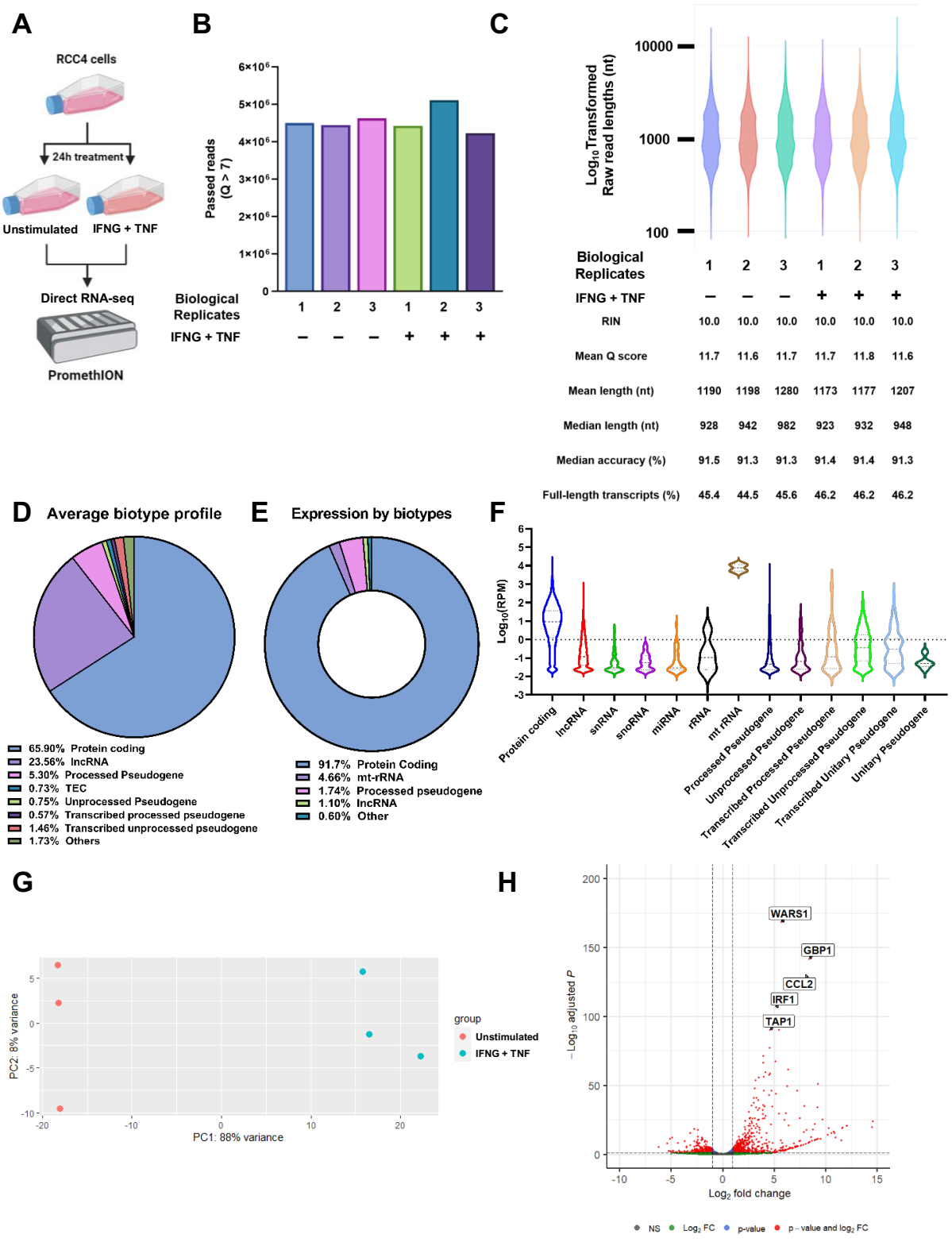


Supplemental Figure S6. Transcript-level mapping of PCS and DRS of ccRCC nephrectomy samples. (A) Volcano plots showing differentially expressed genes (red) between recurrent

and non-recurrent ccRCC tumours from PCS and DRS data using transcript-level mapping method with Ensembl transcriptome reference (cDNA and ncRNA, Ensembl release 105). Dotted lines indicate significance threshold ($|Log_2FoldChange| \geq 2$, padj ≤ 0.1). Names of genes that were validated by qPCR with validation cohort are shown. (B) Correlation between $Log_2FoldChange$ of DEGs identified by either or both transcript-level mapping method with Ensembl transcriptome reference (cDNA and ncRNA, Ensembl release 105) or gene-level mapping with reference genome reference (Ensembl release 105), between recurrent and non-recurrent ccRCC tumours in PCS and DRS. Blue and red dots represent significantly down- and up-regulated genes by either or both PCS and DRS. Diagonal line represent the line of best fit. R2 value was computed to measure goodness-of-fit and p value was generated from F-test, with $p \leq 0.05$ considered statistically significant. (C) Stacked bar graphs representing proportions of CMC1 isoforms in ccRCC tumours using DRS. DRIMseq and DEXSeq padj values are indicated in graph.

