Exploring cancer transcriptomes with long-read Nanopore sequencing

Cheuk Yan Joshua Lee

PhD

University of York

Biology

December 2022

Abstract

Cancer progression is associated with aberrant co- and post-transcriptional processing of RNA transcripts in both tumour and immune cells. This results in complex transcriptional profiles characterised by alternative splicing, 3'UTR and poly(A) tail length alterations, and chemical modifications. Despite advances, accurate profiling of transcript isoform usage and modifications remains challenging due to the limited read length offered by traditional RNA sequencing technologies. Here, we use long-read Nanopore direct RNA sequencing (DRS) and cDNA sequencing (PCS) to explore the transcriptomic profiles of clear cell renal cell carcinoma (ccRCC). Archival nephrectomy tumour samples from patients who experienced disease recurrence and controls were successfully analysed, demonstrating feasibility of the approach. Differential gene expression and transcript usage analysis identified changes in abundance and isoform usage associated with ccRCC recurrence. Gene expression-based cell-type deconvolution showed loss of tumour infiltrating immune cells, specifically $CD8^+$ T cells in the recurrent tumours. Remarkably, using reference-guided transcriptome reconstruction methods, thousands of unannotated transcripts isoforms were identified, including isoforms of clinically important tumour immune checkpoints. DRS analysis of the ccRCC tumour cell line RCC4 revealed that some of these novel isoforms identified were indeed expressed in cancer cells and that exposure to inflammatory cytokines could lead to isoform switch. Notably, differential alterations in poly(A) tail length were also observed in novel and annotated transcripts in response to the exposure of cytokines. Finally, the role of mRNA N^6-Methyladenosine (m^6A) was explored in ccRCC tumour cell line via genetic perturbation of the m^6A writer complex (METTL3 and WTAP) with CRISPR-Cas9 gene editing and siRNA-mediated transient depletion. Overall, our study demonstrates the feasibility of Nanopore long-read sequencing in tumour samples that uncovers cancer transcriptomes at single-transcript resolution and reveals the existence of multiple disease-associated alterations concurrently occurring in the mRNA of clinically relevant targets.

Table of Content

Table of Figures

Table of Tables

Acknowledgement

Working towards this work at York for the last four years has been an absolute privilege. The PhD journey was undoubtedly challenging at times but also deeply fulfilling. Firstly, I would like to thank my supervisor, Dimitris Lagos, for giving me the opportunity to work on this project. The guidance and feedback that you have given me have helped me grow as a scientist, and I will always be grateful for your patience and support throughout this PhD. Many thanks to my co-supervisor, James Hewitson, for his encouragement and guidance. I want to thank my TAP chair Pegine Walrad, for her valuable comments and feedback. In addition, I want to extend my thanks to Naveen Vasudev and Jo Brown from St James's Hospital for making this study possible.

I would also like to thank industrial supervisor Dan Turner for giving me the opportunity to work at ONT during my PhD. A special thanks to Libby Snell for being incredibly helpful and enthusiastic about my research. This work would not have been possible without your support. I am also very grateful to Aino Jarvelin for her guidance with bioinformatics analysis. Everyone in the ONT Apps team has been so kind, and I cannot thank you enough for making my time at Oxford so enjoyable.

To the past and present Lagos lab members, Katie, Mia, Alina, Magnus, Shoumit, and Dan, thank you very much for being so helpful over the years. I would especially like to thank Katie for being kind and supportive since we started. I would also like to thank everyone on Q2 who has made this PhD a lot more fun - I will always cherish the wonderful times we had at the Rook and Gaskill with you all. Many thanks to all my friends at York, especially Katherine and Alex, for all those bad movie nights, drinks by the river and you guys were always there for me. I'm so glad that we got through this together. I'm also grateful to my family for being supportive and always believing in me.

Finally, last but by no means least, I cannot thank Raph enough. You have supported me in all the ways you can, and I am so grateful for your love and kindness. Thank you for believing in me and being my rock throughout this PhD.

Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

Chapter 1

Introduction

1.1 Co- & Post-transcriptional gene regulation

Gene expression is a complex and dynamic process where genetic information is converted into functional products. This flow of information is described as the central dogma of molecular biology, where genetic information stored in DNA is transcribed to messenger RNA (mRNA) molecules and then translated into proteins (Crick, 1970). From the accessibility of chromatin to post-translational modifications of proteins, every step of gene expression is tightly regulated by multitudes of processes to produce the optimal levels of desired gene products in the eukaryotic cell. At the mRNA level, eukaryotes control gene expression via several co- and post-transcriptional RNA regulatory events, including mRNA capping, splicing, generation of alternative 3' untranslated region (UTR) via alternative mRNA polyadenylation, control of mRNA polyadenylation (poly(A)) tail length, mRNA chemical modifications, and mRNA export. These regulatory events dictate the stability, subcellular localisation and translational efficiency of mRNA molecules, as well as the identity of the sequence identity of the final gene products (Corbett, 2018). This chapter will cover the roles of co- & post-transcriptional mRNA regulatory events on gene expression and their relevance in cancer.

1.1.1 mRNA splicing

In eukaryotes, most genes are interweaved by protein-coding (exons) and non-coding regions (introns). Once a gene is transcribed from DNA into nascent precursor mRNA (pre-mRNA), intronic regions are removed, and exonic regions are ligated to form a single translatable mRNA molecule. The median number of exons and introns per human gene are 9 and 8 (with a median length of 131 and 1747 nt, respectively). Removal of introns and joining of consecutive exons is known as consecutive splicing. Alternatively, each mRNA molecule may include or exclude particular exons (Figure 1.1). This process is known as alternative splicing. Alternative splicing events are prevalent in human, where more than 95% of multi-exon genes undergo AS (Piovesan *et al.*, 2019). This

ability to generate multiple gene products from a single gene dramatically expands the diversity of human transcriptome and proteome. As of September 2022, the Ensembl human reference genome assembly and gene annotation contain 44938 genes (19804 protein-coding and 25134 non-protein coding genes) and 251296 transcripts (Cunningham *et al.*, 2022).

Splicing of nascent pre-mRNA is carried out by the spliceosome (Figure 1.2). The spliceosome is a macromolecular complex consisting of 5 small nuclear RNAs (snRNAs) and over 300 proteins currently associated with the complex. At the core of the complex, snRNAs interact with proteins to form small nuclear ribonucleoproteins (snRNPs) and recognises splice sites on pre-mRNAs (Chen and Moore, 2015). Firstly, the U1 snRNP complex binds to the 5' splice site/donor site of the pre-mRNA (GU dinucleotide at the 5' end of the intron sequence). U2 snRNP then binds to an intronic branch site, a short motif 18 – 40 nt upstream of the 3' end of the intron (AG dinucleotide 3' splice site/acceptor site). U4/U6.U5 tri-snRNP complex subsequently joins the two snRNP complexes. This is followed by the rearrangement of the protein complex, which activates the catalytic activities of the spliceosome. The 5' splice site is next cleaved, forming an intron lariat-spliceosome structure with the complex covalently bound to the branch site. The 3' splice site is then excised with the two exons ligated simultaneously. The cleaved intron-spliceosome complex finally disassembles and is released from the mRNA (Gehring and Roignant, 2021). During this process, a multiprotein complex, named the exon junction complex (EJC), is assembled at a conserved 24 nucleotides 5' upstream position from the exon-exon junction. EJC serves as a binding platform for other proteins and influences various mRNA processing events, including splicing, m^6A modification, nonsense-mediated decay and mRNA export. These complexes accompany from the nucleus, through the nuclear pore to the cytoplasm where they are displaced by the ribosomes (Schlautmann and Gehring, 2020).

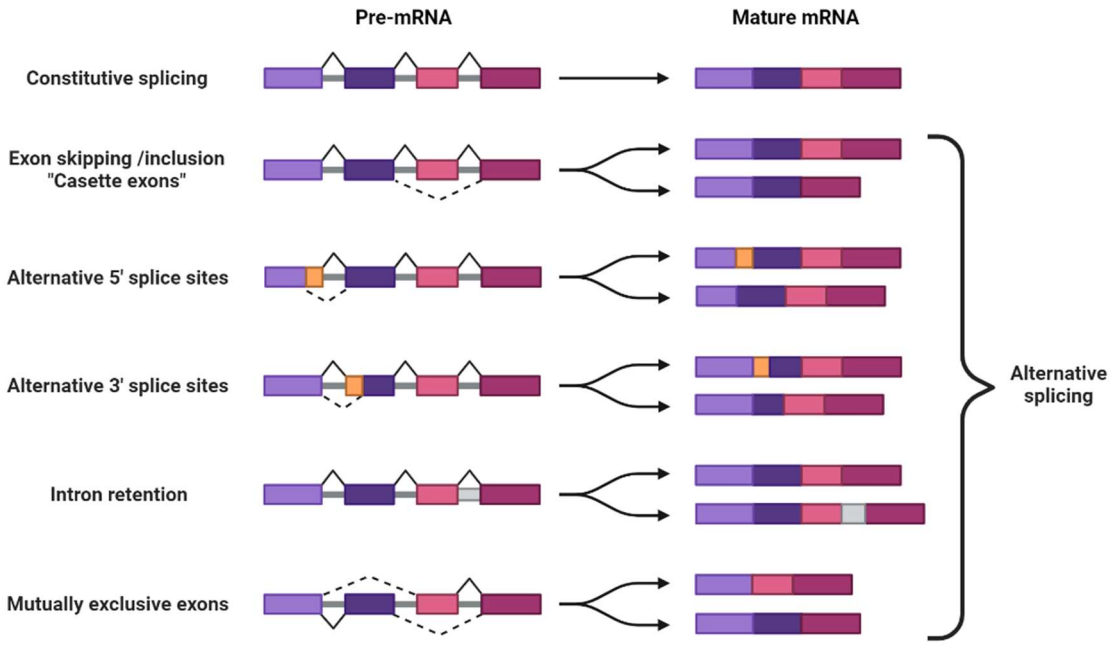


Figure 1.1: Graphical representation of constitutive & alternative splicing patterns

Transcribed pre-mRNA undergoes splicing to remove intronic sequences and ligate exonic sequences. When introns are removed and exons are joined consecutively in the order of which they have been coded, this is known as constitutive splicing. Most multi-exon genes in human undergo alternative splicing, leading to a variety of possible gene products, which can have different biological functions. Some of the most common alternative splicing events include exon skipping/inclusion, alternative 5' or 3' splice sites usage, intron retention, and mutually exclusive exons.

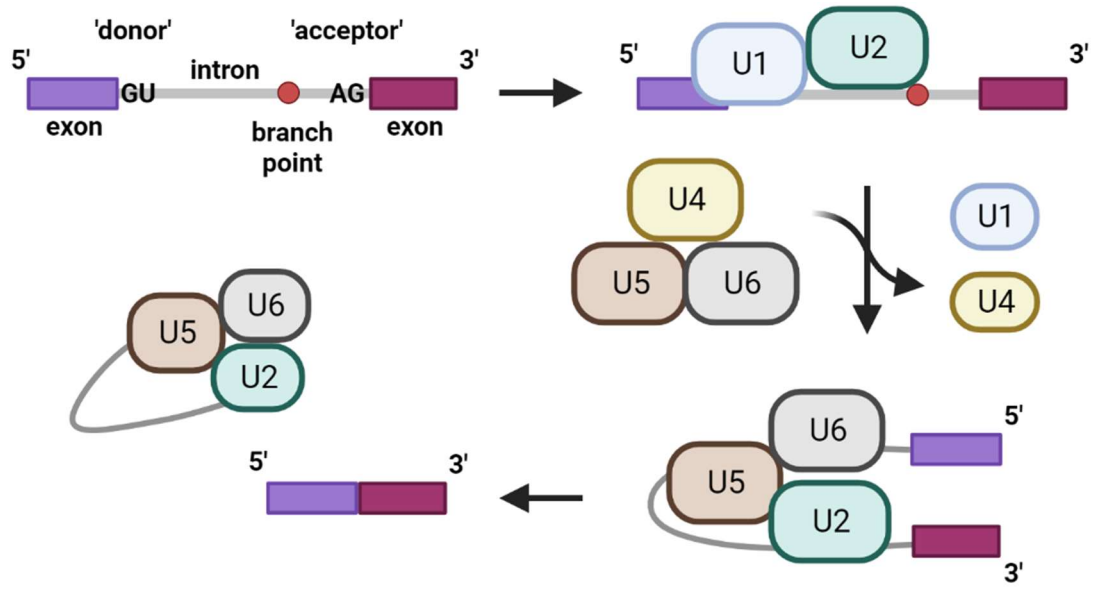


Figure 1.2: Canonical spliceosome assembly

Assembly of spliceosome begins with U1 snRNP recognising and binding to 5' splice site on the pre-mRNA. U2 snRNP then binds to intronic branch site, with auxiliary factors U2AF1 binding to the 3' splice site. U4/U6.U5 tri-snRNP rearranges the complex, leading to the release of U1 and U4 snRNP. The now activated spliceosome cleaves pre-mRNA at 5' splice site first, which creates a branched 'lariat' intron attached to the spliceosome complex at the branch site. Finally, pre-mRNA is cleaved at 3' splice site, with the exons ligated simultaneously. The spliced out intronic RNA is released alongside with U2, U5, and U6.

The decision of whether an exon is included in the matured mRNA is jointly controlled by both *cis*- and *trans*- acting regulatory elements. *Cis*-acting regulatory elements refer to the pre-mRNA sequence features that can influence how pre-mRNA molecules are spliced. In addition to the highly conserved 5' GU/ 3' AG dinucleotides that mark exon-intron boundaries, the sequence and structural context surrounding both 5' and 3' splice sites also directly impact the strength of interactions between snRNPs and nascent pre-mRNA. The 'strength' of the splice sites determines the likelihood of the exons being included in the matured mRNA and thus influences if a gene is constitutively or alternatively spliced (Wachutka *et al.*, 2019).

Other sequence motifs within exons and introns can also act as enhancers or silencers of the splice sites by acting as docking sites for RNA binding proteins (RBPs). The binding of these *trans*- acting RBPs can subsequently aid or disrupt the assembly of spliceosomes on the pre-mRNA. For example, the splicing factor serine and arginine-rich splicing factor 1 (SRSF1) binds to exon enhancer sites, which facilitates the recruitment of U1 snRNP to the 5' splice site and promotes the inclusion of the exon (Cho *et al.*, 2011). Conversely, heterogenous nuclear ribonucleoprotein A1 (hnRNP A1) can bind to intronic silencer sites and inhibit U1 snRNP binding to 5' splice sites. As the adjacent intronic/exonic RNA segments are forced to 'loop out', distant pre-mRNA splice sites can be brought close together. This leads to the exclusion of adjacent exons (Howard *et al.*, 2018). The cooperative and competitive binding of splicing factors to pre-mRNA represents a dynamic mechanism that regulates alternative splicing.

Alternative splicing is heavily influenced by transcriptional regulation. Majority of splicing events in human occur co-transcriptionally. RNA sequencing of nascent RNA purified from human cell lines and tissue shows that between 70-85% of intron removal events occur concurrently with transcription (Neugebauer, 2019). As RNA polymerase II (RNAP II) begins transcribing DNA into pre-mRNA, spliceosome assembles concurrently. U1 snRNP physically associates with RNAP II and captures pre-mRNA molecules at 5'splice sites at the RNAP II pre-mRNA exit site (S. Zhang *et al.*, 2021). The tail-like carboxy-terminal domain (CTD) of the RNAP II catalytic domain serves as a docking site for a wide range of splicing factors. Many interactions between CTD and splicing factors, including U2 snRNP, depend on the CTD's phosphorylation state (Hsin and Manley, 2012). The phosphorylation state of CTD is also critical in regulating other mRNA processing steps such as 5' capping, 3' cleavage and polyadenylation (Gu *et al.*, 2013; Davidson *et al.*, 2014; Noe Gonzalez *et al.*, 2018). Cyclin-dependent kinases (CDKs) and phosphatases dynamically modulate the phosphorylation state of RNAP II CTD as the transcription cycle progresses. Consequently, this changes the profile of splicing factors and resulting splicing patterns.

The speed of transcription elongation heavily influences splice site recognition and usage. A faster transcription rate reveals stronger splice sites more rapidly and 'mask' weaker splice sites. Conversely, a slower transcriptional rate may allow spliceosomes to assemble at weaker splice sites (Carrillo Oesterreich *et al.*, 2016). Many factors, including the gene structural features and sequence identity, control the transcription elongation speed. It was found that the length of the first intron and the number of introns present in a gene positively correlate with gene transcription speed (Fukaya, Lim and Levine, 2017). In contrast, the number of exons in a gene negatively correlates with the speed of RNAP II (Jonkers *et al.*, 2014). As DNA is unwound and transcription begins, DNA, RNA, or DNA: RNA hybrid secondary structures can form and act as 'speed bumps' to the RNAP II, leading to transcription pausing (Saba *et al.*, 2019). For example, the formation of RNA hairpin structures on pre-mRNA splice sites abrogates the assembly of spliceosomes, leading to exon exclusion (Neil and Fairbrother, 2019). RNA secondary structures are also essential recognition features for RBP binding, including members of the serine and Arginine rich (SR) and hnRNP families of proteins. Transcription speed and pre-mRNA splicing are intricately co-regulated by DNA/RNA secondary structure formations and RBPs binding profile.

Interestingly, the formation of functional spliceosome itself is also essential for gene transcription activities in eukaryotes. Upon transcription initiation, RNAP II pauses after 20 – 80 bases downstream in a large proportion (30 – 80% depending on studies) of genes in eukaryotes (Day *et al.*, 2016). Transcriptional activities can only be resumed upon recruitment of positive transcription elongation factor b (p-TEFb) to the promoter, which mediates phosphorylation of RNAP II CTD and other transcription factors (Lu *et al.*, 2016). Recent studies have shown that functional U2 snRNP is required for p-TEFb recruitment and, therefore, for effective transcription elongation. Inhibition of U2 snRNP branch point recognition by small-molecule inhibitors reduces nascent RNA biosynthesis significantly in human. This shows that productive gene expression depends on functional splicing activities (Chathoth *et al.*, 2014; Caizzi *et al.*, 2021).

1.1.2 Aberrant alternative splicing and cancer

Whilst alternative splice site usage generates diversity in mature mRNA transcripts and proteins, aberrant splice site usage can also lead to developmental defects and may have pathological consequences (Wang *et al.*, 2015). Computational analysis of RNA sequencing data from The Cancer Genome Atlas (TCGA) across 32 cancer types reveals that tumour tissues display up to a 30% increase in alternative splicing events compared to matched normal tissues, with exon skipping and alternative 3' splice sites representing 50%+ of all alternative splicing events (Kahles *et al.*, 2018). Somatic alterations in splicing factors are over-represented across different cancer types. Many recurrent mutations of splicing factors have now been linked to tumorigenesis and cancer development (Seiler *et al.*, 2018). For example, U2AF1 is a splicing factor responsible for U2 snRNP binding to the 3' splice site' AG dinucleotides (Figure 1.2). More than 10% of chronic myelomonocytic leukaemia (CMML) patients harbour mutations at the amino acids S34 or Q157, which causes abnormal alternative splicing and impaired haematopoiesis (Okeyo-Owuor *et al.*, 2015).

In addition to genetic mutations, splicing regulation in cancer is also influenced by the differential expression of splicing factors. For example, SRSF1 is highly expressed in many cancer types, including glioblastoma. SRSF1 binds to exons 23 and 24 of myosin 1B (MYO1B) pre-mRNA, which enhances their inclusion in glioma tissues compared to normal tissues. This promotes glioma cell proliferation and a more aggressive tumour phenotype. Furthermore, SRSF1 expression and MYO1B exon inclusion events correlate significantly with glioblastoma patients' prognosis and outcomes (Zhou *et al.*, 2019). hnRNPA1 is also highly expressed in multiple cancer types, and its expression is controlled by the frequently over-expressed oncogenic transcription factor c-MYC (MYC proto-oncogene). In glioma, hnRNP levels are upregulated. hnRNPA1, hnRNPA2 and hnRNPI bind to exon 9 of pyruvate kinase PKM pre-mRNA. This pivots the expression of PKM towards the exon9-excluded PKM2 mRNA isoform, which promotes aerobic glycolysis and tumour cell proliferation (David *et al.*, 2010).

Cell cycle and proliferation signal transduction pathways are often driven via kinase signalling. Hyperactivation of these pathways is also found to influence the phosphorylation state and activities of splicing factors. In cancer cells, the ribosomal protein S6 kinase 2 (S6K2) is often over-expressed and hyperactivated by the pro-survival mitogen activated protein kinase/extracellular signal regulated kinase pathway (MAPK/ERK) and phosphatidylinositol-3-kinase (PI3K) /protein kinase B (AKT) /mechanistic target of rapamycin (mTOR) pathways (Pardo and Seckl, 2013). In colorectal cancer, hnRNPA1 is found to be highly phosphorylated by S6K2, which promotes its binding to PKM pre-mRNA exon9 and increases PKM2 isoform expression. Like glioma, PKM2 promotes metabolic reprogramming in colorectal cancer cells, and its expression levels predict patient prognosis (Sun *et al.*, 2017).

A recent integrated multi-omics study on clear cell renal cell carcinoma (ccRCC) patients (n = 601) reveals 16 ccRCC-specific splice variants (validated against 4365 non-RCC tumours and 71 paired adjacent kidney tissue), where many of their expression significantly correlates with ccRCC disease progression and outcome. One of the ccRCC-specific splice variants is an alternatively 5’ spliced *EGFR* transcript, where its expression levels in patients' tumour tissue are associated with worse disease outcomes (Chang *et al.*, 2022). In ccRCC, high levels of EGFR expression have been linked with hyperactivation of AKT in ccRCC and poor patient prognosis (Q. Zhang *et al.*, 2019). Surprisingly, high gene expression levels of the ccRCC-specific EGFR splice variant were associated with genome-wide DNA hypo-methylation. Whilst the mechanistic links between the splice variant and DNA methylation in ccRCC are yet to be determined, this demonstrates a potential relationship between the splice variant with genome-wide expression remodelling, resulting in tumour progression and poor disease outcomes.

1.1.3 Alternative cleavage and polyadenylation

In addition to alternative splicing, eukaryotic cells can generate additional transcript isoforms with alternative 3' UTRs, via a regulatory mechanism known as alternative cleavage and polyadenylation (APA). Most eukaryotic pre-mRNA molecules are cleaved and polyadenylated at the 3' end by a multi-subunit processing complex known as cleavage and polyadenylation specificity factor (CPSF). CPSF complex recognises and binds to a hexanucleotide polyadenylation signal (PAS). This is facilitated by the recruitment of other multi-proteins complexes, including cleavage factor I (CFI), which binds to the consensus UGUA sequence upstream of PAS, and cleavage stimulation factor (CSTF) at downstream U / GU rich sequence. CPSF complex cleaves pre-mRNA at ~20 nucleotides downstream of PAS. It is now known that 70% of expressed genes exhibit APA (Neve *et al.*, 2017).

The most common human PAS sequence is AAUAAA, representing more than 47% of all PAS sites (Tian and Graber, 2012). Recent transcriptome-wide analysis reveals that more than 70% of all human protein-coding genes utilise multiple PAS, with an average of 2.5 PAS sites per gene (Djebali *et al.*, 2012). The usage of alternative PAS generates multiple mRNA isoforms with variable 3' ends, which are generated in 4 ways (Figure 1.3). Alternative PAS usage can result from differential usage of 3' terminal exon via alternative splicing (3'UTR APA). When APA occurs downstream of the stop coding at the 3'UTR, isoforms with different 3'UTR lengths are generated. Alternative PAS sites at introns and within exons have also been identified, resulting in changes in protein sequences (Elkon *et al.*, 2013). mRNA 3'UTR play essential roles in mRNA stability, localisation, and translation. Multiple post-transcriptional gene expression regulation mechanisms, mediated by RBPs and micro RNA (miRNA), target mRNA 3'UTR specifically (Mayr, 2019). Intronic and premature exonic APA can also result in alternative 3'UTR sequences and changes in amino acid sequences. Like alternative splicing, APA represents a key gene regulation mechanism that enables the generation of multiple isoforms from a single gene.

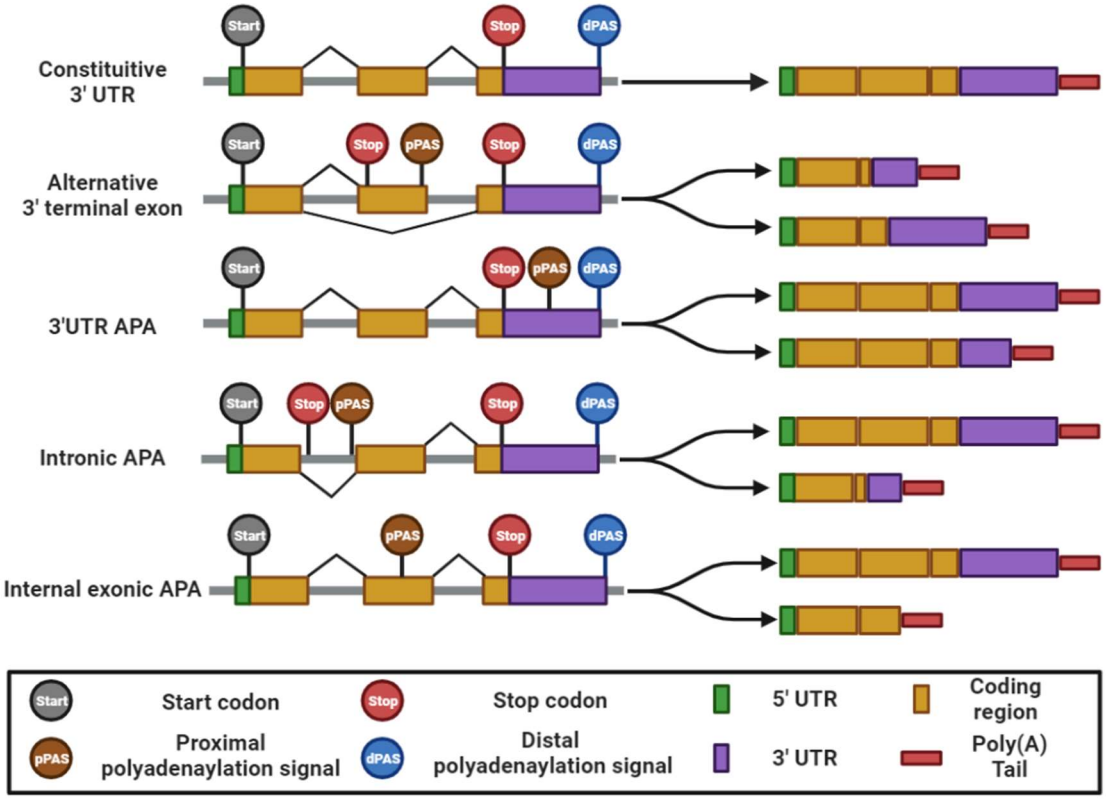


Figure 1.3: Types of APA and mechanisms of differential 3'UTRs generation

The CPSF complex utilises (PAS) to define the 3' end of an mRNA molecule. The majority of human genes contain multiple PAS, which allows APA and generation of different transcript isoforms. AS can produce transcripts with differential 3' terminal exons and PAS usage. Genes may also harbour multiple PAS in its 3'UTR, producing isoforms with identical coding sequence but different lengths of 3'UTR. Usage of cryptic PAS in the intronic and internal exonic regions can also generate transcript isoforms. Presence of a stop codon upstream of the alternative PAS can results in an alternative 3'UTR and the production of truncated proteins.

Like AS events, APA is also regulated by both *cis*- and *trans*- regulatory elements. In addition to the 'canonical' hexameric PAS sequence AAUAAA, 21 other PAS hexamers have been identified to support mRNA cleavage and polyadenylation in 87% of all known human transcripts (Gruber *et al.*, 2016). Different PAS sequences display a varying degree of 'strength', which contribute to prioritising specific isoforms over others, as seen in PAS usage from *in vivo* RNA sequencing data and *in vitro* PAS constructs reporter assays (Bogard *et al.*, 2019). Other RNA motifs, including the downstream U/GU rich sequence and upstream UGUA elements, also contribute to CPSF assembly and efficiencies. Analysis of RNA sequencing data using 3' end enriched mRNA of different human tissues shows that PAS usage and strength vary between tissues (Lianoglou *et al.*, 2013; Leung *et al.*, 2018). The differences in sequence characteristics between proximal and distal PAS sites and the abundance and activation states of CPSF components between cell types contribute to this variation. For example, depletion in the UGUA-binding CPSF5 and CPSF6 expression results in global 3'UTR shortening, whereas reduced expression levels of G/GU-rich binding PCF11 lead to 3'UTR lengthening (Ogorodnikov *et al.*, 2018). Global lengthening and shortening of 3'UTR have been observed in cells in response to cellular stress, differentiation, and pathogenic settings (Mayr and Bartel, 2009; Meng Chen *et al.*, 2018; Zheng *et al.*, 2018). This shows that APA is a dynamic gene expression regulatory mechanism, and cis-regulatory elements alone are insufficient to predict PAS usage.

The rate of transcription elongation and APA events are closely linked. Only the transcribed proximal PAS on pre-mRNA are available for CPSF complex assembly during the time between transcription of proximal PAS and distal PAS. Hence, a high transcription elongation rate favours the usage of distal PAS and the production of full-length mRNA transcripts, whereas slow transcript and transcription pausing promote the usage of proximal PAS and shorter isoforms (Goering *et al.*, 2021). In addition to DNA/RNA structures formed during transcription, RNAP II speed is influenced by chromatin structure and DNA methylation status. CCCTC-binding factor (CTCF) is a

crucial transcription factor that regulates chromatin structure by generating chromatin loops with the cohesin complex (Pugacheva *et al.*, 2020). CTCF binds to more than 50000 GC-rich sites across the genome, and DNA methylation at these CpG sites inhibits CTCF binding (Wang *et al.*, 2012). DNA-bound CTCF-cohesin complex induces RNAP II pausing. Moreover, chromatin loop formation mediated by the complex is also found to promote the usage of proximal PAS and the expression of short mRNA isoforms (Nanavaty *et al.*, 2020). This provides a direct link between the DNA-methylation state and APA events.

Transcription and APA can also be co-regulated by transcription factors. According to the ENCODE DNA functional elements repository, it is currently estimated that there are roughly 1 million enhancers in the human genome, which are short non-coding DNA sequences (average length of 423bp) that regulate gene expression(Abascal *et al.*, 2020). Enhancers contain transcription factor binding sites to help activate gene transcription in a cell-type-dependent manner (Mills *et al.*, 2020). Interestingly, a recent study shows that activation and increased binding of the master transcription factor Nuclear factor kappa B (NF-kB) to *PTEN* enhancer (Penh) promotes the shortening of *PTEN* (Phosphatase and tensin Homolog deleted on Chromosome 10) mRNA 3'UTR. The change in 3'UTR length is lost when the *PTEN* enhancer is removed via the CRISPR-Cas9 system or when function NF-kB transcription factors are depleted (siRNA-mediated knockdown of *RELA*, the gene that encodes NFkB p65 subunit). Using the enhancer-deleted cell line, as well as *in vitro* reporter system, the study confirms that the change in *PTEN* 3'UTR is due to changing preference of PAS, as there is no change in terms of mRNA production levels or differences in mRNA stability between isoforms (Kwon *et al.*, 2022). With the high abundance of enhancers in the human genome, this study demonstrates a potentially critical cell and gene-specific APA regulatory mechanism.

1.1.4 Alternative 3'UTR and cancer

Cancer transcriptomic studies have revealed that across different cancer types, more than 70% of mRNA transcripts exhibit 3'UTR shortening via APA (Xia *et al.*, 2014). Enhanced usage of proximal PAS sites in the 3'UTR removes miRNA and RBP binding sites in mRNAs, which makes the mRNA insensitive to miRNA- and RBP-mediated mRNA stability and translation regulation. Although there is data that suggest global 3'UTR shortening increases mRNA half-life in cancer tissues, its role on mRNA and protein gene expression levels appears to be more nuanced and gene-specific (Mayr and Bartel, 2009; Gruber *et al.*, 2014). Nevertheless, in the cancer setting, perturbation in APA has now been shown to drive tumour progression by promoting the expression of oncogenes and repressing tumour suppressors expression (Park *et al.*, 2018).

Dysregulation of APA is mediated by aberrant expression of proteins that regulate CPA. For example, expression levels of the CPSF complex component CPSF5 are significantly down-regulated in various cancer types. In glioblastoma tumour cells, the shortening of mRNA 3'UTR caused by CPSF5 depletion results in enhanced protein expression of pro-proliferative genes, such as glycogen synthase kinase three beta (GSK3b). Stable RNAi-mediated *CPSF5* gene knockout human glioblastoma tumours show increased growth *in vivo* when transplanted in mice, whereas *CPSF5* overexpressed tumours exhibit reduced proliferation (Masamha *et al.*, 2014). Suppressed CPSF5 expression levels are seen frequently across cancer types and significantly associate with poor patients outcome in glioblastoma, haepatocellular carcinoma and bladder cancer (Nourse *et al.*, 2020). Conversely in glioblastoma, siRNA-mediated depletion of PCF11 promotes 3'UTR lengthening, whereas high level of PCF11 expression is significantly correlated with transcript shortening and poor prognosis (Ogorodnikov *et al.*, 2018).

Other CPA factors regulate 3'UTR lengths in a more gene-specific manner. RNA sequencing analysis of a siRNA-mediated gene depletion screen shows that suppressed expression of RNA splicing factors and all known CPA regulators (174 genes KD in total) affect APA. On average, KD of CPA and splicing factors results in differential PAS usage

in 130 genes. Moreover, KD of most factors causes both an increase and decrease in 3'UTR lengths, and their targets are mainly mutually-exclusive (Ogorodnikov *et al.*, 2018).

CPSF complex components are also co-regulated, as demonstrated by a recent study on PCF11 and CPSF5. Melanoma antigen gene A11 (MAGE-A11) is an oncogene that functions as a substrate adapter for the E3 ubiquitin ligase HUWE1 (Yang, Huang, *et al.*, 2020). MAGE-A11 expression is generally restricted to cells from the testis and placenta, but high levels of expression are also found in different types of cancer. In MAGE-A11 expressing cancer cells, MAGE-A11 facilitates HUWE1-mediated PCF11 ubiquitination and subsequent degradation. However, PCF11 ubiquitination also facilitates the dissociation of CPSF5 from the CPSF complex. Thus, in contrast to 3' UTR lengthening seen in siRNA-mediated PCF11 degradation, ubiquitination/proteasome-mediated degradation of PCF11 results in the unexpected global shortening of 3'UTR. One of the oncogenes with shortened 3'UTR is cyclin D2 (*CCND2*), which plays a critical role in cell cycle regulation. With the loss of 3'UTR sequences, including the miR-191-5p binding site, the shortened *CCND2* 3'UTR upregulates CCND2 protein levels, consequently promoting cell proliferation in MAGE-A11 expressing human brain tumour cells (Yang, Li, *et al.*, 2020).

In summary, APA and 3'UTR length are regulated by both *cis*- and *trans*- regulators. Epigenetics, transcription, and expression levels of CPA regulation factors also control the selection of PAS sites. APA is widely dysregulated in tumour cells. With the many regulatory networks involved and the context-dependent nature of APA, much work is needed to understand its role in tumourigenesis and cancer development.

1.1.5 mRNA poly(A) tail

Most eukaryotic mRNA molecules are polyadenylated at the 3' end. These polyadenylated tracks, known as poly(A) tails, are a crucial determinant of the stability of mRNA molecules. After the mRNA molecule is cleaved by the CPSF complex at the 3' cleavage site, typically around 20 nucleotides downstream of PAS, the poly(A) tail is synthesised by the poly(A) polymerase (PAP). Upon addition of 11 – 14 adenosines, the RBP poly(A) binding nuclear protein 1 (PABPN1) binds to the newly synthesised tail and promotes the rapid, processive synthesis of the growing poly(A) tail (Neve *et al.*, 2017). Newly synthesised poly(A) tail length human is believed to be over ~200 nt long (Kühn *et al.*, 2009). However, poly(A) tail length is also dynamically controlled by cytoplasmic polyadenylation and deadenylation. Recent works suggest that the median poly(A) length of all mRNA molecules in a human cell range between 60 – 100 nt (Chang *et al.*, 2014; Soneson *et al.*, 2019). Curiously, poly(A) tails may also contain non-adenosine residues, albeit at a much lower frequency (~5% non-a residues appearing in ~15% of all mRNA molecules) (Y. Liu *et al.*, 2019; Liu *et al.*, 2021).

Once an mRNA molecule is fully matured and exported from the nucleus to the cytoplasm, its poly(A) tail is typically bound and coated by poly(A) binding cytoplasmic protein 1 (PABPC1). PABPC1 has an RNA binding footprint of ~30 nt, and binding between PABPC1 and poly(A) tail protects the mRNA molecule from deadenylation and degradation from 3' exonucleases. This is exemplified by the 30 nt incremental lengths seen at human poly(A) tails *in vivo* (Nicholson-Shaw *et al.*, 2022). In human, three major 3' exonucleases regulate the length of mRNA poly(A) tails: Pan2-Pan3 complex, CCR4-NOT complex, and poly(A)-specific ribonuclease (PARN) (Wolf and Passmore, 2014; Collart, 2016). Other RBPs mediate associations of Pan2-Pan3 and CCR4-NOT complexes to mRNA molecules. For example, tristetraprolin (TTP) is an RBP that binds to AU-rich elements (AREs) at the 3'UTR and recruits the CCR4-NOT deadenylase complex (Fabian *et al.*, 2013). This allows gene-specific (ARE-dependent) expression regulation in response to environmental cues, which is mediated by TTP expression

levels and the levels of competing AU-rich RNA binding proteins (AUBPs) (Shirai *et al.*, 2014). Other RBPs that recruit CCR4-NOT deadenylase complex include the mRNA m^6A binding protein YTHDF2 (YTH N6-methyladenosine RNA binding protein 2), representing an RNA modification-specific degradation mechanism, which will be discussed later in the chapter.

Regulation of poly(A) tail is closely related to mRNA translation. Mediated by PABPC1, the mRNA poly(A) tail can form a 'close loop' structure by interacting with the mRNA 5' cap. This close loop structure brings multiple translation initiation complexes in close proximity, facilitating their catalytic activities and subsequent recruitment of ribosomal subunits (Passmore and Coller, 2022). Early studies stipulated that poly(A) tail length (thus the number of bound PABPC1 proteins) positively correlates with translation efficiency(Sallés *et al.*, 1994; Sheets *et al.*, 1995). However, recent work suggests that this only applies to the early stages of development since a single poly(A) tail-bound PABPC1 protein is adequate for translation initiation. After the early embryonic stage, the PABPC1 expression level is more than sufficient to coat all mRNA molecules' poly(A) tails (Xiang and Bartel, 2021). Furthermore, poly (A) tail deadenylation rate, and thus the stability of mRNA molecule, depend on mRNA translation rate. A slow translation elongation rate is associated with a fast poly(A) tail deadenylation rate via the CCR4-NOT complex (Buschauer *et al.*, 2020). Determinants of translation rate include (but are not limited to) concentration and availability of sequence corresponding tRNAs, accessibility of coding sequences due to mRNA folding, and encoded amino acids side changes properties (Dana and Tuller, 2012).

Interestingly, highly expressed and translated genes tend to have short poly(A) tails (30-60 nt), which suggests that mRNAs with short poly(A) tails are not necessarily unstable (Lima *et al.*, 2017). Currently, there are conflicting reports on the correlation (or lack of) between poly(A) tail length and mRNA stability (Rissland *et al.*, 2017; Eisen *et al.*, 2020). However, the poly(A) tail is now widely established an essential feature that regulates gene expression by linking translation and environmental cues to mRNA stability.

1.1.6 m^{6}A mRNA modification

Cellular RNAs can be chemically modified in a myriad of ways. To date, more than 170 types of posttranscriptional modifications, including pseudouridine (Ψ), 5-methylcytosine (m^{5}C) and N^6-methyladenosine (m^{6}A), have been identified. m^{6}A methylation in mRNA is a reversible and highly dynamic process involving demethylase and methylation complexes. Approximately 0.1 – 0.4% of mRNA adenosine and 25% of all mRNAs are estimated to be m^{6}A modified, primarily at the consensus sequence 5'-DRACH-3' (D=A/G/U, R=G/A, H=A/C/U) (Nachtergaele and He, 2017). m^{6}A modification is highly dynamic and reversible by m^{6}A methyltransferase protein complex (writers) and demethylases (erasers). Different classes of RNA-binding proteins (RBPs), known as 'readers', can be specifically recruited to mRNAs with m^{6}A (Figure 1.5). In human, m^{6}As in mRNAs are mainly enriched near the 3' untranslated region (UTR), long exons (< 200nt) and near stop codons (Figure 1.4) (Dominissini *et al.*, 2013). m^{6}A is shown to play critical roles in mRNA splicing, stability, localisation and translation via regulation of RNA structure and RNA-RBPs interactions. (C. Zhang *et al.*, 2019).

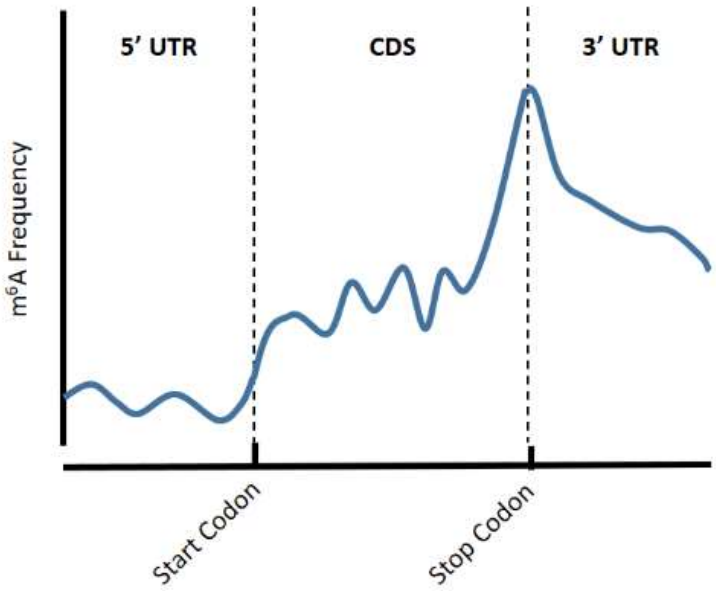


Figure 1.4: Graphical representation of the metagene m^{6}A distribution profile

m^{6}As are highly enriched at 3'UTR and exonic regions near the stop codon. Image adapted from Dominissini *et al.*, 2013.

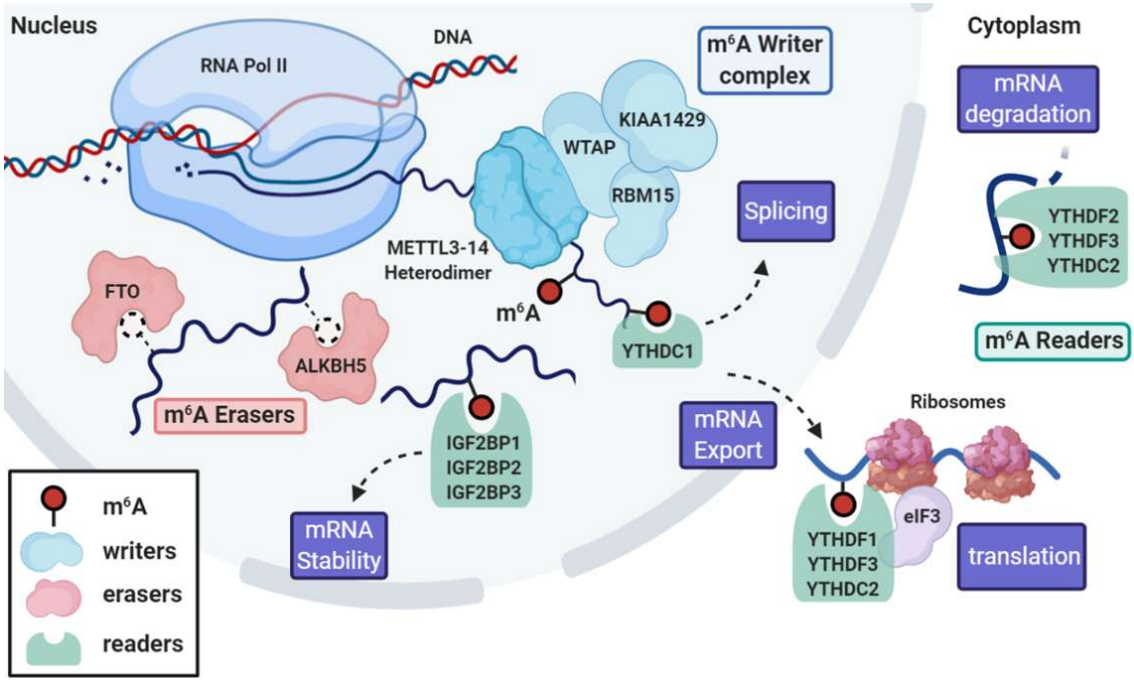


Figure 1.5: An overview of m^{6}A writers, erasers and readers

Deposition of m^{6}A mainly occurs in the nucleus by a methyltransferase 'writer' complex, comprised of the core METTL3 (methyltransferase 3)-METTL14 (methyl transferase 14) heterodimeric complex, as well as other adaptor proteins such as Wilms' tumour-associated protein (WTAP), RBM15, and KIAA1429. Interaction between m^{6}A and m^{6}A reader proteins exert a wide range of effects on the fate of m^{6}A containing mRNAs, such as mRNA localisation, splicing, stability and translation. The m^{6}A mark can also be removed by m^{6}A eraser proteins FTO and ALKBH5.

1.1.6.1 m6A writers

m^{6}A modification is highly dynamic and reversible by m^{6}A methyltransferase protein complex (writers) and demethylases (erasers). Different classes of RNA-binding proteins, known as 'readers', can be specifically recruited to mRNAs with m^{6}A (Figures 1.5 and 1.6).

The core of the m^{6}A writer complex consists of the METTL3-METTL14 heterodimeric complex. METTL3 and METTL14 are highly conserved in mammals and share 43% amino acid sequence identity in human. While METTL3 and METTL14 contain a methyltransferase domain (MTD), only METTL3 has a binding site for S-adenosylmethionine (SAM), the methyl group donor for methylation of adenosine of mRNA. METTL3 also harbours two tandem Cys-Cys-Cys-His (CCCH) zinc finger domains adjacent to the MTD, thus allowing mRNA binding (Wang *et al.*, 2016). METTL14 plays an allosteric role in supporting METTL3 enzymatic activity through structural stabilisation of the complex and RNA substrate recognition via its Arginine-Glycine repeats (RGG) at its C terminal terminus (Schöller *et al.*, 2018). With METTL3 containing a nuclear localisation sequence (NLS) at its N-terminal, the METTL3-METTL14 complex is predominantly found in the nucleus where it forms the m^{6}A writer complex with other 'adaptor' proteins such as WTAP (Wilms' Tumour 1 Associating Protein), RBM15 and KIAA1429. Interestingly, METTL3 is also found in the cytoplasm without forming heterodimers with METTL14 in human cancer cells and promotes translation initiation of selective mRNAs by recruiting eIF3 to the translation initiation complex independent of its methyltransferase activities (Lin *et al.*, 2016). This demonstrates an alternative role of METTL3 in gene expression regulation beyond the catalytic activity of m^{6}A modifications.

Several adaptor proteins of the m^{6}A methyltransferase complex are indispensable in regulating its activities. WTAP is a highly expressed splicing factor initially identified as the protein binding partner of Wilms' Tumour 1 (WT-1) oncoprotein. Deletion of WTAP in mice is embryonically lethal, and it was shown to facilitate G2/M cell cycle transition in

endothelial cells through stabilising cyclin A2 mRNA (Horiuchi *et al.*, 2006). WTAP associates with METTL3-METTL14 heterodimer, specifically the N-terminal part of METTL3, and is essential for the complex to localise in nuclear speckles (Liu *et al.*, 2014). Like METTL3, WTAP binds to the consensus m^{6}A RRACH motif in mRNAs. Depletion of WTAP by siRNA in HEK293-T and HeLa cells results in significantly lower levels of mRNA m^{6}A that are comparable to cells with either METTL3 or METTL14 knockdown (~50%), highlighting its importance in regulating mRNA m^{6}A methylation (Ping *et al.*, 2014).

WTAP also serve as a structural link between METTL3-METTL14 heterodimer and other adaptor proteins by forming direct interactions with RBM15 (and its paralogue RBM15B), KIAA1429 and ZC3H13, each of which when depleted, affects levels of m^{6}A in mRNAs (Schwartz *et al.*, 2014; Lence *et al.*, 2016; Knuckles *et al.*, 2018). RBM15/RBM15B binds to mRNAs preferentially near their m^{6}A sites and facilitates both m^{6}A formation and specific recognition of the long non-coding RNA X-inactive specific transcript (XIST) (Patil *et al.*, 2016). VIRMA, the *Drosophila* homologue of KIAA1429, is essential for m^{6}A deposition in 3' UTR and near stop codon specifically but with no effects on m^{6}A enrichment in long exons (Yue *et al.*, 2018). ZC3H13 also contributes to the specific enrichment of m^{6}A to 3' UTR and controls the nuclear localisation of WTAP (Wen *et al.*, 2018).

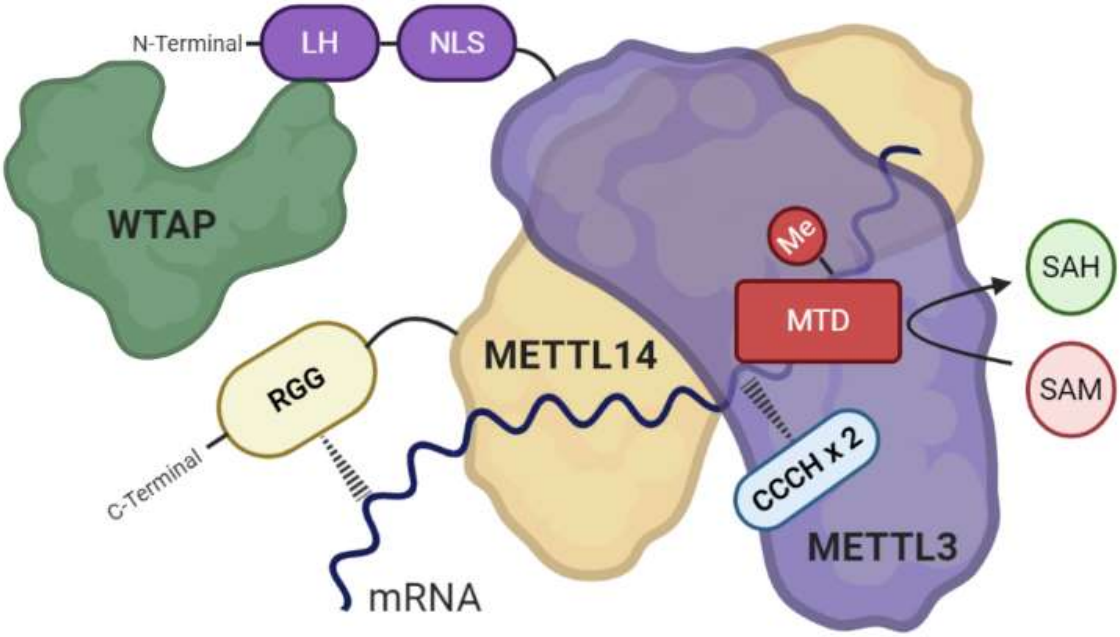


Figure 1.6: Schematic representation of the m^{6}A writer complex

METTL3 (Purple) is the sole catalytic subunit of the m^{6}A methyltransferase complex. Its zinc finger domains (CCCH x2) recognises mRNA molecules and methyltransferase domain (MTD) catalyses methyl group transfer from S-adenosylmethionine (SAM) to target adenosine, which yields S-adenosylhomocysteine (SAH). METTL14 (Yellow) acts as a RNA binding scaffold for the complex with a C-terminal Arginine-Glycine repeats (RGG) domain. WTAP associates with the heterodimer and is required for its localisation to nuclear speckles where pre-mRNA and splicing factors are enriched.

1.1.6.2 m^{6}A erasers

m^{6}A modifications in mRNAs can be removed by m^{6}A demethylases, or 'erasers'. FTO (Fat mass and obesity associated protein) and ALKBH5 (Alkb family member 5) are proteins from the AlkB family of alpha-ketoglutarate-dependent hydroxylase (of 9 human homologues) that have been identified to have RNA m^{6}A demethylation activity. The discovery of FTO's ability to catalyse oxidative demethylation of m6A in RNA coined the term 'epitranscriptome' (Jia *et al.*, 2011). *FTO* knockout mice show lean body mass and m^{6}A enrichment at the 5'UTR in mRNAs (Hess *et al.*, 2013). However, the m^{6}A antibody (Synaptic System) used in this study was found to cross-react with m^{6}Am (N^6, 2' -*O*-dimethyladenosine), a prevalent modification located near the mRNA 5' cap structure that contributes to mRNA stability. Notably, recent studies have also shown that FTO has a 100-fold higher catalytic activity for m^{6}Am compared to m^{6}A and acts as an m^{1}A (N^1 –methyladenosine) demethylase for tRNAs (Mauer and Jaffrey, 2018). FTO can be localised in the cytoplasm and the nucleus, with widely different distribution patterns between cell types that may confer on its substrate specificity (Wei *et al.*, 2018). The precise role of FTO on m^{6}A and mRNAs is still under debate. Contrary to FTO, ALKBH5 is predominantly located in the nucleus with only m6A, not m^{6}Am demethylase activity. Knockdown and knockout of *ALKBH5* results in a subtle, approximately 10 – 20% increase in mRNA m^{6}A levels (Zheng *et al.*, 2013). Knock out of *ALKBH5* in male mice causes infertility, and it was reported that ALKBH5-mediated removal of m^{6}A is essential for correct splicing in mRNAs with longer 3'UTR(Tang *et al.*, 2017).

1.1.6.3 m^{6}A readers

m^{6}A modification has been shown to play a regulatory role in many parts of mRNA metabolism, ranging from mRNA stability, folding, splicing, export, translation and decay(C. Zhang *et al.*, 2019). The main method by which m^{6}A can contribute to such diverse functions is by recruiting RBPs that specifically bind to m^{6}A, also named m^{6}A readers. The most well-characterised m^{6}A binding domain is the YTH domain (YT521-B homology).

YTHDF1, YTHDF2 and YTHDF3 are structurally similar cytoplasmic paralogues in human but differentially expressed across tissues and cell types. Early studies suggest that each paralogue exerts distinct effects on different subsets of methylated transcripts. YTHDF1 promotes m^{6}A-modified mRNAs translation by directly recruiting translation initiation factor eIF3 and bridges eIF4G, which links poly-A binding protein and the cap-binding eIF4E to form a circularised closed-loop mRNA structure (Xiao Wang *et al.*, 2015). YTHDF2 facilitates mRNA degradation with m6A via direct recruitment of the CCR4-NOT deadenylase complex, and YTHDF3 enhances both roles (Shi *et al.*, 2017). A recent interactomics study shows that the three YTHDF paralogues have highly similar protein binding partners and mRNA targets (Zaccara and Jaffrey, 2020). Although knockout of single paralogues has resulted in specific phenotypes, such as YTHDF1 KO mice showing synaptic defects and oocyte maturation arrest in YTHDF2 KO mice, each of these knock-out paralogues is also expressed at a much higher level than the other two paralogues in their respective tissues/cell-lines (Lasman *et al.*, 2020). In cell lines where all three YTHDF proteins are expressed at similar levels, the combined activity of the three paralogues results in the degradation of m^{6}A-modified mRNAs, which can only be ablated when all three proteins are depleted by siRNA.

Aside from the three YTHDF paralogues, YTHDC1 and YTHDC2 are the other m^{6}A readers containing the YTH domain. In contrast to the DFs proteins, YTHDC1 is exclusively localised in the nucleus. Loss of YTHDC1 in mouse oocytes causes alternative splicing defects, resulting in maturation arrest. YTHDC1 can also facilitate the

export of m^{6}A methylated mRNA transcripts from the nucleus by interacting with SRSF3 and the TREX mRNA export complex (Patil *et al.*, 2016; Roundtree *et al.*, 2017) (Patil, Chen, B. F. Pickering, *et al.*, 2016; Roundtree *et al.*, 2017; Lesbirel *et al.*, 2018). YTHDC2 is a perinuclear RNA helicase that positively regulates the translation of mRNAs with m^{6}A modifications by resolving its secondary structures (Hsu *et al.*, 2017). The mechanism by which YTHDC2 increases translational efficiencies of m^{6}A-modified mRNA transcripts and whether it works cooperatively with other m^{6}A readers/METTL3 to promote translation remains unclear.

Another critical role of m^{6}A in mRNAs is promoting mRNA stability through binding IGF2BP (insulin-like growth factor-2 mRNA binding protein) proteins. In mammals, there are 3 IGF2BP paralogues (IGF2BP1, 2 and 3) that directly recognise and bind to m^{6}A via their K-homology (KH) domains (Huang *et al.*, 2018). Moreover, IGF2BP proteins recruit HuR, an RBP that promotes target mRNA stability by competing for AU-rich elements (ARE) occupancy in the 3' UTR against mRNA destabilisers and microRNAs. Lastly, m^{6}A can also infuence RBP binding affinity by remodelling local RNA structure. One of such examples is hnRNP C. hnRNP C binds to *Malat1* in a m^{6}A dependent manner. m^{6}A methylation at the A2577 site exposes an U-tract with 5 contiguous uridines at the hairpin-stem opposing the m^{6}A site, where hnRNP C has a strong-affinity for. Combining CLIP (Crosslinking immunoprecipitation) and transcriptome wide m^{6}A mapping analysis, it was found that KD of *METTL3* and *METTL14* significantly reduced hnRNP C binding to more than 2,500 RNA targets (Liu *et al.*, 2015).

Whilst readers recognise RNA in an m^{6}A dependent manner, different classes of readers share few RNA targets. For example, IGF2BPs and YTHDF2 share only ~1% of their binding partners (Huang *et al.*, 2018). This suggests that the subsets of m^{6}A methylated mRNA transcripts are predetermined to be regulated by different m^{6}A readers based on their sequences and structures. Altogether, the m^{6}A writers, erasers and readers present a complex multi-layered regulatory network that affects numerous aspects of mRNA processing and metabolism, with profound physiological effects if dysregulated.

1.1.7 m^{6}A RNA modifications and cancer

Emerging evidence suggests that m6A writers, erasers, and readers can play oncogenic or tumour-suppressive roles in cancer, with their aberrant expressions tightly associated with tumour progression (Lin *et al.*, 2016; Chen *et al.*, 2019; Wu *et al.*, 2019). For example, in human and mouse Acute Myeloid Leukaemia (AML) cells, METTL3 and METTL14 mRNA and protein levels are expressed at a higher level than normal Haematopoietic Stem and Progenitor Cells (HSPCs). Elevated mRNA m^{6}A levels of important oncogenes such as *c-MYC*, *PTEN* and *BCL2* (B-cell lymphoma 2) result in their higher translational efficiency and overexpression in AML cells and proved essential for AML cell proliferation and differentiation (Barbieri *et al.*, 2017). YTHDF2 is also highly expressed in human AML clinical samples at the protein level. Deletion of *YTHDF2* from mouse leukaemic stem cells (LSCs) and human AML cells reduce their engraftment and propagation capacities, and transcriptomic analysis of *YTHDF2* knockout mouse LSCs shows that YTHDF2 proteins specifically target and degrade a subset of mRNA associated with leukaemogenesis (Paris *et al.*, 2019). Intriguingly, the m^{6}A eraser *ALKBH5* is also found to be aberrantly overexpressed in AML patients compared to normal HSCs, and it is associated with poor prognosis in AML patients. Deletion of *ALKBH5* in murine AML shows that it is selectively required for AML development and propagation. Combining transcriptomics and RIP-seq (RNA immunoprecipitation) analysis, depletion of *ALKBH5* (via gene KO and KD) results in enhanced mRNA transcript stability and heightened expression of *AXL* and *TACC5*, both of which have been shown to promote leukaemogenesis (J. Wang *et al.*, 2020; Shen *et al.*, 2020). The examples mentioned above show that although abnormal expression levels of various m^{6}A-associated genes contribute to tumourigenesis, their expression levels may not tilt towards increased or decreased global mRNA m^{6}A levels. Instead, in tumour cells, m6A writers, readers and erasers likely target subsets of oncogenic/tumour-repressive gene transcripts cooperatively to promote its growth and proliferation.

1.1.8 mRNA export and cancer

After transcription in the nucleus, mRNA molecules need to be exported to the cytoplasm to be translated. This process is highly regulated and plays a critical role in gene expression regulation. Moreover, mRNA export is functionally coupled to other RNA processing steps to avoid generation of aberrant proteins (Wickramasinghe and Laskey, 2015). In human, RBPs involved in mRNA capping, splicing, m^6A modification, polyadenylation and cleavage bind to pre-mRNA molecules co-transcriptionally to form messenger ribonucleoprotein (mRNP) complexes. mRNPs are recognised by transcription-export complex (TREX) through multivalent interactions between EJCs and the TREX subunit Aly/REF export factor (ALYREF) (Pacheco-Fiallos *et al.*, 2023). The TREX-bound mRNPs recruits the heterodimeric mRNA export factors Nuclear RNA export factor 1 (NXF1) and NTF2-related export protein 1 (NXT1). This facilitates the binding of mRNP to the nuclear pore complex and subsequent export of matured mRNA in the 5' – 3' direction to the cytoplasm (Pühringer *et al.*, 2020).

Aberrant RNA export has been linked to many forms of cancer. ALYREF expression was found to be significantly upregulated in many cancer types. In haepatocellular carcinoma patients, high ALYREF expression significantly correlates with worse prognosis and disease outcomes (Xue *et al.* 2021). TREX is also linked to cancer progression through its role in maintaining genome stability. During transcription, nascent RNA can bind to the template DNA strand, leaving the opposing nontemplate DNA single-stranded. These three-stranded structures are known as R-loops. The single stranded DNA exposed by R-loop formation is susceptible to DNA damage. Persistent R-loop generation is now recognised as a major source of genome instability, a hallmark of cancer(Wells *et al*. 2019), In human, the THO subcomplex of TREX recruits DNA-RNA helicases DDX5 and DDX17 to resolve R loops. Depletion of TREX increases genome instability and R-loops accumulation (Polenkowski *et al.* 2023). This demonstrates the intimate interplay between genome stability, transcription and RNA processing pathways, which has proufound implications on cancer development.

1.2 Renal cell carcinoma

This thesis will focus on kidney cancer, one of the leading contributors to the global disease burden, accounting for approximately 2% of all newly diagnosed cancer cases and cancer death worldwide in 2020 (Sung *et al.*, 2021). The most common form of kidney cancer is renal cell carcinoma (RCC), constituting 80 – 95% of all primary renal neoplasms, with the remaining cases coming from transitional cell cancer (TCC) of the renal pelvis (Escudier *et al.*, 2019). RCC originates from kidney tubular epithelial cells, representing a phenotypically and genetically diverse group of malignancies. RCCs are traditionally classified into different subtypes based on their distinct morphological characteristics, including ccRCC, papillary renal cell carcinoma (pRCC), and chromophobe renal cell carcinoma (chRCC), of which ccRCC is the most frequent type accounting for ~75% of all diagnosed RCC cases respectively (Kovacs *et al.*, 1997). While recent advances in cancer genomics and molecular biomarkers discovery have led to reclassifications of tumours by the World Health Organisation, these RCC subtypes have overall displayed distinct genetic alterations and molecular profiles (Moch *et al.*, 2022).

1.2.1 *VHL* loss as an oncogenic driver in ccRCC

The key genetic hallmarks of ccRCC, also known as KIRC (kidney renal clear cell carcinoma), are deletion of the short arm of chromosome 3 (3p) and loss of tumour suppressor gene von Hippel-Lindau (*VHL*) via either inactivating point mutations or deletion (Jonasch *et al.*, 2020). Genomic data from TCGA shows that 91% of ccRCC samples contain an arm-level loss of chromosome 3p. Other common arm-level somatic copy number alterations include the deletion of chromosome 14q (45% of all sequenced samples) and chromosome 5q gain (67% of all sequenced samples) (Creighton *et al.*, 2013). Loss of chromosome 3p results in deletion and loss of heterozygosity (LOH) of multiple tumour suppressive genes, such as *VHL* (3p25), *PBRM1* (Polybromo 1), *SETD2* (SET domain containing 2), *BAP1* (BRCA1 associated protein 1). Intriguingly, they are

also the most significantly mutated genes in various study cohorts, with somatic mutations in ~85%, ~40%, ~15% and ~15% of sequenced ccRCC patients' samples, respectively. Combined genomic-transcriptomic-proteomic approach has subsequently shown that these gene inactivation events result in down regulations of mRNA and protein expression (Clark *et al.*, 2019). Taken together, the loss of the 3p arm, subsequent inactivation and biallelic loss of *VHL* represent the signature ccRCC initiation events (Mitchell *et al.*, 2018).

Prior to the discovery of the connection between the *VHL* gene and ccRCC, the *VHL* gene was first discovered in the 1990s, when the autosomal hereditary Von Hippel-Lindau disease was found to be caused by inheritance of germline mutation (deletion, truncation, or missense mutation) in the gene (Gossage *et al.*, 2014). VHL disease *VHL* encodes pVHL, the substrate recognition component of an E3-ubiquitin ligase (VCB-CR complex consists of Elongin C, Elongin B, Cullin-2 and Rbx1). Under normoxia conditions, the alpha subunits of Hypoxia Inducible Factor (HIF-α, including HIF1α and HIF2α) are hydroxylated by prolyl hydroxylases (PHDs) (Ivan *et al.*, 2001). The VCB-CR complex targets the hydroxylated forms of HIF-α proteins specifically for proteasomal degradation via polyubiquitination (Salceda and Caro, 1997). In hypoxic conditions, hydroxylation reactions of HIF-α proteins by PHDs are inhibited due to the lack of oxygen, allowing HIFα to form a stable heterodimer with HIF-1β (Figure 1.7). The heterodimeric transcription factors then translocate to the nucleus and bind to the promoters of thousands of target genes containing hypoxia response elements (HRE) to activate their transcription, many of which are strongly oncogenic. Loss of *VHL* in ccRCC prevents targeting of HIF-α for proteasomal degradation, which leads to constitutively active HIF signalling and hypoxic response pathway (Schödel *et al.*, 2016).

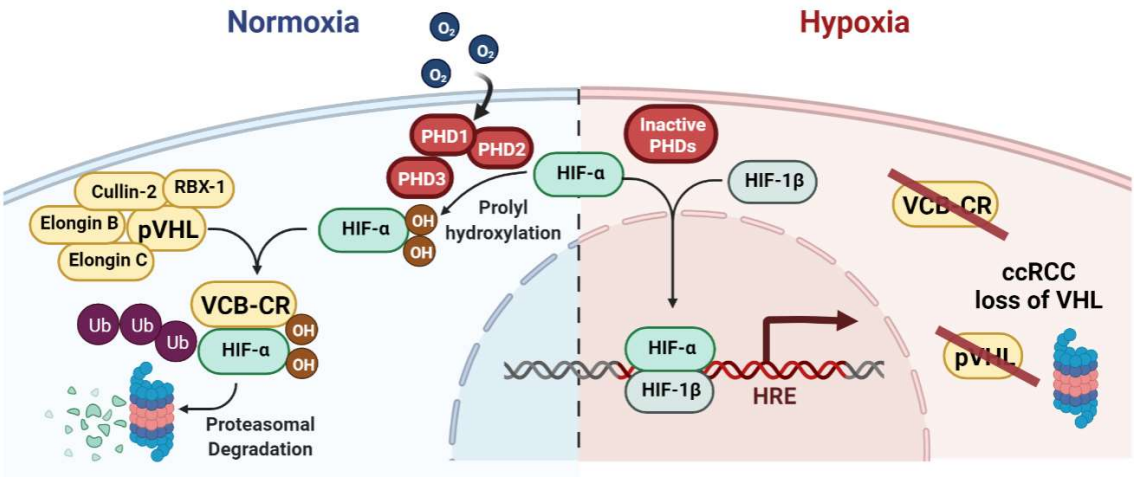


Figure 1.7: VHL/HIF hypoxic response pathway and ccRCC

Cellular hypoxic response is transcriptionally regulated by HIF, which is a heterodimer consists of constitutively expressed HIF-1β and HIF-α (including HIF1A, HIF2A and HIF3A). Under normal physiological oxygen levels, HIF-α proteins are hydroxylated by PHDs and recognised by pVHL of the VCB-CR E3-ubiquitin ligase complex for proteasomal degradation. Under hypoxic condition, PHDs are no longer able to hydroxylates HIF-α due to the lack of oxygen, allowing the formation of HIF heterodimers and induce transcription of genes containing promoters with HRE sequence where HIF binds. In ccRCC, frequent loss of *VHL* leads to accumulation of HIF-α, leading to hyperactive HIF signalling and hypoxic response pathway.

1.2.2 Metabolic rewiring in ccRCC

Pseudo-hypoxic state driven by a constitutively active HIF signalling pathway directly affects other metabolic and signalling pathways, with far-reaching effects on ccRCC progression and aggressiveness. HIFα proteins (HIF1α and HIF2α) are known to regulate the expression of thousands of genes, including a coordinated up-regulation in genes encoding glucose transporters (GLUT) and practically all enzymes in the glycolytic pathway (Downes *et al.*, 2018). Moreover, HIFs activate the expression of pyruvate dehydrogenase kinase 1 (*PDK1*), which inhibits the activities of mitochondrial pyruvate dehydrogenase via phosphorylation. Consequently, usage of the tricarboxylic acid (TCA) cycle and oxidative phosphorylation in HIF-activated cells are suppressed (Papandreou *et al.*, 2006). The rewiring of ATP production from primarily utilising mitochondrial oxidative phosphorylation to cytoplasmic aerobic glycolysis, also known as the Warburg effect, has long been observed and regarded as a hallmark of cancer (Warburg *et al.*, 1927). Many cancer types have since been shown to maintain high levels of oxidative phosphorylation as the primary source of ATP production, nevertheless, and even significantly enhanced in the case of non-small cell lung cancer and glioblastoma (Maher *et al.*, 2012). Recent *in vivo* work by Courtney *et al.* studying ccRCC patients infused with [^{13}C]glucose reveals a high level of lactate labelling (stemmed from aerobic glycolysis), and less than 5% of labelled carbon enters the TCA cycle in ccRCC tumours (Courtney *et al.*, 2018). This confirms that ccRCC tumours are metabolically reliant on the Warburg effect.

The defining morphological characteristic of ccRCC is the cytoplasmic accumulation of glycogen and lipid droplets. In addition to the reduced mitochondrial content due to metabolic rewiring, the cytoplasm of ccRCC tumour cells appears ‘clear’ in histological images (Nilsson *et al.*, 2020). Like the enhanced glycolytic pathway usage, increased glycogen and lipid droplet synthesis levels are also the consequence of the constitutively active HIF signalling pathway. Analysis of TCGA KIRC RNA transcriptomic data indicates that various glycogen synthases (*PGM1*, *GYS1*, *GBE1*) are overexpressed in ccRCC

tumours, although their role in ccRCC tumorigenesis and progression appears to be limited (Xie *et al.*, 2021). In contrast, activation of the HIF pathway suppresses expression levels of *CPT1A* (Carnitine palmitoyltransferase 1A), which is essential for transporting cytoplasmic fatty acid into the mitochondria to go through beta-oxidation. The retained fatty acids in the cytoplasm are subsequently converted into lipid droplets as energy storage, and ablated *CPT1A* expression is essential for ccRCC tumour formation *in vivo* (Du *et al.*, 2017). The constitutively active HIF pathway in ccRCC tumour cells rewires different metabolic pathways, which is essential for ccRCC tumorigenesis and progression.

1.2.3 Dysregulated signalling pathways in ccRCC

To sustain unlimited proliferation, tumour cells must both activate proliferative signalling, as well as deactivate the intrinsic negative feedback loop that prevents uncontrolled growth to sustain unlimited proliferation (Hanahan and Weinberg, 2011). One of the most dysregulated pathways in cancer is the PI3K/AKT/mTOR signalling axis. A graphical illustration of the interplay between PI3K/AKT/mTOR and hypoxic reponse pathways in ccRCC can be found at Figure 1.8.

PI3K/AKT/mTOR pathway is highly activated in various cancer types, including ccRCC. Whilst many tumours activate this via genetic mutations or chromosomal alterations on genes in this pathway, it is thought that PI3K/AKT/mTOR pathway activation in ccRCC is contributed mainly by the highly activated HIF pathway (Guo *et al.*, 2015). Firstly, HIF drives the transcription of various growth factors (VEGFs, IGFs, EGFs, and PDGFs), which promote the activation of the PI3K/AKT/mTOR pathway (Masoud and Li, 2015). In addition to activating PI3K/AKT/mTOR pathway, overexpression of VEGF also plays an essential role in promoting ccRCC angiogenesis by promoting vascular endothelial cell proliferation, migration, and vascular permeability (Pal *et al.*, 2019). Recent studies have also expanded VEGF's role in ccRCC progression by promoting cancer stem cell survival and proliferation and inhibiting immune cell infiltration (Clark *et al.*, 2019; Wang *et al.*, 2021). Moreover, mRNA and protein expression of many RTKs (such as VEGFR,

PDGFR, and EGFR) are also HIF target genes and their expressions are up-regulated in ccRCC tumours (Behbahani *et al.*, 2011). Finally, RTK activation can also drive other cell growth pathways, such as the MAPK/ERK pathway. Together, RTKs at ccRCC tumours are more activated (represented by their phosphorylation state) compared to corresponding adjacent normal kidney tissue (Q. Zhang *et al.*, 2019). This positive feedback loop further enhances the activation of PI3K/AKT/mTOR and the HIF pathway where HIF1α translation is induced by activated mTORC1.

Downstream of RTK activation, various parts of the PI3K/AKT/mTOR pathway are frequently altered in cancer at the genetic, gene expression or post-translational modification level. Whilst direct genetic alteration events on the PI3K/AKT/mTOR pathway are relatively infrequent compared to the near-universal loss of chromosome 3p arm and *VHL* loss, they are consistently pro-activation, either by direct amplification/activation mutation of genes in the pathway or deactivating the negative feedback regulators. Genomic analysis on tumour samples from 25 cancer types in TCGA suggests that on average, 50% of tumours show gene mutation or/and copy number alterations of the PI3K/AKT pathway. In comparison, amongst the 500+ ccRCC tumours surveyed by TCGA, 27.7% of samples harbour gene mutation/copy number alterations in at least 1 PI3K/AKT pathways components (Guo *et al.*, 2015). The genes found with genetic alterations include *mTOR* mutations (~6%) and *PTEN* gene deletions/mutations (~5%). In addition, a point mutation at RheB tyrosine 35 to asparagine (Y35N) was also shown to be sufficient to increase mTORC1 signalling and induce oncogenic transformation in normal cells (Heard *et al.*, 2018). Interestingly, along with *VHL*, gene inactivation of *PBRM1*, *BAP1* and *SETD2* at chromosome 3p have also been shown to individually contribute to the activation of the PI3K/AKT/mTOR pathway (Peña-Llopis *et al.*, 2012; Terzo *et al.*, 2019; Tang *et al.*, 2022).

Negative regulators of the PI3K/AKT/mTOR pathways are often downregulated/disabled in cancer. Multiple studies have suggested that 30-40% of ccRCC tumours display a loss of PTEN protein expression, as measured by immunohistochemical analysis (IHC) and

fluorescent *in situ* hybridisation (FISH) (de Campos *et al.*, 2013; Millis *et al.*, 2016). Decreased PTEN expression also significantly correlates with increased activated (phosphorylated) AKT levels in ccRCC tumour samples (H. Wang *et al.*, 2015). Like HIFα proteins, AKT is hydroxylated by prolyl hydroxylase (prolyl hydroxylases two specifically) under normoxia. VHL protein binds specifically to hydroxylated AKT and inhibits its kinase activity. Under hypoxia, or in ccRCC tumour cells, VHL proteins are depleted, which allows AKT to activate downstream targets and promote cell survival and proliferation (Guo *et al.*, 2016). This represents a non-HIF-dependent role of *VHL* loss in activating the PI3K/AKT/mTOR pathway. Finally, protein expression of the TSC subunits TSC1 and TSC2 are found to be suppressed in ccRCC tumour samples compared to corresponding normal kidney tissue, thus allowing accumulation of RheB-GTP and high levels of mTORC1 activation (Damjanovic *et al.*, 2016). In summary, the characteristically up-regulated PI3K/AKT/mTOR pathway in ccRCC results from VHL loss, multiple genome alterations, aberrant regulation of gene expression, and post-translational modifications.

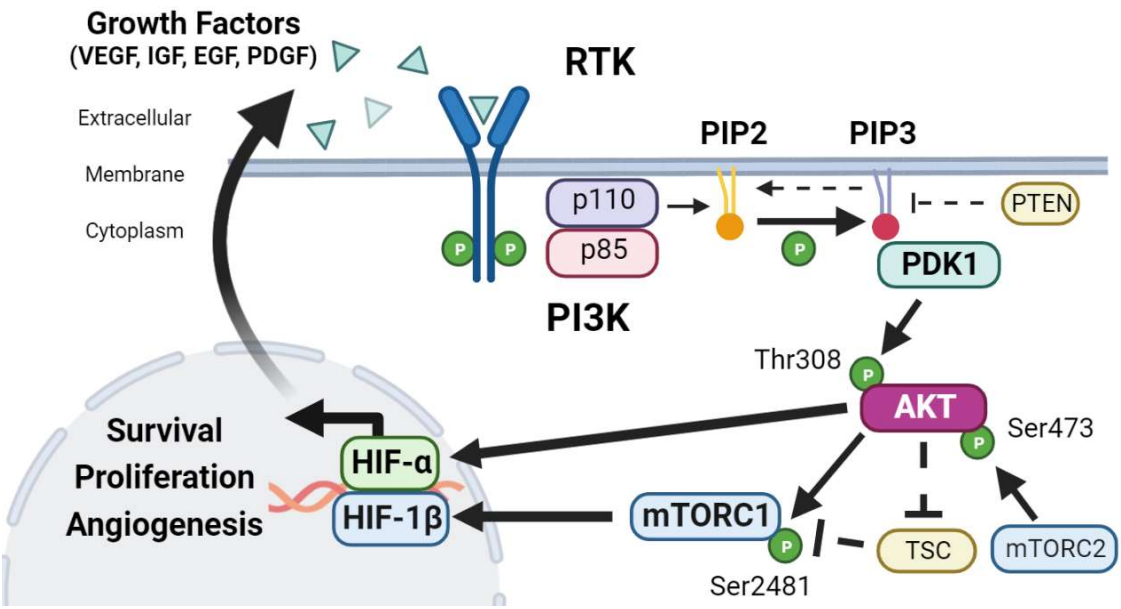


Figure 1.8: PI3K/AKT/mTOR pathway and ccRCC

PI3K/AKT/mTOR pathway is activated by the binding of growth factors (such as VEGF, IGFs, EGFs and PDGFs) to RTKs, resulting in RTK dimerisation and autophosphorylation. Phosphorylated RTK interacts with the p85 regulatory subunit of PI3K, which activates the p110 catalytic PI3K subunit. Activated PI3K complex phosphorylates PIP2 to PIP3, allowing recruitment of PDK1 and AKT. Phosphorylation of AKT by PDK1 and mTORC2 activates its kinase activity, allowing AKT to phosphorylate numerous downstream substrates including mTORC1, and promote cell survival and proliferation. The constitutively active HIF pathway in ccRCC upregulates the expression of growth factors, which in turn activates the PI3K/AKT/mTOR pathway. Multiple RTKs are found to be up-regulated and highly phosphorylated in ccRCC tumours. Genetic alterations on both positive (p110, mTOR) and negative regulators (PTEN) of the pathway are also found in ccRCC tumours, which further enhance PI3K/AKT/mTOR pathway activation. Expression levels of negative regulators of the pathways, such as PTEN, VHL and TSCs, are also suppressed in ccRCC tumours.

1.2.4 ccRCC microenvironment

Tumour tissue is a hugely complex and dynamic environment where different subpopulations of cells with distinct phenotypes and genotypes co-exist (Figure 1.9). The tumour microenvironment (TME) consists of tumour cells and non-malignant stromal fibroblasts, vasculature, infiltrating immune cells, and extracellular matrix (ECM). Interactions between malignant cells and non-malignant components of the TME play crucial roles in tumour progression and cancer treatment outcomes (Anderson and Simon, 2020).

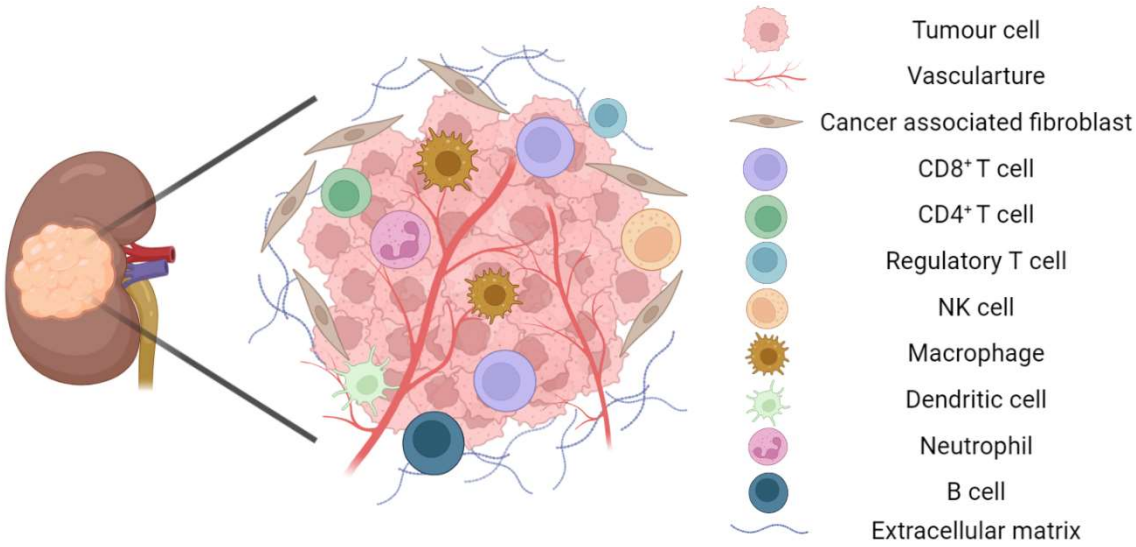


Figure 1.9: Tumour microenvironment (TME)

TME is a complex environment containing tumour cells, stromal cells (vascular endothelial cells, cancer associated fibroblasts), immune cells ($CD8^+$ T cells, $CD4^+$ T cells, regulatory T cells, natural killer (NK) cells, macrophages, dendritic cells (DCs), neutrophils and B cells), and non-cellular components such as the extracellular matrix (ECM). Each component may play tumour-inhibitory/tumour-promoting roles. To support tumour growth and survival, tumour cells manipulate the TME by expression and secretion of growth factors and immunomodulatory molecules (e.g. cytokines and chemokines), metabolic rewiring and ECM alterations.

1.2.4.1 TME: Immune cell populations

Immune cells form an integral part of the TME. Amongst different cancer types, ccRCC tumours present one of the highest median percentages of tumour-infiltrating immune cells at approximately 30% of all cells in the TME, as determined by both flow cytometry, bulk- and single-cell transcriptomic analysis (Aran *et al.*, 2015; Hu *et al.*, 2020). Depending on the tumour type and the TME context, different tumour-infiltrating immune cells can play both pro- and anti-tumour roles and act as fundamental determinants of tumour development and treatment outcome (Giraldo *et al.*, 2019). Different immune cells have been identified in ccRCC TME, including T cells, macrophages, NK cells, DCs, neutrophils, B cells, granulocytes and plasma cells. The two main immune cell populations are T cells (~50%) and macrophages (~30%) (Su *et al.*, 2021). Curiously, unlike most solid tumours, high levels of tumour-infiltrating immune cells in ccRCC, including the generally considered anti-tumour $CD8^+$ T cells, significantly correlates with worse prognosis and disease outcomes (Fridman *et al.*, 2017).

1.2.4.2 TME: $CD4^+$ and $CD8^+$ T cells

T cell-mediated immunity is regarded as the primary immune response against cancer in human. Amongst the T cell populations, the major constituents in TME are the $CD8^+$ and $CD4^+$ T cells. $CD8^+$ T cells, or cytotoxic T lymphocytes (CTLs), represent the key anti-tumour immune response by recognising and direct killing malignant cells. This is primarily achieved by three methods: secretion of the cytokines interferon-gamma (IFNγ) and tumour necrosis factor (TNF), delivery of perforin and granzyme containing cytotoxic granules to malignant cells, and ligation of its FasL (Fas Ligand) with Fas receptor on target cells (Raskov *et al.*, 2021). $CD4^+$ T cells are known as T helper cells (Th) for their critical roles in helping other effector immune cells to activate. Different subsets of effector $CD4^+$ T cells secrete specific cytokines in response to the environment to orchestrate the appropriate adaptive immune response. $CD4^+$ T cells can also play a key role in suppressing the hyperactive immune response, primarily achieved by expressing and releasing effector cytokines, such as IL-10 (Interleukin 10). In tumour immunity, CD4+ T cells help CD8+ T cells by secreting IL-2, a cytokine essential for $CD8^+$ T cell differentiation and effector functions. Like $CD8^+$ T cells, $CD4^+$ T cells can also exert anti-tumour activities by releasing IFNγ and TNF(Tay *et al.*, 2021).

T cells' early progenitors originate from haemopoietic stem cells in the bone marrow. The progenitors first migrate to the thymus, where they undergo various maturation stages, resulting in lineage commitment where each cell commits to differentiate into either a $CD4^+$ or $CD8^+$ T cell (as classified by expression of CD4 or CD8 protein). T cell receptors (TCR) also undergo rearrangement to generate a diverse antigen-binding repertoire (Kumar *et al.*, 2018). The now mature naïve T cells subsequently migrate to the secondary lymphoid organs (such as lymph nodes and spleen) or tertiary lymphoid structures (TLS), frequently organised adjacent to tumours (Sautès-Fridman *et al.*, 2019).

Naïve T cells are activated via a 3-signal system (Figure 1.10). Signal 1 comes from the presentation of antigen peptide on the major hiscapability complex (MHC) class I / II molecules of professional antigen presentation cells (such as DCs in the secondary

lymphoid/TLS) or antigen-presenting cells (APC, such as tumour cells), to the TCR on $CD8^+$ and $CD4^+$ T lymphocytes respectively. HLA-antigen-TCR binding is aided by CD4/CD8 molecules, which act as coreceptors of the TCR complex. CD4/CD8 are essential for amplifying TCR-signalling and productive T cell activation (Mørch *et al.*, 2020).

Signal 2 represent the engagement between APCs and T cells' co-stimulatory signal molecules. For example, binding between T cell CD28 receptors with CD80/CD86 from APC promotes activation of TCR signalling and subsequent T cell activation. Conversely, T cells also express co-inhibitory receptors (also known as immune checkpoints), such as Protein cell death protein 1 (PD-1) and cytotoxic T-lymphocyte-associated protein 4 (CTLA4). T cell activation is restrained upon ligand engagement between these co-inhibitory receptors and their ligands. Expression levels of these co-inhibitory receptors are up-regulated in T cells immediately after TCR engagement. The immune system uses this mechanism to prevent hyperactivation of the immune response. The integration and overall balance of co-stimulatory and co-inhibitory signals dictate the extent of T cells' effector functions (Waldman *et al.*, 2020).

Signal 3 involves binding between APC-produced cytokines and their respective cytokines receptors on T cells. This signal dictates the differentiation phenotype of T cells, which is essential for a productive T cell mediated immune response. Interleukin 12 (IL-12) and Interferon alpha/beta (IFNα/β) are essential for $CD8^+$ T cells' clonal expansion and effector cytotoxic functions. For $CD4^+$ T cells, exposure to specific cytokines and specific transcription factors helps drive their differentiation into different effector subsets(Kalia and Sarkar, 2018; Ruterbusch *et al.*, 2020). Precise integration of all signals is required for an appropriate and effective T cell-mediated immune response.

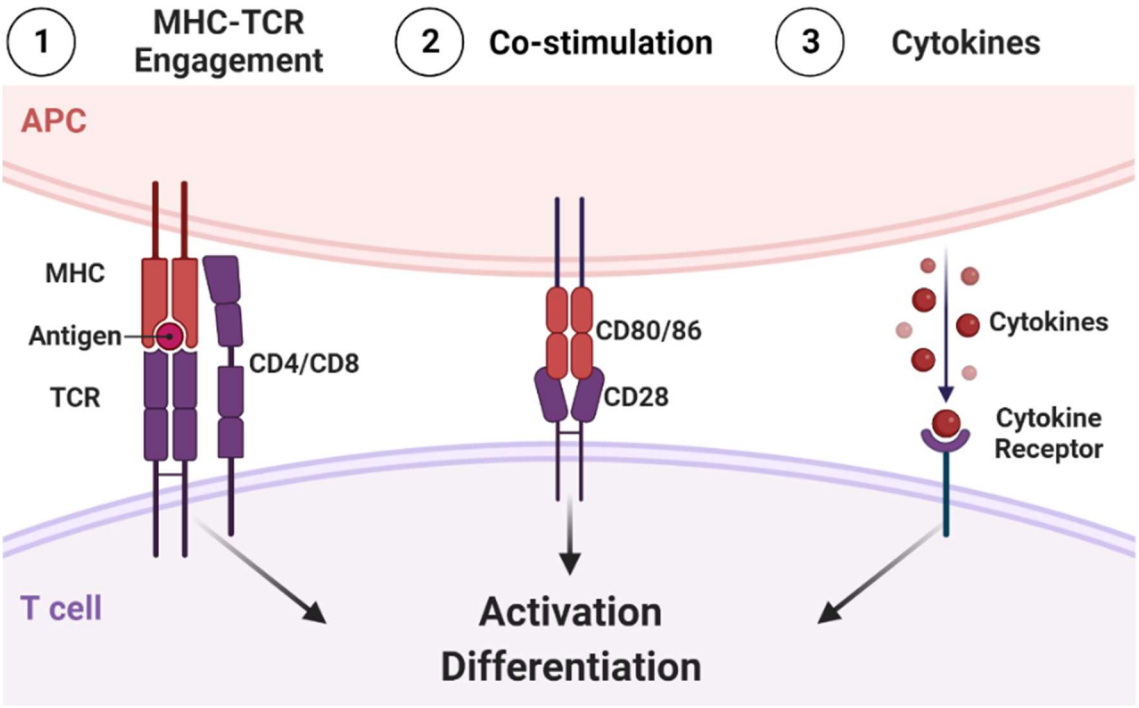


Figure 1.10: T cell activation: 3 signals activation model

Activation of effector T cells is orchestrated by 3 signals. Signal 1 comes from recognition and binding between TCR and antigen peptide bound MHC (also called human leukocyte antigen (HLA) in human). Engagement between CD80/86 and CD28 provides co-stimulatory signal 2. Cytokines (such as IL-12 and IFNα/β) provide signal 3, which influence T cell activation, clonal expansion, and effector cell differentiation.

1.2.4.3 Tumour evasion from T cell mediated immunity

Tumour cells target all parts of the T cell activation pathway to facilitate cancer progression and immune evasion. Cytotoxic $CD8^+$ T cells and $CD4^+$ T cells recognise malignant cells by the tumour-specific or tumour-associated antigen peptides presented on MHC class I/II, respectively. A common immune evasion strategy employed by tumour cells is by down-regulating the expression of antigen presentation pathway members. Across different clinical cohorts, protein expression of MHC class I molecules (measured by IHC) is suppressed in 30% of ccRCC patients' tumour samples (Dhatchinamoorthy *et al.*, 2021). Protein expression of TAP (Transporter associated with antigen processing), the endoplasmic reticulum (ER) transporter protein responsible for loading antigen peptide into ER, is also found to be down-regulated (by IHC) in ccRCC tumour samples compared to normal tissue (Seliger *et al.*, 2003). The loss of antigen presentation pathway in ccRCC is driven by the hyperactivated and overexpressed HIF pathway, particularly HIF-2A. Analysis of the TCGA KIRC dataset shows that mRNA levels of MHC class I (*HLA-A*, *HLA-B*) and MHC class II (*HLA-DMB*, *HLA-DQB2*) negatively correlate with *HIF2A* mRNA expression (Weinstein *et al.*, 2013). A combined proteo-transcriptomic study further confirmed that in the mouse ccRCC model, *VHL/HIF2A* deletion results in enhanced expression of MHC molecules and antigen presentation pathway members (Hoefflin *et al.*, 2020).

Tumour infiltrating T cells are frequently found with impaired anti-tumour effector functions. Many of these T cells enter a state of anergy and fail to create an immune response against tumour cells due to sub-optimal stimulation. This is actively promoted by tumour cells and the TME (Barnet *et al.*, 2018). For example, Programmed death ligand-1 (PD-L1) is a suppressor of T cells that is frequently amplified and overexpressed in many tumours (Escors *et al.*, 2018). Engagement between PD-L1 and immune check point PD-1 on T cells disrupts TCR signal transduction and inhibits functional T cell activation (Mizuno *et al.*, 2019). Expression levels of PD-L1 in many types of solid tumours, including ccRCC, have been shown to correlate with unfavourable prognosis

(Wang *et al.*, 2017; Ueda *et al.*, 2018; M. H. Kim *et al.*, 2021). The significance of the PD1/PD-L1 axis has been well-demonstrated experimentally. For example, cytotoxic $CD8^+$ T cells are found to specifically kill MC38 colorectal carcinoma cells with PD-L1 knocked-out genetically, but not in the PD-L1 expressing wild-type cells (Juneja *et al.*, 2017).

Another form of the dysfunctional state of T cells is known as T cell exhaustion. In the TME, T cells are constantly exposed to tumour antigens. Persistent antigen exposure and activation lead to up-regulation in the expression of inhibitory receptors such as PD-1, CTLA4, T cell immunoreceptor with Ig and ITIM domains (TIGIT), T cell immunoglobulin and mucin domain containing-3 (TIM3) and lymphocyte-activation gene 3 (LAG-3). Exhausted T cells (T_{ex}) also display reduced cytotoxicity and impaired cytokines expression, including IFNγ, TNF and IL-2 (Wherry and Kurachi, 2015).

In an immunosuppressive TME, tumour cells and stromal and immune cells can secrete cytokines and other soluble factors that induce T cell dysfunction. For example, in ccRCC, tumour cells often express high levels of transforming growth factor beta (TGF-β) (Boguslawska *et al.*, 2019). TGF-β mediates suppression of T cell mediated anti-tumour immunity in a multitude of ways. Firstly, TGF-β inhibits the differentiation of naïve T cells into CTLs and Th cells (Oh and Li, 2013). TGF-β can also dampen anti-tumour immunity by up-regulating PD-1 and CTLA-4 whilst suppressing granzyme and IFNγ expression in CTLs (Bao *et al.*, 2021). Finally, TGF-β promotes the differentiation of regulatory T cells (T_{reg}), as well as the polarisation of non-activated macrophages into a pro-tumour, M2 tumour-associated macrophage (TAM) phenotype (Zhang *et al.*, 2016; Liu *et al.*, 2018). Along with tumour cells, T_{reg} and M2 TAM are the other significant sources of immunosuppressive cytokines (such as TGF-β, IL-10, and IL35) and consequently contribute to promoting dysfunctional T cell phenotypes in the TME (Xia *et al.*, 2019).

Recent studies have revealed that instead of a distinct, terminally exhausted T cell population, T cell exhaustion is better represented as a gradient of dysfunctional states (Zheng *et al.*, 2021; Budimir *et al.*, 2022). Recent research has discovered that a subset of T_{ex} populations (progenitor exhausted T cells) can reverse its suppressed effector functions when the immunosuppressive pathways (i.e. PD-1/PD-L1 interaction) are blocked (McCaw *et al.*, 2019; Tabana *et al.*, 2021). Crucially, progenitor T_{ex} is a long-lived and highly proliferative cell population (Im *et al.*, 2016). Reversal of T cell exhaustion is now a major focus of immunotherapy research.

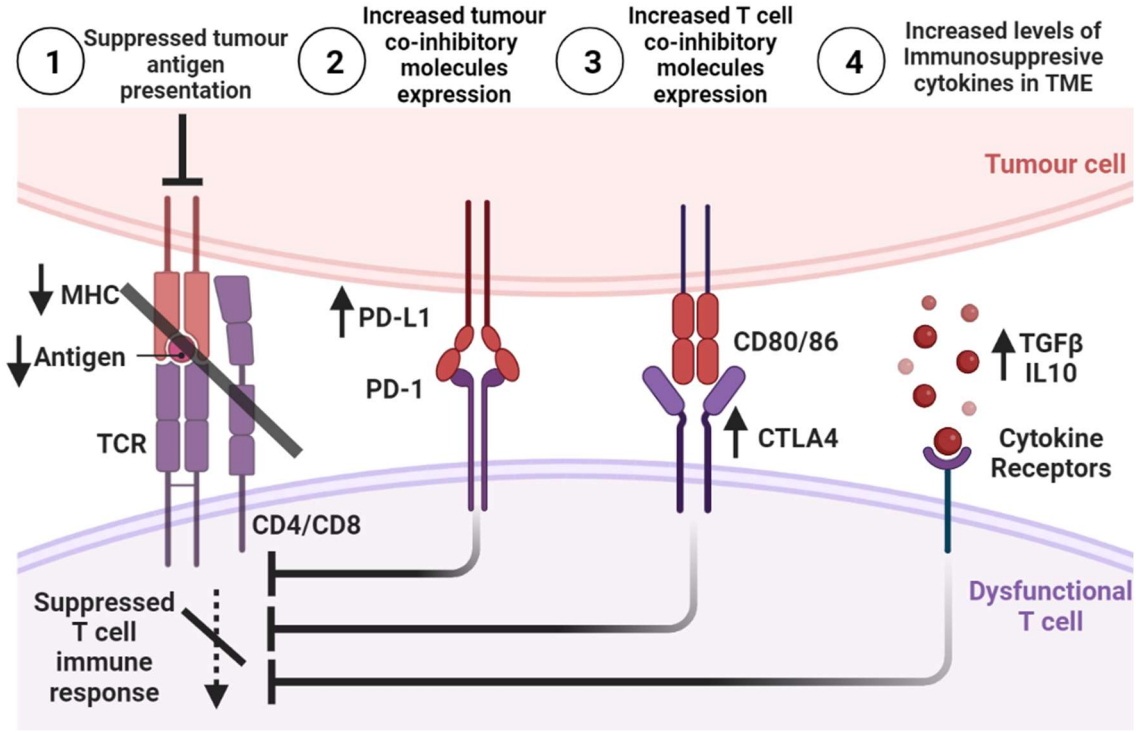


Figure 1.11: Tumour evasion from T cell immunity

Tumour cells evade from T cell mediated anti-tumour immunity in a myriad of ways. Firstly, tumour cells often down regulate expression of antigen presentation pathway members to avoid detection. Tumour cells are also found to up-regulate expression levels of co-inhibitory molecules (such as PD-L1). Persistent activations and exposure to immunosuppressive cytokines (secreted by tumour cells and immunosuppressive immune cell types) also contribute to dysfunctional T cell effector functions.