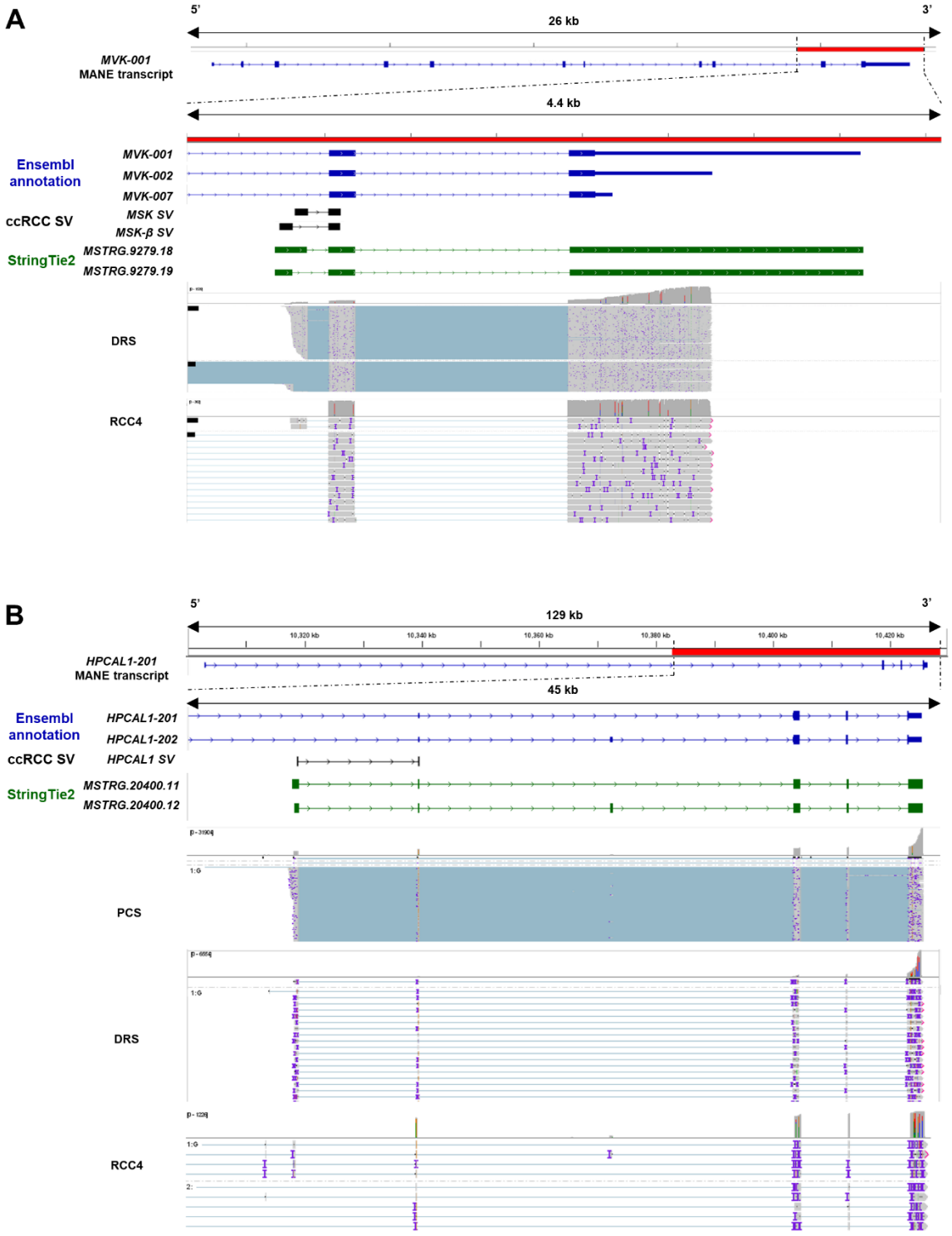


Supplemental Figure S7. Graphical representation of GffCompare transcript class codes. Transcript class codes were categorised into 'Known' ('=' : Complete intron chain match, 'c': Partial intron chain match), 'Novel' ('j' : Multi-exon with at least 1 matched junction', 'k' : Containing reference, 'm' : 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).