1.2.4.4 TME: stromal cells

During tumourigenesis and subsequent stages of tumour progression, tumour cells secrete a wide range of factors (such as growth factors, cytokines, and chemokines) to recruit and transform stromal cells from the resident or nearby tissues. The predominant component of the stromal cells is cancer associated fibroblast (CAF). Fibroblasts are best known for supporting wound healing via chemotaxis-facilitated migration towards the wound area, followed by activation (Foster *et al.*, 2018). Gene expression and secretion of many key chemo-attractant molecules (such as PDGF and TGF-β) are aberrantly up-regulated in different cancer types, including ccRCC (W. Wang *et al.*, 2015; Zhan *et al.*, 2020). Recent single-cell transcriptomic studies have revealed CAFs as a highly heterogeneous cell population with multiple subtypes displaying distinct phenotypes (Richards *et al.*, 2016; Costa *et al.*, 2018; Elyada *et al.*, 2019). However, there is currently a lack of reliable biomarkers to distinguish CAF populations across cancer types. Whilst both tumour-promoting and tumour-suppressing populations have been identified. Thus far, the consensus has been that most CAFs facilitate tumour progression directly by promoting tumour cell growth or indirectly by remodelling the extracellular matrix and creating an immunosuppressive TME (Mao *et al.*, 2021).

Activation of CAFs has been experimentally validated to promote tumour progression in numerous ways. Firstly, activated CAFs are a significant contributor to soluble growth factor secretion in TME. Secretion of VEGF by CAF promotes tumour growth. Moreover, together with the high levels of VEGF expression from malignant ccRCC cells, they contribute to the typically high tumour vascularity (often described as a rich, fishnet-like vascular structure architecture) (T. Liu *et al.*, 2019). TGF-β secretion by CAF promotes activation of the TGF-β/SMAD pathway in tumour cells, which results in elevated expression of epithelial-mesenchymal transition (EMT) related genes with a more aggressive phenotype (Yu *et al.*, 2013; Zhuang *et al.*, 2015). Recent studies have also shown that CAF-derived exosomes (CDE) promote tumour progression by delivering proteins and non-coding RNAs (ncRNA) to neighbouring tumour cells. This results in the

augmentation of tumour cells' metabolic pathways (from oxidative phosphorylation to glycolysis), increased tumour cell motility (activating Notch signalling pathway by delivering the metalloprotease ADAM10), and increased tumour proliferation (by delivering Sonic Hedge Hog (SHH) protein) (Shimoda *et al.*, 2014; Zhao *et al.*, 2016; G. Zhao *et al.*, 2020). Finally, CAFs contribute to creating an immunosuppressive TME by secreting immunomodulatory factors (such as TGF-β and IL-10) and remodelling the tumour ECM (Cohen *et al.*, 2017; Monteran and Erez, 2019).

1.2.4.5 TME & extra-cellular matrix

The extracellular matrix is the critical non-cell constituent of the TME. ECM consists of a network of crosslinked fibrous proteins and proteoglycans. In cancer, tumour cells and CAFs remodel ECM to promote proliferation, suppress anti-tumour immunity and induce angiogenesis (Pickup *et al.*, 2014). For example, ccRCC tumour cells induce the expression of an ECM protein Perostin from the stromal CAFs *in vivo* (Bakhtyar *et al.*, 2013). Periostin forms an integral part of ECM in ccRCC tumours and is highly upregulated in ccRCC tumours compared to adjacent normal tissues (Bond *et al.*, 2021). Periostin promotes cell proliferation by activating ccRCC tumour cell surface integrin-linked kinase (ILK), thereby triggering downstream AKT/mTOR signalling cascade(Jia *et al.*, 2021). Periostin is also shown to promote EMT and metastasis of tumour cells through integrin-binding mediated pathways (Morra and Moch, 2011). Chitinase 3-like 1 (Chi3L1) is another ECM component with pro-tumour effects. Secreted by tumour cells, CAFs, and immune cells, Chi3L1 expression is a marker of poor prognosis, including in ccRCC(Libreros *et al.*, 2013). Secretion of Chi3L1 favours recruitment of pro-tumour M2 TAM but inhibits infiltration of anti-tumour CTLs (Cohen *et al.*, 2017). Moreover, Chi3L1 expression has been shown to promote tumour progression by activating multiple signalling pathways in tumour cells, such as AKT and TGF-β (Qiu *et al.*, 2018;T. Zhao *et al.*, 2020).

1.2.5 ccRCC treatment

Although ccRCC is one of the most common types of cancer with distinct genetic, metabolic and IHC markers, there is currently no public screening implemented across the globe. Most ccRCC cases are discovered incidentally from magnetic resonance imaging (MRI) or computerised tomography (CT) scan (Escudier *et al.*, 2019). After confirming the diagnosis using IHC analysis on biopsy samples, the first-line treatment for localised ccRCC is typically performed by surgical kidney removal (nephrectomy). Depending on the clinical stage of the cancer, partial nephrectomy or radical nephrectomy (removal of the whole kidney) is performed to completely remove ccRCC tumour tissue (Atkins and Tannir, 2018). Nephrectomy is currently the most effective method to treat localised ccRCC tumours. Across different studies, the 5-year overall survival rate for early-stage ccRCC patients post-nephrectomy is higher than 80% (Janssen *et al.*, 2018). However, 20 to 50% of post-nephrectomy ccRCC patients experience disease relapse within five years after the surgery (Capogrosso *et al.*, 2016). Recurrent ccRCC is associated with extremely poor disease outcomes, with less than 30% of patients surviving two years after diagnosis (S. H. Kim *et al.*, 2021).

Whilst early-stage ccRCC can often be treated by nephrectomy alone, and metastasised ccRCC requires systemic treatment. Metastasised ccRCC is highly aggressive. Patients with untreated metastases RCC had a poor 5-year survival rate of 2.7 – 9% (Négrier *et al.*, 2002). ccRCC is insensitive to both radiation and chemotherapy strategies. Until the mid-2000s, the only option to treat advanced ccRCC was by administering human IFN-α and IL-2 cytokines to induce anti-tumour immunity. However, the objective response rate (ORR) to cytokine therapy was poor at ~10% across trials (Ritchie *et al.*, 1999; Atzpodien *et al.*, 2002). In addition, the high dosage of IFN-α and IL-2 needed for enhanced immune response can also cause severe adverse events in up to 50% of patients (Huang and Hsieh, 2020). With the poor treatment outcome from cytokines treatment, there was an urgent need for better and more targeted therapeutic approaches for metastatic ccRCC.

The first targeted therapeutic approach utilises anti-angiogenic tyrosine kinase inhibitors (TKIs). Since its first introduction in 2005, eight different TKIs have been approved by the U.S. food and drug administration (FDA). Sunitinib, an inhibitor of VEGF and PDGF receptors, was found to improve metastatic ccRCC patients' ORR and median overall survival from 10% & 14 months to 30% and 30 months when compared to the cytokines treatment approach (Schmid and Gore, 2016). *In vitro* studies have shown that Sunitinib inhibits tumour cell growth via the inactivation of AKT/mTOR pathways downstream from RTKs (Hudes, 2009). However, *in vivo* studies have suggested that under pharmacologically relevant concentration, the tumour-suppressing ability can be primarily attributed to the suppression of endothelial cell proliferation in the TME (Huang *et al.*, 2010). This demonstrates the clinical importance of non-malignant cells in the TME. Sunitinib is now used as a first-line treatment for metastatic ccRCC worldwide and in the UK (Fontes-Sousa *et al.*, 2022).

Following the success of TKIs, the arrival of immune checkpoint inhibitor (ICI) therapy in the last decade has also fundamentally changed how metastasised ccRCC patients are treated. Tumour infiltrating T cells' effector functions are tightly regulated by co-inhibitory receptor signalling pathways. Human monoclonal antibodies against PD-1 and CTLA4 block the co-inhibitory receptor-ligand interactions, thereby boosting the anti-tumour immunity by tumour infiltrating T cells (Robert, 2020). FDA has approved various ICI, including nivolumab (anti-PD-1 antibody) and ipilimumab (anti-CTLA4 antibody), as the first-line treatment for metastatic ccRCC (Sheng and Ornstein, 2020). Combinatorial treatment of nivolumab and ipilimumab was found to outperform sunitinib monotherapy (ORR 42% vs 27%) (Hammers *et al.*, 2017). Multiple ongoing clinical trials used combinations of ICIs, TKIs, and ICI + TKI, which have shown varying degrees of efficacy, with ORR ranging from 30 - 60% (Rassy *et al.*, 2020). Whilst this is a dramatic improvement from cytokine treatments, no reliable biomarkers can predict metastasised ccRCC responsiveness to ICI or TKI treatments.

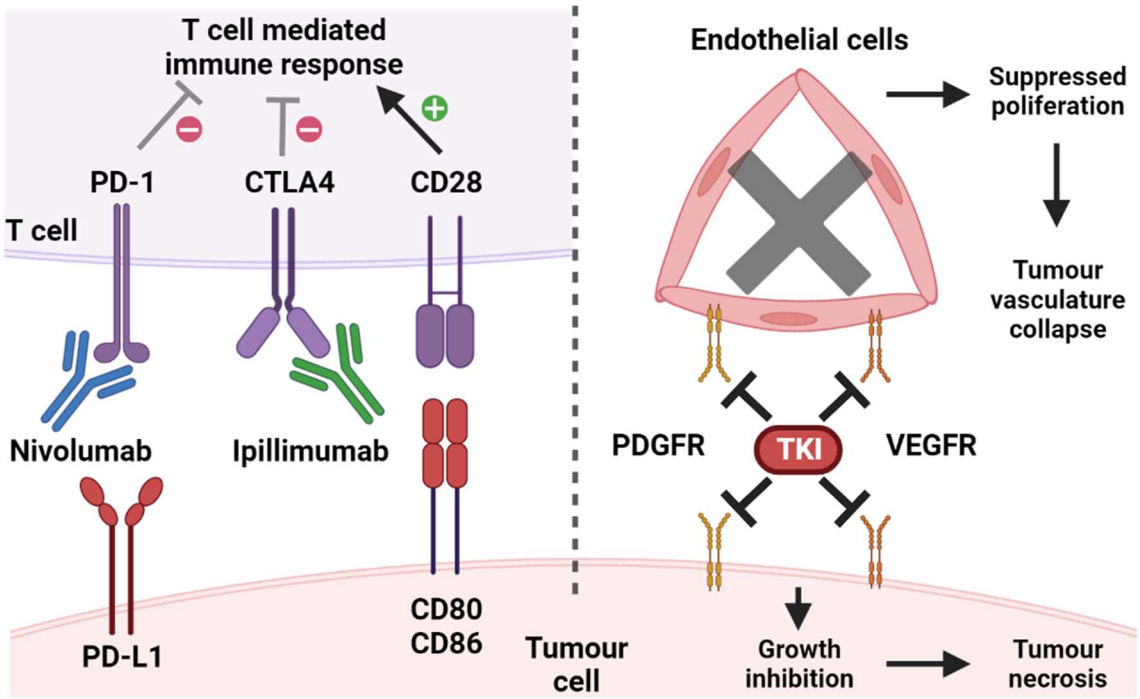


Figure 1.12: Mechanisms of action of Immune checkpoint inhibitor (ICI) and tyrosine kinase inhibitor (TKI) therapy against metastatic ccRCC

ICIs such as Nivolumab and Ipillimumab are synthetic monoclonal antibodies that targets Tumour infiltrating T cells co-inhibitors receptors (PD-1 & CTLA-4 respectively) and block their interactions with co-inhibitory ligands expressed on tumour cells. Successful ICI treatment promotes patients' T cells activation and improve anti-tumour immune response. TKIs are small molecules that inhibit kinase activities of RTKs, such as PDGFRs and VEGFRs. By blocking RTK and downstream signalling pathways, successful TKI therapy can suppress cell growth from both tumour cells and endothelial cells in the TME, which leads to tumour regression.

1.3 RNA sequencing technologies

1.3.1 Next-generation sequencing

Since the late 2000s, various high-throughput RNA sequencing platforms (RNA-seq) have been developed and enable global gene expression profiling (Wang *et al.*, 2009). The application of RNA-seq has provided invaluable insights into RNA biology and a greater understanding of disease development and treatments. With the advancement of RNA-seq technologies, it is now possible to investigate gene expression with single-molecule resolution at the single-cell level.

The vast majority of published transcriptomic studies use next-generation sequencing (NGS) technologies, particularly the Illumina sequencing platform (Stark *et al.*, 2019). The general workflow of Illumina sequencing begins with the generation of complementary DNA (cDNA) libraries, which involves RNA extraction, mRNA enrichment (by oligo(dT) capture or depletion of ribosomal RNA), RNA fragmentation (to under 200 nt), reverse transcription and polymerase chain reaction (PCR) amplification (Figure 1.13). cDNA libraries are subsequently loaded onto a flow cell on an Illumina sequencer. Within the flow cell, cDNA fragments are clustered, further amplified and sequenced by a process called sequencing by synthesis. This is achieved by incorporating fluorescent-tagged nucleotides into the growing DNA strands complementary to cDNA molecules, which allows base-to-base signal detection (Goodwin *et al.*, 2016). Illumina sequencing platform allows high-throughput and cost-effective transcriptome-wide gene expression profiling. However, Illumina RNA sequencing technologies have several technical limitations, the notable being the short length of cDNA molecules compared to the average length of mRNAs (Shi *et al.*, 2021).

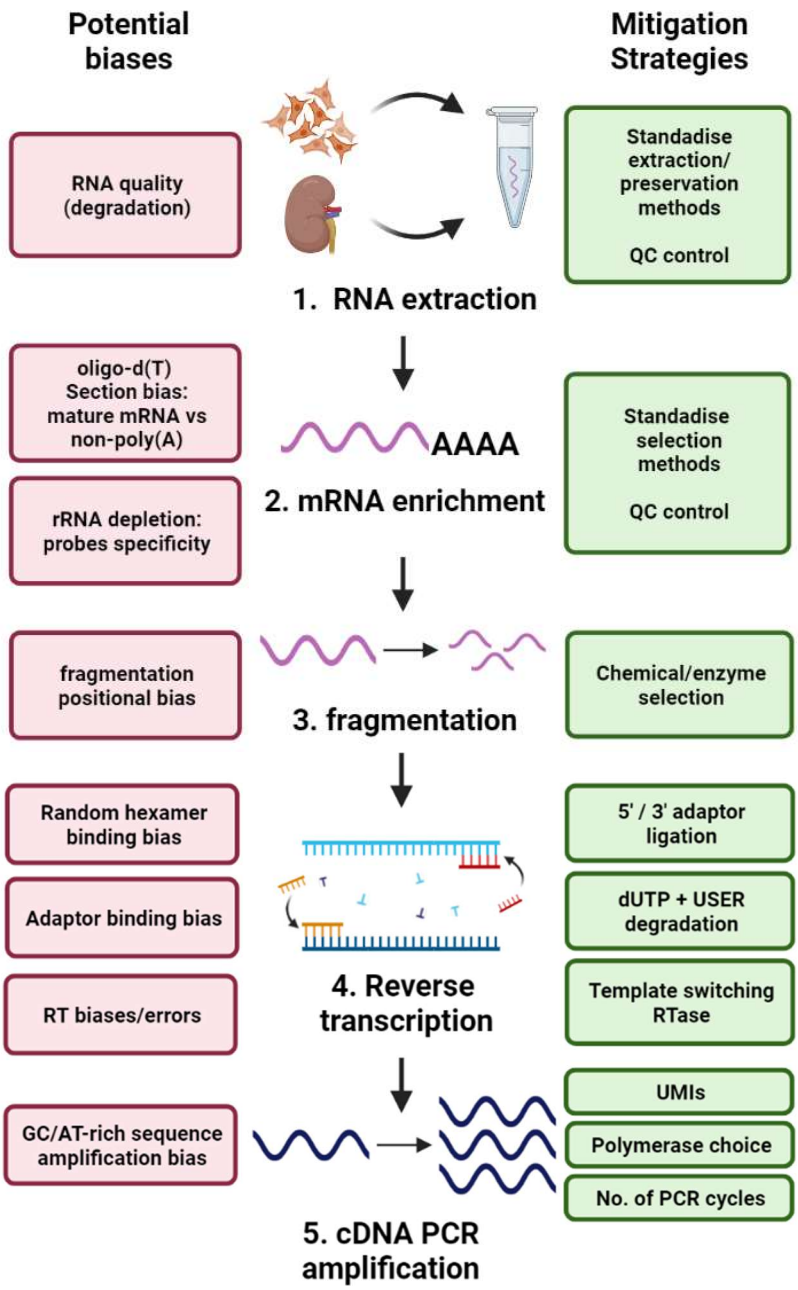


Figure 1.13: Illumina short-read cDNA library construction workflow

Illumina cDNA library generation begins with the extraction of RNA, followed by enrichment of mRNA via either oligo-d(T) pulldown, or depletion of ribosomal RNA (rRNA). RNA molecules are fragmented chemically or enzymatically, and reverse transcribed into cDNAs. cDNA molecules are then amplified via PCR amplification. Each step of the cDNA library construction may introduce biases (red boxes), and mitigation strategies (green) are needed to provide the true representation of the transcriptome.

1.3.2 Nanopore long-read RNA sequencing

To overcome the technological challenges faced by Illumina RNAseq technologies, there are currently two major alternative RNAseq platforms: Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio), each allowing full-length sequencing of native RNA molecules. ONT and PacBio have also developed full-length cDNA library preparation and sequencing pipelines (ONT PCR-cDNAseq and Iso-seq, respectively). ONT developed a sequencing method that directly detects strand-specific, full-length RNA molecules (Direct RNA Seq, or DRS) or cDNA molecules without PCR amplification (Cartolano *et al.*, 2016; Garalde *et al.*, 2018; Grünberger *et al.*, 2022).

Nanopore sequencing depends on membrane-embedded pore-forming proteins (nanopores) and their associated helicase motor proteins. ONT sequencing devices use sequencing flow cells, which contain arrays of nanopores inserted into a polymer membrane connected to a current sensor chip. A constant voltage is applied to the flow cells and their nanopore arrays during the sequencing. When the motor protein recruits and facilitates DNA or RNA molecules to pass through the protein pore, the DNA/RNA molecule disrupts the current. The amplitude of the current disruption and the dwell time of the molecule were found to be characteristic of specific nucleotides. Detection of the current disruptions by the sensor chip is computationally analysed concurrently, which allows accurate single-molecule cDNA and RNA sequencing in real-time (Figure 1.14) (Garalde *et al.*, 2018). ONT also developed a cDNA-based sequencing method that allows the generation of long reads at higher levels of throughput, using a lower amount of input RNA than DRS (Figure 1.15).

Long-read sequencing provides several key advantages over short-read cDNA sequencing. Firstly, the lack of a PCR amplification step from ONT DRS eliminates inherent amplification biases. In addition, Illumina sequencing requires fragmentation of input RNA (or cDNA after RT, depending on the library construction protocol). The latest Illumina RNAseq flow cell (NovaSeq 6000) allows a maximum read length of 150 bp, with the recommended read length standing at 75 bp long (Corchete *et al.*, 2020). Most

human transcriptomic studies using the Illumina platform aim to generate tens to hundreds of million reads per sample, which provide sufficient sequencing depth and coverage for most expressed genes in human (Sheng *et al.*, 2017). However, the length of Illumina reads is far shorter than the average length of mRNA transcripts found in human (~2,100 nt) (Lopes *et al.*, 2021). Thus, whilst short-reads sequencing can reveal gene-level differential expression between samples, their reads cannot span multiple exons and provide isoform-level information. In contrast, the median read lengths generated by ONT DRS typically reach 1000nt, where many of these reads capture mRNA transcripts from end to end (Gleeson *et al.*, 2022). This allows researchers to accurately display complex multi-exonic architecture and dissect transcript isoform expression dynamics without relying on computational isoform modelling for Illumina sequencing.

Another significant advantage of using long-read sequencing technologies over Illumina is its ability to discover novel transcripts. RNAseq gene and isoform expression analysis relies on mapping generated sequence reads to a reference genome or transcriptome. In human and other frequently studied model organisms, reference genome/transcriptomes are assembled using high-throughput short-read sequencing data and curated by databases such as Ensembl, RefSeq, and Gencode (O'Leary *et al.*, 2016; Frankish *et al.*, 2021; Cunningham *et al.*, 2022). Recent long-read sequencing studies by both ONT DRS and PacBio Iso-seq reveal that the human transcriptome is more heterogenous than previously thought, with 30-50% of identified mRNA transcripts being novel and previously unannotated from reference databases (Soneson *et al.*, 2019; Workman *et al.*, 2019; Leung *et al.*, 2021).

Finally, ONT DRS allows concurrent profiling of poly(A) tail lengths and chemical modifications on mRNA molecules. Methods to detect co/post-transcriptional regulatory events in mRNA molecules are outlined and described in the following parts of the chapter.

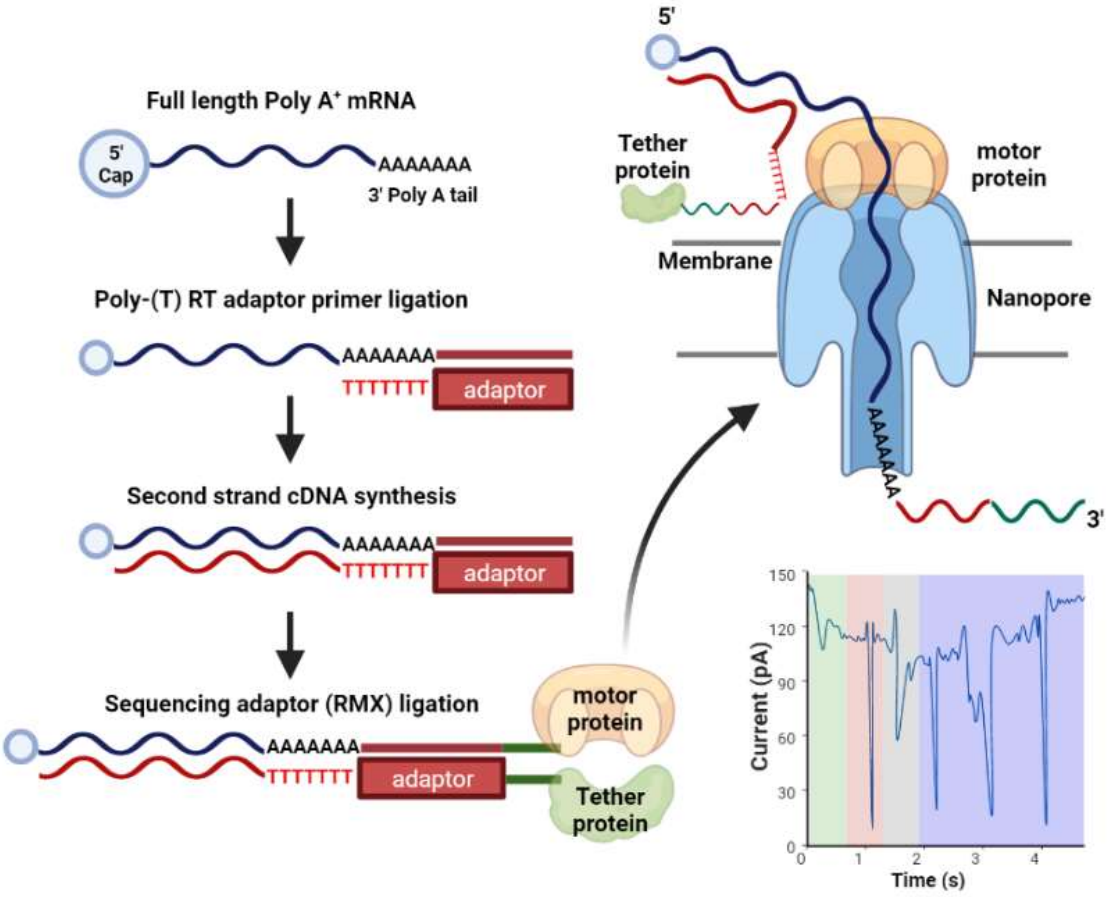


Figure 1.14: ONT direct RNA sequencing (DRS) technology

ONT DRS utilises Poly $(A)^+$ mRNA as input. RNA molecules are first ligated with Poly(T) reverse transcription (RT) adaptor primers, which contain an adaptor sequence at its 5' end. Following the ligation, second strand cDNA synthesis is performed and sequencing adaptors (RMX), consist of motor proteins and tether proteins, are ligated to the RNA: DNA duplex (at the 3' end of RNA and 5' end of cDNA adaptor sequence). Motor protein facilitates unwinding of the duplex and feeds the RNA strand through the nanopore at 90 nucleotides per second. Tether protein keeps cDNA strand away from the pore whilst the motor helicase unwinds the duplex. The pore is embedded on a membrane where electric potential is applied. Different physio-chemical properties of the nucleotides in the pore results in distinctive changes in current signals signatures (also known as squiggles), which can then be used to infer their identities (Green: leader sequence, Red: adaptor sequence, Grey: poly(A) sequence, Blue: mRNA body).

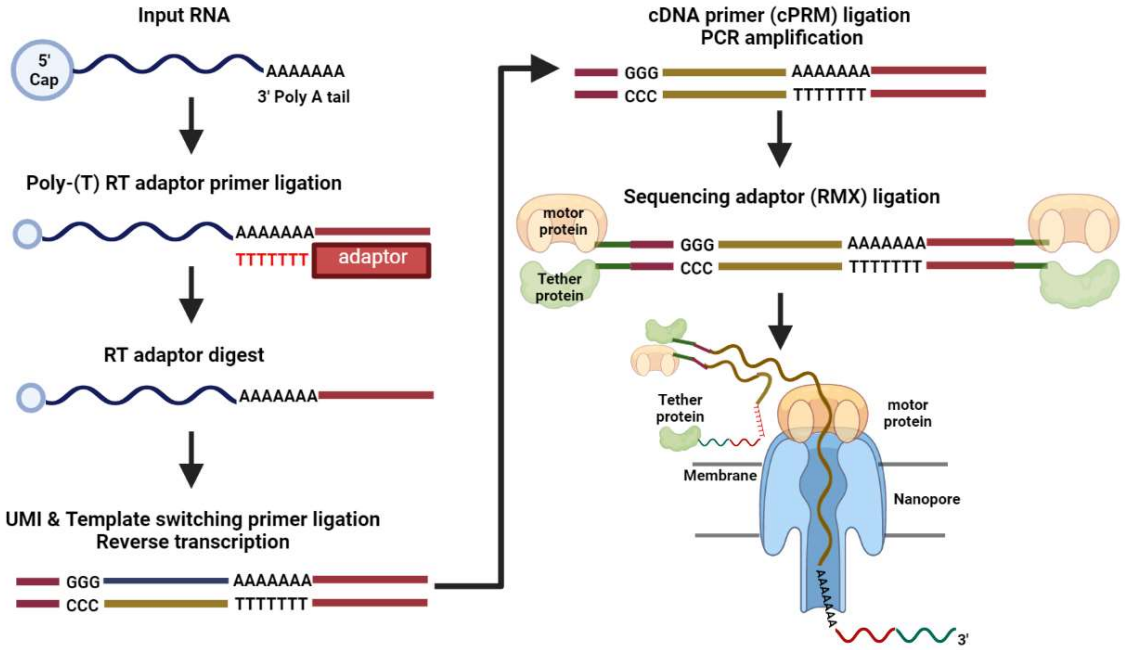


Figure 1.15: ONT PCR-cDNAseq cDNA library generation (PCS111)

poly(A)$^+$ mRNA molecules are enriched by poly(T) reverse transcription (RT) adaptor primers. The adaptors are digested, followed by ligation of unique molecular identifier (UMI) and template switching primers. UMI allows elimination of PCR duplicates and reduce amplification biases. Reverse transcription of mRNA molecules to cDNA is carried out by template switching reverse transcriptase (Maxima H minus) to reduce RT artefacts. cDNA primers (cPRM) are subsequently ligated and facilitates PCR amplification (between 12 – 14 cycles). cDNA duplexes are finally ligated with sequencing adaptors (RMXs) containing motor proteins and tether proteins. cDNA duplexes are unwind by motor protein and thread through nanopore in the sequencer. One strand of cDNA is sequenced at a time. Reads are reorientated, with UMI and other primer sequences removed after the sequencing run.

1.3.3 Detection of alternative splicing and polyadenylation events by RNAseq

Alternative splicing and alternative polyadenylation events contribute to generating multiple transcript isoforms from a single gene. RNAseq analysis is now the gold standard for detecting differential alternative splicing and alternative polyadenylation events between biological conditions. Nevertheless, accurate detection of these events by RNAseq continues to be a challenging bioinformatic problem.

After sequencing runs, the generated reads are typically processed (i.e. removal of any primer sequences and sequence orientation) and aligned to the reference genome or reference transcriptome. Alignment to the reference genome provides gene-level quantification, whilst reference transcriptome alignment allows isoform-level mapping, which can be aggregated subsequently for gene-level expression quantification (Conesa *et al.*, 2016). Different transcript isoforms display structural variations (exon inclusion/exclusion, 3'UTR lengths), and isoform-specific features permit reads to be uniquely mapped. However, most sequencing reads, especially from the Illumina RNAseq platform, only span a short fragment of the isoforms, with the sequence location potentially shared by multiple variants. Moreover, library preparation steps can introduce biases and impact sequence coverage, creating a false picture of alternative splicing and alternative polyadenylation events (Buen Abad Najar *et al.*, 2020). The most used transcript abundance quantifiers, such as Salmon, automatically correct expression levels from sequence and library preparation biases (Patro *et al.*, 2017a). Nevertheless, using long sequencing reads spanning multiple exons remains the most effective way to provide accurate isoform identification.

Many bioinformatic tools have been developed to identify differential isoform usage specifically. For example, DRIMseq is a statistical framework that detects differential isoform usage between biological conditions. In principle, DRIMseq normalises the sum of all isoforms within a sample to 1 and detects if the proportion of each isoform changes significantly between experimental conditions (Robinson and Nowicka, 2016).

Alternatively, DEXseq compares the usage of individual exon within a gene and report potential alternative splicing events independent of isoform assignments (Anders *et al.*, 2012). Both methods are recently integrated into a single pipeline with Salmon and provide p values for differential isoform usage (Love *et al.*, 2018).

DEXseq and DRIMseq rely on the sequence information provided by the reference transcriptome. An alternative approach is therefore needed to discover novel, unannotated isoforms generated by alternative splicing and polyadenylation. This is represented by the Cufflinks pipeline, which relies on *de novo* reconstruction of the transcriptome using sequencing reads generated from experiments (Trapnell *et al.*, 2012). By overlaying reads onto the genome and producing an overlap graph, transcript isoforms are inferred and constructed from the minimum number of 'paths' required to cover all read fragments. Cufflinks assembly can be both guided and unguided by the reference transcriptome. Subsequent mapping of reads and expression analysis can provide information on potential alternative splicing and polyadenylation events between samples. However, the number of novel isoforms using Illumina reads can vary greatly depending on the often arbitrarily chosen stringency and confidence threshold used for Cufflinks assembly. This creates questions regarding the confidence in Cufflinks' performance and if the reconstructed isoforms can be realistically validated (Angelini *et al.*, 2014). Compared to Illumina reads, ONT and PacBio long sequencing reads allow a better quality of transcriptome reconstruction. There are now computational methods available (For example, *SQANTI2*, *FLAME*, *FLAIR*) that integrate long-read sequencing with global maps of transcription start sites (TSS) and polyadenylation sites to identify novel transcript isoforms(Tang *et al.*, 2020; Holmqvist *et al.*, 2021; Leung *et al.*, 2021). These methods have proved to be transformative in novel isoforms characterisation, and their application will provide a far more comprehensive view of alternative splicing and polyadenylation events in the human transcriptome.

1.3.4 Measurement of poly(A) tail length by RNAseq

Poly(A) tail plays a crucial role in mRNA stability, and its length is dynamically controlled throughout the lifetime of an mRNA molecule. Before the development of RNAseq technologies, there were limited ways to determine poly(A) tail lengths. mRNA molecules are treated with and without oligo(dT), followed by RNase H degradation. The hybridisation of oligo(dT) shields mRNA polyA tails from RNase H. Thus, differences in mobility between degraded and non-degraded poly(A) tailed mRNA assessed by northern blotting provide an estimation of the length of poly(A) tails (Murray and Schoenberg, 2008).

Since the mid-2010s, several Illumina RNAseq-based methods, such as TAIL-seq, have been developed to assess global mRNA poly(A) tail lengths. TAIL-seq uses ribodepleted RNA as input. After ligation of a 3' biotinylated DNA adapter sequence, the RNA molecules are partially degraded by Rnase T1, which cleaves after G residues. Intact poly(A) tails fragments are enriched by streptavidin beads pull-down, followed by ligation of a 5' adapter to provide sequence template for reverse transcription. The poly(A) tails are reverse transcribed, amplified by PCR and sequenced using Illumina RNAseq platforms (Chang *et al.*, 2014). However, Illumina sequencing struggles with the accurate detection of homopolymers, such as the poly(A) tail. Instead, Chang *et al.* modified the Illumina sequencer's processing software to extract and analyse raw images of fluorescent-tagged nucleotide incorporation. Using a trained hidden Markov model, the strength of fluorescent signals is used to infer the poly(A) tail lengths. These methods are hugely expensive and technically challenging.

Finally, ONT DRS can provide poly(A) tail length estimation. By reanalysing the raw current data and the dwell time of mRNA transcripts, the length of each transcript's poly(A) tail can be estimated using nanopolish and tailfindR (Krause *et al.*, 2019; Workman *et al.*, 2019). These methods are computationally demanding but offer a unique opportunity to integrate information on poly(A) tail, transcript isoform, and RNA modifications at a single mRNA molecule level.

1.3.5 Detection of mRNA m^{6}A modification

1.3.5.1 Antibody-based enrichment of modified RNAs

Currently, the most widely used techniques for studying RNA modifications are based on methylated RNA-immunoprecipitation (MeRIP), which uses RNA modification targeting antibodies. In 2012, separate groups described methods (MeRIP-Seq or m^{6}A-seq) using anti-m^{6}A antibody to immunoprecipitate ~200bp long m^{6}A enriched mRNA fragment, followed by Illumina sequencing to produce a transcriptome-wide mapping of m^{6}A in human and mouse (Dominissini *et al.*, 2012; Meyer *et al.*, 2012). These studies were the first transcriptome-wide studies of m^{6}A and have revolutionised the study of epitranscriptome. Variations of MeRIP-seq, such as miCLIP-seq, have also been developed to map m^{6}A at single nucleotide resolution. However, due to the lower RNA input requirement (miCLIP studies routinely use 5-20 µg poly(a)$^+$ RNA whereas the current MeRIP-seq protocol can detect m^{6}A peaks well with 2 µg total RNA) and complicated workflow, to date, MeRIP-seq is still the most used method to map m^{6}A.

MeRIP-seq studies typically use either poly (A) enrichment or rRNA-depleted RNA as input (Figure 1.16). Alternatively, total RNA can be used, followed by ribosomal cDNA removal using probes specific to rRNA post reverse transcription and cDNA amplification (SMARTer Stranded Total RNA-Seq Kit, Takara). RNA samples are first fragmented chemically to a size distribution centred at approximately 200nt, followed by immunoprecipitation incubation with anti-m^{6}A antibodies, washes and elution (Zeng *et al.*, 2018). Eluted m^{6}A antibody-bound RNA and a paired input control sample of fragmented mRNA pre-immunoprecipitation are subsequently used to synthesise cDNA libraries and sequenced, typically via Illumina next-generation technology. Detection of m^{6}A methylation peaks is calibrated with paired input control to take account of transcript abundance. Various peak-callers, including MACS2, exomePEAK, MeTPEAK and MoAIMS (Zhang *et al.*, 2008; Cui *et al.*, 2016, 2018; Zhang and Hamada, 2020) have been used to identify m^{6}A peaks from MeRIP-seq data, although with divergent results in terms of the number and locations of statistically-confident unique m^{6}A peaks. Initially

developed for ChIP-seq analysis, MACS2 is currently the most commonly used m^{6}A MeRIP peak caller.

MeRIP-seq has expanded the understanding of m^{6}A in transcriptome substantially. However, there are several limitations. MeRIP-seq performances rely largely on the specificity of the antibody used. It has recently been shown that SySy m^{6}A antibodies cross-react with m^{6}Am, a prevalent modification in 30 – 40% of mRNA (Mauer and Jaffrey, 2018). The resolution of m^{6}A MeRIP-seq is also limited to the fragmentation size. This aspect can be improved by filtering results with the DRACH motif, m^{6}A prediction tools, and previously published datasets.

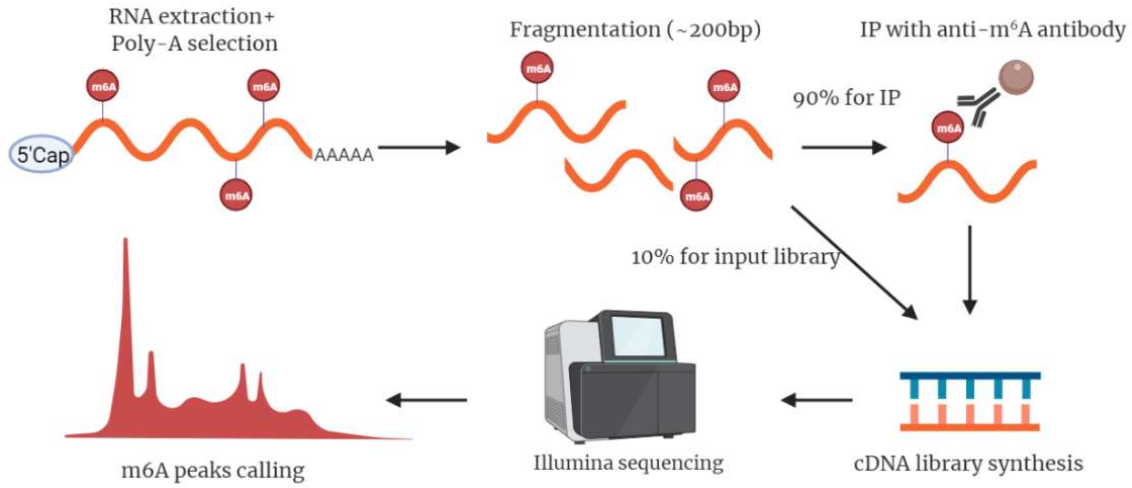


Figure 1.16: Graphical representation of MeRIP-seq workflow

For transcriptomic MeRIP-seq studies, total RNA is typically subjected to Poly-A$^+$ selection, followed by chemical fragmentation to about 200bp long fragments. Paired input control for each MeRIP reaction is kept before immunoprecipitation with anti-m^{6}A antibodies. cDNA libraries are generated using input control and IP eluate, followed by Illuimina next generation sequencing. m^{6}A peaks are identified based on MeRIP samples read frequencies with respect to transcript expression levels from paired input control.

Shortly after the development of MeRIP-seq, protocol was adapted to allow the detection of m^6A in the transcriptome at single nucleotide resolution, namely miCLIP-seq (Figure 1.17) (Linder *et al.*, 2015). For miCLIP-seq, anti-m^6A antibodies are crosslinked to RNA molecules using UV_{254}. Purification steps follow this via protein A/G binding, SDS-PAGE and membrane transfer. Finally, proteinase K digest releases RNA fragments from the bound antibody-protein A/G complex. The peptide fragments on the eluted RNA molecules induce either C->T mutations or truncations during reverse transcription. Interestingly, Linder *et al.* documented different mutation and truncation signatures (position) as well as strength (% conversion of mutation and truncation) for different commercially available antibodies (Figure 1.17).

miCLIP dramatically improves the resolution of m^6A site identification and is now regarded as the gold standard for m^6A site curation and identification. However, since miCLIP is still an antibody enrichment-based method, some drawbacks and limitations persist. Firstly problems concerning antibody specificity (i.e. SySy with m^6Am) can still lead to false positive m^6A signals. A study by Zeng *et al.* compared m^6A peaks using three different antibodies (SySy, NEB and Millipore) and found that only 60% of peaks overlap for all three antibodies (Zeng *et al.*, 2018). Comparing results with MeRIPseq, miCLIP-seq consistently reports a lower number of modified mRNAs (3500 vs 7000 modified sites) (Anreiter *et al.*, 2021). However, without an unbiased method that quantitatively identifies m6A modifications across the transcriptome, it is currently unclear if this disparity comes from a high false-positive rate from MeRIP-seq or a high false-negative/lack of sensitivity from the miCLIP-seq. Finally, whilst both MeRIP-seq and miCLIP-seq may be helpful to compare the number of peaks identified within samples in the same experiment (hence reflects m^6A methylation rate), neither provide precise information on m^6A stoichiometries (i.e. the percentage of m^6A methylated transcript), which is vital for understanding how and to what extent mRNA m^6A modification regulates gene and protein expression.

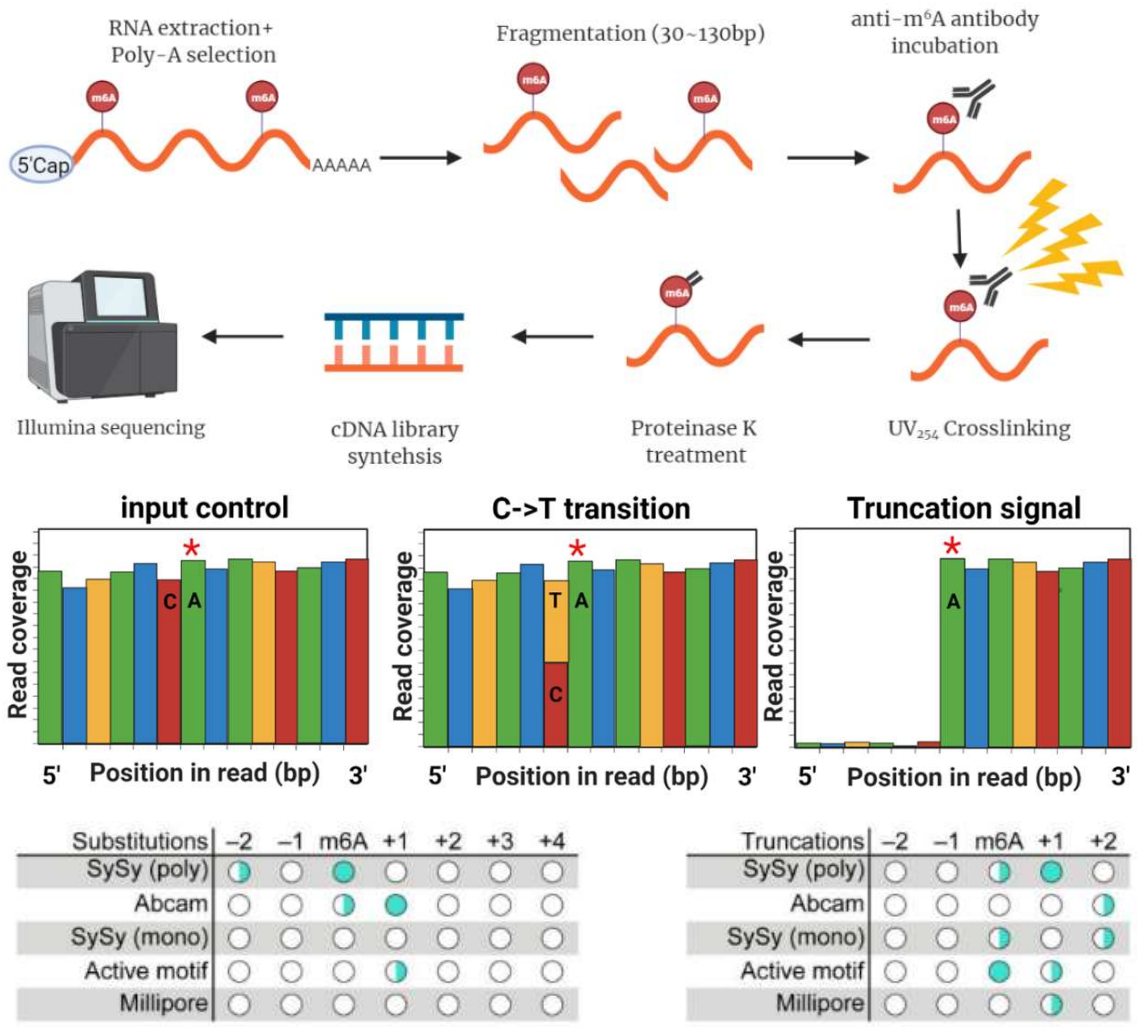


Figure 1.17: Graphical representation of miCLIP-seq workflow

miCLIP-seq begins with poly-A selection of total RNA, followed by chemical fragmentation to 30-130bp long fragments and incubation with anti-m^{6}A antibodies. Anti-m^{6}A antibodies are crossed linked to RNA using UB$_{254}$ and the complex is recovered via protein A/G purification, SDS PAGE and membrane transfer. RNA fragments are released from membrane by proteinase K digestion, and the fragments are subsequently converted into cDNA library for Illumina next generation sequencing. The peptide fragments that remain on eluted RNA after proteinase K treatment results in C-> T transition or truncation at/at the vicinity to the m^{6}A antibody binding site. The site and 'strength' of C-> T transition and truncation signals differ between different commercially available antibodies (figure extracted from Linder *et al* 2015).

1.3.5.2 ONT DRS m^{6}A detection

ONT DRS represents a novel, antibody-free method to detect chemical modifications in RNA. The first published DRS study shows that ONT DRS is sensitive enough to detect differential current signatures between *in vitro* synthesised FireFly Luciferase RNA transcripts, where the adenosines nucleotides are either m^{6}A modified or unmodified (Garalde *et al.*, 2018). In reality, ONT base-calling relies on a trained hidden Markov model to characterise a sliding window of 5 nucleotides, also known as a k-mer. Average current levels for k-mers, with or without m^{6}A modifications, were calculated and showed disparities near the modification sites (Figure 1.18A). A subsequent study also detected changes in signal signatures between modified and unmodified mRNA transcripts in the human poly(A) transcriptome (Figure 1.18B) (Workman *et al.*, 2019). This electric current disruption near modified RNA is now recognised as one of the contributors to the high base calling error rate in DRS (Li *et al.*, 2017).

Tremendous efforts have been made to explore different ways to identify and quantify m^{6}A and other RNA modifications using ONT DRS technology, but it remains technically challenging. The latest ONT base caller Guppy does not support base-calling of modified RNA nucleotides. Another ONT base caller TOMBO succeeded in base-calling the DNA modification 5mC but failed to do so with any RNA modifications, reportedly discarding up to 50% of reads when applied to ONT DRS data (H. Liu *et al.*, 2019). Different research groups have also proposed methods to take advantage of the fact that nucleotides harbouring RNA modifications have a higher sequencing error rate than unmodified bases. *ELIGOS* calculates the percentage differences between native RNA and unmodified RNA by comparing paired DRS and cDNA sequencing with the same input (Jenjaroenpun *et al.*, 2021). Using the DRACH motif to narrow down potential m^{6}A sites, they were able to identify previously validated m^{6}A sites in yeast rRNA sequences. However, this approach cannot truly distinguish distinct types of RNA modifications. Reverse transcription is also known to generate mutations at chemically modified bases,

making accurate modification detection difficult when DRS & PCR-cDNA sequencing results are compared (Helm and Motorin, 2017).

Other DRS m^{6}A detection approaches compare base-calling error rates between samples of interest with baseline non-methylated control samples. These negative controls are either *in vitro* RNA transcripts or RNA extracted from m^{6}A methyltransferase knockout/ knockdown samples (H. Liu *et al.*, 2019; Parker *et al.*, 2020; Price *et al.*, 2020). However, depletion of METTL3 and other m^{6}A writer complex components does not result in 100% depletion of RNA m^{6}A modification, which may result in high levels of false-negative m^{6}A calling. Conversely, *in vitro* RNA transcripts negates all other chemical modifications in physiological mRNAs, which may create false positive m^{6}A base-calling.

An alternative approach to detect m^{6}A is by building a base calling model using raw current data. *Nanocompore* is an analysis pipeline which allows modification calling, using comparative signal signatures between samples of interest and controls with low m^{6}A content (i.e. METTL3 knock-out samples or *in vitro* transcribed sequence) (Leger *et al.*, 2021). *Xpore* compares k-mers signatures from samples with the expected signal of unmodified k-mers from an *in vitro* transcribed human RNA DRS dataset containing all possible k-mers (with an average of 58000 reads per k-mer) (Pratanwanich *et al.*, 2021). It was found that the distribution of the k-mer current signal changes from unimodal to bimodal distributions when it contains a proportion m^{6}A modification. Therefore, the signal peaks from biological samples closest to the *in vitro* transcribed reference k-mer can be assigned as unmodified, with the second peak further away assigned as the modified peak (Figure 1.18C). Whilst these analysis pipelines still require m^{6}A-depleted datasets for validation purposes, they enable quantification of modifications in an antibody-free, unbiased manner that complements current m^{6}A research.

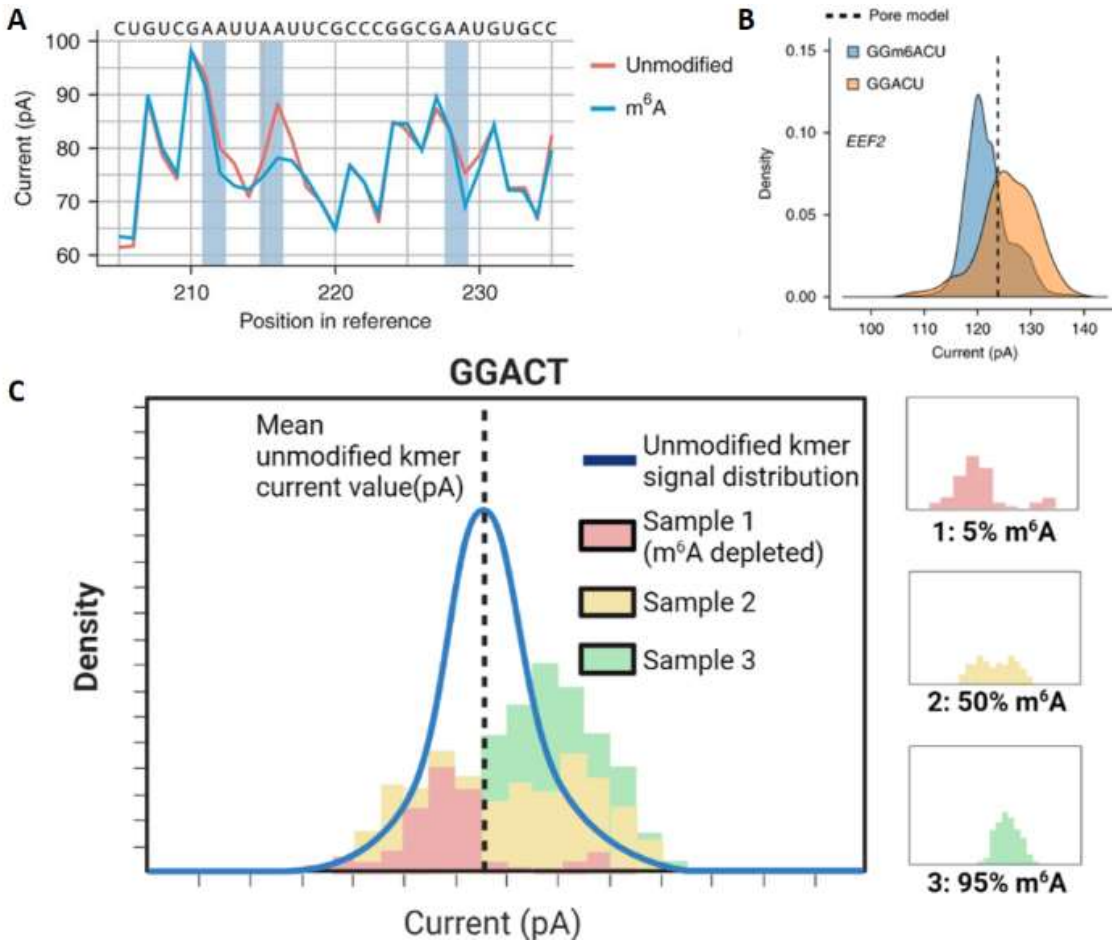


Figure 1.18: m^{6}A detection using ONT DRS technology

A) Geralde *et al.* identified current signal disruption at the vicinity of m^{6}A modified nucleotide using synthetic *in vitro* transcribed RNA fragments. (Figure extracted from Geralde *et al.* 2018) **B)** The first study using ONT DRS on human transcriptome conducted by Workman *et al.* reveals change in raw current peak distribution at a validated putative m^{6}A modified site on *EEF2* mRNA compare to *in vitro* transcribed RNA copy. Dotted line represent mean current amplitude for GGACU 5-mer in ONT model. (Figure extracted from Workman *et al.* 2019) **C)** Graphical representation of modification calling using Xpore. Current signals of 3 samples (Red, yellow and green) are plotted against mean current value from ONT model. The shift from unimodal to bimodal distribution allows quantification of modifications using probablistic model (Adapted from Pratanwanich *et al.* 2020).

1.4 Thesis aims and hypothesis

This work aims to explore the transcriptomic landscape of ccRCC and investigate the existence and co-dependency of multiple co-/post-transcriptional regulatory events using long-read Nanopore sequencing technologies. In addition, this thesis aims to explore the roles of tumour-infiltrating T cell derived cytokines in regulating the expression, isoform usages and post-transcriptional modification of key cancer immune gene transcripts.

The key hypotheses of this study are:

i) ccRCC transcriptome is shaped by tumour-infiltrating T cells and their secreted pro-inflammatory cytokines.

ii) Long-read sequencing represents a novel technology that enables high-resolution characterisation of the ccRCC transcriptomes and the co-/post-transcriptional regulatory events that shape them.

The main aims of this study are as follows:

i) To explore ccRCC transcriptome by long-read sequencing (DRS & PCS) using archival nephrectomy tissues from non-recurrent/recurrent ccRCC patients (Chapter 3 & 4).

ii) To identify key differential expression genes and transcript isoforms between tumours from non-recurrent/recurrent ccRCC patients (Chapter 4).

iii) To compare the immune landscapes between non-recurrent/recurrent ccRCC tumours via RNAseq immune cell-type deconvolution analysis (Chapter 4).

iv) To investigate the roles of pro-inflammatory cytokines (IFNγ & TNF) in shaping the transcriptome of ccRCC tumour cells using DRS (Chapter 5).

v) To characterise roles of m^6A in transcriptomic regulation in ccRCC tumour cells by applying DRS analysis on CRISPR-Cas9 mediated KO of m^6A writer *WTAP* (Chapter 5).

Chapter 2

Materials and Methods

2.1 ccRCC nephrectomy samples and ethics

In this study, 12 ccRCC tumour nephrectomy specimens from 6 non-recurrent and 6 recurrent patients were reviewed and selected by Leeds multidisciplinary research tissue bank covered by regional ethics committee approval for the biobank (Yorkshire & The Humber – Leeds East Research Ethics Committee, reference 15/YH/0080). Following surgical removal, tissue samples were washed in phosphate-buffered saline (PBS) (Gibco), blotted on a tissue before being enveloped in aluminium foil and snap frozen in liquid nitrogen. Samples were weighed before being split into approximately 30 mg tissue blocks with sterile scalpels. Once weighed and divided into smaller tissue blocks, samples were immediately used for RNA extraction without further freeze-thawed cycles. Details of each nephrectomy sample and patient data are outlined in the table below (Table 2.1).

Kidney Number	Age	Sex	Date of operation	Leibovich score	Tumour grade	Cancer stage	Relapsed
135	48	M	11/6/2001	4	3	II	Yes
171	62	M	20/6/2002	5	3	III	Yes
243	56	M	17/12/2004	5	3	III	Yes
254	55	F	9/6/2005	5	3	III	Yes
260	65	M	11/8/2005	5	3	III	Yes
329	51	M	22/1/2008	4	2	II	Yes
273	53	M	8/12/2005	5	3	III	No
314	59	M	21/6/2007	3	3	I	No
318	65	F	26/7/2007	5	3	III	No
320	45	M	13/9/2007	4	3	II	No
382	73	M	4/2/2010	3	3	I	No
395	39	M	24/6/2010	5	3	III	No

Table 2.1 ccRCC nephrectomy tissue samples clinical information

2.2 Human cell lines

HEK293-T and RCC4 cells were obtained from the American Type Culture Collection (ATCC). RCC4-Cas9-GFP and RCC4-WTAP KO clonal cell lines were generated using RCC4 cells, with methods outlined in 2.5 and 2.6. Cell culture maintenance methods are detailed in 2.4.

2.3 Reagents, antibodies, primers, guide RNAs and siRNAs

Reagents used in the thesis are shown on Table 2.2. Antibodies used in this thesis are shown on Table 2.3. Sequences of guide RNAs and qRT-PCR primers are shown on Table 2.4 and 2.5. siRNAs used in this thesis are shown on Table 2.6

Name	Supplier	Catalogue Number
Dulbecco's Modified Eagle's Medium (DMEM)	Gibco	21969-035
Foetal bovine serum (FBS)	Gibco	A5256701
L-Glutamine	Gibco	25030
Penicillin/Streptomycin	Gibco	15140
Trypsin-EDTA	Gibco	12605-010
Phosphate buffered saline (PBS)	Gibco	14190-144
Dimethyl sulfoxide (DMSO)	Sigma	D8418
1M Tris-HCl	Lonza	51237
96-well plate	Corning	3595
96-well plate (V-shaped)	Corning	3894
24-well plate	Corning	3526
6-well plate	Corning	3516
10cm^2 cell culture dish	Corning	CLS430167
T25 flask	Corning	430639
T75 flask	Corning	430641
2 mL Cryo- vials	Corning	430488
PCR tubes	Biologix	60-0088
1.5mL microcentrifuge tube	Starlab	E1415-1510
2.0mL microcentrifuge tube	Starlab	E1420-2010
1.5 mL DNA LoBind tube	Eppendorf	0030108051
OptiMEM reduced serum medium	Gibco	11058-021
Fugene 6 Transfection reagent	Promega	E2691
Lipofectamine 3000	Invitrogen	L3000001
TransIT-siQUEST transfection reagent	Mirus	MIR2114
Qiazol	Qiagen	79306
Chloroform	Sigma	32211
Ethanol	VWR	20821
EZ-10 RNA spin columns	NBS Biologicals	SD5008
RWT Buffer	Qiagen	1038708
RPE Buffer	Qiagen	1018013
RNase-Free DNase I (with RDD)	Qiagen	79254
Nuclease-free water	Cytiva	SH30538/03
Nuclease-free water	Invitrogen	AM9932
Triton X-100	Sigma	T8787
Sodium Chloride (NaCl)	Sigma	S9888
Sodium Deoxycholate	Sigma	89904
Sodium Dodecyl Sulfate	Sigma	71636
Protease inhibitor cocktail	Sigma	P8340
Phosphatase inhibitor cocktail (2)	Sigma	P5746
Phosphatase inhibitor cocktail (3)	Sigma	P0044
Pierce BCA protein assay kit	Thermo Scientific	23225
Glycerol	Sigma	G5516

β-mercaptoethanol	Sigma	M3148
Bromophenol blue	Sigma	B5525
Tween-20	Sigma	P1379
Sodium Azide (10%)	Severn biotech	40-2010-01
PVDF membrane	Millipore	IPVH00010
Amersham Hybond-N+ membrane	GE	RPN2020B
Stratalinker 2400	Stratagene	-
Methanol	Sigma	322415
Tris/Glycine transfer buffer	National diagnostic	EC880
Extra thick Western blot paper	Biorad	703967
Bovine serum albumin	Sigma	A3059
Amersham ECL reagents	GE	RPN2109
Zombie Aqua Viability Kit	BioLegend	423101
Tungsten Carbide Beads (3 mm)	Qiagen	69997
Random hexamers	Promega	C118A
dNTP	Thermo Scientific	R0191
5x First strand buffer	Invitrogen	Y02321
Dithiothreitol (DTT)	Invitrogen	Y00147
RNase OUT	Invitrogen	10000840
SuperScript II Reverse Transcriptase	Invitrogen	18064-014
MicroAmp Fast Optical 96 well plate	Applied Biosystem	4346906
MicroAmp Optical adhesive film	Applied Biosystem	4311971
Fast SYBR Green master mix	Applied Biosystem	4385612
Bioanalyzer RNA Nano kit	Agilent	5067
Qubit RNA HS assay kit	Invitrogen	Q32852
Qubit dsDNA HS assay kit	Invitrogen	Q32851
Dynabeads Oligo$(dT)_{25}$	Invitrogen	61002
Direct RNA sequencing kit	ONT	SQK-RNA002
SuperScript III Reverse transcriptase	Invitrogen	18080093
NEBNext Quick Ligation buffer	NEB	B6058
T4 DNA Ligase	NEB	M0202
RNAClean XP beads	Beckman Coulter	A66514
PCR-cDNA sequencing kit	ONT	SQK-PCS111
AMPure XP beads	Beckman Coulter	A63882
Lambda Exonuclease	NEB	M0262L
LongAmp Hot Start Taq master mix	NEB	M0533S
Maxima H Minus Reverse transcriptase	Thermo Scientific	EP0751
Uracil-specific excision reagent (USER)	NEB	M5505
Exonuclease I	NEB	M0293
PromethION Flow cells (R9.4.1)	ONT	FLO-PRO002
RNA Fragmentation kit	Invitrogen	AM8740
Sodium acetate (pH 5.2)	Sigma	S7899
Glycogen (RNA grade)	Thermo Scientific	R0551
Dynabeads Protein A magnetic beads	Invitrogen	10002D
Bromophenol Blue	Sigma	B5525
Human recombinant IFN γ	Peprotech	300-02
Human recombinant TNF	Peprotech	300-01

Table 2.2: List of reagents, reagent suppliers and catalogue numbers

Name	Species	Class/Isotype	Supplier	Clone/ Catalogue Number	Dilution
Normal rabbit IgG	Rabbit	Polyclonal IgG	Merck	12370	-
m^6A	Rabbit	Polyclonal IgG	Sigma-Aldrich	ABE572	-
WTAP	Mouse	Monoclonal IgG1	Proteintech	60188	1:1000
Cas9	Mouse	Monoclonal IgG_{2b}	Cell Signalling	7A9-3A3	1:1000
CA9	Rabbit	Polyclonal IgG	Proteintech	11071	1:1000
NDUFA4L2	Rabbit	Polyclonal IgG	Proteintech	16480	1:1000
PD-L1	Rabbit	Monoclonal IgG	Cell Signaling	E1L3N	1:1000
GAPDH	Mouse	Monoclonal IgG_{2b}	Proteintech	60004	1:5000
Anti-mouse immunoglobulin-HRP	Goat	IgG	Dako	P0447	1:5000
Anti-rabbit immunoglobulin-HRP	Goat	IgG	Dako	P0448	1:5000
PD-L1 (PE-conjugated)	Mouse	Monoclonal $IgG2_b$	Biolegend	29E.2A3	1:100
PE isotype control	Mouse	Monoclonal $IgG2_b$	Biolegend	MPC-11	1:100

Table 2.3: List of antibodies

Target sequence	Exon	Genomic Location (hg38)	PAM	Catalogue Number
GCATATGTACAAGCTTTGGA	4	Chr6:159742104-126	GGG	CM-017323-01
CTTGGGAAGAGGTTCTTCGT	2	Chr6:159736270-292	TGG	CM-017323-02
CGAAGAACCTCTTCCCAAGA	2	Chr6:159736274-296	AGG	CM-017323-03
TAGGCACTGGGCTGTCACTA	2	Chr14:21503822-844	CGG	CM-005170-01
CTGAAGTGCAGCTTGCGACA	4	Chr14:21501728-750	GGG	CM-005170-02
TCATCTGTCAGGGTCCCATA	6	Chr14:21500564-586	GGG	CM-005170-03

Table 2.4: List of guide RNAs sequences against *WTAP*

Primer names	Target	Primer sequence (5’ – 3’)
WTAP_for	WTAP	TTCCCAAGAAGGTTCGATTG
WTAP_rev	WTAP	TGCAGACTCCTGCTGTTGTT
PDL1_set2_for	PD-L1 (membrane)	GGATTACGTCTCCTCCAAATGTG
PDL1_set2_rev	PD-L1 (membrane)	CATCTTATTATGCCTTGGTGTAGCA
PDL1_common_for	PD-L1 (membrane)	TACAGCTGAATTGGTCATCCCA
PDL1_membrane_rev	PD-L1 (membrane)	TCAGTGCTACACCAAGGCAT
PDL1_soluble_rev	PD-L1 (soluble)	AGGCAGACATCATGCTAGGTG
PDL1_novel_soluble_for	PD-L1 (novel soluble)	CAGTGATTGTTGAATAAATGAATGAA
PDL1_novel_soluble_rev	PD-L1 (novel soluble)	TATTAAGTAACAATATGGTTTGGATGA
NDUFA4L2_for	NDUFA4L2	TTCTACCGGCAGATCAAAAGACA
NDUFA4L2_rev	NDUFA4L2	GGGCGAGTCGCAGCAA
BNIP3_for	BNIP3	TCAGCATGAGGAACACGAGCGT
BNIP3_rev	BNIP3	GAGGTTGTCAGACGCCTTCCAA
GAPDH_for	GAPDH	GGAGTCAACGGATTTGGTCGTA
GAPDH_rev	GAPDH	GGCAACAATATCCACTTTACAGT
SETD7_MeRIP_3UTR_for	SETD7 (m^6A site)	GGGGTTCAGAGACCTGGAAT
SETD7_MeRIP_3UTR_rev	SETD7 (m^6A site)	GCATGGTGAGAGGATGTGAC
SETD7_MeRIP_exon4_6_for	SETD7 (m^6A site)	GAATTGCGTCATTTAAAGCCTAGTT
SETD7_MeRIP_exon4_6_rev	SETD7 (m^6A site)	GTTTCATCCTACCACTCCCAATTAAT
PDL1_MeRIP_exon4_for	PD-L1 (m^6A site)	TATGGTGGTGCCGACTACAA
PDL1_MeRIP_exon4_rev	PD-L1 (m^6A site)	TGCTTGTCCAGATGACTTCG
PDL1_MeRIP_3UTR_for	PD-L1 (m^6A site)	GTGGCATCCAAGATACAAACTCA
PDL1_MeRIP_3UTR_rev	PD-L1 (m^6A site)	ATTTTCAGTGCTTGGGCCTT
PDL1_MeRIP_exon1_3_for	PD-L1 (m^6A site)	GCAGGGCATTCCAGAAAGATG
PDL1_MeRIP_exon1_3_rev	PD-L1 (m^6A site)	ATATAGGTCCTTGGGAACCGTG

Table 2.5: List of qRT-PCR primers

Name	Catalog ID
ON-TARGETplus Human *METTL3* siRNA SMARTPool	L-005170-02-0005
ON-TARGETplus Human *WTAP* siRNA SMARTPool	L-017323-00-0005
ON-TARGETplus Non-targeting Pool	D-001810-10-05

Table 2.6: List of siRNAs

2.4 Cell culture methods

2.4.1 Cell culture maintenance

All cell lines (HEK-293T, RCC4, RCC4-Cas9-GFP and RCC4-WTAP-KO cell lines) used in this study were maintained at 37°C in a humidified atmosphere of 5% CO_2 and grown in complete Dulbecco's Modified Eagle's Medium (DMEM). Complete DMEM media was prepared by using DMEM (Gibco), supplemented with 10% foetal bovine serum (FCS) (Cytiva HyClone), 1% 200 mM L-Glutamine (Gibco) and 1% penicillin/streptomycin (Gibco).

2.4.2 Cell culture passaging

All cells were sub-cultured once they reached 80-90% confluence. Cells were washed once with sterile PBS (Gibco) and detached with Trypsin-EDTA (Gibco) at the 37°C incubator for 5 minutes. Once all adherent cells were confirmed to be detached, three volumes of pre-warmed (37°C water bath) complete DMEM media were added to inactivate trypsin. Cells were centrifuged at 259 x g for 5 minutes. Cell pellets were resuspended in pre-warmed complete DMEM and seeded into new flasks at 1:10 concentration.

2.4.3 Cryopreservation and thawing of cells

Cryopreservation of cell lines was performed when cell populations reached 80 - 90% confluency in either T25 or T75 flasks. Cells were washed, detached, and pelleted following the procedure outlined in 2.4.2. Cells were resuspended in 1mL of freezing media, consisting of 90% foetal bovine serum (FBS) (Gibco) and 10% dimethyl sulfoxide (DMSO) (Sigma), and transferred to 2mL cryo-vials. Cryo-vials were incubated at -80°C for 24 hours before transferring to liquid nitrogen for permanent storage.

From liquid nitrogen storage, cells in cryo-vials were rapidly thawed in a 37°C water bath for 1 minute and resuspended in 9 mL of pre-warmed complete DMEM media. Next, cells were centrifuged to pellet at 259 x g for 5 minutes, resuspended in pre-warmed complete DMEM and seeded in either T25 or T75 tissue culture flasks.

2.5 Generation of RCC4-Cas9-GFP cell line

The RCC4-Cas9-GFP cell line was generated by transducing RCC4 cells with lentiviruses containing an eGFP-tagged Cas9 expression construct, followed by fluorescence-activated cell sorting (FACS). This cell line expresses the Cas9 protein constitutively. Firstly for the production of the lentivirus, 2 x 10^6 HEK293-T cells were seeded in a 10 cm^2 culture dish and incubated overnight before co-transfection of pLenti-Cas9-GFP (Addgene #86145), lentiviral packaging plasmid pCMVΔ8.91 (Brennan *et al.*, 2018), and pCMB-VSV-G envelope plasmid (Addgene #14888). For each 10cm^2 of HEK293-T cells, 2 µg of pLenti-Cas9-GFP, 1.5 µg of pCMVΔ8.91 and 1.5 µg of pCMB-VSV-G plasmids (5 µg of DNA in total) were mixed with 15 uL of Fugene 6 (Promega) and made up to 100 uL with OptiMEM reduced serum medium (Gibco). The transfection mix was incubated at room temperature for 30 minutes. The cell culture media for HEK293-T cells in the 10 cm^2 dish was replaced with 8 mL of optiMEM. After the incubation, the 100 uL of transfection mix was added to the HEK293T cells dropwise across the plate. OptiMEM in the 10cm^2 dish was replaced with 10 mL of pre-warmed complete DMEM. Cell supernatant containing lentiviruses was collected 48 hours post-transfection and filtered through a 0.45 µm syringe filtered. 1 mL of filtered lentiviral medium was applied to a single well of RCC4 cells in a 24-well plate, where 20000 cells were seeded 18 hours before transduction. Unused aliquots of filtered supernatant containing lentiviruses were stored at -80 ^oC. 48 hours after lentiviral transduction, GFP-positive RCC4 cells were isolated using MoFlo Astrios EQ cell sorter (Beckman Coulter). The purity of GFP positive population was validated by flow cytometry (BD Fortessa X-20) and analysed using FlowJo software (Tree Star). The expression of Cas9 protein in the GFP-positive cells was further confirmed via Western blotting. RCC4-Cas9-GFP cells were expanded and frozen down for later use.

2.6 Generation of RCC4-WTAP-KO clonal cell lines

After generating an RCC4 cell line that expresses Cas9 protein constitutively, gene knock-out (KO) cell lines can be achieved by transfection of gene-specific guide RNAs. A pool of 3x synthetic guide RNAs targeting the m^6A writer *WTAP* (Dharmacon) was selected and resuspended in 10mM Tris buffer to make a 10 μM guide RNA stock solution. 20000 RCC4-Cas9-GFP cells per well at a 24-well plate were seeded 18 hours before transfection. One hour before transfection, cell media were replaced with fresh complete DMEM media. The transfection mix was prepared by mixing 1 μL of the 10 μM guide RNA stock, 1 μL of 10 μM transactivating CRISPR RNA (tracrRNA) (U-002005, Dharmacon), and 3 μL of 10mM Tris buffer with 20 μL OptiMEM for each well. In a separate microcentrifuge tube, 1 μL of Lipofectamine 3000 (Invitrogen) was diluted in 24 μL of OptiMEM for each well and mixed by vortexing for 2-3 seconds. The 25 μL transfection mix was mixed gently with Lipofectamine-OptiMEM and incubated at room temperature for 10 minutes. 50 μL of guide RNA-Lipofectamine transfection mix was subsequently added to the RCC4 cells dropwise. Transfected cells were incubated at the 37^{o}C incubator.

After reaching 80 - 90% confluency, transfected cells were detached (using the method outlined in 2.4.2), and 2000 cells were used per 96-well plate for limiting dilution to generate monoclonal populations. 2 x 96 well plates were seeded for each transfected well in 24-well plate. To verify gene editing efficiencies, the remaining unused transfected cells from the 24-well plate were centrifuged and washed with PBS (Gibco) before being lysed for western blotting. Each well was replenished with fresh complete DMEM media every 4-5 days to maintain optimum cell growth conditions. After 2 weeks, wells with single-cell clonal expansion in the 96-well plates were trypsinised and seeded in a 24-well plate for further expansion. Once the wells reached 80-90% confluency, half of the cells were used to seed a single well at a 6-well plate, whilst another half of the cell population was lysed for gene KO validation via western blotting. Confirmed KO clonal lines were expanded and frozen down.

2.7 Cytokines treatment

Cell lines were treated with pro-inflammatory cytokines IFN-γ (peprotech) and TNF (peprotech). For the sequencing experiment in chapter 5, 1 x 10^6 RCC4 Cas9 GFP or WTAP-KO-2H1 cells were seeded in 15 mL of complete DMEM in T 75 flasks. 24 hours after seeding, media were changed into complete DMEM, with or without the addition of IFN-γ (1000U/mL) and TNF (25 ng/mL). Cells were harvested 24 hours later for RNA extraction. 3 flasks of T75s were used for each replicate for the sequencing experiment.

2.8 RNA interference

RNA interference targeting *METTL3*, *WTAP* and non-targeting control (NTC) were purchased from Dharmacon in a pool of 4 siRNAs (SMARTpool) (Table 2.6). RCC4 cells were seeded at 50,000 cells per well in 12 well plates one day before siRNA transfection (50 nM per well). For each well, the transfection mixture was prepared with 4 µL of siRNA (50 nM) and 2 µL of siQUEST transfection reagent (Mirus Bio) with OptiMem (Gibco) to make up to 160 µL in total. The transfection mix was incubated at room temperature for 20 minutes before adding dropwise to cells. After 6 hours, cell media were changed to fresh media. Then, 30 hours post-transfection, cell media were changed again with complete DMEM containing with or without the addition of IFN-γ (1000U/mL) and TNF (25ng/mL) for an additional 24 hours before being harvested.

2.9 Cell lysis and Western blotting

Cells were first washed with ice-cold PBS (Gibco) and lysed in RIPA buffer (150mM NaCl, 1% Triton X-100, 0.5% Sodium Deoxycholate, 0.1% Sodium dodecyl sulfate (SDS) in 50mM Tris buffer pH 7.4) containing protease and phosphatase inhibitors mixture (P8340, P5746, P0044, Sigma) on ice for 10 minutes. Next, cell lysates were collected and centrifuged at 10000g for 15 minutes at 4°C. Cell lysate samples were stored at -20°C.

Protein concentration for each sample was measured with a Pierce BCA assay kit (Thermo Scientific). Cell lysates were defrosted on ice and centrifuged at 10000g for 15 minutes at 4°C. Lysates were diluted with PBS (Gibco) at a 1:5 ratio and loaded on a 96-

well plate along with the BSA protein concentration standard. Each well was mixed well with BCA assay working reagent, and the 96-well plate was incubated at 37ºC for 30 minutes for colour development. The 96-well plate was read on a VersaMax microplate reader (Molecular Devices) at 562nm absorbance. Cell lysate concentrations were determined using a standard curve generated by BSA protein standards. An equal amount of proteins (10 or 20 µg) were loaded for each lane on an SDS-PAGE (polyacrylamide gel electrophoresis) gel.

Cell lysates were diluted in PBS (Gibco), mixed well with loading buffer (4x, 250 mM Tris-HCl (Lonza), 40% glycerol (Sigma), 8% SDS (Sigma), 5% β-mercaptoethanol (Sigma) and 0.05% bromophenol blue (Sigma) and denatured at 95ºC for 10 minutes on a heating block. Samples were then loaded onto a 10% SDS-PAGE gel, electroporated and resolved using a Bio-Rad PowerPac at 120V for 90 minutes. Next, the SDS-PAGE gel was placed between blotting papers (Bio-Rad) with a methanol-activated PVDF membrane (Millipore). Electro-transfer of proteins from SDS-PAGE gel to PVDF membrane was carried out at 0.2A with a maximum voltage of 25V for 90 minutes using a semi-dry Trans-Blot SD (Bio-Rad).

Once proteins from cell lysates were transferred to PVDF membranes, membranes were blocked in blocking buffer (TBS with 1% BSA (Sigma) and 0.1% Tween-20 (Sigma)) for 1 hour at room temperature on a tube-roller. Membranes were then incubated with primary antibodies diluted in blocking buffer overnight at 4ºC on a tube roller. After washing membranes 3 times for 5 minutes with TBS-T (TBS with 0.1% Tween-20 (Sigma)), membranes were incubated with HRP-conjugated secondary antibodies diluted in blocking buffer for 1 hour at room temperature on a tube-roller. Membranes were washed 3 times for 5 minutes with TBS-T and incubated with Amersham ECL (GE) at room temperature for 1 minute before visualisation via ChemiDoc (Bio-Rad). Quantification of protein bands from western blotting was performed using ImageJ (National institutes of health) by normalising protein targets with GAPDH loading control. The antibodies and concentrations used in this thesis are listed in table 2.3.

2.10 Flow cytometry

Cells were first trypsinised, washed with ice-cold PBS (Gibco) twice and transferred to a v-shaped 96-well plate. Cells were then stained with the live/dead marker Zombie Aqua (BioLegend) at 1:1000 in 100µL of PBS for 10 minutes on ice and in the dark before staining with either PE-conjugated anti-PD-L1 antibodies (1:100, Biolegend) or PE-conjugated mouse IgG_{2b} isotype control antibodies (1:100, Biolegend) in FACS buffer (PBS with 0.5% BSA (Sigma) and 0.05% Sodium Azide (Severn Biotech)) for 20 minutes on ice in the dark. After incubation, cells were washed three times with FACS buffer before resuspension in 100 µL of FACS buffer for data acquisition. Data acquisition was performed on Cytoflex LX (Beckman Coulter) and analysed using FCSexpress (De Novo) and CytExpert (Beckman Coulter).

2.11 m^6A dot blot

Total RNA samples were denatured at 95°C for 3 minutes before spotted onto an Amersham Hybond-N^+ membrane (GE) at 2µL per dot. The loaded membrane was then UV-crosslinked to the membrane using a Stratalinker (Stratagene) on auto-crosslink setting for 4 times. Membrane was blocked in TBS-T (TBS with 0.1% Tween-20 (Sigma)) with 1% of BSA (Sigma) at room temperature for 1 hour, followed by overnight incubation with anti-m^6A antibody (1:1000, Abcam) at 4°C. Membrane was washed 3 times for 5 minutes with TBS-T, then incubated with HRP-conjugated secondary antibodes (Dako P0448) for 1 hour at room temperature. Membrane was washed 3 times for 5 minutes with TBS-T and incubated with Amersham ECL (GE) at room temperature for 1 minute before visualisation via ChemiDoc (Bio-Rad). The intensity of dot blot signal was quantified using ImageJ (National institutes of health).

2.12 RNA isolation from ccRCC nephrectomy samples

To extract total RNA from ccRCC nephrectomy samples, tissues were first weighed and divided into ~30 mg sections before RNA extraction. This was performed to avoid exceeding the maximum binding capacity of the EZ-10 RNA-binding spin columns. Tissue sections were transferred to 2 mL nuclease-free microcentrifuge tubes (Starlab) with 700 µL Qiazol (Qiagen) and a stainless steel bead (3 mm diameter) (Qiagen) added for each sample. Tubes were placed in TissueLyser (Qiagen) for 2 minutes at 50 Hz to disrupt tissues and repeated until tissues were homogenised. Samples were centrifuged at 10000x g at 4^{o}C for 5 minutes, and Qiazol solutions were transferred to new 1.5 mL Nuclease-free microcentrifuge tubes (Starlab). 140 µL of chloroform was added per sample and mixed vigorously for 15 seconds. Phenol-chloroform mixtures were incubated at room temperature for 3 minutes before spinning at 12000x g at 4^{o}C for 15 minutes to separate RNA (in the aqueous phase) from DNA, lipids and proteins at the interphase and lower organic phase. For each sample, the upper aqueous phase was transferred to a nuclease-free microcentrifuge tube, where 525 µL of 100% ethanol was added and mixed well. Next, samples were transferred into EZ10 RNA spin columns (NBS Biologicals) 700 µL at a time and centrifuged at 8000x g for 15s. Samples on the columns were washed with 350 µL RWT Buffer (Qiagen). For each column, 80 µL of RNase-free DNase I (in RDD buffer) (Qiagen) was added to the membrane directly and incubated for 15 minutes at room temperature. Spin columns were washed with 350 µL RWT Buffer, followed by 500 µL of RPE buffer (Qiagen). Spin columns were then placed on a new nuclease-free 1.5 mL microcentrifuge tube and centrifuged at 10000x g for 1 minute to remove all residual wash buffers. Finally, the columns were transferred on another nuclease-free 1.5 mL microcentrifuge tube, and 30 µL of nuclease-free water (Cytiva) per spin column was added to the membrane directly. After 1 minute of incubation at room temperature, RNA molecules were eluted by centrifugation at 10000x g for 1 minute at room temperature. Eluted RNA samples were stored at -80 ^{o}C.

2.12 RNA isolation from cell lines

To extract total RNA from cell lines, cells were first washed with ice-cold PBS (Gibco), followed by lysis with 700 µL Qiazol (Qiagen). After incubation at room temperature for 5 minutes, samples were transferred from wells/flasks to a 1.5 mL Nuclease-free microcentrifuge tube by pipetting. From here onward, RNA from cell lines was extracted using the same protocol outlined in 2.9. For other RNA used for other qRT-PCR (quantitative reverse transcription PCR) assays, DNase I treatment was not performed. Columns were directly washed with 700 µL of RWT, followed by the 500 µL RPE wash. For samples that were used for RNAseq or MeRIP-qRT-PCR assays, samples were treated with DNase I, as outlined in 2.9. All samples were eluted in 30 µL of nuclease-free water and stored at -80^{o}C.

2.13 RNA concentration and quality assessment

RNA concentrations were evaluated by using NanoDrop ND 2000 (Thermo Scientific), Qubit 3 Fluorometer with Qubit RNA HS assay kit (Invitrogen), and 2100 Bioanalyzer with Bioanalyzer RNA Nano kit (Agilent). RNA quality was assessed by Nanodrop and Bioanalyzer. For NanoDrop, the quality of RNA samples was estimated using A260/230 (between 2.0-2.2) and A260/280 (~2.0 for pure RNA). Bioanalyzer generates electropherograms which provide an RNA integrity number (RIN) value, where 10 represents intact RNA molecules, and 1 represents completely degraded RNA in the sample. Other information on the samples that Bioanalyzer generated includes the ratio between ribosomal RNA (rRNA) and mRNA and RNA size distribution profiles.

2.14 cDNA synthesis

RNA molecules were reverse transcribed to cDNAs using random hexamers and SuperScript II reverse transcriptase (Invitrogen). For each sample, 1 µL of RNA (with a concentration higher than 10 ng/µL) was mixed with 1 µL of random hexamers (50 ng/µL), 1 µL of dNTP (10 mM) and 10.5 µL of nuclease-free water in a nuclease-free PCR tube (Biologix). PCR tubes were centrifuged briefly before incubating at 65 °C for 5 minutes in a SimpliAmp thermocycler (Applied Biosystems). Samples were then cooled to 4°C on the thermocycler. 4 µL of first strand buffer (Invitrogen), 2 µL of 10 µM DTT (Invitrogen), 1 µL of RNaseOUT RNase inhibitor (Invitrogen), and 0.5 µL of SuperScript II reverse transcriptase (200 U/µL) were added to each sample. Next, PCR tubes were returned to the thermocycler and further incubated at 25°C for 10 minutes, 50°C for 50 minutes and 85°C for 5 minutes. After the reverse transcription reaction, samples (20 µL each) were transferred to nuclease-free 1.5 mL microcentrifuge tubes and stored at -20°C.

2.15 qRT-PCR

To determine mRNA expression levels in samples, qRT-PCR assays were performed using SYBR Green master mix (Applied Biosystem) and transcript-specific primers listed in table 2.5. For each reaction, 10 µL of 2x Fast SYBR Green master mix was added with 7.8 µL of nuclease-free water and 0.6 µL of forward-primers and reverse-primers (10 µM) in a MicroAmp Fast Optical 96 well plate (Applied Biosystem). 1 µL of reverse-transcribed cDNA was then added to each well containing the master mix. Optical 96 well plates were sealed with MicroAmp optical adhesive films (Applied Biosystem) and centrifuged at 1000x g for 60 seconds. qRT-PCR assays were performed on a StepOnePlus Real-Time PCR system (Applied Biosystem) for 40 amplification cycles. All primer sets used in the thesis were validated to produce specific PCR-product via melt-curve analysis, with an efficiency between 80 – 120%. Primer efficiencies were determined by analysing the standard curve generated by serial dilutions of cDNA.GAPDH was used as a loading control, and relative gene expression was calculated by the comparative Ct (cycle threshold) method.

2.16 m^{6}A MeRIP-qRT-PCR

m^{6}A MeRIP-qRT-PCR assay was performed to compare levels of m^{6}A at a specific site of a transcript between samples. For each sample, a near-confluent flask of T75 was harvested with RNA extracted using the protocol outlined in 2.9. After RNA extraction, RNA concentration from each sample was determined by Nanodrop. Next, each sample was diluted to 500 ng/µL and subdivided into 9 µL aliquots. 2 µL of 10x fragmentation buffer (Invitrogen) was added to each aliquot. Samples were incubated at 70^{o}C for 10 minutes on a heating block. At the end of the incubation period, 2 µL of EDTA-based stopping solution was added to terminate the fragmentation reactions. Samples were centrifuged briefly, with 2 µL of sodium acetate (3M), 5 µL of nuclease-free water, and 1 µL of glycogen (Thermo Scientific) added to each sample. 75 µL of ice-cold 100% ethanol was added to each tube, and samples were incubated at -20^{o}C overnight. After incubation, samples were centrifuged at 10000x g for 15 minutes at 4^{o}C. The precipitated RNA pellets were washed with 70% ethanol, kept at -20^{o}C and air-dried. Finally, all RNA pellets originating from the same sample were dissolved in 21 µL of nuclease-free water (Cytiva), where 1 µL of the sample was used for Bioanalyzer RNA concentration and quality checks. Fragmented RNAs should have a size profile of 100-200nt. Fragmented RNA samples were diluted to 40 ng/µL, with 5 µL saved as input control for cDNA synthesis and later qRT-PCR assays.

To perform MeRIP, 30 µL of Dynabead protein A magnetic beads (Invitrogen) per reaction were first blocked in 1% BSA TBS-T at 4^{o}C for 2 hours. Beads were washed twice with IP buffer (250 mM NaCl, 10 mM Tris-HCl (pH 7.4), 0.1% Triton-X 100 (Sigma)), and resuspended in 500 µL of IP buffer with 1 µg anti m^{6}A antibodies (Millipore) or control normal rabbit IgG (Millipore). Antibodies and magnetic beads were conjugated for 6 hours at 4 ^{o}C and mixed using an end-over-end rotor. Next, beads were washed twice with IP buffer. For each reaction, beads were resuspended with 50 µL of fragmented RNA sample (previously diluted to 40 ng/µL, 2 µg in total for each reaction), 5 µL of RNaseOUT (Invitrogen), 345 µL of nuclease-free water and 100 µL 5x IP buffer (750 mM

NaCl, 50 mM Tris-HCl (pH 7.4), 0.5% Triton-X 100). Samples were placed on an end-over-end rotator and mixed overnight at 4°C.

After the RNA-binding step, beads were washed twice for 10 minutes on a rotator at 4°C using 1 mL of IP buffer per sample. Next, samples were washed twice for 10 minutes using 1 mL of low-salt wash buffer (50 mM NaCl, 10 mM Tris-HCl (pH 7.4), 0.1% Triton-X 100), and finally twice with 1 mL of high-salt wash buffer (500 mM NaCl, 10 mM Tris-HCl (pH 7.4), 0.1% Triton-X 100) on an end-over-end rotator at 4°C.

After m^6A-RNA-immunoprecipitation, antibodies-bound-RNA molecules were eluted using 700 µL of Qiazol. After adding Qiazol, samples were placed on an end-over-end tube mixer at room temperature for 5 minutes. Then, Qiazol samples were separated from the magnetic beads using a magnet and transferred to new nuclease-free tubes. Next, RNA extraction was performed using the protocol detailed in 2.12 and eluted in 30 µL nuclease-free water. cDNA synthesis (detailed in 2.14) was performed using both fragmented RNA input (1 µL, 40ng / µL) and pulled-down RNA (6 µL, representing 20% of all pulled-down RNA). qRT-PCR assays were performed using m^6A site-specific primers. m^6A methylated transcripts were quantified by comparing Ct values of m^6A-immunoprecipitated samples with standard curves generated via titration of corresponding input RNA samples.

2.17 poly(A)$^+$ mRNA enrichment using magnetic beads immobilised oligo d(T)$_{25}$

Poly(A)$^+$ mRNA molecules can be isolated from total RNA by affinity purification using magnetic beads coupled with oligo-(dT)$_{25}$. Firstly, 75 µg of total RNA extracted from samples was diluted to 100 µL nuclease-free water (Cytiva). Samples were placed on a 65^{o}C heating block for 2 minutes to disrupt secondary RNA structures and then placed on ice. For each sample, 200 µL of oligo-(dT)$_{25}$ conjugated magnetic beads (Invitrogen) were washed once with 100 µL of binding buffer (20 mM Tris-HCl (pH 7.5), 1 M LiCl, 2 mM EDTA), and resuspended in 100 µL binding buffer in a 1.5 mL microcentrifuge tube. RNA samples were transferred to tubes containing conjugated beads and mixed thoroughly by tube rotator for 5 minutes at room temperature. After mRNA-binding, tubes were placed on a magnet to remove the supernatant. Beads were washed twice with wash buffer (10 mM Tris-HCl (pH 7.5), 150 mM LiCl, 2 mM EDTA). Finally, poly(A)$^+$ mRNAs were eluted from beads with 11 µL nuclease-free water. Samples were heated to 65 ^{o}C for 2 minutes, and eluates were transferred to new nuclease-free microcentrifuge tubes. Poly(A)$^+$ mRNA samples were stored at -80^{o}C before sequencing library preparations.

2.18 Direct RNAseq sequencing library preparation (RNA002)

Sequencing libraries used for DRS were generated using the SQK-RNA002 sequencing kit, and an outline of the protocol can be found in figure 1.14. For each sequencing library, either 500 ng of cell line extracted, poly(A) enriched mRNA or 2 µg of total RNA from ccRCC nephrectomy were used as input. Firstly, the volume of RNA samples was adjusted to 9.5 µL with nuclease-free water (Invitrogen) in a PCR tube and mixed with 3 µL NEBNext quick ligation reaction buffer (NEB), 1 µL of RT adapter (RTA), and 1.5 µL T4 DNA Ligase (NEB). Samples were incubated for 10 minutes at room temperature for RTAs to ligate with the 3' end of RNA molecules. Next, 9 µL of nuclease-free water, 2 µL of 10 mM dNTPs (Thermo Scientifics), 8 µL of 5x first-strand buffer (Invitrogen) and 4 µL of 0.1M DTT were added to each sample, followed by the addition of 2 µL of SuperScript III reverse transcriptase (Invitrogen). Samples were mixed thoroughly, and PCR tubes were incubated in a thermal cycler (Bio-Rad) at 50ºC for 50 minutes, then 70ºC for 10 minutes, and finally down to 4ºC.

After second-strand cDNA synthesis, samples were transferred to fresh DNA LoBind tubes. 72 µL of resuspended Agencourt RNAClean XP beads (Beckman Coulter) were added to the samples and mixed on a hula mixer (Thermo Fisher Scientific) for 5 minutes at room temperature. Samples were briefly centrifuged and pelleted on a magnet. Supernatants were removed, and magnetic beads were washed with 150 µL of 70% ethanol. After removing all residual 70% ethanol, beads were resuspended in 20 µL of nuclease-free water and incubated for 5 minutes at room temperature. Beads were pelleted on a magnet, and eluates were transferred to fresh 1.5 mL DNA LoBind tubes.

Next, 8 µL of NEBNext quick ligation reaction buffer, 6 µL of RNA Adapter (RMX), 3 µL of nuclease-free water and 3 µL of T4 DNA ligase were added to each sample and mixed thoroughly by pipetting. Samples were incubated at room temperature for 10 minutes. After ligation of RMX, 16 µL of RNAClean XP beads were added to each sample and mixed for 5 minutes using a hula mixer at room temperature. Samples were then briefly centrifuged and pelleted on a magnet. Beads were washed with 150 µL of wash buffer

(WSB) twice and resuspended in 41 µL of Elution buffer, followed by 10 minutes of incubation at room temperature. 1 µL of the eluate was used to assess the yield of the library prep using the Qubit fluorometer DNA HS assay (Invitrogen). The remaining 40 µL of eluates were used as DRS input and loaded into PromethION flow cells (method detailed in 2.20).

2.19 PCR-cDNAseq sequencing library preparation (PCS-111)

Sequencing libraries used for PCR-cDNAseq in this thesis were generated using SQK-PCS-111 kit with 200ng of ccRCC nephrectomy extracted total RNA as input (Figure 1.15). RNA samples were first thawed on ice, with 200 ng of total RNA adjusted to 10 µL in volume with nuclease-free water (Invitrogen). RNA samples were then mixed well with 1 µL of cDNA reverse transcription adapter (cRTA) (ONT) and 1 µL of annealing buffer (AB) (ONT) per reaction in a 0.2 mL PCR tube. Samples were incubated in a thermal cycler (Bio-Rad) at 60 ^{o}C for 5 minutes, then cooled for 10 minutes at room temperature. Next, 3.6 µL of NEBNext quick ligation reaction buffer (NEB), 1.4 µL of T4 DNA ligase (2 x 10^{6}U /mL) (NEB) and 1 µL RNaseOUT (Invitrogen) were added to each PCR tube and incubated for 10 minutes at room temperature. After cRTAs ligation, 1 µL of Lambda Exonuclease (NEB) and 1 µL of Uracil-specific excision reagent (USER) were added to each PCR tube. Samples were incubated at 37^{o}C for 15 minutes at a thermocycler (Bio-Rad) to remove nucleotides on the doubled-stranded overhangs.

After incubation, samples were transferred to DNA LoBind tubes (Eppendorf), and 36 µL of RNAClean XP beads (Beckman Coulter) were added per sample. Samples were incubated on a hula mixer (Thermo Fisher Scientific) for 5 minutes. Samples were briefly centrifuged and pelleted on a magnet. Supernatants were removed, and magnetic beads were washed with 100 µLof short fragment buffer (SFB) (ONT) twice. After removing all residual buffer and air-dried for 30 seconds, magnetic beads were resuspended in 12 µL of nuclease-free water for elution at room temperature for 10 minutes. Eluates were transferred to fresh 0.2 mL PCR tubes.

For each PCR tube, 1 µL of RT primer (RTP) (ONT) and 1 µL of 10 mM dNTPs (Thermo Scientific) were added. Samples were incubated at room temperature for 15 minutes. Next, 4.5 µL of Maxima H Minus 5x RT Buffer (Thermo Scientific), 1 µL of RNaseOUT (Invitrogen), and 2 µL of Strand switching primer II (SSPII) were added to each sample. PCR tubes were first incubated at 42 ^{o}C for 2 minutes in a thermal cycler. 1 µL of Maxima H Minus Reverse Transcriptase (Thermo Scientific) was added to each sample, and PCR

tubes were returned to the thermocycler again and further incubated at 42ºC for 90 minutes, 85ºC for 5 minutes and then cooled to 4ºC.

After reverse transcription, reverse-transcribed samples were divided into 4x PCR tubes with 5 µL each. In each tube, 1.5 µL cDNA primer (cPRM) (ONT), 18.5 µL nuclease-free water and 25 µL of 2x LongAmp Hot Start Taq master mix (NEB) were added and mixed thoroughly by pipetting. PCR tubes were placed at a thermocycler and amplified with the following conditions: denaturation at 95ºC for 30 seconds, followed by 14 cycles of denaturation at 95ºC for 15 seconds, annealing at 62ºC for 15 seconds, extension at 65ºC for 6 minutes, a final extension step at 65ºC for 6 minutes and finally cooled to 4ºC.

Following PCR amplification, each tube was mixed with 1 µL of Exonuclease I (NEB) and digested at 37ºC for 15 minutes, followed by an inactivation step at 80ºC for 15 minutes in a thermal cycler. All 4 PCR reactions were then pooled into a fresh 1.5 DNA LoBind tube (Eppendorf), and 160 µL of AMPure XP beads (Beckman Coulter) were added to each tube. Samples were incubated on a Hula mixer for 5 minutes at room temperature and pelleted on a magnet. Beads were washed twice with 500 µL 70% ethanol, pelleted by a magnet and airdried for 30 seconds. 12 µL of elution buffer (EB) (ONT) was added to each sample and incubated for 10 minutes at room temperature. Eluates were separated from beads using a magnet and transferred to fresh 1.5 mL DNA LoBind tubes. Using the Qubit fluorometer DNA HS assay (Invitrogen), 1 µL from each sample was used to check the yield of cDNA amplification. 20 ng of amplified cDNA per sample was used for the next steps and adjusted to 11 µL using EB, with the rest of the cDNA stored at -80ºC. Finally, 1 µL of Rapid Adapter T (RAP T) was added to each sample and samples were incubated at room temperature for 5 minutes. cDNA libraries were now able to be loaded on PromethION flow cells (method detailed in 2.20).

2.20 ONT PromethION flow cell loading and sequencing

PromethION flow cells (R9.4.1) were used in all sequencing experiments in this thesis. Before sequencing runs, flow cells were inserted in the PromethION sequencer, and the number of available pores was checked. All sequencing runs were performed using PromethION flow cells with at least 5000 available pores before the start of the run.

Following hardware checks, flow cells were primed with the priming mix (1170 µL of Flush buffer (FB) and 30 µL of Flush tether (FLT)). For each PromethION flow cell, 500 µL of priming mix was first flushed into the inlet port. After 5 minutes of incubation, another 500 µL of priming mix was added. Next, the 40 µL RNA libraries (DRS or PCS) prepared in 2.18 / 2.19 were mixed with 35 µL of nuclease-free water (Invitrogen) and 75 µL of RNA Running Buffer (RRB) and loaded into the flow cell by pipetting drop-wise through the inlet port. Sequencing runs were then started and lasted for 72h each.

2.20 Mapping of sequencing data

Sequencing reads generated from Direct RNAseq, and PCR-cDNAseq with a minimum read quality score of 7 were used for mapping and downstream analysis. For Direct RNAseq, FASTA files generated from sequencing runs were concatenated using catfishq (version 1.3.0) and aligned using an index built by either human genome (Ensembl release 105, Genome assembly version: GRCh38) or transcriptome (Ensembl release 105 cDNA reference) using minimap2 (version 2.22), with recommended parameters for DRS data (-ax splice –uf –k14: spliced mapping mode, forward transcript strand only, kmer size of 14, minimum mapping score of 10) (Li, 2018). Aligned reads were sorted, merged and indexed to BAM files with samtools (version 1.13) (Li *et al.*, 2009).

For PCR-cDNAseq, FASTA files were first concatenated using catfishq (version 1.3.0). Reads were first processed by pychopper (version 2.5.0). Pychopper re-orientates reads, filters for reads with 5' and 3' sequencing adapters for full-length reads and finally trims off sequencing adapters sequences. Poly(A) tails from reads were identified and trimmed from reads by cutadapt (version 4.1) (Martin, 2011). Processed reads were subsequently aligned to the genome/transcriptome using an index built by either the human genome (Ensembl release 105, Genome assembly version: GRCh38) or transcriptome (Ensembl release 105 cDNA reference) using minimap2 (version 2.22), with recommended parameters (-ax splice –uf –k14: spliced mapping mode, forward transcript strand only, kmer size of 14, minimum mapping score of 10). Aligned reads were sorted, merged and indexed to BAM files with samtools (version 1.13). Mapping Data quality and statistics of DRS and PCS data were analysed and visualised using Nanoplot (De Coster *et al.*, 2018). Analysis by BamSlam was used to provide statistics on reference transcriptome-mapped read lengths and coverage of annotated transcripts (Gleeson *et al.*, 2022).