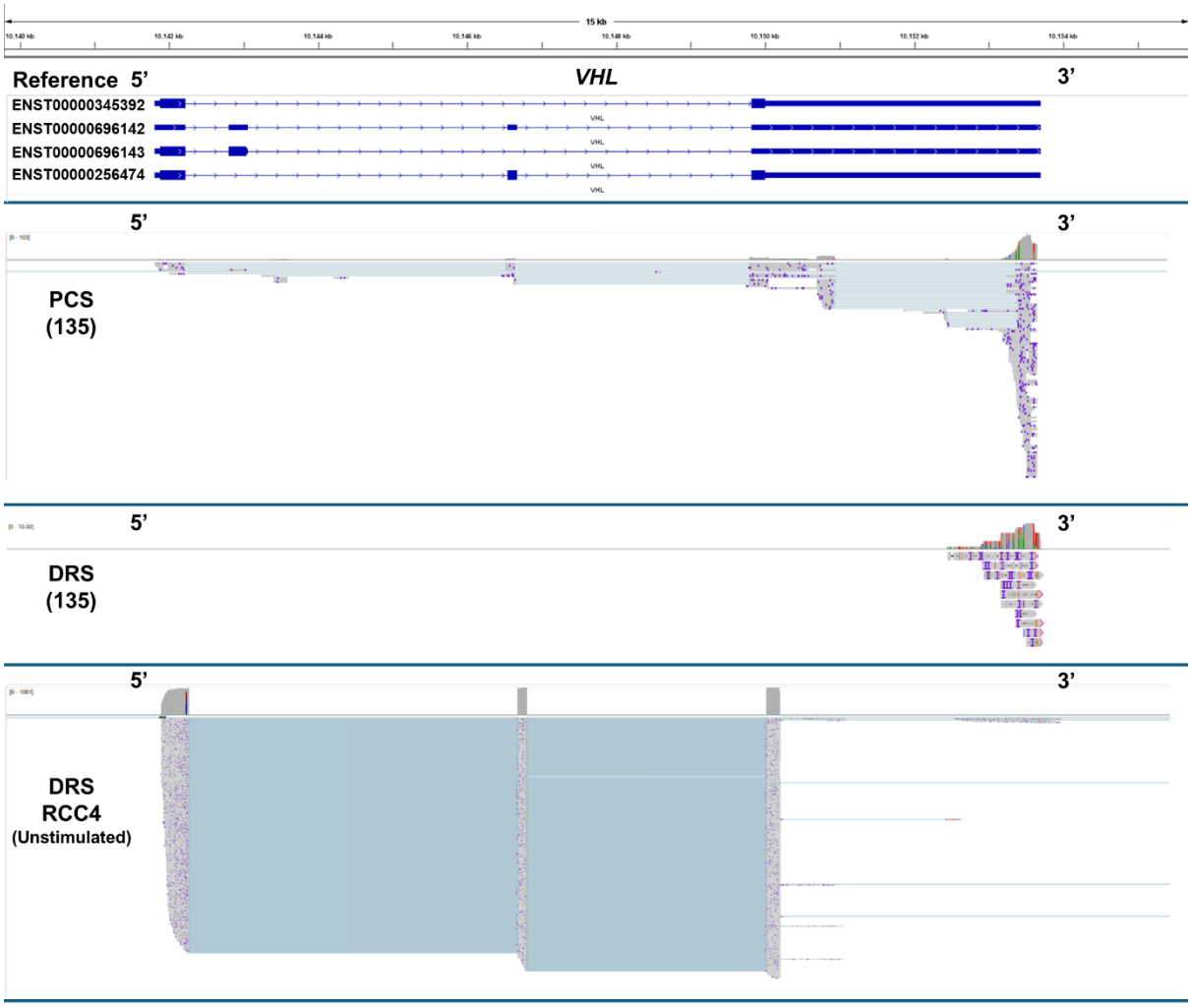
Supplemental Figure S8. Sequencing statistics of IFNG + TNF treated and untreated RCC4 cells. (*A*) Summary workflow for DRS of RCC4 cells. Figure made with Biorender. (*B*)

Bar chart showing the number of sequencing reads generated by DRS of RCC4 cells that passed the quality filter (Q score >7). (*C*) Violin plot showing $\log_{10}$ transformed raw read lengths of passed reads generated by DRS of RCC4 cells. RIN score, mean read Q score, mean and median read length for each sequencing dataset are listed in the table below violin graph. (*D*) Pie chart depicting the average proportions of RNA biotypes of genes mapped by DRS of RCC4 cells. (*E*) Pie chart showing the average proportions of RNA biotypes of mapped genes by expression levels from DRS of RCC4 cells. (*F*) Violin plot depicting the distribution of gene expression levels (Log_{10}RPM) of mapped genes by biotypes from DRS of RCC4 cells, with first, third quartiles and median shown as horizontal line within each plot. (*G*)) DESeq2 generated PCA plots using expression data from DRS of RCC4 cells, showing PCA of unstimulated and IFNG+TNF treated groups. (H) Volcano plot showing differentially expression genes (red) between unstimulated and IFNG+TNF treated RCC4 cells from DRS data. Dotted lines indicate significance threshold ($|Log_2FoldChange| \geq 2$, padj ≤ 0.1). Names of top 5 most significantly differentially expressed genes (by padj) are shown.

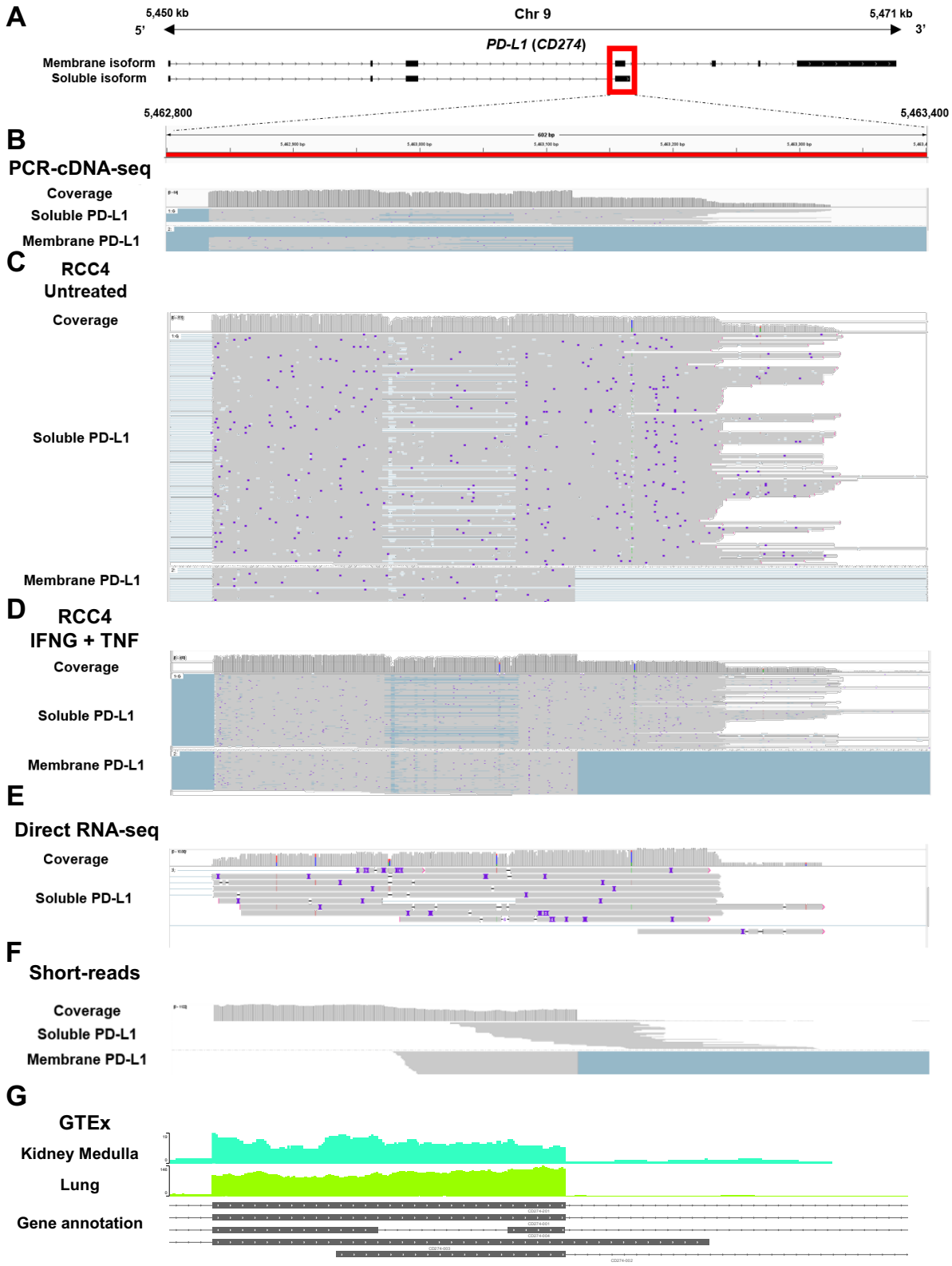


Supplemental Figure S9. Visualisation of long-read RNAseq reads mapping to ccRCC-specific splice junctions. (*A*) IGV visualisation of *MVK* reference annotations (blue), ccRCC specific *MVK* splice junctions (black), StringTie2 assembled novel transcripts (green), DRS of ccRCC tumour samples (labelled as DRS) and RCC4 coverage tracks (grey) and sequencing

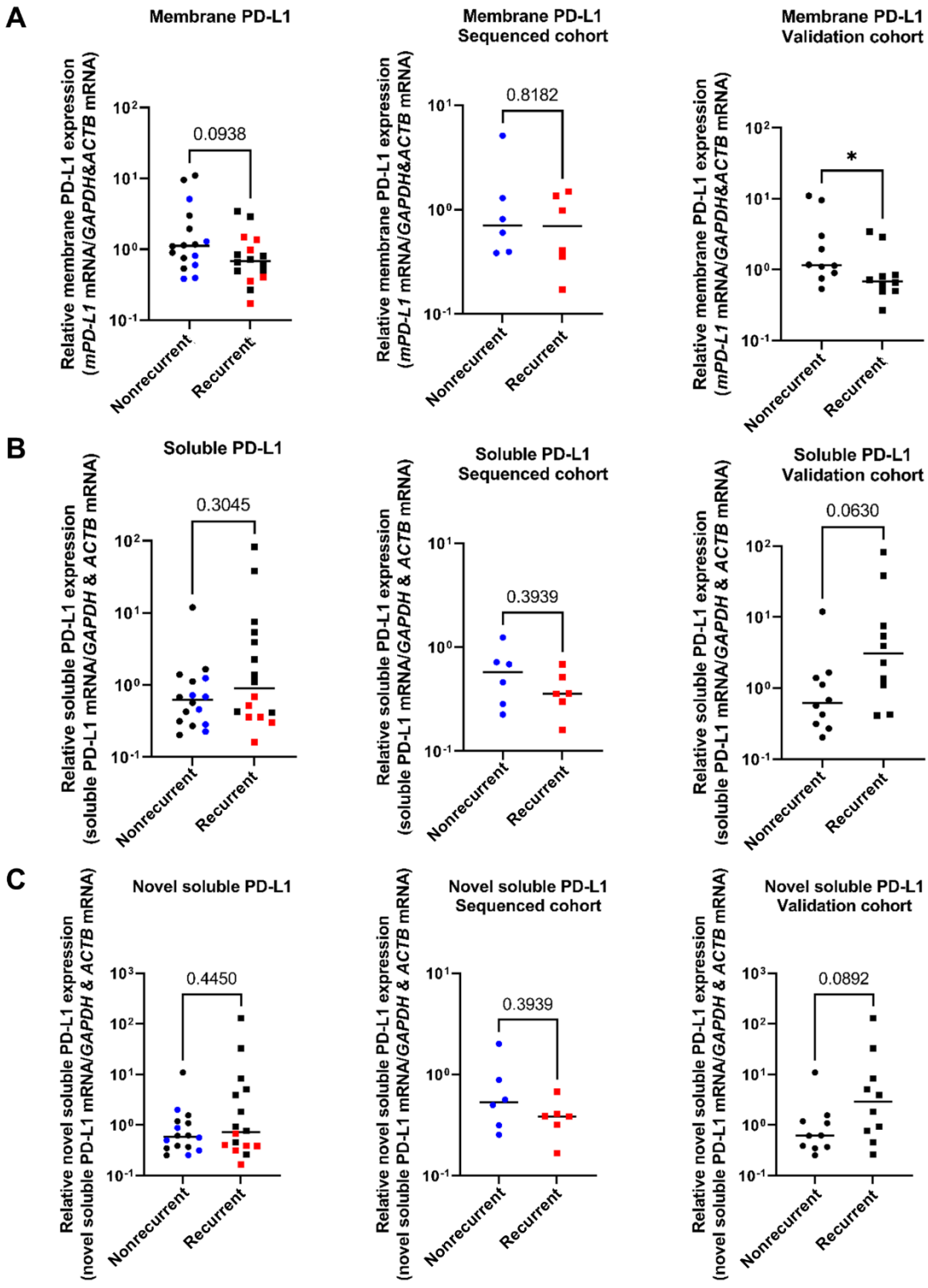
reads aligned to the reference genome in the region of interest (red bar, hg38 Chr12:109,594,200 – 109,598,600). (*B*) IGV visualisation of *HPCAL1* reference annotations (blue), ccRCC specific *MVK* splice junctions (black), StringTie2 assembled novel transcripts (green), PCS and DRS of ccRCC tumour samples (labelled as DRS) and RCC4 coverage tracks (grey) and sequencing reads aligned to the reference genome in the region of interest (red bar, hg38 Chr2:10,300,100-10,429,600).