2.21 Differential gene expression

FeatureCounts performed gene expression counts of genome-aligned DRS sequencing data (Liao *et al.*, 2014). Transcript counts of transcriptome-aligned DRS data were quantified with Salmon (version 1.5.2) with the error model and length correction disabled (--noErrorModel, -noLengthCorrection) using human reference transcriptome (Ensembl release 105, cDNA reference) (Patro *et al.*, 2017b). Count matrices were imported to R with tximport (version 1.22, R-version 4.2.1), followed by normalisation and identification of differentially expressed genes (padj value < 0.1 and $\log_2$Fold change > 2) using DESeq2 (version 3.15, R-version 4.2.1) (Love *et al.*, 2014; Soneson *et al.*, 2016). P values were adjusted by the Benjamini-Hochberg Method.

For expression levels of novel transcripts, read counts were generated using samtools view command (version 1.13) for the specified genomic region, as noted in the figure. Read counts were scaled to reads per million with the number of aligned reads per sample, using mapping statistics generated by Nanoplot (Section 2.19).

2.22 Gene set enrichment analysis (GSEA) of RNA-seq data

Gene set enrichment analysis of transcriptomic data was performed using *clusterProfiler* (v4.4.4, R-version 4.2.1) (Wu *et al.*, 2021). Gene ontology (GO) biological process (BP), cellular component (CC), molecular function (MF) and Kyoto Encyclopaedia of Genes and Genomes (KEGG) databases were used for functional enrichment analysis of differentially expressed genes from RNA-seq. Parameters used for GO and KEGG 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). Graphs were plotted using ggplot2 (v3.3.6, R-version 4.2.1).

2.23 Estimation of stromal and tumour-infiltrating immune cell population abundance by *ESTIMATE* and *xCell*

Tumour purity, stromal and tumour-infiltrating immune cell population abundance was estimated using two gene signature-based algorithms: ESTIMATE (v1.0.13, R-version 4.2.1) and xCell (v1.1.0, R-version 4.2.1). ESTIMATE produces stromal and immune scores based on expression levels of 130 stromal/immune gene signatures (Yoshihara *et al.*, 2013). xCell uses 489 gene signatures to infer 64 immune and stromal cell types and produce an overall immune and stromal score based on gene expression levels (Aran *et al.*, 2017). These scores represent the presence and abundance of stromal cells and tumour-infiltrating immune cell populations in tumour samples. Entrez gene IDs for DRS/PCS mapped genes, and corresponding genome mapped DESeq2 normalised expression for PCS data of ccRCC nephrectomy samples were used as input for ESTIMATE. Entrez gene IDs for mapped genes were retrieved using biomaRt (v4.2, R-version 4.2.1). Genome-mapped DESeq2 normalised expression for PCS data of ccRCC tumour samples was used as input for xCell.

2.24 Tumour-infiltrating immune cell type deconvolution using CIBERSORTx and EPIC

DRS and PCS data from ccRCC nephrectomy samples were computationally deconvoluted using CIBERSORTx and *EPIC* (Racle *et al.*, 2017; Newman *et al.*, 2019). Using these algorithms, the abundance of tumour-infiltrating immune cells was estimated for each ccRCC tumour sample.

For CIBERSORTx, files containing gene IDs and DESeq2 normalised expression for genome-reference mapped PCS/DRS data were uploaded to the CIBERSORTx online analysis platform (cibersortx.standford.edu). Using the LM22 gene signature matrix (547 genes), CIBERSORTx was used to impute 22 different immune cell types. CIBERSORTx

analysis was performed using absolute mode, with batch correction and quantile normalisation options disabled and permutations set at 1000.

For EPIC, files containing gene IDs and DESeq2 normalised expression for genome-reference mapped PCS/DRS data were uploaded to the *EPIC* online analysis platform (epic.gfellerlab.org). The EPIC Tumour infiltrating cells gene signature matrix (98 genes) was used to deconvolute seven different immune cell types. Both CIBERSORTx and EPIC produce a p-score for confidence in the deconvolution of each cell type. Only cell types with $p < 0.1$ across all 12 tumour samples were discussed in this thesis.

2.25 Differential transcript usage using reference transcriptome-mapped data

Differential transcript usage (DTU) analysis of DRS and PCS data was performed using the R library rnaseqDTU (version 3.14, R-version 4.2.1) (Love *et al.*, 2018). Normalised, scaled transcript-per-million (scaledTPM) from Salmon quantification output files were imported using tximport (version 1.22, R-version 4.2.1), followed by transcript-to-gene mapping of human reference transcriptome (Ensembl release 105, cDNA reference) using the R library GenomicFeatures (Lawrence *et al.*, 2013). Transcripts were filtered using a minimum transcript expression of 3 (scaled TPM) across all samples, 5% of total gene expression in at least half of the samples, and removal of transcripts where only one isoform was identified. Genes with differential transcript usage between cell lines and treatment conditions were identified using either DRIMSeq or DEXSeq, followed by stageR statistical package where genes with padj value < 0.1 were considered significant (Anders *et al.*, 2012; Robinson and Nowicka, 2016; Van den Berge *et al.*, 2017).

2.26 Transcriptome assembly by *StringTie2*

To explore potential novel transcripts from ccRCC transcriptomic data, *StringTie2* was used to generate *de novo* assembled transcriptomes from DRS/PCS data, which were then compared with reference gene annotation to define novel isoforms. Using genome-aligned (Ensembl release 105 human genome reference) DRS/PCS data (.bam files), transcripts were assembled by *StringTie2* using long-reads processing mode (-L) without using reference gene annotation files to guide the assembly. Transcriptome assemblies from all tumour samples were merged using the --merge option, and the generated transcript annotation files (.gtf) from merged assemblies were compared to Ensembl reference GRCh 38 gene annotation (with -r option) using GffCompare (v0.12.6) (Pertea and Pertea, 2020).

GffCompare provides transcript classification codes for each assembled transcript from the StringTie2 assemblies. The transcript classification codes indicate the relationship between assembled transcript in question and the closest related transcript from the reference gene annotation. The transcript classification codes can be broadly categorised into three groups: known transcripts (=, c), novel transcripts (j, k, m, n, o, i, x, y, u), and potential artefacts (p, e, s, r) (Gleeson *et al.*, 2022). Classification codes and what they represent are listed below in Figure 2.1. StringTie2 assembled transcript annotation files were converted to .bed files for visualisation on the integrative genomics viewer (IGV).

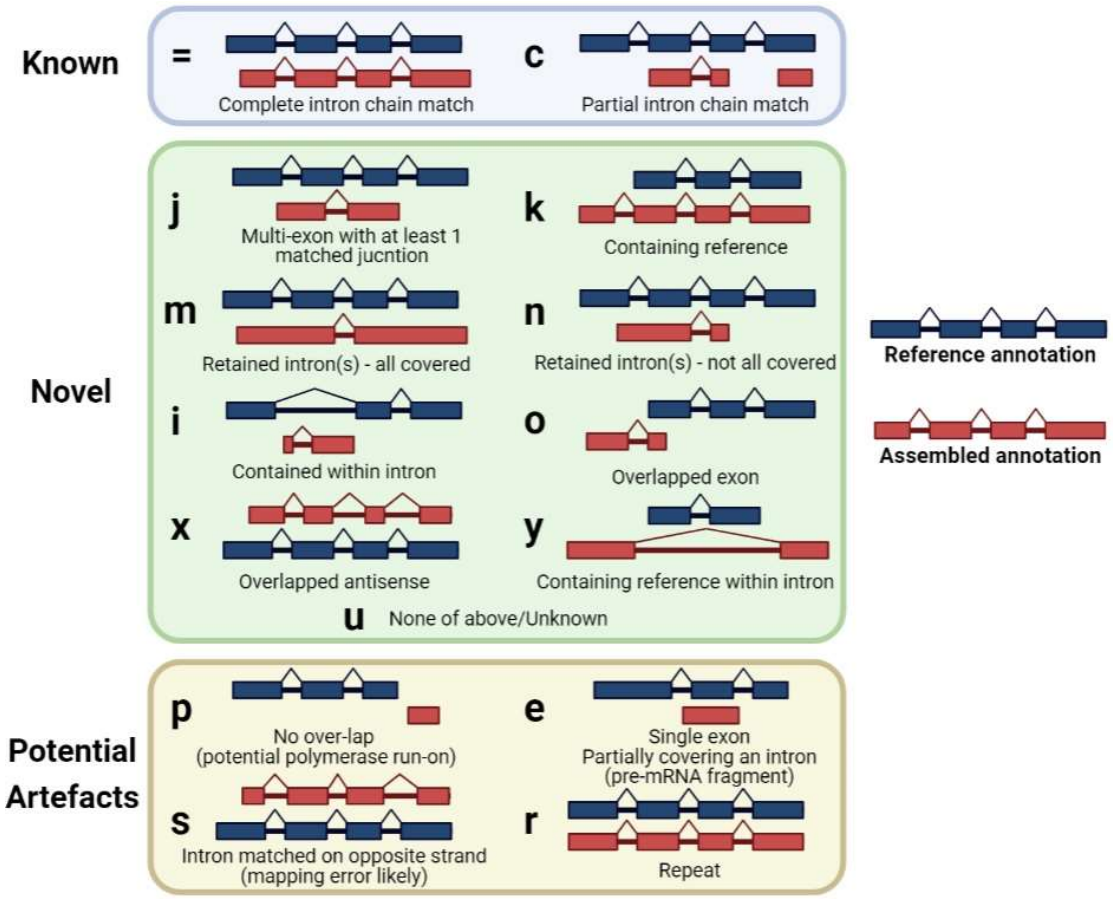


Figure 2.1: GffCompare transcript classification code

GffCompare provides a class code for assembled annotation transcripts compared to the most-related reference transcript annotation. Figure adapted from Pertea *et al. 2020*.

2.27 Transcriptome assembly and novel isoforms discovery by FLAIR

FLAIR is another transcriptome-assembly method used in this study to discover novel isoforms (Tang *et al.*, 2020). Genome-aligned (Ensembl release 105 human genome reference) DRS/PCS data (.bam files) were inputted. Misaligned splice sites from mapped transcripts were first corrected using the 'flair correct' command. Next, high-confidence isoforms were defined from corrected transcripts using the 'flair collapse' command, with the long-read optimised option selected (--trust_ends). In addition, human CAGE-seq data from the FANTOM5 consortium was used to define the 5' start sites and to filter out truncated isoforms that are erroneously marked as novel isoforms (Noguchi *et al.*, 2017). Corrected transcripts must be within the 100 nt range from the closest annotated transcription start sites, as annotated by the CAGE-seq data. The '--stringent' option was specified to ensure corrected transcripts used for flair collapse span at least 80% of any annotated reference transcript and have at least 25 nt overlaps with the first and last exon. Generated transcript annotation files (.gtf) from flair-collapse were compared to Ensembl reference gene annotation (with -r option) using GffCompare (v0.12.6). Finally, flair-generated transcript annotation files were converted to .bed files for visualisation on the IGV.

2.28 Estimation of poly(A) tail length by nanopolish

Estimation of poly(A) tail lengths of Direct RNAseq reads was performed by nanopolish (v0.14.0). Firstly, raw sequence reads (FASTQ) were indexed (nanopolish index) with ONT raw data files (FAST5) and the sequencing summary file so that each read could be directly linked with the electric signals used to base-call the sequence using a unique read id. Next, the length of the poly(A) tail from each read was estimated (nanopolish polya). Poly(A) tail length estimations were associated with transcripts using information from reference transcriptome aligned bam files, where the Ensembl transcript ID and read id for each uniquely mapped read (primary alignment) was extracted using samtools view -F (version 1.13). For novel transcripts where Ensembl transcript ID is unavailable, read IDs aligned to the novel transcript-specific region were extracted from reference genome mapped files using samtools view (version 1.13). Reads with poly(A) tail length estimations were then associated with the reference gene IDs and RNA biotypes using bedtools (v2.28) and biomaRt (v4.2, R-version 4.2.1). Lengths of poly(A) tail are plotted using R ggplot2 (v3.3.6, R-version 4.2.1) and GraphPad Prism 9.

2.29 Analysis of publicly available datasets

Clinical, genomic, and transcriptomic data of 510 ccRCC patients from the TCGA KIRC dataset were obtained from cbioportal (Weinstein *et al.*, 2013). Genomic data was used for gene copy number variations (CNV) identification and correlations between CNV and the overall survival of ccRCC patients. Transcriptomic data was used for gene expression analysis and correlations between high and low target gene expression (based on median expression level) and overall survival of patients. Kaplan-Meier plots were generated for survival analysis using GraphPad Prism 9.

2.30 Statistical analysis

Statistical analysis was performed using GraphPad Prism 9 or RStudio. Two-tailed unpaired T-tests with Welch's correction were used to compare gene or transcript expression levels between experimental groups, with $p < 0.05$ considered statistically significant. R^2 (coefficient of determination) was used to calculate the goodness of fit between datasets, and P values were generated from F-test, with $p < 0.05$ 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 are considered to be significantly differentially expressed. Differential transcript usage analysis by Drimseq and DEXseq performs likelihood ratio statistics based on the Dirichlet-multinomial and general linear models, followed by false discovery rate correction by the Benjamini-Hochberg Method. padj values were given for both genes and transcripts. Genes are considered to have statistically significant differential transcript usage when padj values at both gene and transcript levels are less than 0.1. Comparisons of poly(A) tail length were analysed by nested two-tailed nested T test which accounts for intra-dataset (sequencing run) variance, before comparing the conditions. $P < 0.05$ was considered statistically significant. Differences in patients' overall survival were assessed using a log-rank (Mantel-Cox) test, where $p < 0.05$ was considered statistically significant.

Chapter 3

Assessment of long-read transcriptome sequencing approaches to profile archival tumour tissue samples

3.1 Introduction

RNA sequencing using next-generation sequencing technologies has revolutionised cancer research, allowing in-depth, high-resolution global assessment of the tumour transcriptome on an unprecedented scale. Illumina-based RNAseq technologies have enabled the identification of gene signatures that predict cancer prognosis and treatment outcomes. These biomarkers are crucial in developing strategies for patient stratification and targeted cancer therapy (Büttner *et al.*, 2022). Advancements in single-cell RNA sequencing technologies (scRNAseq) and bulk-RNAseq cell-type deconvolution methods have allowed quantitative characterisation of intra-tumoural heterogeneity and the tumour microenvironment, which is now widely recognised to have profound implications for disease progression and clinical outcome (Dagogo-Jack and Shaw, 2018). With short-read-based methods, aberrant regulation in splicing, alternative polyadenylation and mRNA chemical modification have all been reported to play critical roles in cancer development (Chen *et al.*, 2019; Y. Zhang *et al.*, 2021; Yuan *et al.*, 2021). Whilst short-read sequencing is now ubiquitously used, long-read sequencing methods offer promising opportunities to further advance and integrate transcriptome-wide gene expression, splicing and epitranscriptomic profiling at the single mRNA molecule level.

The field of long-read RNA sequencing is currently in rapid development. Yet, at the time of writing, only a limited number of published works have utilised long-read RNAseq technologies using clinical tumour samples. Using RNAseq on clinical tumour samples, especially archival samples, poses several technical challenges. One of the significant challenges is RNA degradation. Archival tumour samples are commonly preserved as either fresh-frozen tissues (snap-frozen in liquid nitrogen) or formalin-fixed paraffin-embedded (FFPE) tissues. Fresh frozen tissues are considered more suitable for RNAseq experiments than FFPE tissues since chemical fixation by FFPE tissue preparation and their routine storage at room temperature lead to RNA degradation(Liu *et al.*, 2022). Typically, RINs (RNA Integrity Numbers) from fresh frozen tumours RNA range between 6.0 – 8.0, whereas RINs from RNA extracted from FFPE tissues are often

lower than 2.0 (Kap *et al.*, 2014; Lalmahomed *et al.*, 2017; Marczyk *et al.*, 2019). Recently it has been shown that FFPE samples can be used for transcriptomic analysis using Illumina technologies, which allows the integration between gene expression profiles with spatial information via tissue section staining (Gracia Villacampa *et al.*, 2021). However, it was also reported that a large proportion of RNA from FFPE samples failed to be captured by poly(A) tail enrichment due to the highly degraded or absence of poly(A) tail (Pennock *et al.*, 2019). The first step of both ONT DRS and PCS libraries requires ligating sequencing adaptor primers with poly-d(T) overhang to capture mRNAs. Thus highly degraded mRNA samples may not be suitable input for ONT sequencing libraries preparation. All assessed tumour samples in this study were snap-frozen in liquid nitrogen. No work has been published to demonstrate the relationships between sequencing output levels (number of reads), average sequencing read lengths, and RNA quality (by RIN score) using ONT DRS or PCS.

Another major challenge of RNA sequencing of archival tumour samples is the limited yield of extracted RNA. For Illumina-based sequencing libraries and ONT PCR-cDNAseq (PCS111), the low input of RNA per sample is manageable since libraries are PCR-amplified. ONT PCR cDNAseq (PCS111) library requires 4 ng of poly(A) enriched RNA or 200ng of total RNA with 14 amplification cycles. In comparison, without PCR amplification, direct RNAseq requires a much higher amount of input RNA. Previous studies have typically used 50 – 500 ng of poly(A) enriched RNA as input for library preparation (Jain *et al.*, 2022). With only 1 – 5 % of total cellular RNA being polyadenylated mRNA, up to 5 – 50 µg of total RNA per sample would be needed if published protocols are followed. This amount of RNA is almost always unachievable from tumour samples. An alternative input quantity is thus needed.

At the time of writing, a published studies have applied long-read Nanopore RNA sequencing technologies on clinical tumour samples. Oka *et al.* used ONT cDNA sequencing technologies and identified aberrant splicing isoforms from 22 non-small cell lung cancer (NSCLC) cell lines, seven clinical lung cancer samples and seven

corresponding non-cancer controls. Full-length cDNA libraries were first generated from the clinical samples, and 1.5 µg of cDNA was used as input for sequencing library generation (LSK308). Sequencing libraries were subsequently sequenced using the MinION sequencer and R9 MinION flow cells. On average, 2.69 million reads were generated for each clinical lung tumour and control samples, and 448 novel transcript isoforms were identified from the tumours. The authors subsequently analysed the Illumina sequencing data from TCGA and genotype-tissue expression (GTEx) database and validated the expression of several novel isoforms. This study demonstrates the feasibility of using nanopore technologies to capture novel transcripts. However, it is important to note that the LSK308 library kit used in this study is optimised for the sequencing of genomic DNA. Unlike the DRS and PCS library preparation kits, the genomic DNA kit is designed to enrich long DNA fragments, which would present significant selection biases towards long transcripts. This length selection bias will likely interfere with any differential gene expression analysis. Moreover, after filtering for reads quality (read quality < 20), the average number of reads per sample dropped from 2.69 million raw reads to 520,000, which may not provide sufficient depth to characterise the diversity of transcript isoforms in tumour samples (Oka *et al.*, 2021).

After the generation of raw sequencing reads, reads must align to reference genome or transcriptome before quantification for gene expression analysis. For ONT long-read sequencing results, the aligner *minimap2* is universally recommended and used in all published studies due to its option to align long, relatively high error 'noisy' reads to the reference genome/transcriptome (Li, 2018). Comparing matched sequencing results of HEK293 and HAP1 cell lines using ONT direct RNA seq (RNA001) and direct cDNAseq (DCS108), Soneson *et al*. concluded that a maximum number of identified genes can be obtained by mapping via *minimap2*, followed by gene-level quantification (Soneson *et al.*, 2019). However, this study provided no direct comparisons between the number of uniquely identified genes and their RNA biotypes after aligning to reference genome or reference transcriptome, nor did it compare sequencing results between DRS and PCS.

There are significant gaps in knowledge concerning the use of ONT long-read sequencing on clinical samples, as well as a lack of characterisation studies on the similarities and differences between DRS and PCS results.

3.2 Chapter aims

The overarching aim of this chapter is to **assess the feasibility of using ONT long-read sequencing technologies (DRS and PCS) to profile archival tumour samples**. Other aims of this chapter include:

i) Examination of relationships between RNA quality, the number of DRS/PCS generated sequencing reads, and the length of sequencing reads

ii) Characterisation of RNA biotypes of the reference genome and reference transcriptome aligned DRS & PCS reads from tumour samples

iii) Assessment of gene expression levels of different RNA biotypes in DRS and PCS of tumour samples

iv) Correlations and comparisons of gene expression profiles between the reference genome and reference transcriptome aligned DRS & PCS of tumour samples.

3.3 Results

3.3.1 Yield and quality of RNA extracted from ccRCC tumours

Total RNA was successfully isolated from the 12 ccRCC tumour samples. Summary data of the tumour samples, RNA quantities and qualities can be found in table 3.1.

No.	ccRCC Sample Number	Weight (g)	Number of extractions	RIN scores	Nanodrop	Bioanalyzer	Qubit RNA HS
					RNA conc. (ng/µL)		
1	135	0.0119	1	7.4	460	520	276
2	171	0.0783	3	8.5	555	570	306
				8.7	750	685	391
				7.9	742	690	399
3	243	0.0312	1	7.2	487	540	336
4	254	0.1076	3	7.6	617	585	371
				7.7	590	560	332
				6.5	480	435	308
5	260	0.0416	1	8.2	694	810	339
6	273	0.0852	3	8.5	739	905	540
				8.2	924	975	800
				6.7	815	930	670
7	314	0.034	1	8.3	328	445	322
8	318	0.0394	1	8.6	641	545	388
9	320	0.0399	1	7.4	654	765	437
10	329	0.0969	3	8.0	527	665	350
				-	398	402	332
				6.6	336	375	280
11	382	0.059	2	8.1	438	505	331
				8.1	565	635	373
12	395	0.1059	3	8.5	1059	1050	530
				8.4	409	615	340
				8.2	554	660	446

Table 3.1: ccRCC tumour sample sizes, RNA concentrations and integrity

ccRCC sample number, tumour sizes, RNA integrity numbers (RIN), and RNA concentrations as determined by Nanodrop, Bioanalyzer RNA Nano, and Qubit RNA HS assays. Samples labelled in red were used for transcriptome-profiling.

RNA concentrations from each extraction (n = 23 tumour pieces from 12 tumours) were determined by Nanodrop, Bioanalyzer and Qubit RNA HS assay, with median concentrations of 565 ng/μL (range: 328 – 1059 ng/ μL), 615 ng/μL (range: 378 – 1050 ng/ μL) and 350 ng/μL (range: 276 – 540 ng/ μL), measured by respective methods. The elution volume used for all extractions is 30 μL. Published protocols have used RNA concentration from both Bioanalyzer and Qubit. Correlative analysis was thus performed here to assess the suitability of the quantification methods. Whilst there are high levels of variation in the measured RNA concentration levels, subsequent correlation analysis demonstrates significant correlations between the quantification methods. RNA concentration levels of the extracted ccRCC tumour RNA samples measured by Qubit RNA HS assay were strongly correlated with Nanodrop (R^2 = 0.5844 and p = <0.0001) and Bioanalyzer (R^2 = 0.6711 and p = <0.0001) (Figure 3.1A and B). A high degree of concordance was also found between RNA concentrations measured by Nanodrop and Bioanalyzer (R^2 = 0.8218 and p = <0.0001) (Figure 3.1C).

Next, the relationship between tumour weight and yield of extracted total RNA was examined by producing scatter plots for correlation analysis. Strong evidence of correlation was found between the weight of tumours and the extracted RNA yield (combined yield from all extractions for each tumour block) as determined by Qubit RNA HS (R^2 = 0.6291 and p = 0.0012), Bioanalyzer (R^2 = 0.6803 and p = 0.0010) and Nanodrop (R^2 = 0.7099 and p = <0.0006) (Figure 3.1D - F).

The RIN score for each RNA sample was determined by Bioanalyzer. RIN scores for the tumour samples ranged between 6.7 and 8.7, with a median RIN score of 8.1. RNA sample with the highest RIN score (and highest RNA concentration by Qubit RNA HS where there is a tie) for each tumour sample was used for Nanopore RNA sequencing. Sequenced RNA samples are highlighted in red in table 3.1. The Bioanalyzer gel image showing the size distribution of RNA fragments from the sequenced samples is displayed in Figure 3.2.

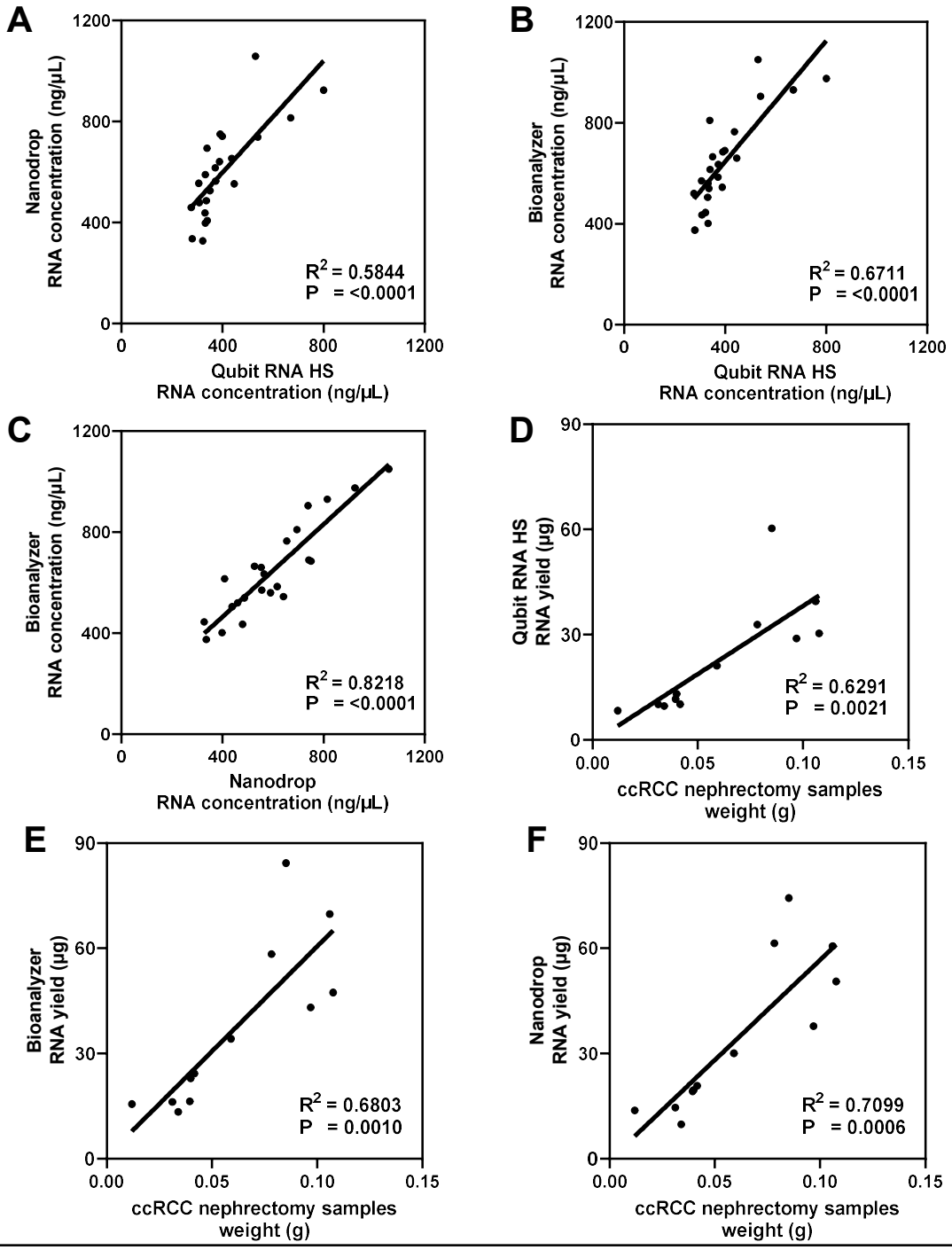


Figure 3.1: Assessment of RNA concentration and yield from ccRCC samples

Extracted total RNA (n = 23) from ccRCC samples (n = 12) were quantified using Nanodrop, Bioanalyzer (RNA Nano) and Qubit (RNA HS). Scatter plots showing correlation between RNA concentrations determined by **A)** Nanodrop & Qubit, **B)** Bioanalyzer & Qubit, and **C)** Bioanalyzer and Nanodrop. Scatter plots showing correlations between ccRCC nephrectomy sample weights and total RNA yield measured by **D)** Qubit, **E)** Bioanalyzer and **F)** Nanodrop. Throughout, diagonal lines represent the line of best fit. R^2 values were computed to measure goodness-of-fit, and P values generated from F-test, with $p < 0.05$ considered statistically significant.

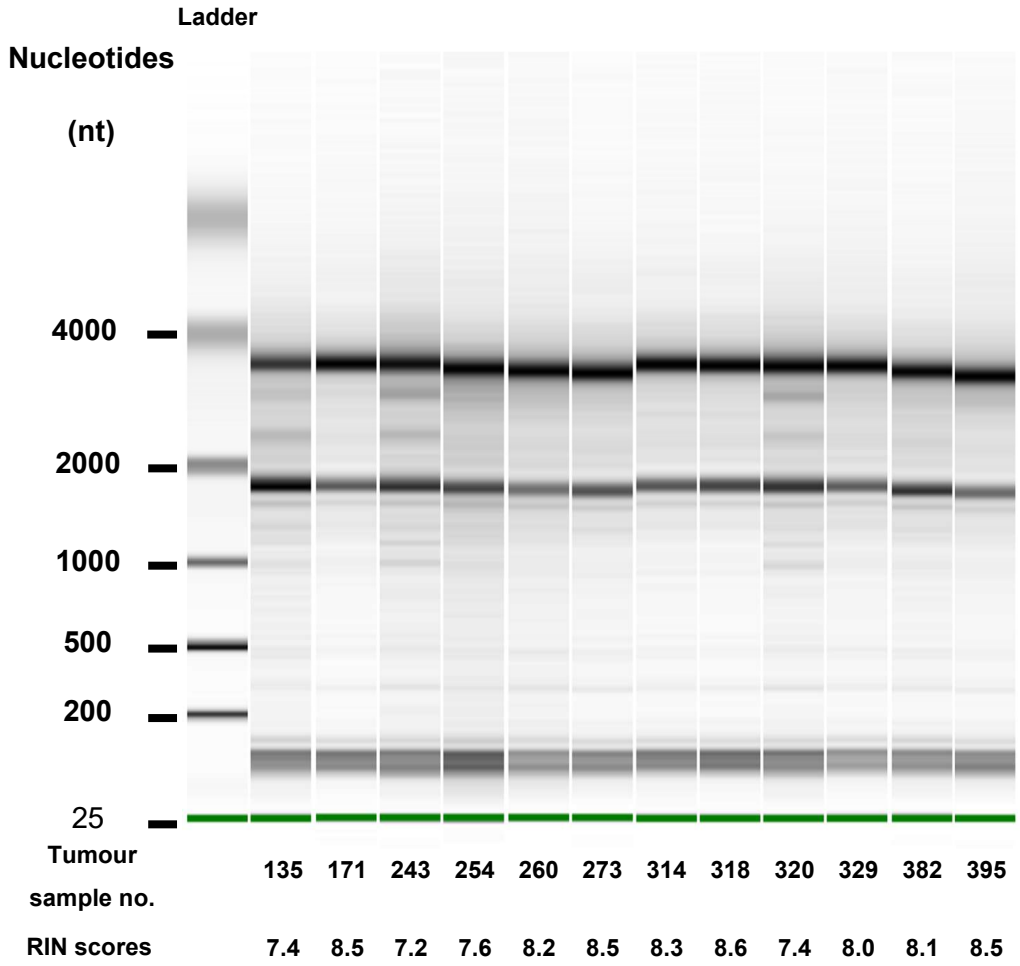


Figure 3.2: ccRCC tumour RNA analysis by Agilent 2100 bioanalyzer

Analysis of extracted total RNA from ccRCC nephrectomy samples on an Agilent 2100 bioanalyzer using the RNA 6000 Nano kit. Bioanalyzer gel image of the extracted RNA samples is shown, with visible 28S and 18S rRNA bands. Ratio of 28S:18S bands were used to assess integrity of RNA samples (where 10 represent intact RNAs and 1 represent completely degraded RNAs). RIN scores indicated below sample numbers.

3.3.2 Evaluation of sequencing output from DRS and PCS of ccRCC tumour samples

ccRCC tumour samples were sequenced on a PromethION sequencer using PromethION flow cells (R9.4.1), using DRS and PCS libraries prepared from 2 µg and 200 ng of total RNA, respectively. The workflow for the sequencing experiment is outlined in Figure 3.3. After 72 hours of sequencing, DRS generated 1.9 – 4.7 million sequencing reads per tumour sample that passed sequence quality control (read quality Q score above 7), with a median of 3.2 million passed reads per sample (Figure 3.4A). PCS generated 43.9 – 72.8 million sequencing reads per tumour sample with Q above 7, with a median of 56.6 million passed reads per tumour sample (Figure 3.4B).

To understand the relationship between the quantity of generated sequencing reads and the quality of RNA samples used for sequencing library preparation, scatter plots between the number of DRS- and PCS-generated passed reads, as well as the RIN scores of RNA samples, were generated. A borderline non-significant negative correlation between the number of DRS- and PCS-generated passed reads was observed (R^2 = 0.3016 and p = 0.0644) (Figure 3.4C). No correlation was found between the number of DRS-generated passed reads and the corresponding RNA sample RIN score (R^2 = 0.0089 and p = 0.7708) (Figure 3.4D). In contrast, a strong negative correlation was observed between the number of PCS-generated passed reads and the corresponding RNA sample RIN score (R^2 = 0.5172 and p = 0.0084) (Figure 3.4E).

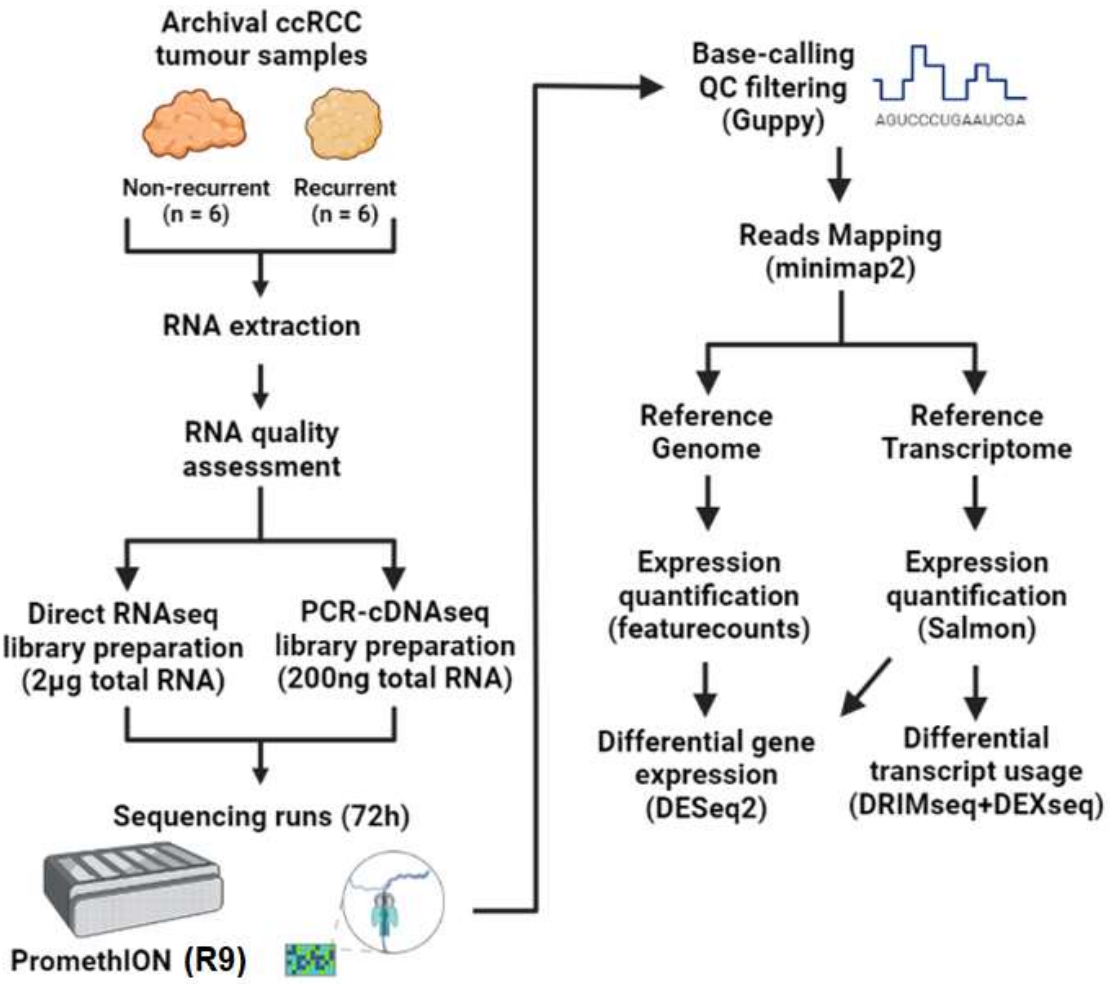


Figure 3.3: Summary of clinical samples DRS and PCS workflow

RNA was extracted from Fresh frozen archival ccRCC tumour samples, followed by quality assessment by Agilent bioanalyzer assay. 2 µg and 200 ng of total RNA were used per sample to prepare DRS and PCS library respectively. Libraries were loaded in PromethION R9 flow cells, and each sequencing run lasted 72 hours. Reads were base called concurrently by Guppy, where Q score > 7 were kept as passed reads. Reads were subsequently mapped to either reference genome or transcriptome via minimap2, using nanopore sequencing specific setting. Gene expression levels were determined by featurecounts and Salmon, followed by differential gene expression analysis by DESeq2, and differential transcript usage analysis performed by DRIMseq and DEXseq.

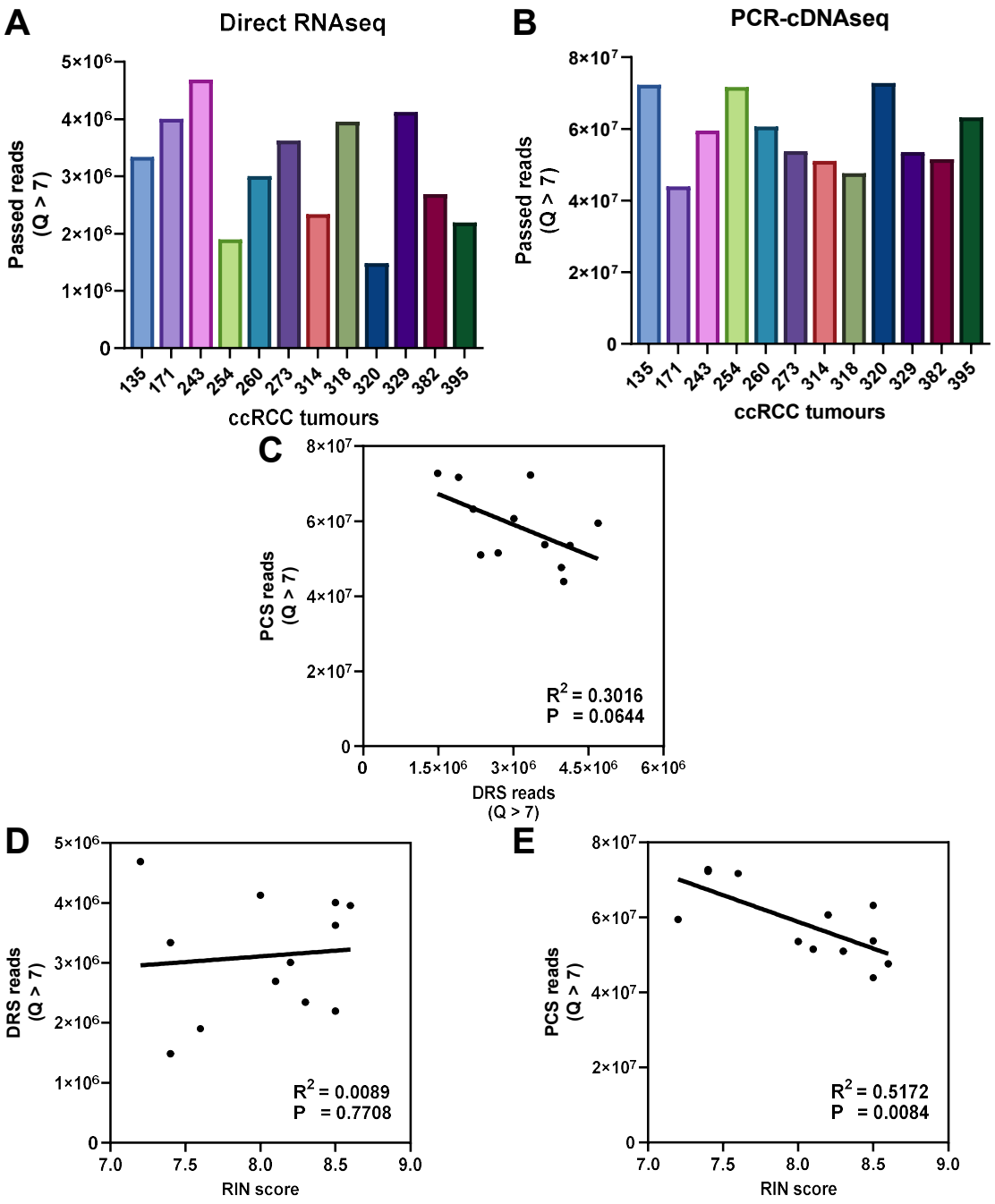


Figure 3.4: Summary of DRS and PCS reads generated from ccRCC samples

A) Bar graphs showing the number of passed reads (Q score > 7) generated by direct RNAseq (RNA002). **B)** Bar graphs showing the number of passed reads (Q score > 7) generated by direct RNAseq (RNA002) PCR-cDNaseq (PCS111) for each ccRCC nephrectomy samples (n = 12) using PromethION flow cells (R 9.4.1). **C)** Correlation between number of PCS reads and DRS reads, **D)** Correlation between number of DRS reads and sample RIN scores **E)** Correlation between number of PCS reads and sample RIN scores. For C-E diagonal lines represent the line of best fit. R^2 values were computed to measure goodness-of-fit, and P values generated from F-test, with $p< 0.05$ considered statistically significant.

3.3.3 Evaluation of read lengths from DRS & PCS of ccRCC tumours

The read length distributions of passed reads generated by DRS and PCS were analysed and visualised by *Nanoplot*. Violin plots showing Log_{10} transformed raw read lengths from DRS (Figure 3.5A) and PCS (Figure 3.5B) were plotted, with corresponding RNA sample RIN score, mean read quality Q score, and the mean and median read length for each sample indicated below the graphs. Q score (or Phred-scaled quality score) is calculated by $Q = -10\log_{10}P$, where P denotes the probability of base-calling error. Thus, a Q score of 10 represents 90% base-calling accuracy. Mean Q scores for DRS reads ranged between 9.9 and 10.7 (median: 10.45). Mean and median DRS passed-reads read length ranged between 287 – 515nt and 248 – 447nt, with median lengths of 420nt and 365.5nt, respectively. Mean Q scores for PCS reads ranged between 10.1 and 12.7 (median: 12.05). Mean and median PCS passed-reads read length ranged between 564 – 959nt and 410 – 794nt, with median lengths of 754nt and 657nt, respectively.

To evaluate similarities and differences between DRS- and PCS-generated reads, a scatter plot between the mean length of DRS- and PCS-generated reads from each tumour sample was produced, with no significant correlation found (R^2 = 0.0431 and P = 0.5175) (Figure 3.6A). Furthermore, no significant correlations were observed between mean read Q scores and RNA sample RIN scores for either DRS (R^2 = 0.0457 and P = 0.5045) or PCS (R^2 =0.0356 and P = 0.5568) (Figure 3.6B).

Next, relationships between DRS and PCS read lengths and corresponding sample Q scores and RIN scores were assessed. Whilst there was no significant degree of concordance observed between mean lengths of DRS reads and corresponding RNA sample RIN scores (R^2 = 0.0782 and P = 0.3788) (Figure 3.6C), a strong positive correlation was found between mean lengths of PCS reads and corresponding RNA sample RIN scores (R^2 = 0.6651 and P = 0.0012) (Figure 3.6D). Finally, a significant correlation was observed between mean DRS read lengths read Q scores (R^2 = 0.6234 and P = 0.0023) (Figure 3.6E). However, no significant correlation was found between mean PCS read lengths and Q scores (R^2 = 0.2208 and P = 0.1232) (Figure 3.6F).

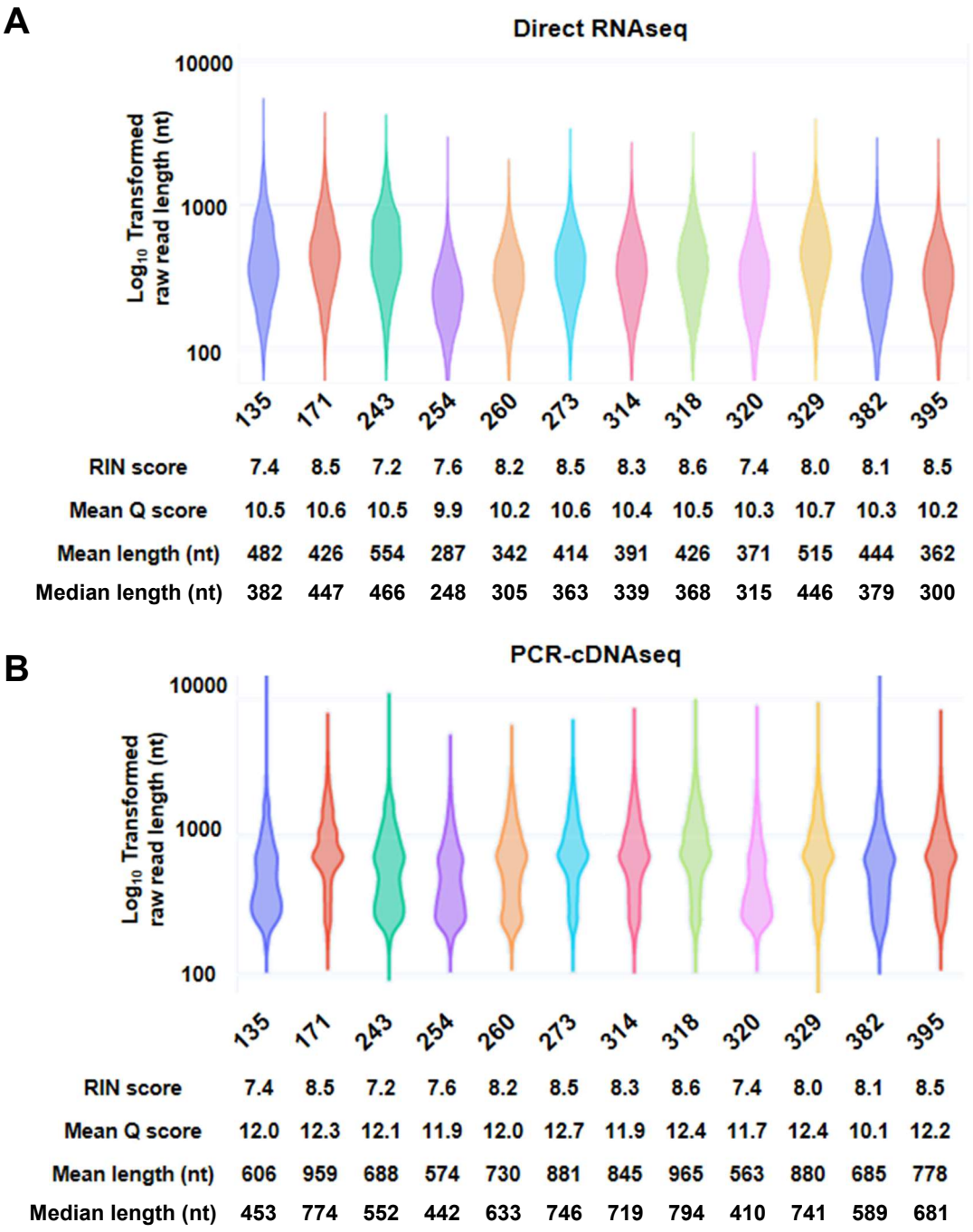


Figure 3.5: Distribution of raw read lengths from DRS and PCS of ccRCC tumours

A) Violin plot showing Log10 transformed raw read lengths from Direct RNAseq. RIN score, mean read Q score (Read basecall quality score) and mean read length for each sequencing dataset are listed in the tables below violin graphs.

B) As in **A** but for PCR-cDNAseq sequencing data.

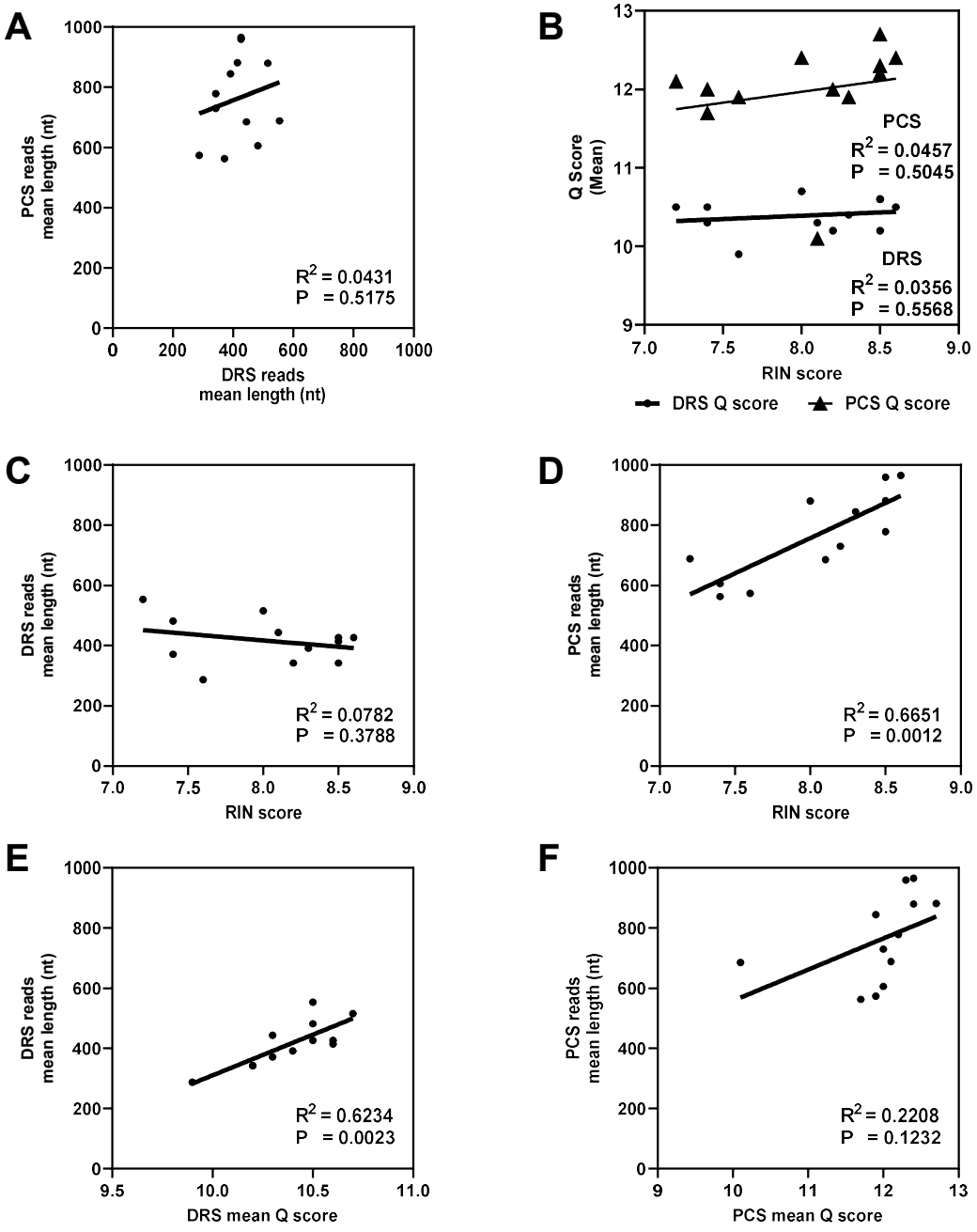


Figure 3.6: Relationships between PCS and DRS read lengths, Q scores and RIN scores

A) Correlation between mean PCS and DRS read lengths, **B)** correlation between mean PCS and DRS read basecall quality scores (Q scores) and sample RIN scores, **C)** correlation between mean DRS read lengths and sample RIN scores, **D)** correlation between mean PCS read lengths and sample RIN scores, **E)** correlation between mean DRS read lengths and mean read Q scores, **F)** correlation between mean PCS read lengths and mean read Q scores. Throughout, diagonal lines represent the line of best fit. R^2 values were computed tomeasure goodness-of-fit, and P values generated from F-test, with $p< 0.05$ considered statistically significant.

3.3.4 Statistics of the reference genome and transcriptome aligned ccRCC tumour DRS and PCS reads

After processing and filtering raw reads, DRS and PCS reads were aligned to either reference genome (Ensembl release 105, Genome assembly version: GRCh38) or reference transcriptome (Ensembl release 105, cDNA reference) by the sequence mapping and aligner minimap2. The read length distributions of genome-aligned reads from DRS and PCS of ccRCC tumours were analysed and visualised by Nanoplot. Violin plots showing Log_{10} transformed genome-aligned read lengths from DRS (Figure 3.7A) and PCS (Figure 3.7B) were plotted, with mean and median read lengths for each sample indicated below the graphs. Mean and median reference genome-aligned DRS read lengths ranged between 342 – 595nt and 301 – 507nt, with median lengths of 466nt and 404.5nt, respectively. Mean and median reference genome-aligned PCS read lengths ranged between 482 – 779nt and 372 – 621nt, with median lengths of 613nt and 507.5nt, respectively.

Reference transcriptome-aligned reads were analysed by bamslam, which provided summary alignment statistics for each DRS and PCS sample (Tables 3.2 and 3.3). The median length of reference transcriptome-aligned reads ranged between 258 – 470nt (median: 387nt) for DRS and 446 – 616nt (median: 517nt) for PCS. Median read mapping accuracy (base identity) for DRS was ~90%, and ~95% for PCS, except 382 at 91.97%. Between 2.7 – 18% of the reference transcriptome aligned DRS reads and 21.0 – 37.1% PCS reads represent full-length transcript (95%+ coverage of the length of aligned reference transcript). The median transcript coverage per aligned DRS read ranged between 20.9 – 40.6% and 41.7 – 77.4 % for each aligned PCS read. Data here shows that despite RNA degradation in the samples, Nanopore long-read RNAseq can still detect full length transcripts and provide high levels of transcript coverage.

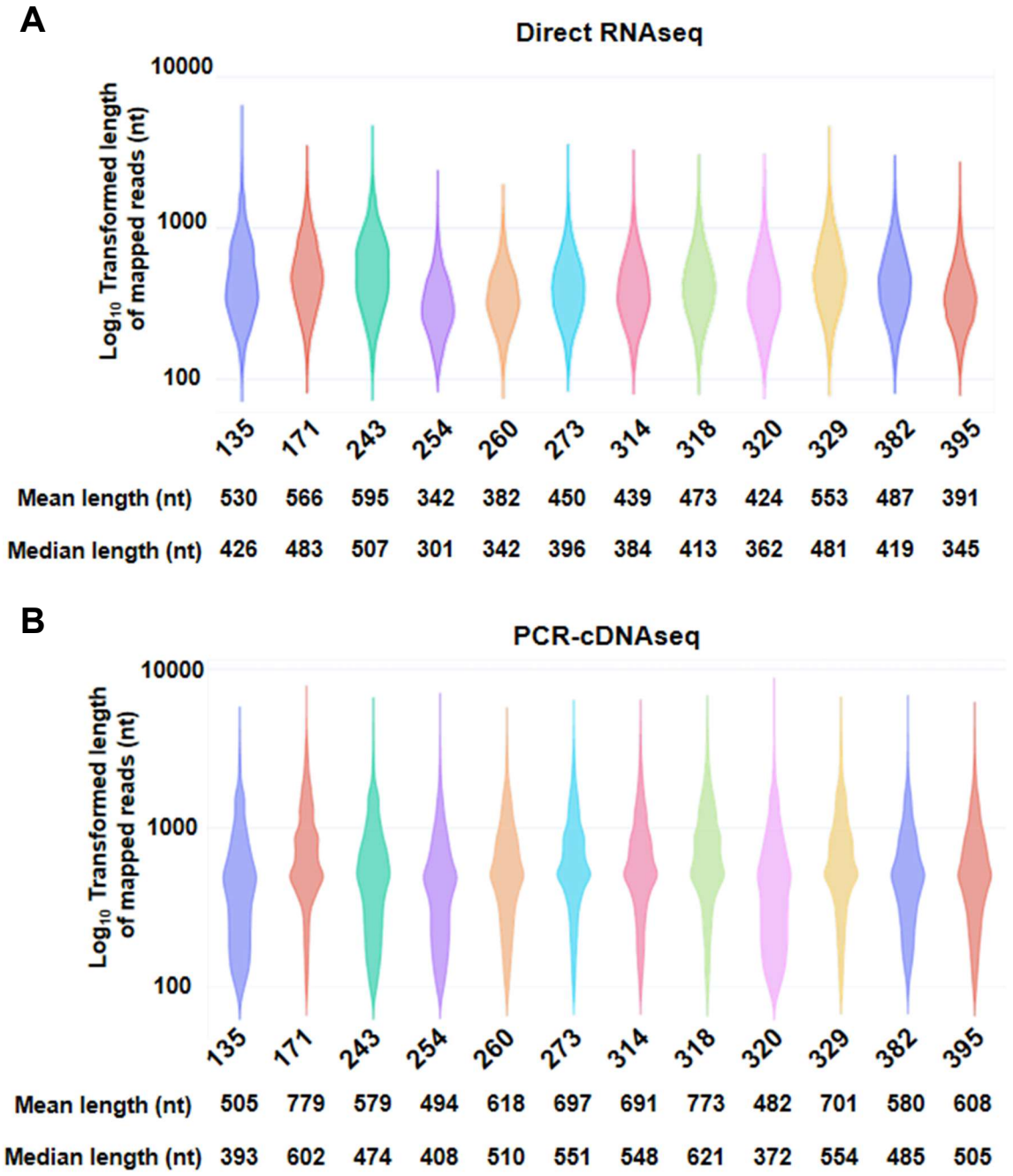


Figure 3.7: Distribution of reference genome aligned read lengths from DRS and PCS of ccRCC tumours

A) Violin plots showing Log10 transformed reference genome aligned (Ensembl release 105, Genome assembly version: GRCh38) read lengths for Direct RNAseq. Mean and median read lengths for each sequencing dataset are listed in the tables below violin graphs.

B) As in **A**, but for PCR-cDNAseq.

Direct RNAseq	135	171	243	254	260	273	314	318	320	329	382	382
Median read-alignment lengths (nt)	455	470	519	258	299	381	352	394	346	460	400	400
Median accuracy of read alignments (%)	90.09	91.01	89.82	90.74	90.75	90.95	90.65	90.48	90.00	90.63	90.53	90.53
Reads representing full-length transcripts (%)	15.82	11.34	18.08	2.76	3.15	7.2%	5.94	7.40	7.80	10.34	8.14	8.14
Median coverage of transcript per read (%)	40.61	28.98	45.55	20.87	22.12	26.82	27.72	29.26	34.68	36.45	31.51	31.51

Table 3.2: Read alignment statistics from Direct RNAseq of ccRCC tumours

Statistics related to alignment of Direct RNAseq generated reads to the reference transcriptome (Ensembl release 105, cDNA reference), including median lengths of read-alignment, median accuracy of reads, percentage of reads which represent full-length transcripts (covering at least 95% of annotated transcript where the read was aligned), and median coverage of annotated transcript per mapped read.

PCR-cDNAseq	135	171	243	254	260	273	314	318	320	329	382	382
Median read-alignment lengths (nt)	461	554	519	447	515	552	616	539	446	552	490	510
Median accuracy of read alignments (%)	95.28	95.59	95.20	95.88	95.52	95.97	95.49	95.00	95.21	95.52	91.97	95.41
Reads representing full-length transcripts (%)	22.70	37.13	25.15	21.77	25.08	34.22	36.36	30.30	21.02	31.74	25.37	26.40
Median coverage of transcripts per read (%)	46.59	84.28	54.52	41.74	56.01	72.03	77.35	66.50	42.92	68.51	54.34	58.30

Table 3.3: Read alignment statistics from PCR-cDNAseq of ccRCC tumours

Statistics related to alignment of PCR-cDNAseq generated reads to the reference transcriptome (Ensembl release 105, cDNA reference), including median lengths of read-alignment, median accuracy of reads, percentage of reads which represent full-length transcripts (covering at least 95% of annotated transcript where the read was aligned), and median coverage of annotated transcript per mapped read.

3.3.5 Relationship between DRS and PCS read alignment lengths and transcript coverage

To evaluate the relationship between reference transcriptome alignment by DRS and PCS-generated reads, a scatter plot between median DRS and PCS reference transcriptome aligned read lengths from each tumour sample was plotted, with no significant correlation found (R^2 = 0.0614 and P = 0.4375) (Figure 3.8A).

Next, correlations between DRS and PCS reference transcriptome aligned read lengths and percentage of transcript coverage were assessed. Firstly, proportions of reads that represent full-length transcript (95%+ coverage of the length of aligned reference transcript) were found to positively correlate with both the median DRS aligned read lengths (R^2 = 0.8409 and P < 0.0001) and median PCS aligned read lengths (R^2 = 0.8381 and P < 0.0001) (Figure 3.8B – C). Furthermore, the median transcript coverage per aligned read was also found to be strongly correlated with both the median aligned read lengths of DRS (R^2 = 0.6818 and P = 0.0009) and PCS (R^2 = 0.8323 and P < 0.0001) (Figure 3.8 D – E).

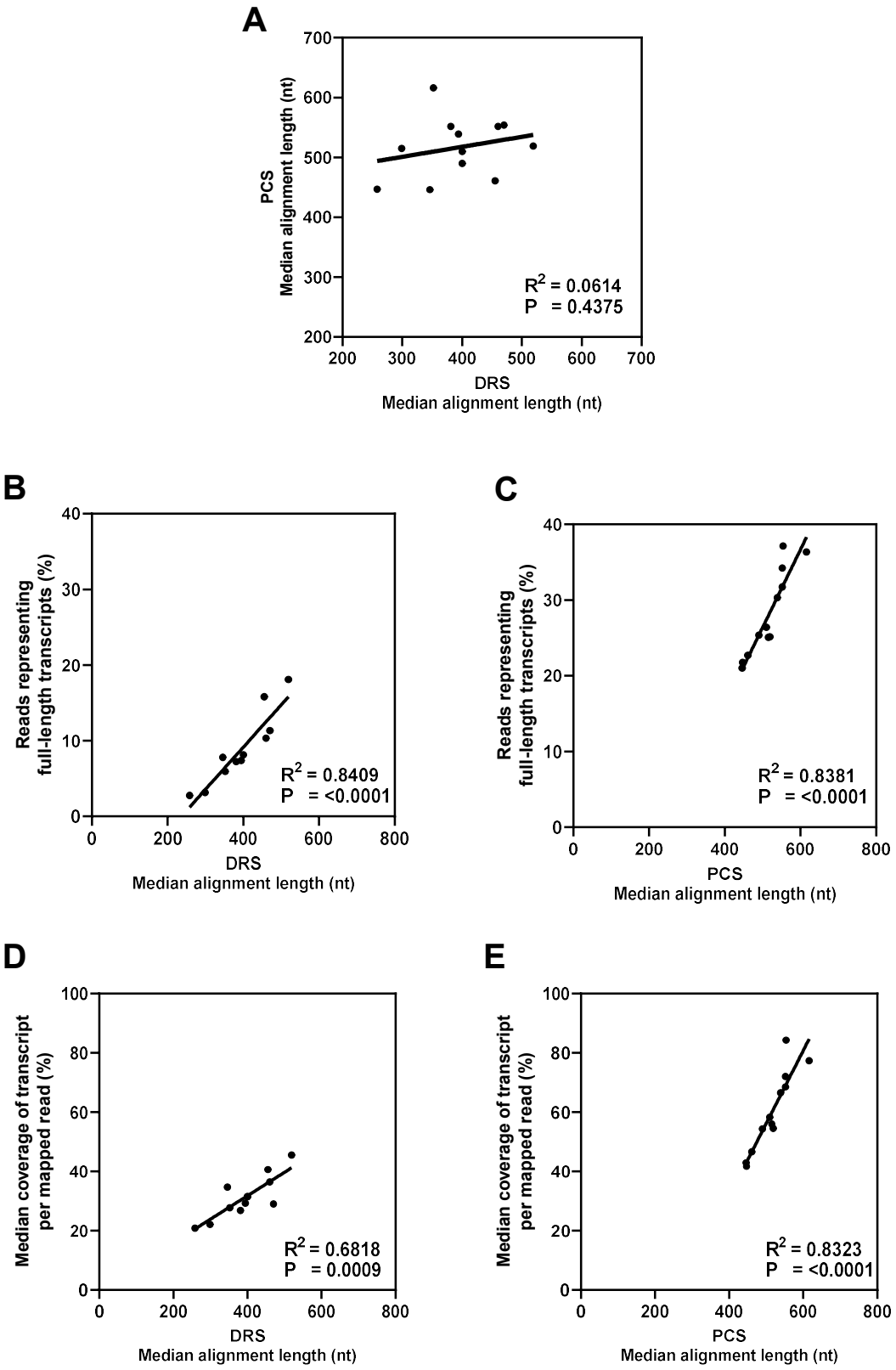

A
PCS
Median alignment length (nt)
DRS
Median alignment length (nt)
R2 = 0.0614
P = 0.4375
B
Reads representing full-length transcripts (%)
DRS
Median alignment length (nt)
R2 = 0.8409
P = <0.0001
C
Reads representing full-length transcripts (%)
PCS
Median alignment length (nt)
R2 = 0.8381
P = <0.0001
D
Median coverage of transcript per mapped read (%)
DRS
Median alignment length (nt)
R2 = 0.6818
P = 0.0009
E
Median coverage of transcript per mapped read (%)
PCS
Median alignment length (nt)
R2 = 0.8323
P = <0.0001

Figure 3.8: Correlations between read alignment lengths and coverage

A) Correlation between median PCS and DRS read alignment lengths, **B)** correlation between median DRS read alignment lengths and percentage of reads which represent full-length transcripts (covering at least 95% of annotated transcript), **C)** correlation between median PCS read alignment lengths and percentage of reads which represent full-length transcripts (covering at least 95% of annotated transcript), **D)** correlation between median DRS read alignment lengths and median coverage of annotated transcript per mapped read, **E)** correlation between median PCS read alignment lengths and median coverage of annotated transcript per mapped read. Throughout, diagonal lines represent the line of best fit. R^2 values were computed to measure goodness-of-fit, and P values generated from F-test, with $p < 0.05$ considered statistically significant.

3.3.6 Composition of RNA Biotypes from reference genome aligned DRS and PCS of ccRCC tumour

Tumour tissues are comprised of numerous cell types with highly varied transcriptome profiles. To evaluate the ability of long-read sequencing to capture transcriptomic diversity using RNA extracted from snap frozen tissue samples, bioinformatic analysis on the RNA biotypes (Ensembl) of DRS- and PCS-identified unique genes from ccRCC tumours was performed.

Firstly, pie charts depicting the proportions of RNA biotypes of genes discovered by reference genome (Ensembl release 105, GRCh38) aligned DRS reads from each ccRCC tumour sample were produced (Figure 3.9A). Tumour samples exhibited similar composition of RNA biotypes, and the average RNA biotype proportion of identified unique genes across the 12 tumour tissues is displayed in Figure 3.9B. The majority of identified genes are classified as protein-coding (78.57%), followed by lncRNA (15.39%), processed pseudogenes (3.11%) and transcribed unprocessed pseudogenes (1.25%). The total number of unique genes identified per sample ranged from 15,854 to 19,371, with a median of 18,057.

Next, pie charts were generated to illustrate the RNA biotype profile of genes discovered by reference genome-aligned PCS reads from each ccRCC tumour sample and the average profile between tumours (Figure 3.10A - B). Similar to DRS, RNA biotype profiles are highly similar between tumours. However, compared to DRS, reference genome-aligned PCS reads discovered a higher number of non-protein-coding genes. On average, 60.25% of identified genes are protein-coding, followed by lncRNA (25.35%), processed pseudogenes (6.64%), transcribed unprocessed pseudogenes (1.61%), and unprocessed pseudogenes (1.04%) (Figure 3.10B). The numbers of unique genes identified per tumour sample were also higher by PCS, ranging between 25207 and 27071, with a median of 26203 genes (Figure 3.10C).

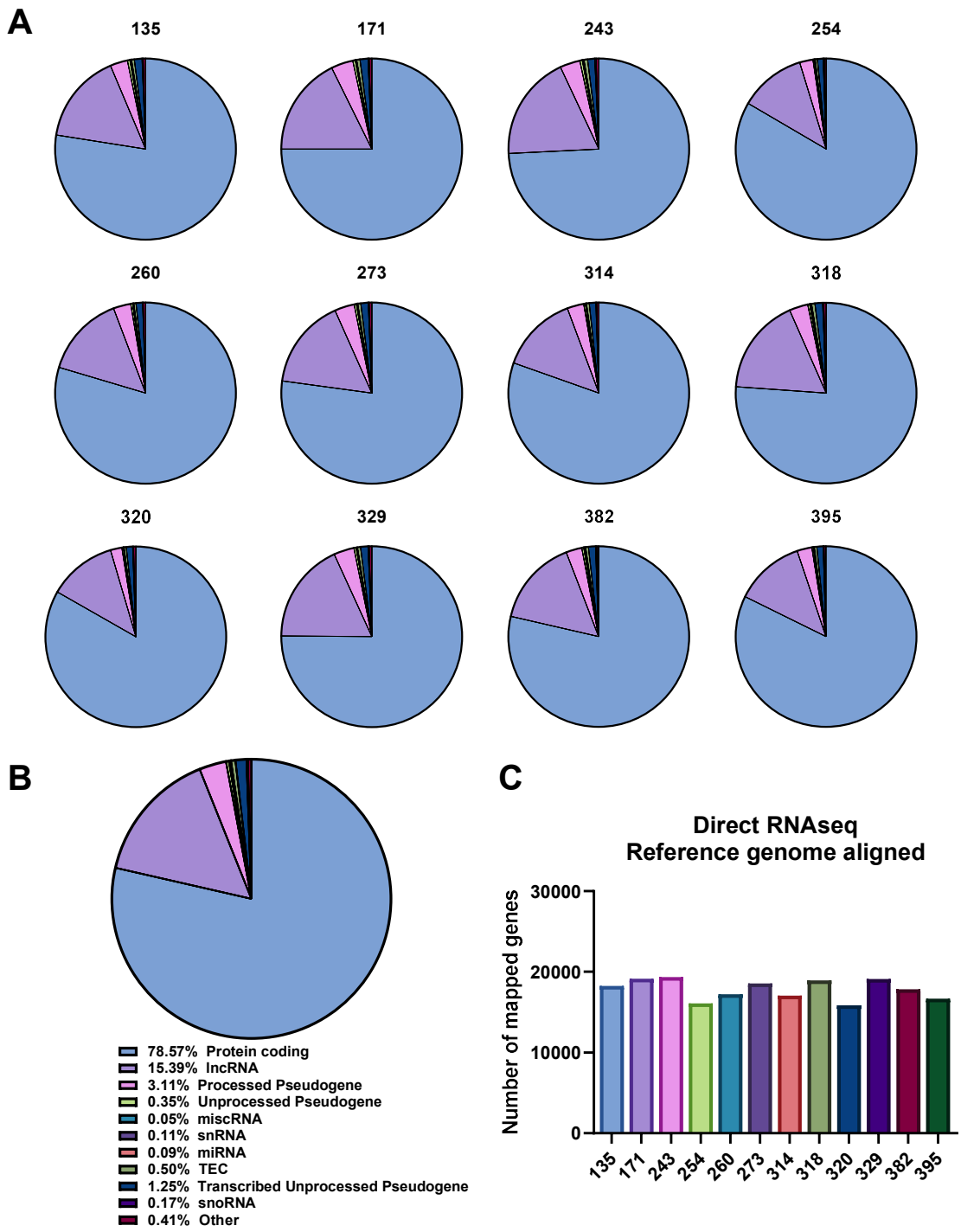


Figure 3.9: RNA biotype composition of ccRCC tumours by DRS aligned to the human reference genome

A) Pie charts depicting the proportions of gene biotypes of reference genome (Ensembl release 105, GRCh38) aligned DRS reads from each ccRCC tumour sample. **B)** Pie chart depicting the average proportions of gene biotypes of reference genome mapped DRS reads from all ccRCC tumour sample. **C)** Bar graphs showing the number of unique genes identified from reference genome aligned DRS of ccRCC tumours.

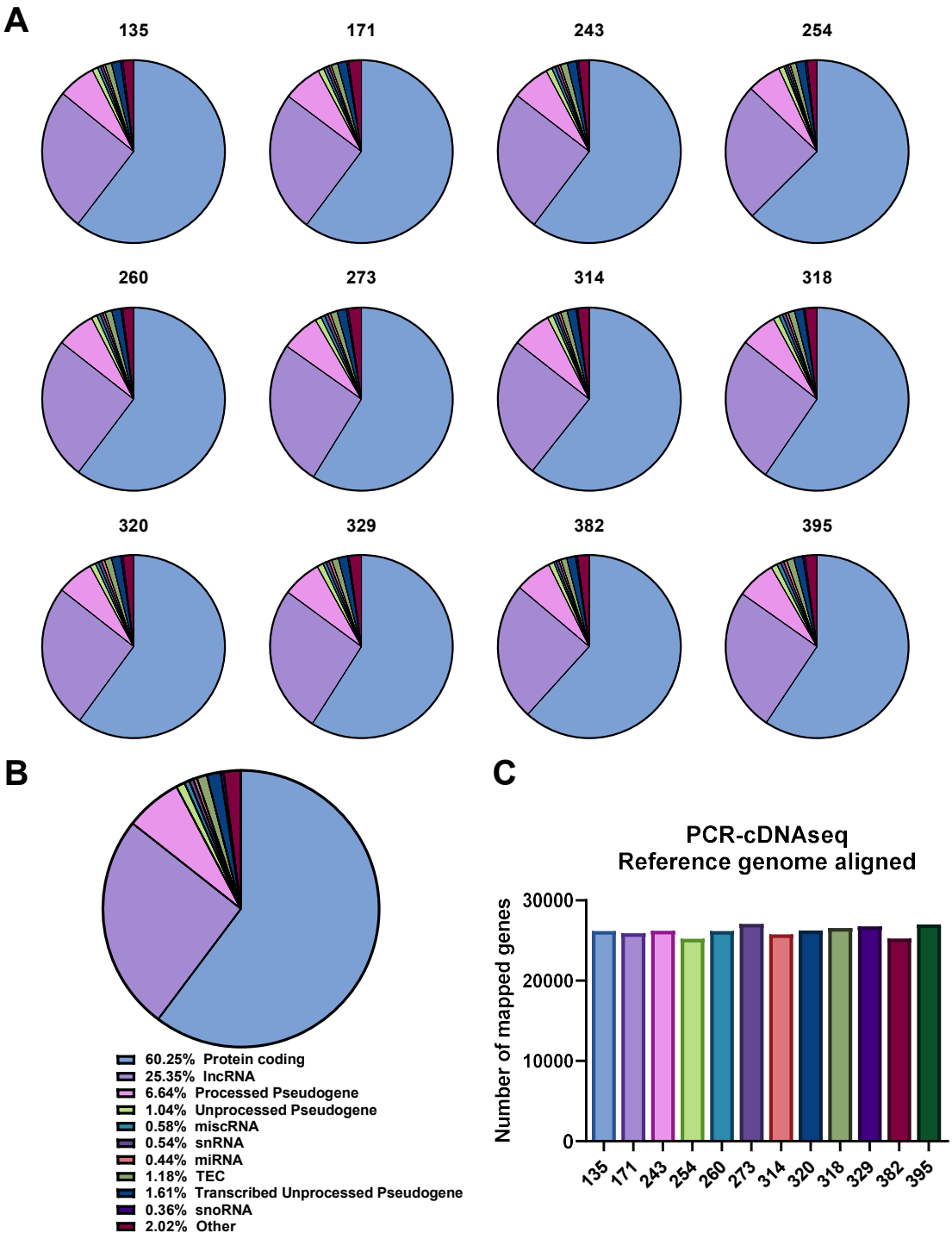


Figure 3.10: RNA biotype composition of ccRCC tumours by PCS aligned to the human reference genome

A) Pie charts depicting the proportions of gene biotypes of reference genome (Ensembl release 105, GRCh38) aligned PCS reads from each ccRCC tumour sample. **B)** Pie charts depicting the average proportions of gene biotypes of reference genome aligned PCS reads from all ccRCC tumour sample. **C)** Bar graphs showing the number of unique genes identified from reference genome aligned PCS of ccRCC tumours.

3.3.7 Composition of RNA Biotypes from reference transcriptome aligned DRS and PCS of ccRCC tumour

Following the assessment of reference genome-aligned DRS and PCS, the biotype profiles of reference transcriptome-aligned DRS and PCS of ccRCC tumours were explored. Reference transcriptome includes more than 200,000 curated transcripts compiled from transcriptional evidence from cDNA expression libraries, expressed sequence tags and RNAseq data (Aken *et al.*, 2016). Recent advancement in bioinformatics tools has enabled the quantification of transcript isoform-level expression by aligning reads against the reference transcriptome. To understand the composition of mapped transcripts, pie charts depicting the proportions of RNA biotypes of genes mapped by reference transcriptome (Ensembl release 105, cDNA reference) aligned DRS reads from each ccRCC tumour sample and the average profile were produced (Figure 3.11A - B). Like genome-aligned DRS, tumour samples exhibited similar composition of RNA biotypes (Figure 3.11A). However, since the reference cDNA is devoid of non-coding RNA, the majority of identified genes are protein-coding (94.63% on average), followed by transcribed unprocessed pseudogenes (1.99% on average) and processed pseudogenes (1.53% on average) (Figure 3.11B). The number of unique genes identified per sample ranged between 9,165 and 13,194 (Figure 3.11C).

For PCS, Pie charts depicting the proportions of RNA biotypes of genes discovered by reference transcriptome (Ensembl release 105, cDNA reference) aligned reads from each ccRCC tumour and the average biotype profile are displayed in Figure 3.12A – B. Whilst more non-protein-coding genes were discovered by PCS compared to DRS, the proportions are substantially smaller than reference genome aligned PCS or DRS. On average, 86.93% of identified genes are classified as protein-coding, 6.20% are processed pseudogenes, 2.84% are transcribed unprocessed pseudogenes, 1.38% are unprocessed pseudogenes, and 1.11% are transcribed processed pseudogenes (Figure 3.12B). The number of unique genes identified per tumour sample was also higher by PCS, ranging between 18,151 and 29,289, with a median of 18,677 genes (Figure 3.12C).

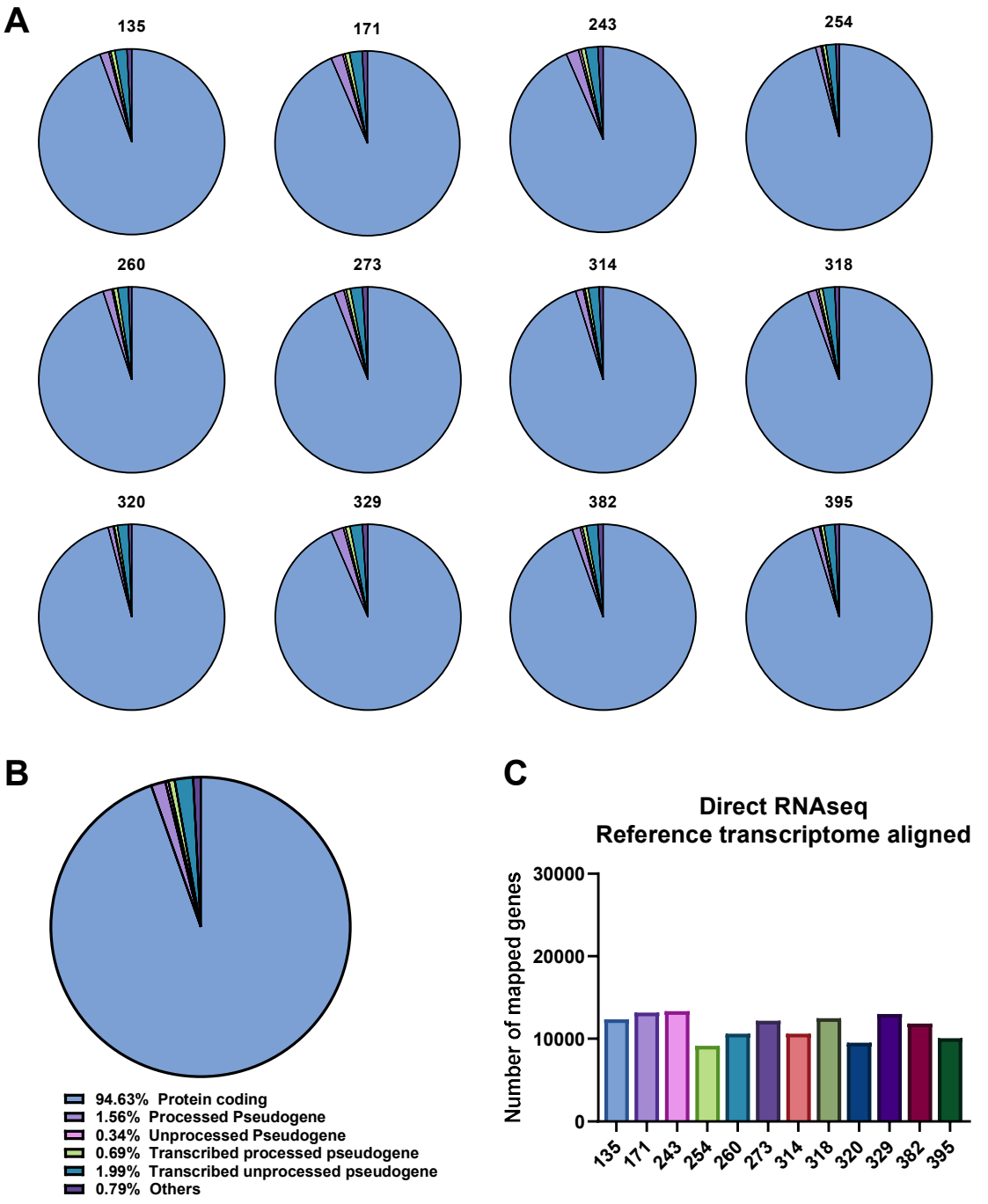


Figure 3.11: RNA biotype composition of ccRCC tumours by DRS aligned to the human reference transcriptome

A) Pie charts depicting the proportions of gene biotypes of reference transcriptome (Ensembl release 105, cDNA reference) aligned DRS reads from each ccRCC tumour sample. **B)** Pie chart depicting the average proportions of gene biotypes of reference genome aligned DRS reads from all ccRCC tumour sample. **C)** Bar graphs showing the number of unique genes identified from reference transcriptome aligned DRS of ccRCC tumours.

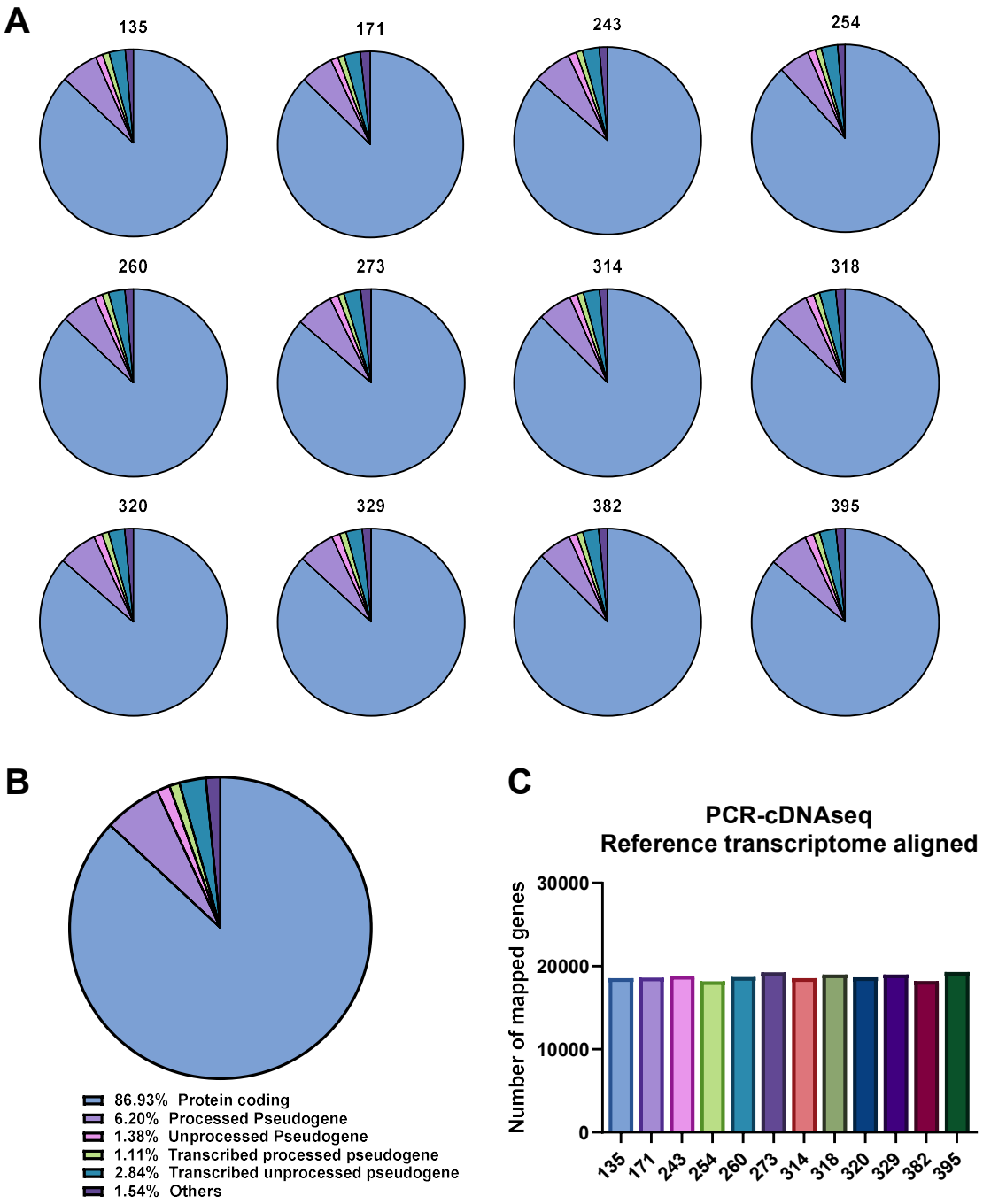


Figure 3.12: RNA biotype composition of ccRCC tumours by PCS aligned to the human reference transcriptome

A) Pie charts depicting the proportions of gene biotypes of reference transcriptome (Ensembl release 105, cDNA reference) aligned PCS reads from each ccRCC tumour sample. **B)** Pie chart depicting the average proportions of gene biotypes of reference genome aligned PCS reads for all ccRCC tumour samples. **C)** Bar graphs showing the number of unique genes identified from reference transcriptome aligned PCS of ccRCC tumours.

3.3.8 Abundance of RNA biotypes in ccRCC tumours

Next, bioinformatics analysis was conducted to investigate the breakdowns of DRS- and PCS-identified RNA biotypes in ccRCC tumours by gene expression levels. Expression levels of identified genes from both reference genome and transcriptome-aligned DRS/PCS were normalised to the library size and expressed as reads per million mapped reads (RPM). RPM was used instead of reads per kilobase of transcript per million reads mapped (RPKM). RPKM is helpful for gene expression normalisation for short-read RNAseq since generated reads usually fail to span a substantial part of a transcript. Hence, transcript length must be considered to compare gene expression between genes and samples fairly. However, in the long-read RNAseq protocols used in this study, reads always span from poly(A) tail, from 3' to 5'. Thus, each read from DRS and PCS represents one mRNA molecule, and gene length normalisation is unsuitable for estimating gene expression.

Pie charts depicting the averaged proportions of RNA biotypes of the reference genome and transcriptome-aligned DRS and PCS reads by expression levels (RPM) were constructed (Figure 3.13A – D). When aligned with the reference genome, most DRS and PCS reads were mapped to protein-coding genes (91.7% and 89.41%, respectively), followed by mitochondrial ribosomal RNA (mt-rRNA), at 4.66% for DRS and 5.35% for PCS. Only 1.10% and 1.71% of expressed genes from DRS and PCS reads were classified as lncRNA (Figure 3.13A, C). In contrast, more than 98% of reference transcriptome-aligned DRS and PCS reads were mapped to protein-coding genes. The next largest biotype group from both reference transcriptome-aligned DRS and PCS is ribosomal RNA pseudogene, constituting 0.36% and 0.88% of reads by expression levels, respectively (Figure 3.13B, D).

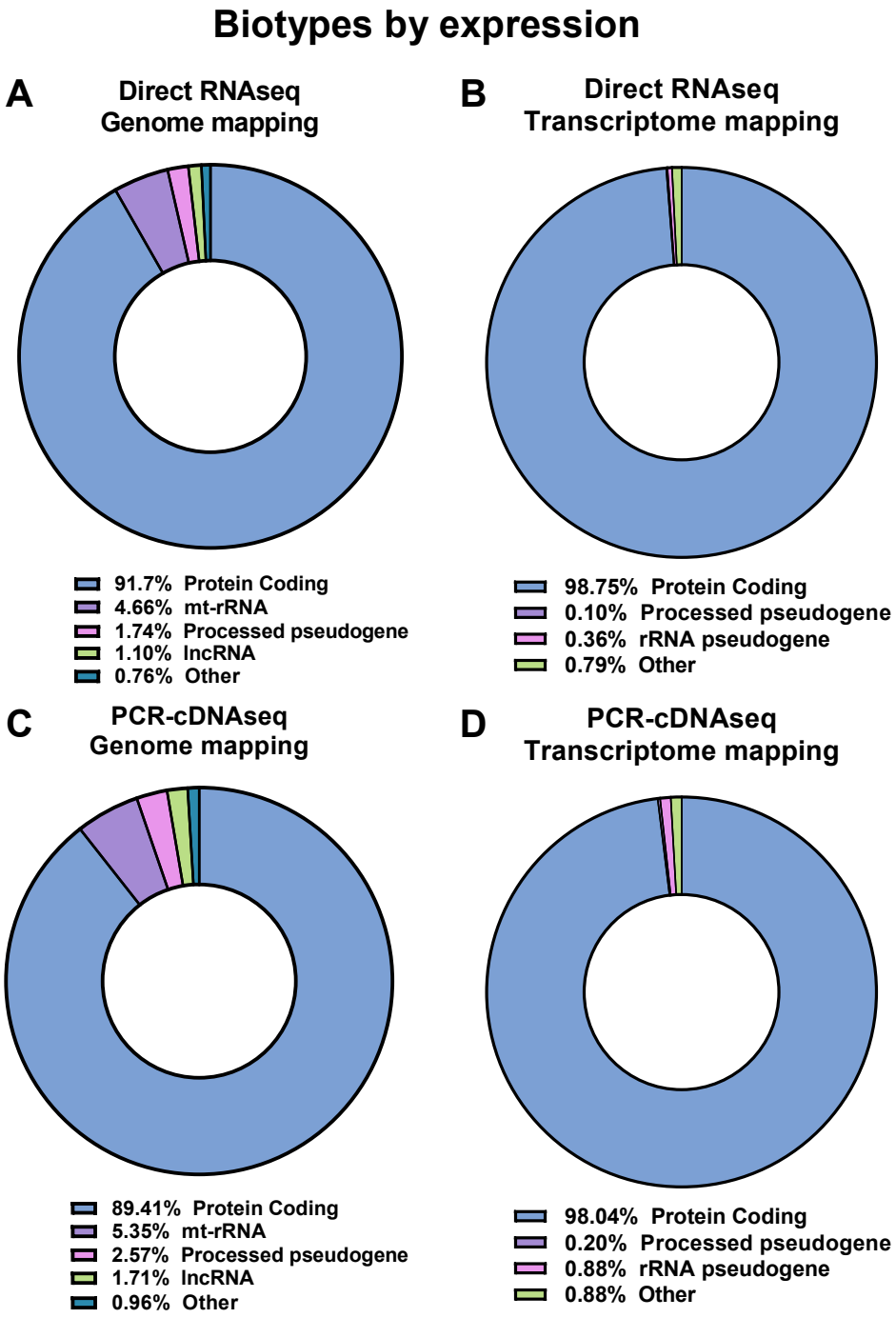


Figure 3.13: RNA biotype composition of ccRCC tumours by expression levels

A) Pie chart depicting the average proportions of gene biotypes of reference genome (Ensembl release 105, GRCh38) mapped DRS reads by expression levels (scaling to library size, RPM) from ccRCC tumour sample. **B)** Pie chart depicting the average proportions of gene biotypes of reference transcriptome (Ensembl release 105, cDNA reference) mapped DRS reads by expression levels (scaling to library size, RPM) from ccRCC tumour sample. **C)** As in **A**, but for PCR-cDNAseq. **D)** As in **B**, but for PCR-cDNAseq.

3.3.9 Overlapping genes identified from reference genome aligned DRS and PCS of ccRCC tumours

Combining all unique genes identified by reference genome-aligned DRS and PCS of the 12 ccRCC tumours, pie charts depicting the proportions of RNA biotypes for DRS and PCS were produced (Figure 3.14A – B). A total number of 26,457 genes were discovered by DRS, where 32.48% of genes are classified as protein coding, 28.81% are lncRNAs, and 16.5% are processed pseudogenes. In comparison, PCS identified 39115 unique genes, with 45.47% being protein-coding genes, 29.48% being lncRNAs, and 12.34% being processed pseudogenes. Surprisingly, 3.10% and 3.05% of DRS genes and 1.32% and 1.15% of PCS genes are small nuclear RNAs (snRNAs) and microRNAs (miRNAs), which are not usually polyadenylated when matured.

Next, the extent of overlap between reference genome-aligned DRS- and PCS-identified genes from ccRCC tumours was determined. Of the 26,457 unique genes identified by DRS, 25,692 genes were also found by PCS (Figure 3.14C). 765 genes were identified only by DRS, and 13423 genes were identified exclusively by PCS, as shown in the Venn diagram in Figure 3.14C. The majority of the genes commonly identified by both reference genome-aligned DRS and PCS are protein-coding (63.08%), followed by lncRNA (24.21%) and processed pseudogenes (6.36%) (Figure 3.14E). Of the 765 genes that DRS exclusively identified, 67.59% are lncRNA, and 16.54% are processed pseudogenes (Figure 3.14D). Of the 13,423 genes PCS exclusively found, 39.60% are lncRNA, 23.83% are processed pseudogenes, and 11.67% are protein-coding genes (Figure 3.14F). Finally, to understand if genes were mapped exclusively by DRS/PCS due to sequencing depth, violin plots of gene expression level profiles of DRS-exclusive, commonly found, and PCS-exclusive genes were produced (Figure 3.14G – I). Genes that were only found by either reference genome-aligned DRS or PCS are substantially lower expressed than genes that both DRS and PCS identified. The median RPM of DRS-exclusive and PCS-exclusive genes are 0.071 and 0.011, respectively, compared to 3.104 by DRS and 3.068 by PCS for commonly-identified genes.