Supplemental Figure S10. Visualisation of long-read RNAseq reads mapping to *VHL*. IGV visualisation of *VHL* reference annotations (blue), coverage tracks and sequencing reads of PCS and DRS of archival tumour sample 135, and DRS of RCC4 (Unstimulated, replicate 1), aligned to the reference genome in the region of interest (hg38 Chr3:10,140,100-10,155,000).

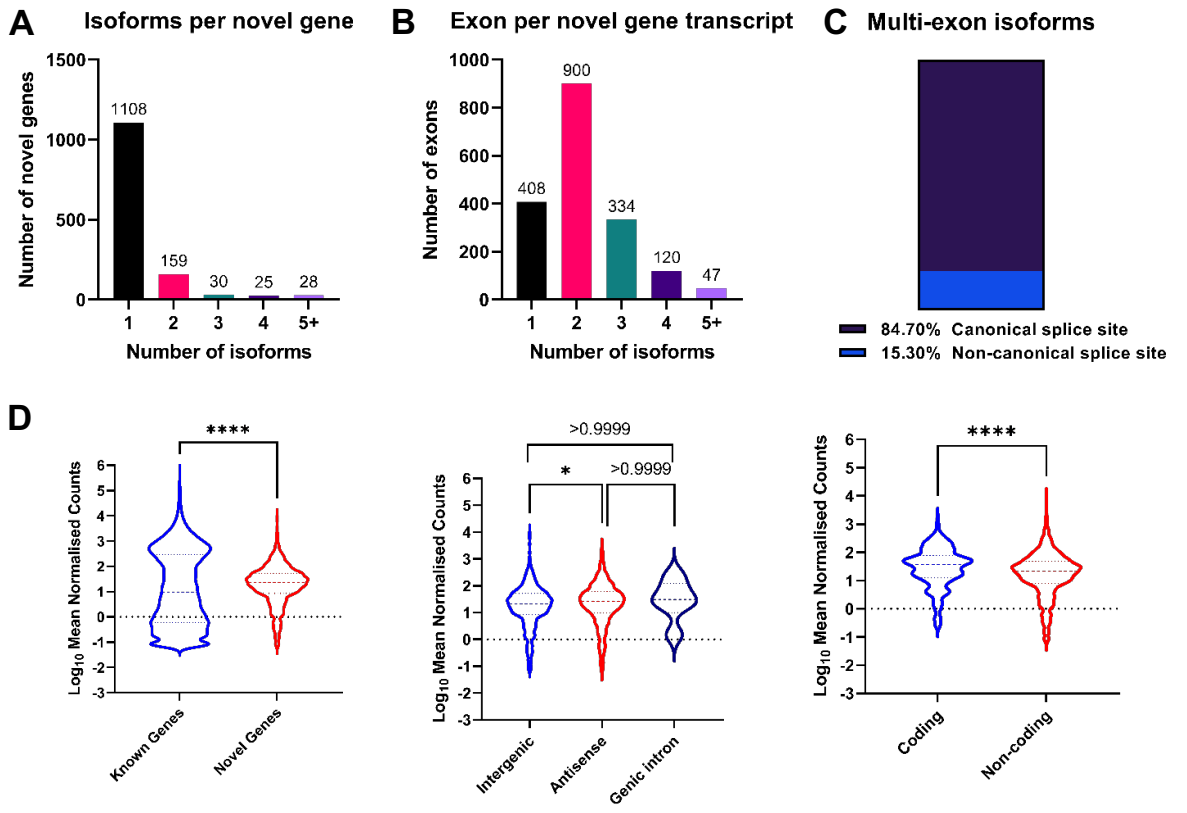

A
5,450 kb
Chr 9
5,471 kb
5'
3'
PD-L1 (CD274)
Membrane isoform
Soluble isoform
5,462,800
5,463,400
B
PCR-cDNA-seq
Coverage
Soluble PD-L1
Membrane PD-L1
C
RCC4
Untreated
Coverage
Soluble PD-L1
Membrane PD-L1
D
RCC4
IFNG + TNF
Coverage
Soluble PD-L1
Membrane PD-L1
E
Direct RNA-seq
Coverage
Soluble PD-L1
F
Short-reads
Coverage
Soluble PD-L1
Membrane PD-L1
G
GTEx
Kidney Medulla
Lung
Gene annotation