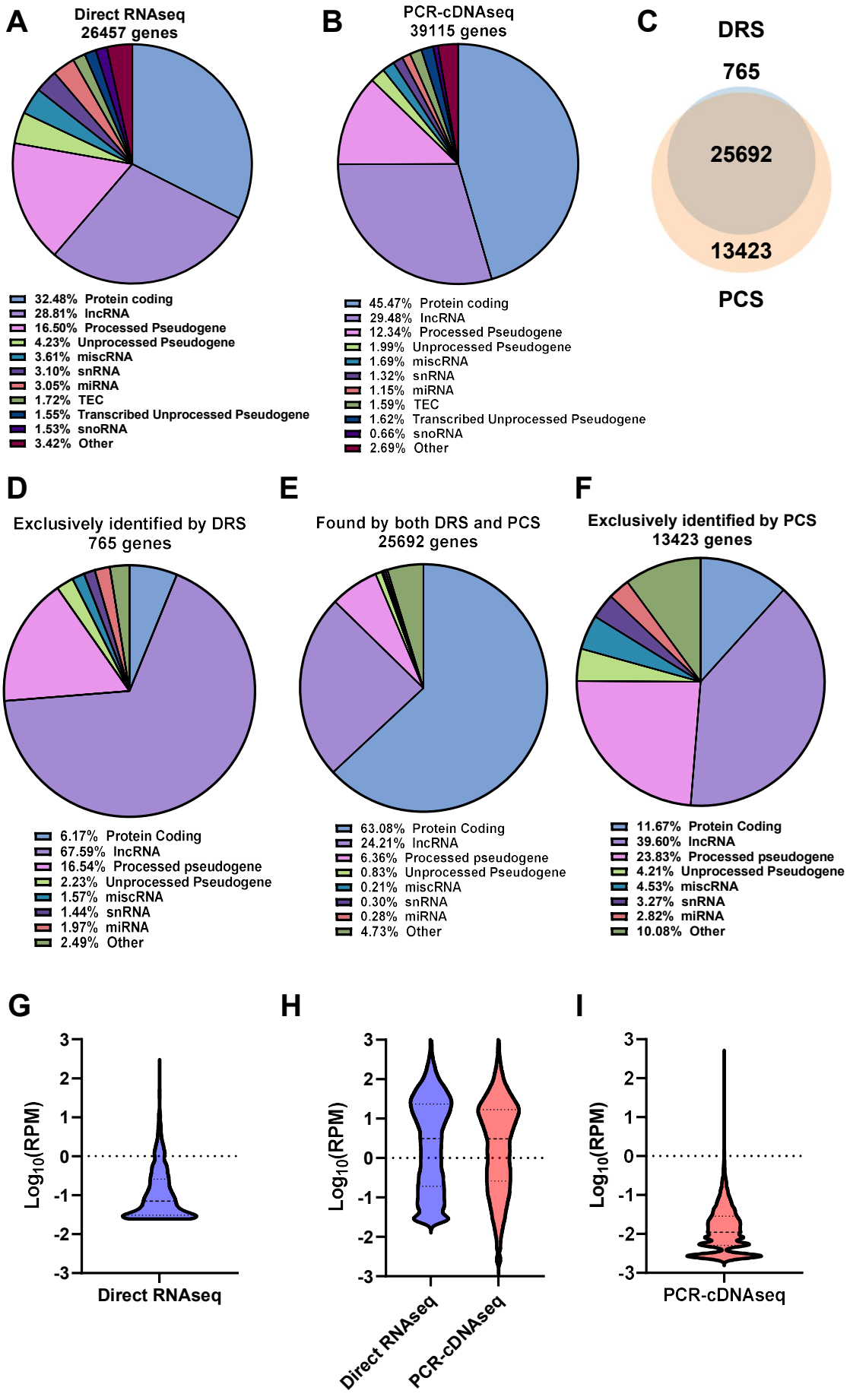

A
Direct RNAseq
26457 genes
32.48% Protein coding
28.81% lncRNA
16.50% Processed Pseudogene
4.23% Unprocessed Pseudogene
3.61% miscRNA
3.10% snRNA
3.05% miRNA
1.72% TEC
1.55% Transcribed Unprocessed Pseudogene
1.53% snoRNA
3.42% Other
B
PCR-cDNAseq
39115 genes
45.47% Protein coding
29.48% lncRNA
12.34% Processed Pseudogene
1.99% Unprocessed Pseudogene
1.69% miscRNA
1.32% snRNA
1.15% miRNA
1.59% TEC
1.62% Transcribed Unprocessed Pseudogene
0.66% snoRNA
2.69% Other
C
DRS
765
25692
13423
PCS
D
Exclusively identified by DRS
765 genes
6.17% Protein Coding
67.59% lncRNA
16.54% Processed pseudogene
2.23% Unprocessed Pseudogene
1.57% miscRNA
1.44% snRNA
1.97% miRNA
2.49% Other
E
Found by both DRS and PCS
25692 genes
63.08% Protein Coding
24.21% lncRNA
6.36% Processed pseudogene
0.83% Unprocessed Pseudogene
0.21% miscRNA
0.30% snRNA
0.28% miRNA
4.73% Other
F
Exclusively identified by PCS
13423 genes
11.67% Protein Coding
39.60% lncRNA
23.83% Processed pseudogene
4.21% Unprocessed Pseudogene
4.53% miscRNA
3.27% snRNA
2.82% miRNA
10.08% Other
G
Log10(RPM)
3
2
1
0
-1
-2
-3
Direct RNAseq
H
Log10(RPM)
Direct RNAseq
PCR-cDNAseq
I
Log10(RPM)
PCR-cDNAseq

Figure 3.14: Differences and common genes identified in ccRCC tumours by reference genome mapped reads from DRS and PCS

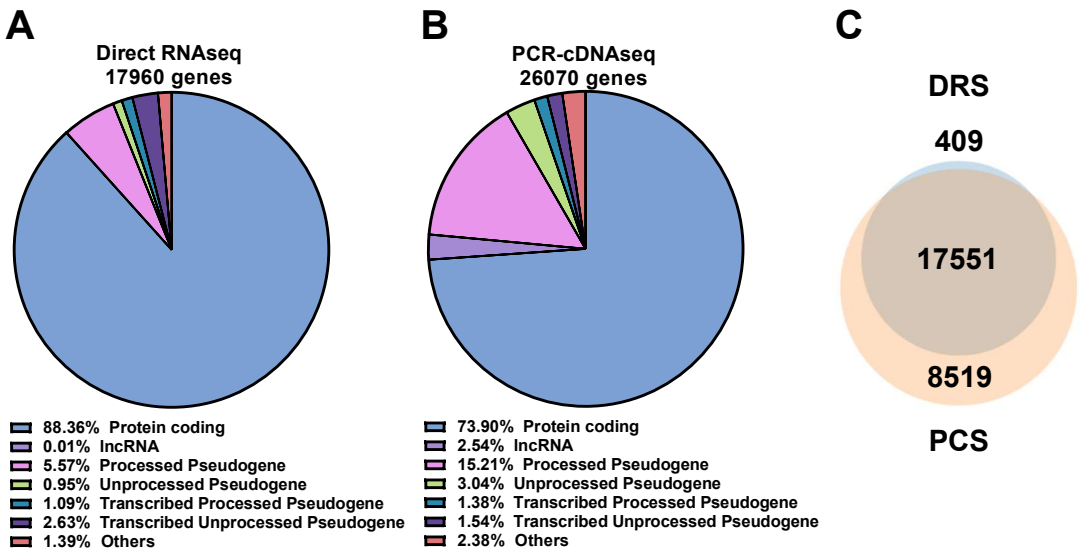
A) Pie chart depicting the proportions of gene biotypes of all reference genome (Ensembl release 105, GRCh38) mapped DRS reads from all ccRCC tumour sample. **B)** As in **A** but for PCS reads. **C)** Venn diagram showing the overlap between reference genome mapped DRS and PCS identified genes. **D)** Pie chart depicting the proportions of RNA biotypes of genes detected by reference genome-mapped DRS reads exclusively. **E)** As in **D**, but for genes identified by both DRS and PCS. **F)** As in **D**, but for genes detected by genome-mapped PCS reads exclusively. **G)** Violin plot depicting the distribution of gene expression levels (Log_{10} RPM) of genes detected by reference genome-mapped DRS reads exclusively. **H)** As in **G**, but for genes identified by both DRS and PCS. **I)** As in **G**, but for genes detected by reference genome-mapped PCS reads exclusively.

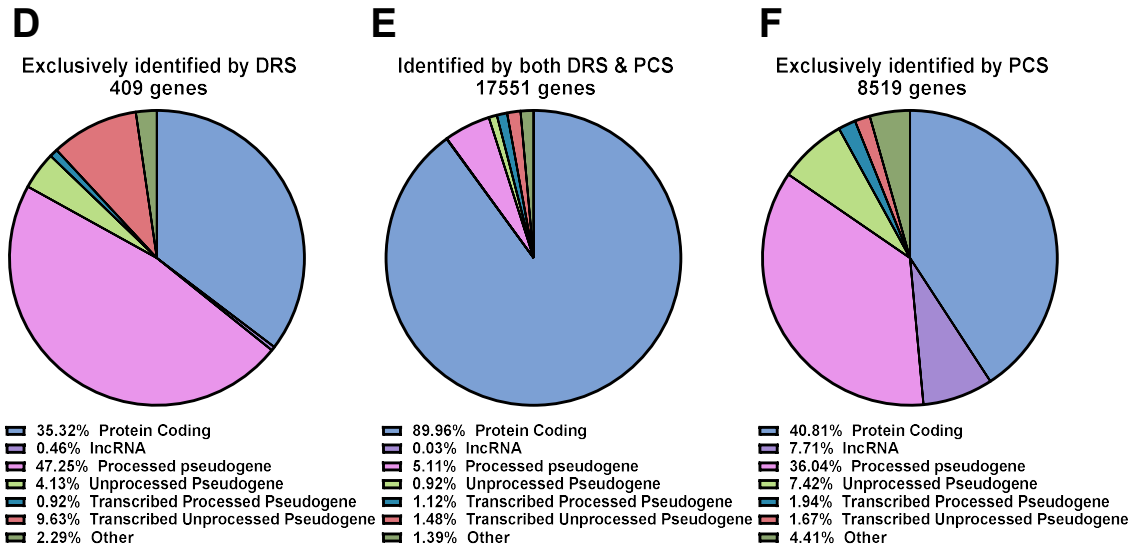
3.3.10 Overlapping genes identified from reference transcriptome aligned DRS and PCS of ccRCC tumours

Next, to evaluate the similarities and differences between the two sequencing methods, unique genes identified by reference transcriptome-aligned DRS and PCS of the 12 ccRCC tumours were profiled. Pie charts depicting the proportions of RNA biotypes for DRS and PCS-identified genes were generated (Figure 3.15A – B). 17,960 genes were discovered by reference transcriptome aligned DRS of ccRCC tumours, where 88.36% are protein-coding genes, and 5.57% are processed pseudogenes. Only 0.01% of all DRS-identified genes were classified as lncRNA. PCS identified 26,070 unique genes from ccRCC tumour samples, where 73.90% of identified genes are protein-coding genes, 15.21% are processed pseudogenes, and 2.54% are lncRNA. Of the 17,960 genes identified by reference transcriptome-aligned DRS, 17,551 were also identified in PCS. 409 genes were identified exclusively by DRS, and 13,423 genes were found only by PCS (Figure 3.15C).

Most of the genes identified by both reference transcriptome-aligned DRS and PCS are protein-coding (89.96%), with 5.11% of commonly found genes classified as processed pseudogenes (Figure 3.15E). In contrast, for genes that DRS exclusively identified, 47.25% are classified as processed pseudogenes, and 35.32% are protein-coding genes (Figure 3.15D). For genes that only PCS found, 40.81% are protein-coding genes, 36.04% are classified as processed pseudogenes, and 7.42% are lncRNAs.

Similar to what was observed in reference genome-aligned DRS and PCS, genes that were only identified by either reference transcriptome-aligned DRS or PCS are markedly lower expressed than genes identified by both methods. Violin plots of gene expression level profiles of DRS-exclusive, commonly found, and PCS-exclusive genes were generated (Figure 3.15G – I). The median RPM of DRS-exclusive and PCS-exclusive genes are 0.155 and 0.020, respectively, compared to 2.602 by DRS and 3.507 by PCS for commonly-identified genes.


A
Direct RNAseq
17960 genes
88.36% Protein coding
0.01% lncRNA
5.57% Processed Pseudogene
0.95% Unprocessed Pseudogene
1.09% Transcribed Processed Pseudogene
2.63% Transcribed Unprocessed Pseudogene
1.39% Others
B
PCR-cDNAseq
26070 genes
73.90% Protein coding
2.54% lncRNA
15.21% Processed Pseudogene
3.04% Unprocessed Pseudogene
1.38% Transcribed Processed Pseudogene
1.54% Transcribed Unprocessed Pseudogene
2.38% Others
C
DRS
409
17551
8519
PCS


D
Exclusively identified by DRS
409 genes
35.32% Protein Coding
0.46% lncRNA
47.25% Processed pseudogene
4.13% Unprocessed Pseudogene
0.92% Transcribed Processed Pseudogene
9.63% Transcribed Unprocessed Pseudogene
2.29% Other
E
Identified by both DRS & PCS
17551 genes
89.96% Protein Coding
0.03% lncRNA
5.11% Processed pseudogene
0.92% Unprocessed Pseudogene
1.12% Transcribed Processed Pseudogene
1.48% Transcribed Unprocessed Pseudogene
1.39% Other
F
Exclusively identified by PCS
8519 genes
40.81% Protein Coding
7.71% lncRNA
36.04% Processed pseudogene
7.42% Unprocessed Pseudogene
1.94% Transcribed Processed Pseudogene
1.67% Transcribed Unprocessed Pseudogene
4.41% Other

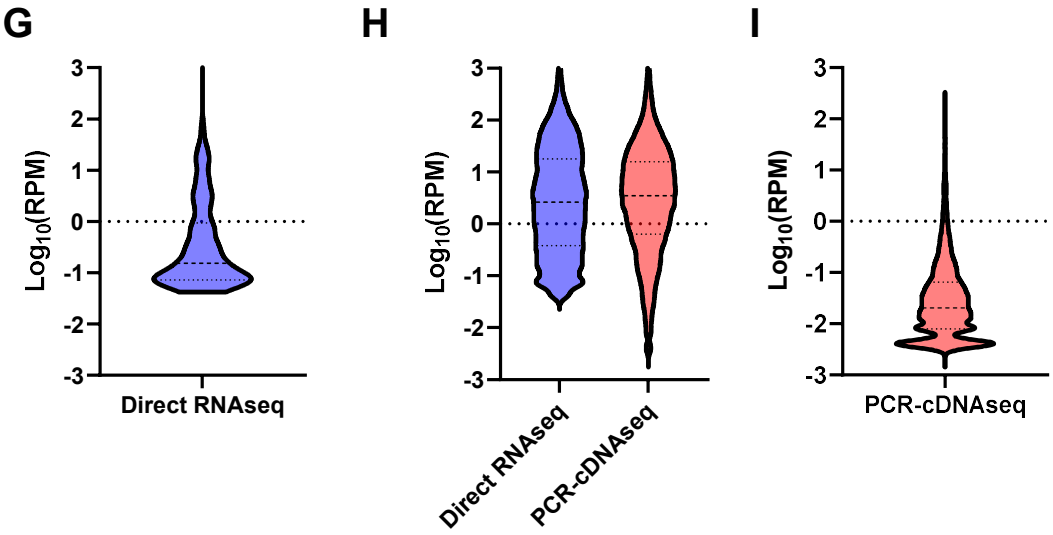

G
Log10(RPM)
3
2
1
0
-1
-2
-3
Direct RNAseq
H
Log10(RPM)
Direct RNAseq
PCR-cDNAseq
I
Log10(RPM)
PCR-cDNAseq

Figure 3.15: Differences and common genes identified in ccRCC tumours by reference transcriptome mapped reads from DRS and PCS

A) Pie chart depicting the proportions of gene biotypes of all reference genome (Ensembl release 105, cDNA reference) mapped DRS reads from all ccRCC tumour sample. **B)** As in A but for PCS reads. **C)** Venn diagram showing the overlap between reference transcriptome mapped DRS and PCS identified genes. **D)** Pie chart depicting the proportions of RNA biotypes of genes detected by reference genome-mapped DRS reads exclusively. **E)** As in **D**, but for genes identified by both DRS and PCS. **F)** As in **D**, but for genes detected by reference transcriptome-mapped PCS reads exclusively. **G)** Violin plot depicting the distribution of gene expression levels (Log10 Reads per million (RPM)) of genes detected by reference transcriptome-mapped DRS reads exclusively. **H)** As in **G**, but for genes identified by both DRS and PCS. **I)** as in **G**, but for genes detected by reference transcriptome-mapped PCS reads exclusively.

3.3.11 Overlapping genes identified by DRS and PCS in ccRCC tumour samples

Taking all four sets of genes that were identified by DRS and PCS, aligned to either reference genome or transcriptome, a Venn diagram was created to identify overlapping genes (Figure 3.16A). 15,811 unique genes were found. Amongst the 15811 unique genes, 92.24% are identified as protein-coding genes, 3.85% are processed pseudogenes, 0.57% are unprocessed pseudogenes, 0.82% are transcribed processed pseudogenes, and 1.21% are transcribed unprocessed pseudogenes (Figure 3.16B).

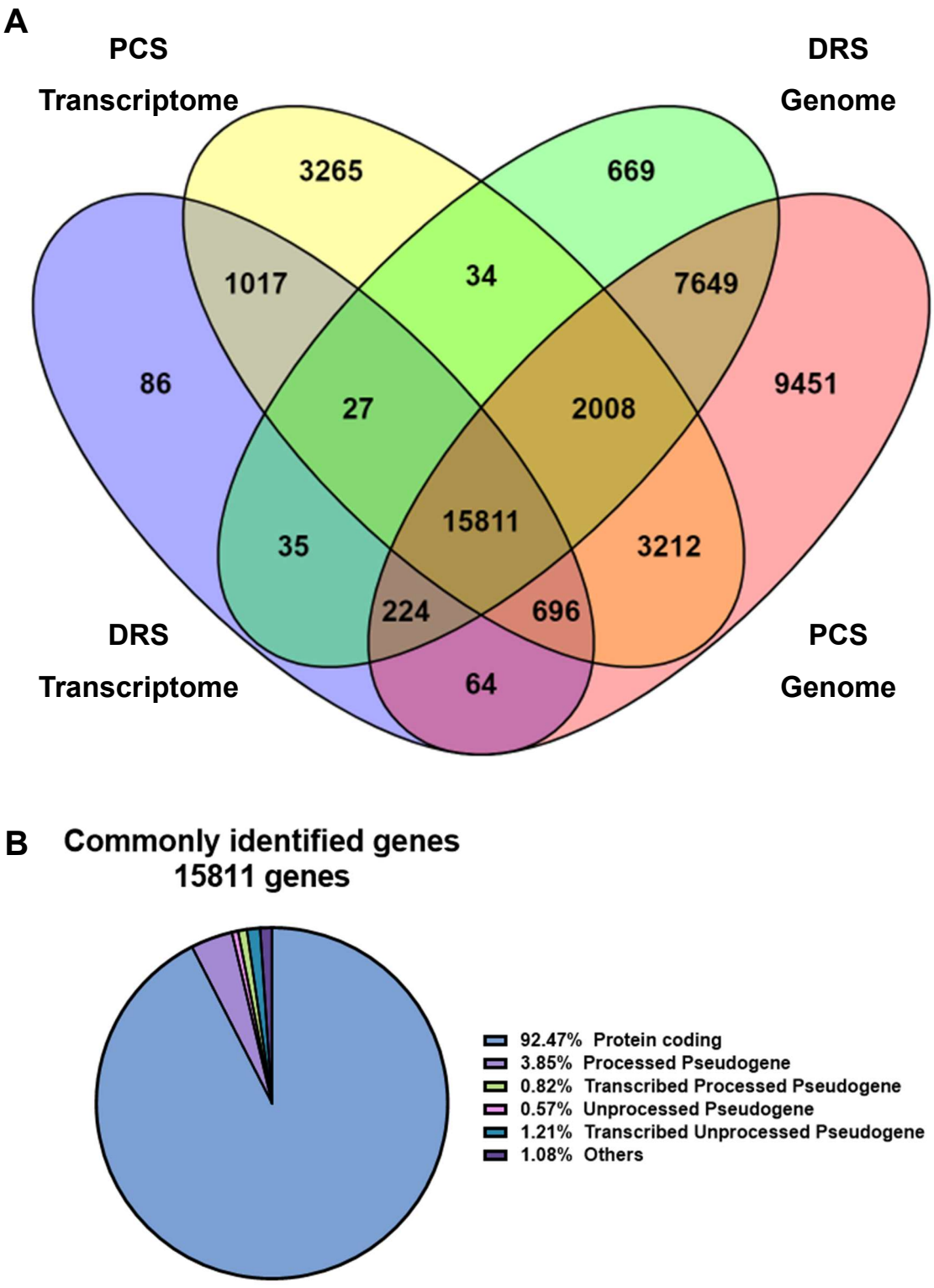


Figure 3.16: Common genes identified by DRS and PCS in ccRCC tumours

A) Venn diagram showing the overlaps between reference genome and transcriptome aligned DRS and PCS identified genes. **B)** Pie chart depicting the proportions of gene biotypes of overlapping genes from both reference genome and transcriptome aligned DRS and PCS (n = 15811).

3.3.12 Characterisation of gene expression levels per RNA biotype in ccRCC tumours

After identifying RNA biotypes in the ccRCC tumours, the distributions of gene expression levels for each biotype were analysed. Violin plots were generated to show the distribution of gene expression levels (Log_{10} RPM) of detected genes by biotypes for genome-aligned DRS and PCS (Figure 3.17A – B), as well as transcriptome-aligned DRS and PCS (Figure 3.18A – B).

The highest expressing biotype for reference genome-aligned DRS and PCS is mt-rRNA (n = 2), with mean RPM at 23,285 and 26,741, respectively. Compared to mt-rRNA, the expression of rRNA (n = 8) is at a much lower level, with median RPM at 0.9711 for DRS and 0.07121 for PCS. The second most highly expressed biotype is protein-coding genes, with median RPM at 14.76 for DRS and 10.80 for PCS. Although lncRNA is the second largest group of uniquely identified genes behind protein-coding genes in reference genome aligned DRS and PCS of ccRCC tumours, the median RPM for lncRNA are 0.2082 and 0.3466 for DRS and PCS, respectively (Figure 3.17A – B).

For reference transcriptome-aligned DRS and PCS, mt-rRNAs are not found in the sequencing data. Instead, the highest expressing biotype is protein coding, with a median RPM of 3.448 for DRS and 2.496 for PCS. Previous analysis has shown that the rRNA pseudogenes are the second-highest expressed group of genes in reference transcriptome aligned DRS and PCS (Figure 3.12B, D). Whilst most of the rRNA pseudogenes expressed at a low level (median for DRS: 0.1892, median for PCS: 0.01623), the distributions are skewed. The RPM of the highest expressing rRNA pseudogene in reference transcriptome-aligned DRS and PCS are 2619 and 8388. Lastly, although processed pseudogenes represent the second largest number of unique genes in reference transcriptome-aligned DRS and PCS, they are typically very lowly expressed, with the DRS median RPM at 0.1230 and PCS median RPM at 0.01636 (Figure 3.18A – B).

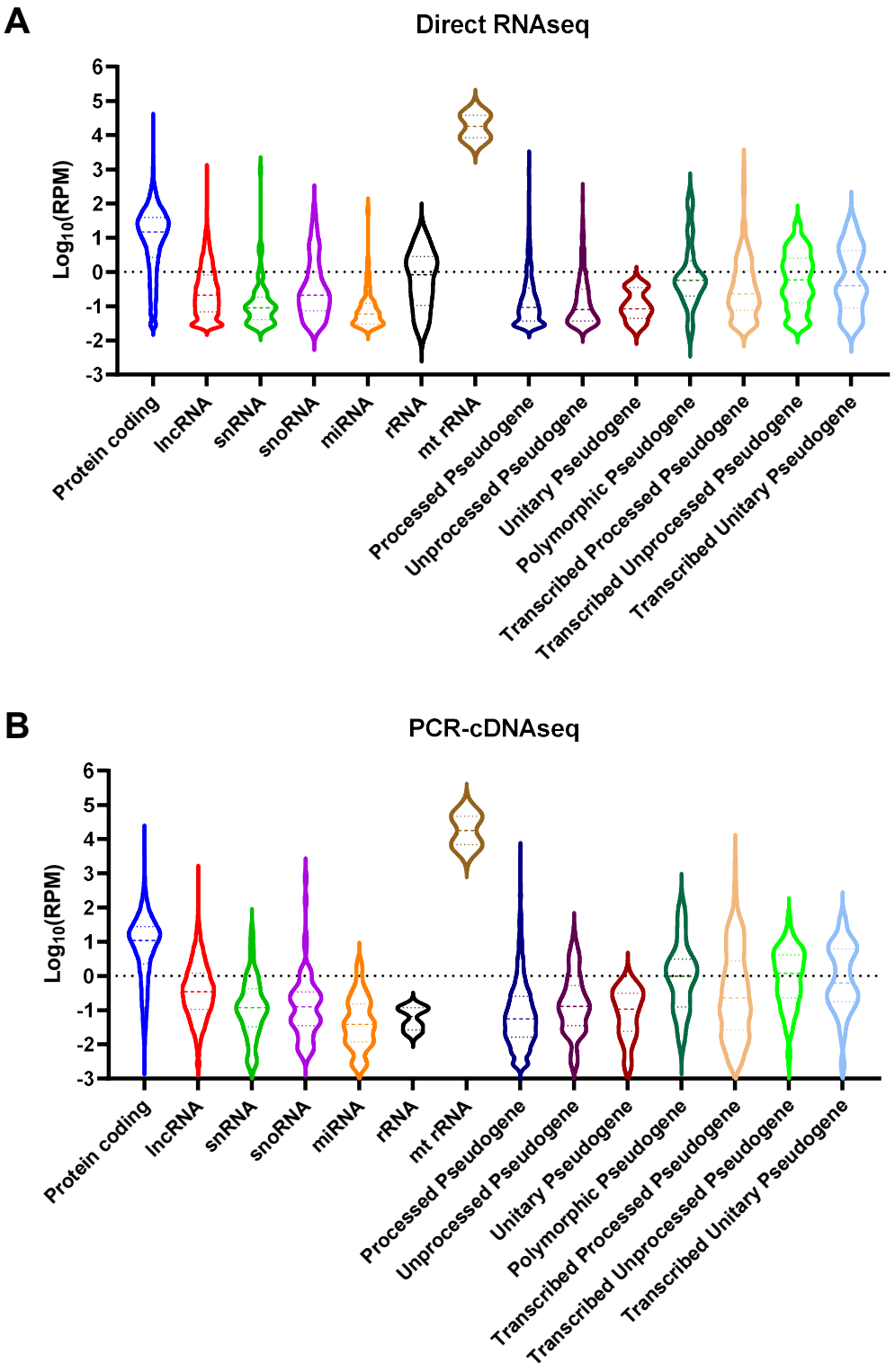


Figure 3.17: Expression levels of genes identified in ccRCC tumours by reference genome aligned reads by biotypes

A) Violin plot depicting the distribution of gene expression levels (Log_{10} Reads per million (RPM)) of genes detected by reference genome aligned DRS reads (Ensembl release 105, GRCh38) by biotypes. **B)** As in A, but for PCS.

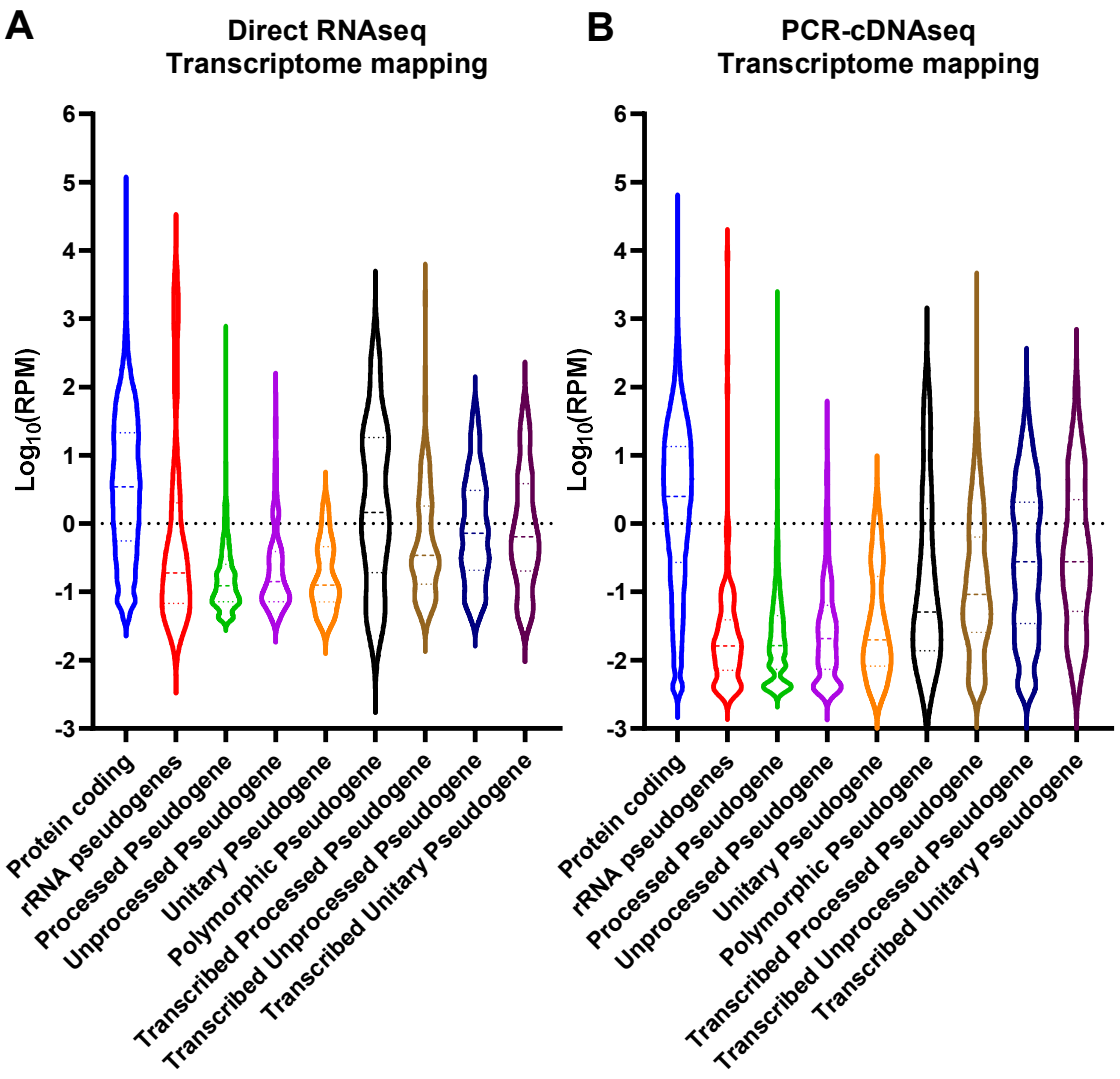


Figure 3.18: Expression levels of genes identified in ccRCC tumours by reference transcriptome aligned reads by biotypes

A) Violin plot depicting the distribution of gene expression levels (Log_{10} Reads per million (RPM)) of genes detected by transcriptome aligned DRS reads (Ensembl release 105, cDNA reference) by biotypes. **B)** As in A, but for PCS.

3.3.13 Correlations of gene expression per biotype between DRS and PCS of ccRCC tumour samples

To test the comparability of gene expression levels between DRS- and PCS-generated data, expression levels of genes that both DRS and PCS identified from different biotypes were compared. For both reference genome-aligned data (Figure 3.19A – F) and reference transcriptome-aligned data (Figure 3.20A – F), scatter plots between PCS and DRS expression levels (Log_{10} transformed RPM) of commonly found protein-coding genes, lncRNA, processed pseudogenes, unprocessed pseudogenes, transcribed processed pseudogenes and transcribed unprocessed pseudogenes were generated.

For reference genome-aligned data, expression levels of commonly identified genes show a high degree of concordance. For example, the strongest correlation can be found between DRS and PCS expression levels for protein-coding genes (n = 16203, R^2 = 0.9230 and P = <0.0001), whilst a relatively weak correlation can be observed amongst lower expressed biotypes, albeit still statistically significant, such as processed pseudogenes (n = 1633, R^2 = 0.6837 and P = <0.0001) and unprocessed pseudogenes (n = 213, R^2 = 0.5831, P = <0.0001) (Figure 3.19A - F).

For reference transcriptome-aligned data, expression levels of DRS and PCS data exhibit a lower degree of concordance than reference genome-aligned data. Significant, positive correlations can be found between DRS and PCS expression levels of protein-coding genes (R^2 = 0.7787, P = <0.0001), processed pseudogenes (R^2 = 0.3367, P = <0.0001), unprocessed pseudogenes (R^2 = 0.5615, P = <0.0001), and transcribed processed pseudogenes (R^2 = 0.6152, P = <0.0001). However, no significant correlations were observed for the expression levels of lncRNA (R^2 = 0.1962, P = 0.4551) and transcribed unprocessed pseudogenes (R^2 = 0.0098, P = 0.1113) (Figure 3.20A – F).

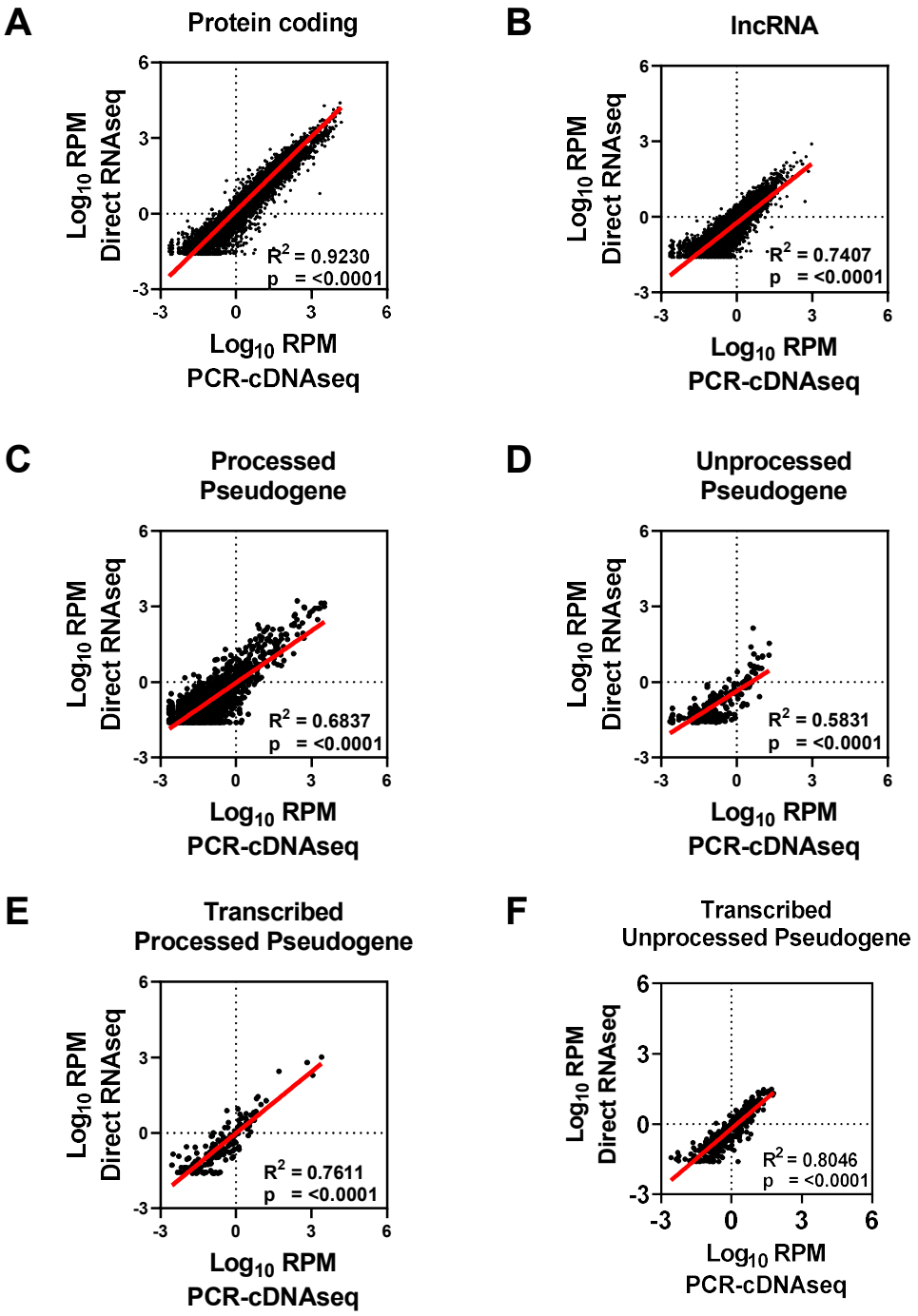


Figure 3.19: Correlations of gene biotype expression levels between reference genome aligned DRS and PCS of ccRCC tumour samples

Correlation between gene expression levels (Log_{10} Reads per million (RPM)) of **A)** protein coding genes, **B)** lncRNA, **C)** processed pseudogenes, **D)** unprocessed pseudogenes, **E)** transcribed processed pseudogenes, **F)** transcribed unprocessed pseudogenes detected by genome aligned (Ensembl release 105, GRCh38) DRS and PCS 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 generated from F-test, with $p < 0.05$ considered statistically significant.

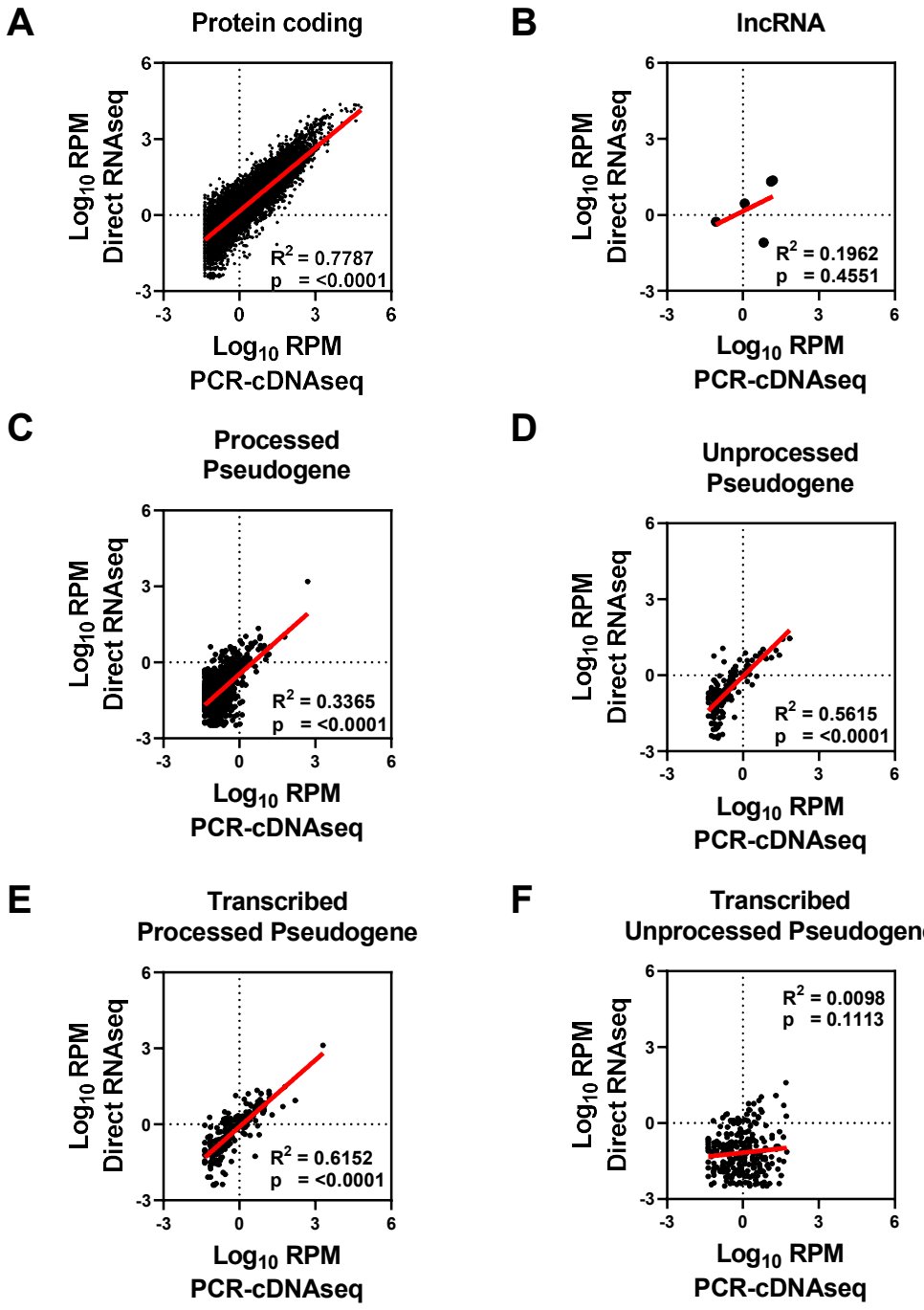


Figure 3.20: Correlations of gene biotype expression levels between reference transcriptome aligned DRS and PCS of ccRCC tumour samples

Correlation between gene expression levels (Log_{10} Reads per million (RPM)) of **A)** protein coding genes, **B)** lncRNA, **C)** processed pseudogenes, **D)** unprocessed pseudogenes, **E)** transcribed processed pseudogenes, **F)** transcribed unprocessed pseudogenes detected by transcriptome aligned (Ensembl release 105, cDNA reference) DRS and PCS 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 generated from F-test, with $p < 0.05$ considered statistically significant.

3.3.14 Characterisation of gene expression levels between DRS and PCS of ccRCC tumour samples

Comparisons of expression levels ($Log_{10}RPM$) of all commonly identified genes between DRS and PCS show strong and statistically significant correlations for both reference genome-aligned genes (R^2 = 0.8966 and P = <0.0001) and reference transcriptome-aligned genes (R^2 = 0.7570 and P = <0.0001) (Figure 3.21A, 3.22A). However, many genes still show disparities in expression levels between DRS and PCS. Differences between gene expression levels measured by DRS and PCS ($log_{10}RPM_{PCS}$ – $log_{10}RPM_{DRS}$) via reference genome or transcriptome alignment were plotted as violin graphs and shown in Figures 3.21B and 3.22B, respectively. Expression levels between DRS and PCS showed higher similarities when mapped to the reference genome, with the 75% and 25% quartile at 0.15 and -0.25, compared to 0.36 and -0.30 for reference transcriptome-aligned data. Summary descriptive statistics are outlined below the violin graphs.

To evaluate the profile of genes that show higher expression levels (or enriched) by either DRS or PCS, genes with more than 10-fold differences in expression levels were identified ($|\Delta log_{10}RPM| > 1$). For reference genome-aligned data, 409 genes were found to express ten times more in DRS than PCS ($log_{10}RPM_{PCS} – log_{10}RPM_{DRS} < -1$), where 51.34% of DRS enriched genes are classified as processed pseudogenes, 17.60% as lncRNA and 12.96% as protein-coding genes (Figure 3.21C). 306 genes were enriched in PCS, with 46.25% of identified genes classified as lncRNA and 21.50% as protein-coding genes (Figure 3.21D). For reference transcriptome-aligned data, 461 genes were enriched by DRS, of which 30.37% are classified as protein coding, 31.02% are processed pseudogenes, and 29.28% are transcribed unprocessed pseudogenes (Figure 3.22C). 621 genes were enriched by PCS, where 95.15% of identified genes are protein-coding genes (Figure 3.22D).

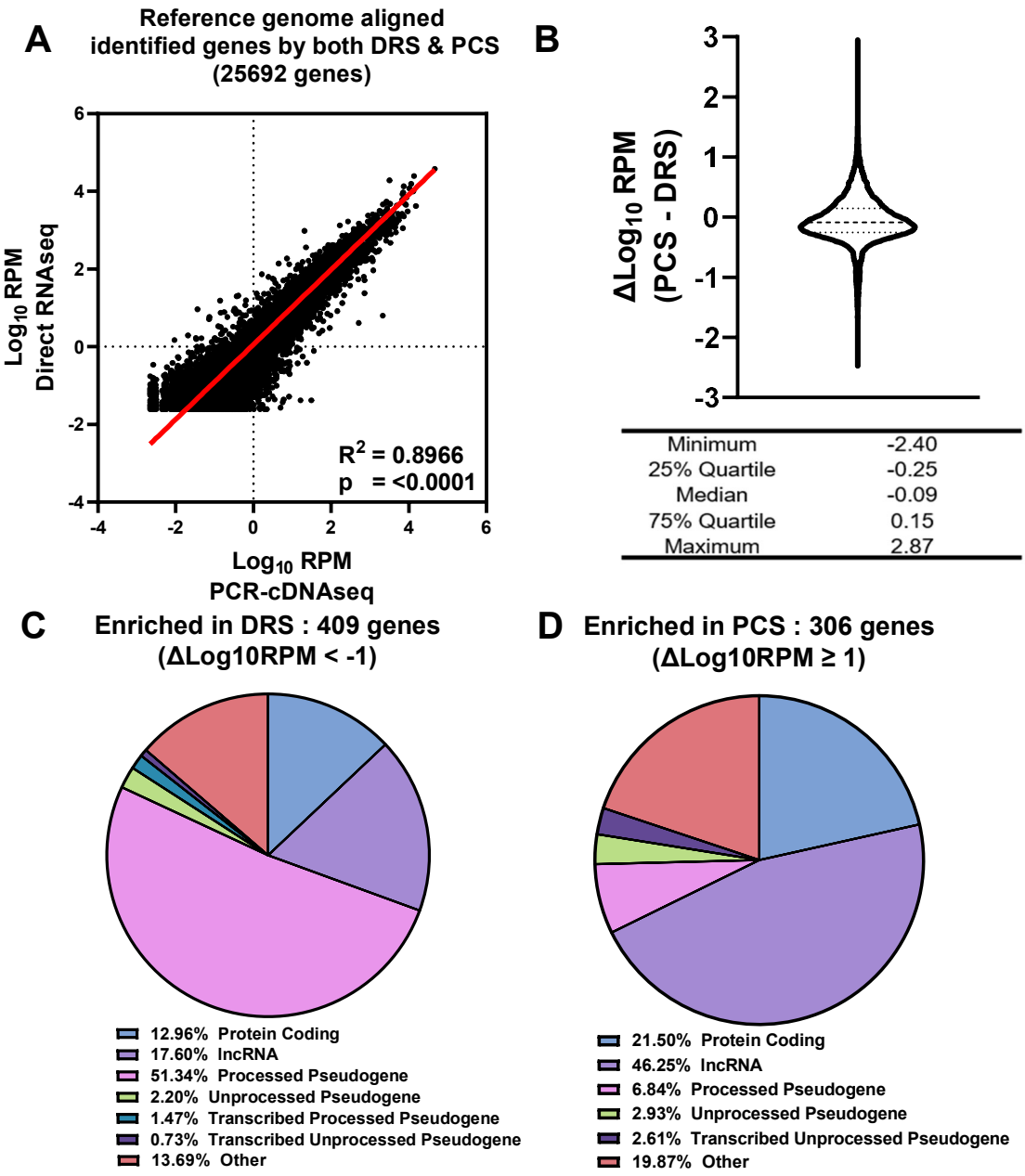


Figure 3.21: Biases in gene expression level between reference genome mapped DRS and PCS of ccRCC tumour samples

A) Correlation between gene expression levels (Log_{10} Reads per million (RPM)) of all genes detected by both reference genome mapped reads (Ensembl release 105, GRCh38) from DRS and PCS. Diagonal line represents the line of best fit. R^2 value was computed to measure goodness-of-fit, and P value generated from F-test, with $p< 0.05$ considered statistically significant. **B)** Violin plot depicting the distribution of expression level differences ($\Delta Log_{10}RPM$) between DRS and PCS, with descriptive statistics outlined below graph. **C)** Pie chart showing the proportions of biotypes of DRS enriched genes, where gene expression is 10 times higher than PCS ($\Delta Log_{10}RPM$ (PCS – DRS) < -1). **D)** As in **C**, but for PCS enriched genes.

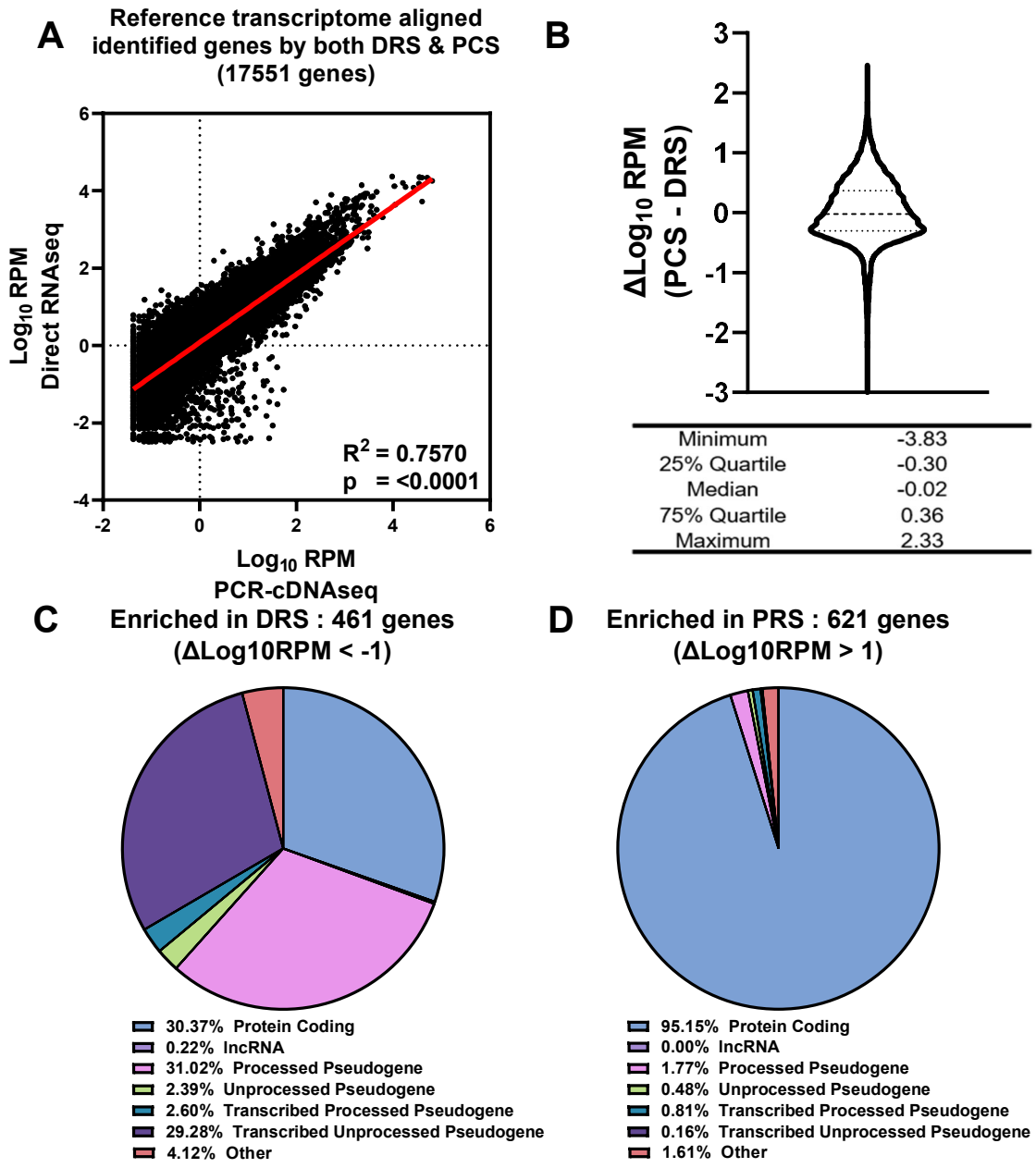


Figure 3.22: Biases in gene expression level between reference transcriptome mapped DRS and PCS of ccRCC tumour samples

A) Correlation between gene expression levels (Log_{10} Reads per million (RPM)) of all genes detected by both reference transcriptome mapped reads (Ensembl release 105, cDNA reference) from DRS & PCS. Diagonal line represents the line of best fit. R^2 value was computed to measure goodness-of-fit, and P value generated from F-test, with $p< 0.05$ considered statistically significant. **B)** Violin plot depicting the distribution of expression level differences ($\Delta Log_{10}RPM$) between DRS and PCS, with descriptive statistics outlined below graph. **C)** Pie chart showing the proportions of biotypes of DRS enriched genes, where gene expression is 10 times higher than PCS ($\Delta Log_{10}RPM$ (PCS – DRS) < -1). **D)** As in **C**, but for PCS enriched genes.

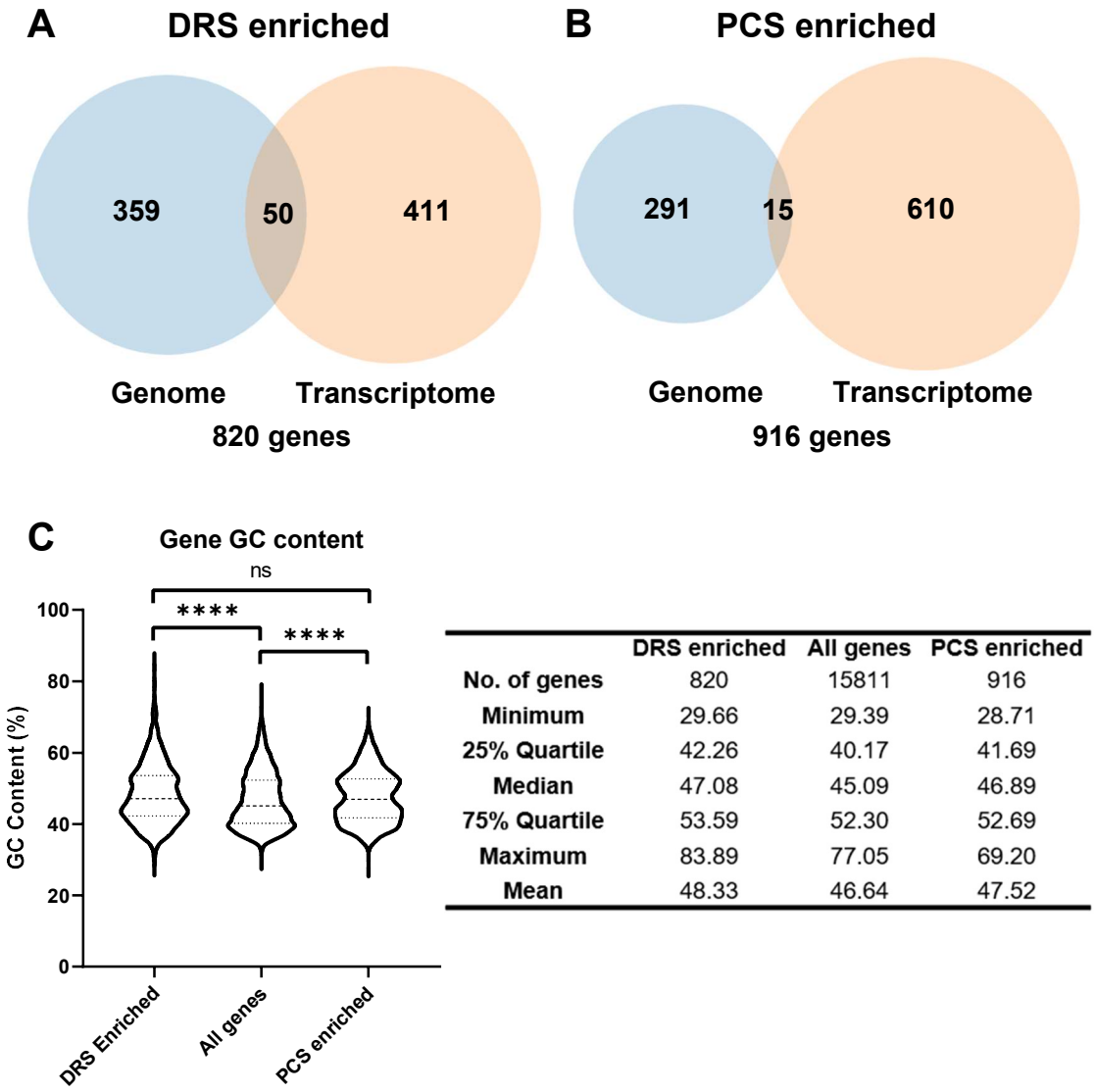
3.3.15 Characterisation of DRS and PCS expression enriched genes

To understand the potential reason for enrichment, Venn diagrams were first created to identify overlapping DRS and PCS enriched genes ($|\Delta log_{10}RPM| > 1$) between the reference genome and transcriptome alignment data (Figure 3.23 A – B). 820 unique DRS-enriched genes were discovered, with 50 genes enriched in both reference genome and transcriptome-aligned data (Figure 3.23A). In addition, 916 PCS genes were also identified, with 15 overlapping genes enriched in both reference genome and transcriptome-aligned data (Figure 3.23B).

GC content bias is one of the main factors which can introduce bias in PCR amplification during sequencing library preparation. To assess if GC content explains the enrichment in DRS (failure in amplification by PCR) or PCS (enhanced amplification by PCR), violin plots were generated to display the distribution of GC content (%) of DRS-enriched genes (n = 820), PCS enriched genes (n = 916), as well as all commonly mapped genes (n = 15811) between DRS and PCS (Figure 3.22C). GC contents are significantly higher in both DRS-enriched genes (47.08%) and PCS-enriched genes (46.89%) compared to commonly mapped genes (45.09%). However, no significant differences were found between the GC contents of DRS-enriched genes and PCS-enriched genes.

3.3.16 Correlations of gene expression between reference genome or transcriptome aligned DRS & PCS

Finally, to evaluate the correlations of gene expression levels between reference genome or transcriptome aligned DRS & PCS, scatter plots between reference genome-aligned and reference transcriptome aligned DRS and PCS expression data (Log_{10} transformed RPM) for all identified genes and protein-coding genes were generated (Figure 3.24A - D). Strong and statistically significant positive correlations were found in both DRS ($R^2 = 0.4454$ and $P = <0.0001$ for all genes, $R^2 = 0.3775$ and $P = <0.0001$ for protein-coding genes) and PCS ($R^2 = 0.5223$ and $P = <0.0001$ for all genes, $R^2 = 0.4742$ and $P = <0.0001$ for protein-coding genes).



	DRS enriched	All genes	PCS enriched
No. of genes	820	15811	916
Minimum	29.66	29.39	28.71
25% Quartile	42.26	40.17	41.69
Median	47.08	45.09	46.89
75% Quartile	53.59	52.30	52.69
Maximum	83.89	77.05	69.20
Mean	48.33	46.64	47.52

Figure 3.23: Characterisation of DRS/PCS enriched genes

A) Venn diagram showing the overlaps of DRS enriched (10 folds higher than PCS, $\Delta Log_{10}RPM$ (PCS – DRS) < -1) genes between reference genome and transcriptome alignment methods. **B)** As in **A**, but for PCS enriched genes. **C)** Violin plots showing distribution of GC content (%) of DRS/PCS enriched genes, as well as all commonly mapped genes (n = 15811) between DRS and PCS using both genome and transcriptome alignment methods. Descriptive statistics of GC contents for each gene sets were outlined next to violin plots. Statistical analysis was performed in C using a non-parametric Kruskal-Wallis one-way ANOVA test. Asterisks indicate statistical significance levels (**** = $p < 0.0001$, ns = not significant).

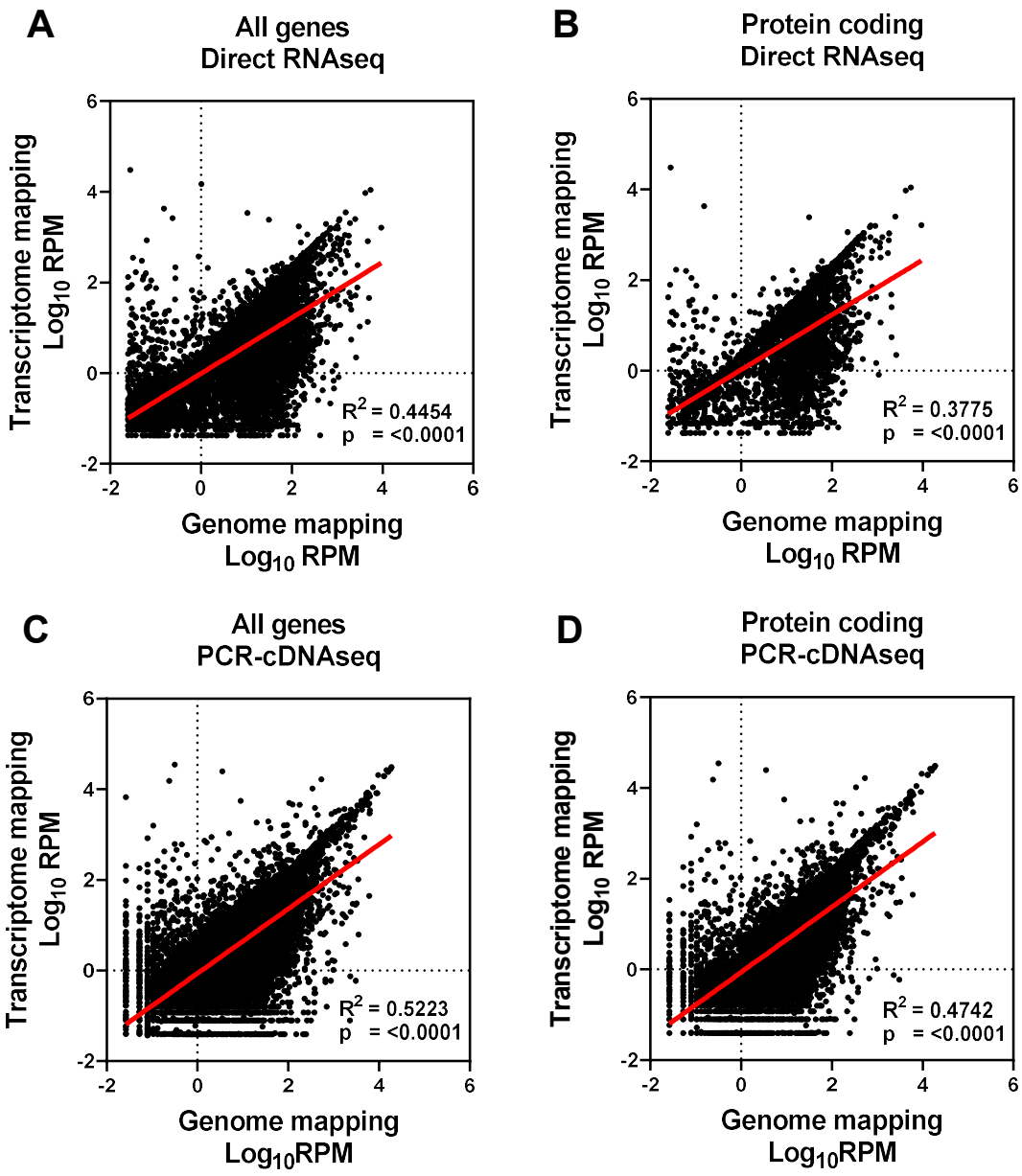


Figure 3.24: Correlations of gene expression levels between reference genome or transcriptome aligned DRS & PCS of ccRCC tumours

Correlation between gene expression levels (Log_{10} Reads per million (RPM)) of **A)** all genes and **B)** protein coding genes detected by genome-mapped (Ensembl release 105, GRCh38) or transcriptome-mapped (Ensembl release 105, cDNA reference) reads from DRS of ccRCC tumour samples. **C)** As in **A**, and **D)** as in **B** but for PCS 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 generated from F-test, with $p < 0.05$ considered statistically significant.

3.4 Discussion

This chapter comprehensively assessed the application of ONT long-read sequencing technologies (DRS and PCS) on archival tumour samples. Results in this chapter provide proof-of-concept data demonstrating the feasibility of the long-read sequencing approach for transcriptome profiling of archival tumour samples. To the best of our knowledge, this is the first work that described the usage of the ONT long-read RNA sequencing library on fresh frozen tumour specimens, as well as the first ccRCC transcriptome profiling study using the long-read RNA sequencing approach.

Firstly, RNA was successfully extracted from all archival tumour samples, with RNA yield (9.8 – 74.3 µg) sufficient for DRS and PCS library preparation. Although RNA from samples was found to be partially degraded, this is consistent with other studies using archival fresh frozen tumour tissues where RIN typically range between 6.0 – 8.0 (Lalmahomed *et al.*, 2017; Bossel Ben-Moshe *et al.*, 2018). Our results show significant correlations between the RNA concentrations recorded by the 3 RNA quantification methods (Nanodrop, Bioanalyzer and Qubit), but the RNA concentrations differ substantially. The three approaches utilise fundamentally different methods to estimate RNA concentration. Nanodrop is a UV spectrometer that calculates RNA concentration by the intensity of absorbance peak at 260 nm (by the RNA heterocyclic rings) (Desjardins and Conklin, 2010). Bioanalyzer (Agilent 2100) performs rapid electrophoresis on microfluidic chips and determines RNA concentration by assessing the distribution and intensity of RNA fragments in samples against the RNA reference control ladder (Davies *et al.*, 2016). Qubit measures RNA concentration by utilising RNA binding dyes, which emit fluorescence only when bound (Li *et al.*, 2015). Previous work has suggested that the presence of 260nm absorbing contaminants and degraded free nucleic acid bases contribute to inaccurate and inflated RNA concentration measurements by the UV-spectrometer-based Nanodrop (Garcia-Elias *et al.*, 2017). The presence of short RNA fragments in degraded RNA samples has also been attributed to overestimated RNA concentration using a bioanalyzer (Becker *et al.*, 2010). In addition,

reliable concentration measurements by bioanalyzer depend entirely on the consistency of RNA reference control loading on each microfluidic chip. Previous research has shown that RNA concentration measured by Qubit exhibit a higher accuracy and reproducibility level than Bioanalyzer (Davies *et al.*, 2016). Here, results show higher levels of RNA concentration measurement by Nanodrop and Bioanalyzer compared to Qubit. Thus, to ensure adequate input was used for library generation, RNA concentrations measured by Qubit were used in this study.

Compared to the ng – pg range of total RNA input requirement for Illumina RNAseq library preparation, previous works 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 reasoned that since the first step of DRS and PCS library generation involves binding an adapter primer containing a poly-d(T) sequence, a prior step for the enrichment of poly(A) RNA molecules should not be necessary. Indeed, a recent work published in July 2022 has demonstrated that poly(A) selection is not only unnecessary for DRS input, but it can also introduce a potential bias towards mRNAs with longer poly(A) tails (Viscardi and Arribere, 2022). Using prepared sequencing libraries, DRS and PCS successfully sequenced archival ccRCC tumours, with a median of 3.2 and 56.6 million passed reads (Q > 7), respectively. Interestingly, whilst there is no significant correlation between the number of DRS-generated reads and sample RIN score, the number of generated PCS reads strongly correlates with the RIN score. A likely reason for this significant correlation is that the number of cDNA molecules in the PCS library was not the limiting factor in the number of reads generated. Instead, it was the speed at which reads were passed through pores. Therefore, the shorter the cDNAs, the higher number of reads that could be generated. This result also indicates that a lower number of PCR amplification cycles can be used in the PCS library generation protocol. It is long recognised that PCR amplification can introduce sequencing bias in RNA sequencing experiments, and one of the most effective ways to reduce bias is by minimising the number of PCR cycles (Aird *et al.*, 2011). Comparing the 200ng total RNA input with 14

amplification cycles recommended by ONT PCS library generation protocol, Illumina sequencing libraries, such as Takara SMARTseq v4, can now utilise input as low as 10pg (single cell equivalent yield) with 17-18 PCR amplification cycles (Sarantopoulou *et al.*, 2019). Results here show the potential of lowering both the input RNA levels and the number of amplification cycles used in the current PCS library protocol.

Next, the read-length and alignment-length profiles of DRS and PCS of tumour samples were evaluated. On average, raw reads generated by PCS were longer than DRS reads. The difference in length is partially influenced by the fact that raw reads from PCS have additional ligated reverse transcription, PCR amplification primer and unique molecular identifier (UMI) sequences compared to DRS reads. Once aligned to the reference genome, the difference between PCS and DRS read lengths reduced. Data from reference-transcriptome aligned reads show that PCS provides a higher percentage of transcripts that represent full-length transcripts. Overall, PCS generate reads that are longer than DRS.

To our surprise, no significant correlations were found between DRS and PCS read lengths for each tumour sample (Figure 3.6A). Whilst strong concordance was observed between RIN score and PCS read lengths, no significant correlation was found between RIN score and DRS read lengths (Figure 3.6). This lack of correlation suggests that additional factors impacted the selection of RNA molecules sequenced using the constructed DRS libraries. Recent work has examined whether ONT DRS sequence shorter RNA molecules preferentially. Using *in vitro* transcribed spike-in RNA controls with various lengths (from 150 – 2500 nt), the study concluded no length-based selection bias (Ibrahim *et al.*, 2021). Other DRS experiments conducted in this thesis (Chapters 5) utilise non-degraded poly(A) enriched RNA (RIN = 10). No comparisons can thus be drawn between the experiments. To the best of our knowledge, no study has explored the effects of RNA degradation on DRS and PCS read lengths. It would be highly valuable to characterise this relationship systematically so that the broader research

community can make informed decisions on using long-read or short-read sequencing for partially degraded RNA from clinical samples.

The longer PCS reads can be attributed to two main reasons: firstly, PCS library generation uses strand switching method with Maxima H Minus reverse transcriptase, which adds three additional protruding cytosines (CCC) at the 3' end of the reverse transcribed cDNAs (5' end of the mRNA molecules), allowing a template switching oligo (TSO) with three riboguanosines to anchor. It has been shown that reverse transcriptase exhibits higher strand switching efficiency for 5' m^7G capped mRNA than uncapped RNAs (Wulf *et al.*, 2019). This enrichment enhances the proportion of full-length capped mRNAs in PCS compared to DRS. Secondly, the translocation speed through the nanopore is significantly slower for RNA molecules (~70 bases/s) than DNA (~450 bases/s). The slower speed increases the probability of read stalling and pore blocking, resulting in more truncated short reads in DRS compared to PCS (Ibrahim *et al.*, 2021). Overall, DRS and PCS from clinical ccRCC tumour samples produce longer reads than traditional sequencing methods.

The second step of transcriptomic analysis is mapping reads against the reference genome or transcriptome. The decision to select alignment against a reference genome or transcriptome depends on multiple factors. Reference databases such as Ensembl provide curated reference genomes and reference cDNA for transcriptome mapping (Cunningham *et al.*, 2022). The most challenging aspect of reference alignment is deciding the identity of reads mapped to multiple genes and transcripts, also known as multi-mapped reads (Deschamps-Francoeur *et al.*, 2020). For reference genome alignment, differentiating parent genes and processed pseudogenes, which are intronless mRNAs that are reverse transcribed into the genome, has proved difficult for aligners (Raplee *et al.*, 2019). For reference transcriptome alignment, reads can align to multiple genes that share a similar identity and potentially multiple transcript isoforms that share exon structures. The read-mapping rate also tends to be lower than reference genome alignment, since reads from unannotated transcripts have no sequence to map

to (Conesa *et al.*, 2016). Characterisation on the quantity and biotype profile of identified genes from the reference genome and reference transcriptome aligned DRS and PCS read forms the next focus of the chapter.

PCS identified a higher number of unique genes when compared to DRS, which is expected since PCS generated nearly 20 times more reads than DRS. When aligned to the reference genome, the median number of genes identified by PCS and DRS is higher than when reads were aligned to the reference transcriptome. This discrepancy can be attributed to the lack of annotated non-coding genes in the reference transcriptome, as reflected in the biotype profiles for reference genome and reference transcriptome mapped samples (Figure 3.14 – 15). Regarding the proportion of mapped reads per biotype in tumour samples, in both reference genome and transcriptome-mapped DRS/PCS data, the vast majority of reads (89-98% of all reads) are mapped to protein-coding genes. However, combining all unique genes identified across the tumour samples, in reference genome-aligned DRS and PCS, 32.48% and 45.47% of genes are classified as protein coding, and 28.81% and 29.48% of genes are lncRNAs. Other biotypes were also discovered, with over 20% and 15% of all unique genes found from reference genome-aligned DRS and PCS being pseudogenes. These findings demonstrate the diversity of genes being expressed in the ccRCC transcriptome and the varieties of transcripts that long-read sequencing technologies can profile.

Looking at the genes that were identified by both reference genome and reference transcriptome aligned DRS and PCS, although most genes were either identified by both DRS and PCS or exclusively by PCS, there is a significant number of genes that were only found via DRS. This is despite the considerably higher number of reads acquired through PCS and the greater sequencing depth. Subsequent gene expression analysis shows that the expression levels of the DRS-exclusive and PCS-exclusive genes are significantly lower than the genes identified by both DRS and PCS. Many of these genes, especially PCS-exclusive genes, may be absent from the other dataset due to the lack of coverage and depth. It is also possible that many of the DRS-exclusive genes do not

amplify well by PCR, hence excluded from the PCS data. Conversely, PCS-exclusive genes may represent a subset of preferentially amplified genes. The inclusion of UMI in the PCS111 library allows the correction of amplification bias after sequencing, and the pre-mapping read QC module pychopper (v2.7.2, Oct 2022) has recently been updated to pass UMI information to the later mapping step. It will be of great interest to characterise genes that are heavily influenced by PCR amplification bias.

Analysis of the gene expression levels across RNA biotypes reveals that despite the lack of poly(A) RNA enrichment or ribodepletion, rRNAs are only detected at a low level when aligned to the reference genome and undetectable when aligned to reference transcriptome in both PCS and DRS (Figure 3.17 – 18). The lack of rRNA mapping further suggests that poly(A) enrichment before library preparation is not a requirement for ONT long-read sequencing. As expected, protein-coding genes are expressed at a higher level than lncRNA and pseudogenes, which was reflected previously. The highest expressed biotype in both reference genome-aligned DRS and PCS data is mitochondrial rRNA, which is polyadenylated once matured (Chang and Tong, 2012). However, regarding the total number of aligned reads, mt-rRNA only represents 4.66% and 5.35 % of reference genome-aligned DRS and PCS reads (Figure 3.13). Commercial ribodepletion kits are designed to target both cytoplasmic and mitochondrial rRNAs (Herbert *et al.*, 2018). Benchmarking experiments comparing biotype composition and gene expression levels between ONT sequencing libraries generated from poly(A) enriched RNA, ribodepleted RNA and total RNA will be useful for the research community. One of the main benefits of using ribodepletion over poly(A) enrichment is the ability to isolate non-poly(A) transcripts, including numerous functional protein-coding and non-coding RNAs (Zhao *et al.*, 2014). The design and sequence of the poly-d(T) primer from ONT library preparation will have to be amended to capture this diverse population of transcripts.

Though gene expression levels strongly correlate between DRS and PCS, hundreds of genes with various biotypes exhibit at least 10-fold differences in gene expression levels

between the two methods (Figure 3.21 – 22). Similar to the genes that DRS and PCS exclusively mapped, these 'enriched' genes may be influenced by variability linked to low expression levels. Venn diagram shows that only small subsets of genes are 'enriched' by either DRS or PCS when both reference genome and reference transcriptome aligned data were overlapped. The lack of overlapping genes suggests that at least in part, the observed variations may be attributed to the alignment method used or gene expression level estimation by *featurecount* or *Salmon*. Indeed, whilst gene expression levels between the reference genome and reference transcriptome aligned DRS and PCS are significantly correlated, the correlations are far from perfect (R^2 = 0.4454 for DRS, R^2 = 0.5523 for PCS) (Figure 3.24). Intriguingly, the median GC content of DRS and PCS-enriched genes were significantly higher than the average GC content of all DRS/PCS detected genes (Figure 3.23). It is known that both high GC content (<55%) and low GC content (>40%) in RNA fragments cause inefficient PCR amplification (Benjamini and Speed, 2012). This analysis assessed the GC content of the whole mapped parent gene due to difficulties tracing the specific isoform and the length of read fragments. However, it would be interesting to assess if these transcripts share similar characteristics with the DRS- and PCS-exclusively mapped genes and whether UMI corrections can drastically reduce the gene-expression discrepancies between DRS and PCS.

3.5 Evaluation of key objectives

- **Feasibility of using ONT LRS to profile archival tumour samples**

  Clinical tumour samples can be sequenced using DRS and PCS. Semi-degraded RNA samples (RIN > 7) can still result in reads that represent full-length transcript isoforms. The ability of using total RNA instead of poly(A)$^+$ RNA can substantially reduce the amount of input RNA needed for DRS/PCS library preparation. In addition, this will lower the cost related to RNA sample processing, as well as prevent RNA degradation during poly(A)$^+$ RNA isolation.

- **Comparisons of sequencing reads and output between ONT DRS & PCS**

  PCS provided more than 10x depth and on average 100nt+ longer reads compared to DRS. Therefore, with the current technologies, PCS is the better choice for studying gene expression, isoform usage and identification of novel transcripts. However, DRS can preserve RNA-modification signals and poly-A tail length estimation using nanopolish, allowing more opportunities to study the relationships between gene expression and co-/post-transcriptional mRNA regulations.

- **Assessment of different RNA biotypes' gene expression levels in DRS & PCS**

  Both DRS and PCS captured a diverse profile of RNA biotypes from ccRCC tumours. Strikingly, despite using total RNA as input, ribosomal RNA was largely depleted from DRS/PCS output. For the vast majority of mapped genes, expression levels were highly concordant between DRS and PCS data. Nevertheless, more than 1,000 genes were found to be enriched by 10-folds by either method. Additional work is needed to understand the discrepencies in gene expression levels.

3.6 Summary

This chapter examined the feasibility of using ONT DRS and PCS on archival tumour samples. Tumours were comprehensively sequenced, with substantially longer reads than traditional NGS methods. Genes from a wide variety of RNA biotypes were identified, and the differences in biotype profiles of genes identified by DRS and PCS and by reference genome and reference transcriptome alignment were demonstrated. Expression levels were highly correlated between DRS and PCS and the alignment methods. However, subsets of genes have also been found to be exclusively identified by either DRS or PCS. Hundreds of genes were also found to show high variance in gene expression levels estimated by both sequencing methods.

Chapter 4

Comprehensive analysis of ccRCC tumours using long-read sequencing technologies

4.1 Introduction

The use of whole genome sequencing and transcriptome sequencing has transformed the fundamental understanding of ccRCC over the past decade. ccRCC is now known to exhibit many distinct characteristics, including the near-universal loss of chromosome 3p arm and *VHL*, dysregulated PI3K/AKT/mTOR signalling pathway and metabolic rewiring (Qi *et al.*, 2021). With the introduction of TKI and ICI therapy, precision medicine has become a promising therapeutic approach that can revolutionalise ccRCC treatment. However, several key challenges remain that prevent the stratification of optimal systemic therapy for ccRCC patients (Signoretti *et al.*, 2018).

Localised, early-stage ccRCC is currently highly treatable via nephrectomy, with more than 80% of patients surviving five years post-surgery. However, more than 20% of patients experience local or distant recurrent post-nephrectomy (Janssen *et al.*, 2018). In addition, once metastasised, ccRCC patient prognosis is extremely poor, with a 5-year survival rate under 10% (Padala *et al.*, 2020). A recent retrospective ccRCC study has shown that amongst patients who developed recurrent ccRCC (286 out of 1265 ccRCC patients, 22.6%), 54.2% (n = 155) were defined as incurable, with only 33.4 % of recurrent ccRCC patients surviving longer than 24 months after diagnosis (Dabestani *et al.*, 2019). Therefore, reliable predictive and prognostic biomarkers for ccRCC recurrence and metastasis are urgently needed.

Leibovich score is the most widely used clinical predictive model for ccRCC recurrence risk after nephrectomy in the UK (Vasudev *et al.*, 2020). The score is calculated based on TNM (tumour size, lymph node status, metastasis) classification, Fuhrman grade (size and morphology of tumour cell nucleus), and histologic tumour necrosis (Leibovich *et al.*, 2003). In addition, a multi-genes signature-based RT-PCR assay has also been developed and validated to assess the risk of post-surgery ccRCC recurrence (Rini *et al.*, 2015). Both methods are robust and reliable, with follow-up screening protocols being adopted. However, no benefits have been found in early systemic treatment compared to delayed treatment. Moreover, no prediction model has been developed to stratify front-

line systemic therapy for recurrent and non-recurrent ccRCC (Escudier *et al.*, 2019). Thus, a deeper understanding of the underlying biology of ccRCC disease recurrence is needed to improve current treatment strategies.

Gene expression-based recurrent risk assay reveals differences between recurrent and non-recurrent ccRCC tumours. Rini *et al.* published a 16-gene RT-PCR-based recurrence prediction model with 11 cancer biomarkers and five reference genes. The biomarkers can be broadly separated into four categories: vascular, cell growth, immune response and inflammation. Suppressed angiogenic dependency and immune response, high levels of cell division and inflammation are characteristics of aggressive recurrent tumours (Rini *et al.*, 2015). Interestingly, ccRCC tumours have one of the most heterogeneous cell populations within the TME amongst all cancer types. Analysis of TCGA data showed that on average ccRCC tumours have the third lowest tumour purity (64.6%, n = 542), behind lung adenocarcinoma and lung squamous cell carcinoma (Aran *et al.*, 2015).

The rapid development of single-cell RNAseq and computational cell-type deconvolution method from bulk-RNAseq have played a crucial role in characterising the immune landscape in ccRCC TME. It is now widely recognised that ccRCC disease progression treatment outcomes are strongly linked with the profile of tumour infiltrating immune cells. In most solid cancer types, such as non-small cell lung cancer (NSCLC) and breast cancer and melanoma, high levels of $CD8^+$ T cell infiltration correlate with favourable overall survival and better immune checkpoint inhibitor treatment outcome (Li *et al.*, 2019; F. Li *et al.*, 2021). For ccRCC, $CD8^+$ T cell infiltration does not confer improved immunotherapy treatment outcomes. In addition, high CD8+ T cell infiltration levels correlate with worse overall patient survival (Giraldo *et al.*, 2017; Braun *et al.*, 2020). Recent studies using single-cell RNAseq have discovered that ccRCC tumour infiltrating $CD8^+$ T cells exhibit hugely diverse transcriptomic profiles. However, in advanced ccRCC tumours, most infiltrating $CD8^+$ T cells are dysfunctional and display an exhausted phenotype (marked by upregulation of the immune checkpoints) (Hu *et al.*, 2020; Braun

et al., 2021). This was an exciting finding since various reports have shown T cell exhaustion as a heterogenous and potentially reversible state that can be specifically targeted as a therapeutic approach (Budimir *et al.*, 2022). Overall, tumour-infiltrating immune cells play critical roles in ccRCC disease progression. Using a transcriptomic approach, a better understanding of the ccRCC immune landscapes could potentially identify predictive biomarkers for immunotherapy efficacy and better ccRCC treatment stratification.

Aside from differential gene expression analysis, the composition of expressed transcript isoforms (or differential transcript usage (DTU)) can also be analysed using transcriptome data. Differential splicing events and alternative polyadenylation (APA) are two main contributors to DTU. Dysregulated splicing events are oncogenic drivers that can promote cancer progression (Y. Zhang *et al.*, 2021). For example, abnormal splicing and DTU of the metabolic enzyme *PKM* play a critical role in tumour growth in multiple cancer types, including ccRCC. In ccRCC, polypyrimidine tract-binding protein 1 (PTBP1) mediates the isoform switch towards the expression of the pro-oncogenic transcript isoform *PKM2* (Jiang *et al.*, 2017). High *PKM2* expression in ccRCC is linked to unfavourable overall survival, whereas expression of the other *PKM* transcript isoforms ENST00000389093 and ENST00000568883 significantly correlate with better survival outcomes (X. Li *et al.*, 2021).

APA is another critical regulatory mechanism that diversifies mRNA isoforms. Multiple studies have reported global shortening of 3'UTRs via APA in cancer cells compared to normal cells (Mayr and Bartel, 2009; Zingone *et al.*, 2021). 3'UTR shortening allows proto-oncogenes to escape from miRNA- and RBP-mediated gene expression regulation, activating oncogenic pathways (Yuan *et al.*, 2021).

With longer reads and the possibility to cover the entire transcript with a single read, long-read sequencing provides an opportunity to study transcript structure accurately. In addition to studying splicing pattern changes of known transcripts, previous studies using long-read sequencing methods have discovered tens of thousands of novel isoforms,

representing a wealth of opportunities to explore previously uncharacterised splicing events(Workman *et al.*, 2019; Tang *et al.*, 2020). With each read profiled from the poly(A) tail towards the 5' end, and every read representing one mRNA molecule, DRS and PCS can provide the clear poly(A) site for every transcript. However, no study has explored differential splicing and APA between recurrent and non-recurrent ccRCC using long-read sequencing methods.

ONT DRS also enables the characterisation of poly(A) tail lengths with nanopolish and tailfindR (Krause *et al.*, 2019; Workman *et al.*, 2019). Poly(A) tail plays a crucial role in mRNA stability. In human cells, the lengths of poly(A) tails vary widely, with the median poly(A) length ranging between 50 – 100 nucleotides, depending on the type of poly(A) tail profiling method used (Chang *et al.*, 2014). Using data generated from DRS of human chronic myelogenous leukaemia cell line HAP1, Soneson *et al.* demonstrated that nanopolish and tailfindR results show high levels of concordance, with a similar profile of poly(A) length compared to results from TAILseq and PALseq(Soneson *et al.*, 2019). Interestingly, transcripts from different RNA biotypes displayed differential poly(A) tail profiles. Moreover, whilst poly(A) tail length profiles have been profiled globally using yeast, tumour cell lines, animal oocytes/embryos and mouse tissues, transcriptome-wide poly(A) tail length profile has not been studied using human tumour tissues.

4.2 Chapter aims

The chapter aims to comprehensively **characterise the transcriptomes of ccRCC tumours and identify disease recurrence-associated signatures, using data generated from DRS and PCS of archival samples**. The specific aims of this chapter include the following:

i) Using unsupervised methods (PCA and hierarchical clustering) to explore and characterise gene expression patterns in relation to ccRCC disease recurrence status and other clinical features

ii) Identify ccRCC recurrence-associated differential expressed genes

iii) Identify genes which display differential transcript usage in recurrent and non-recurrent ccRCC tumours

iv) Discover novel transcript isoforms using transcriptome assembly methods

v) Profile global poly(A) tail lengths of mRNA transcripts from ccRCC tumours

vi) Characterise activated and suppressed pathways and processes in recurrent ccRCC tumours

vii) Profile tumour immune infiltrate landscapes in recurrent and non-recurrent ccRCC tumours

4.3 Results

4.3.1 Evaluation of ccRCC tumour transcriptomes by PCA

After the alignment of DRS and PCS reads to reference genome and transcriptome via minimap2, gene expression levels were first quantified by featurecounts and Salmon, respectively, followed by library size normalisation and differential gene expression analysis. Reference genome-aligned data represent expression profiles that include protein-coding genes, pseudogenes and polyadenylated non-coding RNAs, whilst reference transcriptome-mapped data represent expression profiles that primarily consist of protein-coding genes and pseudogenes. Principal component analysis (PCA) was performed to assess the global gene expression patterns of ccRCC tumour samples. PCA is a commonly used data dimensionality reduction technique in transcriptomic analysis, which provides information on data variability and helps identify clusters of samples that share similar expression profiles (Conesa *et al.*, 2016).

For reference-genome aligned DRS of ccRCC tumours, the first principal component (PC1, x-axis) explains 28% of data variation, and the second principal component (PC2, y-axis) explains 16% of data variations (Figure 4.1A). For reference-genome aligned PCS, PC1 (x-axis) explains 26% of data variation, and PC2 (y-axis) explains 18% of data variations (Figure 4.1B). Scatter plots of PCA results showed no visually distinguishable clusters that correlate with patients and clinical information (ccRCC recurrence status, gender and cancer stage) in either DRS or PCS data (Figure 4.1A – F).

PCA of reference-transcriptome aligned DRS and PCS data display similar levels of variance explained by the first two principal components. For DRS, PC1 (x-axis) explains 22% of data variation and PC2 (y-axis) explains 15% of data variations (Figure 4.2A). For PCS, PC1 (x-axis) explains 21% of data variation, and PC2 (y-axis) explains 14% of data variations (Figure 4.2B). Like reference genome-aligned data, when including reference transcriptome-aligned data, scatter plots of PCA results did not show sample clusters that correlate with ccRCC recurrence status, patient gender and cancer stages in either DRS or PCS data (Figure 4.2A – F).

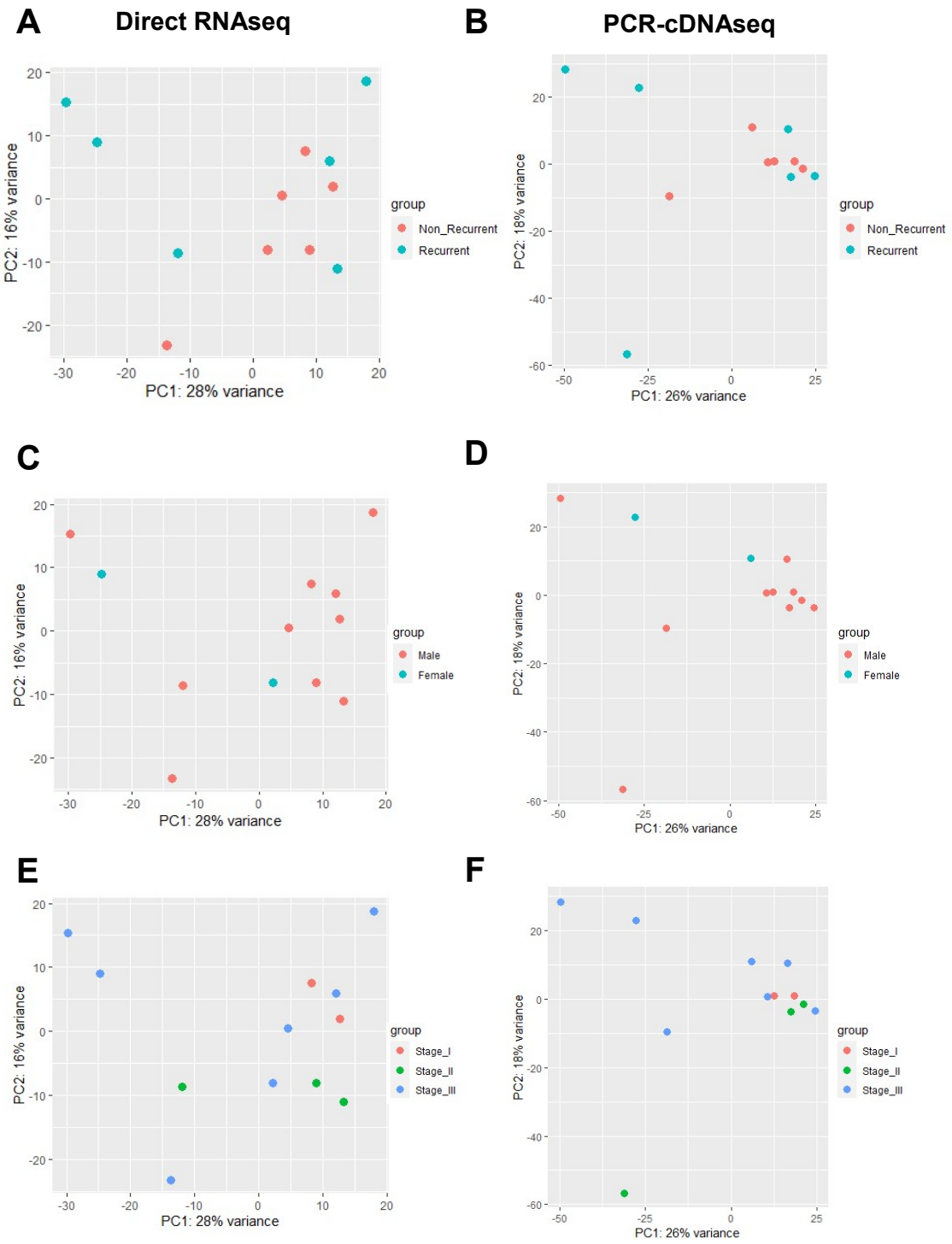


Figure 4.1: PCA analysis of ccRCC tumour transcriptome profiles using reference genome aligned DRS and PCS

Principal component analysis (PCA) on ccRCC tumours gene expression data illustrating variations between samples (dots, n = 12). DESeq2 generated plots using reference genome (Ensembl release 105, GRCh38) aligned **A)** DRS and **B)** PCS data showing PCA of recurrent vs non-recurrent ccRCC groups; **C)** DRS and **D)** PCS data showing PCA of male vs female tumour samples; **E)** DRS and **F)** PCS data showing PCA of stage I, stage II and stage III cancer samples.

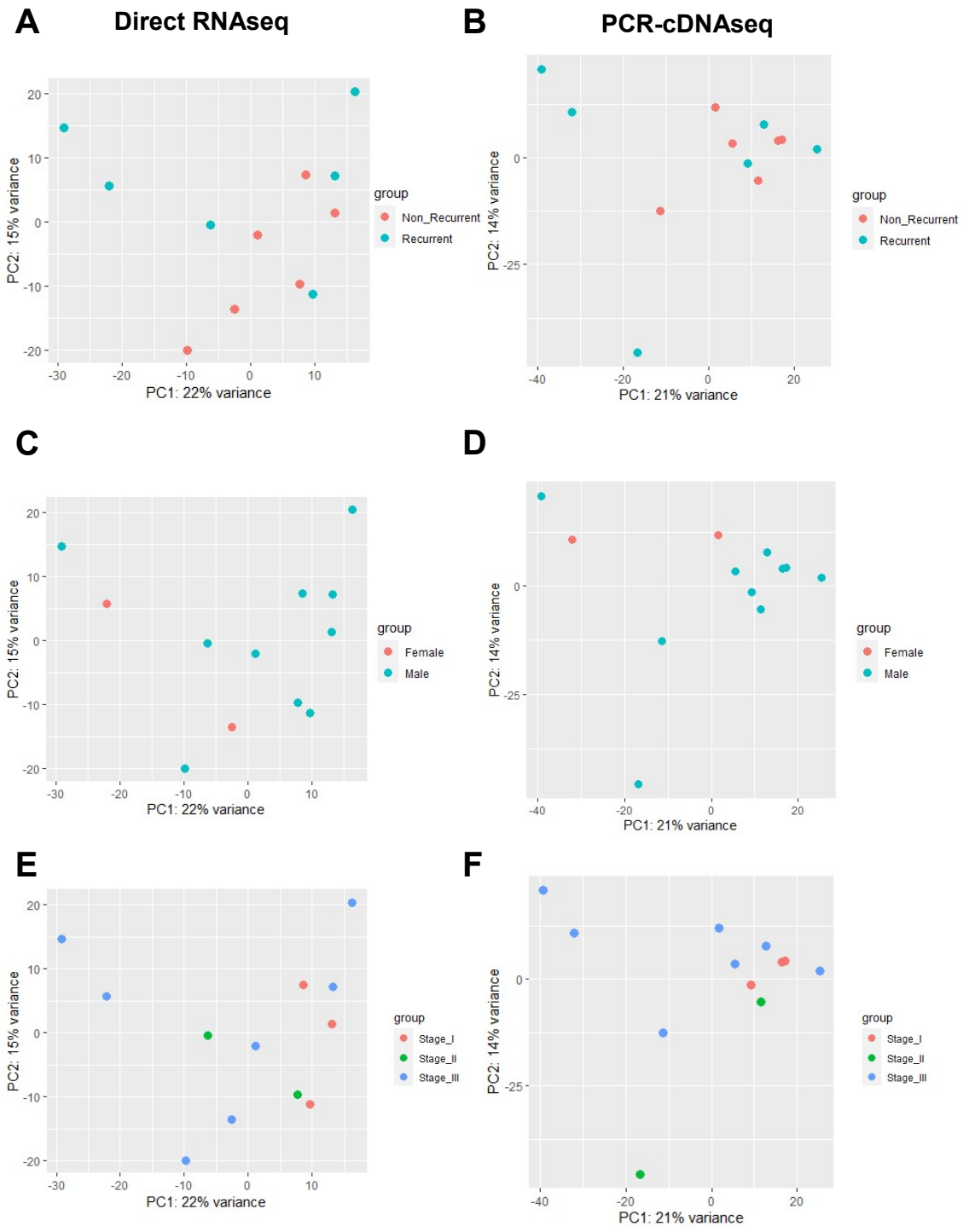


Figure 4.2: PCA analysis of ccRCC tumour transcriptome profiles using reference transcriptome aligned DRS and PCS

Principal component analysis (PCA) on ccRCC tumours gene expression data illustrating variations between samples (dots, n = 12). DESeq2 generated plots using reference transcriptome (Ensembl release 105, cDNA reference) aligned **A)** DRS and **B)** PCS data showing PCA of recurrent vs non-recurrent ccRCC groups; **C)** DRS and **D)** PCS data showing PCA of male vs female tumour samples; **E)** DRS and **F)** PCS data showing PCA of stage I, II and III cancer samples.

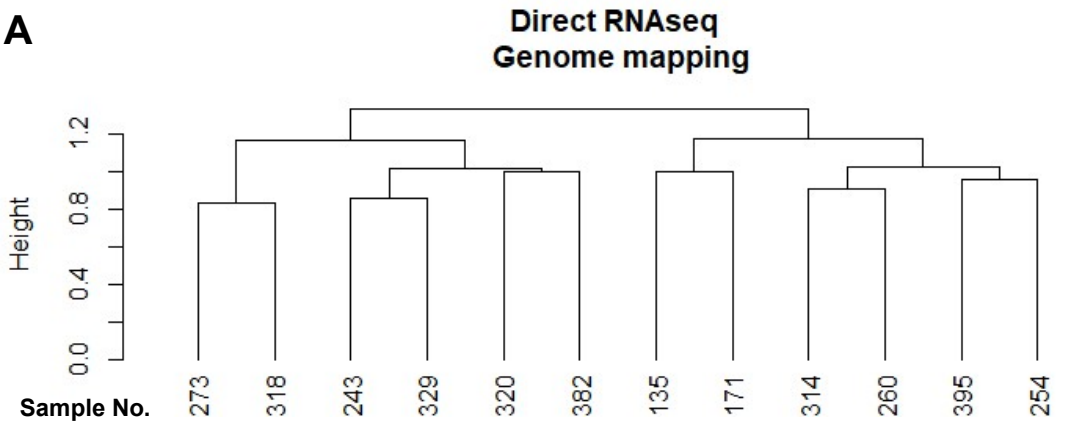
4.3.2 Unsupervised hierarchical clustering of ccRCC tumour expression profiles

Next, to evaluate if ccRCC tumour expression profiles could be grouped by their clinical features, unsupervised hierarchical clustering analysis was performed, and dendrograms were constructed to identify sample clusters based on spearman rank correlations of gene expression levels. Here, samples with the highest degree of similarities in gene expression profile were grouped and sequentially merged with other clusters. Finally, reference genome-aligned and transcriptome-aligned data were analysed to assess if the inclusion of non-coding RNAs affects the clustering of tumour samples.

The hierarchical clustering analysis showed that gene expression profiles cluster differently between DRS and PCS data. For example, in reference genome aligned DRS data, the sample with the highest degree of similarity in gene expression with tumour 135 is 171, followed by 395, 254, 314 and 260 (Figure 4.3A). In reference genome aligned PCS, tumour 135 is clustered with 395, followed by 320 and 243, and finally, 382 and 254 (Figure 4.3 B). In contrast, gene expression profiles clustered similarly when reference genome-aligned and reference transcriptome-aligned dendrograms were compared. All six initial sample clusters for PCS are paired identically between the reference genome and reference transcriptome-aligned data. Furthermore, the most significant clusters of 6 tumours are again matched for reference genome and reference transcriptome mapped PCS data (Figure 4.3B, 4.4B).

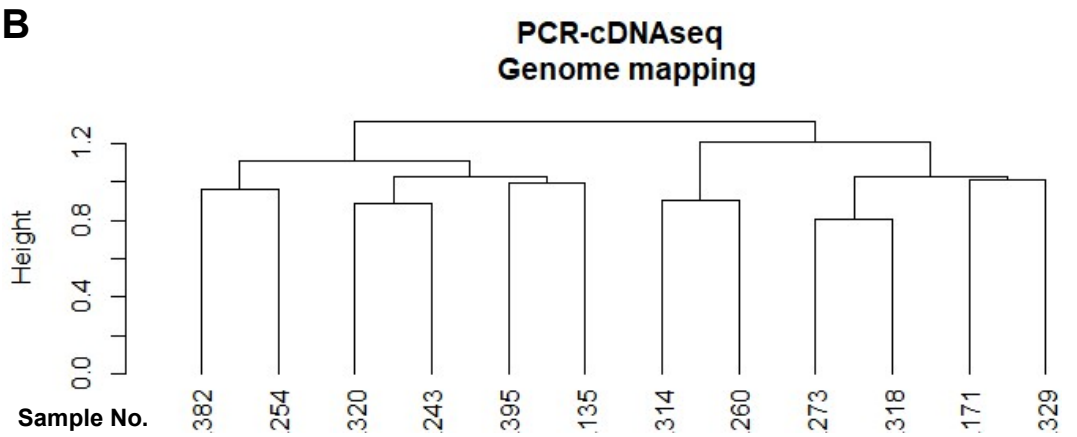
Analogous to observations from PCA scatter plots, hierarchical clustering did not result in segregated clusters of tumour samples that match with ccRCC recurrence status, patient gender, cancer stage, Fuhrman grade and Leibovich score for either reference genome or reference transcriptome aligned DRS and PCS data (Figure 4.3 – 4.4).

A



Recurrent	N	N	Y	Y	N	N	Y	Y	N	Y	N	Y
Sex	M	F	M	M	M	M	M	M	M	M	M	F
Stage	III	III	III	II	II	I	II	III	I	III	II	III
Grade	3	3	3	2	3	3	3	3	3	3	3	3
Leibovich Score	5	5	5	4	4	3	4	5	3	5	5	5

B

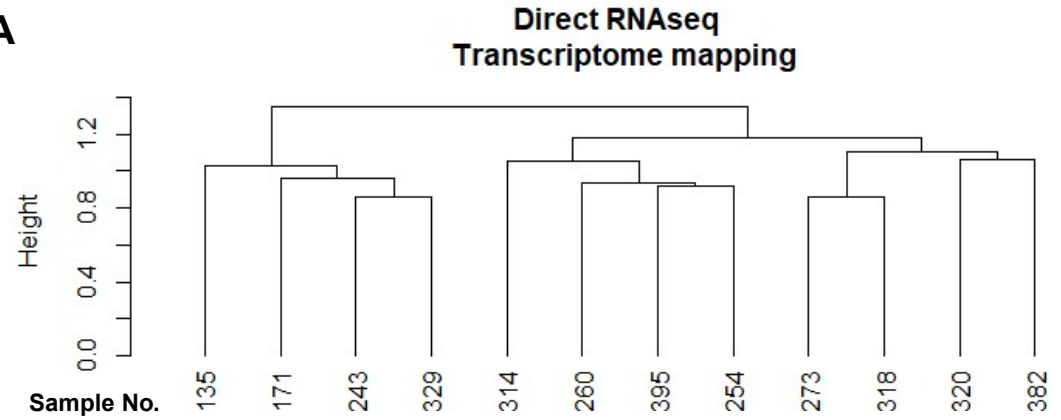


Recurrent	N	Y	N	Y	N	Y	N	Y	N	N	Y	Y
Sex	M	F	M	M	M	M	M	M	M	F	M	M
Stage	I	III	II	III	II	II	I	III	III	III	III	II
Grade	3	3	3	3	3	3	3	3	3	3	3	2
Leibovich Score	3	5	4	5	5	4	3	5	5	5	5	4

Figure 4.3: Hierarchical clustering of ccRCC transcriptome profile by reference genome mapped DRS and PCS

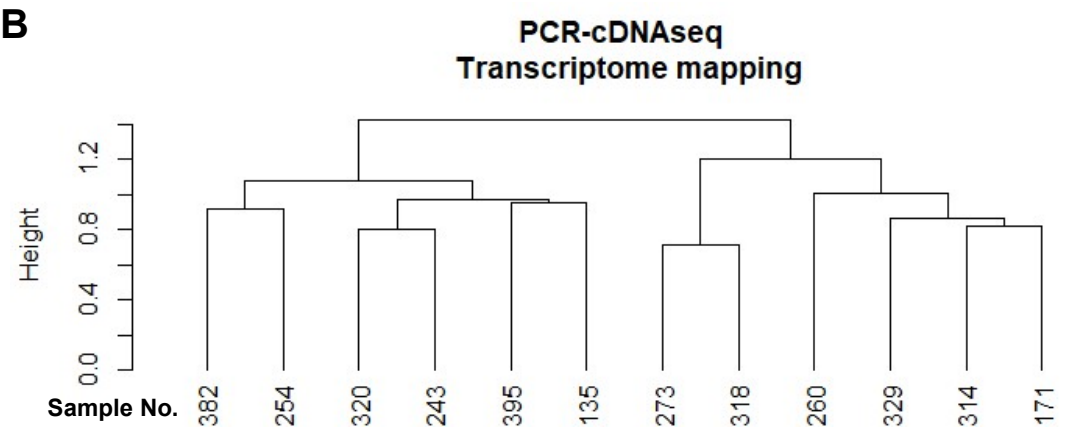
A) Dendrogram for hierarchical clustering of reference genome mapped ccRCC tumour DRS transcriptomes based on spearman rank correlations of gene expression levels. Patient information and clinical ccRCC features are listed below dendrogram. **B)** As in **A**, but for PCS.

A



Recurrent	Y	Y	Y	Y	N	Y	N	Y	N	N	N	N
Sex	M	M	M	M	M	M	M	F	M	F	M	M
Stage	II	III	III	II	I	III	II	III	III	III	II	I
Grade	3	3	3	2	3	3	3	3	3	3	3	3
Leibovich Score	4	5	5	4	3	5	5	5	5	5	4	3

B



Recurrent	N	Y	N	Y	N	Y	N	N	Y	Y	N	Y
Sex	M	F	M	M	M	M	M	F	M	M	M	M
Stage	I	III	II	III	II	II	III	III	III	II	I	III
Grade	3	3	3	3	3	3	3	3	3	2	3	3
Leibovich Score	3	5	4	5	5	4	5	5	5	4	3	5

Figure 4.4: Hierarchical clustering of ccRCC transcriptome profiles by reference transcriptome mapped DRS and PCS

A) Dendrogram for hierarchical clustering of reference transcriptome mapped ccRCC tumour DRS transcriptomes based on spearman rank correlations of gene expression levels. Patient information and clinical ccRCC features are listed below dendrogram. **B)** As in **A**, but for PCS.

4.3.3 Identification of differential expressed genes associated with ccRCC recurrence

Next, DRS and PCS-generated gene expression profiles were analysed using DESeq2 to identify differentially expressed genes between recurrent and non-recurrent ccRCC tumours (Love *et al.*, 2014). Genes with log_2FoldChange ≤ -2 or ≥ 2 and p_{adj} ≤ 0.1 are considered significant. Results of differential gene expression analysis between recurrent and non-recurrent ccRCC tumours were plotted as volcano plots in Figure 4.5. The top fifteen differentially expressed genes (ranked by p_{adj} values) in DRS and PCS (by reference genome and transcriptome alignment methods) are listed in Tables 4.1 and 4.2, respectively. Overall, both DRS and PCS identified DEGs. However, PCS identified a higher number of DEGs compared to DRS. Reference genome alignment also resulted in more DEGs than reference transcriptome alignment.

68 and 219 significant DEGs were discovered for reference genome aligned DRS and PCS, respectively, including 24/111 upregulated and 44/108 downregulated genes. For DRS, the top 3 differentially expressed genes in recurrent ccRCC tumours (by p_{adj} values) are *SORCS3* (Sortilin related VPS10 domain containing receptor 3, Log_2FC: 5.9758, p_{adj} = 2.32 x 10^{-8}), *TIGIT* (T cell immunoreceptor with Ig and ITIM domains, Log_2FC: -2.6793, p_{adj} = 5.25 x 10^{-4}) and *NKG7* (Natural killer cell granule protein 7). For reference genome-aligned PCS, the top 3 differentially expressed genes are lncRNAs that are not yet well-characterised: ENSG00000229740 (novel transcript, Log_2FC = 22.7129, p_{adj} = 5.73 x 10^{-10}), ENSG00000248515 (novel transcript, Log_2FC = 22.6095, p_{adj} = 5.73 x 10^{-10}) and ENSG00000276241 (novel transcript, Log_2FC = -4.8518, p_{adj} = 1.05 x 10^{-8}).

Reference transcriptome-aligned DRS and PCS identified 34 and 126 significant DEGs, including 10/57 upregulated and 24/69 downregulated genes. For reference transcriptome aligned DRS, the top 3 differentially expressed genes in recurrent ccRCC tumours (by p_{adj} values) are *SORCS3* (Log_2FC = 6.2563, p_{adj} = 4.89 x 10^{-8}), *NKG7* (Log_2FC = -2.7936, p_{adj} = 4.21 x 10^{-4}) and *GZMK* (Granzyme K, Log_2FC = -3.0141, p_{adj} = 7.38 x 10^{-4}). For PCS, the top 3 differentially expressed genes are *MUC17* (Mucin 17,

Log_2FC = 24.1211, p_{adj} = 1.59 x 10^{-11}), *CD8B* (CD8 antigen beta polypeptide, Log_2FC = -3.5878, p_{adj} = 1.66 x 10^{-4}) and *FAM83F* (Family with sequence similarity 83 member F, Log_2FC = -5.5527, p_{adj} = 3.65 x 10^{-4}). Comprehensive lists of all identified significant DEGs by DRS and PCS, aligned with reference genome and reference transcriptome, can be found in Appendix Tables 7.1 – 4.

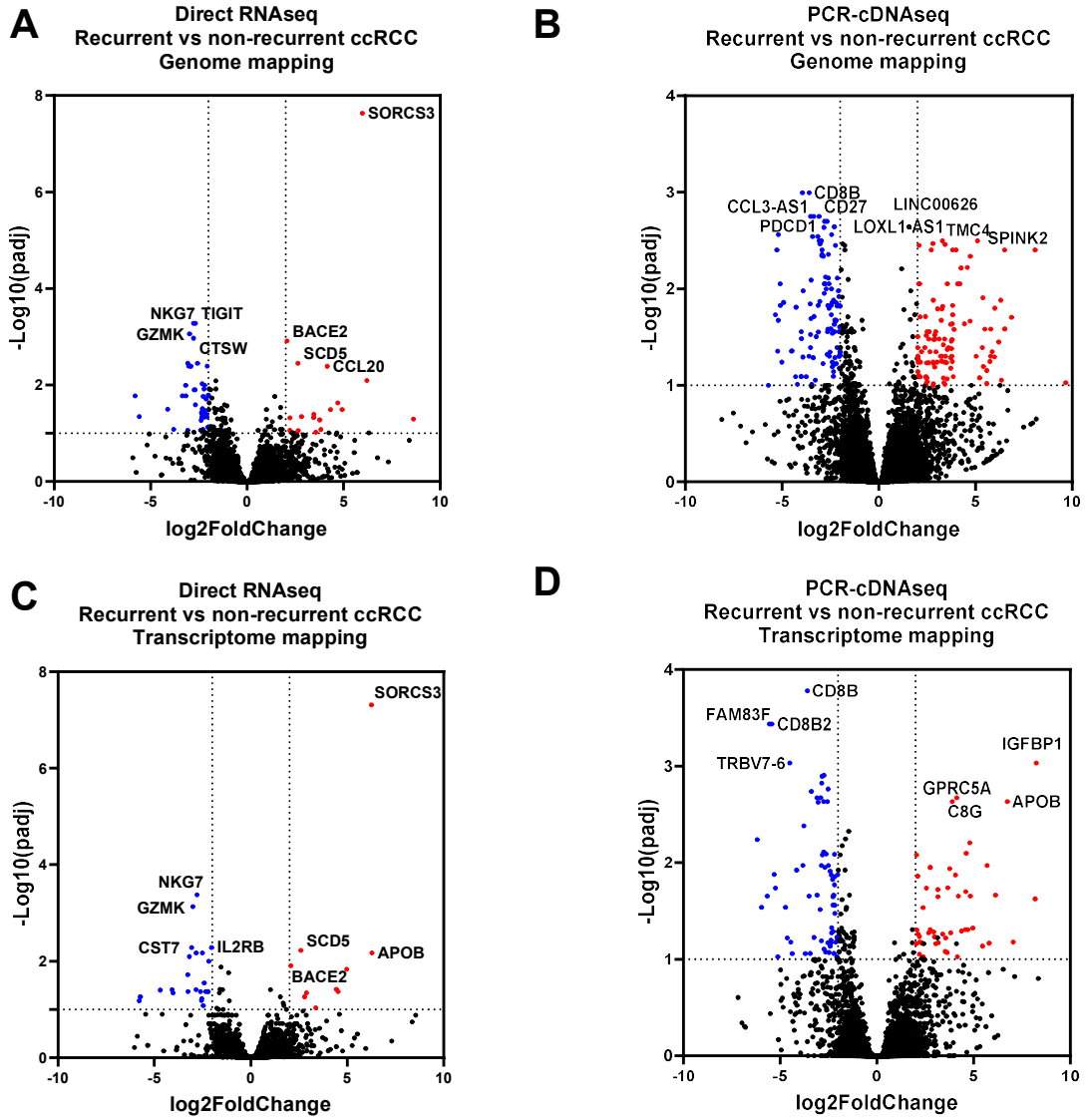


Figure 4.5: DGEs between recurrent and non-recurrent ccRCC tumours

Volcano plots showing differentially expressed genes between recurrent and non-recurrent ccRCC tumours (n = 6 per group) profiled by **A)** DRS by reference genome alignment, **B)** PCS by reference genome alignment, **C)** DRS by reference transcriptome alignment and **D)** PCS by reference transcriptome alignment. Dotted lines indicate significance threshold ($p_{adj} \leq 0.1$, $|log_2FoldChange| > 2$). Significantly upregulated genes are in red and downregulated genes are in blue. Names of top 4 most significantly up/down regulated annotated genes (by padj) are shown.

A

DRS: Reference genome mapping				
ENSEMBL ID	**Gene Name**	**Biotype**	**Log_2FoldChange**	**padj**
ENSG00000156395	SORCS3	protein_coding	5.98	2.32E-08
ENSG00000181847	TIGIT	protein_coding	-2.68	5.25E-04
ENSG00000105374	NKG7	protein_coding	-2.77	5.25E-04
ENSG00000113088	GZMK	protein_coding	-2.98	8.71E-04
ENSG00000172543	CTSW	protein_coding	-2.78	1.06E-03
ENSG00000182240	BACE2	protein_coding	2.07	1.22E-03
ENSG00000163508	EOMES	protein_coding	-3.07	3.54E-03
ENSG00000145284	SCD5	protein_coding	2.64	3.54E-03
ENSG00000277734	TRAC	TR_C_gene	-2.57	3.54E-03
ENSG00000204252	HLA-DOA	protein_coding	-2.07	4.06E-03
ENSG00000077984	CST7	protein_coding	-3.02	4.06E-03
ENSG00000115009	CCL20	protein_coding	4.15	4.07E-03
ENSG00000153563	CD8A	protein_coding	-2.92	4.07E-03
ENSG00000084674	APOB	protein_coding	6.21	8.01E-03
ENSG00000101082	SLA2	protein_coding	-2.32	9.37E-03

B

DRS: Reference transcriptome mapping				
ENSEMBL ID	**Gene Name**	**Biotype**	**Log_2FoldChange**	**padj**
ENSG00000156395	SORCS3	protein_coding	6.26	4.89E-08
ENSG00000105374	NKG7	protein_coding	-2.79	4.21E-04
ENSG00000113088	GZMK	protein_coding	-3.01	7.38E-04
ENSG00000077984	CST7	protein_coding	-3.07	5.24E-03
ENSG00000100385	IL2RB	protein_coding	-2.04	5.24E-03
ENSG00000145284	SCD5	protein_coding	2.59	5.99E-03
ENSG00000084674	APOB	protein_coding	6.28	6.74E-03
ENSG00000163564	PYHIN1	protein_coding	-2.83	6.74E-03
ENSG00000277734	TRAC	TR_C_gene	-2.51	6.74E-03
ENSG00000139193	CD27	protein_coding	-3.18	8.06E-03
ENSG00000145649	GZMA	protein_coding	-2.18	1.00E-02
ENSG00000182240	BACE2	protein_coding	2.07	1.25E-02
ENSG00000115009	CCL20	protein_coding	4.96	1.46E-02
ENSG00000089692	LAG3	protein_coding	-3.26	1.90E-02
ENSG00000147138	GPR174	protein_coding	-2.42	2.81E-02

Table 4.1: Top 15 differentially expressed genes between recurrent and non-recurrent ccRCC tumours by DRS

A) List of top 15 differentially expressed genes (by p_{adj} values) between recurrent and non-recurrent ccRCC tumours profiled by reference genome mapped DRS. **B)** As in **A** but for reference transcriptome mapped DRS.

A

PCS: Reference genome mapping				
ENSEMBL ID	**Gene Name**	**Biotype**	**Log_2FoldChange**	**padj**
ENSG00000229740		lncRNA	22.71	5.73E-10
ENSG00000248515		lncRNA	22.61	5.73E-10
ENSG00000276241		lncRNA	-4.85	1.05E-08
ENSG00000172116	CD8B	protein_coding	-3.60	1.01E-03
ENSG00000234261		lncRNA	-3.95	1.01E-03
ENSG00000188389	PDCD1	protein_coding	-3.38	1.78E-03
ENSG00000139193	CD27	protein_coding	-3.10	1.78E-03
ENSG00000277089	CCL3-AS1	lncRNA	-3.52	1.78E-03
ENSG00000163508	EOMES	protein_coding	-2.81	2.00E-03
ENSG00000163606	CD200R1	protein_coding	-2.68	2.00E-03
ENSG00000277632	CCL3	protein_coding	-2.30	2.28E-03
ENSG00000105374	NKG7	protein_coding	-2.84	2.31E-03
ENSG00000147138	GPR174	protein_coding	-2.45	2.75E-03
ENSG00000133477	FAM83F	protein_coding	-5.19	2.75E-03
ENSG00000049249	TNFRSF9	protein_coding	-3.15	2.88E-03

B

PCS: Reference transcriptome mapping				
ENSEMBL ID	**Gene Name**	**Biotype**	**Log_2FoldChange**	**padj**
ENSG00000169876	MUC17	protein_coding	24.12	1.59E-11
ENSG00000172116	CD8B	protein_coding	-3.59	1.66E-04
ENSG00000133477	FAM83F	protein_coding	-5.55	3.65E-04
ENSG00000229295	HLA-DPB1	protein_coding	10.19	3.65E-04
ENSG00000254126	CD8B2	protein_coding	-5.43	3.65E-04
ENSG00000146678	IGFBP1	protein_coding	8.24	9.28E-04
ENSG00000211727	TRBV7-6	TR_V_gene	-4.50	9.28E-04
ENSG00000147570	DNAJC5B	protein_coding	-2.73	1.24E-03
ENSG00000105374	NKG7	protein_coding	-2.82	1.27E-03
ENSG00000113088	GZMK	protein_coding	-2.85	1.50E-03
ENSG00000211772	TRBC2	TR_C_gene	-2.52	1.72E-03
ENSG00000177494	ZBED2	protein_coding	-3.38	1.83E-03
ENSG00000013588	GPRC5A	protein_coding	4.12	2.13E-03
ENSG00000117560	FASLG	protein_coding	-2.90	2.13E-03
ENSG00000139193	CD27	protein_coding	-3.10	2.13E-03

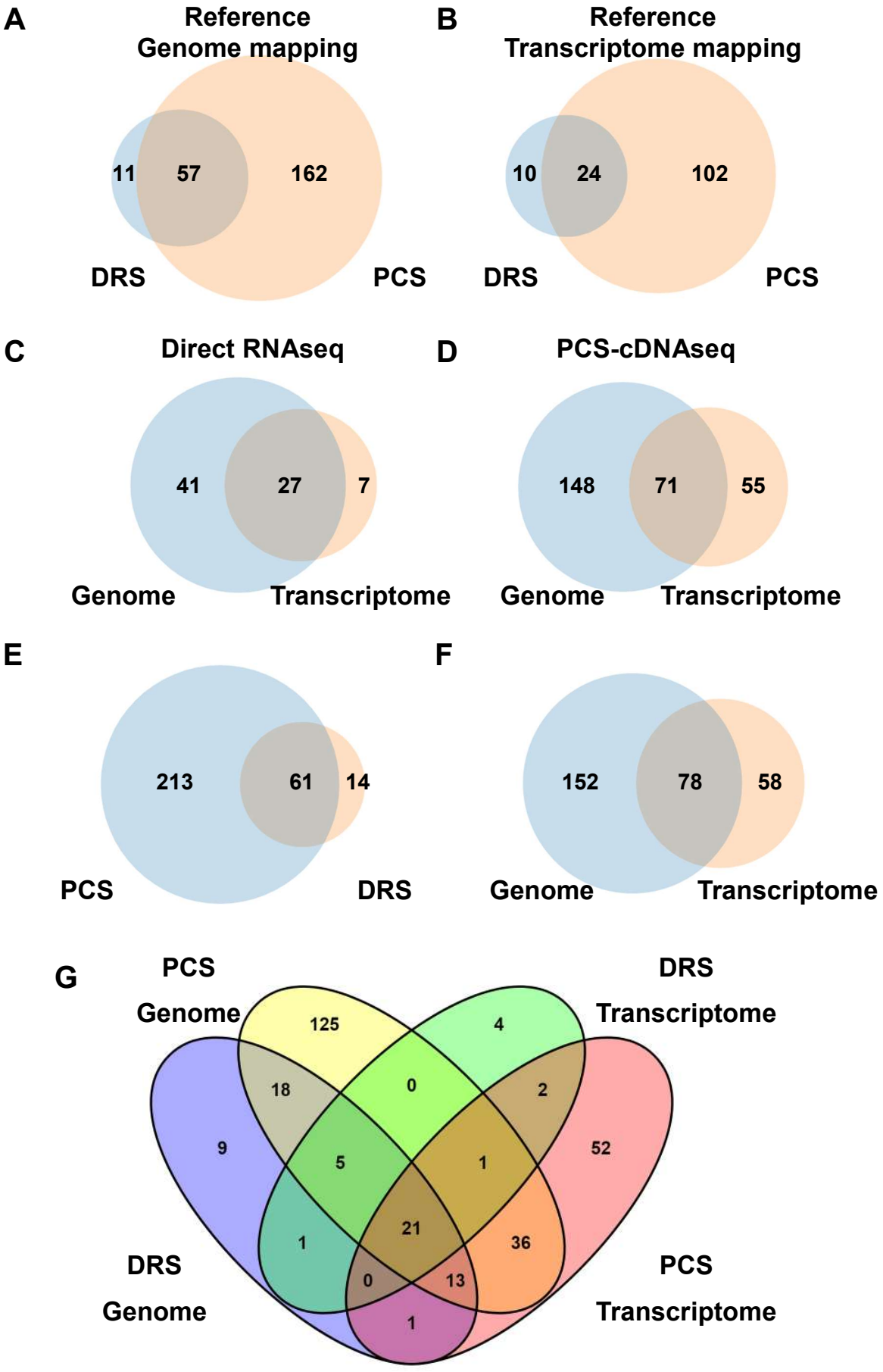
Table 4.2: Top 15 differentially expressed genes between recurrent and non-recurrent ccRCC tumours by PCS

A) List of top 15 differentially expressed genes (by p_{adj} values) between recurrent and non-recurrent ccRCC tumours profiled by reference genome mapped PCS. **B)** As in **A** but for reference transcriptome mapped PCS.

4.3.5 Characterisation of differential expressed genes associated with ccRCC recurrent status

To understand the data similarity of DEG analysis from DRS/PCS and reference genome/transcriptome, Venn digrams were constructed for the overlaps between recurrent vs non-recurrent ccRCC DEGs identified by DRS and PCS via reference genome and reference transcriptome alignment. Of the 68 DEGs identified by reference genome aligned DRS, 57 genes were also differentially expressed in PCS (Figure 4.6A). For the 34 DEGs identified by reference transcriptome-mapped DRS, 24 genes were also found to be differentially expressed in PCS (Figure 4.6B). A high level of overlaps was also found between DEGs discovered by reference genome or reference transcriptome alignment. For DRS, of the 34 DEGs found in reference transcriptome-mapped data, 27 were also identified as DEGs in reference genome-mapped data (Figure 4.6C).

Similarly, for PCS, of the 126 DEGs found in reference transcriptome-mapped data, 71 genes were found to be differentially expressed in reference genome-mapped data (Figure 4.6D). PCS identified 274 DEGs between recurrent and non-current ccRCC tumour samples, and DRS identified 75 DEGs. Of the 75 DRS-identified DEGs, 61 were also identified in PCS (Figure 4.6E). Reference genome alignment from DRS and PCS identified 230 DEGs, and reference transcriptome alignment from DRS and PCS identified 136 DEGs. 78 DEGs were determined by both alignment methods (Figure 4.6F). Overall, 21 DEGs were found in both reference genome and transcriptome-aligned DRS and PCS data (Figure 4.6G).



Total no. of DEGs: 288
Common DEGs : 21

Figure 4.6: Common disease recurrence associated DEGs identified by DRS and PCS of ccRCC tumours

A) Venn diagram showing overlaps of DEGs ($p_{adj} \leq 0.1, |log_2FoldChange| > 2$) identified by reference genome aligned DRS and PCS. **B)** as in **A**, but aligned with reference transcriptome. **C)** Venn diagram showing overlaps of DEGs identified by reference genome or transcriptome aligned DRS. **D)** as in **C**, but for PCS. **E)** Venn diagram showing overlaps of all DEGs identified by DRS and PCS. **F)** Venn diagram showing overlaps of all DEGs identified by reference genome and reference transcription aligned DRS and PCS. **G)** Venn diagram showing overlaps of DEGs identified by both reference genome and transcriptome aligned DRS and PCS.

Previously, mapping data illustrated that PCS identified a substantially higher number of mapped genes (39115 genes) compared to DRS (26457 genes) (Figure 3.14A - B). In addition, reference genome mapping resulted in more uniquely mapped genes than the reference transcriptome (Figure 3.15A). Also, both reference genome and transcriptome-aligned DRS and PCS data contain genes uniquely mapped by each method (Figure 3.16A). Therefore, to assess if the discrepancies in DEG numbers identified were due to the ability to detect the genes, stacked bar graphs were generated to show the number of DEGs mapped by both and solely by each method. The results showed that all 75 DEGs identified by DRS and 264 out of 274 identified by PCS (via both reference genome and transcriptome alignment) were mapped by both DRS & PCS (Figure 4.7A).

In contrast, of the 230/196 DEGs identified by reference genome or transcriptome-mapped DRS and PCS, only 154/136 DEGs were mapped by both methods. Finally, biotypes of identified DEGs were assessed. The results indicate that most DEGs identified by reference genome and transcriptome-mapped DRS and PCS are protein-coding genes. Whilst reference genome aligned PCS identified 64 differentially expressed lncRNAs, only six lncRNAs were identified in reference genome aligned DRS (Figure 4.7C). Of the 76 DEGs that were exclusively mapped via reference genome mapping, 65 genes are lncRNAs. In contrast, of the 60 exclusively reference transcriptome-mapped DEGs, 51 genes are protein-coding (Figure 4.7D), demonstrating their differential potential in identifying DEGs.

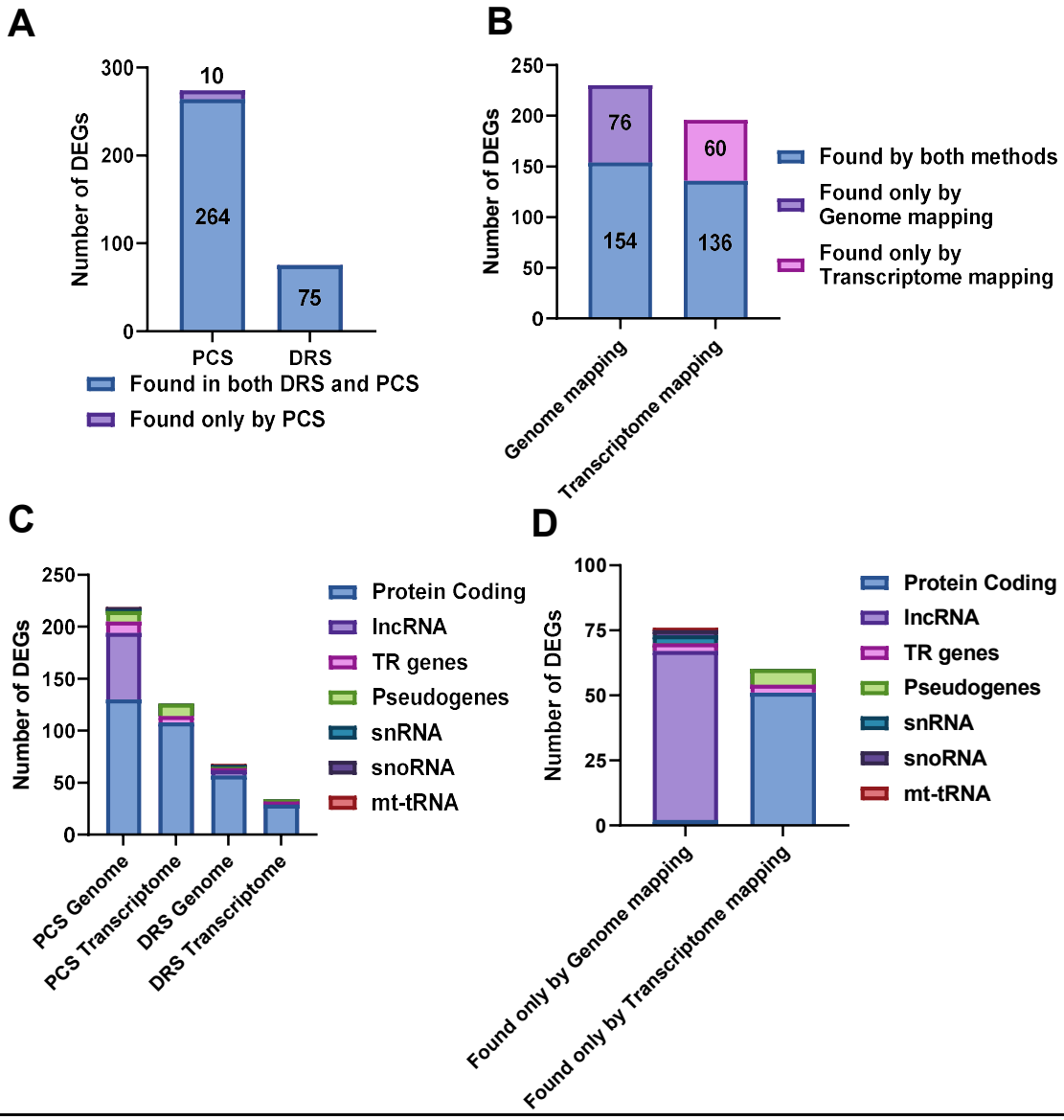


Figure 4.7: Biotype characterisation of DEGs between recurrent and non-recurrent ccRCC identified by DRS and PCS

A) Stacked bar graph showing number of DEGs identified by PCS and DRS, and the proportion of genes are present in both DRS and PCS (blue) or present only in PCS data (purple) **B)** Stacked bar graph showing proportion of DEGs identified by either reference genome or reference transcriptome mapping of DRS and PCS data. Proportion of DEGs present in both reference genome and transcriptome mapped DRS/PCS data is in blue, DEGs that are only present in reference genome mapped DRS/PCS data is in purple, and DEGs that are only present in reference transcriptome mapped DRS/PCS data in in pink. **C)** Stacked bar graph showing number of DEGs identified by reference genome/reference transcriptome aligned DRS and PCS and biotypes distribution. **D)** Stacked bar graph showing number of DEGs that are present only when DRS and PCS data aligned to reference genome/reference transcriptome, and the distribution of their biotypes.

4.3.6 Concordant gene expression patterns of ccRCC recurrence associated DEGs between alignment and sequencing methods

Following analysis of the relationship between gene mapping and identification of DEGs, the gene expression patterns of identified DEGs were evaluated. Both the effects of reference genome vs transcriptome alignment and between DRS and PCS were analysed. Firstly, for the reference alignments, of the 75 DEGs between non-recurrent and recurrent ccRCC tumours that DRS identified, 57 were identified by both reference genome and reference transcriptome alignment (Figure 4.8A). For PCS, of the 274 DEGs, 182 genes were identified by both reference genome and reference transcriptome mapping (Figure 4.8B).

The Log_2FoldChange in DEGs gene expression levels from reference genome-aligned significantly correlated with reference transcriptome-aligned DRS data, as analysed by DESeq2 ($R^2 = 0.8819$, $p < 0.0001$) (Figure 4.8C). Likewise, a high degree of concordance can be found between Log_2FoldChange in DEGs from reference genome-aligned and reference transcriptome-aligned PCS data ($R^2 = 0.8286$, $p < 0.0001$) (Figure 4.8D). Although several genes showed substantial levels of Log_2FoldChange by one reference alignment method yet close to zero for another, most DEGs exhibit the same directionality in their Log_2FoldChange in gene expression. Only one outlier, the ribosomal pseudogene *RPL7P60,* was found to be downregulated in recurrent ccRCC when aligned with reference transcriptome but upregulated when aligned with reference genome in PCS data (Figure 4.8D).

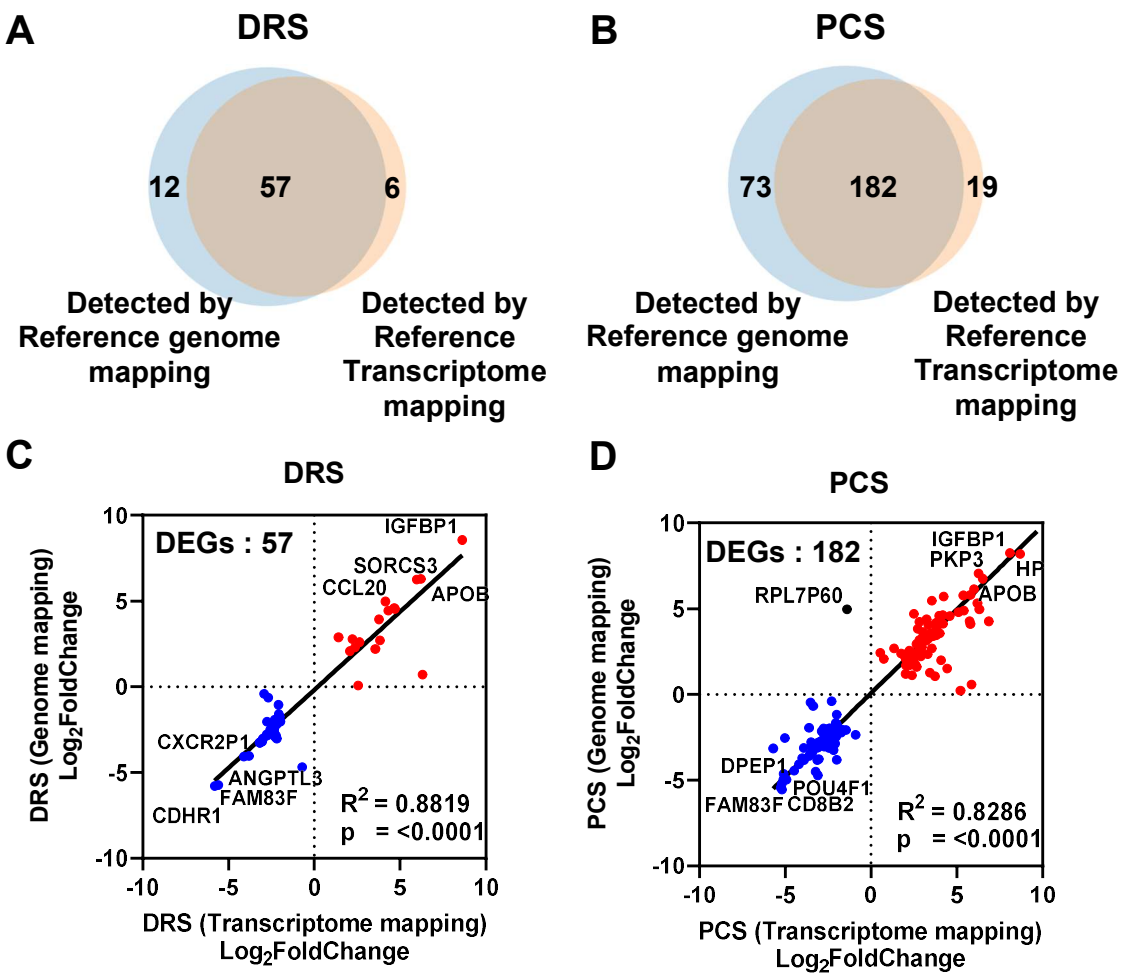


Figure 4.8: Evaluation of DEGs expression levels profiled by reference genome and reference transcriptome alignment

A) Venn diagram showing overlaps of DEGs ($p_{adj} \leq 0.1, |log_2FoldChange| > 2$) that can be found by reference genome and reference transcriptome aligned DRS. **B)** Ss in **A**, but for PCS. **C)** Correlation between $log_2FoldChange$ of commonly found DEGs (recurrent vs non recurrent ccRCC) from reference genome mapped- and reference transcriptome mapped- DRS. Top 4 up regulated and downregulated genes by averaged $Log_2FoldChange$ are indicated. **D)** Correlation between $log_2FoldChange$ of commonly found DEGs (recurrent vs non recurrent ccRCC) from reference genome mapped- and reference transcriptome mapped- PCS. Top 4 up regulated and downregulated genes by averaged $Log_2FoldChange$ are indicated. Throughout, diagonal lines represent the line of best fit. R^2 values were computed to measure goodness-of-fit, and P values generated from F-test, with $p \leq 0.05$ considered statistically significant.

Next, the effects of differential sequencing methods (DRS and PCS) on DEGs' expression patterns were evaluated. Strikingly, for the DEGs that were identified by either reference genome or reference transcriptome mapping, the majority of genes were detected by both DRS and PCS. Of the 230 DEGs identified by reference genome mapping, 223 were detected by both DRS and PCS (Figure 4.9A). For the 137 DEGs identified via reference transcriptome alignment, 132 were detected by both DRS and PCS (Figure 4.9B).

Log_2FoldChange in DEGs gene expression levels are highly concordant between PCS and DRS. Strong correlations of DEGs Log_2FoldChange between reference genome aligned DRS and PCS (R^2 = 0.8710, p = < 0.0001) and between reference transcriptome-aligned DRS and PCS (R^2 = 0.7196, p = < 0.0001) (Figure 4.9C – D). Interestingly, in contrast to the correlation dot plots between reference genome vs transcript alignment, here more genes spread towards x = 0 (Log_2FoldChange for DRS) for both upregulated DEGs (red) and downregulated DEGs (blue). This suggests that many of these DEGs are only shown to be differentially expressed in PCS, not DRS. Amongst all the DEGs identified by both PCS and DRS, all genes except for two exhibited the same directionality of Log_2FoldChange. For reference genome aligned data, the mitochondrial tRNA *MT-TS1* was found to be differentially downregulated in recurrent ccRCC tumours by DRS but not by PCS (Figure 4.9C). For reference transcriptome aligned data, the chemokine *CCL5* was identified to be downregulated in recurrent ccRCC tumours by DRS (log_2FoldChange = 0.537, p_{adj} = 0.863), and not by PCS (Log_2FoldChange = -2.684, p_{adj} = 0.0112) (Figure 4.9D).

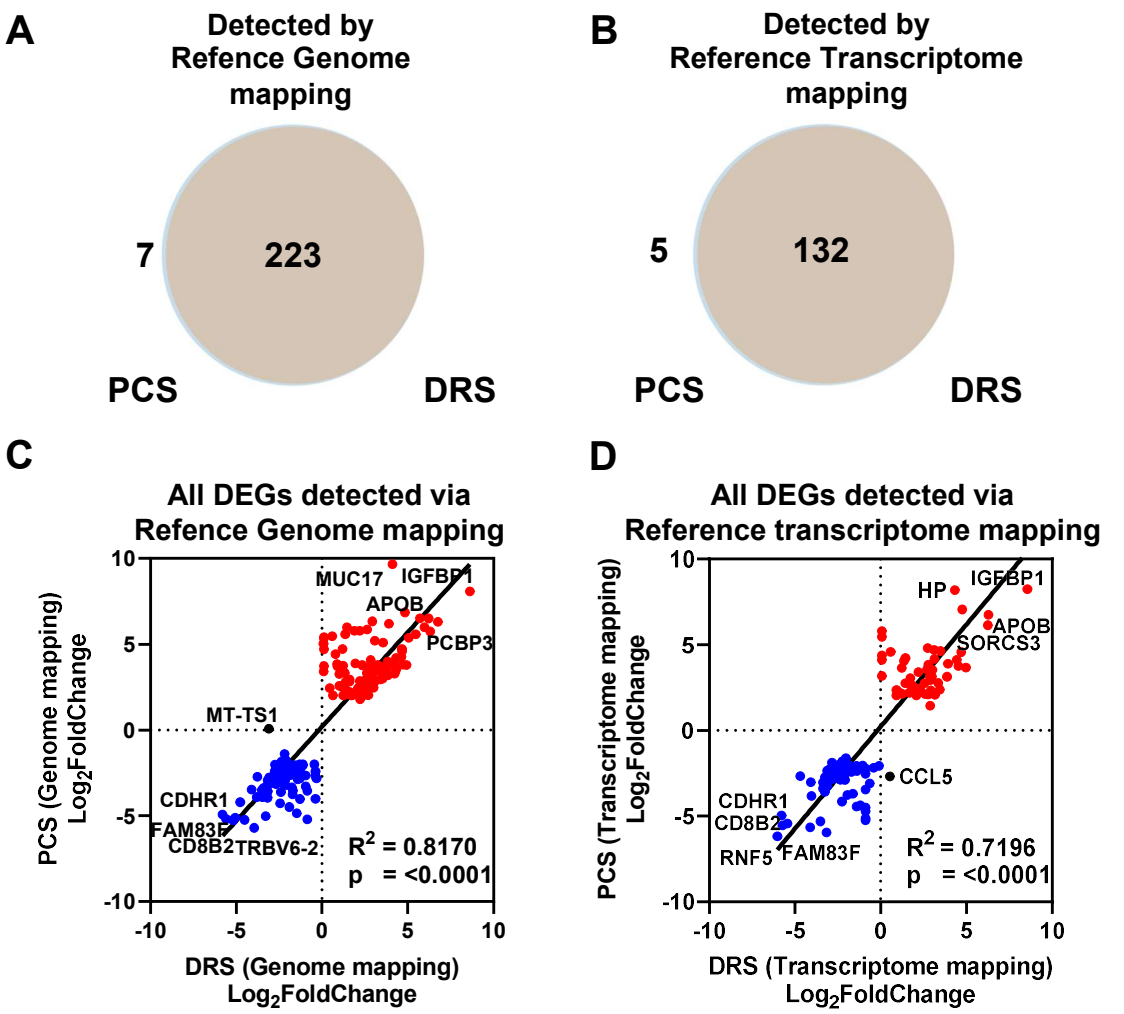


Figure 4.9: Characterisation of DEGs expression levels profiled by DRS and PCS

A) Venn diagram showing overlaps of DEGs ($p_{adj} \leq 0.1, |log_2FoldChange| > 2$) that can be found by reference genome mapped PCS and DRS. **B)** As in **A**, but PCS and DRS were aligned to reference transcriptome. **C)** Correlation between $log_2FoldChange$ of commonly found DEGs (recurrent vs non recurrent ccRCC) from reference genome aligned DRS and PCS. Top 4 up regulated and downregulated genes by averaged $Log_2FoldChange$ are indicated. **D)** Correlation between $log_2FoldChange$ of commonly found DEGs (recurrent vs non recurrent ccRCC) from reference transcriptome mapped PCS and DRS. Top 4 up regulated and downregulated genes by averaged $Log_2FoldChange$ are indicated. Throughout, diagonal lines represent the line of best fit. R^2 values were computed to measure goodness-of-fit, and p values generated from F-test, with $p \leq 0.05$ considered statistically significant.

4.3.7 GSEA GO BP analysis reveals suppression of adaptive immune response-related pathways in recurrent ccRCC tumours

Gene set enrichment analysis (GSEA) by clusterprofiler (v4.0) was conducted to systematically evaluate differences in biological processes and pathways between recurrent and non-recurrent ccRCC tumours (Wu *et al.*, 2021). Genes were ranked using log_2Foldchange values of DESeq2 normalised gene expression from all detected genes from the ccRCC tumours (n = 12), profiled by reference transcriptome aligned DRS and PCS. GSEA using the gene ontology (GO) biological processes (BP) database reveals that the top 10 most significantly enriched (by padj values) GO BP terms, both DRS and PCS are immune system related, with a high degree of overlapping genes between enriched gene sets as demonstrated by the linkages in the GO BP enrichment maps (Figure 4.10A, 4.11A). The overlapping terms between the top 10 enriched GO BP from DRS, and PCS include regulation of leukocyte differentiation, T cell differentiation, regulation of T cell activation and T cell activation. The list of significantly enriched GO BP terms for DRS and PCS data can be found in Appendix tables 7.5 - 6, respectively.

The top 10 enriched GO BP terms from the results of GSEA between recurrent and non-recurrent ccRCC tumours are also shown in the dot plots in Figure 4.10B and 4.11B for DRS and PCS, respectively. Dot sizes in the dot plot signify the number of overlapping genes between differentially expressed genes and respective GO BP terms, and the x-axis shows the proportion of represented genes from the GO BP term. Strikingly, all top 10 enriched GO BP terms (by p_{adj}) from both DRS and PCS had negative normalised enrichment scores (NES), signifying that the gene sets were significantly suppressed in recurrent ccRCC tumours compared to non-recurrent counterparts (Appendix table 7.6). The same trend was also shown by GSEA enrichment plots for the top 5 enriched GO BP terms (by p_{adj}) in DRS and PCS data, where represented genes were concentrated towards the lower end of the ranked gene list (Figure 4.10C, 4.11C). Overall, GSEA reveals significant suppression of adaptive immune response-related pathways in recurrent ccRCC tumours compared to non-recurrent tumours.

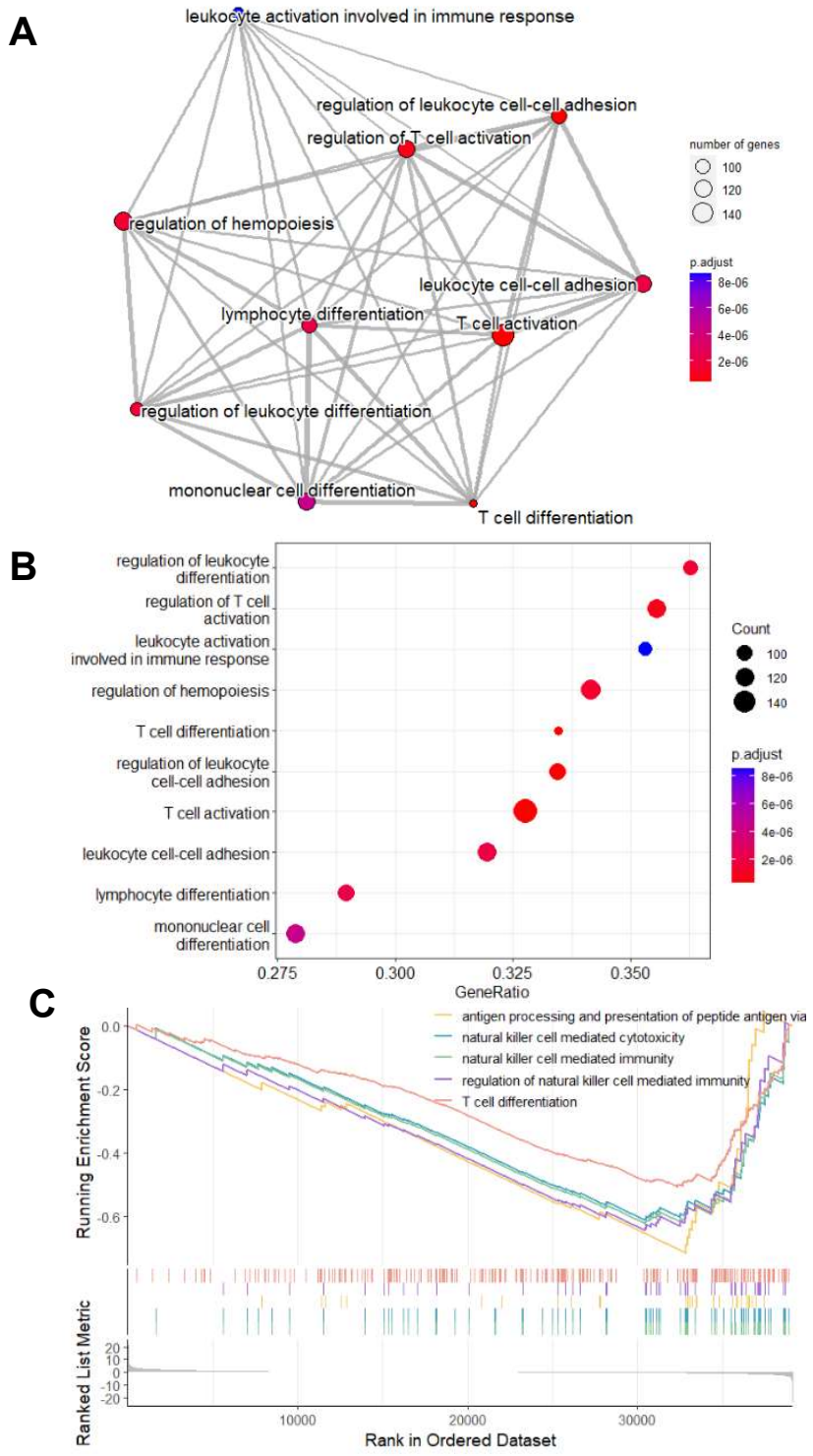


Figure 4.10: Gene Ontology Biological Process (GO:BP) GSEA for ccRCC recurrence associated differential gene expression profiled by DRS

A) GO:BP enrichment map showing top 10 enriched terms associated ccRCC recurrence associated differential gene expression profiled by reference transcriptome mapped DRS. **B)** Dot plot showing top 10 enriched GO:BP terms, with dot size representing gene count per term and colour reflecting p_{adj} value. **C)** GSEA enrichment plot for the top 5 enriched GO:BP terms. The x-axis shows genes represented in each pathway, and the y-axis shows enrichment scores.

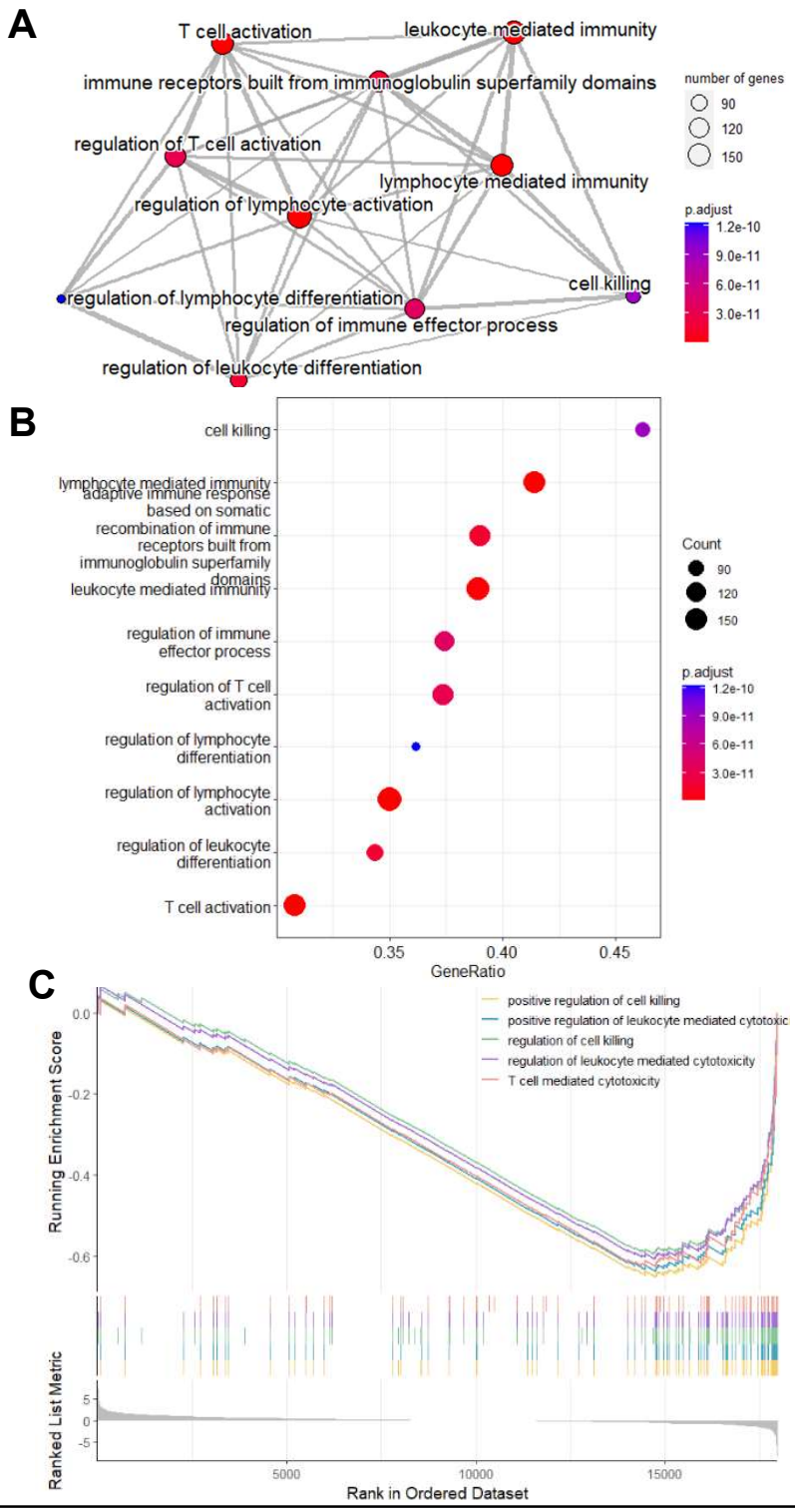


Figure 4.11: Gene Ontology Biological Process (GO:BP) GSEA for ccRCC recurrence associated differential gene expression profiled by PCS

A) GO:BP enrichment map showing top 10 enriched terms associated ccRCC recurrence associated differential gene expression profiled by reference transcriptome mapped PCS. **B)** Dot plot showing top 10 enriched GO:BP terms, with dot size representing gene count per term and colour reflecting p_{adj} value. **C)** GSEA enrichment plot for the top 5 enriched GO:BP terms. The x-axis shows genes represented in each pathway, and the y-axis shows enrichment scores.

4.3.8 GSEA GO MF and GO CC analysis show repression of antigen presentation pathways in recurrent ccRCC tumours

Next, GSEA was performed with GO Molecular Function (MF) and Cellular Compartments (CC) terms to interrogate differentially expressed pathways from recurrent ccRCC tumours compared to non-recurrent controls. GSEA using the GO MF database illustrated that differentially expressed genes between recurrent and non-recurrent ccRCC tumours are highly enriched with proteins involved in antigen presentation pathways for both DRS and PCS data. Amongst the top ten most enriched GO MF terms (by p_{adj}) from reference transcriptome-aligned DRS data, eight are related to MHC binding and antigen presentation pathways, including peptide antigen binding, immune receptor activity, antigen binding, MHC class II receptor activity, peptide binding, TAP binding, MHC protein complex binding and MHC class II protein complex binding (Figure 4.12A, Appendix table 7.7). GSEA GO MF also identified several significantly enriched and activated pathways, including extracellular matrix structure constituent and cadherin binding (Figure 4.12B).

Similar results were observed when GSEA was conducted using the GO CC database, where the top 10 most enriched (by p_{adj} values) GO CC terms are all antigen presentation pathway related (Figure 4.13A, Appendix tables 7.9 – 10). Like findings from GO MF analysis, these enriched pathways were found to be significantly suppressed in the recurrent ccRCC tumours compared to non-recurrent tumours, as shown in the dot plots in Figure 4.12B and 4.13B, as well as GSEA enrichment plots in Figure 4.12C and 4.13C. GSEA GO MF and GO CC enrichment analysis demonstrated a potential differential antigenic landscape between recurrent and non-recurrent ccRCC tumours. The complete list of significantly enriched GO MF and GO CC terms for both DRS and PCS data can be found in Appendix tables 7.7 – 10.

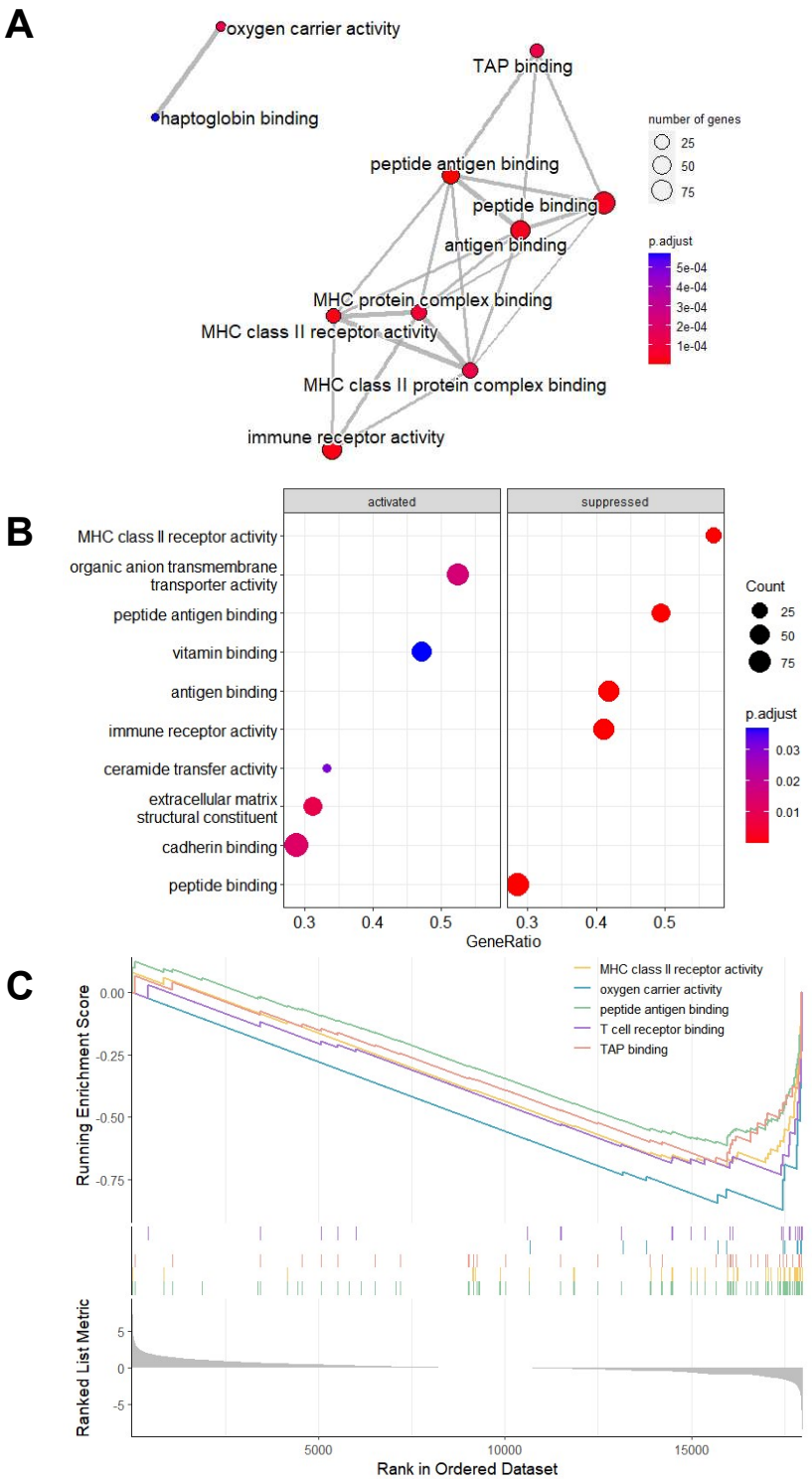


Figure 4.12: Gene Ontology Molecular Function (GO:MF) GSEA for ccRCC recurrence associated differential gene expression

A) GO:MF enrichment map showing top 10 enriched terms associated ccRCC recurrence associated differential gene expression profiled by reference transcriptome mapped DRS. **B)** Dot plot showing top 10 enriched GO:MF terms, with dot size representing gene count per term and colour reflecting p_{adj} value. **C)** GSEA enrichment plot for the top 4 enriched GO:MF terms. The x-axis shows genes represented in each pathway, and the y-axis shows enrichment scores.

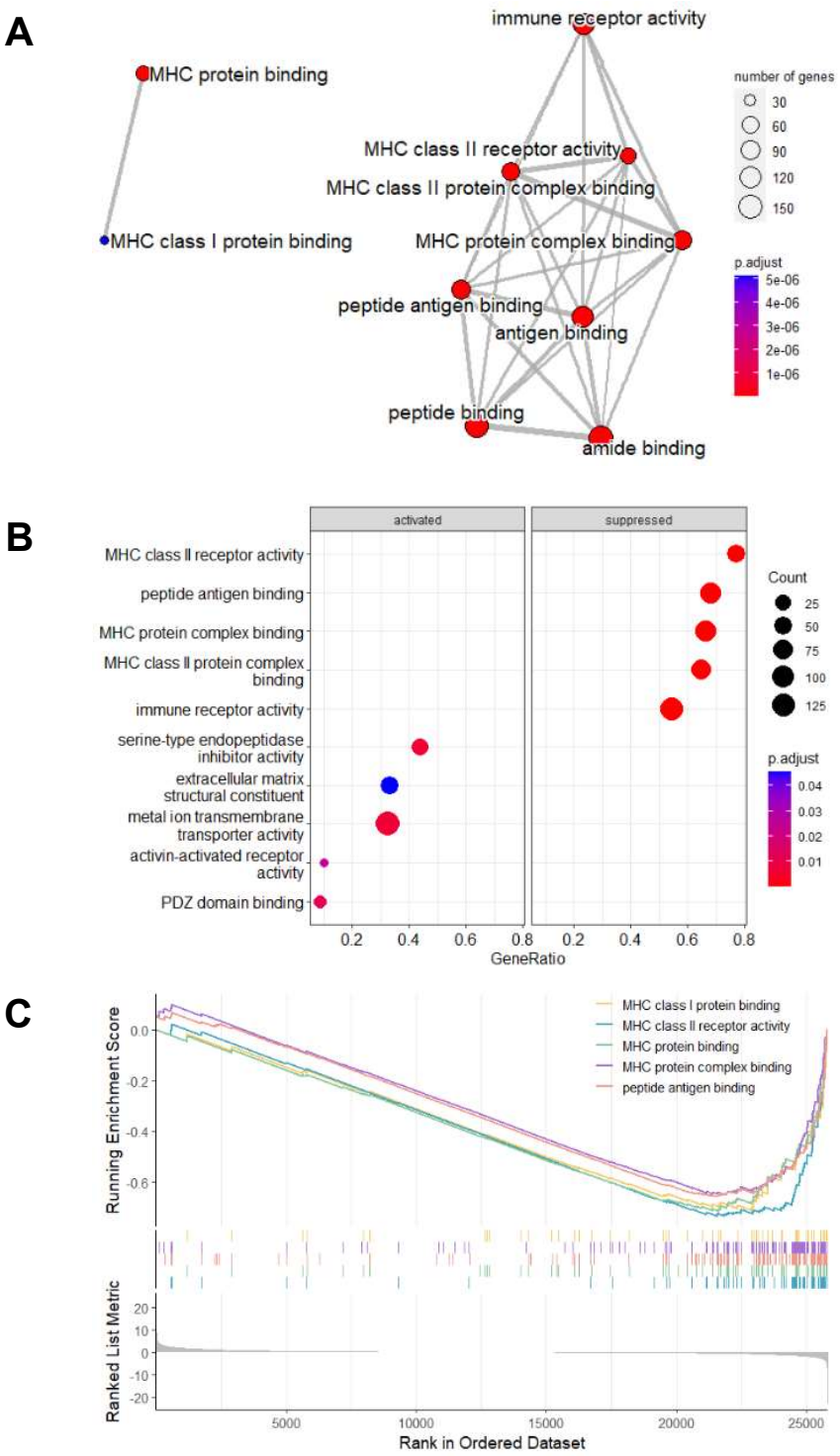


Figure 4.13: Gene Ontology Cellular Compartment (GO:CC) GSEA for ccRCC recurrence associated differential gene expression

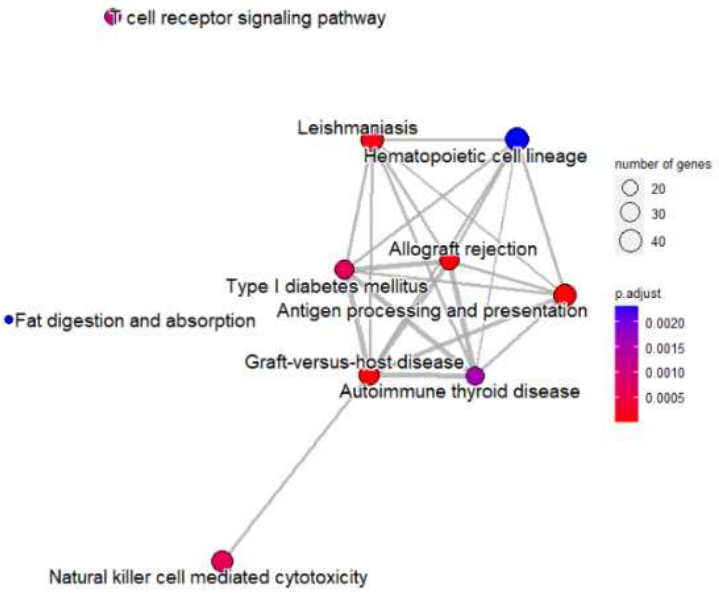
A) GO:CC enrichment map showing top 10 enriched terms associated ccRCC recurrence associated differential gene expression profiled by reference transcriptome mapped PCS. **B)** Dot plot showing top 10 enriched GO:CC terms, with dot size representing gene count per term and colour reflecting p_{adj} value. **C)** GSEA enrichment plot for the top 5 enriched GO:CC terms. The x-axis shows genes represented in each pathway, and the y-axis shows enrichment scores.

4.3.9 GSEA KEGG pathway enrichment analysis indicates differential lipid metabolism in recurrent ccRCC

GSEA KEGG pathway analysis identified 9 and 26 significantly enriched pathways from the differentially expressed genes between recurrent and non-recurrent ccRCC tumours in reference transcriptome-aligned DRS and PCS, respectively. Similar to the results from previous GO BP, MF and CC analyses, many of the significantly enriched KEGG pathways are related to either Immune cell function (e.g. T cell receptor signalling pathway, Th1 and Th2 cell differentiation, Th17 cell differentiation), or antigen processing and presentation related (Figure 4.14A). These pathways were significantly suppressed in the recurrent ccRCC tumours, as shown in the ridge plot in Figure 4.14B and from Appendix tables 7.11 and 12. Interestingly, the KEGG pathway 'PD-L1 expression and PD-1 checkpoint pathway' is also found to be significantly suppressed (NES = -1.994, $p_{adj} = 6.21 \times 10^{-3}$).

The top enriched KEGG pathway activated in recurrent ccRCC tumours was 'Fat digestion and absorption'. This pathway was significantly activated in both DRS and PCS data. Visualisation of pathway enrichment is shown in the ridge plot in Figure 4.14B. Amongst the core enriched genes, *APOB* (Apolipoprotein B) was previously identified as one of the most differentially upregulated genes in recurrent ccRCC tumours compared to non-recurrent tumours (Figure 4.5 C – D). Other differentially upregulated genes in the pathway include *APOA4* (Apolipoprotein A4), *PLPP2* (Phospholipid Phosphatase 2) and the ATP-binding cassette (ABC) transporters *ABCG5* and *ABCG8*.

A



B

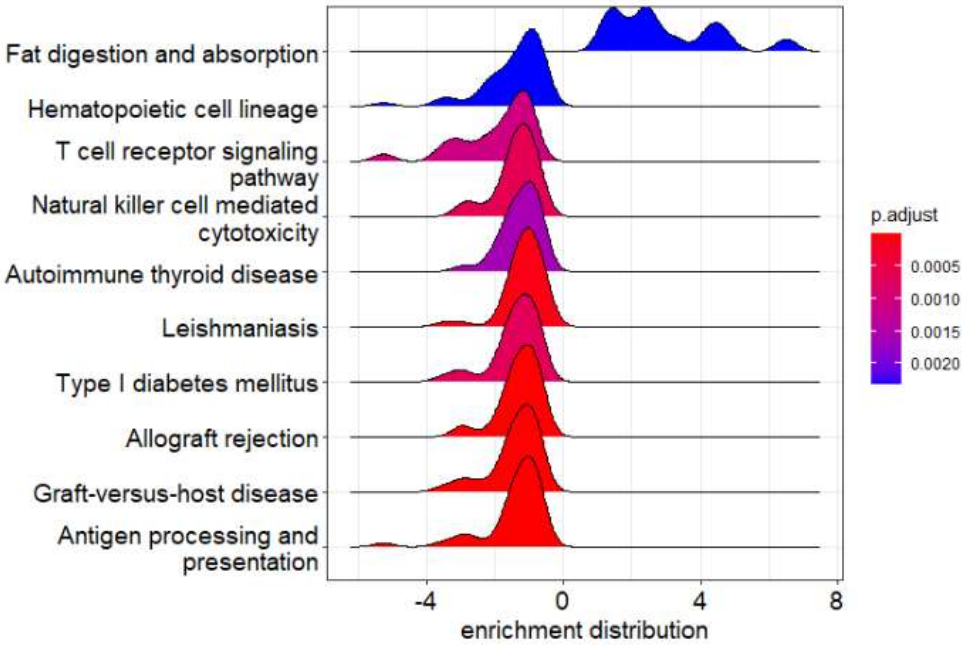


Figure 4.14: KEGG pathways GSEA for ccRCC recurrence associated differential gene expression

A) KEGG pathways enrichment map showing top 10 enriched pathways associated ccRCC recurrence associated differential gene expression profiled by reference transcriptome mapped PCS. **B)** Ridge plot of enriched KEGG pathways, with x axis showing enrichment distribution and ridge colour representing p_{adj} values.

4.3.10 Hierarchical clustering identified a subset of non-recurrent ccRCC with distinct T cell markers

Next, to evaluate if suppression in T cell-associated immune pathways is a universal feature across all recurrent ccRCC tumours, hierarchical clustering of tumour samples was conducted using z-score normalised expression levels from genes in the GO BP gene set 'T cell activation' (n = 549). The clustering result demonstrated that samples did not separate into two groups based on their recurrent status. Notably, three non-recurrent ccRCC tumours (273, 318, 320) displayed distinctive gene expression patterns compared to the rest of the tumour samples (Figure 4.15A). The other three non-recurrent ccRCC tumours were clustered broadly with the six recurrent ccRCC tumours, with no clear characteristic gene expression patterns observed.

To further evaluate the T cell characteristics of this sample group compared to other tumours, a 23-gene panel of $CD8^+$ T cell exhaustion gene markers was selected based on existing literature (Wherry and Kurachi, 2015; Zheng *et al.*, 2021). Intriguingly, hierarchical clustering using the $CD8^+$ T cell exhaustion gene panel again resulted in the distinct clustering of the three non-recurrent ccRCC tumours (273, 318, 320) (Figure 4.15B). These tumour samples were found to highly express T cell exhaustion markers, such as *PDCD1*, *CTLA4*, and *LAG3*, compared to the other tumours. Although GSEA GO MF and GO CC results demonstrated differential MHC expression profiles between non-recurrent and recurrent ccRCC, hierarchical clustering analysis using MHC signatures (GO:0042611) did not separate into clusters corresponding to recurrence status. Tumour samples 273, 318 and 320 also did not form a distinct group (Figure 4.16).

Overall, hierarchical clustering results illustrated the heterogeneity in T cell immune activation status across ccRCC tumours, and a subset of non-recurrent ccRCC tumours display a distinct upregulated expression in $CD8^+$ T cell exhaustion markers. It is important to note that since only 12 tumours were profiled, data here should be treated as preliminary findings. Additional validation cohorts will be needed to confirm findings here.

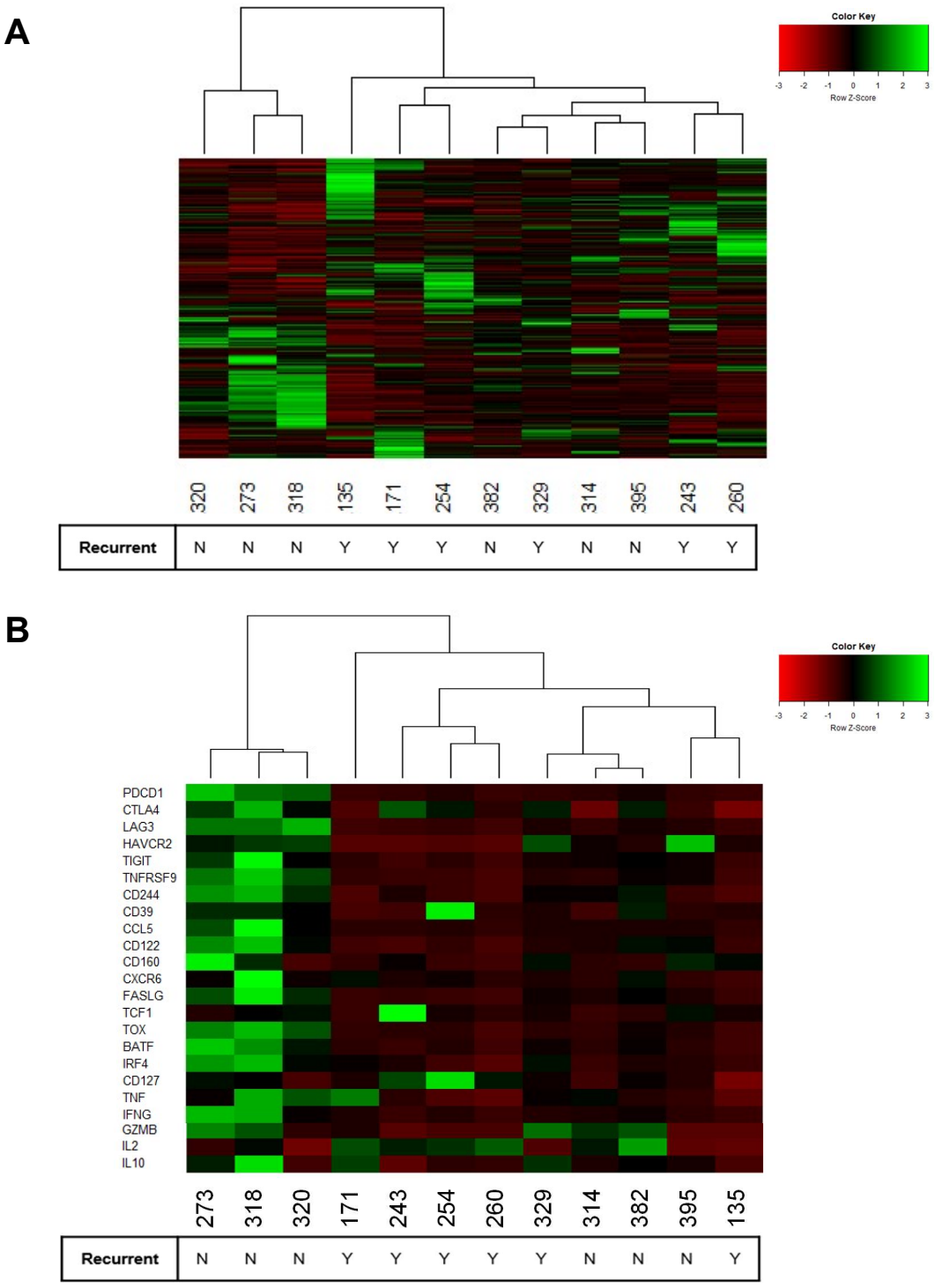


Figure 4.15: Hierarchical clustering analysis of T cell activation and exhaustion markers in ccRCC tumours

A) Z-score hierarchical heatmap based on spearman rank correlations of gene expression levels of the Gene Ontology Biological Process (GO:BP) T cell activation (GO:0042110) genes (n = 549) in ccRCC tumour samples (n = 12), profiled by reference transcriptome mapped PCS. **B)** As in **A**, but for T cell exhaustion markers (n = 23).

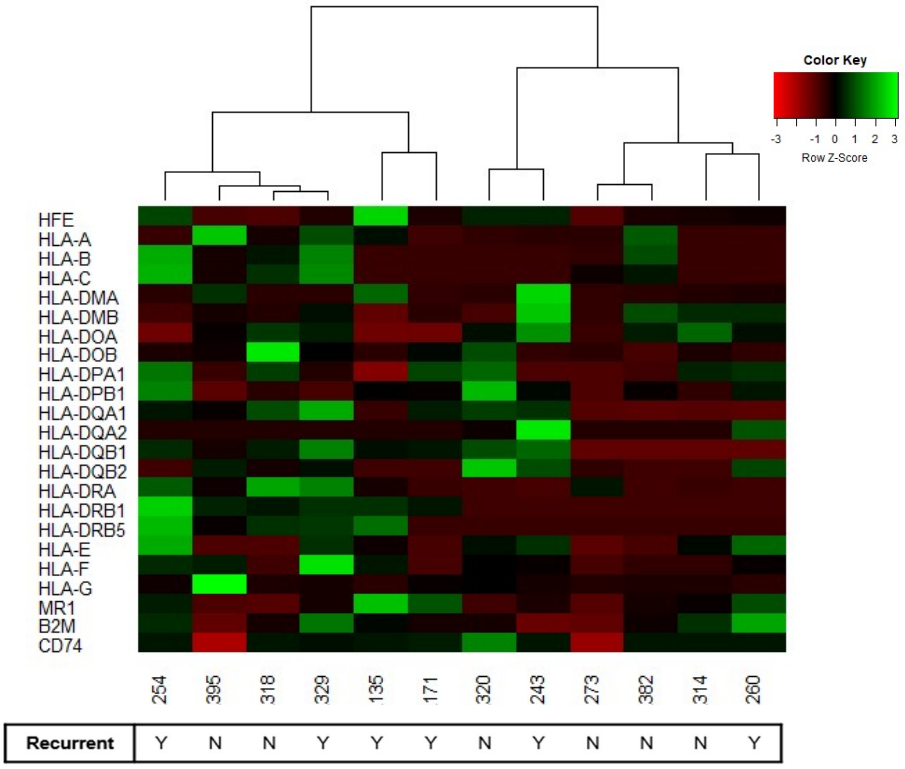


Figure 4.16: Hierachical clustering analysis of MHC protein complex genes expression in ccRCC tumours

Z-score hierarchical heatmap based on spearman rank correlations of gene expression levels of the Gene Ontology Cellular Compartment (GO:CC) MHC protein complex (GO:0042611) genes (n = 37) in ccRCC tumour samples (n = 12), profiled by reference transcriptome mapped PCS.

4.3.11 Recurrence of ccRCC is associated with lower tumour immune infiltration

To explore the relationship between ccRCC recurrence and tumour immune infiltrate populations, the gene expression signature-based algorithm ESTIMATE (Estimation of Stromal and Immune cells in malignant tumour using expression data) was used to infer the abundance of immune cells and stromal cells in the TME (Yoshihara *et al.*, 2013). No significant difference was observed between stromal scores of recurrent and non-recurrent ccRCC tumours when DRS-generated gene expression data was used (Figure 4.17A). Using PCS gene expression data, a borderline non-significant trend towards reduction in stromal score was found in recurrent ccRCC tumours compared to non-recurrent samples ($p = 0.619$) (Figure 4.17B). ESTIMATE immune scores were significantly lower in recurrent ccRCC tumours compared to non-recurrent counterparts using PCS data (Figure 4.17D). DRS data displayed a borderline non-significant trend towards a decrease in immune score in recurrent ccRCC tumours ($p = 0.0881$) (Figure 4.17C). A high degree of concordance was found between immune scores inferred by PCS or DRS expression data ($R^2 = 0.87$, $p = < 0.0001$) (Figure 4.17I). ccRCC tumour purity levels (Combined stromal score and immune score results) were significantly lower in non-recurrent ccRCC tumours compared to recurrent ccRCC tumours when PCS gene expression data was analysed. In contrast, no significant difference was observed between recurrent and non-recurrent ccRCC tumours using DRS gene expression data.

To further test the above findings, a second gene expression signature-based algorithm xCell was used to confirm the difference in the immune infiltrate abundance between recurrent and non-recurrent ccRCC tumours found by ESTIMATE. DRS and PCS data indicated a significantly lower immune score for recurrent ccRCC tumours than for non-recurrent tumours (Figure 4.17G – H). Similar to the ESTIMATE immune scores, xCell immune scores generated from DRS and PCS were strongly correlated ($R^2 = 0.9714$, $p = < 0.0001$) (Figure 4.17J).

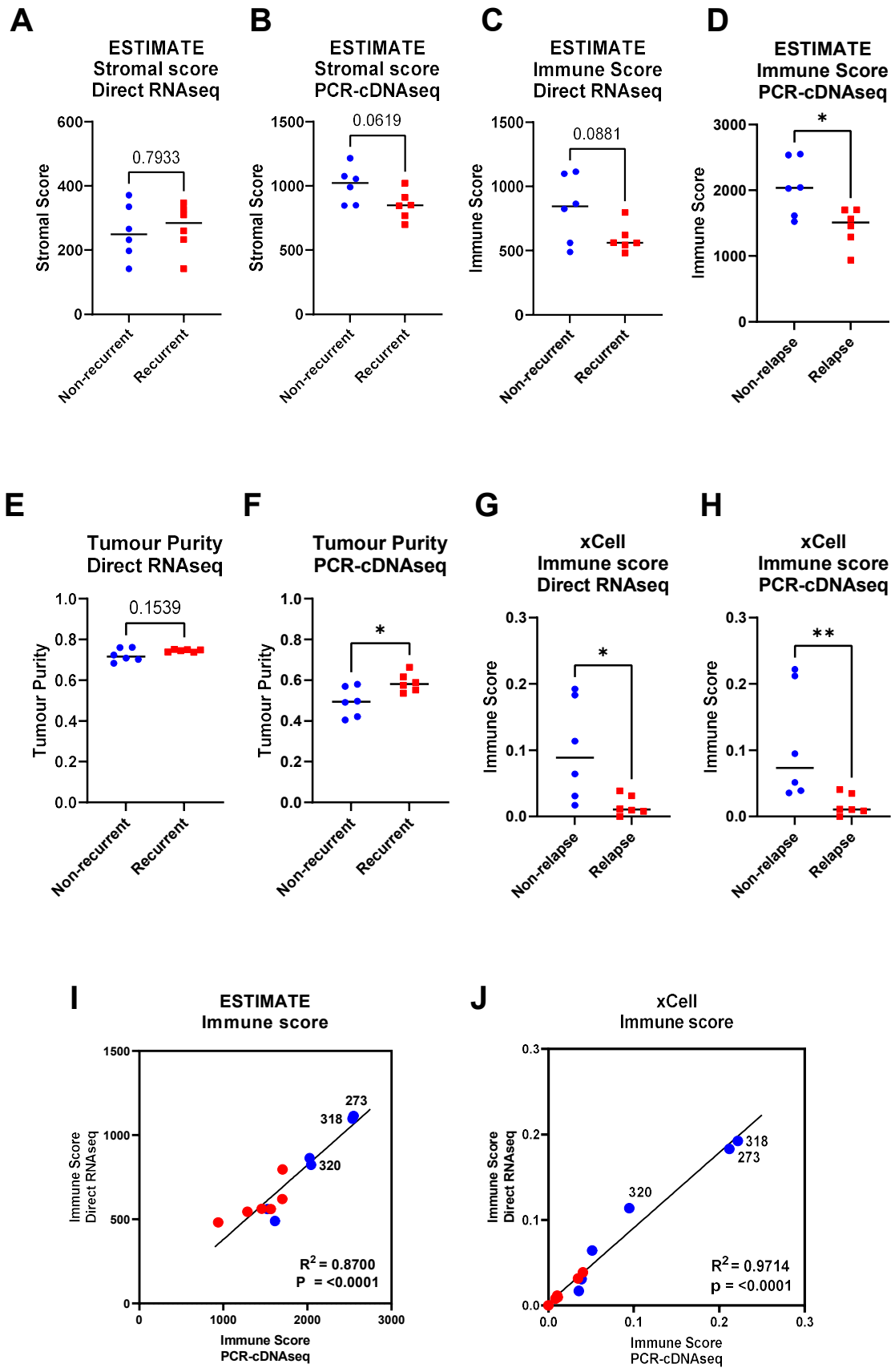

A
ESTIMATE
Stromal score
Direct RNAseq
Stromal Score
600
400
200
0
0.7933
Non-recurrent
Recurrent
B
ESTIMATE
Stromal score
PCR-cDNAseq
Stromal Score
1500
1000
500
0
0.0619
Non-recurrent
Recurrent
C
ESTIMATE
Immune Score
Direct RNAseq
Immune Score
1500
1000
500
0
0.0881
Non-recurrent
Recurrent
D
ESTIMATE
Immune Score
PCR-cDNAseq
Immune Score
3000
2000
1000
0
*
Non-relapse
Relapse
E
Tumour Purity
Direct RNAseq
Tumour Purity
1.0
0.8
0.6
0.4
0.2
0.0
0.1539
Non-recurrent
Recurrent
F
Tumour Purity
PCR-cDNAseq
Tumour Purity
1.0
0.8
0.6
0.4
0.2
0.0
*
Non-recurrent
Recurrent
G
xCell
Immune score
Direct RNAseq
Immune Score
0.3
0.2
0.1
0.0
*
Non-relapse
Relapse
H
xCell
Immune score
PCR-cDNAseq
Immune Score
0.3
0.2
0.1
0.0
**
Non-relapse
Relapse
I
ESTIMATE
Immune score
Immune Score
Direct RNAseq
1500
1000
500
0
273
318
320
R2 = 0.8700
P = <0.0001
0
1000
2000
3000
Immune Score
PCR-cDNAseq
J
xCell
Immune score
Immune Score
Direct RNAseq
0.3
0.2
0.1
0.0
318
273
320
R2 = 0.9714
p = <0.0001
0.0
0.1
0.2
0.3
Immune Score
PCR-cDNAseq

Figure 4.17: Estimation of stromal and immune cells in ccRCC tumours using DRS and PCS gene expression data

A) Grouped dot plot showing estimated immune score of non-recurrent and recurrent ccRCC tumours by the ESTIMATE algorithm, using reference genome aligned DRS gene expression data. **B)** As in **A**, but with PCS. **C)** Grouped dot plot showing estimated stromal score of non-recurrent and recurrent ccRCC tumours by the ESTIMATE algorithm, using reference genome aligned DRS gene expression data. **D)** As in **C**, with PCS data used instead. **E)** Grouped dot plot showing estimated tumour purity of non-recurrent (blue) and recurrent (red) ccRCC tumours by the ESTIMATE algorithm, using reference genome aligned DRS gene expression data. **F)** As in **E**, but with PCS data. **G)** Grouped dot plot showing estimated immune score of non-recurrent (blue) and recurrent (red) ccRCC tumours by xCell algorithm, using reference genome aligned DRS gene expression data. **H)** As in **G**, with PCS data used instead. **I)** Correlation between ESTIMATE immune scores of non-recurrent (blue) and recurrent (red) ccRCC tumours, generated by DRS and PCS gene expression data. **J)** Correlation between xCell immune score of non-recurrent (blue) and recurrent (red) ccRCC tumours, generated by DRS and PCS gene data. For **A – H**, two-tailed unpaired T-tests with Welch's correction were used, with $p \leq 0.05$ considered statistically significant. * ≤ 0.05, ** ≤ 0.01. P values of non-significant results are indicated in graphs. Centre line represents median for each group. For **I – J**, diagonal lines represent the line of best fit. R^2 values were computed to measure goodness-of-fit, and P values generated from F-test, with $p \leq 0.05$ considered statistically significant.

4.3.12 CD8$^+$ T cell populations are depleted in recurrent ccRCC

Having found that the quantity of tumour immune infiltrates was significantly lower in recurrent ccRCC tumours, the relationship between ccRCC recurrent status and immune cell-type profiles was next explored. To determine the relative proportions of different immune infiltrates, gene expression profiles (DRS and PCS) were analysed by the CIBERSORTx algorithm. Recurrent and non-recurrent ccRCC tumours showed substantial differences in immune infiltrate profiles. In non-recurrent ccRCC tumours, cytotoxic CD8$^+$ T cells represented the largest proportion of immune infiltrates, followed by the immune suppressive M2 macrophages. On average, 41.4%/34.4% (DRS/PCS) of all immune infiltrates in non-recurrent ccRCC tumours were estimated to be CD8$^+$ T cells. On average, M2 macrophages were estimated to represent 24.1%/27.4% (DRS/PCS) of immune infiltrates in non-recurrent ccRCC tumours (Figure 4.18 A – B).

In recurrent ccRCC tumours, 17.8%/15.6% (DRS/PCS) of immune infiltrates on average were CD8$^+$ T cells, a significantly lower proportion compared to non-recurrent ccRCC tumours (Figure 4.19 A – B). The cell type deconvolution tool EPIC found similar results, where a significant decrease in the proportion of CD8$^+$ T cells in recurrent ccRCC tumours was observed with both DRS and PCS data (Figure 4.19 C – D). Estimated CD8$^+$ T cell proportions from DRS and PCS are strongly correlated using CIBERSORTx and EPIC (Figure 4.19 E–F). Previously, hierarchical clustering results identified a cluster of non-recurrent tumours (318, 273, 320) showing a different expression profile of T cell activation markers and T cell exhaustion markers (Figure 4.15). *PDCD1* (encodes for PD-1) and *TOX* (Thymocyte selection associated high mobility group box) are established exhaustion signatures, with their expression levels positively correlated with levels of CD8$^+$ T cell exhaustion (Zheng *et al.*, 2021). The expression levels of *PDCD1* and *TOX* (Normalised counts from reference genome mapped PCS) significantly correlated with the proportion of CD8$^+$ T cell population in the tumours. Furthermore, tumours 318, 273 and 320, previously clustered using T cell exhaustion markers (Figure 4.15), were shown with distinctly high proportion of CD8$^+$ T cells (Figure 4.19 G – H).

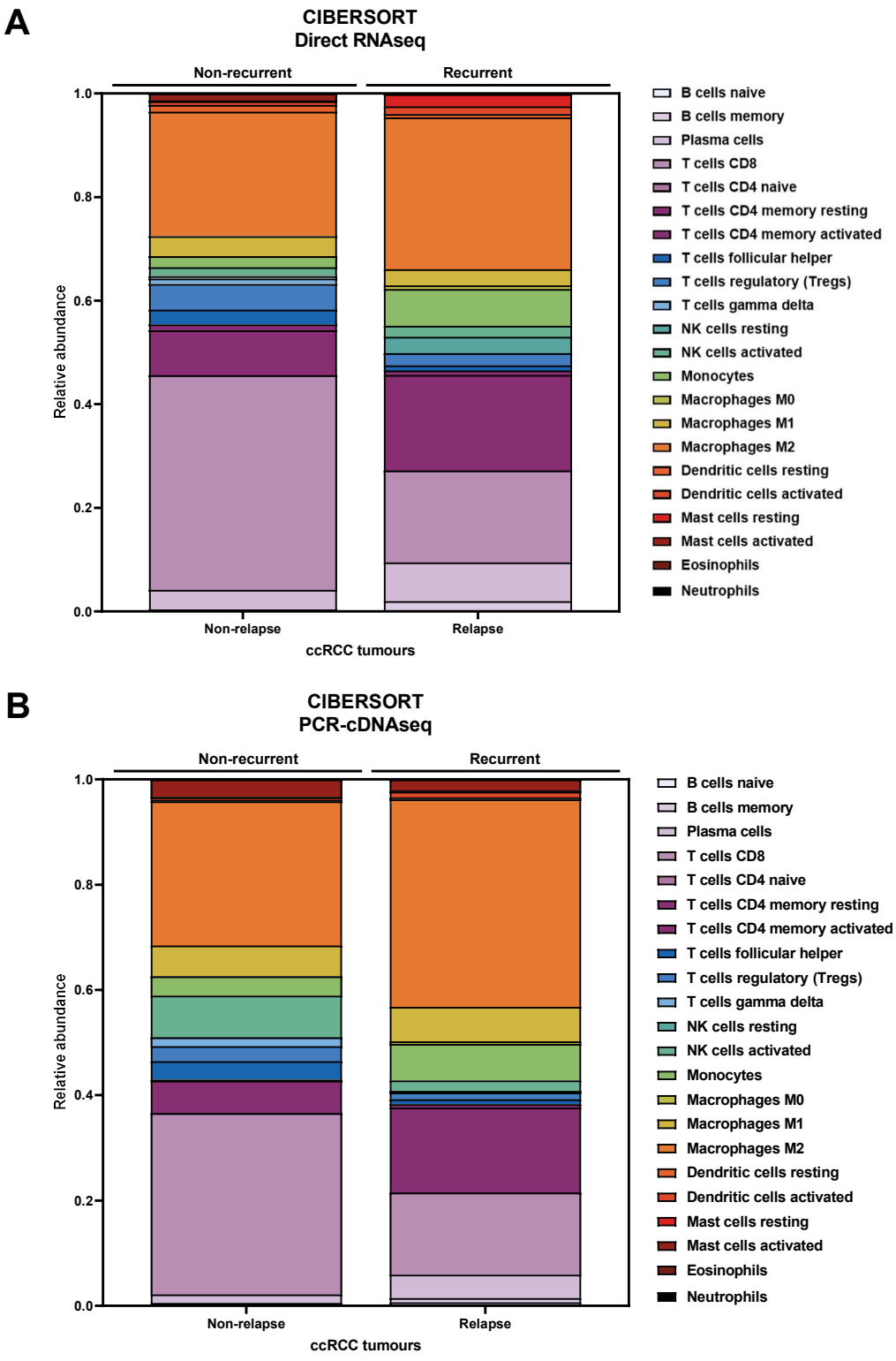


Figure 4.18: Immune infiltration landscape in ccRCC tumours estimated with CIBERSORT

A) Stacked bar chart showing averaged cell type composition of tumour infiltrating immune cells in non-recurrent and recurrent ccRCC tumour samples, analysed by CIBERSORT using reference transcriptome aligned DRS gene expression data. **B)** As in **A**, but with PCS gene expression data.

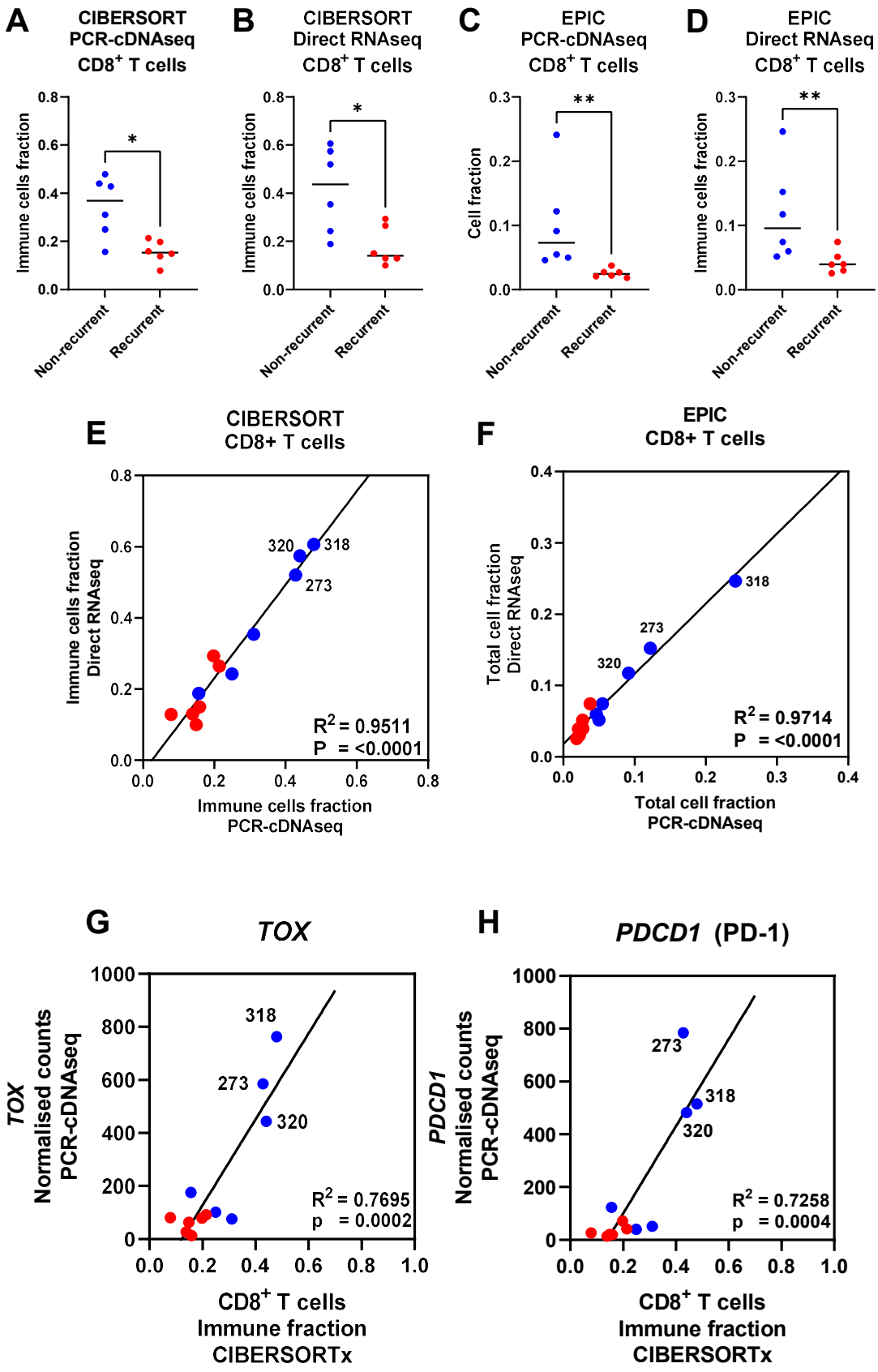

A
CIBERSORT
PCR-cDNAseq
CD8+ T cells
Immune cells fraction
Non-recurrent
Recurrent
B
CIBERSORT
Direct RNAseq
CD8+ T cells
C
EPIC
PCR-cDNAseq
CD8+ T cells
Cell fraction
D
EPIC
Direct RNAseq
CD8+ T cells
E
CIBERSORT
CD8+ T cells
Immune cells fraction
Direct RNAseq
Immune cells fraction
PCR-cDNAseq
320
318
273
R2 = 0.9511
P = <0.0001
F
EPIC
CD8+ T cells
Total cell fraction
Direct RNAseq
Total cell fraction
PCR-cDNAseq
R2 = 0.9714
P = <0.0001
G
TOX
TOX
Normalised counts
PCR-cDNAseq
CD8+ T cells
Immune fraction
CIBERSORTx
R2 = 0.7695
p = 0.0002
H
PDCD1 (PD-1)
PDCD1
Normalised counts
PCR-cDNAseq
CD8+ T cells
Immune fraction
CIBERSORTx
R2 = 0.7258
p = 0.0004

Figure 4.19 Depletion in CD8⁺ T cells in recurrent ccRCC tumours compared to non-recurrent CRCC tumours

A) Grouped dot plot showing relative population of $CD8^+$ T cells within immune infiltrates of non-recurrent (blue) and recurrent (red) ccRCC tumours estimated by CIBERSORT using reference genome aligned PCS data. **B)** As in **A**, but with DRS data. **C)** Grouped dot plot showing proportions of $CD8^+$ T cells within immune infiltrates of non-recurrent and recurrent ccRCC tumours estimated by EPIC, using reference genome aligned PCS data. **D)** As in **C**, but with DRS. **E)** Correlation between CIBERSORT estimated $CD8^+$ T cells fraction amongst immune infiltrates in non-recurrent (blue) and recurrent (red) ccRCC tumours, generated by DRS and PCS gene expression data. **F)** Correlation between EPIC estimated $CD8^+$ T cells fraction amongst immune infiltrates in non-recurrent and recurrent ccRCC tumours, generated by DRS and PCS gene expression data. **G)** Correlation between normalised expressions of *TOX* from reference genome mapped PCS and proportions of $CD8^+$ T cells within immune infiltrates of ccRCC tumours estimated by CIBERSORT. **H)** As in **G**, but for *PDCD1*. For **A – D**, two-tailed unpaired T-tests with Welch's correction were used, with $p \leq 0.05$ considered significant. * ≤ 0.05, ** ≤ 0.01. Centre line represents median for each group. For **E – H**, diagonal lines represent the line of best fit. R^2 values were computed to measure goodness-of-fit, and P values generated from F-test, with $p \leq 0.05$ considered statistically significant.

For other immune cell types, CIBERSORTx and EPIC presented conflicting results. For M2 macrophages, CIBERSORTx estimated an average proportion of 29.3/39.5% (DRS/PCS) immune infiltrates in recurrent ccRCC tumours (Figure 4.20 A – B). Like other Macrophage subtypes (M0, M1), the M2 macrophage proportions were not changed significantly. In contrast, EPIC estimated a significant decrease in the proportions of the macrophage population in recurrent tumours (Figure 4.20 C–D). CIBERSORTx estimated a substantial drop for activated NK cells using PCS, but only 5 out of 12 samples could identify the cell type using DRS (Figure 4.20 E-F). Similarly, NK cells were only detected in 2 samples using EPIC (Figure 4.20 G-H).