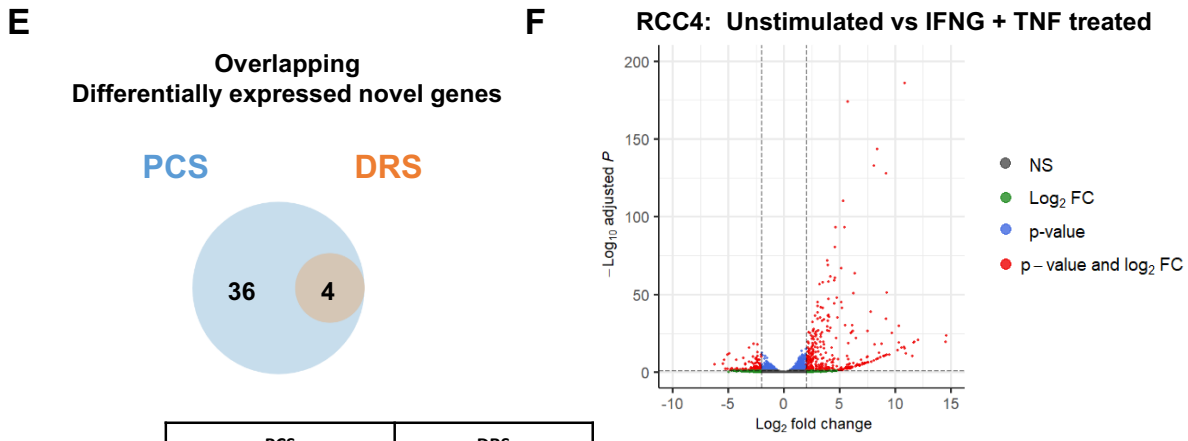
Supplemental Figure S11. Sequencing reads aligned to *PD-L1* exon 4. (*A*) IGV visualisation of reference annotation of *mPD-L1* isoform (ENST00000381577) and *sPD-L1* (black, NM_001314029). (*B*) ccRCC tumours PCR-cDNA sequencing coverage plot and reads aligned to *PD-L1* exon 4 region (Chr9:5,462,800 – 5,463,400). (*C*) Untreated RCC4 DRS coverage plot and reads aligned to *PD-L1* exon 4 region (Chr9:5,462,800 – 5,463,400). (*D*) IFNG + TNF treated RCC4 DRS coverage plot and reads aligned to *PD-L1* exon 4 region (Chr9:5,462,800 – 5,463,400). (*E*) ccRCC tumours DRS sequencing coverage plot and reads aligned to *PD-L1* exon 4 region (chr9:5,462,800 – 5,463,400). (*F*) IFNG + TNF treated RCC4 Illumina short reads coverage plot and reads aligned to *PD-L1* exon 4 region (Chr9:5,462,800 – 5,463,400). (*G*) GTEx Kidney Medulla (Cyan) and Lung (Green) short reads coverage plot at the *PD-L1* exon 4 region (Chr9:5,462,800 – 5,463,400), and GENCODE gene annotation (black) tracks showing annotated *PD-L1* transcripts' structure.



Supplemental Figure S12. Validation of *PD-L1* sequencing results and novel *sPD-L1* isoform via qRT-PCR. (*A*) *mPD-L1*, (B) *sPD-L1* and (C) novel *sPD-L1* mRNA levels measured by qRT-PCR in recurrent and non-recurrent tumours from sequenced cohort (blue and red, middle, n = 12) and validation cohort (black, right, n = 20) relative to average mRNA levels in non-

recurrent tumours. mRNA levels were normalised to *GAPDH* and *ACTB*. Plots showing data from both sequenced cohort and validation cohort (left) replicate content of Fig. 5 from the main body of the paper to provide a clearer visual representation of data. Two-tailed Mann-Whitney U tests were used with $p \leq 0.05$ considered significant. * = $p < 0.05$, p value of non-significant results is indicated in graph. Centre line represents median for each group.