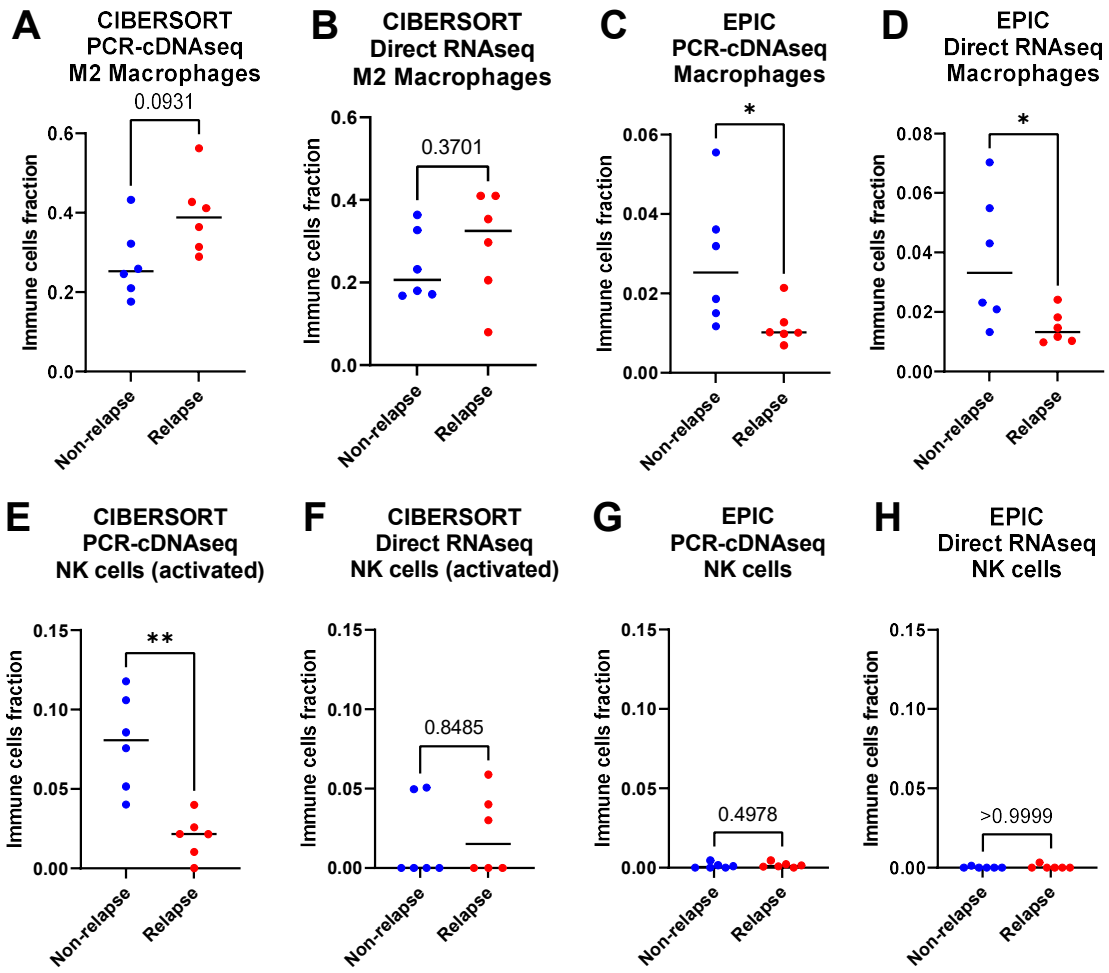


Figure 4.20: Estimation of Macrophage and NK cell fractions in immune infiltrates of ccRCC tumours by CIBERSORT and EPIC

A) Grouped dot plot showing proportions of M2 Macrophages within immune infiltrates of non-recurrent (blue) and recurrent (red) ccRCC tumours, estimated by CIBERSORT using reference genome aligned PCS gene expression data. **B)** As in **A**, but with reference genome aligned DRS gene expression data. **C)** Grouped dot plot showing proportions of Macrophages within immune infiltrates of non-recurrent and recurrent ccRCC tumours, estimated by EPIC using PCS data. **D)** As in **C**, but with DRS data. **E)** Grouped dot plot showing proportions of activated NK cells within immune infiltrates of non-recurrent and recurrent ccRCC tumours, estimated by CIBERSORT using PCS data. **F)** As in **E** but with DRS. **G)** Grouped dot plot showing proportions of NK cells within immune infiltrates of non-recurrent and recurrent ccRCC tumours, estimated by EPIC using PCS data. **H)** As in **G** but with DRS gene expression data. Throughout, two-tailed unpaired T-tests with Welch's correction were used, with $p \leq 0.05$ considered statistically significant. * ≤ 0.05, ** ≤ 0.01. P values of non-significant results are indicated in graphs. Centre line represents median for each group.

4.3.13 Identification of ccRCC recurrent associated differential transcript isoform usage events

After identifying genes that undergo differential expression in recurrent ccRCC tumours, differential transcript usage analysis was carried out using a bioinformatic pipeline integrating DRIMseq and DEXseq (Love *et al.*, 2018). Analysis results from PCS data identified six genes that displayed isoform switching in recurrent ccRCC tumours compared to non-recurrent tumours (Figure 4.21A). These genes include *EFHC2* (EF-hand domain containing 2), *ABI1* (Abl interactor 1), *NRP2* (Neuropilin 2), *CMC1* (C-X9-C motif containing 1), *MIB2* (MIB E3 Ubiquitin protein ligase 1), and *ST6GALNAC6* (ST6 N-Acetylgalactosaminide alpha-2,6-Sialyltransferase 6). All six genes were identified by DRIMseq, amongst which three genes (*CMC1*, *MIB2*, *ST6GALNAC6*) were also identified by DEXseq (Figure 4.21B). *CMC1* was used as an example to demonstrate detected DTU events in recurrent ccRCC tumours.

CMC1 encodes for a mitochondrial protein which regulates the assembly of cytochrome c oxidase, or complex IV (Bourens and Barrientos, 2017). PCS gene expression data (reference transcriptome aligned) demonstrated that whilst not labelled as a significant DEG, expression levels of *CMC1* are lower in recurrent ccRCC tumours compared to non-recurrent ccRCC ($Log_2FoldChange$ = -1.20, p_{adj} = 0.0743) (Figure 4.21C). Four *CMC1* transcript isoforms were mapped in PCS data (Figure 4.21 E- F). In non-recurrent ccRCC tumours, average proportions for ENST00000423894, ENST00000466830, ENST00000468330 and ENST00000495428 were 0.140, 0.312, 0.206 and 0.342, respectively. In contrast, for recurrent ccRCC, the average proportions were 0.338, 0.553, 0.0321 and 0.0765 (Figure 4.21D). The shift in transcript usage was shown visually by the IGV coverage tracks after reference genome alignment. Combined coverage tracks from PCS data demonstrate a proportional drop in coverage where the arrows are highlighted (Figure 4.21F), which correlates with the decreased expression of ENST00000468330 in recurrent ccRCC. The same observation was also found in DRS data, albeit with non-significant DRIMseq and DEXseq p values (p_{adj} = 1) (Figure 4.21G).

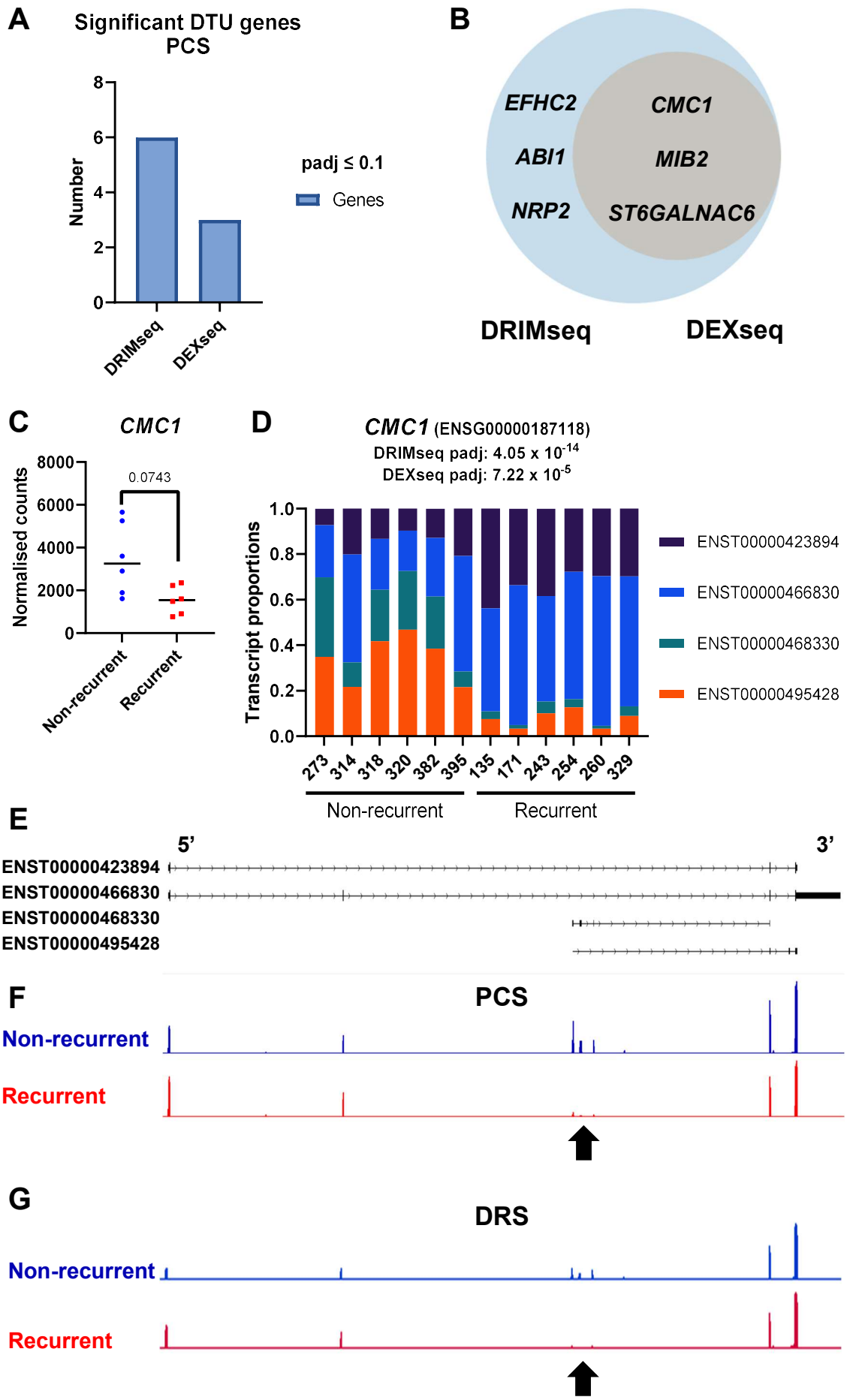

A
Significant DTU genes
PCS
Number
8
6
4
2
0
DRIMseq
DEXseq
padj ≤ 0.1
Genes
B
EFHC2
ABI1
NRP2
CMC1
MIB2
ST6GALNAC6
DRIMseq
DEXseq
C
CMC1
Normalised counts
8000
6000
4000
2000
0
0.0743
Non-recurrent
Recurrent
D
CMC1 (ENSG00000187118)
DRIMseq padj: 4.05 x 10-14
DEXseq padj: 7.22 x 10-5
Transcript proportions
1.0
0.8
0.6
0.4
0.2
0.0
273
314
318
320
382
395
135
171
243
254
260
329
Non-recurrent
Recurrent
ENST00000423894
ENST00000466830
ENST00000468330
ENST00000495428
E
5'
3'
ENST00000423894
ENST00000466830
ENST00000468330
ENST00000495428
F
PCS
Non-recurrent
Recurrent
G
DRS
Non-recurrent
Recurrent

Figure 4.21: DTU analysis using ONT PCS identifies isoform switching events associated with ccRCC recurrence

A) Bar graph showing the number of genes that display significant differential transcript usage ($p_{adj} \leq 0.1$) between recurrent and non-recurrent ccRCC tumours, as analysed by DRIMseq and DEXseq using reference transcriptome aligned PCS data. **B)** Venn diagram showing the overlaps of DRIMseq and DEXseq identified genes that display DTU between recurrent and non-recurrent ccRCC tumours. **C)** Grouped dot plot showing PCS DESeq2 normalised *CMC1* gene expression in non-recurrent (blue) and recurrent (red) ccRCC tumours. **D)** Stack bar graphs representing proportions of *CMC1* isoforms in ccRCC tumours using PCS data. DRIMseq and DEXseq p_{adj} values for DTU of *CMC1* are indicated in graph. **E)** Graphical representation of *CMC1* transcripts Ensembl reference annotations (Ensembl release 105, GRCh38) in Integrated genomics viewer (IGV). **F)** IGV visualisation of combined PCS reads coverage tracks for non-recurrent (blue) and recurrent (red) ccRCC tumours at *CMC1* locus. Locations of exon exclusion events by recurrent ccRCC are highlighted by arrows. **G)** IGV visualisation of combined DRS reads coverage tracks for non-recurrent (blue) and recurrent (red) ccRCC tumours at *CMC1* locus**.** For **C**, p_{adj} value was generated by Wald test followed by Benjamini-Hochberg correction. p_{adj} value was indicated in graph. Centre line represents median for each group.

Focusing on ENST00000423894 and ENST00000466830, the two transcripts differ in their exon two inclusion/skipping and their 3' ends, where ENST00000466830 has a substantially longer 3' UTR (Figure 4.21E). Although both transcripts were mapped across the tumour samples, the coverage track shows that most reads adopt the short 3'UTR structure of ENST00000423894 (or ENST00000495428). This indicates that the decision by minimap2 to assign reads between the isoforms lies in the coverage of exon 2, despite the 3'UTR of ENST00000466830 being 5500nt long and exon 2 being 90nt in length.

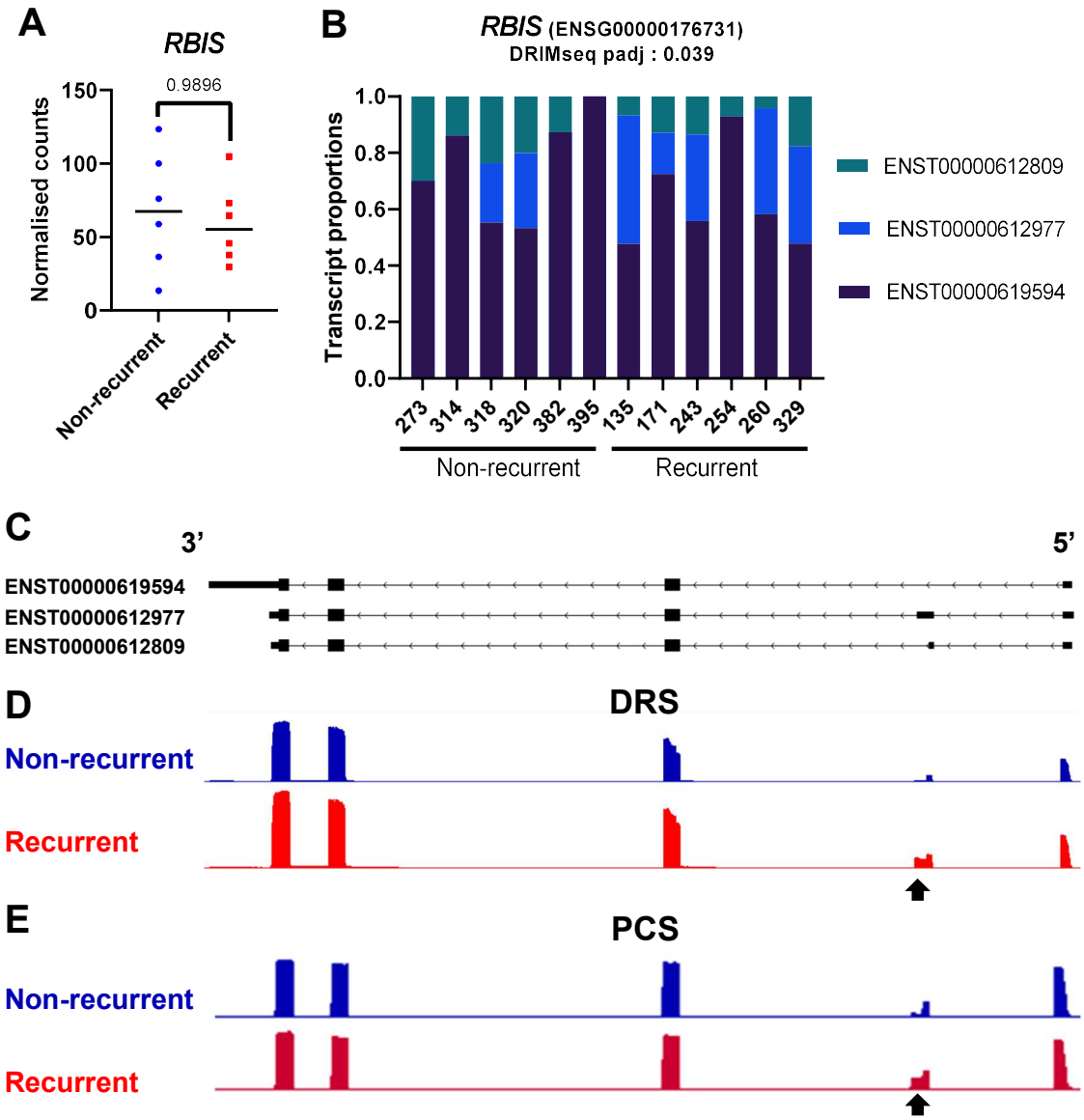


Figure 4.22: DTU analysis revealed ccRCC reccurrence associated DTU of *RBIS*

A) Grouped dot plot showing DRS DESeq2 normalised *RBIS* gene expression in non-recurrent (blue) and recurrent (red) ccRCC tumours. **B)** Stack bar graphs representing proportions of *RBIS* isoforms in ccRCC tumours using DRS data. DRIMseq and DEXseq p_{adj} values for DTU of *RBIS* are indicated in graph. **C)** Graphical representation of *RBIS* transcripts Ensembl reference annotations (Ensembl release 105, GRCh38) in Integrated genomics viewer (IGV). **D)** IGV visualisation of combined DRS reads coverage tracks for non-recurrent (blue) and recurrent (red) ccRCC tumours at *RBIS* locus. Locations of exon exclusion events by recurrent ccRCC are highlighted by arrows. **E)** IGV visualisation of combined PCS reads coverage tracks for non-recurrent (blue) and recurrent (red) ccRCC tumours at *RBIS* locus**.** For **A**, p_{adj} value was generated by Wald test followed by Benjamini-Hochberg correction by DESeq2. p_{adj} value was indicated in graph. Centre line represents median for each group.

For DRS, only one gene (*RBIS*, Ribosomal biogenesis factor) showed significant differential transcript usage by DRIMseq and DEXseq. *RBIS* showed no differential expression between recurrent and non-recurrent ccRCC tumour samples ($Log_2FoldChange$ = -0.2017, p_{adj} = 0.9896). On average, recurrent ccRCC express a higher level of ENST0000612977 at 27.1%, compared to non-recurrent ccRCC tumours at 7.953% (Figure 4.22B). The isoform switching event was visually assessed by DRS IGV coverage tracks (reference genome aligned), with the arrow indicating the unique 5' region ENST00006162977 encodes for (Figure 4.22D). This DTU event was also visible in PCS. Reference genome-aligned PCS IGV coverage tracks showed higher coverage of the ENST00006162977 exclusive 5' region for recurrent ccRCC tumour samples (Figure 4.22E).

Finally, survival analysis using TCGA KIRC *CMC1* and *RBIS* mRNA expression and patient data shows that high *CMC1* and low *RBIS* expression in ccRCC tumours significantly correlate with better overall survival (Figure 4.23A – B). No transcript-level expression data was available to construct isoform-specific survival analysis.

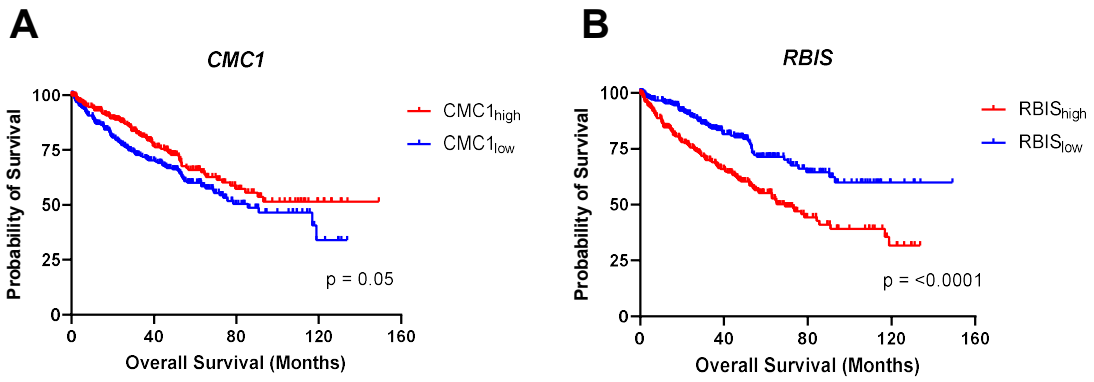


Figure 4.23: Kaplan-Meier survival curves of high- and low- *CMC1* and *RBIS* expression in TCGA KIRC cohort

A) Kaplan-Meier survival curve of overall survival in TCGA KIRC cohort patients (n = 510), with patients grouped by high- (red, n = 255) and low-*CMC1* (blue, n = 255) expression groups based on median gene expression. **B)** As in **A**, but for *RBIS*.

Throughout, p-values were calculated using log-rank (Mantel-Cox) test, with $p \leq 0.05$ considered statistically significant.

4.3.14 Reference-guided transcriptome assembly from read alignments identified novel isoforms from ccRCC tumours

To systematically identify sequencing reads representing unannotated novel transcript isoforms, transcriptome assembly pipeline StringTie2 and FLAIR were performed using reference genome-aligned PCS data. In brief, read alignments were 'collapsed' after alignment to the reference genome into high-confidence isoforms that explain the alignments. Next, isoforms were collated and compared to reference gene annotation by gffcompare, which assigns a class code depending on their relationship with the closest matching reference isoform (Figure 2.1). Analysis of gffcompare results from StringTie2 and FLAIR showed that thousands of mapped reads in PCS represent novel isoforms.

For StringTie2, 32.19% (17160 isoforms) of assembled transcripts represent 'known transcripts', with matched intron/exon junctions with at least one reference transcript annotation. 64.62% (34450 isoforms) of assembled transcripts represent 'novel transcripts'. Most novel transcripts have a gffcompare class code of 'j' (40.58% of all assembled transcripts). 'j' indicates a multi-exon transcript with at least one matched exon junction with the reference transcript annotation, thus likely to be a novel spliced variant. Another source of novel spliced variants were classed as 'k' (3.25% of all assembled transcripts), where the assembled transcript contains all elements of a reference transcript annotation but with additional sequence/exons compared to the reference. 'm' and 'n' are assembled transcripts with retained introns. They represent 3.73% and 3.60% of all assembled transcripts. 7.41% of all StringTie2 assembled transcripts were classed as 'o', which share exonic structure with existing reference transcript annotation but not completely. 'o' represents another source of novel spliced variants or isoforms with alternative polyadenylation sites. The final major source of novel transcripts is 'x' (4.34% of all StringTie2 assembled PCS transcripts), where the transcript sequence overlaps with reference transcript annotation but on the opposite strand. 'x' represents a potential anti-sense transcript that is currently not in the reference transcript annotation (Table 4.3).

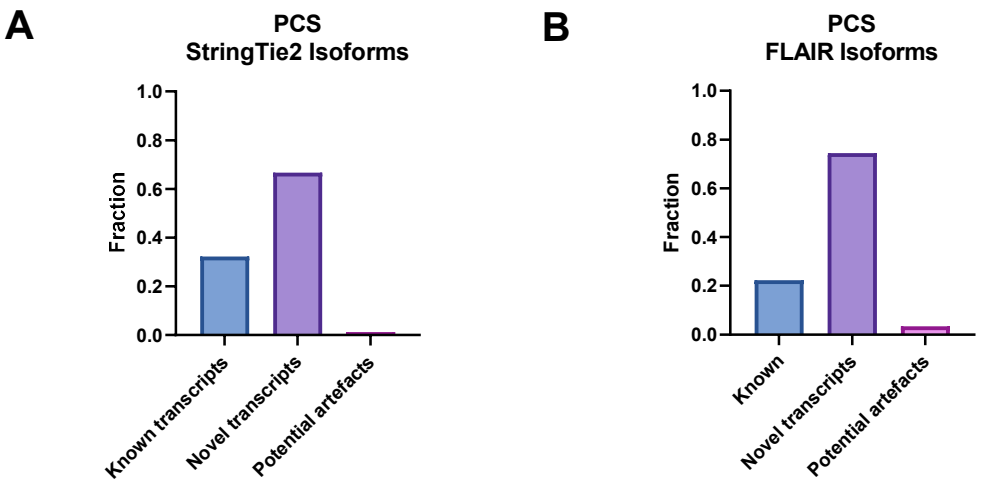


Figure 4.24: Novel transcripts identification by StringTie2 and FLAIR transcriptome assembly using ccRCC tumours PCS data

A) Bar chart representing proportion of StringTie2 assembled transcripts from PCS of ccRCC tumours that represent 'known transcripts', 'novel transcripts' and 'potential artefacts', compared to Ensembl gene annotation (GRCh38, release version 105) using gffcompare. **B)** As in **A**, but for FLAIR assembled transcripts. Known: '=', 'c'; Novel: 'j','k','m','n','i','o','x','y'; Potential artefacts: 'p','e','s','r','u'.

			StringTie2		FLAIR	
	Gffcompare class code	Description	Count	%	Count	%
Known	=	Complete intron chain match	15099	28.32	35095	15.7
	c	Partial intron chain match	2061	3.87	14592	6.54
Novel	j	Multi-exon gene with at least 1 matched exon junction	21635	40.58	112476	50.4
	k	Containment of reference transcript	1732	3.25	2723	1.22
	m	Retained intron(s)	839	1.57	7490	3.36
	n	Retained intron(s), not all exon junctions matched	1987	3.73	22162	9.94
	i	Fully contained within reference intron	1917	3.60	12465	5.59
	o	Same strand as reference, with overlapping exons	3951	7.41	1443	0.65
	x	Overlap with reference exon on the opposite strand	2314	4.34	3079	1.38
	y	Contains reference gene within its intron	75	0.14	1	0.00
Potential artefacts	p	No overlap with reference	163	0.31	885	0.40
	e	Single exon, partially covering an intron	453	0.85	6588	2.96
	s	Intron(s) matched with reference on opposite strand	2	0.01	0	0.00
	r	Repeated sequence	0	0.00	0	0.00
	u	Unknown, not classified	1084	2.03	3978	1.78

Table 4.3: Classification of assembled transcript isoforms predicted by StringTie2 and FLAIR

Statistics related to StringTie2 and FLAIR identified transcripts and corresponding transcript classification code compared to Ensembl gene annotation (GRCh38, release version 105) using gffcompare.

Compared with StringTie2 assembly, FLAIR transcriptome assembly with PCS data resulted in far higher numbers of known and novel transcripts. For FLAIR assembly, 22.28% of all assembled transcripts (49687) are 'known transcripts'. 74.36% of FLAIR assembled transcripts (161839) are novel transcripts. Similar to StringTie2, most novel transcripts have a class code of 'j' (112476, 50.4% of all assembled transcripts), representing potential splice variants. A higher number of retained introns transcripts (both 'm' and 'n', 7490 and 22162) were discovered in FLAIR assembly compared to StringTie2 (Table 4.3, Figure 4.26B – C).

Whilst the number of 'known transcripts' identified by StringTie2 and FLAIR transcript assembly were dwarfed by the number of novel transcripts (Figure 4.24), the number of assembled 'known transcripts' were substantially lower than the number of uniquely mapped transcript isoforms when PCS data was aligned to the reference transcriptome. When reads generated from PCS of ccRCC tumours were aligned to the reference transcriptome, a total number of 26,070 unique genes were mapped (Figure 3.15B). At the transcript isoform levels, 124,741 unique transcript isoforms from the Ensembl reference annotation were mapped. In comparison, only 17160 and 49687 StringTie2 and Flair 'known transcripts' were reconstructed (Figure 4.26A). Comparing the total number of unique transcripts, FLAIR identified a higher number of unique transcripts (222,977) than the total number of reference transcriptome-mapped transcripts. However, even when combined with novel transcripts, StringTie2 had a lower number of identified unique transcripts (53,222) than the total number of reference transcriptome-mapped unique transcripts.

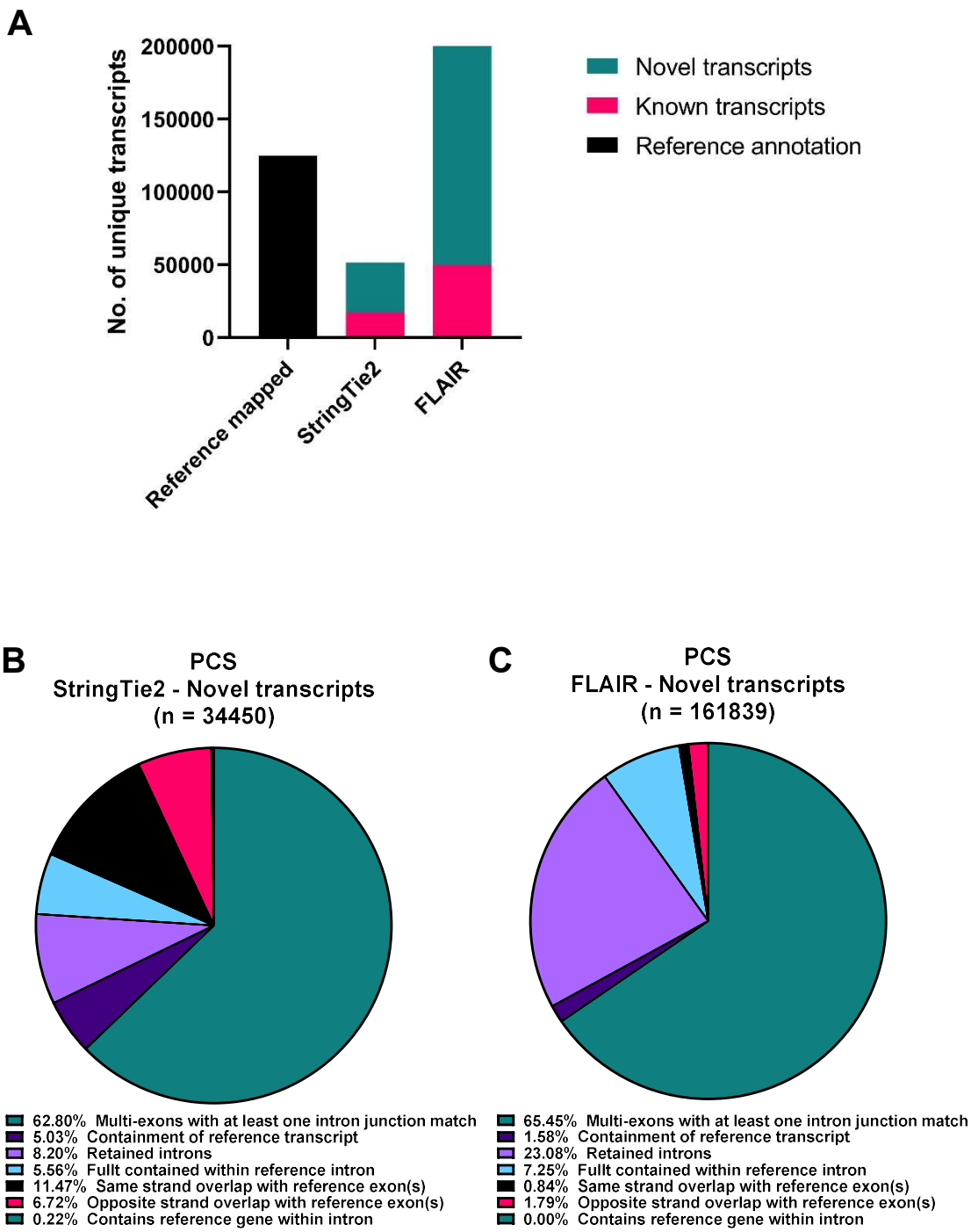


Figure 4.25: Characterisation of StringTie2 and FLAIR assembled transcripts

A) Bar chart representing the number of unique transcripts mapped by minimap2 using reference transcriptome (Ensembl release 105, cDNA reference), 'Known' StringTie2 assembled transcripts and 'Known' FLAIR transcripts from PCS of ccRCC tumours. **B)** Pie chart depicting the proportions of gffcompare classes of novel transcripts (compared to Ensembl gene annotation (GRCh38, release version 105)) from StringTie2 assembled ccRCC tumour transcriptomes (PCS). **C)** As in **B**, but for FLAIR. Known: '=', 'c'; Novel: 'j','k','m','n','i','o','x','y'.

4.3.15 Characterisation of a novel *CTLA4* isoform with an alternative 3'UTR structure

With a large amount of potential novel transcripts in the sequencing data, a particular focus was put on immune checkpoints that play an essential role in modulating tumour immunity. CTLA4 is an immune checkpoint receptor expressed in T cells. Upon engagement with CD80/CD86 molecules at the surface of antigen-presenting cells, CTLA4 negatively regulates T cell activation (Kalbasi and Ribas, 2020). *CTLA4* transcripts were identified across the sequenced ccRCC tumours by DRS and PCS, as shown by the IGV reads coverage track (Figure 4.26A – B). Coverage tracks were next compared to the Ensembl gene annotation. The majority of *CTLA4* locus-mapped transcripts matches the 'full length' *CTLA4* transcript (ENST00000648405), indicated by the coverage of transcripts throughout the full-length 3'UTR structure (Figure 4.26A - C). FLAIR assembly identified three novel *CTLA4* transcripts. The first novel FLAIR transcript resembles the 'full length' *CTLA4* transcript but with a retained intron between exon 2 and 3. The second FLAIR transcript resembles the 'full length' *CTLA4*, with exon 3 exclusion. Finally, the third novel FLAIR transcript has a short 3'UTR that matches the reference transcript ENST00000648406 (Figure 4.26D).

Looking at the gene coverage tracks and the visual evidence from aligned reads, many *CTLA4* locus-mapped transcripts were found with a short 3'UTR (chr2:203867771 - 203872856) compared to the 'full length' *CTLA4*. Raw reads of both 'full-length *CTLA4*' (ENST00000648405) and the 'short 3'UTR *CTLA4*' transcripts can be seen in Figure 4.26F. Focussing on the 3'UTR region, the 'short 3'UTR' reads were found to have a 3' transcript end 100nt and 140nt away from the closest annotated transcripts (Figure 4.27B - D). The short 3'UTR FLAIR assembled transcript has adopted the 3'UTR structure of the reference annotation, which has a 3' end that is 100nt 5' upstream compared to the sequencing reads (Figure 4.27E). StringTie2 did not report any transcript isoforms with a short 3'UTR (Figure 4.26E, 4.27E).

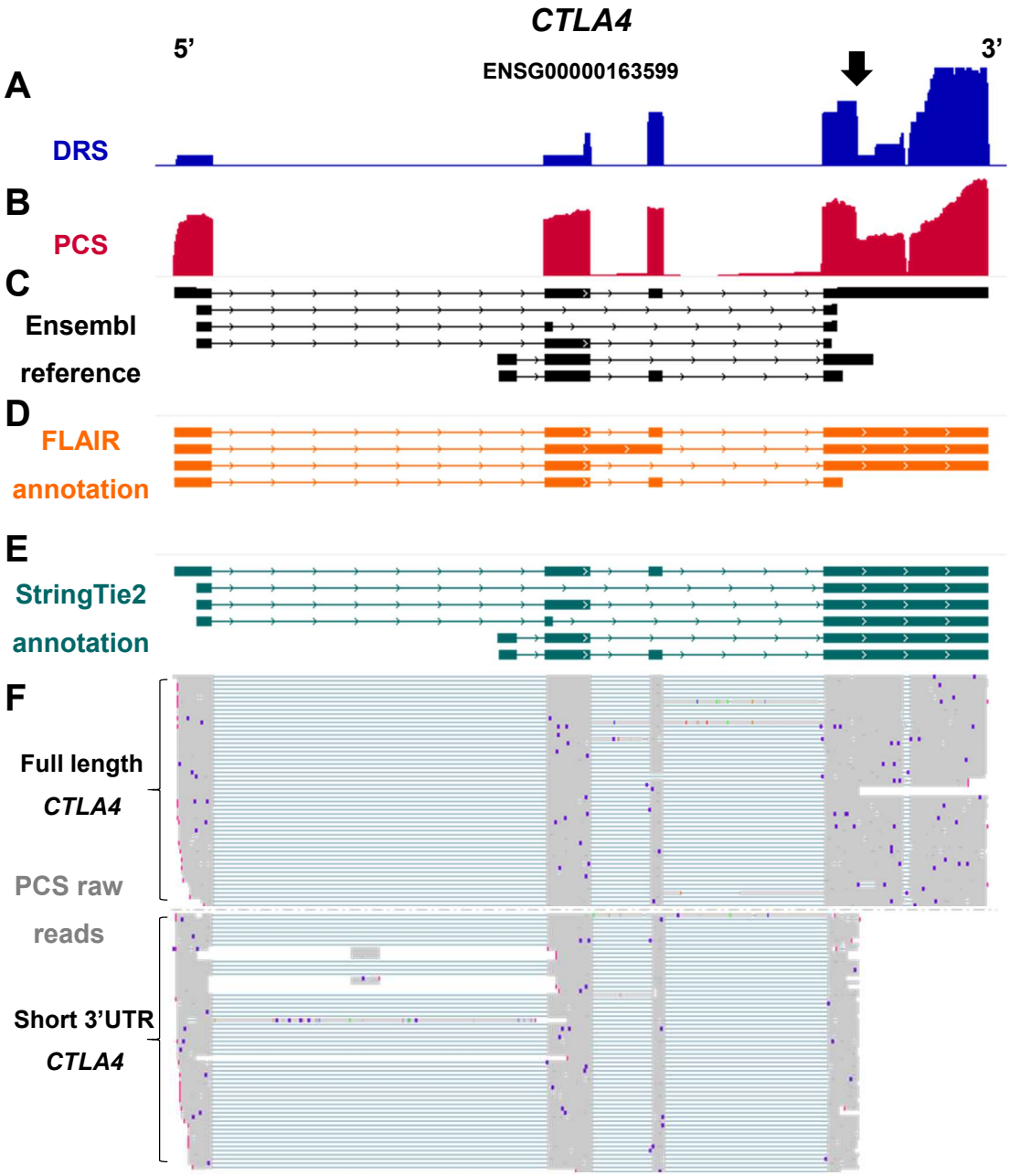


Figure 4.26: Identification of novel *CTLA4* isoform from ccRCC tumours

A) IGV visualisation of combined DRS reads coverage track (Blue) for all sequenced ccRCC tumours in the region of the *CTLA4* gene. **B)** IGV visualisation of of combined PCS reads coverage track (Red) for all sequenced ccRCC tumours in the region of the *CTLA4* gene. **C)** Graphical representation of *CTLA4* transcripts from Ensembl reference annotations (Ensembl release 105, GRCh38). **D)** Graphical representation of *CTLA4* transcripts from *FLAIR* transcriptome assembly annotation. **E)** Graphical representation of *CTLA4* transcripts from *StringTie2* transcriptome assembly annotation. **F)** IGV visualisation of combined PCS raw read alignment tracks at *CTLA4* locus. Reads were representing 'full length' *CTLA4* transcripts (ENST00000648405) and 'Short 3'UTR' *CTLA4* transcripts are grouped separately.

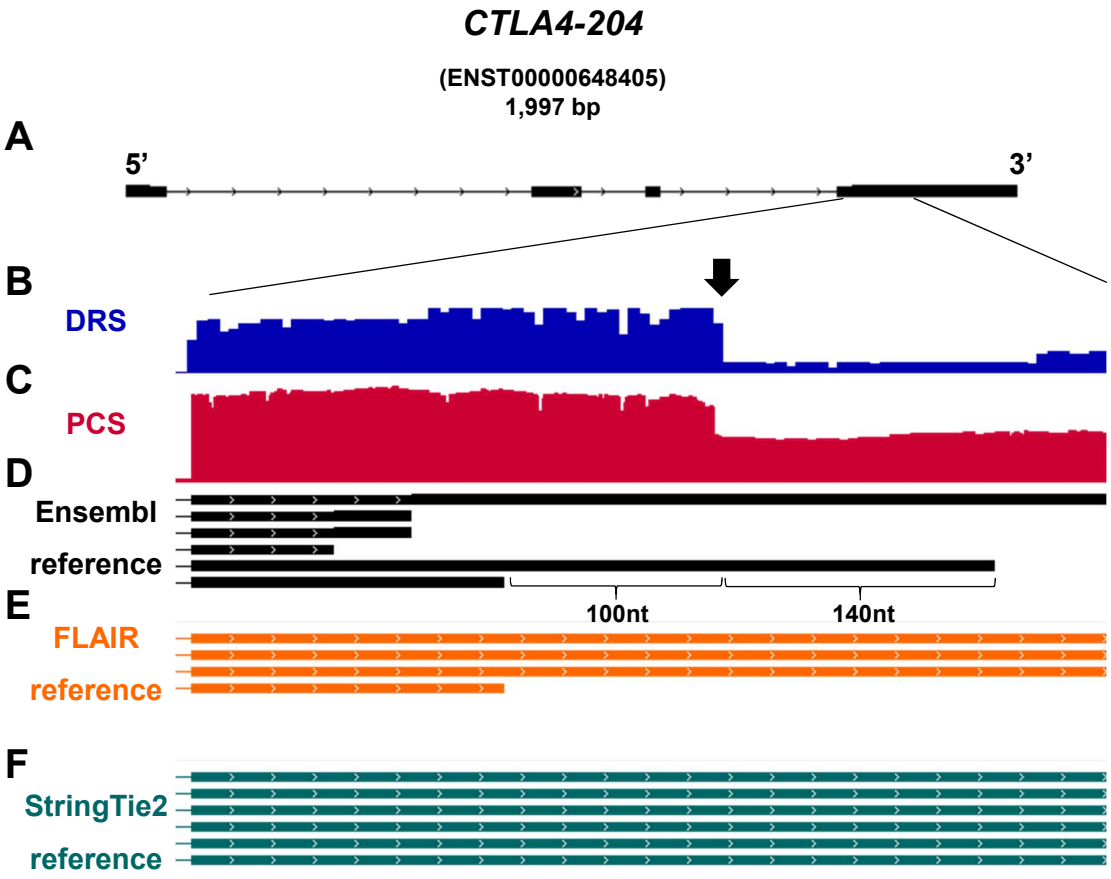


Figure 4.27: Long-read sequencing allows accurate annotation of novel *CTLA4* transcript isoforms at high-resolution

A) Graphical representation of *CTLA4* transcript (ENST00000648405) Ensembl reference gene annotation (GRCh38). **B)** IGV visualisation of combined DRS reads coverage track (Blue) for all sequenced ccRCC tumours in the region of ENST00000648405 3' end (hg38 chr2:203867771:203873965). Black arrow indicates the 3' end of *CTLA4* 'short 3'UTR' transcripts. **C)** IGV visualisation of of combined PCS reads coverage track (Red) for all sequenced ccRCC tumours in the region of ENST00000648405 3' end. **D)** Graphical representation of the 3'end region of *CTLA4* transcripts from Ensembl reference gene annotations (GRCh38). **E)** Graphical representation of the 3' ends of *CTLA4* transcripts (orange) from *FLAIR* transcriptome assembly annotation. **F)** Graphical representation of the 3' ends of *CTLA4* transcripts (green) from *StringTie2* transcriptome assembly annotation.

4.3.16 Identification and validation of a novel soluble PD-L1 transcript isoform in ccRCC tumours

PD-L1 is another crucial immune checkpoint that is critical in regulating tumour immunity. A soluble PD-L1 isoform (truncated with exon 5, 6, 7 absent) has recently been described in the literature (Figure 4.28F) (Ng *et al.*, 2019). Expression of soluble PD-L1 is now associated with disease prognosis and immunotherapy treatment outcomes(Han *et al.*, 2021; Vajavaara *et al.*, 2021). To characterise the soluble PD-L1 expression levels in the ccRCC tumours, reference genome aligned DRS and PCS IGV coverage tracks were visually inspected and compared with the Ensembl gene annotation, as well as using FLAIR and StringTie 2 assembled transcriptomes.

The majority of *PD-L1* transcripts aligned with an isoform (ENST0000381577) that encodes for membrane PD-L1 (Figure 4.28A – B). The transcript for soluble PD-L1 is not registered in the Ensembl gene annotation, but it has been described in the NCBI GenBank database (NM_001314029) (Figure 4.28F). Another transcript in the Ensembl gene annotation (ENST00000474218) partially overlaps with soluble *PD-L1* transcript, and it has been used as a surrogate for mapping soluble *PD-L1* (Ng *et al.*, 2019) (Figure 4.28F). FLAIR identified two novel transcripts resembling soluble *PD-L1* transcript (where one of the transcripts displayed exon 2 skipping), whereas StringTie2 assembled a transcript corresponding to the 'surrogate' soluble PD-L1 transcript from the Ensembl gene annotation (ENST00000474218).

Upon closer inspection of the raw reads, two different soluble *PD-L1* isoforms with varying 3'UTR lengths were found (Figure 4.28G, 4.29D). The novel soluble *PD-L1* isoform has an extra 100nt of 3'UTR compared to the annotated soluble *PD-L1* transcript isoform. Graphical representation of the exon4/3'UTR structures for the membrane, soluble and novel soluble *PD-L1*, can be found in Figure 4.29A. IGV coverage tracks and evidence from raw reads track showed that the novel soluble *PD-L1* isoform could be seen by both DRS and PCS (4.29B – D). Comparing the assembled transcripts from FLAIR and StringTie2, the FLAIR soluble PD-L1 transcripts have a 3'UTR representing

the known soluble *PD-L1* transcripts with a short 3'UTR (Figure 4.29F). In contrast, the StringTie2 assembled transcript represents the 3'UTR structure of the novel soluble *PD-L1* transcript (Figure 4.29G). However, the StringTie2 soluble transcript retained the 'surrogate' soluble *PD-L1* transcript structure, with exons 1, 2 and 3 missing/truncated (Figure 4.28E).

Next, the expression levels of membrane *PD-L1* and the soluble *PD-L1* transcripts were profiled. The number of reads mapped to each isoform-specific region was registered for all tumour samples and subsequently normalised to the library size (expressed as TPM). The expression levels of membrane *PD-L1* transcripts were significantly lower in the recurrent ccRCC tumours compared with non-recurrent tumours (Figure 4.29H). In contrast, neither short 3'UTR nor long 3'UTR soluble *PD-L1* transcripts showed significant differences in expression levels between recurrent and non-recurrent ccRCC tumours (Figure 4.29I – J). qRT-PCR analysis was next performed using primer pairs specifically targeting membrane *PD-L1* transcripts, soluble PD-L1 transcripts (both isoforms) and novel soluble *PD-L1* transcripts. Whilst no significant differential expression of the *PD-L1* transcripts was found between recurrent and non-recurrent ccRCC tumours, their expression in the ccRCC tumours was validated, including the novel soluble isoform (Figure 4.29L – M).

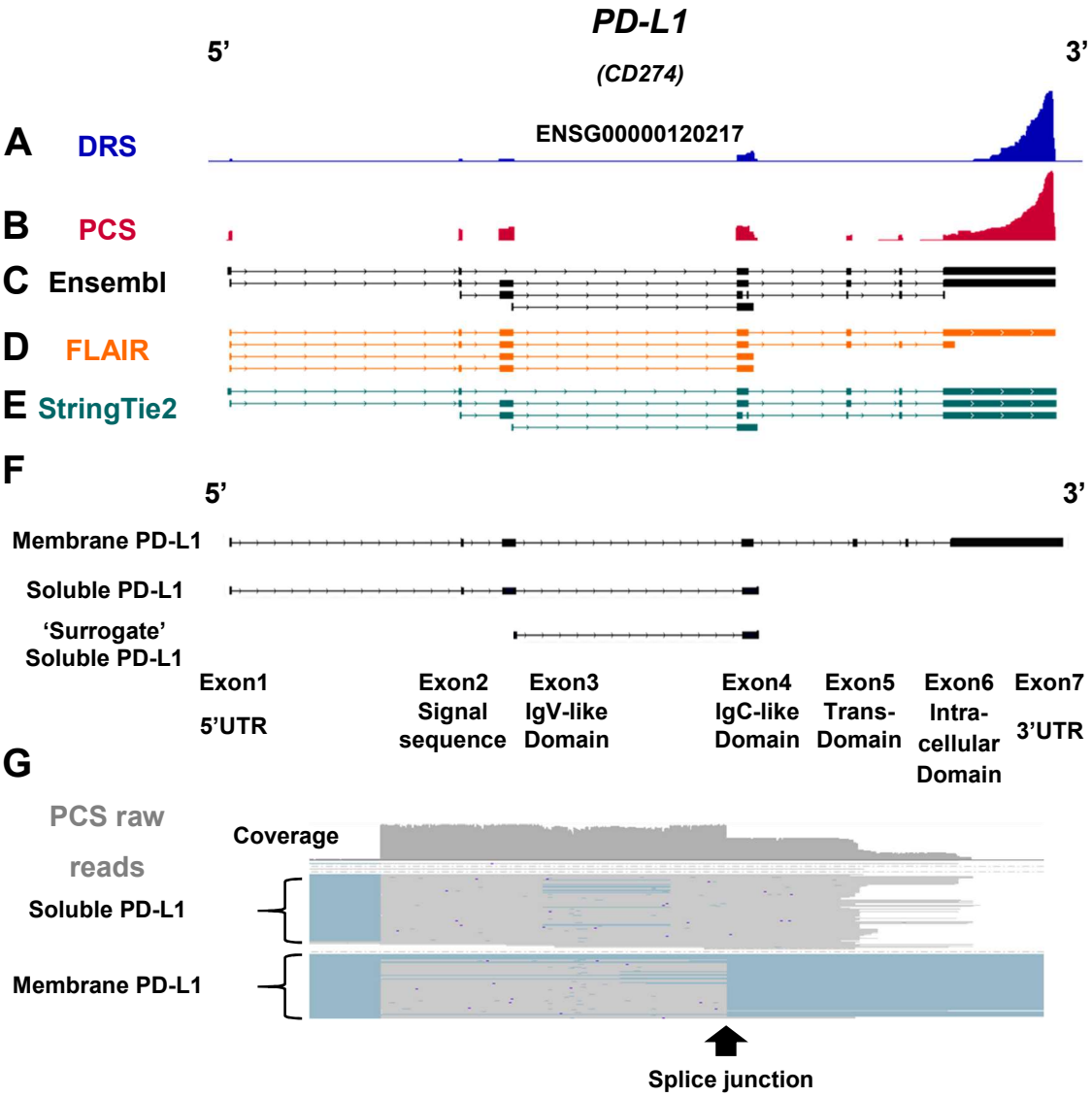


Figure 4.28: Long-read sequencing enable detection of soluble *PD-L1* expression in ccRCC tumours

A) IGV visualisation of combined DRS reads coverage track (Blue) for all sequenced ccRCC tumours in the region of the *PD-L1* gene. **B)** IGV visualisation of combined PCS reads coverage track (Red) for all sequenced ccRCC tumours in the region of the *PD-L1* gene. **C)** Graphical representation of *PD-L1* transcripts from Ensembl reference annotation (GRCh38) **D)** Graphical representation of *PD-L1* transcripts (orange) from *FLAIR* transcriptome assembly annotation. **E)** Graphical representation of *PD-L1* transcripts (green) from StringTie2 transcriptome assembly annotation. **F)** Graphical representation of *PD-L1* transcript ENST00000381577 (membrane PD-L1) from Ensembl reference annotations and NM_001314029 (soluble PD-L1) from NCBI GenBank reference database. **G)** IGV visualisation of PCS coverage track and combined PCS raw read alignment tracks at *PD-L1* locus. Reads representing Soluble *PD-L1* and membrane *PD-L1* transcripts (ENST00000381577) are grouped separately.

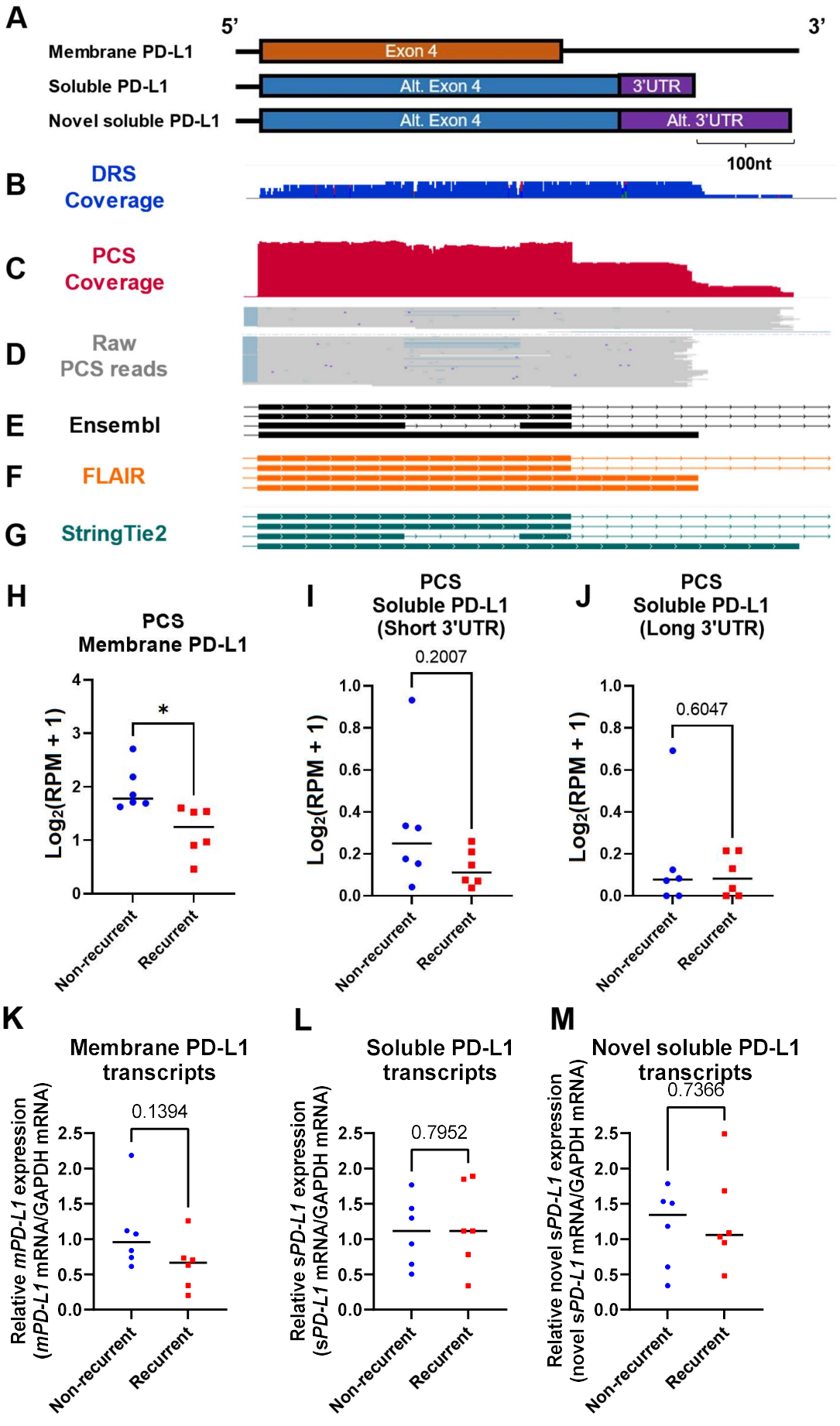

A
5'
3'
Membrane PD-L1
Exon 4
Soluble PD-L1
Alt. Exon 4
3'UTR
Novel soluble PD-L1
Alt. Exon 4
Alt. 3'UTR
100nt
B
DRS Coverage
C
PCS Coverage
D
Raw PCS reads
E
Ensembl
F
FLAIR
G
StringTie2
H
PCS Membrane PD-L1
Log2(RPM + 1)
*
Non-recurrent
Recurrent
I
PCS Soluble PD-L1 (Short 3'UTR)
0.2007
Log2(RPM + 1)
Non-recurrent
Recurrent
J
PCS Soluble PD-L1 (Long 3'UTR)
0.6047
Log2(RPM + 1)
Non-recurrent
Recurrent
K
Membrane PD-L1 transcripts
0.1394
Relative mPD-L1 expression (mPD-L1 mRNA/GAPDH mRNA)
Non-recurrent
Recurrent
L
Soluble PD-L1 transcripts
0.7952
Relative sPD-L1 expression (sPD-L1 mRNA/GAPDH mRNA)
Non-recurrent
Recurrent
M
Novel soluble PD-L1 transcripts
0.7366
Relative novel sPD-L1 expression (novel sPD-L1 mRNA/GAPDH mRNA)
Non-recurrent
Recurrent

Figure 4.29: Identification and validation of a novel soluble *PD-L1* transcript expressed in ccRCC tumours

A) Graphical representation of *PD-L1* transcript isoforms structures near the exon 4 region (hg38 chr9:5450542 – 5463350), with exonic and UTR regions for 'membrane PD-L1', 'soluble PD-L1' and 'novel soluble PD-L1' highlighted. **B)** IGV visualisation of combined DRS reads coverage track (Blue) for all sequenced ccRCC tumours at *PD-L1* locus between hg38 chr9:5450542 – 5463350. **C)** IGV visualisation of combined PCS reads coverage track (Red) for all sequenced ccRCC tumours at *PD-L1* locus between hg38 chr9:5450542 – 5463350. **D)** IGV visualisation of combined PCS raw read alignment tracks at *PD-L1* locus. Reads representing soluble *PD-L1* (NM_001314029) and novel soluble *PD-L1* transcripts with elongated 3'UTRs are separately shown. **E)** Graphical representation of *PD-L1* transcripts from Ensembl reference annotation (GRCh38) at *PD-L1* locus between chr9:5450542 – 5463350. **F)** As in **E**, but for ccRCC PCS FLAIR transcriptome assembly annotation. **G)** As in **F**, but for StringTie2 transcriptome assembly annotation. **H**) Grouped dot plot showing PCS normalised membrane *PD-L1* (defined by reads mapping to chr9:5,468,000 – 5,470,600) transcript isoform expression (RPM) in non-recurrent (blue) and recurrent (red) ccRCC tumours. **I)** Grouped dot plot showing PCS normalised soluble *PD-L1* (defined by reads mapping to chr9:5,462,830 – 5,463,330 and not a membrane *PD-L1* transcript) transcript isoform expression (RPM) in non-recurrent (blue) and recurrent (red) ccRCC tumours. **J)** Grouped dot plot showing PCS normalised novel soluble PD-L1 (defined by reads mapping to chr9: 5,463,280 – 5,463,330 and not a membrane PD-L1 transcript) transcript isoform expression (RPM) in non-recurrent and recurrent ccRCC tumours. **K**) Membrane *PD-L1* mRNA levels measured by qRT-PCR in non-recurrent and recurrent ccRCC tumours (n = 12), relative to average membrane *PD-L1* mRNA levels in non-recurrent ccRCC tumours. Membrane *PD-L1* mRNA levels were normalised to *GAPDH*. **L)** As in **K**, but for soluble *PD-L1* transcripts (all soluble isoforms), **M)** as in **L**, but for novel soluble *PD-L1* transcripts (Long 3'UTR). For **H – M**, two-tailed unpaired T-tests with Welch's correction were used, with $p \leq 0.05$ considered significant. P values of non-significant results are indicated in graphs. Centre line represents median for each group.

4.3.17 Long-read RNA sequencing allows accurate inference of *IDO1* isoforms

IDO1 (Indoleamine 2,3-dioxygenase 1) is an immune checkpoint protein that modulates immune response through the catabolism of tryptophan in the tumour environment. IDO1 is overexpressed in tumour cells across many cancer types, including ccRCC (Lucarelli *et al.*, 2019). Here, most *IDO1* transcripts profiled by DRS and PCS of ccRCC tumours were mapped to the protein-coding ENST00000253237 (IDO1-202). However, coverage tracks from sequencing data revealed that the 3'UTR of most of the expressed *IDO1* transcripts matches that of ENST0000253513 (IDO-201), which was annotated as a nonsense-mediated decay (NMD) transcript due to exon 5 skipping and the presence of a premature stop codon (Figure 4.30A-C). FLAIR identified a novel transcript (highlighted with an asterisk) which has the 3'UTR structure of the NMD transcript but without exon 5 skipping (Figure 4.30D). StringTie2 did not report any assembled transcripts with the short 3'UTR structure of the NMD transcript (Figure 4.30E).

Looking at the reference genome aligned reads from PCS of ccRCC tumours, both exon5 included, and exon skipped *IDO1* transcripts exhibited the two 3'UTR structures (Figure 4.30F). This result highlighted the ability of long-read RNA sequencing to distinguish reads between different transcript isoforms. Moreover, it also demonstrated the potential of misattribution in the reference gene annotation, which can significantly impact any bioinformatics-based study.

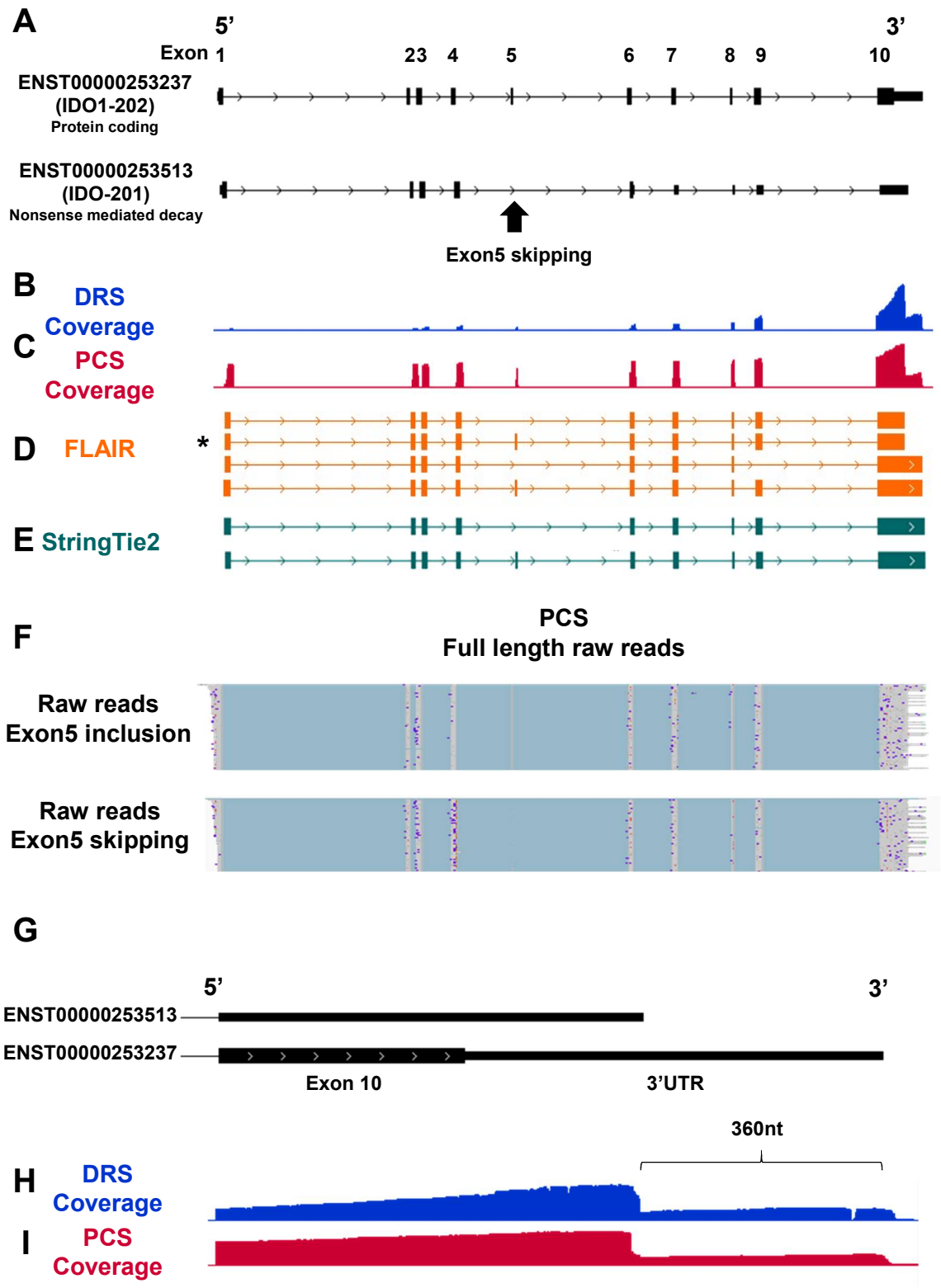

A
5'
3'
Exon 1 23 4 5 6 7 8 9 10
ENST00000253237
(IDO1-202)
Protein coding
ENST00000253513
(IDO-201)
Nonsense mediated decay
Exon5 skipping
B DRS Coverage
C PCS Coverage
D FLAIR
*
E StringTie2
F
PCS
Full length raw reads
Raw reads
Exon5 inclusion
Raw reads
Exon5 skipping
G
5'
3'
ENST00000253513
ENST00000253237
Exon 10
3'UTR
360nt
H DRS Coverage
I PCS Coverage

Figure 4.30: Short 3'UTR is not a hallmark for exon 5 skipping events for *IDO1*

A) Graphical representation of the protein coding *IDO1* transcript (ENST00000253237, IDO1-202) and nonsense mediated decay *IDO1* transcript (ENST00000253513, IDO1-201) from Ensembl reference gene annotations (GRCh38). Exon numbers are indicated above IDO1-202. **B)** IGV visualisation of combined DRS reads coverage track (Blue) for all sequenced ccRCC tumours in the region of the *IDO1* gene. **C)** IGV visualisation of of combined PCS reads coverage track (Red) for all sequenced ccRCC tumours in the region of the *IDO1* gene. **D)** Graphical representation of *IDO1* transcripts (orange) from FLAIR transcriptome assembly annotation. Asterisk highlights novel non-NMD *IDO1* transcript with short 3'UTR **E)** Graphical representation of *IDO1* transcripts (green) from StringTie2 transcriptome assembly annotation. **F)** IGV visualisation of combined PCS raw read alignment tracks at *IDO1* locus. Reads representing transcripts with exon 5 inclusion and exon 5 eskpping are separately shown. **G)** Graphical representation of the 3'end region of *IDO1* transcripts (ENST00000253237 and ENST00000253513) from Ensembl reference gene annotations (GRCh38). **H)** IGV visualisation of combined DRS reads coverage track for all sequenced ccRCC tumours in the region of ENST00000253237 and ENST00000253513 3' end (hg38 chr8 39913891: 39928790). **I)** IGV visualisation of combined PCS reads coverage track for all sequenced ccRCC tumours in the region of the 3' end of ENST00000253237 and ENST00000253513 (chr8 39913891: 39928790).

4.3.18 Identification of a novel *CD24* transcript from ccRCC tumours

CD24 is an anti-phagocytic immune checkpoint protein that shields tumour cells from macrophages when over-expressed (Barkal *et al.*, 2019). *CD24* was expressed in all ccRCC tumours by DRS and PCS, where most transcripts displayed a 3'UTR structure that matches ENST00000619133 and ENST00000606017 (Figure 4.31A – C). Curiously, both DRS and PCS coverage tracks and raw reads tracks showed the presence of *CD24* transcripts with a shorter 3'UTR compared to ENST00000619133 and ENST00000606017 (Figure 4.31B-C, F).

FLAIR assembled transcripts from PCS data showed several novel transcripts with the short 3'UTR (Figure 4.31D). However, the 5' ends of these novel transcripts were not represented at notable levels in the transcriptomic mapping data, as shown in the coverage tracks (Figure 4.31B – C). StringTie2 assembled transcripts were adopted from the reference annotations of ENST00000606017 and ENST00000619869 and overall showed no novel transcripts (Figure 4.31E).

Next, focussing on the 3' end of the short *CD24* 3'UTR, reference-genome aligned reads showed that the short 3'UTR ends at chr6:106971300, which is 60 nt away from the next closest Ensembl annotated reference transcript (ENST00000620405, ENST00000610952). The assembled transcripts from FLAIR and StringTie2 with short 3'UTRs displayed the same 3'end positions as ENST00000619869, which were approximately 60nt 5' upstream from the aligned reads' 3'end (Figure 4.31I-J). In summary, a novel *CD24* transcript with a previously unannotated 3'UTR structure has been discovered from ccRCC tumours using DRS and PCS. Transcriptome assembly methods provided a list of novel transcripts but failed to recapitulate the exact transcript structures shown from mapped reads.

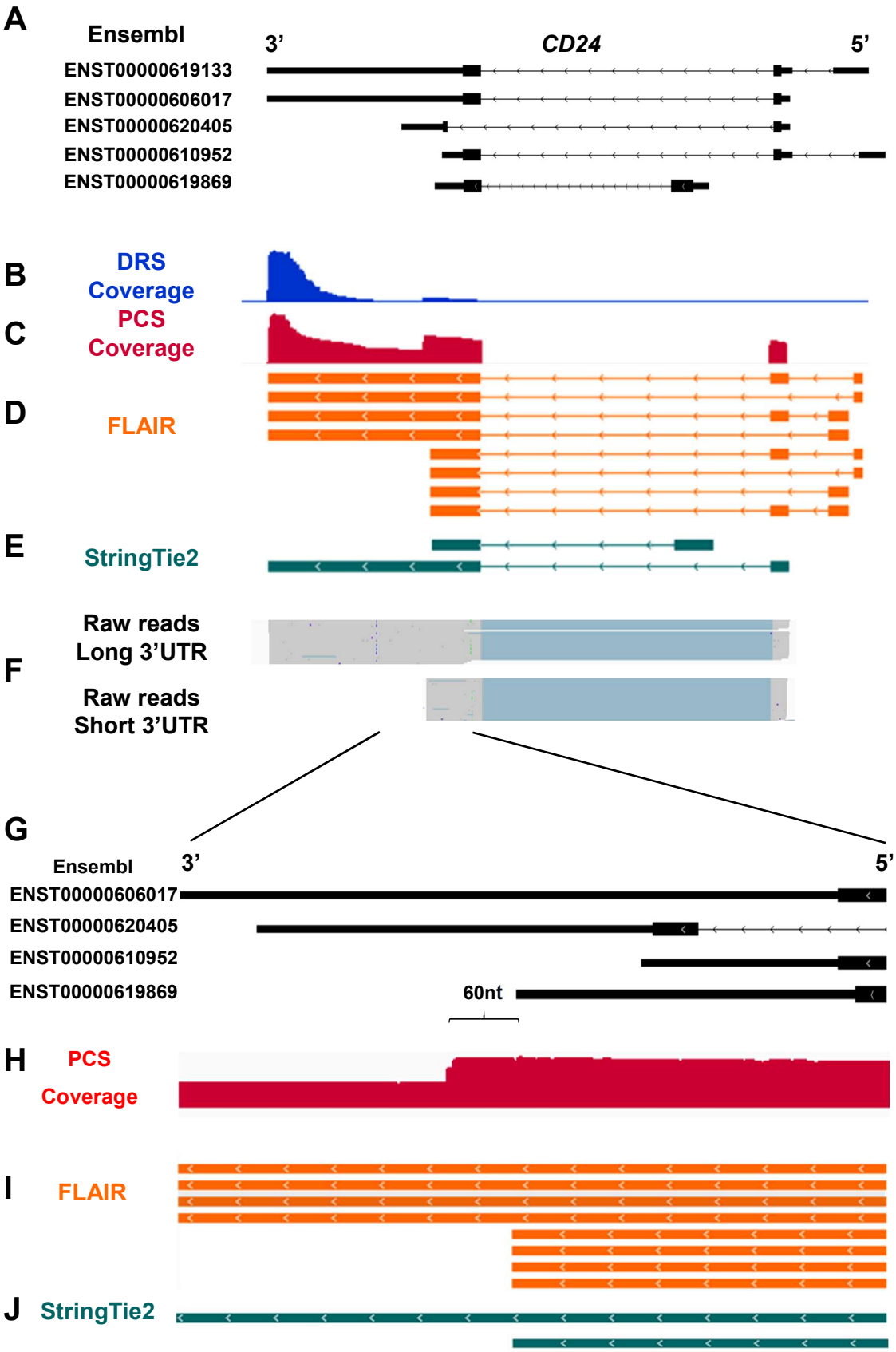

A
Ensembl
3'
CD24
5'
ENST00000619133
ENST00000606017
ENST00000620405
ENST00000610952
ENST00000619869
B
DRS
Coverage
C
PCS
Coverage
D
FLAIR
E
StringTie2
F
Raw reads
Long 3'UTR
Raw reads
Short 3'UTR
G
Ensembl
3'
5'
ENST00000606017
ENST00000620405
ENST00000610952
ENST00000619869
60nt
H
PCS
Coverage
I
FLAIR
J
StringTie2

Figure 4.31: Long-read sequencing identifies a novel *CD24* transcript isoform expressed in ccRCC tumours

A) Graphical representation of *CD24* transcripts from Ensembl reference annotation (GRCh38)**. B)** IGV visualisation of combined DRS reads coverage track (Blue) for all sequenced ccRCC tumours in the region of the *CD24* gene. **C)** IGV visualisation of combined PCS reads coverage track (Red) for all sequenced ccRCC tumours in the region of the *CD24* gene. **D)** Graphical representation of *CD24* transcripts (orange) from FLAIR transcriptome assembly annotation. **E)** Graphical representation of *CD24* transcripts (green) from StringTie2 transcriptome assembly annotation. F**)** IGV visualisation of combined PCS raw read alignment track for all sequenced ccRCC tumours at *CD24* locus. Reads representing *CD24* transcripts with 'long 3'UTR' (hg38 chr6:106969831 – 106971834) and 'short 3'UTR (chr6: 106971300 - 106971834). **G)** Graphical representation of *CD24* transcripts from Ensembl reference annotation (GRCh38) at *CD24* locus between chr6:106971000 - 106971750. **H)** IGV visualisation PCS reads coverage track in the region of chr6:106971000 – 106971750. **I)** Graphical representation of *CD24* transcripts (orange) from FLAIR transcriptome assembly annotation in the region of chr6:106971000 - 106971750. **J)** Graphical representation of *CD24* transcripts from StringTie2 transcriptome assembly annotation in the region of chr6:106971000 - 106971750.

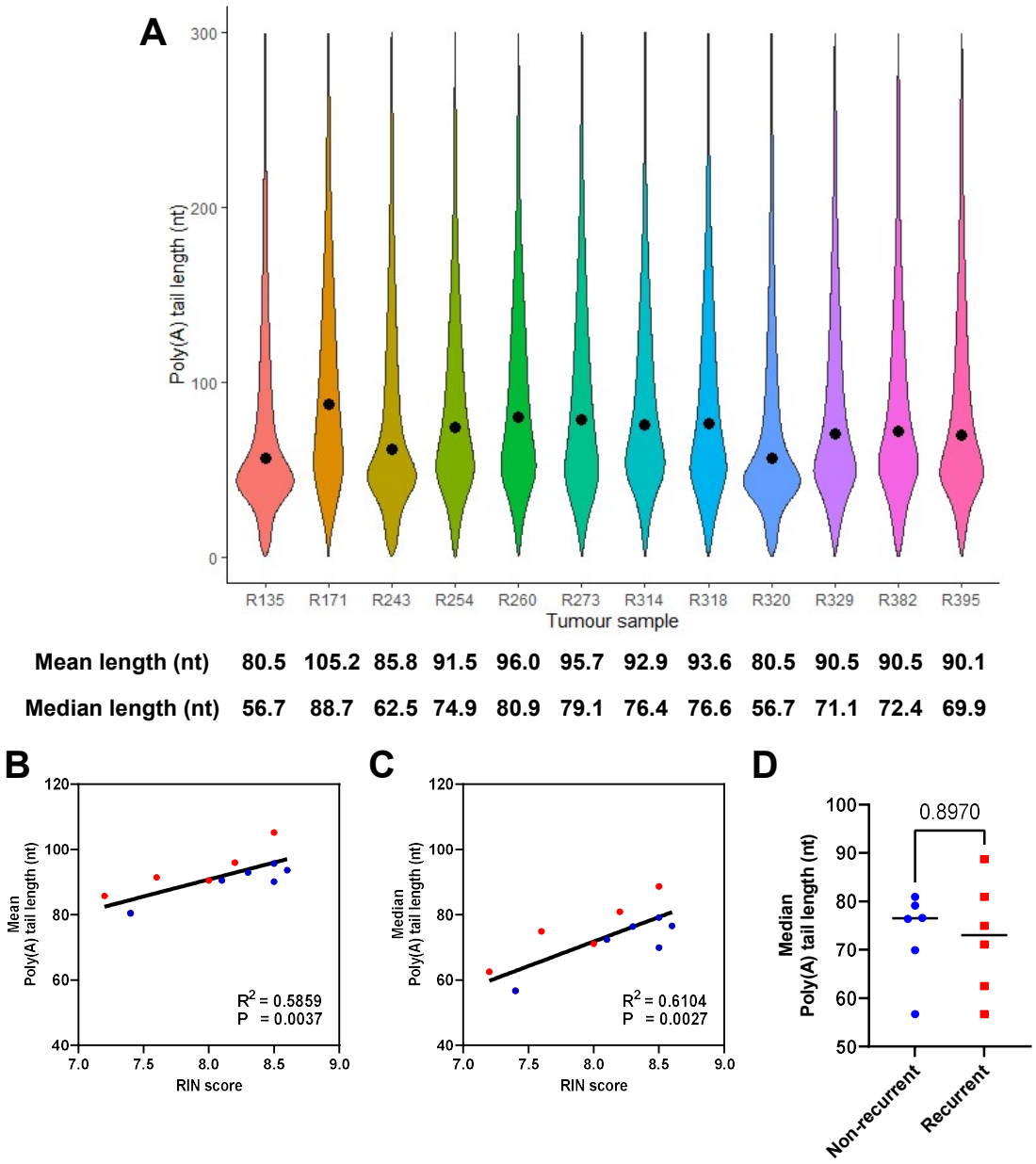
4.3.19 Estimation of poly(A) tail lengths from DRS of ccRCC tumours by nanopolish

Modulation in mRNA polyadenylation is a dynamic process that plays a crucial role in mRNA stability (Passmore and Coller, 2022). Using current signals from DRS, nanopolish was used to characterise poly(A) tail profile in ccRCC tumours (Workman *et al.*, 2019). The mean poly(A) tail lengths across the ccRCC tumours ranged between 80.5 to 105.2 nt, with an average of 91.1 nt. The median poly(A) tail lengths ranged between 56.7 nt and 88.7 nt, with an average of 72.2 nt (Figure 4.32 A).

Seeing that there were variations in the global profiles of poly(A) tail across the tumour samples, the impact of RNA sample degradation was investigated. Significant correlations were found between tumour sample RIN numbers and the mean poly(A) tail lengths (R^2 = 0.5859, p = 0.0037) and also between RIN numbers and median poly(A) tail lengths (R^2 = 0.6104, p = 0.0027) (Figure 4.32B - C). These results suggest that RNA degradation substantially impacts estimated poly(A) tail lengths.

Next, the global poly(A) tail profiles between recurrent and non-recurrent ccRCC tumours were compared. Analysis suggested no significant differences in the median mRNA poly(A) tail lengths between non-recurrent and recurrent ccRCC tumours (p = 0.8970) (Figure 4.32D).

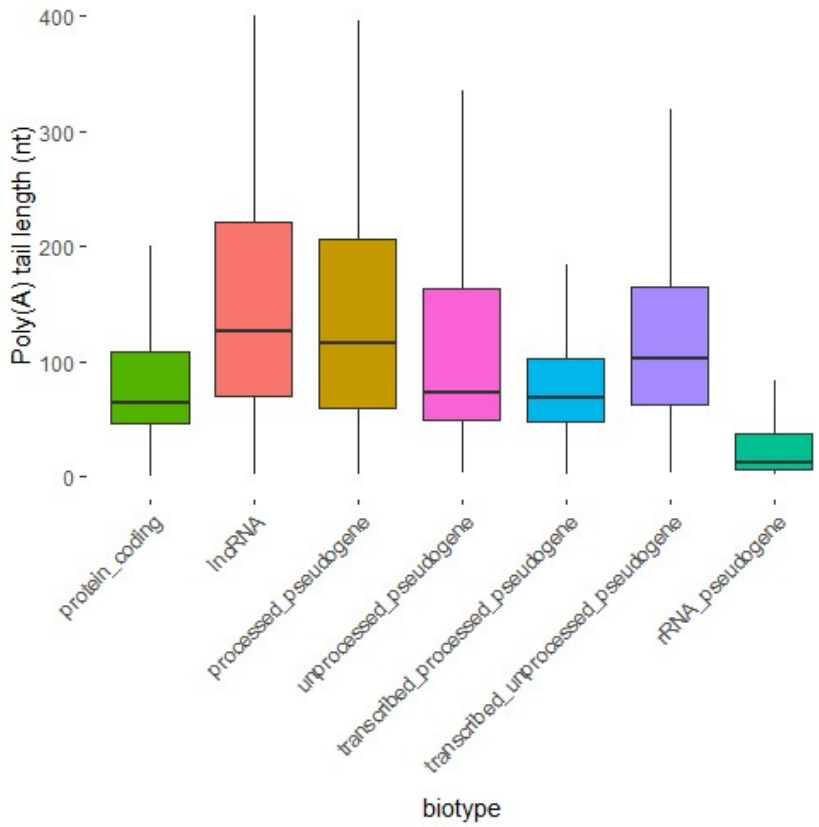
Finally, poly(A) tail lengths of transcripts from different RNA biotypes were examined. Median poly(A) tail lengths of protein-coding genes' transcripts (64.45nt, n = 6,881,791) were found to be substantially shorter than lncRNA (128.46nt, n = 9,948), processed pseudogenes (117.02nt, n = 2,361), as well as transcribed unprocessed pseudogenes (102.93nt, n = 814). The poly(A) tail lengths of rRNA pseudogene transcripts were found to be the shortest amongst the biotypes that were investigated, with a median length of 11.50 nt (n = 39) (Figure 4.33).



Mean length (nt)	**80.5**	**105.2**	**85.8**	**91.5**	**96.0**	**95.7**	**92.9**	**93.6**	**80.5**	**90.5**	**90.5**	**90.1**
Median length (nt)	**56.7**	**88.7**	**62.5**	**74.9**	**80.9**	**79.1**	**76.4**	**76.6**	**56.7**	**71.1**	**72.4**	**69.9**

Figure 4.32: Poly(A) tail length profiling in ccRCC tumours by nanopolish

A) Violin plots showing poly(A) tail lengths of transcripts from ccRCC tumours, estimated by nanopolish using DRS data. Dot within violin represents median. Mean and median poly(A) tail lengths are indicated below graph. **B)** Correlation between mean poly(A) tail lengths from each tumour sample and corresponding RIN score. **C)** Correlation between median poly(A) tail lengths from each tumour sample and corresponding RIN score. **D)** Grouped dot plot showing median poly(A) tail lengths of transcripts from non-recurrent and recurrent ccRCC tumours. For **B – C**, diagonal lines represent the line of best fit. R^2 values were computed tomeasure goodness-of-fit, and P values generated from F-test, with $p \leq 0.05$ considered significant. For **D**, two-tailed unpaired T-tests with Welch's correction were used, with $p \leq 0.05$ considered significant. P values of non-significant results are indicated in graphs. Centre line represents median for each group.



Biotype	Min.	1st Quartile.	Median	Mean	3rd Quartile	Max
Protein coding	0.00	46.68	64.45	85.45	108.52	918.70
lncRNA	1.59	70.91	128.46	152.52	224.15	619.46
Processed pseudogene	2.27	60.46	117.02	140.81	210.30	573.12
Unprocessed pseudogene	3.56	48.59	73.17	117.36	166.25	531.92
Transcribed processed pseudogene	1.02	48.00	68.27	86.21	102.52	544.73
Transcribed unprocessed pseudogene	3.61	63.03	102.93	128.67	168.19	477.54
rRNA pseudogene	1.88	5.41	11.50	45.83	37.56	328.17

Figure 4.33: Poly(A) tail length profiles per biotype in ccRCC tumours

Boxplots showing poly(A) tail lengths of transcripts of different biotypes from ccRCC tumours, estimated by nanopolish using DRS data. For each box, boxed area represent lower and upper quartile, with black line showing the median. Extended whiskers show the highest and lowest values within 1.5 times of respective interquartile ranges. Summary statistics are displayed in table below graph.

4.3.20 Differential poly(A) tail lengths found between different immune checkpoint transcript isoforms

Focussing on tumour immune checkpoints, the poly(A) tail lengths of *IDO1*, *CD24* and *PD-L1* isoforms were characterised and compared. For *IDO1*, transcript isoforms with 'long 3'UTR' (3'UTR ending at chr8: 39928790) and 'short 3'UTR' (3'UTR ending at chr8: 39928444) were compared. Poly(A) tail lengths of transcripts with 'short 3'UTR' (n = 389, mean = 118.2nt) were significantly longer compared to transcripts with 'long 3'UTR' (n = 184, mean = 100.1nt) (Figure 4.34A). *IDO1* transcripts' poly(A) tail lengths were significantly longer in recurrent ccRCC tumours (n = 252, mean = 102.6nt) compared to non-recurrent tumours (n = 319, mean = 120.6nt) (Figure 4.34B). The IDO1 transcripts with short 3'UTR drove this difference. *IDO1* transcripts with long 3'UTR showed no significant differences between non-recurrent and recurrent ccRCC tumours (Figure 4.34C). For *IDO1* transcripts with short 3'UTR, the analysis revealed significantly longer poly(A) tail lengths in recurrent ccRCC tumours (n = 167, mean = 128.1nt) compared to non-recurrent tumours (n = 222, mean = 107.7nt) (Figure 4.34D).

Across all 12 ccRCC tumours, poly(A) tail lengths of long 3'UTR *CD24* transcripts (3'UTR ending at chr6:106975465, n = 13562, mean = 102.9nt) were significantly longer than the novel short 3'UTR *CD24* transcripts (3'UTR ending at chr6: 106971300 - 106971834, n = 2601, mean = 88.22nt) (Figure 4.34E). No statistically significant differences in *CD24* mRNA poly(A) length were found between non-recurrent and recurrent ccRCC tumours (Figure 4.34F – H).

Finally, for *PD-L1* (*CD274*) transcripts, a borderline non-significant trend (p = 0.0565) was observed between poly(A) lengths of transcripts that encode for membrane PD-L1 and soluble PD-L1 (Both short and long 3'UTR) (Figure 4.34I). Poly(A) tails of membrane *PD-L1* transcripts (n = 36) have a mean length of 152.9nt and a median length of 142.4nt, whereas mean and median poly(A) tail lengths of soluble PD-L1 transcripts (n = 9) are 98.5nt and 89.3nt respectively.

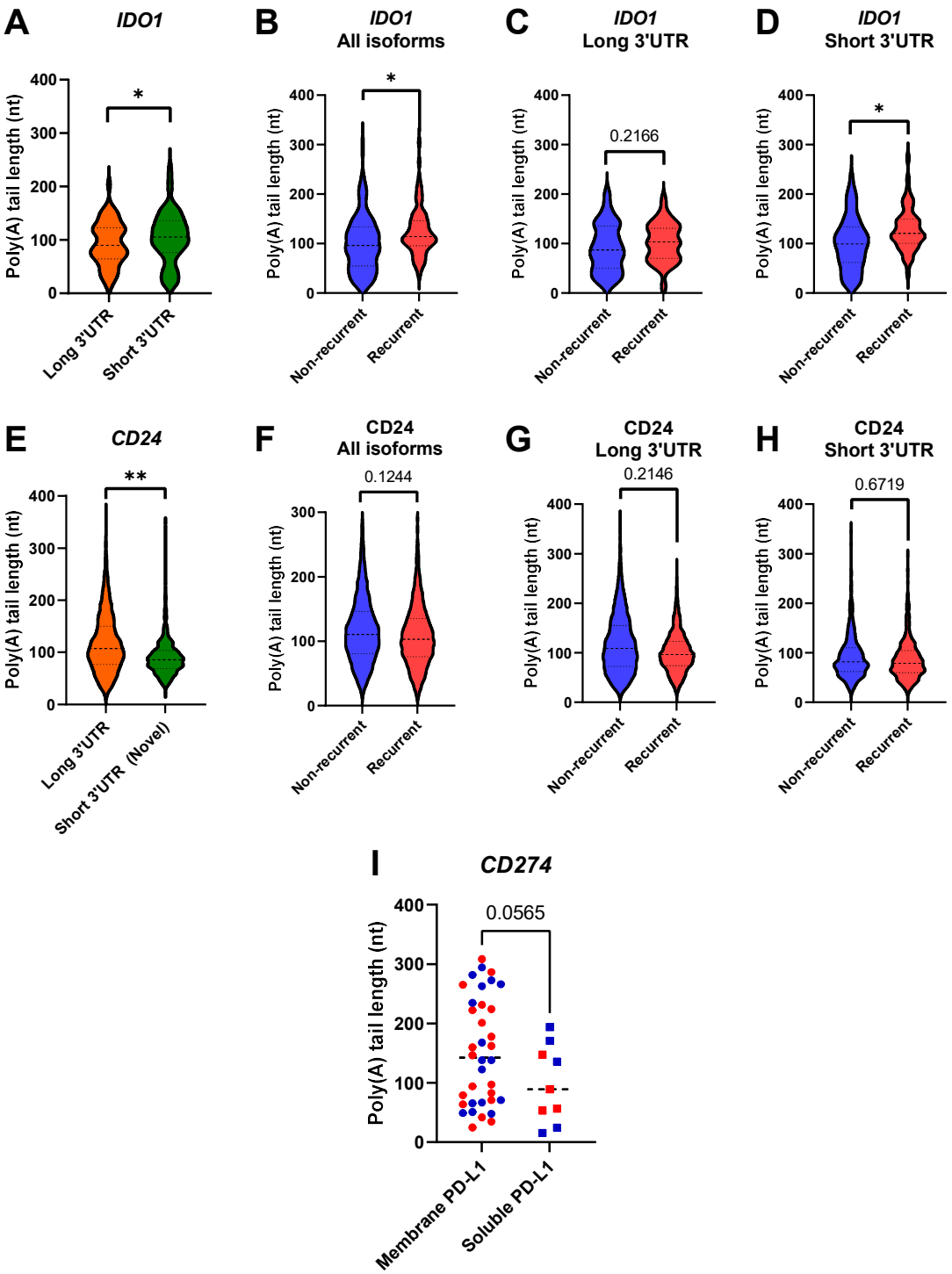

A
IDO1
Poly(A) tail length (nt)
400
300
200
100
0
*
Long 3'UTR
Short 3'UTR
B
IDO1
All isoforms
*
Non-recurrent
Recurrent
C
IDO1
Long 3'UTR
0.2166
D
IDO1
Short 3'UTR
*
E
CD24
**
Long 3'UTR
Short 3'UTR (Novel)
F
CD24
All isoforms
0.1244
G
CD24
Long 3'UTR
0.2146
H
CD24
Short 3'UTR
0.6719
I
CD274
0.0565
Membrane PD-L1
Soluble PD-L1

Figure 4.34: Tumour immune checkpoint isoforms display differential poly(A) tail lengths

A) Violin plot showing nanopolish estimated poly(A) tail lengths of *IDO1* transcripts with 'long 3'UTRs' (Final exon and 3'UTR = chr8: 39927830 - 39928790) and 'short 3'UTRs' (Final exon and 3'UTR = chr8: 39927830- 39928444) from DRS data of all 12 ccRCC tumours. **B)** Violin plot showing poly(A) tail lengths of all *IDO1* transcripts between non-recurrent and recurrent ccRCC tumours. **C)** Violin plot showing poly(A) tail lengths of *IDO1* transcripts with long 3'UTR from non-recurrent and recurrent ccRCC tumours. **D)** Violin plot showing poly(A) tail lengths of *IDO1* transcripts with short 3'UTR from non-recurrent and recurrent ccRCC tumours. **E)** Violin plot showing nanopolish estimated poly(A) tail lengths of *CD24* transcripts with 'long 3'UTRs' (Final exon and 3'UTR = chr6: 106969831 – 106971834) and 'short 3'UTRs' (Final exon and 3'UTR = chr6: 106971300 - 106971834) from DRS data of all 12 ccRCC tumours. **F)** Violin plot showing poly(A) tail lengths of all *CD24* transcripts between non-recurrent and recurrent ccRCC tumours. **G)** Violin plot showing poly(A) tail lengths of *CD24* transcripts with long 3'UTR from non-recurrent and recurrent ccRCC tumours. **H)** Violin plot showing poly(A) tail lengths of *CD24* transcripts with short 3'UTR from non-recurrent and recurrent ccRCC tumours. **I)** Grouped dot plot showing poly(A) tail lengths of *CD274* (*PD-L1*) transcripts encode for membrane PD-L1 and soluble PD-L1, from DRS data of all 12 ccRCC tumours. Blue dots represent transcripts coming from non-recurrent ccRCC tumours, and red dots for recurrent ccRCC tumours. Throughout, two-tailed nested t-tests were used, with $p \leq 0.05$ considered significant. * ≤ 0.05, ** ≤ 0.01. P values of non-significant results are indicated in graphs.

4.4 Discussion

This chapter characterised the transcriptomic profiles of ccRCC tumours sequenced by PCS and DRS. Results in this chapter provide insights into the gene expression profiles of recurrent and non-recurrent ccRCC tumours. Global gene expression profiles of the 12 tumours were assessed using the unsupervised methods: PCA and hierarchical clustering (Chapter 4.3.1 – 4.3.2). Unfortunately, neither method could identify tumour clusters correlating with patients and clinical information due to the low number of analysed samples per group. PCA is a commonly used method that reduces the dimensions of transcriptomic data and identifies 'principal components (PC) that can explain the variances between datasets (Son *et al.*, 2018). The low percentages of variance explained by the top PCs (Figure 4.1) demonstrate the high dimensionality nature of tumour transcript expression profiles and the complex nature of the tumour transcriptomes.

The inability in linking clinical and patient information by PCA can be explained by two factors: the number of samples and the size of the phenotype-associated effect. Firstly the number of samples sequenced by this study (n =12) is comparatively low compared to data generated by microarray, single-cell sequencing and large-scale bar-coded short-read RNAseq experiments. PCA has been extensively used to identify outliers from a few transcriptomic profiles, which can be attributed to library preparation, sequencing or bioinformatics analysis errors (Conesa *et al.*, 2016). However, if the size of the phenotype-associated effect is small, a more significant number of samples and a higher number of PCs are needed to identify clusters confidently. For example, using microarray data from 208 human B-cell lymphoma samples, Lenz *et al.* showed that the majority of sex-related information (118 male and 90 female) were contained between PCs 20 to 33, despite dramatic changes in the expression levels of chromosome Y genes and *XIST* (Lenz *et al.*, 2016).

Similar to results from PCA, hierarchical clustering did not result in any precise match between clinical information and tumour samples (Chapter 4.3.2). The agglomerative

clustering method that was used clusters samples with the most similar gene expression profiles (based on Spearman correlations here) and subsequently merged with the next most similar group of profiles. Here, whilst clustering patterns were different between PCS and DRS, samples were clustered similarly between the reference genome and reference transcriptome-aligned data (Figure 4.3 – 4.4). This result supported previous data that showed highly correlated gene expression levels between the reference genome and reference transcriptome-aligned DRS and PCS data (Figure 3.21, 3.22).

Though global transcriptome profile characterisation by PCA and hierarchical clustering did not identify obvious distinctions between recurrent and non-recurrent ccRCC tumours, differential gene expression analysis identified hundreds of DEGs associated with ccRCC recurrence. Combining results from both reference genome and reference transcriptome aligned data, DRS and PCS identified 75 and 274 significantly DEGs, respectively (Figure 4.6). Strikingly, many of the significantly downregulated genes such as *CD8B* (T cell surface glycoprotein CD8 beta chain)*, NKG7* (Natural Killer Cell Granule 7), *GZMK* (granzyme K), *TRAC* (T cell receptor alpha constant) are known to be highly or exclusively expressed by immune cells. Together with subsequent GSEA and immune deconvolution results, sequencing data agree with recent studies which showed reduced tumour immune infiltration as a hallmark of recurrent ccRCC tumours (Ghatalia *et al.*, 2019; Peng *et al.*, 2022).

Among the 21 significant DEGs in recurrent ccRCC identified by DRS and PCS, seven genes were significantly upregulated via both reference genome and reference transcriptome alignment (Figure 4.6G). Furthermore, supporting the aggressive phenotype that is displayed by recurrent ccRCC, 3 out of the six upregulated genes, namely *CCL20* (C-C motif chemokine ligand 20), *KLF15* (Kruppel like factor 15), and *SCD5* (stearoyl-CoA desaturase 5) have been demonstrated to associate with poor prognosis (Tian *et al.*, 2019).

CCL20 is a chemokine secreted by various cells, including macrophages, B cells, neutrophils, natural killer cells, T_H17 cells, and endothelial cells(Lee *et al.*, 2013). CCL20

exclusively binds to CCR6 (C-C chemokine receptor 6). CCR6 is expressed in immune cells ($CD8^+$ T cells, T_H17, T_{regs}, B cells) and also aberrantly upregulated in tumour cells in various cancers, including ccRCC(Kadomoto, Izumi and Mizokami, 2020). An *in vitro* co-culture study using ccRCC cell lines and human macrophages showed that macrophages promote tumour cell migration and invasion by releasing CCL20, which activates AKT upon CCR6 binding (Kadomoto, Izumi, Hiratsuka, *et al.*, 2020). In addition, using flow cytometry and immunohistochemical staining, CCL20 expression was previously shown to recruit immunosuppressive T_{regs} to ccRCC tumour tissue, with T_{reg} infiltration associated with poorer patient prognosis (Oldham *et al.*, 2012). This was not observed in cell-type deconvolution analysis of the tumours. Further gene expression characterisation with a validation cohort of recurrent and non-recurrent ccRCC tumours will be helpful in establishing these disease recurrence-associated signatures.

Most DEGs were detected in both DRS and PCS, with highly correlated log_2FoldChange observed between DRS and PCS. This suggests that the differences in the number of significant DEGs are due to read depth, where the higher number of mapped reads in PCS provided higher statistical power for differential expression analysis by DESeq2. In contrast, when the lists of DEGs identified by reference genome and reference transcriptome alignment were compared, a substantial number of genes were identified solely by one method and not the other. For reference genome-aligned data, most of the exclusive DEGs were lncRNAs, which was expected since the reference transcriptome has few registered ncRNAs for mapping. It is also important to note that although most of the differentially expressed lncRNAs were found in both reference genome aligned DRS and PCS, they were only registered as significantly differentially expressed using PCS. This highlights the importance of having high sequencing depth to identify significant DEGs.

Most of the 60 reference transcriptome alignment exclusive DEGs are protein-coding genes, including *CCL4* (Chemokine ligand 4), *CCL5* (Chemokine ligand 5), *HLA-DPB1* (Major histocompatibility complex class II, DP Beta 1), *TRBC1* (T cell receptor beta

constant 1) and *TRBC2* (T cell receptor beta constant 2) (Figure 4.7D). Only one protein-coding DEG was exclusively mapped by reference genome alignment: *PDCD1* (ENSG00000188389), which encodes for the immune checkpoint protein PD-1. Using reference genome alignment, *PDCD1* was found to be a highly significant DEG between recurrent and non-recurrent ccRCC tumours with a high averaged RPM across tumours. Many of these DEGs are critical immune genes, and failure to identify these genes would present a loss of valuable information on the tumour immune landscape. The usage of a reference genome or reference transcriptome for transcriptomic analysis directly impacts the number of mapped genes and gene expression levels (Srivastava *et al.*, 2020). Read alignment tools, including minimap2, typically generate a mapping quality score for each read based on how closely the read aligns with the reference (Li, 2018). It would be interesting to investigate if the reads mapped to reference transcriptome-exclusive DEGs failed to map to the reference genome or mapped to genes that closely resemble the DEGs. These results demonstrate the importance of using multiple reference alignment methods to analyse RNAseq data.

In agreement with existing literature, GSEA showed significantly suppressed immune cell and antigen presentation pathways in recurrent ccRCC tumours. The suppression in immune cell pathways was further supported by the results from tumour immune infiltrate deconvolution algorithms, where lower levels of tumour infiltrating immune cells were found in recurrent ccRCC tumours. Reduced antigen presentation in recurrent ccRCC contributes to immune surveillance evasion and the loss of tumour-infiltrating immune cells in the TME (de Charette *et al.*, 2016). Gene expression of antigen presentation machinery, including MHC class I molecules, correlates with better ccRCC patient prognosis (Matsushita *et al.*, 2016; Sekar *et al.*, 2016). Transcriptional activity of MHC class I genes is mainly regulated in three ways: i) Binding of NF-kB to the gene Enhancer A region, ii) Engagement of Interferon regulatory factor 1 (IRF1) to the Interferon stimulated response element (ISRE), and iii) occupancy of the IRF1-responsive transcription factor NLRC5 (NLR Family CARD Domain Containing 5) at the promoter

(René *et al.*, 2016). Gene expression levels of *IFNG*, *IRF1* and *NLRC5* in PCS data were all found to downregulated in recurrent ccRCC tumours compared to non-recurrent counterparts. The reduced mRNA levels of *IFNG* reflected the altered tumour immune landscape in recurrent ccRCC tumours, which may have contributed to the suppressed antigen presentation pathways. The loss of antigen presentation pathways can also be the results of the dysregulated HIF pathway in ccRCC tumours. Activation of the HIF pathway in ccRCC suppresses the expression levels of MHC class I molecules and other antigen presentation proteins (*TAP1*, *TAP2*, *LMP7*) and contributes to reduced recognition and killing from $CD8^+$ T cells (Sethumadhavan *et al.*, 2017). The degree of HIF activation was previously shown to correlate with ccRCC recurrence and poorer prognosis (Schödel *et al.*, 2016). Although GSEA and DGE analysis showed no significant changes in the activation or suppression of the HIF pathway, HIF activity is also highly regulated by post-translational modifications (Albanese *et al.*, 2021). Therefore, an integrated multi-omics approach is needed to comprehensively characterise the causal relationship between tumour cell molecular phenotypes and TME immune landscapes in ccRCC.

Fat metabolism is another highly dysregulated pathway impacted by the hyperactivated HIF pathway in ccRCC. Here, GSEA of KEGG pathways revealed significantly upregulated Fat digestion and absorption pathways in recurrent ccRCC tumours (Figure 4.14B). One of the most highly upregulated genes in recurrent ccRCC tumours was *APOB*. Apolipoprotein B (ApoB) is a major protein component of lipoproteins and facilitates lipid transportation (Ren *et al.*, 2019). A study has shown that ApoB is highly accumulated in ccRCC, and the level of ApoB protein positively correlates with ccRCC disease progression. However, the source of *APOB* expression in ccRCC tumours is currently unknown: neither ccRCC cell lines (RCC4, 786-O) nor ccRCC patients-derived ccRCC cells were found to express *APOB* (Velagapudi *et al.*, 2018). In ccRCC, *APOB* gene expression level was shown to confer with a worse treatment response to the TKI sunitinib but a better response to the PD-1 inhibitor Nivolumab (Puzanov, 2022).

Investigation into the potential role of ApoB in regulating ccRCC tumour immunity and identifying the cell origin of *APOB* expression in ccRCC via single-cell experiments may improve ccRCC therapy stratification.

The use of computational methods to estimate tumour purity and infer cell-type proportions using bulk RNAseq has advanced our understanding of the TME substantially. In agreement with the literature, tumour purity calculated by ESTIMATE showed low levels of tumour purity across the ccRCC tumour. In addition, ESTIMATE and xCell showed significantly lower levels of tumour immune infiltrates in recurrent ccRCC tumours compared to non-recurrent ccRCC tumours. Interestingly, whilst DRS and PCS immune scores showed a significant correlation, the raw immune score values from PCS were consistently higher than DRS. ESTIMATE generates a stromal score and an immune score by performing single-sample GSEA based on the gene expression levels of 141 gene signatures each. Tumour purity is inferred by combining the two scores (ESTIMATE Score) against a validated regression model for purity using TCGA data (Yoshihara *et al.*, 2013). Since library-scaled, normalised gene expression data was used as input, estimated tumour purity levels, stromal scores and immune scores should be comparable between sequencing runs and experiments regardless of the sequencing depth. However, read-depth can have an impact on the detection of gene signatures. Recommended guideline on minimum read-depth for genome sequencing methods has previously been published (Jennings *et al.*, 2017). However, consensus on RNAseq and long-read sequencing remain unclear. *In silico* study on the impact of read-depth (for both short- and long-read) on tumour purity estimation would be helpful for future sequencing experimental design.

In addition to the depletion of tumour infiltrating immune cell population, the composition of immune cell types was also altered in recurrent ccRCC tumours. In line with a recently published single-cell sequencing study, recurrent ccRCC tumours showed a significant reduction in $CD8^+$ T cells compared to non-recurrent tumours, as shown by both CIBERSORTx and EPIC (Figure 4.19) (Peng *et al.*, 2022). Both CIBERSORT and EPIC

were developed for Illumina gene expression data. Benchmarking studies using DRS and PCS will be immensely useful for estimating the depth required for detecting different tumour infiltrating immune cells accurately. A recent study using single-cell short-read sequencing on the 10x platform with bar-coded long-read PCR-cDNAseq showed differential isoform usage between B cells, T cells and monocytes (Volden and Vollmers, 2022). Identifying cell-type-specific transcript isoform represents a promising strategy for cell-type deconvolution.

The difference in the $CD8^+$ T cell population between the tumours was primarily driven by three non-recurrent tumours (278, 318, 320) that showed distinctly high proportions of $CD8^+$ T cells (Figure 4.19E – F). These tumours also exhibited high levels of immune scores via ESTIMATE and xCell (Figure 4.17I, J). $CD8^+$ T cell exhaustion markers were highly elevated in the three non-recurrent ccRCC tumours (Figure 4.15B). Expression levels of exhaustion markers *TOX* and *PDCD1* (PD-1) were also highly upregulated in the three tumours (Figure 4.19I-J). Since the expression of exhaustion markers was not scaled to the population of $CD8^+$ T cells, it is unclear if there are more or less exhausted CD8+ T cells in the three non-recurrent ccRCC tumours proportionally. However, these results indicate the presence of $CD8^+$ T cells showing exhausted phenotypes. Results here also showed that the other three non-recurrent ccRCC tumours have similar levels of tumour infiltrating immune cells (including $CD8^+$ T cells), expression patterns of $CD8^+$ exhaustion markers and MHC proteins with the non-recurrent ccRCC tumours.

T cell exhaustion is the result of chronic antigen presentation. At the mRNA expression level, MHC proteins were not specifically upregulated in 278,318 and 320, nor did they form a distinct cluster of MHC expression pattern (Figure 4.16). Persistent TCR activation promotes nuclear localisation of the nuclear factor of activated T cells (NFAT) and promotes the transcription of *TOX* (Seo *et al.*, 2019). siRNA-mediated knockdown of *TOX* in $CD8^+$ T cells results in the depletion of immune checkpoint proteins such as PD-1, whereas overexpression of *TOX* enhances PD-1 expression(Kim *et al.*, 2020). Remarkably, $CD8^+$ T cells with conditional deletion of *TOX* can still generate functional

effector T cells but not exhausted T cells(Khan *et al.*, 2019). Therefore, inhibition of TOX by small molecules is a promising therapeutic approach to revive exhausted $CD8^+$ T cells where intense efforts are currently being made (Agrawal *et al.*, 2019; Radaeva *et al.*, 2021). Combinatorial TOX inhibitors and anti-PD1 therapy in tumours with high levels of TOX_{hi} $PD1_{hi}$ $CD8^+$ T cells, such as the three non-recurrent ccRCC tumours, may represent a viable therapeutic approach.

Deconvolution of other immune cell types in the ccRCC tumours presented mixed results. CIBERSORT analysis showed that M2 macrophages were the predominant type in the ccRCC tumours using DRS and PCS data (Figure 4.18A – B). Data from EPIC suggested a significant decrease in the proportion of macrophages in recurrent ccRCC tumours, whereas CIBERSORT showed no significant changes between the tumours (Figure 4.20A – D). Using PCS data, CIBERSORT showed a significant suppression in activated NK cells in recurrent ccRCC tumours. However, neither CIBERSORT using DRS data nor EPIC (PCS and DRS) could detect NK cells sufficiently for meaningful comparisons (Figure 4.20E - H). These results show the current limitation in cell-type deconvolution methods, where validation by flow cytometry and immunohistochemistry may be required.

Next, DRIMseq and DEXseq analysis identified seven genes that showed differential transcript usage between recurrent and non-recurrent ccRCC tumours. *CMC1* was shown to express four transcripts in ccRCC tumours and displayed recurrence-associated DTU. CMC1 protein plays a crucial role in the biogenesis of mitochondrial complex IV, which is critical for ATP synthesis via oxidative phosphorylation (Bourens and Barrientos, 2017). *CMC1* was not significantly differentially expressed between recurrent and non-recurrent tumours, albeit a borderline non-significant decrease trend was observed in recurrent ccRCC tumours. Decreased complex IV assembly may indicate the metabolic shift from mitochondrial oxidative phosphorylation to cytoplasmic aerobic glycolysis. Low *CMC1* expression in ccRCC patients significantly correlates with worse patient prognosis (Figure 4.23A - B). Data showed that in non-recurrent ccRCC tumours, significantly higher levels and proportions of ENST00000468330 and

ENST00000495428 were expressed compared to recurrent ccRCC tumours, where ENST00000423894 and ENST00000466830 were the predominant isoforms of *CMC1*. All four transcripts display different coding sequences and express unique proteins when translated, which may confer differential functional properties. It is also important to note that the DTU of *CMC1* may result from the differential levels and cell types of infiltrating cells in ccRCC tumours.

Through closer inspection into ENST00000423894 and ENST00000466830, it was clear that minimap2 prioritise exon structure as the definitive 'trait' of a transcript over 3'UTR structures. For short transcripts like *CMC1*, most reads span the entirety of the transcript, so there was no need to 'split ties' to assign the closest resembling reference transcript. However, for longer transcripts, such as *CD24,* where many reads do not represent the full-length transcript, reads often match with multiple reference transcripts, and aligners may rely on read coverage percentage for transcript assignment probabilistically.

One of the recurrent findings of long-read RNA sequencing experiments is the high number of novel transcripts compared to the reference gene annotation. Using nanopore PCR-cDNA sequencing of human tissue samples, a recent study showed that 77% of mapped transcripts (93,718) were characterised as novel transcripts by FLAIR (Glinos *et al.*, 2022). Here, akin to the findings from the study, 65% and 74% of StringTie2 and FLAIR assembled transcripts from PCS of ccRCC tumours were identified as novel transcripts (Figure 4.24, Table 4.3). Only PCS data was analysed here using FLAIR and StringTie2 due to the higher sequencing depth, longer reads and higher read accuracies compared to DRS data. Nevertheless, it will be of interest to systematically identify and characterise novel transcripts from ccRCC tumour DRS data.

FLAIR is a reference-guided transcriptome assembler that first aligned reads to the reference genome using minimap2, followed by splice junctions correction by reference gene annotation, and finally, identification of high-confident transcripts. The largest group of novel transcripts found here was classified as 'j' under the gffcompare class code. This represents 'multi-exon isoforms with at least one splice junction match', which can also

be interpreted as a novel splice variant that shares at least part of an existing annotated transcript's exon/intron structure. Unfortunately, whilst nanopore sequencing long reads can span the complete length of many transcript isoforms, it also inherently shows a 3' bias due to RNA degradation, read stalling and pore blocking (Soneson *et al.*, 2019). This presents a potential danger in overestimating the number of novel splice variants from assembled transcriptome data by FLAIR and StringTie2.

FLAIR analysis uses human CAGE-seq data from the FANTOM5 consortium as a guide to define potential 5' start sites globally to mitigate the over-estimation problem. CAGE (cap analysis gene expression) is a method that allows mapping of the 5' ends of capped RNA via cap biotinylation and streptavidin-based pulldown (Takahashi *et al.*, 2012). However, although most translated eukaryotic mRNAs (90%) have 5' caps, uncapped transcripts can also be translated via internal ribosome entry sites (Shatsky *et al.*, 2018). Interestingly, a recent study has shown that through APA at proximal polyadenylation sites, the downstream uncapped mRNA transcripts extending to the distal polyadenylation site can be translated into functional proteins (Malka *et al.*, 2022). Thus, using CAGE-seq 5' annotation as a filter may discard valuable information.

StringTie2 employs a different route to define high-confident isoforms. StringTie2 assembles sequencing reads into super-reads by extending sequencing reads in both directions with overlapping reads and their unique coverage. Subsequently, splice graphs are constructed based on the read coverage levels within each super-read. With guidance from reference gene annotation, StringTie2 identifies transcript structures that best explain the maximum read coverage within each super-read until all reads have been assigned (Kovaka *et al.*, 2019). Compared to FLAIR, StringTie2 identified a substantially lower number of transcripts from PCS of ccRCC tumours (Table 4.3).

Focusing on immune checkpoints that play crucial roles in tumour immunity, novel transcript isoforms of *CTLA4*, *PD-L1*, *IDO1* and *CD24* were discovered by FLAIR and StringTie2. In particular, the identified novel transcripts display differential 3'UTR structure compared to existing reference gene annotation. 3'UTR is vital in regulating

mRNA stability, localisation, and translation. An $(AT)_n$ microsatellite polymorphism in the *CTLA4* gene, located in the 3'UTR region of the full-length *CTLA4* transcript (ENST00000648405), has been shown as a genetic marker for the susceptibility of type 1 diabetes (De Jong *et al.*, 2016). The length of $(AT)_n$ repeat inversely correlates with *CTLA4* transcript stability and protein expression levels (Malquori *et al.*, 2008). Truncation in the 3'UTR of membrane *PD-L1* mRNA (ENST00000381577) is a widespread structural variation across cancer types. A landmark study has shown that the shortening of 3'UTR in *PD-L1* transcripts leads to enhanced stability and elevated PD-L1 protein expression *in vivo* (Kataoka *et al.*, 2016). These studies demonstrate the importance of establishing the biological roles that these novel 3'UTRs may play in regulating the expression of immune checkpoints.

Both FLAIR and StringTie2 in novel isoform identification were able to identify the presence of novel isoforms in the inspected immune checkpoints. For *IDO1*, short 3'UTR was previously annotated as a feature exclusive to an exon 5 skipping NMD transcript (Figure 4.30A). However, long reads from DRS and PCS showed that mRNAs with both long and short 3'UTR could display exon 5 skipping events (Figure 4.30F). The transcript isoforms were accurately identified by FLAIR (Figure 4.30D). Accurate reference gene annotation is crucial for transcriptomic analysis. For Illumina data, with the current reference gene annotation, all reads that span the final 100nt of *IDO1 3'UTR* would be aligned to the protein-coding full-length *IDO1* transcript (ENST00000253237). Many of these mRNA transcripts might represent NMD transcripts which would not result in IDO1 protein expression. This finding demonstrated the power of long-read sequencing and transcript assembly to detect novel transcripts and infer isoform identity accurately.

In some cases, FLAIR and StringTie2 were found to 'overcorrect the assembled transcripts to the adjacent annotated splice junctions. For *CTLA4*, FLAIR correctly identified a novel isoform with the exon structure of *CTLA4* but with a shorter 3'UTR, as supported by evidence from raw reads. However, FLAIR adopted the 3'UTR end of the short *CTLA4* for the novel isoform, with the 3'UTR end 100nt 5' upstream of where raw

reads mapping ends (Figure 4.27). Similarly, in *PD-L1*, a novel isoform was identified to encode for soluble *PD-L1* protein but with a longer 3'UTR than the current annotation. StringTie2 was able to identify the 3' end of the novel soluble *PD-L1* transcript. However, the exon structure of the novel transcript was adopted from ENST00000474218, which was unsubstantiated by raw reads evidence. As a method, StringTie2 is designed to prioritise using the minimum number of assembled transcripts to explain the maximum number of reads. This design feature was reflected in all immune checkpoints examined here, where StringTie2 assembled transcripts tend to adopt the maximum 3'UTR lengths to capture all reads. These results suggest that FLAIR and StringTie2 can indicate the existence of novel transcripts, but accurate annotation still requires mannual inspection of raw reads and reference genome alignments. Instead of using reference genome-guided assemblers, de novo transcriptome assembly is a viable option to avoid overcorrections from reference gene annotation. However, unlike FLAIR and StringTie2, which were optimised to permit transcript reconstruction using long reads with relatively high error rates, established *de novo* assemblers such as Trinity, Oases and rnaSPAdes were developed based on low-error short reads from Illumina sequencing (Schulz *et al.*, 2012; Haas *et al.*, 2013; Bushmanova *et al.*, 2019). Currently, one solution for using a *de novo* transcriptome assembler on long-read sequencing data is by sequencing RNA with both DRS and Illumina sequencing to 'correct' the high error rates in DRS reads (Fu *et al.*, 2018). No benchmarking report has been published to examine the performance of *de novo* assemblers on PCS data, where reads have a lower error rate than DRS. A systematic comparison of the performance between reference-guided assemblies demonstrated here with reference-free transcriptome assembly tools using DRS and PCS data would be highly valuable to the broader research community.

Next, using nanopolish on DRS data, global poly(A) tail profiles from ccRCC tumours were produced. This is the first global poly(A) tail profiling of archival tumour samples. Here, the poly(A) tails were found to be highly variable within each tumour sample, but the global profiles (median lengths: 56.7 – 88.7 nt) from the tumour samples were

consistent with published TAILseq, PAIseq and DRS nanopolish analysis (Chang *et al.*, 2014; Workman *et al.*, 2019; Eisen *et al.*, 2020). The intra- and inter-sample variabilities present a technical challenge in comparing poly(A) tail lengths between samples. Nested T-tests were used where possible to account for the intra-sample variations (Krzywinski *et al.*, 2014). However, no published method was found to normalise the impact of RNA degradation on global poly(A) tail length. A benchmarking study to correlate RNA sample RIN number with poly(A) tail profile across different profiling methods would be extremely valuable.

The poly(A) tail profiles of recurrent and non-recurrent tumours showed no significant difference. Interestingly, nanopolish results showed that protein-coding genes have, on average shorter poly(A) tails than lncRNA and most pseudogenes (Figure 4.33). A previous poly(A) tail profiling study using DRS and nanopolish showed no significant differences in the length of poly(A) tail lengths between protein-coding genes, lncRNA and pseudogenes in the human HAP1 cell line (Soneson *et al.*, 2019). Here, instead of a homogenous cell population, ccRCC tumours represent a complex mix of cells which may display differential poly(A) tail profile.

Recent studies have also shown that global poly(A) tail lengths can vary dramatically depending on environmental cues. A study found that upon lipopolysaccharides (LPS) exposure, activation of human macrophages (THP-1 cell line) induced rapid cytoplasmic re-adenylation and poly(A) tail lengthening in 1,500 genes within one hour (Kwak *et al.*, 2022). With the complex, heterogeneous composition of cytokines, chemokines and growth factors in the TME, mRNA poly(A) tail lengths in ccRCC tumours are likely to be regulated in a dynamic and cell-type dependent manner.

For the immune checkpoints IDO1, CD24 and PD-L1, different transcript isoforms of the same gene had significantly different poly(A) tail lengths (Figure 4.34). Poly(A) tail length was found to negatively correlates with gene expression levels (Lima *et al.*, 2017). Poly(A) tail lengths also did not correlate with the rate of transcription or mRNA stability (Tudek *et al.*, 2021). Recent works have demonstrated that mRNAs with the same poly(A) tail

length but from different genes can exhibit different decay rates up to 1000 folds (Subtelny *et al.*, 2014; Eisen *et al.*, 2020). Other mRNA features, like mRNA 5' cap and *cis*-elements (such as cytoplasmic polyadenylation element (CPE)), may act synergistically as contributing factors to determine mRNA stability. Though the precise biological role of differential poly(A) tail lengths is yet to be determined, data here showed interesting co-dependencies between splicing and poly(A) tail length regulation in immune checkpoint transcripts.

4.5 Evaluation of key objectives

- **Characterisation of ccRCC tumour gene expression profiles using unsupervised methods**

  PCA and hierarchical clustering methods were unable to cluster ccRCC tumours based on patients' clinical characteristics. This may be due to the limited number of samples and the size of the phenotype-associated effect on gene expression.

- **Identification of ccRCC recurrence-associated differentially expressed genes and differential transcript usage events**

  DRS and PCS data identified dozens of significant DEGs and DTU events between recurrent and non-recurrent ccRCC tumours. Large number of significant DEGs were identified to be immune cell related genes.

- **Detection of activated and suppressed pathways in recurrent ccRCC tumours**

  GSEA revealed suppression of adaptive immune response and antigen presentation pathways and activated lipid metabolism pathways in recurrent ccRCC tumours. These pathways are highly linked to ccRCC aggressiveness and patient outcome.

- **Comparisons of tumour microenvironment between recurrent and non-recurrent ccRCC tumours**

  Cell type deconvolution analysis showed a significant reduction in the number of tumour infiltrating immune cells as well as the proportion of $CD8^+$ T cells in recurrent ccRCC tumours. A sub-group of non-recurrent ccRCC tumours was identified to have an elevated level of $CD8^+$ T cells with high expression levels of exhaustion markers.

- **Identification of novel transcript isoforms in ccRCC tumours**

  Using StingTie2 and FLAIR, 65 - 74% of assembled transcripts from ccRCC tumours were classified as novel transcripts. Both methods discovered novel transcript isoforms for clinically important immune checkpoints (such as *PD-L1*). However, accurate annotation still requires confirmation by visual inspection of the read-alignment tracks.

- **Estimation of mRNA poly(A) tail lengths in ccRCC tumours**

  mRNA poly(A) tail lengths from tumour DRS data were measured by nanopolish, with similar profiles to published works. Several immune checkpoint genes were found to display differential poly(A) tail lengths depending on transcript isoform, suggesting co-dependencies between splicing and poly(A) tail regulation.

4.6 Conclusion

This chapter systematically characterised the transcriptome profiles of ccRCC archival tumours. Using data from DRS and PCS of archival samples, ccRCC recurrence-associated differentially expressed genes and differential transcript usage were identified. GSEA revealed that recurrent ccRCC tumours showed suppressed immune cell pathways, decreased gene expression in antigen presentation pathways, and augmented fat metabolism. Cell-type deconvolution analysis showed significant decreases in both tumour-infiltrating immune cells and the proportion of $CD8^+$ T cells in recurrent ccRCC tumours. A subgroup of non-recurrent ccRCC tumours showed significantly high levels of $CD8^+$ T cells and high expression levels of exhausted $CD8^+$ T cell markers, such as *TOX* and *PDCD1* (PD-1). Tens of thousands of novel transcripts were discovered using reference-guided transcriptome assemblers, including transcripts from the immune checkpoint genes *CTLA4*, *PD-L1*, *IDO1* and *CD24*. Global transcript poly(A) tail profiles of the ccRCC tumours were characterised, and different isoforms of the same immune checkpoint genes were found to display differential poly(A) tail lengths. Results here demonstrated the ability of ONT LRS technologies in providing in-depth characterisation of transcriptomic profiles from archival tumour samples.

Chapter 5

Exploring the effects of m^6A machinery disruption on the transcriptome of kidney cancer cells

5.1 Introduction

Since the recent technological breakthroughs in transcriptome-wide m^{6}A detection methods, m^{6}A mRNA modification has been recognised as an essential layer of the post-transcriptional gene expression regulatory network (Chapter 1.1.6). For mRNA molecules, m^{6}A is deposited through the m^{6}A writer complex, comprised of multiple subunits such as METTL3, METTL14 and WTAP. mRNA m^{6}A can also be dynamically removed by the erasers: ALKBH5 and FTO. Furthermore, by recruiting RBPs with a specific affinity towards m^{6}A-modified transcripts, known as the m^{6}A readers, the modification has been shown to exert regulatory effects on mRNA splicing, stability, localisation, and translation (Murakami and Jaffrey, 2022).

Dysregulation in m^{6}A regulation is associated with tumourigenesis and tumour progression in ccRCC. Transcriptomic data analyses from the TCGA KIRC and other independent cohorts showed that almost all m^{6}A regulators were significantly differentially expressed in ccRCC tumours compared to adjacent normal tissues (Chen *et al.*, 2021). A recent study showed that the protein expression levels of m^{6}A writer METTL14 by IHC were significantly suppressed in ccRCC tumours compared to paired adjacent normal tissues. Consequently, the global m^{6}A levels of ccRCC tumours were also significantly decreased. Reduced METTL14 protein expression was associated with poorer ccRCC prognosis (Shen *et al.*, 2022).

In contrast, protein expression levels of the m^{6}A eraser FTO by IHC were significantly upregulated in ccRCC tumours compared to normal adjacent tissues. Furthermore, inhibition of *FTO* via CRISPR-Cas9-mediated gene deletion and siRNA-mediated gene silencing in *VHL*-deficient ccRCC cell lines significantly reduced proliferation and colony formation (Xiao *et al.*, 2020). With the increasing number of studies attributing both oncogenic and tumour-suppressive roles to m^{6}A regulators, their biological functions are now widely recognised as cell-type and tumour-type dependent (Gao *et al.*, 2021). However, the role of m^{6}A in regulating ccRCC gene expression is still poorly understood.

One of the defining features of ccRCC is the characteristically complex TME. Within the TME, ccRCC tumour cells constantly interact with various types of immune and stromal cells and soluble factors like cytokines. As the critical mediators of inter-cellular communication within the TME, the immunomodulatory cytokines are responsible for orchestrating immune responses in the TME, with far-reaching impact on the gene expression profiles of both immune and tumour cells (Kartikasari *et al.*, 2021).

IFNγ and TNF are two of the major cytokines found in the TME. IFNγ is primarily secreted by activated $CD4^+$ Th1 cells, $CD8^+$ T cells and NK cells, whereas TNF is mainly secreted by activated macrophages and, to a lesser degree NK cells and T cells (Castro *et al.*, 2018; Laha *et al.*, 2021). IFNγ activates the JAK (Janus kinase)/STAT1 (Signal transducer and activator of transcription 1) pathway in cancer cells upon binding to interferon-gamma receptors on the cell surface. Exposure to a high IFNγ dosage triggers apoptosis in cancer cells by activating the downstream JAK-STAT1-dependent caspase activities. (Owen *et al.*, 2019). The binding of TNF with the cell surface receptor TNFR1 on tumour cells activates the downstream apoptotic pathway through TNFR1-associated death domain protein and caspase 8, leading to cell death (Josephs *et al.*, 2018).

In addition to their cytotoxic effects, IFNγ and TNF can play an anti-tumour role by modifying the tumour immune cell population. As a positive feedback mechanism, IFNγ promotes the differentiation of naïve $CD4^+$ T cells into Th1 cells, amplifying the release of IFNγ (Castro *et al.*, 2018). Exposure to IFNγ has also been shown to enhance cytotoxic T cells' cytotoxicity and cell-killing ability (Bhat *et al.*, 2017). Both IFNγ and TNF polarise the differentiation of macrophages towards the anti-tumour M1 state, which also serves as a significant source of TNF secretion (Laha *et al.*, 2021).

IFNγ and TNF can also enhance the gene expression of MHC molecules and other components of the antigen presentation pathway, such as *TAP* (Transporter associated with antigen processing). Upregulation of the antigen presentation pathway in tumour cells results in enhanced T cell recognition and killing (S. Zhang *et al.*, 2019). However, IFNγ and TNF can also exert pro-tumour effects. IFNγ and TNF synergistically induce

expression of the immune checkpoints PD-L1 and IDO1, negatively impacting tumour immune response(Robinson *et al.*, 2003; Li *et al.*, 2018). IDO1 is a cytosolic enzyme which catalyses the conversion of tryptophan to kynurenine. IDO1-mediated depletion of tryptophan in the TME inhibits proliferation and the effector functions of tumour-infiltrating immune cells. In addition, the downstream kynurenine metabolites induce the differentiation of the immunosuppressive regulatory T cells (Hornyák *et al.*, 2018). The IFNγ and TNF-induced expression of PD-L1 and IDO1 represent a negative feedback regulation in preventing excessive inflammation. However, it also promotes immune escape, leading to tumour progression.

Aside from regulating immune gene expression in tumour cells, combinatorial treatment of IFNγ and TNF was also found to promote tumourigenesis. Long-term IFNγ and TNF treatment upregulates the expression of *c-Fos* and *c-Myc* in mesenchymal stem cells, resulting in an increased rate of malignant transformation (Wang et al., 2013). For tumour cells, exposure to low IFNγ dosage by NSCLC tumour cells leads to activation of PI3K-Akt and Notch1 pathways, resulting in enhanced tumour stemness and proliferative abilities (Song *et al.*, 2019). Similarly, in breast cancer, TNF treatment can promote tumour cell proliferation and stem-cell-like phenotype via activation of the NF-kB pathway, contributing to tumorigenesis and tumour progression (Liu *et al.*, 2020). It is currently unclear if, and in what context, IFNγ and TNF may play synergistic roles in promoting tumour growth.

Overall, the effects of IFNγ and TNF on tumour progression are complex and involve both direct and indirect mechanisms. Further research is needed to fully understand the role of IFNγ and TNF in regulating ccRCC tumour progression. Although the interaction between tumour cells with IFNγ and TNF has tremendous effects on tumour immunity, the role of m^6A in modulating cytokines-induced gene expression changes, particularly in ccRCC tumour cells, is yet to be elucidated.

5.2 Chapter aims

The main aims of this chapter are to **understand the role of cytokines (IFNγ and TNF) and mRNA m^{6}A modifications in regulating ccRCC gene expression**. The specific aims of this chapter are as follows:

i) Analysis of the potential role of m^{6}A on ccRCC gene expression using publicly available genomic data, including somatic mutation and copy number variations analysis.

ii) Generation of CRISPR-Cas9-mediated m^{6}A writers KO ccRCC cell lines

iii) Characterising the gene expression profiles of parental cell line and m^{6}A writer-KO cell lines, both at basal, unstimulated cells and IFNγ and TNF stimulated cells.

iv) Profile activated and suppressed pathways associated with m^{6}A writer-KO and IFNγ + TNF combinatorial treatment.

v) Identify genes that display differential transcript usage after IFNγ and TNF treatment

vi) Characterise the effects of m^{6}A on mRNA and protein expression of PD-L1, using both CRISPR-Cas9-mediated gene deletion and siRNA-mediated gene silencing approaches.

5.3 Results

5.3.1 Frequent genomic copy number variations of m^{6}A regulators genes in ccRCC

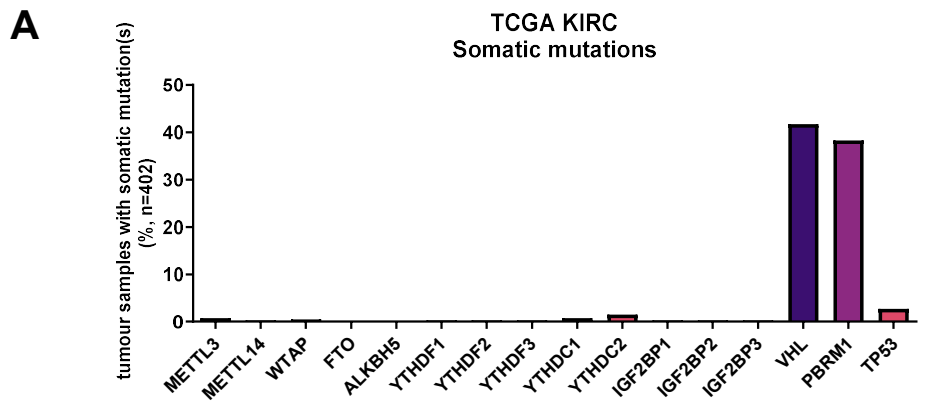
To understand the potential role of m^{6}A in ccRCC tumours, the genetic alterations of key m^{6}A regulators (writer, erasers and readers) in ccRCC patients' tumour samples were surveyed using the TCGA KIRC cohort genomic data (Weinstein *et al.*, 2013). KIRC somatic mutation rate and copy-number analysis results were extracted from cBioPortal. Copy number level for each gene in each sample was defined by one of the following five categories: i) deep deletion, which indicates possible homozygous gene deletion; ii) shallow deletion, which indicates potential heterozygous gene deletion; iii) diploid; iv) low-level gain, representing genes with few additional copies across the genome; and v) high-level amplification, which indicates a high number of extra gene copies focally (Cerami *et al.*, 2012).

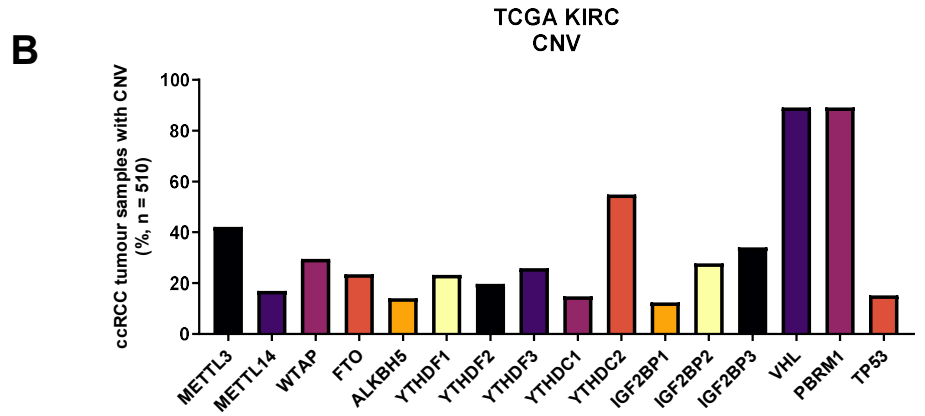
In ccRCCs, the most frequently mutated genes are *VHL* and *PBRM1*, at 41.79% and 38.31% of all ccRCC tumours with somatic mutation data in the KIRC cohort (n = 402). In contrast, m^{6}A regulators were found to be infrequently mutated. Amongst all of the m^{6}A regulators that were surveyed, *YTHDC2* was mutated at the highest frequencies in ccRCC tumours at 1.49% (Figure 5.1A).

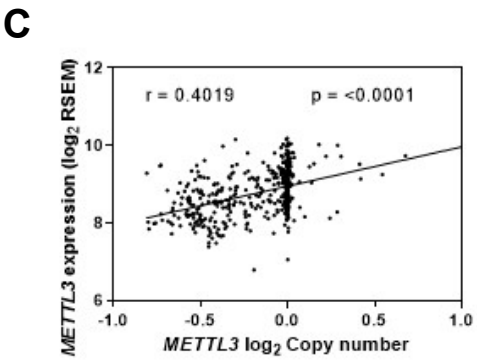
Copy number analysis demonstrated that most ccRCC tumours experience copy number variation (CNV) with *VHL* and *PBRM1* (89.15% of all samples for both genes). Compared to the rate of somatic mutations, CNVs of m^{6}A regulators were much more prevalent. *METTL3* and *WTAP* were found with CNV In 42.21% and 29.58% of ccRCC tumours with copy number data in the KIRC cohort (n = 507) (Figure 5.1B).

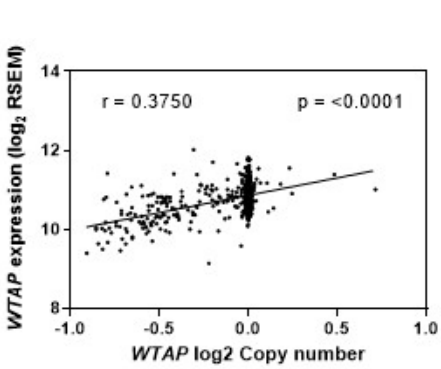
Next, the relationship between gene copy numbers and gene expression levels was evaluated. Statistically significant positive correlations were observed between the gene expression levels (Log_2RSEM (RNA-Seq by Expectation-Maximisation) normalised counts) and gene copy number (Log_2CopyNumber) of *METTL3* and *WTAP* in TCGA KIRC cohort ccRCC tumour samples (Figure 5.1C – D).

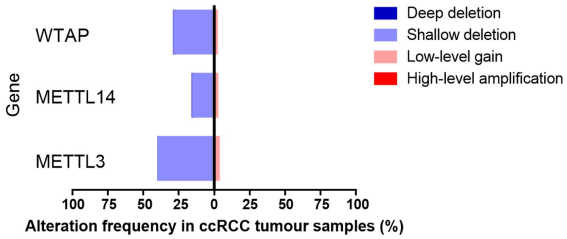
Interestingly, for m^6A writers, almost all CNVs were found to be shallow deletions in the ccRCC tumours (Figure 5.1E). On the other hand, for m^6A erasers, CNVs of *FTO* were mostly low-level gain, whilst more CNVs of *ALKBH5* were deleterious rather than gain (Figure 5.1F). Heterogeneous patterns were observed in the types of CNV in m^6A reader genes. CNVs of the most highly altered m^6A regulator, *YTHDC2*, were almost exclusively gain of gene copies. *YTHDF1* and *IGF2BP3* genes were also found with mostly low-level gains, whereas CNVs of *YTHDF2* in ccRCC tumours were mostly shallow deletions (Figure 5.1G). Data here suggest that the gene copy number and gene expression of m^6A regulators are frequently altered in ccRCC tumours. Moreover, CVN of the core components of m^6A writers *METTL3*, *METTL14* and *WTAP* were near-exclusively heterozygous gene deletions.

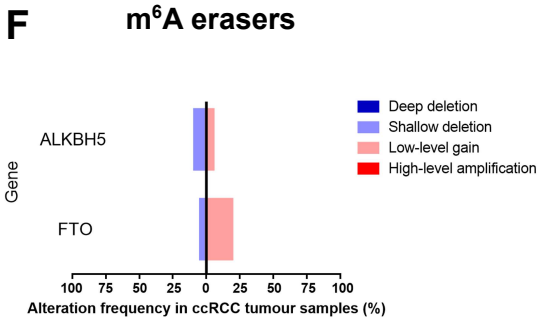

A
TCGA KIRC
Somatic mutations
tumour samples with somatic mutation(s) (%, n=402)
50
40
30
20
10
0
METTL3
METTL14
WTAP
FTO
ALKBH5
YTHDF1
YTHDF2
YTHDF3
YTHDC1
YTHDC2
IGF2BP1
IGF2BP2
IGF2BP3
VHL
PBRM1
TP53


B
TCGA KIRC
CNV
ccRCC tumour samples with CNV (%, n = 510)
100
80
60
40
20
0
METTL3
METTL14
WTAP
FTO
ALKBH5
YTHDF1
YTHDF2
YTHDF3
YTHDC1
YTHDC2
IGF2BP1
IGF2BP2
IGF2BP3
VHL
PBRM1
TP53


C
r = 0.4019
p = <0.0001
METTL3 expression (log2 RSEM)
METTL3 log2 Copy number


D
r = 0.3750
p = <0.0001
WTAP expression (log2 RSEM)
WTAP log2 Copy number


E
m6A writers
WTAP
METTL14
METTL3
Gene
Deep deletion
Shallow deletion
Low-level gain
High-level amplification
Alteration frequency in ccRCC tumour samples (%)


F
m6A erasers
ALKBH5
FTO
Gene
Deep deletion
Shallow deletion
Low-level gain
High-level amplification
Alteration frequency in ccRCC tumour samples (%)

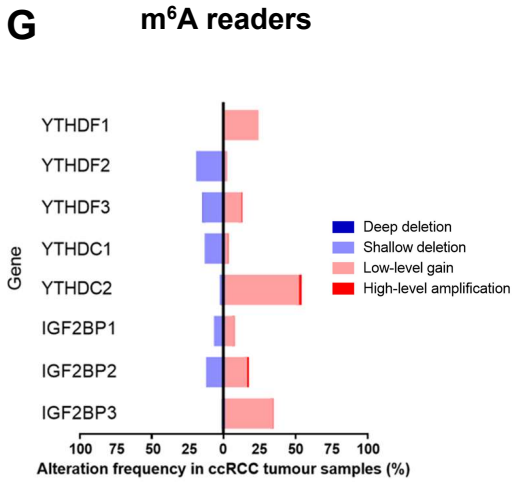
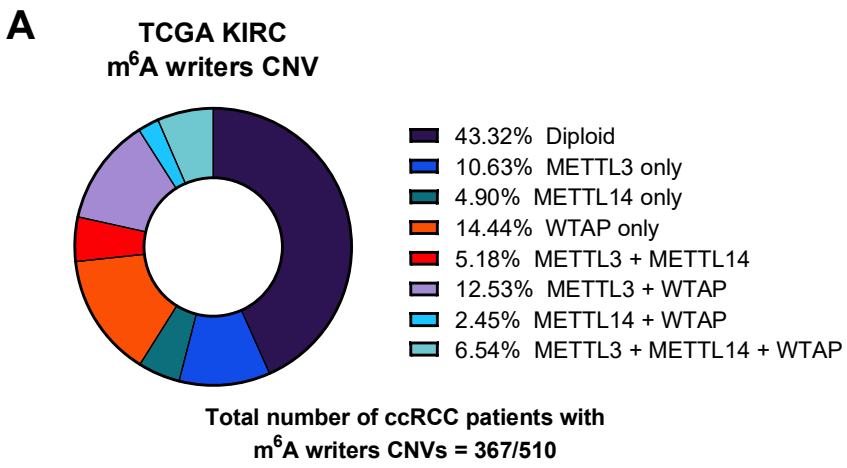

G
m6A readers
YTHDF1
YTHDF2
YTHDF3
YTHDC1
YTHDC2
IGF2BP1
IGF2BP2
IGF2BP3
Gene
Deep deletion
Shallow deletion
Low-level gain
High-level amplification
Alteration frequency in ccRCC tumour samples (%)

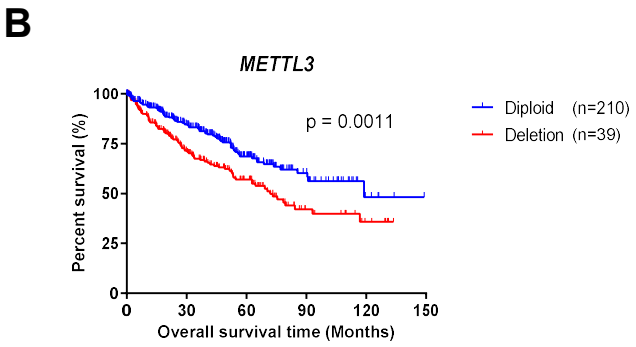
Figure 5.1: Genetic alterations of m^{6}A regulators in ccRCC tumours

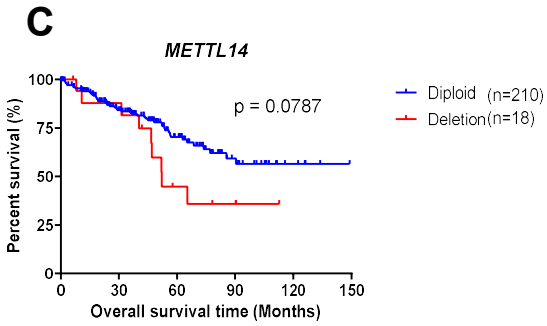
A) Bar graph showing the frequency of somatic mutations in tumours samples from TCGA KIRC cohort (n = 402). **B**) Bar graph showing the frequency of copy number variations (CNV) in ccRCC tumour samples from TCGA KIRC cohort (n = 510). **C**) Correlation between *METTL3* gene expression (Log_2RSEM (RNA-Seq by Expectation-Maximisation) normalised counts) and gene copy number (Log_2CopyNumber) in TCGA KIRC cohort ccRCC tumour samples (n = 510). **D**) As in **C**, but for *WTAP.* **E**) Stacked bar graph showing the frequency and types of CNV in m^{6}A writer genes in ccRCC tumour samples from TCGA KIRC cohort (n = 510). **F**) As in **E**, but for m^{6}A eraser genes. **G**) As in **E**, but for m^{6}A readers genes. For **C – D**, Diagonal line represents the line of best fit. Spearman's rank correlation coefficients were calculated to measure the strength and direction of the correlations. P values were generated from F-test, with $p \leq 0.05$ considered statistically significant.

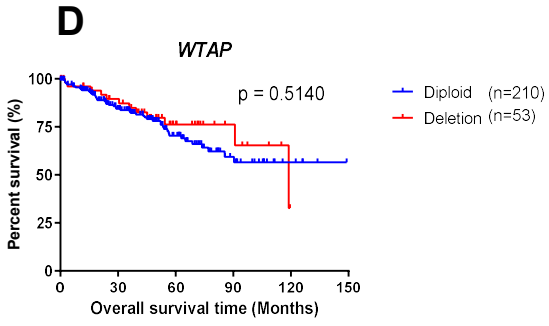
5.3.2 Copy number deletion of m^{6}A writers negatively correlates with overall survival of ccRCC patients

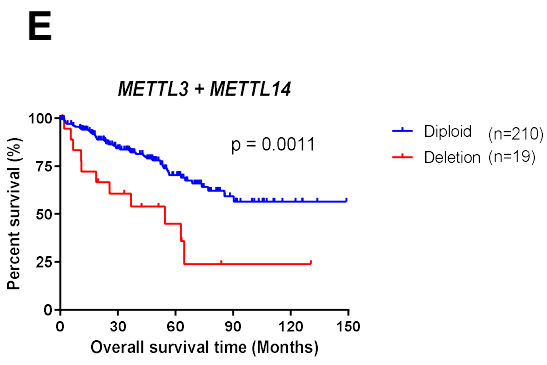
Next, survival analyses were conducted to understand the role of m6A writers, and in extension m^{6}A, on ccRCC patients' prognosis. Firstly, m^{6}A writers' CNV events, which were almost all deleterious in ccRCC tumours, were not found to be mutually exclusive. Only 42.32% of all ccRCC tumours in the KIRC cohort harboured no CNVs in any of the m^{6}A writers. In addition, in many ccRCC patients, more than one writer was found with CNVs, with 6.54% (24 / 510) of tumours having CNVs in all three writers (Figure 5.2A). Survival analysis showed that the deletion of *METTL3* significantly correlated with worse overall survival in ccRCC patients (p = 0.0011) (Figure 5.2B). A near-significant trend was identified between *METTL14* deletion and worse overall survival (p = 0.0787) (Figure 5.2C). For *WTAP*, no significant correlation was found (Figure 5.2D). Lastly, copy number losses of multiple m^{6}A writers (*METTL3* + *METTL14*, *METTL3* + *METTL14* + *WTAP*) in ccRCC patients were significantly associated with worse overall survival (Figure 5.2E – F). The results suggest a significant association between genetic alterations in m^{6}A writer genes and poor prognosis.


A
TCGA KIRC
m6A writers CNV
43.32% Diploid
10.63% METTL3 only
4.90% METTL14 only
14.44% WTAP only
5.18% METTL3 + METTL14
12.53% METTL3 + WTAP
2.45% METTL14 + WTAP
6.54% METTL3 + METTL14 + WTAP
Total number of ccRCC patients with
m6A writers CNVs = 367/510


B
METTL3
p = 0.0011
Diploid (n=210)
Deletion (n=39)
Percent survival (%)
Overall survival time (Months)


C
METTL14
p = 0.0787
Diploid (n=210)
Deletion(n=18)
Percent survival (%)
Overall survival time (Months)


D
WTAP
p = 0.5140
Diploid (n=210)
Deletion (n=53)
Percent survival (%)
Overall survival time (Months)


E
METTL3 + METTL14
p = 0.0011
Diploid (n=210)
Deletion (n=19)
Percent survival (%)
Overall survival time (Months)

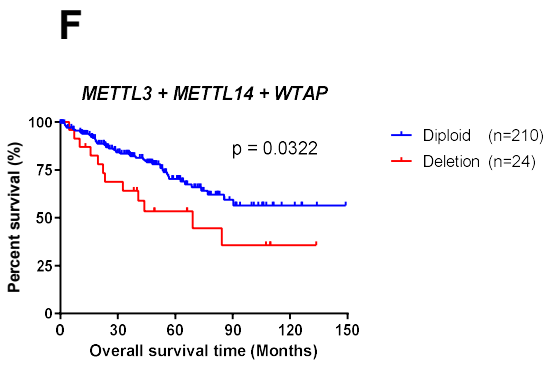

F
METTL3 + METTL14 + WTAP
p = 0.0322
Diploid (n=210)
Deletion (n=24)
Percent survival (%)
Overall survival time (Months)

Figure 5.2: Copy number loss of m^{6}A writers confer unfavourable survival outcomes in ccRCC

A) Pie chart showing the types and distribution of m^{6}A writers' CNVs in ccRCC tumour samples from TCGA KIRC cohort (n = 510). Kaplan-Meier survival plots comparing overall survival time of TCGA KIRC cohort ccRCC patients with diploid, unaltered m^{6}A writer genes (n = 210) and patients with copy number deletions in **B)** *METTL3* exclusively (n = 39), **C)** *METTL14* exclusively (n=18), **D)** *WTAP* exclusively (n=53), **E)** both *METTL3* and *METTL14* (n=19), and **F)** *METTL3, METTL14* and *WTAP* (n= 24). Differences in survival rates were assessed using Log-rank (Mantel-Cox) test with $p \leq 0.05$ considered statistically significant.

5.3.3 Generation of CRISPR-Cas9-mediated genetic knock out of m^{6}A writers in the ccRCC cell line RCC4

As an essential subunit of the m^{6}A complex, depletion of *METTL3* and *WTAP* via CRISPR-Cas9 mediated gene knockout and siRNA-mediate gene knockdown was previously shown to decrease mRNA m^{6}A levels globally in mammalian cells (Ping *et al.*, 2014; Ge *et al.*, 2021). Therefore, to understand the role of m^{6}A in regulating the gene expression landscape of ccRCC tumour cells, the m^{6}A writers *METTL3* and *WTAP* were targeted for CRISPR-Cas9-mediated gene knockout, with the aim of further characterisation by DRS analysis.

Firstly, RCC4 cells were stably transduced with lentiviruses containing an eGFP-tagged Cas9 expression construct. Then, successfully transduced, GFP-positive RCC4 cells were isolated by FACS using the MoFlo Astrios EQ cell sorter. The purity of the GFP-positive population was confirmed by flow cytometry, where 96.5% of the cell population expressed GFP (Figure 5.3A – D). Finally, the expression of Cas9 protein in the RCC4 Cas9 GFP cells was validated via western blotting (Figure 5.3E, G).

Single-cell derived clonal *METTL3*-KO and *WTAP*-KO cell lines were isolated by transfection of guide RNAs followed by limiting dilution. Western blot analysis demonstrated the efficiency of guide RNAs, with decreased protein expression of METTL3 and WTAP in guide RNA transfected cells (48 hours post-transfection) compared to mock-transfected RCC4 Cas9 GFP cells (Figure 5.3E, 5.3G). Finally, gene KO in single-cell derived clonal *METTL3* KO (Figure 5.3F) and *WTAP* KO (Figure 5.3H) cell lines were validated via western blot analysis. Isolated *METTL3* KO lines were not viable after passaging. In contrast, *WTAP* KO lines were viable, with clone 2H1 (WTAP KO 2H1) expanded for subsequent DRS analysis. Finally, RNA m^{6}A dot blot analysis demonstrated downregulated global RNA m^{6}A levels in extracted total RNA from WTAP KO cell lines (2E6 and 2H1) compared to the parental RCC4 Cas9 GFP cells (Figure 5.3I).

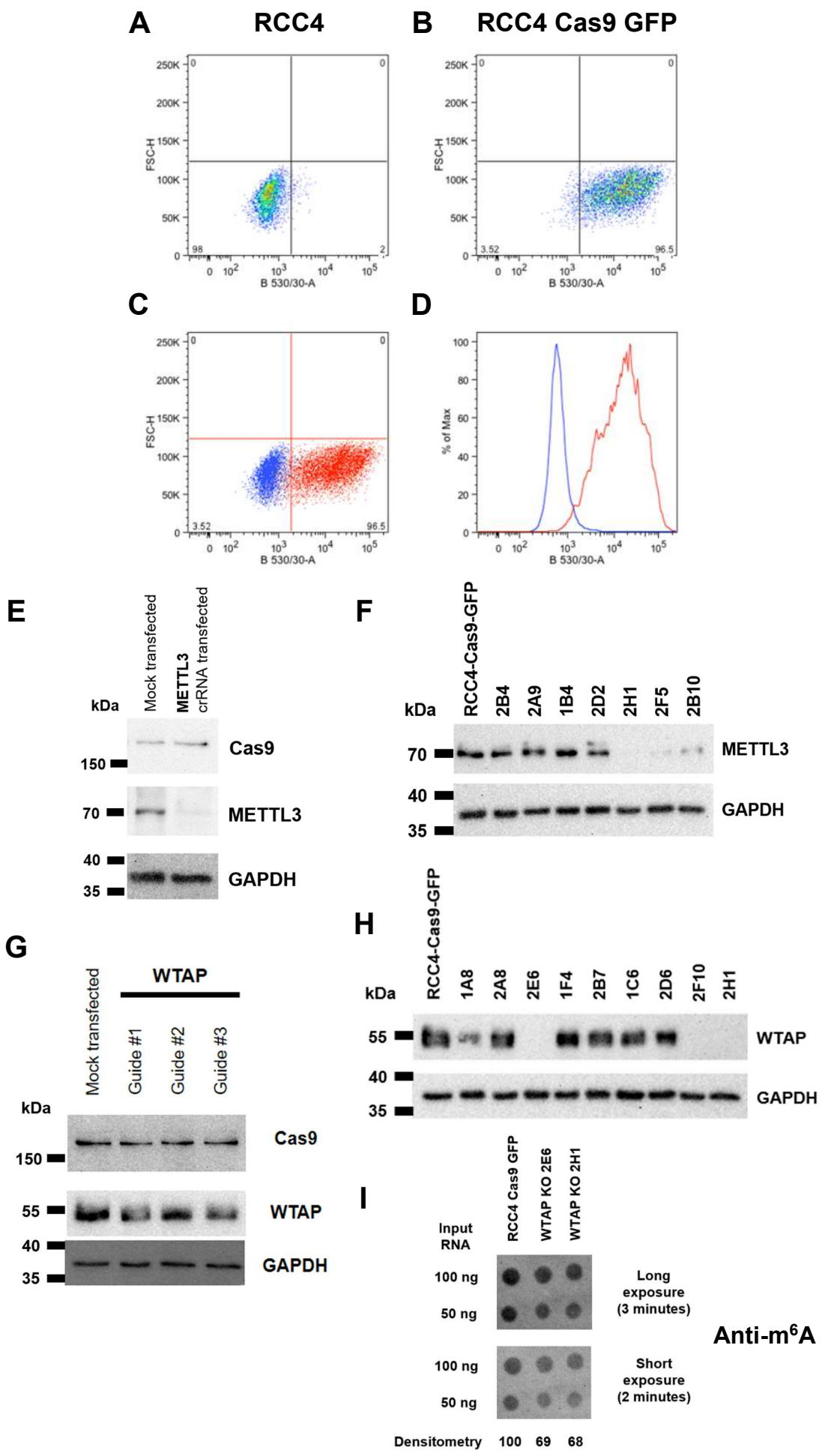

A
RCC4
B
RCC4 Cas9 GFP
FSC-H
B 530/30-A
98
2
3.52
96.5
C
D
% of Max
E
Mock transfected
METTL3 crRNA transfected
kDa
150
Cas9
70
METTL3
40
35
GAPDH
F
RCC4-Cas9-GFP
2B4
2A9
1B4
2D2
2H1
2F5
2B10
G
WTAP
Guide #1
Guide #2
Guide #3
55
H
1A8
2A8
2E6
1F4
2B7
1C6
2D6
2F10
2H1
I
Input RNA
RCC4 Cas9 GFP
WTAP KO 2E6
WTAP KO 2H1
100 ng
50 ng
Long exposure (3 minutes)
Short exposure (2 minutes)
Anti-m6A
Densitometry
100
69
68

Figure 5.3: Generation of CRISPR-Cas9 mediated *METTL3* and *WTAP* KO ccRCC clonal cell lines

Flow cytometry dot plots of GFP expression (x-axis, B 530/30A) versus forward scatter-Height (y-axis, FSC-H) for **A)** RCC4 cells, **B)** RCC4 Cas9 GFP cells, and **C)** overlay of RCC4 cell population (blue) and RCC4 Cas9 GFP cell population (red). **D)** Representative flow cytometry histogram of GFP expression (x-axis, B 530/30A) across RCC4 (blue) and RCC4 Cas9 GFP cell population (red). **E)** Western blot analysis of Cas9 protein, METTL3 and GAPDH (loading control) in mock transfected and *METTL3* crRNA-pool transfected RCC4 Cas9 GFP cells at 48 hours after transfection. **F)** Western blot analysis of METTL3 and GAPDH (loading control) in RCC4 Cas9 GFP and METTL3-KO clonal cell lines. **G)** Western blot analysis of Cas9 protein, WTAP and GAPDH (loading control) in mock transfected and *WTAP* crRNA #1, #2 and #3 transfected RCC4 Cas9 GFP cells at 48 hours after transfection. **H)** Western blot analysis of WTAP and GAPDH (loading control) in RCC4-Cas9-GFP and WTAP KO clonal cell lines. **I)** RNA m^6A dot blot analysis of RCC4-Cas9-GFP, WTAP KO 2E6 and WTAP KO 2H1 total RNA (100 ng and 50 ng). RNA m^6A levels of WTAP KO cell lines were relative to RCC4 Cas9 GFP, quantified by densitometry analysis and averaged between 100ng and 50ng input RNA.

5.3.4 DRS of RCC4 Cas9 GFP and WTAP KO 2H1 cells

Using DRS, the transcriptomic profiles of RCC4 Cas9 GFP and WTAP KO 2H1 under both untreated and 24 hours IFNγ + TNF treatment conditions (n = 3 for each cell line at each condition) were elucidated. After the treatment time point, total RNA was extracted, with RNA quality assessed by Agilent 2100 Bioanalyzer. Bioanalyzer results demonstrated that the extracted RNA was not degraded, with a RIN score of 10 across all samples. The Bioanalyzer gel image showing the size distribution of RNA fragments from the sequenced samples is displayed in Figure 5.6. Poly(A)$^+$ RNA was isolated from total RNA using poly d(T) magnetic beads, with 500ng of poly(A)$^+$ RNA used as input for DRS library generation (RNA002). The DRS libraries were sequenced on a PromethION sequencer using PromethION flow cells (R9.4.1). The DRS library generation protocol, sequencing parameters and analysis pipelines used here were the same as in previous DRS of ccRCC archival tumour samples. A summary workflow for the experimental setup and analysis pipeline is shown in Figure 5.5.

After 72 hours of sequencing, DRS generated 2.2 – 6.4 million sequencing reads per sample that passed sequence quality control (read quality Q score above 7), with a median of 4.8 million passed reads per sequencing run (Figure 5.7A). Mean Q scores for DRS reads ranged between 11.0 and 12.0 (median: 11.3). Mean and median DRS passed-reads read length ranged between 1,149 – 1,313 nt and 928 – 1,036 nt, with median lengths of 1,245 nt and 997nt, respectively.

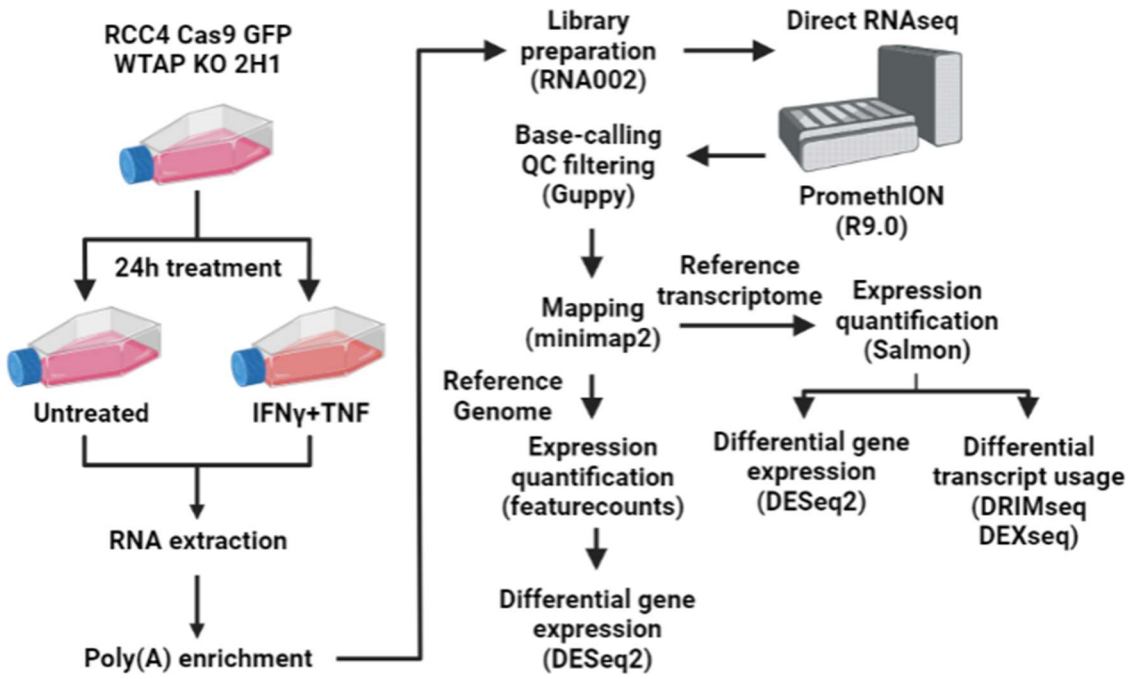


Figure 5.5: Summary workflow for DRS of RCC4 Cas9 GFP and WTAP KO 2H1

Total RNA was extracted from untreated and IFNγ + TNF treated (24 hours) RCC4 Cas9 GFP and WTAP KO 2H1 cells. Poly(A)$^+$ RNAs were enriched by poly d(T) beads. 500 ng of poly(A)$^+$ RNA were used per sample to prepare DRS library (RNA002). Libraries were loaded in PromethION R9 flow cells, and each sequencing run lasted 72 hours. Reads were base called concurrently by Guppy, where Q score > 7 were kept as passed reads. Reads were subsequently mapped to either reference genome or transcriptome via minimap2, using nanopore sequencing specific setting. Gene expression levels were determined by featurecounts and Salmon, followed by differential gene expression analysis by DESeq2, and differential transcript usage analysis performed by DRIMseq and DEXseq.

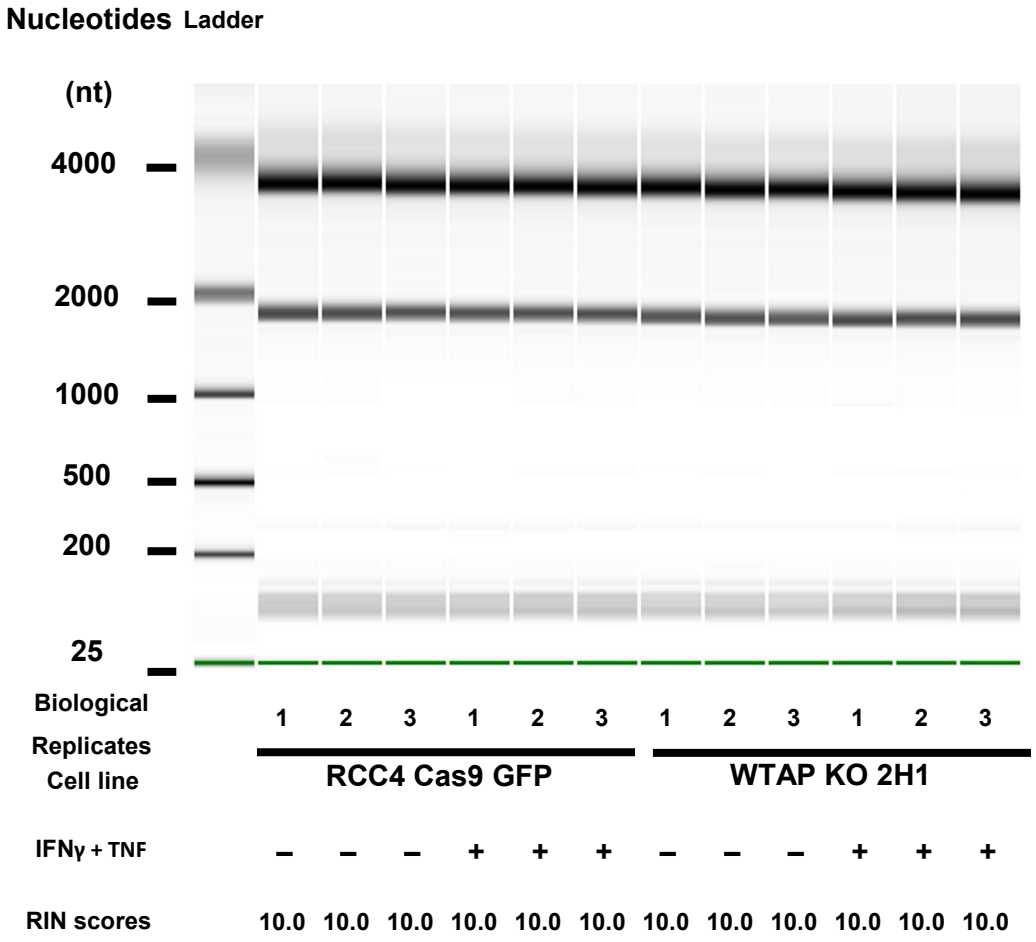


Figure 5.6: RCC4 Cas9 GFP and WTAP KO 2H1 RNA analysis by Agilent 2100 bioanalyzer

Analysis of extracted total RNA from RCC4 Cas9 GFP and WTAP KO 2H1 on an Agilent 2100 bioanalyzer using the RNA 6000 Nano kit. Bioanalyzer gel image of the extracted RNA samples is shown, with visible 28S and 18S rRNA bands. Ratio of 28S:18S bands were used to assess integrity of RNA samples (where 10 represent intact RNAs and 1 represent completely degraded RNAs).

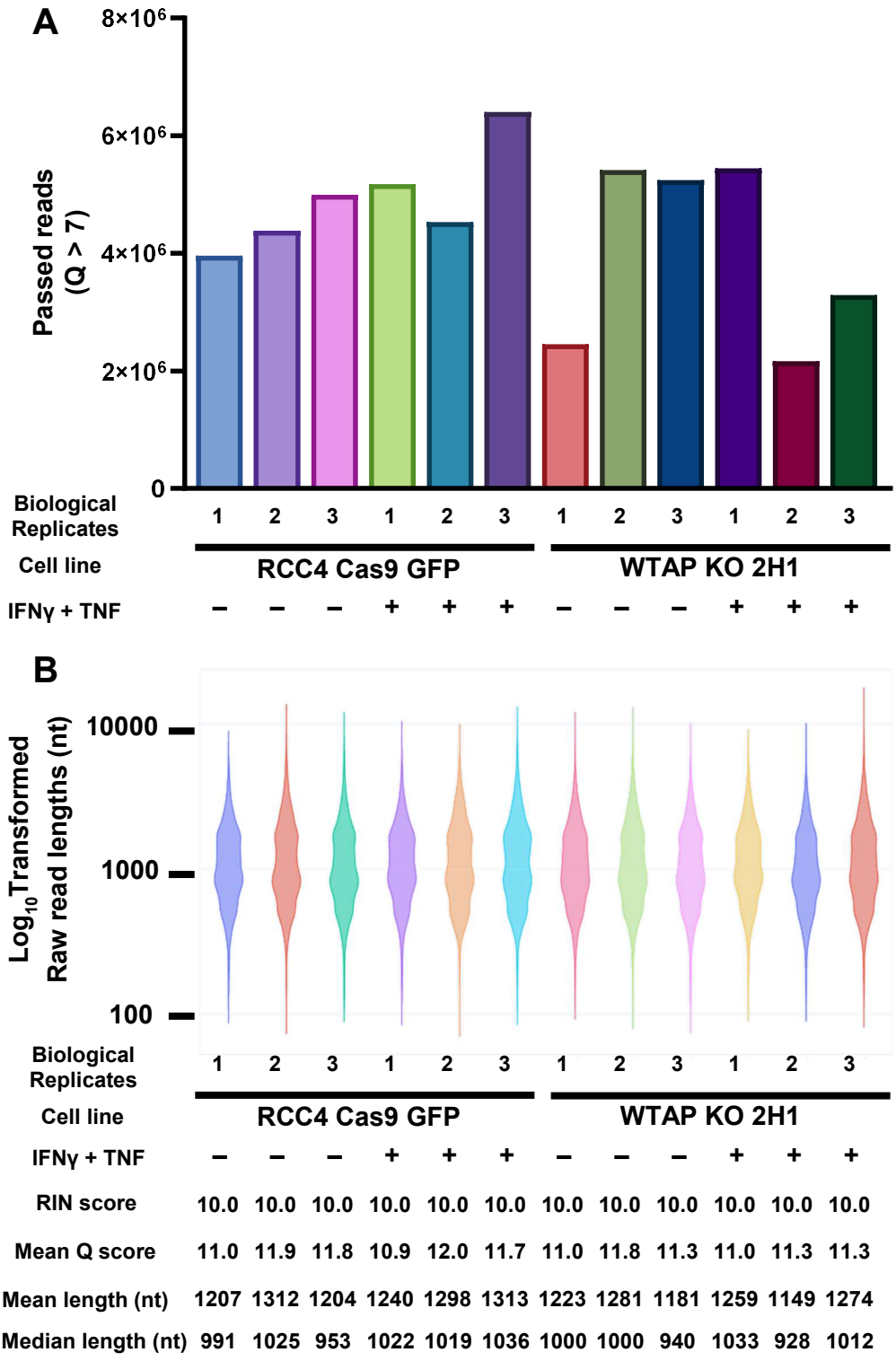


Figure 5.7: Summary of DRS reads generated from RCC4 Cas9 GFP and WTAP KO 2H1

A) Bar graphs showing the number of passed reads (Q score > 7) generated by direct RNAseq (RNA002). **B)** Violin plot showing Log10 transformed raw read lengths. RIN score, mean read Q score, mean and median read length for each sequencing dataset are listed in the tables below violin graphs.

5.3.5 DRS of RCC4 Cas9 GFP and WTAP KO 2H1 produces long reads representing full-length transcripts

DRS reads were aligned to the reference genome (Ensembl release 105, Genome assembly version: GRCh38) or reference transcriptome (Ensembl release 105, cDNA reference) by the sequence mapping and aligner *minimap2*. The read length distributions of reference genome-aligned reads from DRS of RCC4 Cas9 GFP and WTAP KO 2H1 were analysed and visualised by *Nanoplot*. Violin plots showing Log_{10} transformed genome-aligned read lengths from DRS were plotted, with mean and median read lengths for each sample indicated below the graphs (Figure 5.8A). Mean and median reference genome-aligned DRS read lengths ranged between 1,080 – 1,296 nt and 942 – 1,039 nt, with median lengths of 1,177 nt and 1,015 nt, respectively.

Reference transcriptome-aligned reads were analysed by *bamslam*, which provided summary alignment statistics for each sample (Tables 5.1). The median length of reference transcriptome-aligned reads ranged between 869 – 938 nt (median: 908.5 nt). The median read mapping accuracy (base identity) for the samples was 90.56%. Between % of the reference transcriptome aligned reads represent full-length transcript (95%+ coverage of the length of aligned reference transcript), with the median transcript coverage per aligned read ranging between 35.16 – 47.79%.

The relationship between read lengths and transcript coverage was next explored. The median raw read lengths were found to be significantly correlated with the median read alignment lengths (R^2 = 0.8419, p = <0.0001) (Figure 5.9A). However, unlike results from PCS and DRS of archival tumour samples (Figure 3.8B – C), no significant correlation was found between the median alignment lengths and the percentage of full-length transcript reads (Figure 5.9B).

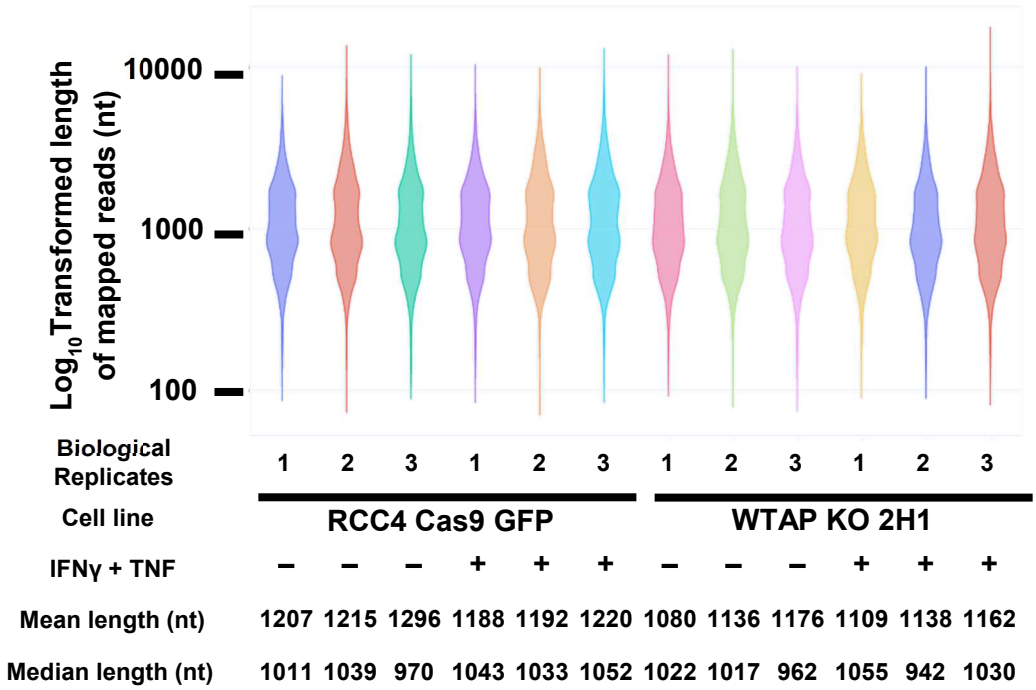


Figure 5.8: Distribution of reference genome aligned read lengths from DRS of RCC4 Cas9 GFP and WTAP KO 2H1

Violin plots showing Log_{10}transformed reference genome aligned (Ensembl release 105, GRCh38) read lengths from Direct RNAseq. Mean and median read lengths for each sequencing dataset are listed in the tables below violin graphs.

Cell line	RCC4 Cas9 GFP						WTAP KO 2H1					
Biological replicate	1	2	3	1	2	3	1	2	3	1	2	3
IFNγ + TNF	–	–	–	–	–	–	+	+	+	+	+	+
Median read-alignment lengths (nt)	895	911	877	937	913	927	915	907	884	938	869	929
Median accuracy of read alignments (%)	89.34	91.61	91.65	89.10	91.96	91.34	89.47	91.54	90.37	89.58	90.45	90.29
Reads representing full-length transcripts (%)	41.02	46.17	47.12	38.66	45.97	47.79	35.16	44.93	43.89	40.50	41.53	43.03

Table 5.1: Alignment statistics from DRS of RCC4 Cas9 GFP and WTAP KO 2H1

Statistics related to alignment of Direct RNAseq generated reads to the reference transcriptome (Ensembl release 105, cDNA reference), including median lengths of read-alignment, median accuracy of reads, percentage of reads which represent full-length transcripts (covering at least 95% of annotated transcript where the read was aligned).

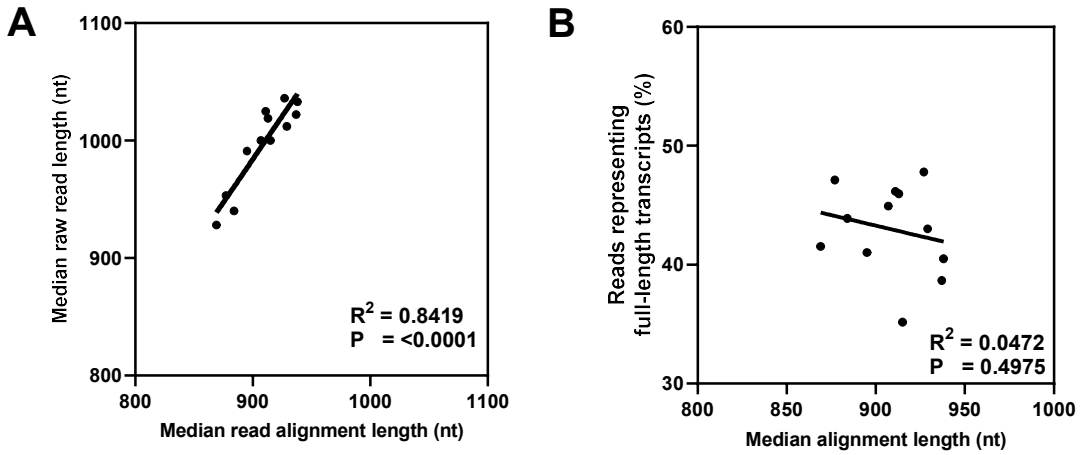


Figure 5.9: Correlations between DRS read alignment lengths and coverage

A) Correlation between median raw read lengths and median reference transcriptome alignment lengths**. B)** Correlation between median reference transcriptome alignment lengths and percentage of reads representing full-length transcripts (covering at least 95% of annotated transcript). Throughout, diagonal lines represent the line of best fit. R^2 values were computed to measure goodness of fit, and P values generated from F-test, with $p < 0.05$ considered significant.

5.3.6 Statistics of unique genes identified by reference genome and reference transcriptome aligned DRS data

Next, the number of unique genes identified in the sequencing experiment was surveyed to assess the breadth of transcriptome coverage. For reference genome aligned data, the median number of unique genes identified per sample was 14,948, with 20,522 unique genes identified across all samples (Figure 5.10A). For reference transcriptome-aligned data, the median number of unique genes found per sample was 12,652, and 16,244 unique genes were identified across all sequenced samples (Figure 5.10B). For both reference genome and reference transcriptome aligned data, the number of detected unique genes showed significant positive correlations with the number of passed reads (Q > 7) generated for each sample. Results here demonstrate the importance of generating a high number of reads to ensure comprehensive transcriptome coverage.

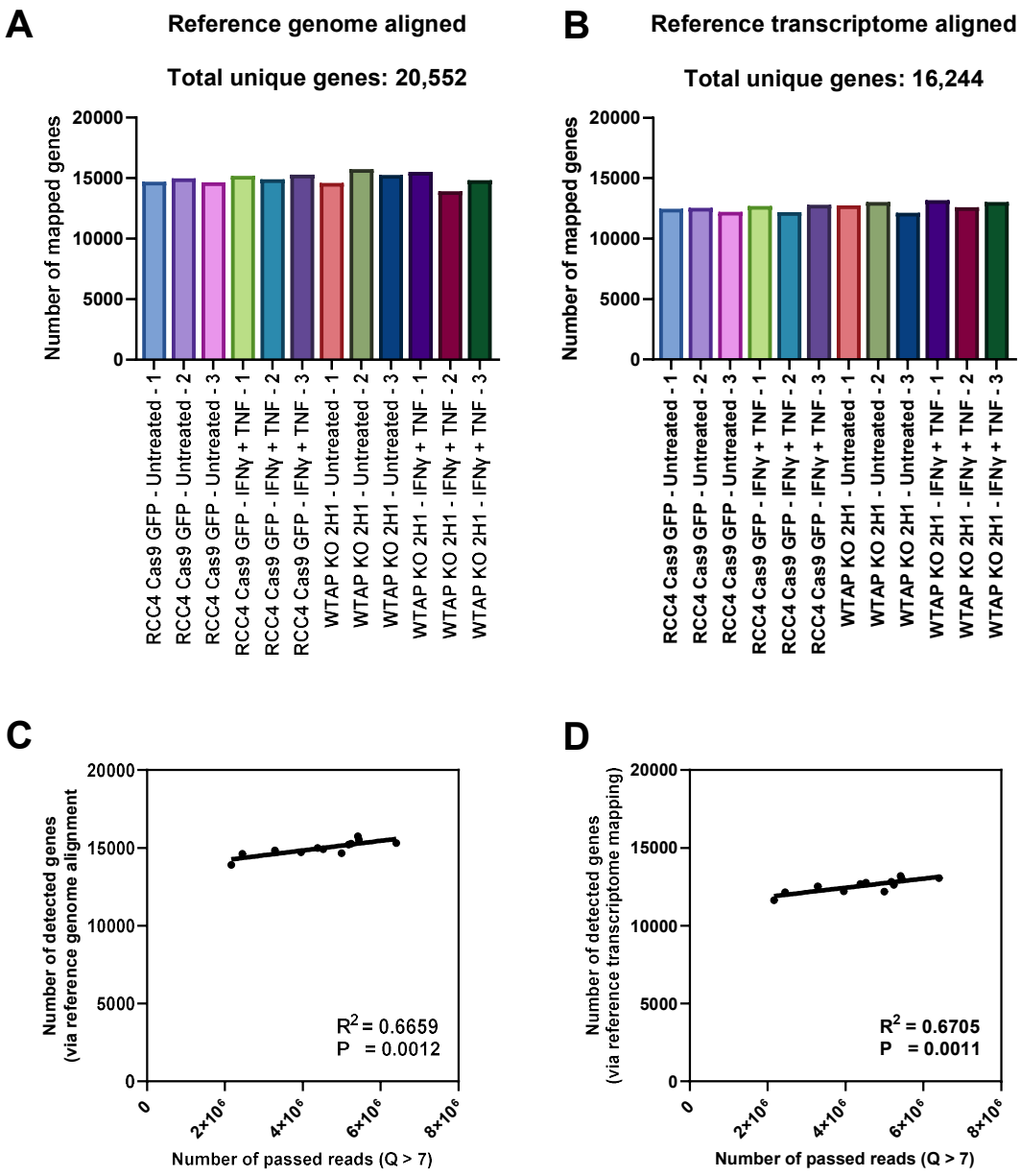


Figure 5.10: Unique genes identified from reference genome and reference transcriptome aligned DRS of RCC4 Cas9 GFP and WTAP KO 2H1

A) Bar graph demonstrating the number of unique genes identified from reference genome (Ensembl release 105, GRCh38) aligned DRS of RCC4 Cas9 GFP and WTAP KO 2H1. **B)** As in **A**, but for reference transcriptome aligned DRS. **C)** Correlation of the number of genes detected via reference genome alignment and the number of DRS passed reads (Q > 7) generated. **D)** As in **C**, but for reference transcriptome alignment. Throughout, diagonal lines represent the line of best fit. R^2 values were computed to measure goodness of fit, and P values generated from F-test, with $p < 0.05$ considered significant.

5.3.7 Composition of RNA biotypes from DRS of RCC4 Cas9 GFP and WTAP KO 2H1

To further evaluate the transcriptomic diversity captured in the sequencing experiment, the RNA biotypes of the reference genome (Ensembl release 105, GRCh38) and reference transcriptome (Ensembl release 105, cDNA reference) aligned transcripts were evaluated. For reference genome alignment data, the average RNA biotype proportion of identified unique genes across the 12 samples is displayed in Figure 5.11A. The majority of identified genes were classified as protein-coding (66.67%), followed by lncRNA (21.75%), processed pseudogenes (6.21%) and transcribed unprocessed pseudogenes (1.48%). For reference transcriptome alignment data, the average RNA biotype proportion of identified unique genes is displayed in Figure 5.11B. On average, 90.41% of identified genes are protein-coding, followed by processed pseudogene (3.73%) and lncRNA (2.35%).

Next, biotype composition by transcript abundance, pie charts depicting the averaged proportions of RNA biotypes of the reference genome and transcriptome-aligned data by expression levels (Reads per million (RPM)) across the 12 samples were constructed (Figure 5.11C – D). For reference genome aligned data, 93.71% of mapped reads were classified as protein coding, followed by 3.34% processed pseudogene, 1.43% mt-rRNA and 0.72% lncRNA. rRNA constituted a negligible proportion of mapped reads. In contrast, 99.65% of reads were mapped to protein-coding genes when aligned to reference transcriptome, with only 0.14% of aligned reads mapped to lncRNA.

Similar to the gene expression levels profile of the RNA biotype described in chapter 3.3.12, the highest expressing biotype for reference genome-aligned data on average across samples was mt-rRNA (n = 2, RPM: 10,752 and 3,577), followed by protein-coding genes (n = 13702, median RPM: 11.07). mt-rRNA were not mapped in reference transcriptome-aligned data, where protein-coding genes showed the highest expression level across biotypes (n = 14686, median RPM: 6.358), followed by lncRNA (n = 381, median RPM: 0.4879).

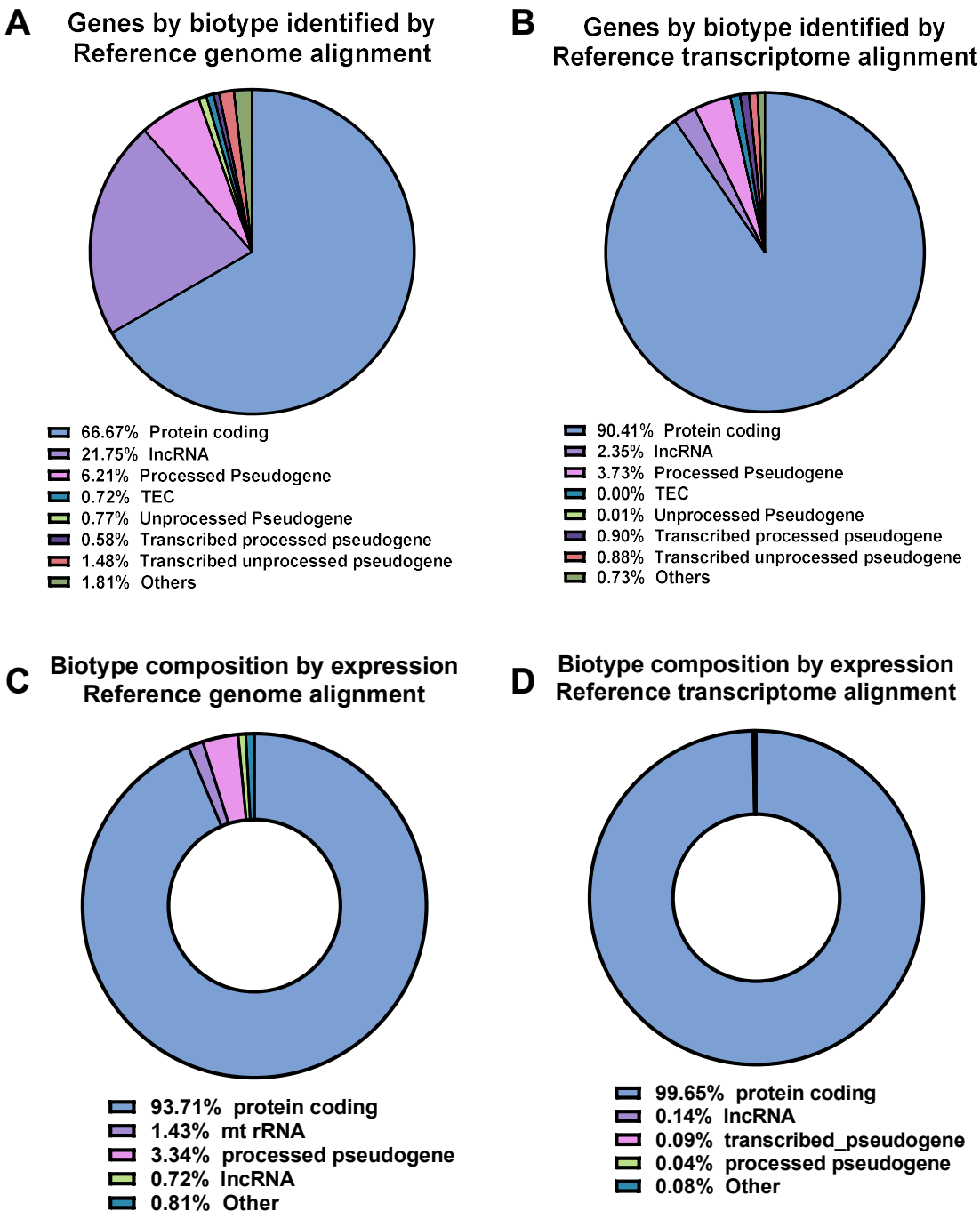


Figure 5.11: RNA biotype composition of RCC4 Cas9 GFP and WTAP KO 2H1 by DRS

A) Pie chart depicting the average proportions of RNA biotypes of reference genome (Ensembl release 105, GRCh38) aligned DRS reads from RCC4 Cas9 GFP and WTAP KO 2H1. **B)** As in **A** but aligned to the reference transcriptome (Ensembl release 105, cDNA reference). **C)** Pie chart depicting the average proportions of gene biotypes of reference genome aligned DRS reads by expression levels (scaled to library size, reads per million (RPM)). **D)** As in **C**, but aligned to the reference transcriptome.

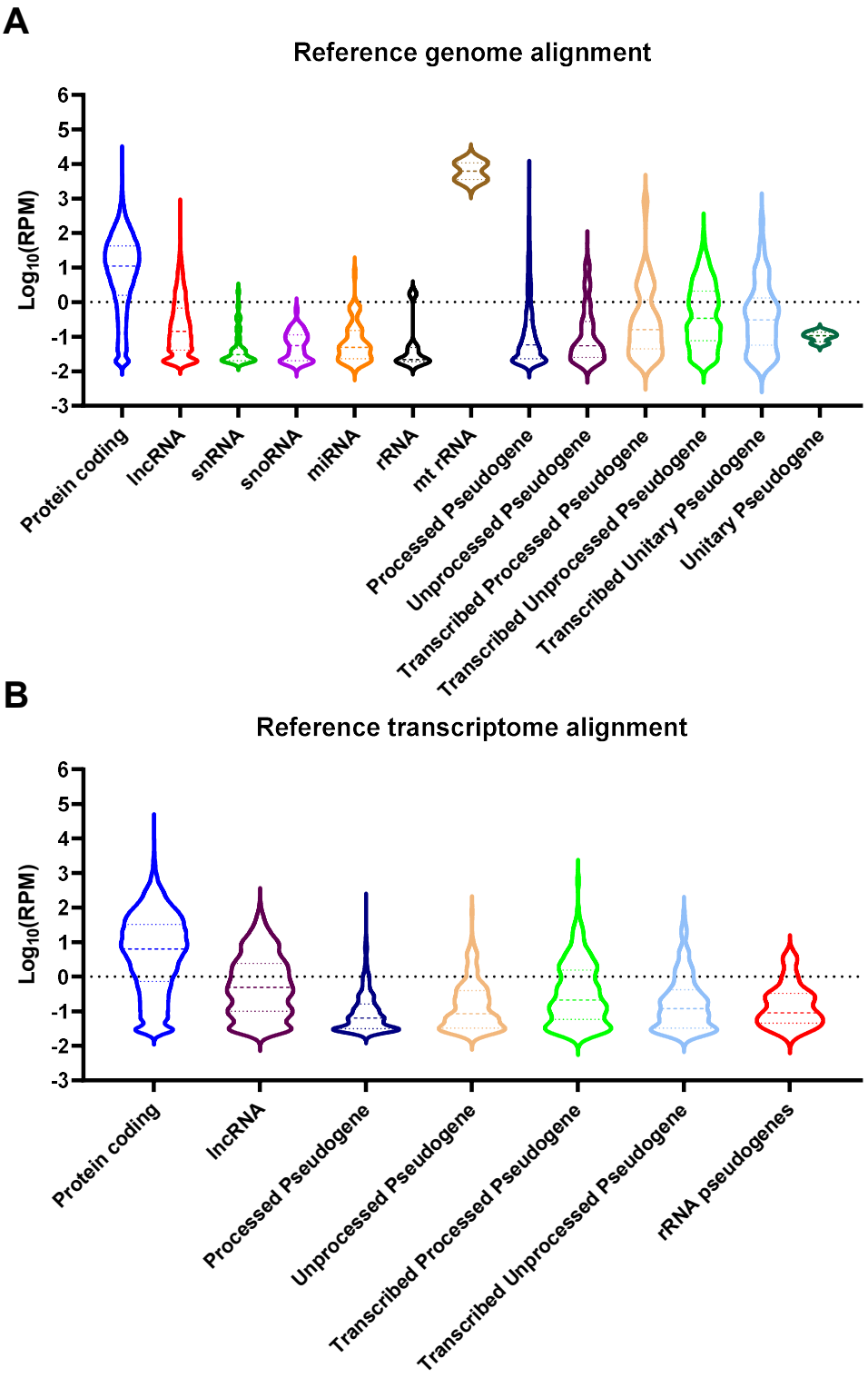


Figure 5.12: Expression levels of genes identified in RCC4 Cas9 GFP and WTAP KO 2H1 by biotypes

A) Violin plot depicting the distribution of gene expression levels (Log_{10} Reads per million (RPM)) of genes detected by reference genome aligned (Ensembl release 105, GRCh38) DRS by biotypes. **B)** As in **A**, but for reference transcriptome aligned DRS data.

5.3.8 Overlapping genes identified from the reference genome and reference transcriptome aligned DRS data

Collating unique genes identified by reference genome alignment and reference transcriptome alignment, the extent of overlap was determined using a Venn diagram (Figure 5.13A). A total of 13,657 genes were identified by both reference genome alignment and reference transcriptome alignment. Reference genome alignment identified 6,895 genes that were not identified by reference transcriptome alignment, whereas reference transcriptome alignment found 2,587 genes that were not in reference genome-aligned data.

The biotype composition of the reference genome exclusive, reference transcriptome exclusive and genes commonly mapped by both methods is demonstrated as a stacked bar chart in Figure 5.13B. Of the 13,657 genes identified by both methods, 95.5% (13,047) were classified as protein-coding, and 4.3% (591) were pseudogenes. Only three lncRNA were commonly found by both reference genome and reference transcriptome alignment. For reference genome-exclusive genes, 64.8% (4,467) were lncRNA, 19.4% (1,336) were pseudogenes, and 9.5% (656) were protein-coding genes. For reference transcriptome-exclusive genes, the majority were protein-coding genes at 63.3% (1,639), followed by pseudogenes at 21.8% (563) and lncRNA at 14.6% (378).

Focusing on genes identified by both alignments, a scatter plot between the reference genome and reference transcriptome-aligned gene expression data was generated (Figure 5.13C). Gene expression levels were highly correlated between the two methods ($R^2 = 0.6515$, $p < 0.0001$).

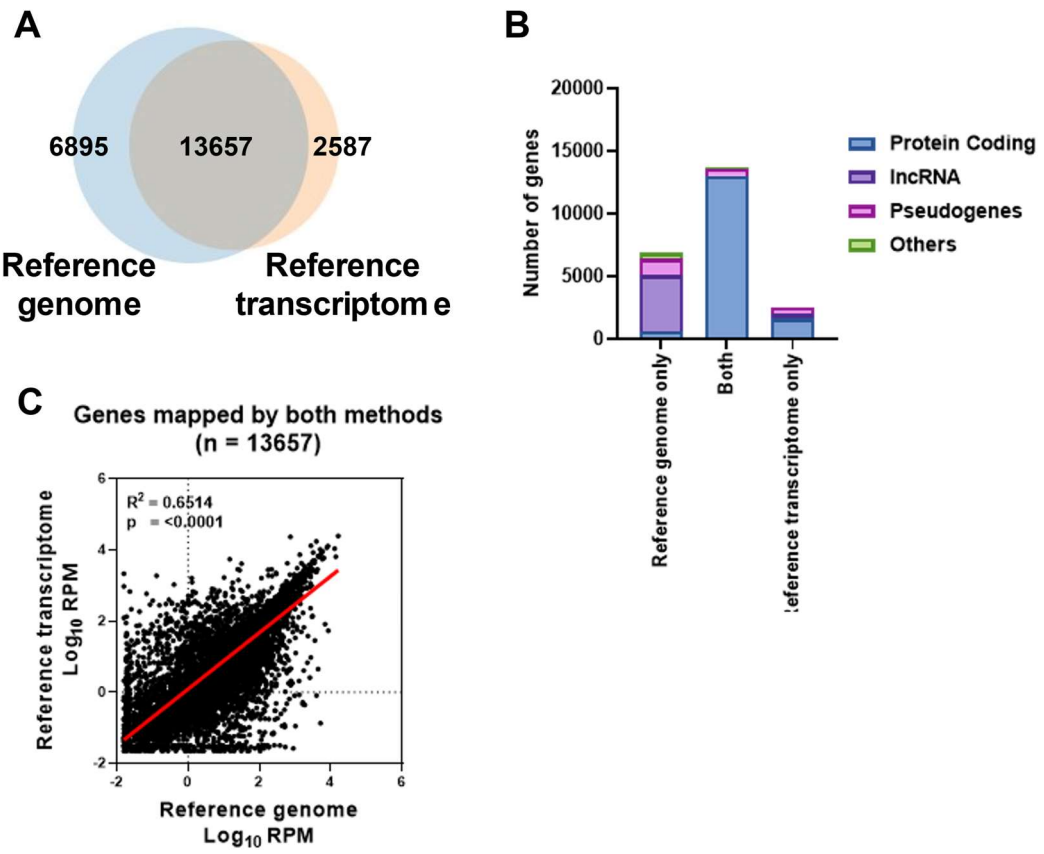


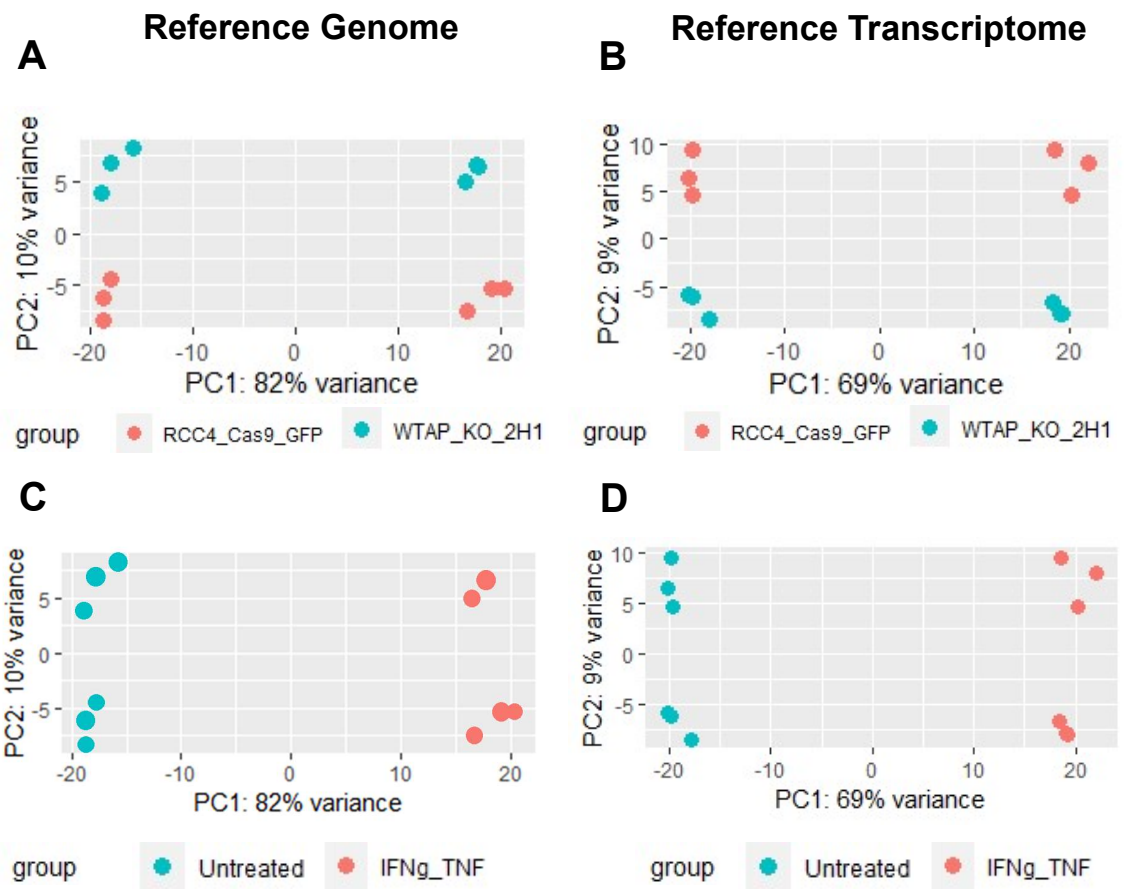
Figure 5.13: Differences and common genes identified in RCC4 Cas9 GFP and WTAP KO 2H1 via reference genome and reference transcriptome alignment

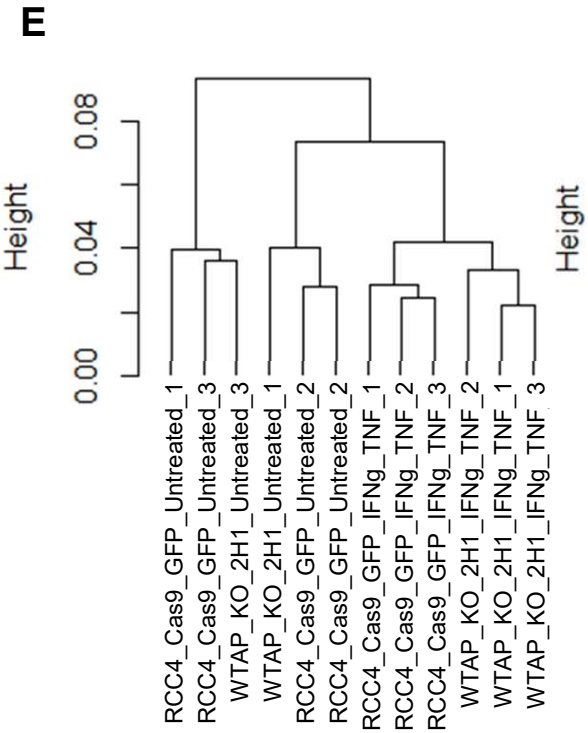
A) Venn diagram showing the overlap between reference genome aligned and reference transcriptome aligned DRS identified genes from RCC4 Cas9 GFP and WTAP KO 2H1. **B)** Stacked bar chart showing the number and proportions of RNA biotypes of genes detected by reference genome exclusively, reference transcriptome exclusively, or detected by both. **C)** Correlation between gene expression levels (Log_{10} Reads per million (RPM)) of all genes detected by both reference genome alignment (Ensembl release 105, GRCh38) and reference transcriptome alignment (Ensembl release 105, cDNA reference) of DRS of RCC4 Cas9 GFP and WTAP KO 2H1. 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.