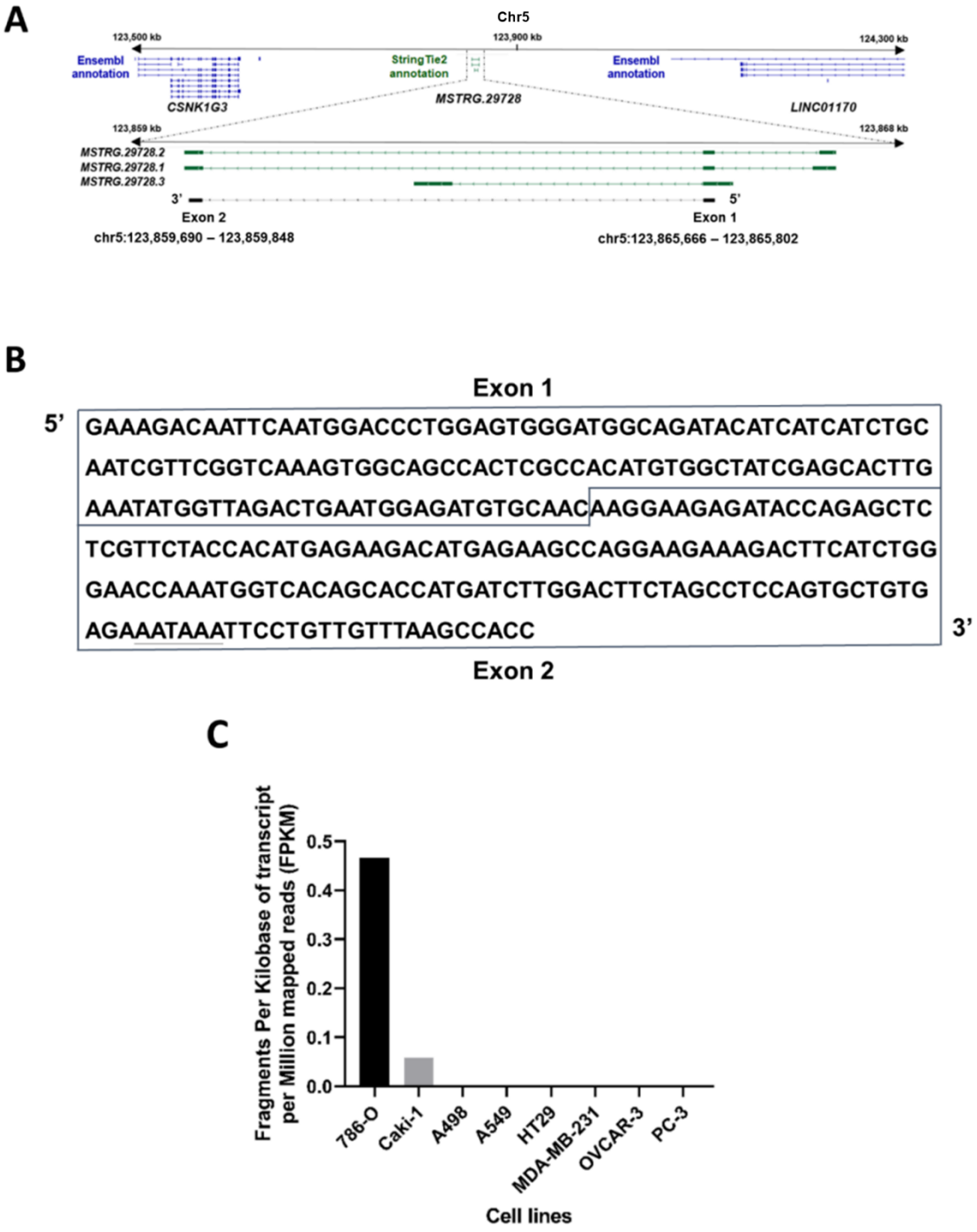


	PCS		DRS	
Novel Genes	Log_2FC	p_{adj}	Log_2FC	p_{adj}
MSTRG.1938	-2.550	0.055	-2.989	0.031
MSTRG.21132	5.227	$1.88x10^{-7}$	5.994	0.035
MSTRG.29728	2.951	0.051	3.441	0.067
MSTRG.38727	4.821	0.099	5.099	0.092

	Downregulated	Upregulated
Known	111	324
Novel	4	9

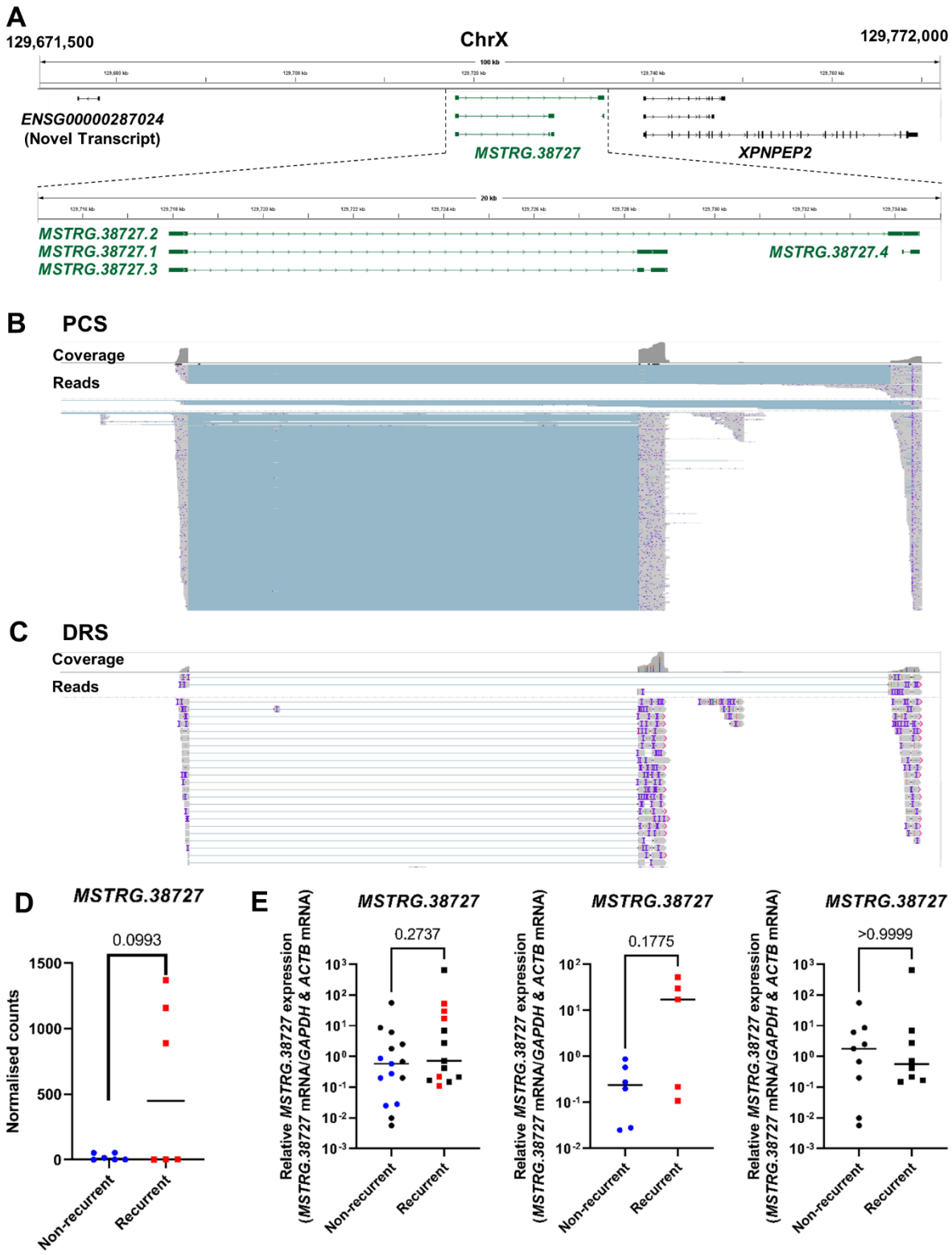
Supplemental Figure S13. Characterisation of novel genes. (*A*) Bar chart showing the distribution of the number of isoforms per PCS mapped novel gene (n = 1,350). (*B*) Bar chart showing the distribution of the number of exons per isoform from mapped novel genes (n =

1,350). (*C*) Bar chart showing the proportion of multi-exon novel gene isoform (n = 758) that consists of canonical splice sites (purple) or non-canonical splice sites (blue). (*D*) Violin plots depicting the distribution of gene expression levels (Log_{10}RPM) of novel genes versus known genes (Ensembl annotation 105); Novel genes that are classified by SQANTI3 as intergenic versus antisense and genic intron; Novel genes that are classified as coding versus non-coding. Two-tailed Mann-Whitney U tests (left, right) and Kruskal-Wallis test (centre) were used with $p \leq 0.05$ considered significant. **** = $p < 0.0001$, * = $p < 0.05$. p value of non-significant results is indicated in graph. (*E*) Venn diagram showing the number of overlapping differentially expressed novel genes between PCS and DRS of ccRCC tumour samples. Log_2FoldChanges and padj values of the overlapping differentially expressed novel genes by DESeq2 analysis are shown in table below the diagram. (*F*) Volcano plots showing differentially expressed genes (red) between unstimulated and IFNG+TNF treated RCC4 cells using StringTie2 assembled reference. Number of differentially expressed novel and known genes are shown in table below plots.



Supplemental Figure S14. *MSTRG.29728* expression in cancer cell lines of different tissue origins. (*A*) IGV visualisation of *MSTRG.29728* isoforms StringTie2 reference annotation (green), 2-exon *MSTRG.29728* isoform (black) and the closest neighbouring genes (*LINC01170* and *CSNK1G3*) in the Ensembl reference annotation (Ensembl release 105) at Chr5:123,500,000-124,300,000. Genomic locations (hg38) for the 2-exon *MSTRG.29728* isoform are indicated below isoform structure. (*B*) FASTA RNA sequence for the 2-exon

MSTRG.29728 isoform, with putative poly(A) signal underlined. (*C*) Bar chart showing the estimated expression levels of *MSTRG.29728* 2-exon isoform in ccRCC (786-O, Caki-1, A498), Lung squamous cell carcinoma (A549), Colorectal adenocarcinoma (HT29), Triple-negative breast cancer (MDA-MB-231), Ovarian cancer (OVCAR-3) and Prostate cancer (PC-3) cell lines by LocExpress.



Supplemental Figure S15. Visualisation and expression validation of novel gene *MSTRG.38727*. (*A*) IGV visualisation of *MSTRG.38727* (green) and the closest neighbouring genes (*ENSG00000287024 (novel transcript)* and *XPNPEP2*) in the Ensembl reference

annotation (Ensembl release 105) at chrX:129,671,500 – 129,772,000 (Top tracks); *MSTRG.38727* isoforms structures (*MSTRG.38727.1*, *MSTRG.38727.2*, *MSTRG.38727.3*, *MSTRG.38727.4*) at ChrX: 129,715,000 – 129,735,000 (Green, bottom tracks). (*B*) IGV coverage track and reference genome aligned reads from PCS of ccRCC tumour samples in the region of interest (ChrX: 129,715,000 – 129,735,000). (*C*) IGV coverage track and reference genome aligned reads from PCS of ccRCC tumour samples in the region of interest (ChrX: 129,715,000 – 129,735,000). (*D*) Grouped dot plot showing reference DESeq2 normalised *MSTRG.38727* expression in non-recurrent (blue) and recurrent (red) tumours' PCS data. DESeq2 p_{adj} value is shown in graph. Centre line represents median for each group. (*E*) *MSTRG.38727* mRNA levels measured by qRT-PCR in recurrent and non-recurrent tumours from sequenced cohort (blue and red, middle, n = 12), validation cohort (black, right, n = 20) and both cohorts relative to average mRNA levels in non-recurrent tumours. mRNA levels were normalised to *GAPDH* and *ACTB*. Two-tailed Mann-Whitney U tests were used with $p \leq 0.05$ considered significant. p value of non-significant results is indicated in graph. Centre line represents median for each group.