5.3.9 Evaluation of RCC4 Cas9 GFP and WTAP KO 2H1 transcriptomes by PCA and hierarchical clustering

To understand the effects of WTAP-KO and IFNγ+TNF treatment on the transcriptome-wide gene expression profiles, non-supervised PCA and hierarchical clustering analysis were conducted. For PCA of reference genome-aligned data, the first principal component (PC1, x-axis) explained 82% of data variation, and the second principal component (PC2, y-axis) explained 10% of data variation (Figure 5.14A, 5.14C). For PCA of reference transcriptome-aligned data, PC1 (x-axis) explained 69% of data variation, and PC2 (y-axis) explained 9% of data variation (Figure 5.14B, 5.14D). Four distinct clustering of samples were observed. PC1 (x-axis) was explained by IFNγ+TNF treatment-induced differential gene expression, whereas PC2 (y-axis) represented WTAP KO status-related gene expression variation. Overall, combining the two PCs explained 92% and 78% of gene expression variance in the reference genome and reference transcriptome-aligned data.

Dendrograms of hierarchical clustering analysis were constructed based on Spearman rank correlation coefficients of gene expression levels between samples (Figure 5.14E – F). In agreement with PCA results, samples from the same cell line and the same treatment were clustered initially. Untreated and IFNγ + TNF treated samples clustered separately in most cases, except for 'RCC4 Cas9 GFP untreated 1' via reference transcriptome alignment (Figure 5.14F). Taken together, both PCA and hierarchical clustering results demonstrated the distinct gene expression profiles between RCC4 Cas9 GFP and WTAP KO 2H1 cell lines, with and without IFNγ+TNF treatment.


Reference Genome
Reference Transcriptome
A
B
PC2: 10% variance
PC1: 82% variance
PC2: 9% variance
PC1: 69% variance
group
RCC4_Cas9_GFP
WTAP_KO_2H1
C
D
PC2: 10% variance
PC1: 82% variance
PC2: 9% variance
PC1: 69% variance
group
Untreated
IFNg_TNF


E
Height
0.00
0.04
0.08
RCC4_Cas9_GFP_Untreated_1
RCC4_Cas9_GFP_Untreated_3
WTAP_KO_2H1_Untreated_3
WTAP_KO_2H1_Untreated_1
RCC4_Cas9_GFP_Untreated_2
RCC4_Cas9_GFP_IFNg_TNF_1
RCC4_Cas9_GFP_IFNg_TNF_2
RCC4_Cas9_GFP_IFNg_TNF_3
WTAP_KO_2H1_IFNg_TNF_2
WTAP_KO_2H1_IFNg_TNF_1
WTAP_KO_2H1_IFNg_TNF_3

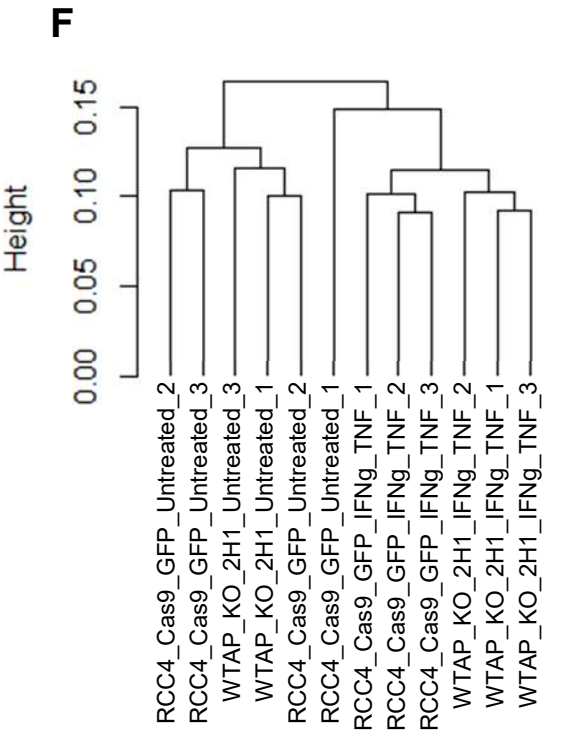

F
Height
0.00
0.05
0.10
0.15
RCC4_Cas9_GFP_Untreated_2
RCC4_Cas9_GFP_Untreated_3
WTAP_KO_2H1_Untreated_3
WTAP_KO_2H1_Untreated_1
RCC4_Cas9_GFP_Untreated_2
RCC4_Cas9_GFP_Untreated_1
RCC4_Cas9_GFP_IFNg_TNF_1
RCC4_Cas9_GFP_IFNg_TNF_2
RCC4_Cas9_GFP_IFNg_TNF_3
WTAP_KO_2H1_IFNg_TNF_2
WTAP_KO_2H1_IFNg_TNF_1
WTAP_KO_2H1_IFNg_TNF_3

Figure 5.13: PCA and hierarchical clustering of RCC4 Cas9 GFP and WTAP KO 2H1 transcriptome profiles

Principal component analysis (PCA) on RCC4 Cas9 GFP and WTAP KO 2H1 gene expression data illustrating variations between samples (dots, n = 12). DESeq2 generated plots using **A)** Reference genome and **B)** Reference transcriptome aligned data showing PCA of RCC4 Cas9 GFP vs WTAP KO 2H1 samples. **C)** Reference genome and **D)** Reference transcriptome aligned data showing PCA of untreated vs IFNγ and TNF treated samples. **E)** Dendrogram for hierarchical clustering of reference genome mapped RCC4 Cas9 GFP and WTAP KO 2H1 DRS transcriptomes based on spearman rank correlations of gene expression levels. **F)** As in **E** but for reference transcriptome aligned data.

5.3.10 Identification of DEGs between unstimulated and IFNγ+TNF treated RCC4 Cas9 GFP and WTAP KO 2H1

Gene expression profiles of untreated and IFNγ+TNF treated RCC4 Cas9 GFP and WTAP KO 2H1 were compared and analysed by DEseq2. Genes with log_2FoldChange ≤ -2 or ≥ 2 and p_{adj} ≤ 0.1 are considered significant DEGs. Results of differential gene expression analysis (Untreated vs IFNγ+TNF treated RCC4 CAS9 GFP, untreated vs IFNγ+TNF treated WTAP KO 2H1, untreated RCC4 Cas9 GFP vs WTAP KO 2H1, IFNγ+TNF treated RCC4 Cas9 GFP vs WTA KO 2H1) by reference genome alignment and reference transcriptome alignment were plotted as volcano plots in Figure 5.14 and 5.15, respectively. The top fifteen differentially expressed genes (ranked by p_{adj} values) from the comparisons by reference genome alignment are listed in Tables 5.2 – 5.5. Comprehensive lists of DEGs can be found in Appendix tabled 7.15 – 22.

Comparing gene expression levels of untreated versus IFNγ+TNF treated RCC4 Cas9 GFP, 441 significant DEGs, including 308 upregulated and 131 downregulated genes were found using reference genome aligned data (Figure 5.14A). The top fifteen DEGs by p_{adj} values were all found to be upregulated in IFNγ+TNF treated cells, with the top three being *WARS1* (tryptophanyl-tRNA synthetase 1, Log_2FoldChange = 4.9355, p_{adj} = 9.46 x 10^{-114}), *SOD2* (superoxide dismutase 2, Log_2FoldChange = 4.5742, p_{adj} = 8.25 x

10^{-99}), *CCL2* (C-C motif chemokine ligand 2, $Log_2FoldChange = 9.6200$, $p_{adj} = 4.52 \times 10^{-90}$) (Table 5.2).

For untreated versus IFNγ+TNF treated WTAP KO 2H1, 327 significant DEGs, including 248 upregulated and 79 downregulated genes were found (Figure 5.14B). Like RCC4 Cas9 GFP, the top fifteen DEGs by p_{adj} values also consisted entirely of upregulated genes. As shown by the spread of data points at the volcano plots, IFNγ+TNF induced a large number of upregulated genes in both cell lines. The top three most significantly upregulated genes (by p_{adj} values) were *GBP1* (Guanylate binding protein 1, $Log_2FoldChange = 10.3937$, $p_{adj} = 7.66 \times 10^{-131}$), *ICAM1* (Intercellular adhesion molecule 1, $Log_2FoldChange = 5.5519$, $p_{adj} = 3.13 \times 10^{-115}$), and *CCL2* ($Log_2FoldChange = 9.6430$, $p_{adj} = 6.76 \times 10^{-111}$) (Table 5.3).

Significant DEGs associated with WTAP KO were also identified. Comparing untreated RCC4 Cas9 GFP and WTAP KO 2H1, 48 DEGs (18 upregulated genes and 30 downregulated genes) were found in reference genome alignment data (Figure 5.14C). The top 10 most significant DEGs were exclusively downregulated genes, with the top three being *NDUFA4L2* (NDUFA4 mitochondrial complex associated like 2, $Log_2FoldChange = -9.4549$, $p_{adj} = 1.15 \times 10^{-53}$), *CA9* (Carbonic anhydrase 9, $Log_2FoldChange = -3.9148$, $p_{adj} = 8.32 \times 10^{-35}$), and *EGLN3* (egl-9 family hypoxia inducible factor 3, $Log_2FoldChange = -6.6649$, $p_{adj} = 1.24 \times 10^{-18}$) (Table 5.4).

For IFNγ+TNF treated RCC4 Cas9 GFP vs WTAP KO 2H1, 65 significant DEGs (32 upregulated genes and 33 downregulated genes) were discovered in reference genome alignment data (Figure 5.14D). Similar to untreated comparisons, top DEGs were mostly downregulated genes. The top three most significant DEGs (by p_{adj} values) were *NDUFA4L2* ($Log_2FoldChange = -9.4352$, $p_{adj} = 2.48 \times 10^{-58}$), *BNIP3* (BCL2 interacting protein 3, $Log_2FoldChange = -2.3804$, $p_{adj} = 2.80 \times 10^{-50}$), and *CA9* ($Log_2FoldChange = 5.5873$, $p_{adj} = 1.16 \times 10^{-49}$) (Table 5.5). Volcano plots from reference transcriptome aligned data were shown in Figure 5.15, with the similarities and differences between the DEG analyses from the two alignment methods discussed in the next section (5.3.11).

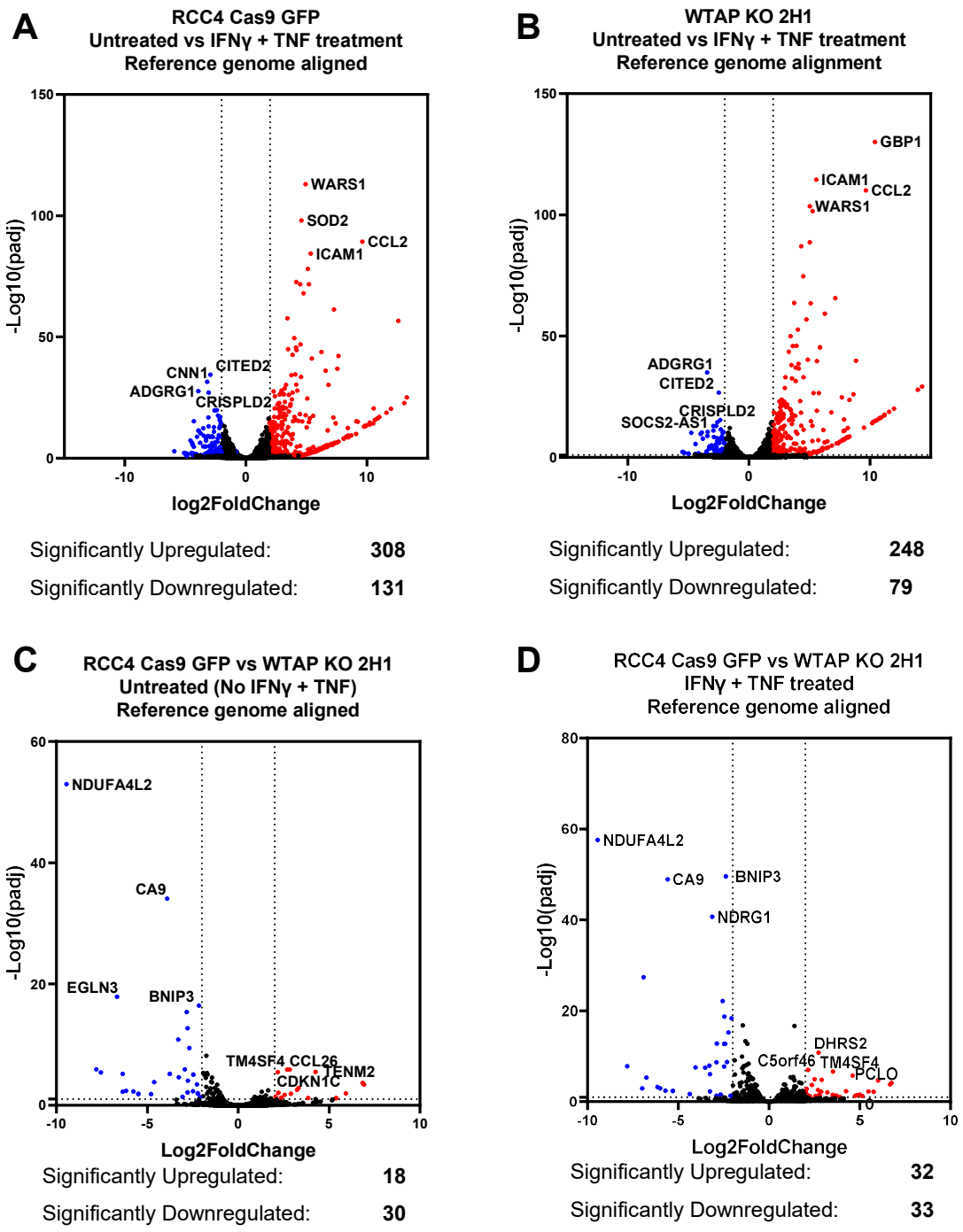


Figure 5.14: DEGs between untreated and IFNγ + TNF treated RCC4 Cas9 GFP and WTAP KO 2H1 by reference genome alignment

Volcano plots showing differentially expressed genes between **A)** untreated and IFNγ + TNF treated RCC4 Cas9 GFP, **B)** untreated and IFNγ + TNF treated WTAP KO 2H1, **C)** untreated RCC4 Cas9 GFP vs WTAP KO 2H1 and **D)** IFNγ + TNF treated RCC4 Cas9 GFP vs WTAP KO 2H1 profiled by reference genome alignment. Dotted lines indicate significance threshold ($p_{adj} \leq 0.1$, $|\log_2 FoldChange| > 2$). Significantly upregulated genes are in red and downregulated genes are in blue. Names of top 4 most significantly up/down regulated genes (by padj) are shown.

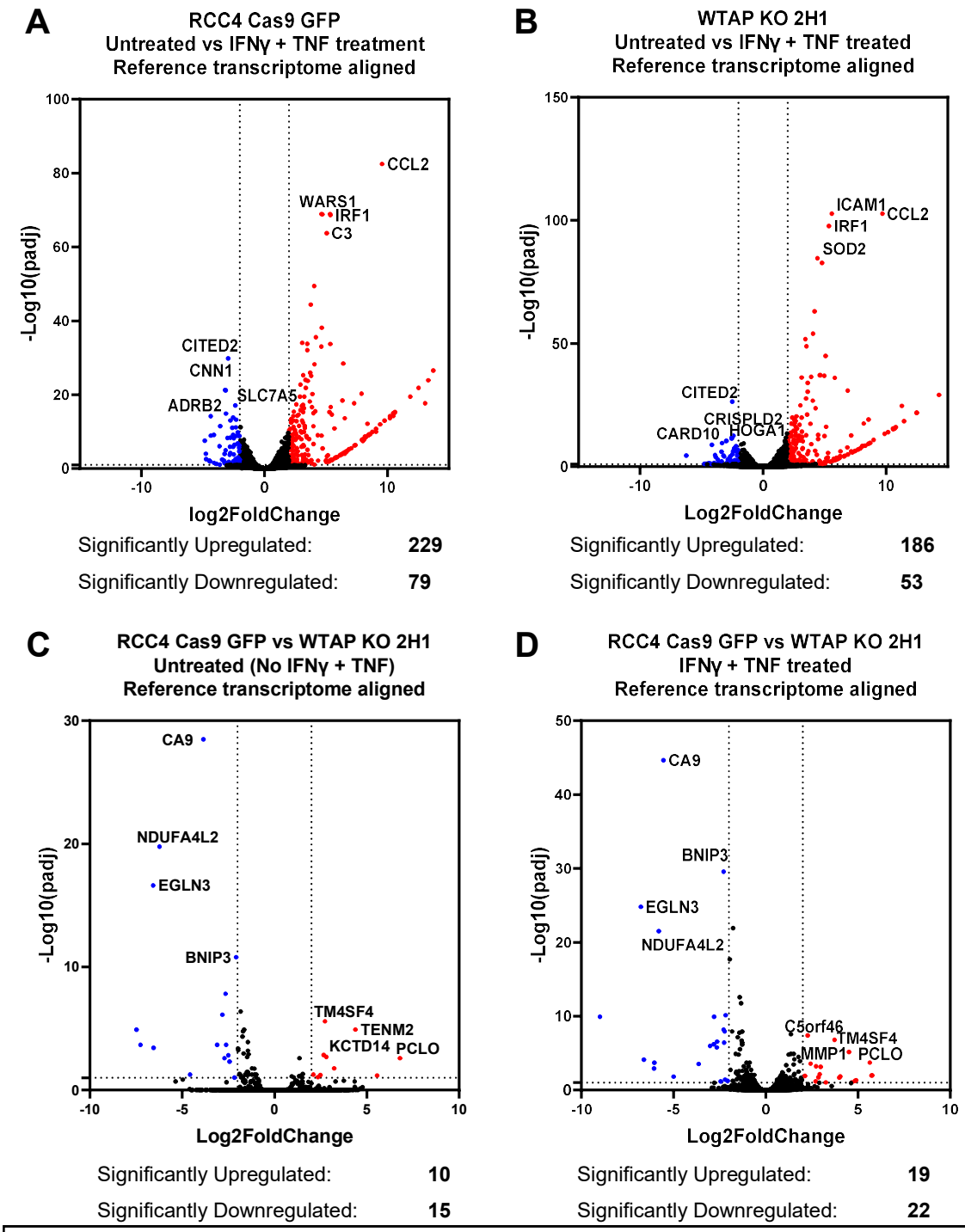


Figure 5.15: DEGs between untreated and IFNγ + TNF treated RCC4 Cas9 GFP and WTAP KO 2H1 by reference transcriptome alignment

Volcano plots showing differentially expressed genes between **A)** untreated and IFNγ + TNF treated RCC4 Cas9 GFP, **B)** untreated and IFNγ + TNF treated WTAP KO 2H1, **C)** untreated RCC4 Cas9 GFP vs WTAP KO 2H1 and **D)** IFNγ + TNF treated RCC4 Cas9 GFP vs WTAP KO 2H1 profiled by reference transcriptome alignment. Dotted lines indicate significance threshold ($p_{adj} \leq 0.1$, $|\log_2 FoldChange| > 2$). Significantly upregulated genes are in red and downregulated genes are in blue. Names of top 4 most significantly up/down regulated genes (by padj) are shown.

RCC4 Cas9 GFP - untreated vs IFNγ + TNF stimulation (Genome mapping)				
ENSEMBL ID	Gene Name	Biotype	Log_2FoldChange	padj
ENSG00000140105	WARS1	protein_coding	4.94	9.46E-114
ENSG00000112096	SOD2	protein_coding	4.57	8.25E-99
ENSG00000108691	CCL2	protein_coding	9.62	4.52E-90
ENSG00000090339	ICAM1	protein_coding	5.36	3.47E-85
ENSG00000125730	C3	protein_coding	5.11	8.12E-79
ENSG00000156587	UBE2L6	protein_coding	4.14	2.20E-73
ENSG00000125347	IRF1	protein_coding	5.20	1.75E-72
ENSG00000168394	TAP1	protein_coding	4.48	1.90E-72
ENSG00000240065	PSMB9	protein_coding	4.76	9.94E-69
ENSG00000137496	IL18BP	protein_coding	7.28	4.31E-62
ENSG00000234745	HLA-B	protein_coding	3.44	2.03E-58
ENSG00000117228	GBP1	protein_coding	12.59	2.56E-57
ENSG00000111331	OAS3	protein_coding	3.99	3.27E-50
ENSG00000068079	IFI35	protein_coding	4.50	1.24E-47
ENSG00000006210	CX3CL1	protein_coding	4.14	2.12E-46

Table 5.2 Top differentially expressed genes after IFNγ and TNF stimulation in RCC4 Cas9 GFP cells

Top 15 differential expressed genes in IFNγ + TNF treated compared to untreated, in RCC4 Cas9 GFP cells as analysed by DESeq2, ranked by padj values.

WTAP KO 2H1 - untreated vs IFNγ + TNF stumulation (Genome mapping)				
ENSEMBL ID	Gene Name	Biotype	Log_2FoldChange	padj
ENSG00000117228	GBP1	protein_coding	10.39	7.66E-131
ENSG00000090339	ICAM1	protein_coding	5.55	3.13E-115
ENSG00000108691	CCL2	protein_coding	9.64	6.76E-111
ENSG00000140105	WARS1	protein_coding	5.02	2.80E-104
ENSG00000125347	IRF1	protein_coding	5.26	3.24E-102
ENSG00000240065	PSMB9	protein_coding	5.01	1.93E-89
ENSG00000112096	SOD2	protein_coding	4.32	8.62E-88
ENSG00000168394	TAP1	protein_coding	4.50	2.13E-75
ENSG00000182326	C1S	protein_coding	7.13	2.41E-66
ENSG00000234745	HLA-B	protein_coding	3.74	1.84E-64
ENSG00000089127	OAS1	protein_coding	5.09	2.70E-64
ENSG00000119917	IFIT3	protein_coding	6.25	5.08E-60
ENSG00000006210	CX3CL1	protein_coding	4.74	1.41E-57
ENSG00000125730	C3	protein_coding	4.03	2.00E-53
ENSG00000184371	CSF1	protein_coding	3.42	1.17E-50

Table 5.3 Top differentially expressed genes after IFNγ and TNF stimulation in WTAP KO 2H1 cells

Top 15 differential expressed genes in IFNγ + TNF treated compared to untreated, in WTAP KO 2H1 cells as analysed by DESeq2, ranked by padj values.

RCC4 Cas9 GFP vs WTAP KO 2H1 without IFNγ + TNF stimulation (Genome mapping)				
ENSEMBL ID	Gene Name	Biotype	Log_2FoldChange	padj
ENSG00000185633	NDUFA4L2	protein_coding	-9.45	1.15E-53
ENSG00000107159	CA9	protein_coding	-3.91	8.32E-35
ENSG00000129521	EGLN3	protein_coding	-6.66	1.24E-18
ENSG00000176171	BNIP3	protein_coding	-2.16	3.78E-17
ENSG00000104419	NDRG1	protein_coding	-2.84	4.43E-16
ENSG00000168209	DDIT4	protein_coding	-2.79	2.07E-13
ENSG00000109107	ALDOC	protein_coding	-3.30	1.57E-11
ENSG00000114268	PFKFB4	protein_coding	-2.69	3.71E-10
ENSG00000234745	HLA-B	protein_coding	-1.76	6.72E-09
ENSG00000247095	MIR210HG	lncRNA	-2.94	1.23E-06
ENSG00000123146	ADGRE5	protein_coding	-7.81	1.23E-06
ENSG00000169903	TM4SF4	protein_coding	2.69	1.34E-06
ENSG00000006606	CCL26	protein_coding	2.82	1.41E-06
ENSG00000145934	TENM2	protein_coding	4.24	3.33E-06
ENSG00000129757	CDKN1C	protein_coding	2.17	3.58E-06

Table 5.4 Top differentially expressed genes between RCC4 Cas9 GFP and WTAP KO 2H1 cells without IFNγ+TNF stimulation

Top 15 differential expressed genes between unstimulated RCC4 Cas9 GFP and WTAP KO 2H1 cells as analysed by DESeq2, ranked by padj values.

RCC4 Cas9 GFP vs WTAP KO 2H1 with IFNγ + TNF stimulation (Genome mapping)				
ENSEMBL ID	Gene Name	Biotype	Log_2FoldChange	padj
ENSG00000185633	NDUFA4L2	protein_coding	-9.44	2.48E-58
ENSG00000176171	BNIP3	protein_coding	-2.38	2.80E-50
ENSG00000107159	CA9	protein_coding	-5.59	1.16E-49
ENSG00000104419	NDRG1	protein_coding	-3.14	2.06E-41
ENSG00000129521	EGLN3	protein_coding	-6.92	3.85E-28
ENSG00000100342	APOL1	protein_coding	-2.56	7.13E-23
ENSG00000168209	DDIT4	protein_coding	-2.47	1.86E-19
ENSG00000113739	STC2	protein_coding	-2.07	4.72E-19
ENSG00000234745	HLA-B	protein_coding	-1.45	1.58E-17
ENSG00000146674	IGFBP3	lncRNA	1.40	2.07E-17
ENSG00000181458	TMEM45A	protein_coding	-2.24	5.48E-16
ENSG00000102144	PGK1	protein_coding	-1.31	4.78E-14
ENSG00000114268	PFKFB4	protein_coding	-2.88	1.66E-13
ENSG00000134333	LDHA	protein_coding	-1.20	1.92E-13
ENSG00000114023	FAM162A	protein_coding	-2.47	1.92E-13

Table 5.5 Top differentially expressed genes between RCC4 Cas9 GFP and WTAP KO 2H1 cells with IFNγ+TNF stimulation

Top 15 differential expressed genes between unstimulated RCC4 Cas9 GFP and WTAP KO 2H1 cells as analysed by DESeq2, ranked by padj values.

5.3.11 Characterisation of DEGs between unstimulated and IFNγ+TNF treated RCC4 Cas9 GFP and WTAP KO 2H1

After identifying DEGs between unstimulated and IFNγ + TNF stimulated RCC4 Cas9 GFP and WTAP KO 2H1, both reference genome and transcriptome-aligned data were collated. 496 and 367 unique genes were significantly differentially expressed after IFNγ + TNF treatment in RCC4 Cas9 GFP and WTAP KO 2H1 (Figure 5.16A – B). 54 and 71 unique genes were significantly differentially expressed between RCC4 Cas9 GFP and WTAP KO 2H1 under no treatment or IFNγ + TNF stimulation, respectively (Figure 5.16 C – D). In addition, Venn diagrams showed between 37.0% (Unstimulated RCC4 Cas9 GFP vs WTAP KO 2H1) to 54.2% (WTAP KO 2H1 untreated vs IFNγ + TNF treated) of DEGs were identified by both reference genome and reference transcriptome alignment for all DEG analysis.

The differential expression levels (Log2FoldChange) of DEGs between alignment methods were further analysed. Scatter plots were plotted with Log2FoldChange values of DEGs detected in both reference genome mapping and reference transcript mapping (Figure 5.17). For untreated vs IFNγ + TNF treated RCC4 Cas9 GFP and WTAP KO 2H1, 400/496 and 299/367 DEGs were detected in both reference genome and reference transcriptome aligned data. For untreated and IFNγ + TNF treated RCC4 Cas9 GFP vs WTAP KO 2H1, 40/54 and 54/71 DEGs were detected by both alignment methods. Expression patterns between the alignment methods showed high levels of concordance across all four comparisons. Whilst most genes exhibit similar expression trends between the methods (blue and red dots), several outliers were also identified (black dots). For example, with reference genome alignment, *CTSO* (Cathepsin O) was identified as significantly upregulated in both RCC4 Cas9 GFP ($Log_2FoldChange = 2.79$, $p_{adj} = 5.49 \times 10^{-5}$) and WTAP KO 2H1 ($Log_2FoldChange = 2.43$, $p_{adj} = 2.17 \times 10^{-2}$) upon IFNγ + TNF treatment. However, *CTSO* was not identified as a DEG when aligned to reference transcriptome in neither RCC4 Cas9 GFP ($Log_2FoldChange = -0.42$, $p_{adj} = 0.85$) nor WTAP KO 2H1 ($Log_2FoldChange = -0.6744$, $p_{adj} = 0.99$) (Figure 5.17 A – B).

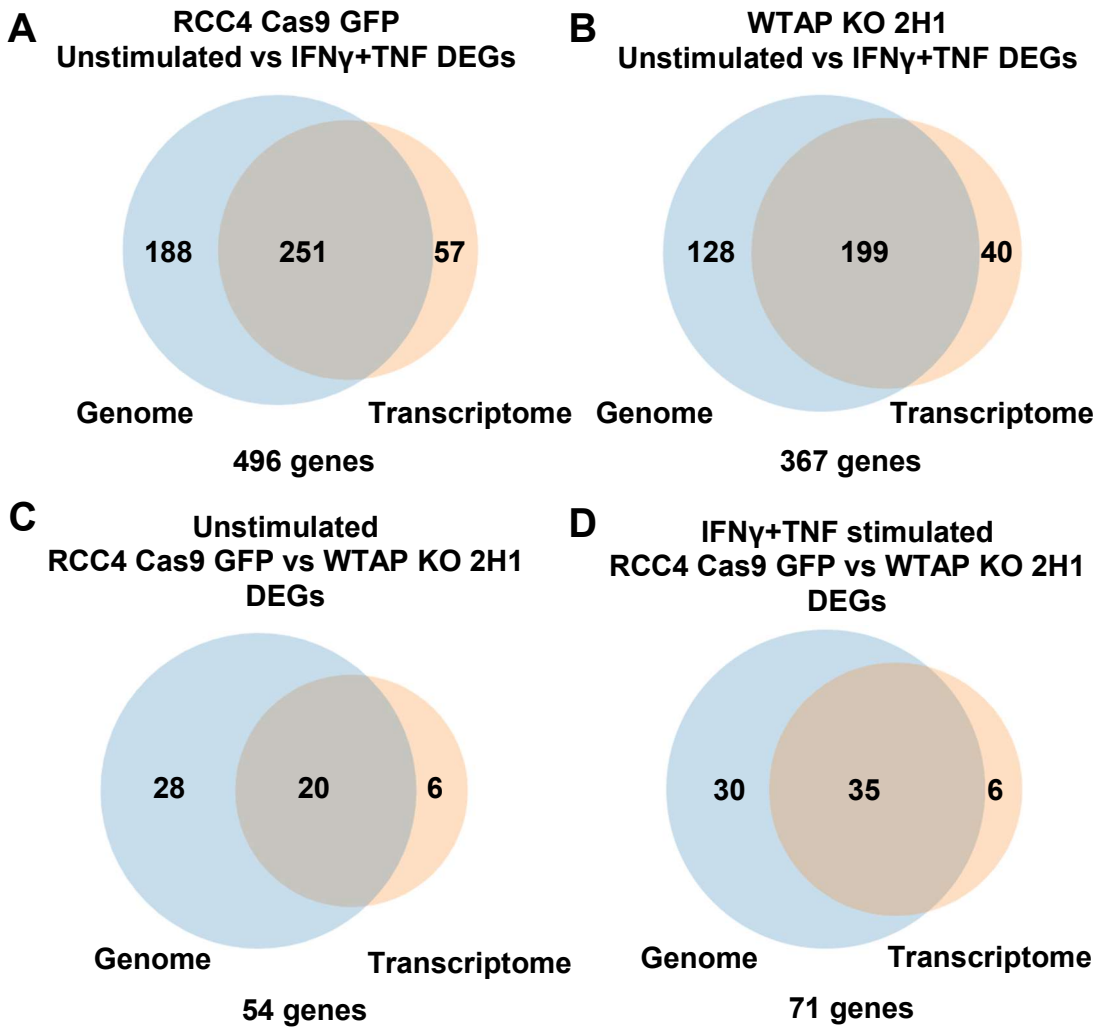


Figure 5.16: Common DEGs identified by reference genome and reference transcriptome aligned DRS of RCC4 Cas9 GFP and WTAP KO 2H1

A) Venn diagram showing overlaps of DEGs ($p_{adj} \leq 0.1, |\log_2 FoldChange| > 2$) between unstimulated and IFNγ + TNF stimulated RCC4 Cas9 GFP by reference genome and reference transcriptome alignment. **B)** As in **A**, but for WTAP KO 2H1. **C)** Venn diagram showing overlaps of DEGs ($p_{adj} \leq 0.1, |\log_2 FoldChange| > 2$) between unstimulated RCC4 Cas9 GFP and WTAP KO 2H1. **D)** As in **C**, but for IFNγ + TNF stimulated RCC4 Cas9 GFP and WTAP KO 2H1. Throughout, the total number of unique DEGs identified by either alignment methods are shown below the Venn diagrams.

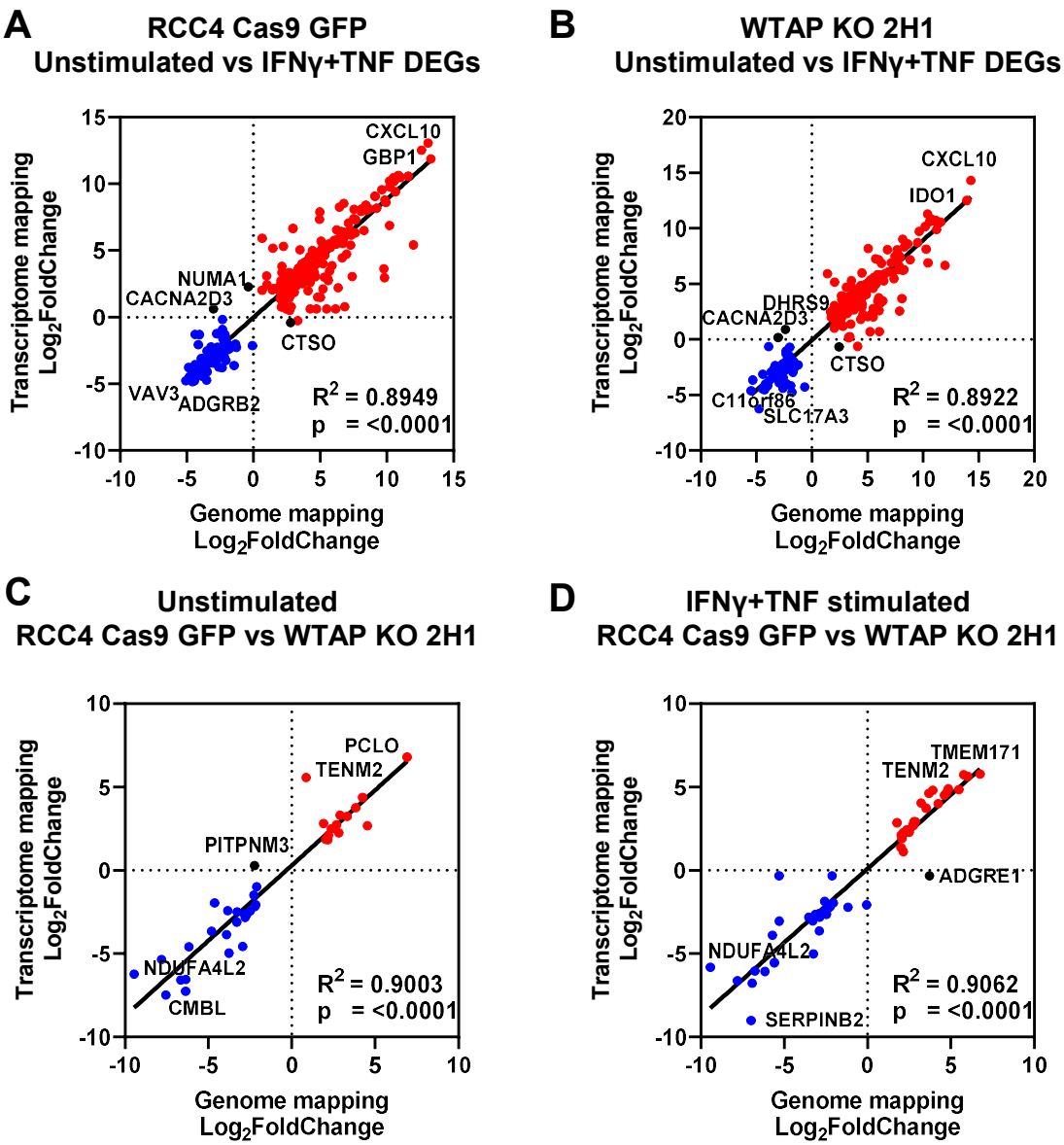


Figure 5.17: Evaluation of DEGs expression levels profiled by reference genome and reference transcriptome alignment

A) Correlation of Log_2FoldChange of commonly found unstimulated vs IFNγ + TNF stimulated RCC4 Cas9 GFP DEGs (n = 400, $p_{adj} \leq 0.1$, $|log_2FoldChange| > 2$) between reference genome and reference transcriptome alignment. **B)** As in **A**, but for unstimulated vs IFNγ + TNF stimulated WTAP KO 2H1 DEGs (n = 299). **C)** As in **A**, but for unstimulated RCC4 Cas9 GFP vs WTAP KO 2H1 DEGs (n = 40). **D)** As in **A**, but for IFNγ + TNF stimulated RCC4 Cas9 GFP vs WTAP KO 2H1 DEGs (n = 54). Red dots signify significantly upregulated genes. Blue dots signify significantly downregulated genes. Black dots signify DEGs which showed opposing expression patterns between alignment methods. 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.

5.3.12 Impact of *WTAP* KO on IFNγ+TNF stimulated DEGs

Transcriptomic data here have shown that IFNγ + TNF stimulations induced dramatic changes in the transcriptomic profiles of both RCC4 Cas9 GFP and WTAP KO 2H1. To assess the impact of WTAP on IFNγ + TNF stimulation-induced differential gene expression, lists of DEGs from RCC4 Cas9 GFP untreated vs IFNγ + TNF treated, and WTAP KO 2H1 untreated vs IFNγ + TNF were compared using a Venn diagram (Figure 5.18A). Of the 496 DEGs from RCC4 Cas9 GFP and 367 DEGs from WTAP KO 2H1 (combining both reference genome and transcriptome aligned DEGs, as shown in Figure 5.16A – B), 305 genes were identified as DEGs in both cell lines. Amongst the overlapping genes, 237 genes were significantly upregulated, and 68 genes were significantly downregulated. 191 IFNγ + TNF stimulation-induced DEGs were identified exclusively in RCC4 Cas9 GFP, where 111 were upregulated and 80 were downregulated. 62 IFNγ + TNF stimulation-induced DEGs were found in WTAP KO 2H1 exclusively, with 36 upregulated and 26 downregulated genes.

Next, focusing on the cell type specific DEGs, scatter plots were constructed with Log2FoldChange values from RCC4 Cas9 GFP and WTAP KO 2H1 (Figure 5.18B – C). Although the plotted DEGs were only identified as differentially expressed in one of the two cell lines, their expression patterns were significantly correlated between the cell lines (RCC4 Cas9 GFP exclusive DEGs: R^2 = 0.7900, p = <0.0001; WTAP KO 2H1 exclusive DEGs: R^2 = 0.8277, p = <0.0001). For the 191 RCC4 Cas9 GFP-exclusive DEGs, 22 genes showed at least a four-fold difference in gene expression levels ($|\log_2 FoldChange| > 2$) between the two cell lines (Figure 5.18B). Of the 62 WTAP KO 2H1 exclusive DEGs, 12 genes exhibited at least a four-fold difference in gene expression levels between the two cell lines (Figure 5.18C). Data presented here suggest that the effects of IFNγ+TNF stimulated on significant differential gene expression were consistent in both RCC4 Cas9 GFP and WTAP KO 2H1, with a low number of genes failing to up- or downregulated upon cytokine treatment of WTAP KO 2H1 cells.

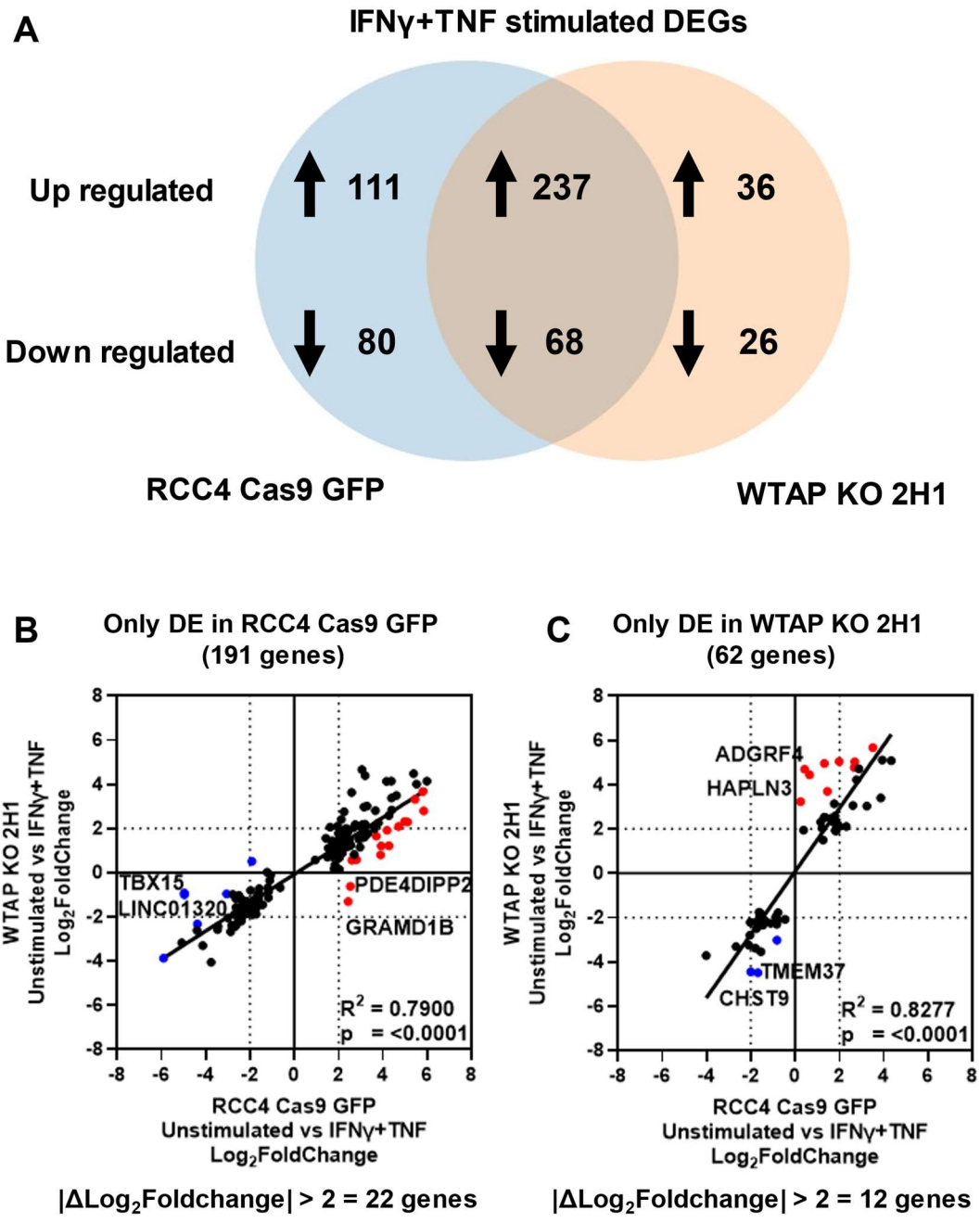

A
IFNγ+TNF stimulated DEGs
Up regulated
Down regulated
111
237
36
80
68
26
RCC4 Cas9 GFP
WTAP KO 2H1
B
Only DE in RCC4 Cas9 GFP
(191 genes)
WTAP KO 2H1
Unstimulated vs IFNγ+TNF
Log2FoldChange
TBX15
LINC01320
PDE4DIPP2
GRAMD1B
R2 = 0.7900
p = <0.0001
RCC4 Cas9 GFP
Unstimulated vs IFNγ+TNF
Log2FoldChange
|ΔLog2Foldchange| > 2 = 22 genes
C
Only DE in WTAP KO 2H1
(62 genes)
ADGRF4
HAPLN3
TMEM37
CHST9
R2 = 0.8277
p = <0.0001
|ΔLog2Foldchange| > 2 = 12 genes

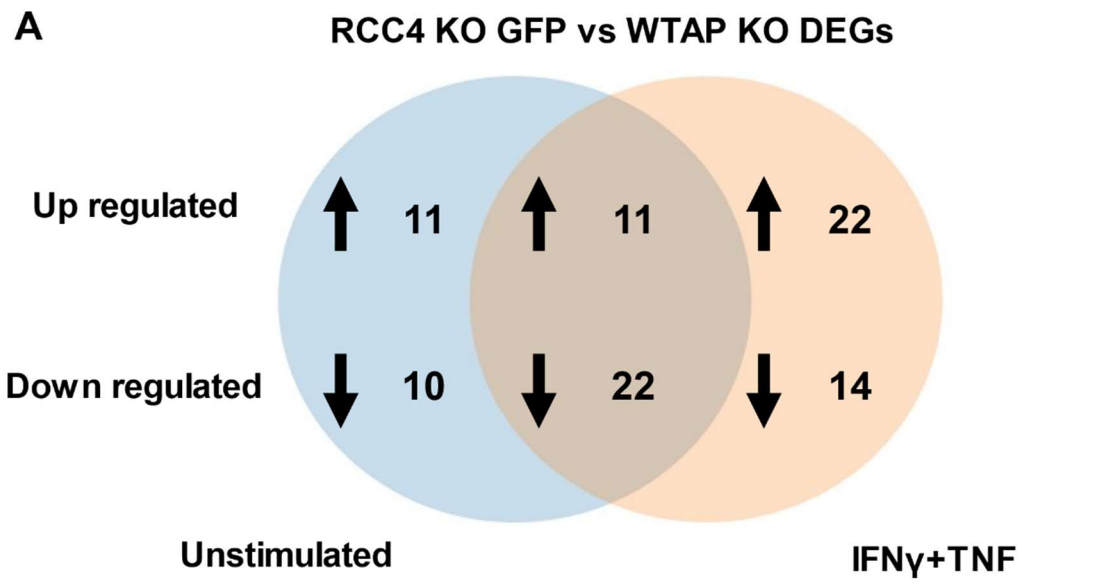
Figure 5.18: Differential IFNγ + TNF stimulation induced DEGs between RCC4 Cas9 GFP and WTAP KO 2H1

A) Venn diagram showing the numbers of IFNγ + TNF stimulation induced significantly upregulated and downregulated genes exclusively in RCC4 Cas9 GFP, exclusively in WTAP KO 2H1, and in both, using DEGs found from both reference genome and reference transcriptome alignment. **B)** Correlation of Log_2FoldChange of unstimulated vs IFNγ + TNF stimulated RCC4 Cas9 GFP exclusive DEGs ($p_{adj} \leq 0.1$, $|log_2FoldChange| > 2$), between RCC4 Cas9 GFP and WTAP KO 2H1. Log_2FoldChange averaged from both reference genome and transcriptome alignment data. Blue dots represent genes with ΔLog_2Foldchange (RCC4 Cas9 GFP – WTAP KO 2H1) < 2. Red dots represent genes with ΔLog_2Foldchange (RCC4 Cas9 GFP – WTAP KO 2H1) > 2. **C)** As in **B** but for WTAP KO 2H1 exclusive DEGs. 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.

5.3.13 Characterisation of WTAP KO-associated DEGs

Previously, differential gene expression analyses have identified 54 and 71 DEGs between RCC4 Cas9 GFP and WTAP KO 2H1, with no treatment and under IFNγ + TNF treatment, respectively (Figure 5.16C – D). To find out the extent of WTAP KO-associated differential gene expression patterns that were IFNγ + TNF stimulation dependent, a Venn diagram was generated using the lists of DEGs between RCC4 Cas9 GFP and WTAP KO 2H1, with and without IFNγ + TNF stimulation. 34 DEGs were identified with and without IFNγ + TNF treatment, representing 61.1% and 47.8% of all unstimulated and IFNγ + TNF stimulated DEGs, respectively (Figure 5.19A).

To obtain more information on the gene expression patterns of the treatment condition-specific WTAP KO-associated DEGs, Scatter plots were constructed using the treatment-specific DEGs' Log2FoldChange values of RCC4 Cas9 GFP vs WTAP KO 2H1, with and without IFNγ + TNF treatment (Figure 5.19B – C). Expression patterns were significantly correlated between untreated and IFNγ + TNF treated conditions, despite being highlighted as a DEG in only one of the conditions. Amongst the 57 condition-specific RCC4 Cas9 GFP vs WTAP KO 2H1 DEGs, ten genes exhibited at least a four-fold difference in gene expression levels between the treatment conditions. In addition, results here showed that WTAP KO induced significant differential gene expression in 47 genes, regardless of IFNγ + TNF treatment.


A
RCC4 KO GFP vs WTAP KO DEGs
Up regulated
Down regulated
11
11
22
10
22
14
Unstimulated
IFNγ+TNF

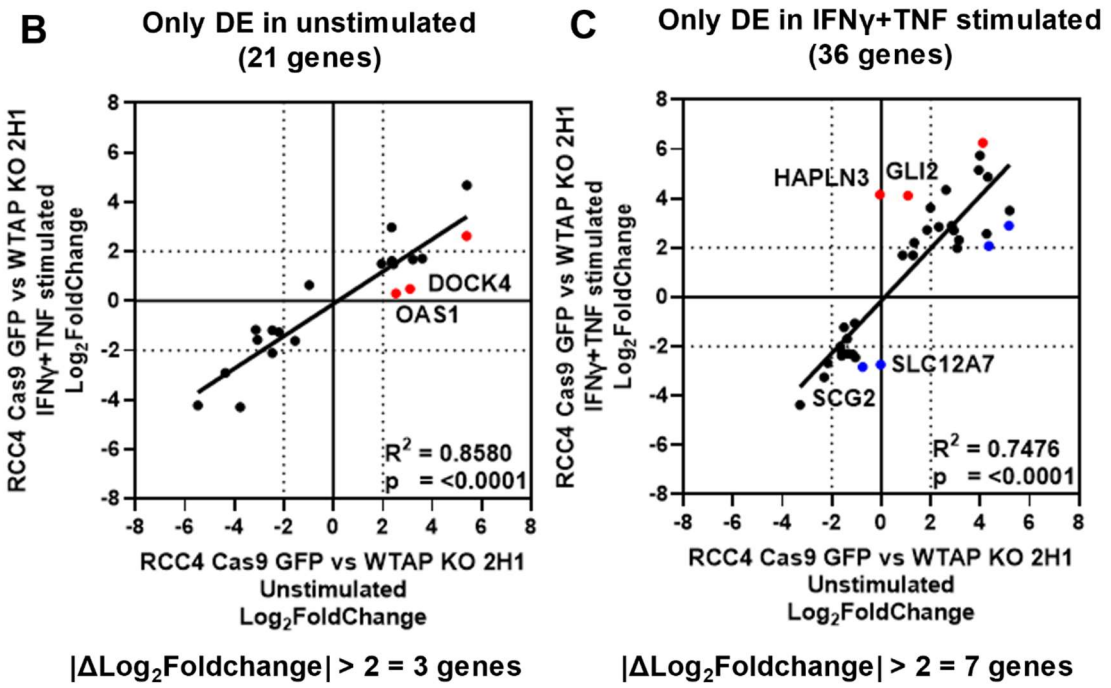

B
Only DE in unstimulated
(21 genes)
RCC4 Cas9 GFP vs WTAP KO 2H1
IFNγ+TNF stimulated
Log2FoldChange
RCC4 Cas9 GFP vs WTAP KO 2H1
Unstimulated
Log2FoldChange
DOCK4
OAS1
R2 = 0.8580
p = <0.0001
|ΔLog2Foldchange| > 2 = 3 genes
C
Only DE in IFNγ+TNF stimulated
(36 genes)
HAPLN3
GLI2
SLC12A7
SCG2
R2 = 0.7476
p = <0.0001
|ΔLog2Foldchange| > 2 = 7 genes

Figure 5.19: The majority of *WTAP* KO associated DEGs show similar expression trends with and without IFNγ + TNF exposure

A) Venn diagram showing the numbers of significantly upregulated and downregulated genes between RCC4 Cas9 GFP and WTAP KO 2H1 exclusively under no treatment, exclusively under IFNγ + TNF, and in both conditions. DEGs found from both reference genome and reference transcriptome alignment were used. **B)** Correlation of Log_2FoldChange of RCC4 Cas9 GFP vs WTAP KO 2H1 untreated-exclusive DEGs ($p_{adj} \leq 0.1$, $|log_2FoldChange| > 2$), between untreated and IFNγ + TNF treated condition. Log_2FoldChange averaged from both reference genome and transcriptome alignment data. Blue dots represent genes with ΔLog_2Foldchange (RCC4 Cas9 GFP – WTAP KO 2H1) < 2. Red dots represent genes with ΔLog_2Foldchange (RCC4 Cas9 GFP – WTAP KO 2H1) > 2. **C)** As in **B** but for IFNγ + TNF treated-exclusive DEGs. 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.

5.3.14 GSEA GO MF analysis reveals similar IFNγ + TNF induced and suppressed pathways in RCC4 Cas9 GFP and WTAP KO 2H1

Next, GSEA was conducted to assess the impact of IFNγ + TNF treatment on cellular pathways in RCC4 Cas9 GFP and WTAP KO 2H1 cells using clusterprofiler (v4.0). Here, log_2Foldchange values of DESeq2 normalised gene expression from all detected genes across samples were used as input for GSEA. GSEA with the GO: MF database showed the molecular pathways activated and suppressed in both RCC4 Cas9 GFP and WTAP KO 2H1 after IFNγ + TNF treatment. In both cell lines, enrichment maps demonstrated that most of the top 10 GO: MF terms were related to chemokine/cytokine receptor activities or the antigen presentation binding pathway (Figure 5.20A, 5.21A). The top 5 activated GO MF pathways in RCC4 Cas9 GFP by p_{adj} values were 'chemokine activity', 'peptide antigen binding', 'chemokine receptor binding', 'antigen binding' and 'cytokine receptor binding' (Figure 5.20B - C). Four GO: MF terms also appeared on the top five activated pathways list in WTAP KO 2H1, as demonstrated by both the dot plot and GSEA enrichment plot (Figure 5.21B – C).

GSEA GO MF analysis also revealed significantly suppressed pathways by IFNγ + TNF treatment. In RCC4 Cas9 GFP, the top 5 suppressed pathways were 'gated channel activity', 'ion channel activity', 'voltage-gated channel activity', 'voltage-gated ion channel activity' and 'potassium channel activity'. The downregulation in channel protein-related pathways was also observed in WTAP KO 2H1, where the top suppressed pathways were 'metal ion transmembrane transporter activity', 'voltage-gated anion channel activity', 'ion channel activity', and 'gated channel activity' (Figure 5.21B). Overall, GSEA GO: MF analysis showed similar profiles of activated and suppressed molecular pathways for RCC4 Cas9 GFP and WTAP KO 2H1 cells after IFNγ + TNF treatment. Comprehensive lists of significant GO: MF terms enriched in RCC4 Cas9 GFP and WTAP KO 2H1 upon IFNγ + TNF treatment can be found in Appendix tabled 7.23 and 7.25, respectively.

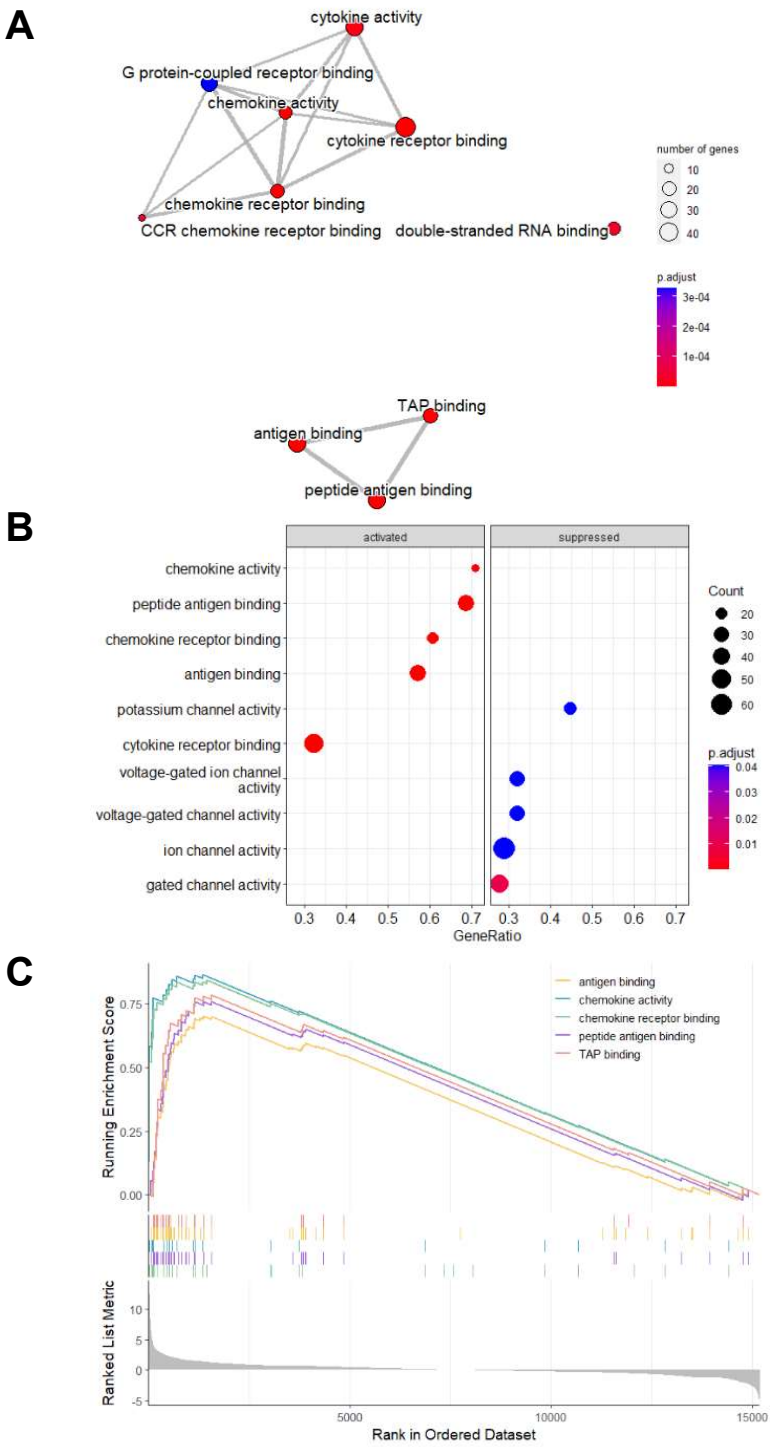


Figure 5.20: Gene Ontology Molecular Function (GO:MF) GSEA for IFNγ + TNF stimulation induced pathway enrichment in RCC4 Cas9 GFP

A) GO:MF enrichment map showing top 10 enriched terms associated IFNγ + TNF stimulation induced pathway enrichment in RCC4 Cas9 GFP. **B)** Dot plot showing top 10 enriched GO:MF terms, with dot size representing gene count per term and colour reflecting p_{adj} value. **C)** GSEA enrichment plot for the top 5 enriched GO:MF terms. The x-axis shows genes represented in each pathway, and the y-axis shows enrichment scores.

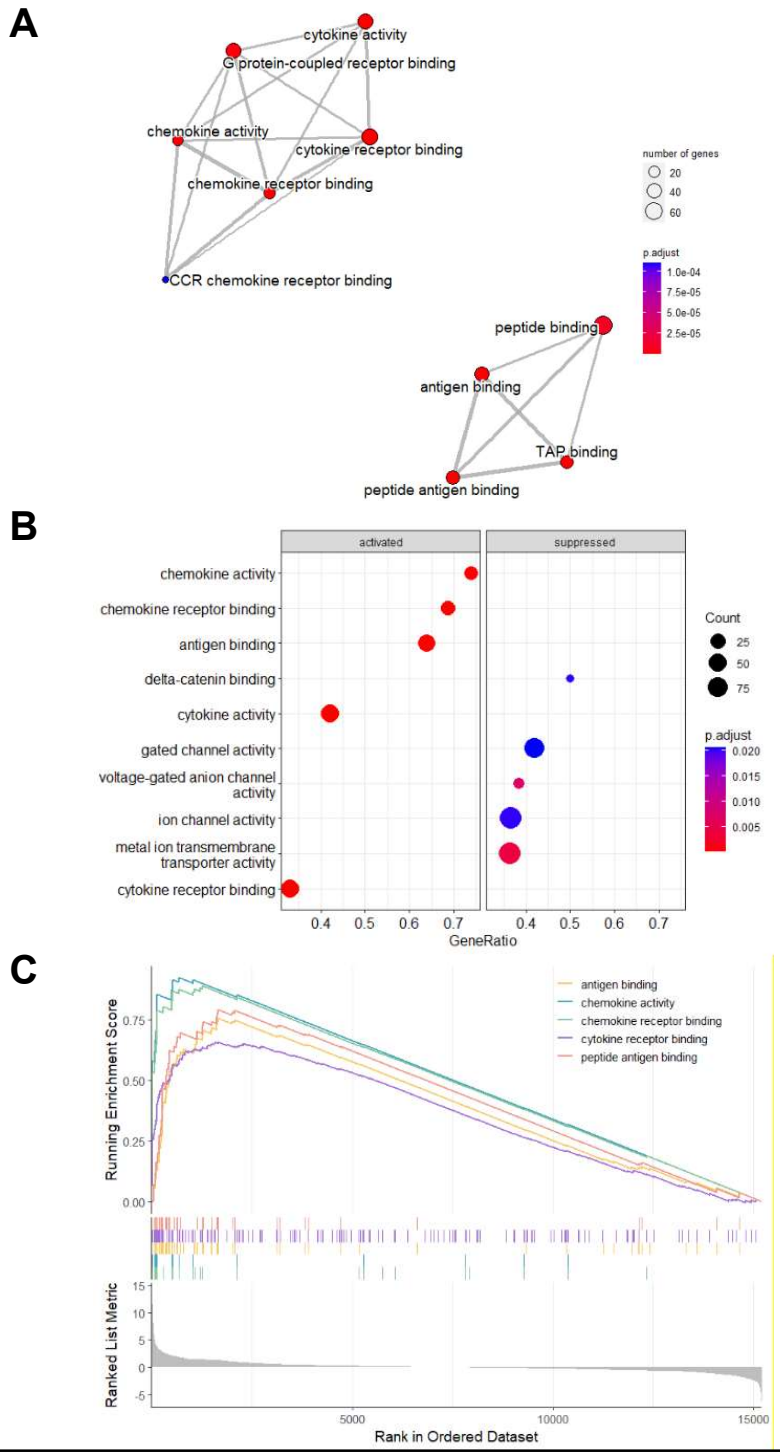


Figure 5.21: Gene Ontology Molecular Function (GO:MF) GSEA for IFNγ + TNF stimulation induced pathway enrichment in WTAP KO 2H1

A) GO:MF enrichment map showing top 10 enriched terms associated IFNγ + TNF stimulation induced pathway enrichment in WTAP KO 2H1. **B)** Dot plot showing top 10 enriched GO:MF terms, with dot size representing gene count per term and colour reflecting p_{adj} value. **C)** GSEA enrichment plot for the top 5 enriched GO:MF terms. The x-axis shows genes represented in each pathway, and the y-axis shows enrichment scores.

5.3.15 GSEA Hall mark geneset analysis identified suppressed glycolysis and hypoxic pathways in WTAP KO 2H1

To assess the molecular pathways that exhibited augmented expression with the KO of WTAP, Log_2FoldChange values from untreated and IFNγ + TNF treated RCC4 Cas9 GFP vs WTAP KO 2H1 were profiled by GSEA GO: BP and GO: MF databases (clusterprofiler v4.0). No significant GO term was identified.

Next, the GSEA of hallmark genesets (extracted from the molecular signatures database (MSigDb)) was performed using fgsea (Korotkevich and Sukhov, 2016). For untreated RCC4 Cas9 GFP vs WTAP KO 2H1, two hallmark genesets, ' Hypoxia' and 'Glycolysis', were identified to be significantly suppressed in WTAP KO 2H1 (Figure 5.22A, 5.22C). Similarly, for IFNγ + TNF treated RCC4 Cas9 GFP vs WTAP KO 2H1, 'Hypoxia' and 'Glycolysis' were the only hallmark gene sets that were significantly suppressed. (Figure 5.22B, 5.22D).

RCC4 is a *VHL*-defective ccRCC cell line with a constitutively active HIF pathway with high protein expression levels of HIF-1α (Ruf *et al.*, 2016). In the context of cancer, HIF-1α mediates transcription activation of multiple oncogenic pathways, such as the PI3K-Akt-mTOR pathway (Masoud and Li, 2015). As a well-studied transcription factor, HIF-1α target genes have been characterised and validated experimentally (Benita *et al.*, 2009). Here, Venn diagrams showed that amongst the DEGs from both untreated and IFNγ + TNF treated RCC4 Cas9 GFP vs WTAP KO 2H1, seven and nine DEGs, respectively, were also targets of HIF-1α (Figure 5.22E, F). Some of the overlapping genes include *BNIP3*, *CA9* and *EGLN3*.

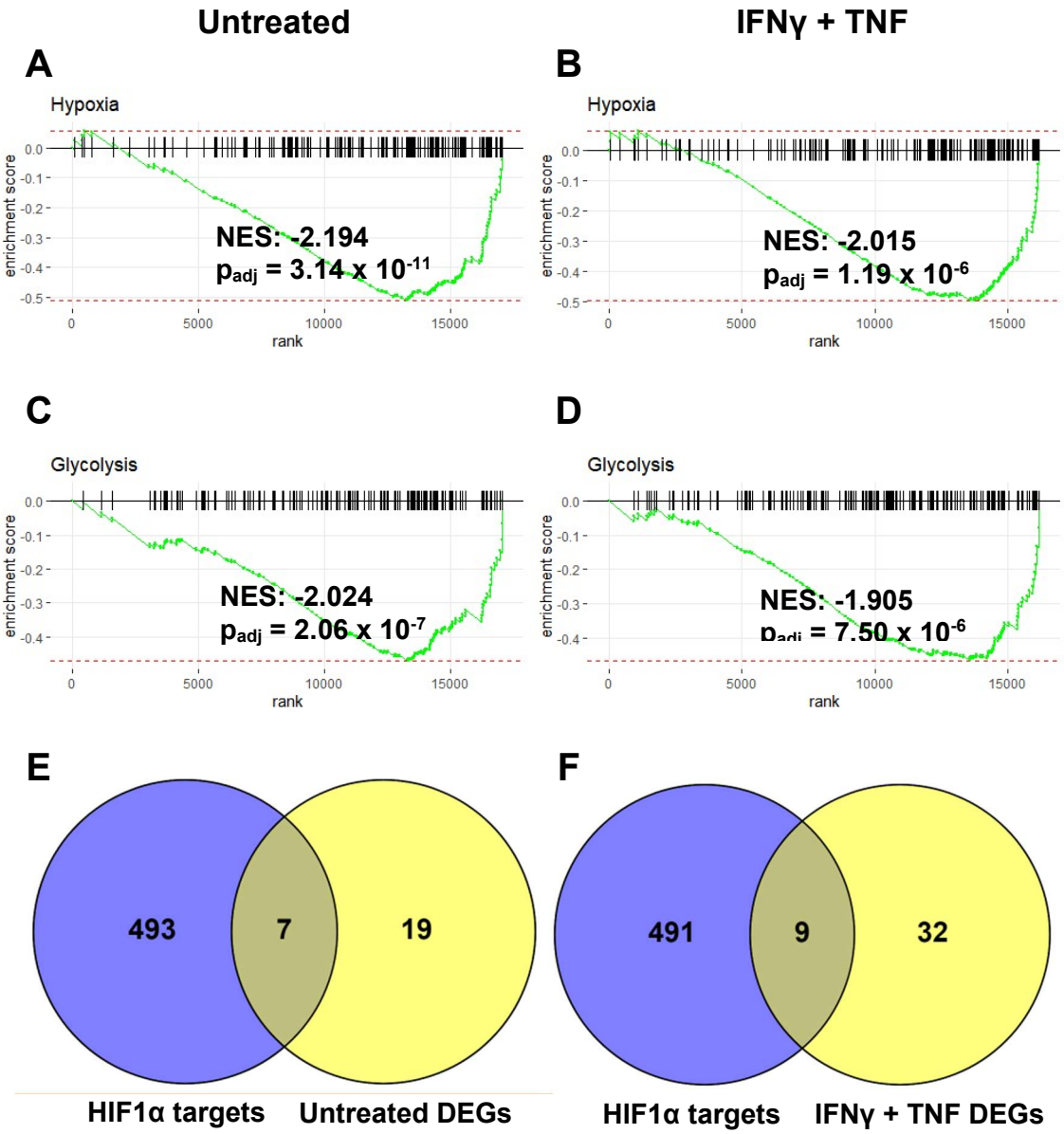


Figure 5.22: WTAP KO 2H1 exhibit suppressed hypoxia & glycolysis pathways

GSEA enrichment plots of the hall mark gene set 'Hypoxia' for differential gene expression between RCC4 Cas9 GFP and WTAP KO 2H1 when **A)** untreated and **B)** under IFNγ + TNF treatment. **C)** As in **A** and **D)** as in **B**, but for the hall mark gene set 'Glycolysis.The x-axis shows genes represented in each pathway, and the y-axis shows enrichment scores. Normalised enrichment scores (NES) and p_{adj} values are indicated in each plot. **E)** Venn diagram showing overlaps of untreated RCC4 Cas9 GFP vs WTAP KO 2H1 DEGs and HIF1 target genes (n = 500, Benita *et al.* 2009). **F)** As in **E** but for IFNγ + TNF treated RCC4 CAs9 GFP vs WTAP KO 2H1 DEGs.

5.3.16 Validation of DRS identified DEGs between RCC4 Cas9 GFP and WTAP KO 2H1

To confirm DRS gene expression results, mRNA and protein levels of significant DEGs from RCC4 Cas9 GFP and WTAP KO clonal cell lines (2E6, 2H1 and 3A1) were assessed. Transcript levels of *NDUFA4L2* and *BNIP3* were measured using qRT-PCR. Previously, *NDUFA4L2* (and *BNIP3* were identified as two of the most significantly downregulated genes in both untreated and IFNγ + TNF treated RCC4 Cas9 GFP vs WTAP KO 2H1 (Tables 5.4 – 5.5). Using qRT-PCR, both *NDUFA4L2* and *BNIP3* mRNA levels were found to be significantly suppressed in WTAP KO 2E6 and WTAP KO 2H1, under both untreated and IFNγ + TNF treated conditions (Figure 5.23A – D). These results validated gene expression analysis from DRS. However, in contrast to the other two WTAP KO clonal cell lines, mRNA expression levels of *NDUFA4L2* in WTAP KO 3A1 were found to be significantly upregulated compared to RCC4 Cas9 GFP in both untreated and IFNγ + TNF treated conditions (Figure 5.23A – B). Moreover, no significant suppression of *BNIP3* mRNA levels was observed (Figure 5.23C – D).

Next, protein expression levels of *CA9* and *NDUFA4L2* were examined using western blotting analysis. Like *NDUFA4L2*, *CA9* was one of the most highly suppressed genes identified using DRS (Tables 5.4 – 5.5). Similar to the observations from qRT-PCR experiments, western blotting analysis revealed downregulation in protein expression of CA9 and NDUFA4L2 in WTAP KO 2E6 and WTAP KO 2H1 cell lines compared to RCC4 Cas9 GFP. However, no suppression in CA9 and NDUFA4L2 expression was found in the WTAP KO 3A1 cell line (Figure 5.23E). Although both qRT-PCR and western blotting analysis validated the DRS gene expression analysis between RCC4 Cas9 GFP and WTAP KO 2H1, the inconsistencies in validation experiments across WTAP KO cell lines indicated potential off-target effects from the gene editing process.

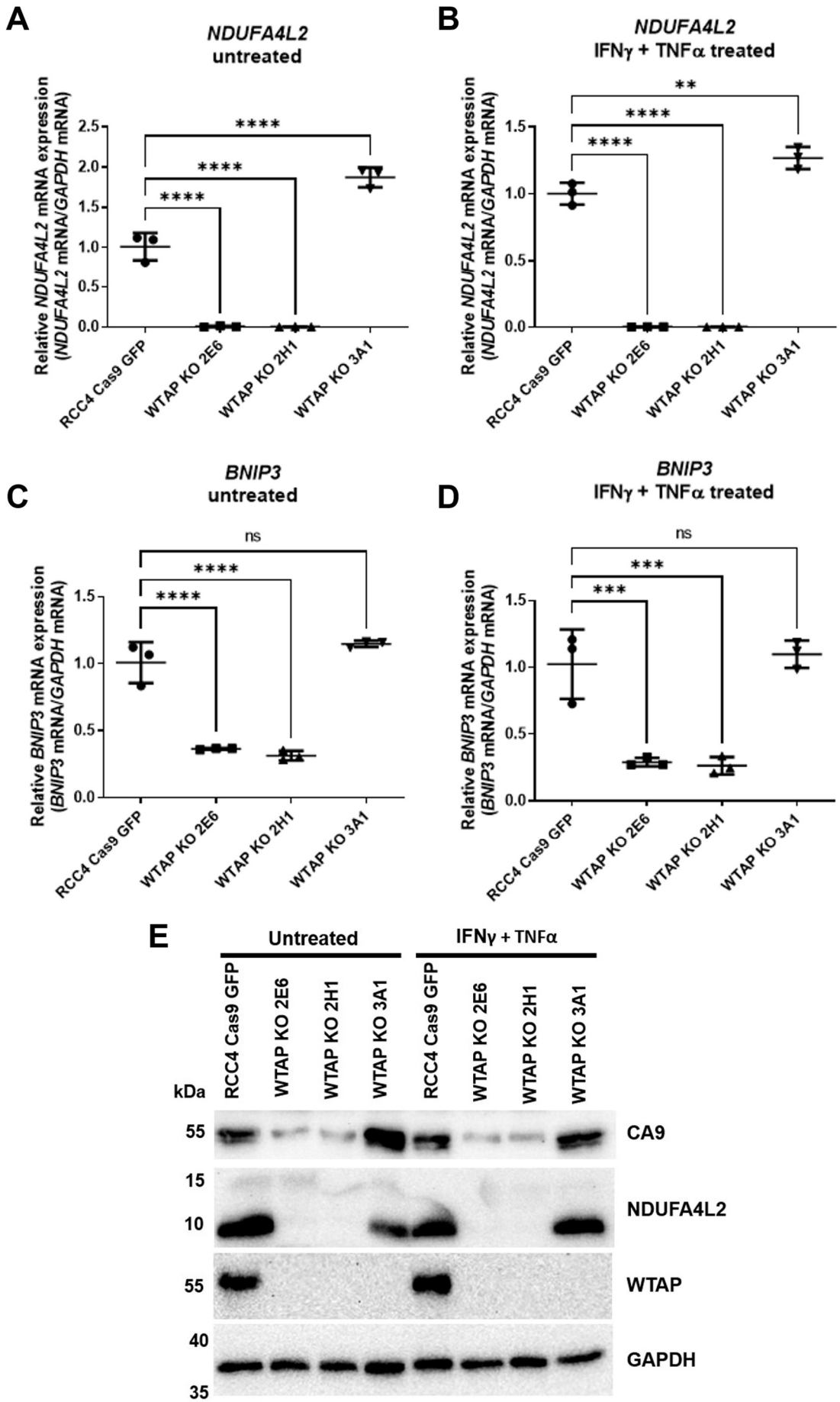

A
NDUFA4L2
untreated
Relative NDUFA4L2 mRNA expression
(NDUFA4L2 mRNA/GAPDH mRNA)
2.5
2.0
1.5
1.0
0.5
0.0
****
****
****
RCC4 Cas9 GFP
WTAP KO 2E6
WTAP KO 2H1
WTAP KO 3A1
B
NDUFA4L2
IFNγ + TNFα treated
Relative NDUFA4L2 mRNA expression
(NDUFA4L2 mRNA/GAPDH mRNA)
1.5
1.0
0.5
0.0
**
****
****
RCC4 Cas9 GFP
WTAP KO 2E6
WTAP KO 2H1
WTAP KO 3A1
C
BNIP3
untreated
Relative BNIP3 mRNA expression
(BNIP3 mRNA/GAPDH mRNA)
1.5
1.0
0.5
0.0
ns
****
****
RCC4 Cas9 GFP
WTAP KO 2E6
WTAP KO 2H1
WTAP KO 3A1
D
BNIP3
IFNγ + TNFα treated
Relative BNIP3 mRNA expression
(BNIP3 mRNA/GAPDH mRNA)
1.5
1.0
0.5
0.0
ns
***
***
RCC4 Cas9 GFP
WTAP KO 2E6
WTAP KO 2H1
WTAP KO 3A1
E
Untreated
IFNγ + TNFα
RCC4 Cas9 GFP
WTAP KO 2E6
WTAP KO 2H1
WTAP KO 3A1
RCC4 Cas9 GFP
WTAP KO 2E6
WTAP KO 2H1
WTAP KO 3A1
kDa
55
CA9
15
10
NDUFA4L2
55
WTAP
40
35
GAPDH

Figure 5.23: Validation of RNAseq results via qRT-PCR and western blotting

A) *NDUFA4L2* mRNA levels measured by qRT-PCR in untreated RCC4 Cas9 GFP, WTAP KO 2E6, 2H1, 3A1, and RCC4 cell lines, relative to averaged *NDUFA4L2* mRNA levels in RCC4 Cas9 GFP. *NDUFA4L2* mRNA levels were normalised to *GAPDH*. **B)** As in **A,** but for IFNγ + TNF treated (24 hours) cells. **C)** *BNIP3* mRNA levels measured by qRT-PCR in untreated RCC4 Cas9 GFP, WTAP KO 2E6, 2H1, 3A1, and RCC4 cell lines (n = 3), relative to averaged *BNIP3* mRNA levels in RCC4 Cas9 GFP. *BNIP3* mRNA levels were normalised *GAPDH*. **D)** As in **C**, but for IFNγ + TNF treated (24 hours) cells. **E)** Western blot analysis of CA9, NDUFA4L2, WTAP and GAPDH (loading control) in untreated and in IFNγ and TNFα stimulated (24 hours) RCC4 Cas9 GFP, WTAP KO 2E6, WTAP KO 2H1 and WTAP KO 3A1 cell lines (n = 1). For **A** – **D**, one-way ANOVA tests were performed, with $p \leq 0.05$ considered statistically significant. Asterisks indicate statistical significance levels (**** = $p < 0.0001$, *** = $p < 0.001$, ** = $p < 0.01$, ns = $p > 0.05$, not significant).

5.3.17 Orthogonal validation of *WTAP* gene knockout associated DEGs by siRNA-mediated knockdown of m^{6}A writers

Having observed inconsistent effects of *WTAP* gene knockout in the clonal KO cell lines, siRNA-mediated gene silencing was used to validate the impact of m^{6}A writers on the gene expression of DRS-identified *WTAP* KO-associated DEGs. siRNAs (pool of four) against *WTAP*, *METTL3* and non-targeting control (NTC) were transfected into RCC4 cells, with and without subsequent 24 hours IFNγ + TNF treatment (30 hours post-transfection).

Transfection of siRNAs against *WTAP* (siWTAP) and *METTL3* (siMETTL3) resulted in depletion of their respective mRNA levels in RCC4 cells, compared to siNTC transfected cells in both unstimulated and IFNγ + TNF stimulated cells (n = 2, Figure 5.23A – B, 5.23E – F). Depletion of *WTAP* mRNA expression levels in RCC4 cells did not result in changes in the *METTL3* mRNA levels (n = 2, Figure 5.23C – D). Similarly, no compensatory effect on *WTAP* expression was observed in siMETTL3 transfected RCC4 cells (n = 2, Figure 5.23G – H).

In both siWTAP and siMETTL3 transfected RCC4 cells, *NDUFA4L2* mRNA levels were not found to be significantly suppressed in both untreated and IFNγ + TNF treated conditions, compared to siNTC transfected controls (Figure 5.24I – J). Similarly, neither the transfection of siWTAP nor siMETTL3 resulted in changes in *BNIP3* mRNA levels, compared to siNTC transfected controls (Figure 5.24M – P). Results here did not support the effects shown in DRS results, where KO of *WTAP* significantly suppressed *NDUFA4L2* and *BNIP3* mRNA levels.

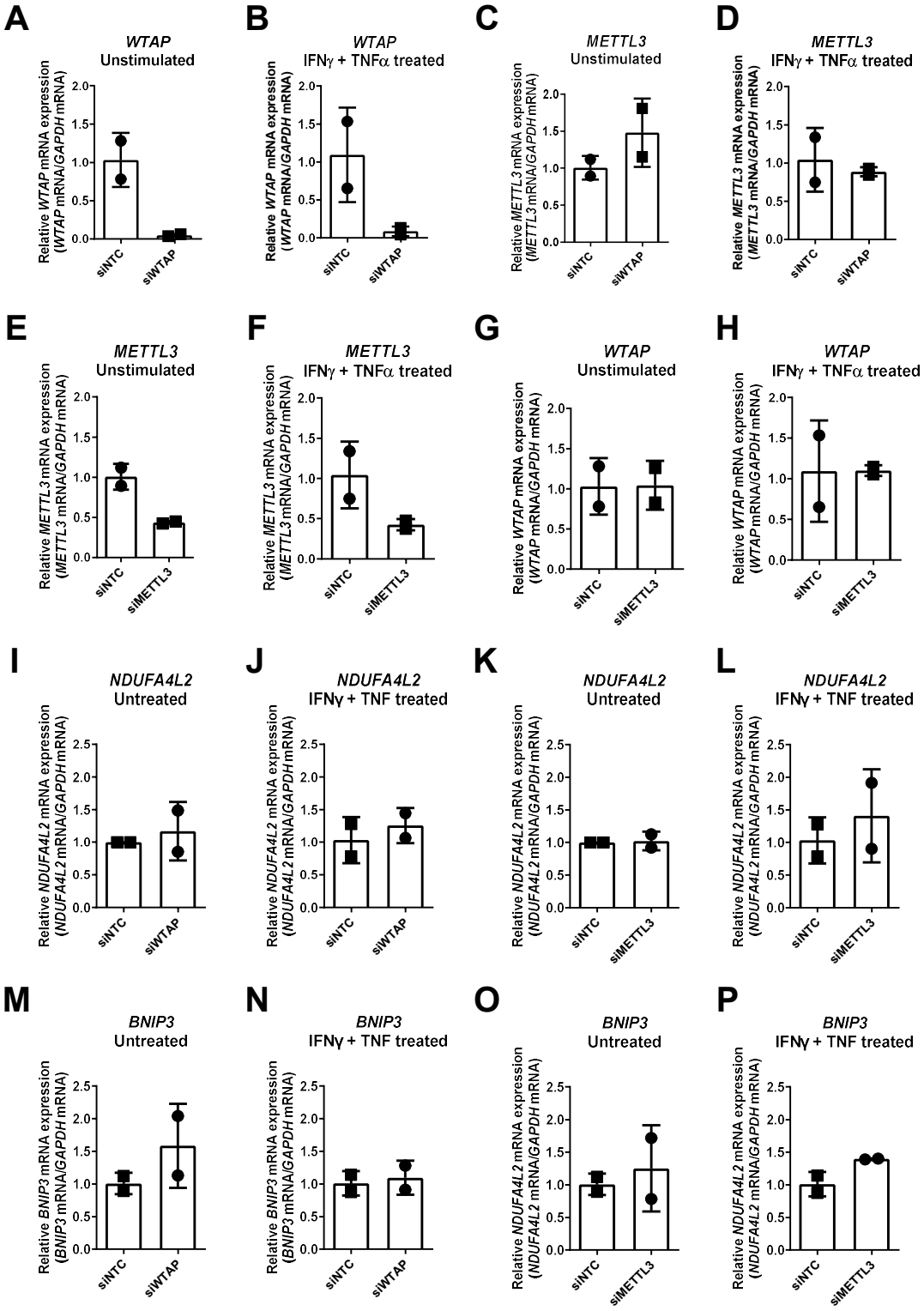

A
WTAP
Unstimulated
Relative WTAP mRNA expression (WTAP mRNA/GAPDH mRNA)
siNTC
siWTAP
B
WTAP
IFNγ + TNFα treated
C
METTL3
Unstimulated
siMETTL3
D
METTL3
IFNγ + TNFα treated
E
METTL3
Unstimulated
F
METTL3
IFNγ + TNFα treated
G
WTAP
Unstimulated
H
WTAP
IFNγ + TNFα treated
I
NDUFA4L2
Untreated
Relative NDUFA4L2 mRNA expression (NDUFA4L2 mRNA/GAPDH mRNA)
J
NDUFA4L2
IFNγ + TNF treated
K
NDUFA4L2
Untreated
L
NDUFA4L2
IFNγ + TNF treated
M
BNIP3
Untreated
Relative BNIP3 mRNA expression (BNIP3 mRNA/GAPDH mRNA)
N
BNIP3
IFNγ + TNF treated
O
BNIP3
Untreated
P
BNIP3
IFNγ + TNF treated

Figure 5.23: Orthogonal validation of *WTAP* gene knockout associated differential gene expression via siRNA-mediated knockdown of m^{6}A writers

A) *WTAP* mRNA levels measured by qRT-PCR in unstimulated siNTC (non-targeting control) and siWTAP transfected RCC4 cells, relative to averaged *WTAP* mRNA levels in siNTC transfected RCC4 cells from 2 independent experiments. **B)** As in **A**, but for IFNγ + TNF treated cells. **C)** *METTL3* mRNA levels measured by qRT-PCR in unstimulated siNTC and siWTAP transfected RCC4 cells, relative to averaged *METTL3* mRNA levels in siNTC transfected RCC4 cells from 2 independent experiments. **D)** As in **C**, but for IFNγ + TNF treated cells. **E)** *METTL3* mRNA levels measured by qRT-PCR in unstimulated siNTC and siMETTL3 transfected RCC4 cells, relative to averaged *METTL3* mRNA levels in siNTC transfected RCC4 cells from 2 independent experiments. **F)** As in **E**, but for IFNγ + TNF treated cells. **G)** *WTAP* mRNA levels measured by qRT-PCR in unstimulated siNTC and siMETTL3 transfected RCC4 cells, relative to averaged *METTL3* mRNA levels in siNTC transfected RCC4 cells from 2 independent experiments. **H)** As in **G** but for IFNγ + TNF treated cells. **I)** *NDUFA4L2* mRNA levels measured by qRT-PCR in untreated siNTC and siWTAP transfected cells, relative to averaged *NDUFA4L2* mRNA levels in siNTC transfected RCC4 cells from 2 independent experiments. **J)** As in **I**, but for IFNγ + TNF treated cells. **K)** As in **I**, and **L)** as in **J**, but with siNTC and siMETTL3 transfected RCC4 cells. M) *BNIP3* mRNA levels measured by qRT-PCR in untreated siNTC and siWTAP transfected cells, relative to averaged *BNIP3* mRNA levels in siNTC transfected RCC4 cells from 2 independent experiments. **N)** As in **M**, but for IFNγ + TNF treated cells. **O)** As in **M**, and **P)** as in **N**, but with siNTC and siMETTL3 transfected RCC4 cells. Throughout, mRNA levels were normalised to *GAPDH*. Centre lines represent mean for each group. Error bars represent standard deviation.

5.3.18 Identification of DTU events between unstimulated and IFNγ+TNF treated RCC4 Cas9 GFP and WTAP KO 2H1

The effects of IFNγ + TNF treatment and WTAP-KO in ccRCC tumour cells were further assessed through differential transcript usage analysis using DRIMseq and DEXseq (Love *et al.*, 2018). Comparing stimulated and IFNγ + TNF stimulated RCC4 Cas9 GFP, DRIMseq identified 50 genes that underwent DTU, whereas DEXseq found 53 genes which displayed DTU (Figure 5.25A). For WTAP KO 2H1, comparing stimulated and IFNγ + TNF stimulated cells, DRIMseq found seven genes which displayed DTU, and DEXseq identified 43 genes which showed DTU (Figure 5.25B). 61 and 45 unique genes were found to display DTU in RCC4 Cas9 GFP and WTAP KO 2H1, respectively, after IFNγ + TNF treatment. Among the DTUs identified, 13 genes overlap between the two cell lines (Figure 5.25 C).

DTU events were also identified between RCC4 Cas9 GFP and WTAP KO 2H1. In unstimulated RCC4 Cas9 GFP vs WTAP KO 2H1, 69 genes showed significant DTU. Under IFNγ + TNF treatment, 32 genes were found to display DTU. Seven genes were found to display DTU regardless of IFNγ + TNF treatment condition (Figure 5.25D). Comprehensive lists of DTU genes between unstimulated and IFNγ + TNF stimulated RCC4 Cas9 GFP, unstimulated and IFNγ + TNF stimulated WTAP KO 2H1, and between the cell lines at both conditions can be found in the Appendix Tables 7.26 – 29.

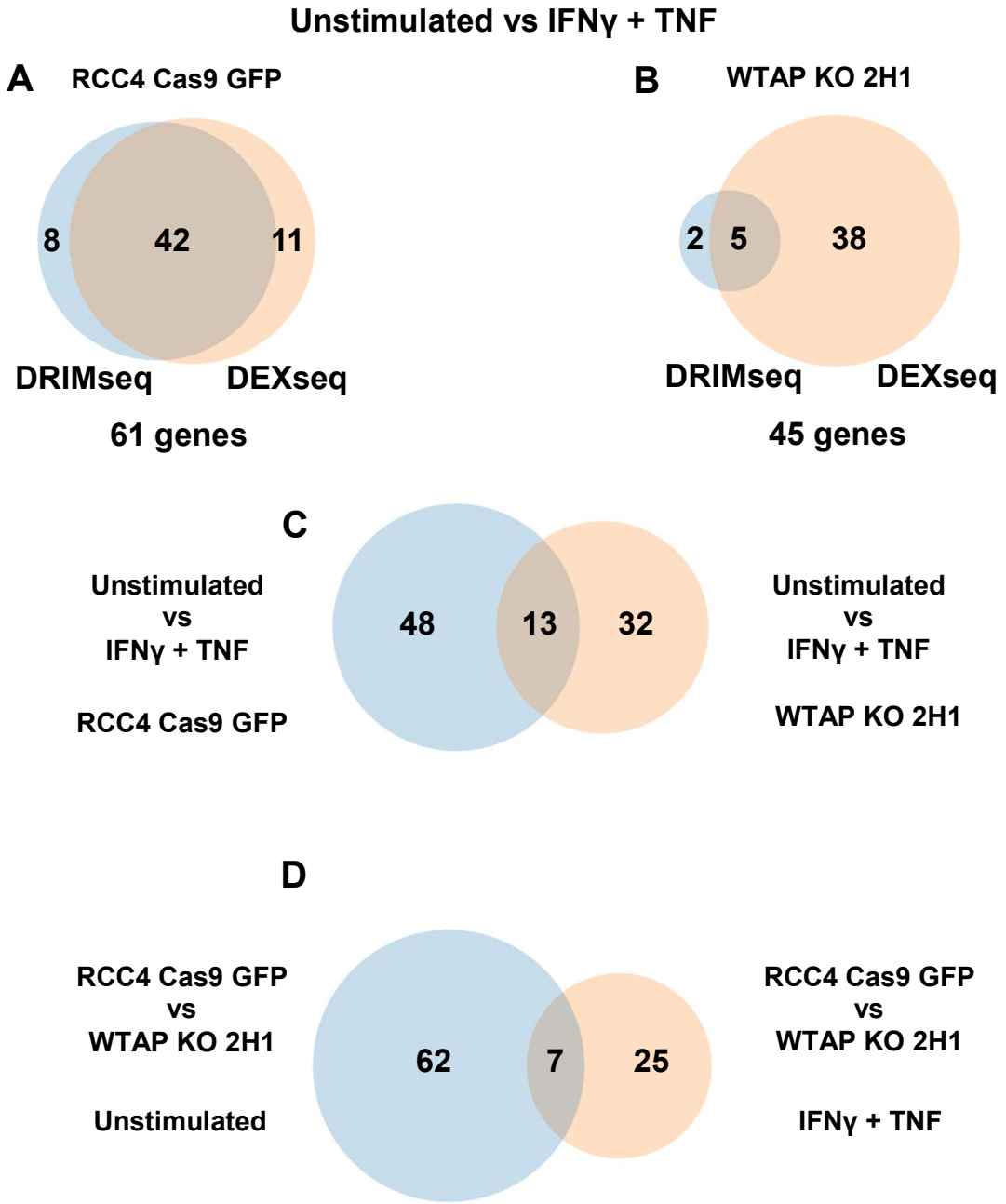


Figure 5.25: Characterisation of IFNγ + TNF induced DTU events in RCC4 Cas9 GFP and WTAP KO 2H1 cells

A) Venn diagram showing the overlap between DRIMseq and DEXseq identified genes that displayed significant DTU between untreated and IFNγ + TNF treated RCC4 Cas9 GFP cells. **B)** As in **A**, but for WTAP KO 2H1 cells. **C)** Venn diagram showing the overlap between IFNγ + TNF treatment induced-DTU genes identified in RCC4 Cas9 GFP and WTAP KO 2H1. **D)** Venn diagram showing the overlap of significant DTU genes between unstimulated and IFNγ + TNF stimulated RCC4 Cas9 GFP vs WTAP KO 2H1.

5.3.19 *CD24* displays DTU in RCC4 Cas9 GFP and WTAP KO 2H1 after IFNγ and TNF stimulation

CD24 was one of the genes which displayed DTU in both RCC4 Cas9 GFP and WTAP KO 2H1 when stimulated with IFNγ + TNF (Figure 5.26J). Two transcripts were mapped in RCC4 Cas9 GFP: ENST00000619133 and ENST00000606017. Both transcripts share the same coding sequence, with ENST00000619133 having an extended 5'UTR sequence (Figure 5.26A). IGV coverage tracks of RCC4 Cas9 GFP and WTAP KO 2H1 in both unstimulated and IFNγ+TNF stimulated conditions confirmed coverage of the extended 5'UTR, and upon visual inspection, increased representations of the extended 5'UTR were observed in both RCC4 Cas9 GFP and WTAP KO 2H1 (Figure 5.26B – C). Moreover, *CD24* transcripts with the characteristically short 3'UTR previously seen in the ccRCC tumours (by both DRS and PCS, Figure 4.31B – C) were also found here, as highlighted by the arrow (Figure 5.26C).

At the gene level, nonsignificant downregulation trends in *CD24* expression levels were observed between unstimulated and IFNγ + TNF stimulated RCC4 Cas9 GFP, with both reference transcriptome alignments (median TPM: 925.7 vs 187.5) and reference genome alignments (median TPM: 2711 vs 467.9) (Figure 5.26D – E). Reference transcriptome data suggested that the decrease in *CD24* levels upon IFNγ + TNF treatment was primarily driven by the reduction in the ENST00000606017 (short 5'UTR transcript) (median TPM: 819.4 vs 105.2), whereas ENST0000619133 (long 5'UTR transcript) only experienced a modest reduction (median TPM: 106.3 vs 82.3) (Figure 5.26F – G). From the reference transcriptome aligned data, the proportion of the long 5'UTR transcript in RCC4 Cas9 GFP increased from an average of 14.86% in unstimulated samples to 42.15% in IFNγ + TNF treated cells (Figure 5.26J). Both DRIMseq (p_{adj} = 0.0043) and DEXseq (p_{adj} = 0.053) identified the differential transcript usage of *CD24* after IFNγ + TNF treatment in RCC4 Cas9 GFP. In WTAP KO 2H1 cells, DTU of *CD24* upon IFNγ + TNF treatment was recognised by DEXseq (p_{adj} = 0.0027) but not DRIMseq (p_{adj} = 0.754).

Similar trends in the increased proportion of the long 5'UTR *CD24* transcript after IFNγ + TNF stimulation in RCC4 Cas9 GFP and WTAP were found using reference genome alignment, as shown by the coverage tracks in Figure 5.26B – C. Since reference genome alignment only provides gene-level counts, the levels of short 5'UTR (ENST00000606017) and long 5'UTR (ENST0000061913) *CD24* transcripts were calculated based on *CD24* reads with overlaps with the extended 5'UTR sequence. No distinction between 3'UTR structures was taken into account here. After IFNγ + TNF treatment, an 82% decrease in levels of short 5'UTR transcript (median TPM: 398.8 vs 73.2) was identified, whereas long 5'UTR transcript (ENST0000061913) expression levels reduced by 54% on average (median TPM: 19.72 vs 9.08) (Figure 5.26H – I). However, the transcripts' proportions at unstimulated and IFNγ + TNF stimulated conditions were vastly different here compared to reference transcriptome-aligned data. At unstimulated conditions, reference genome alignment showed that 95.28% of CD24 transcripts were short 5'UTR isoforms on average. Only 4.72% of the transcripts displayed the extended 5'UTR. After IFNγ + TNF stimulation, on average, 89.8% of *CD24* transcripts in RCC4 Cas9 GFP represented the short 5'UTR *CD24* isoforms, whereas 10.2% of the *CD24* were the long 5'UTR isoforms (Figure 5.26K). Data here showed that although the DTU trend remained, the proportions of *CD24* transcripts assigned to specific isoforms via reference transcriptome alignment were not reciprocated from reference genome-aligned data.

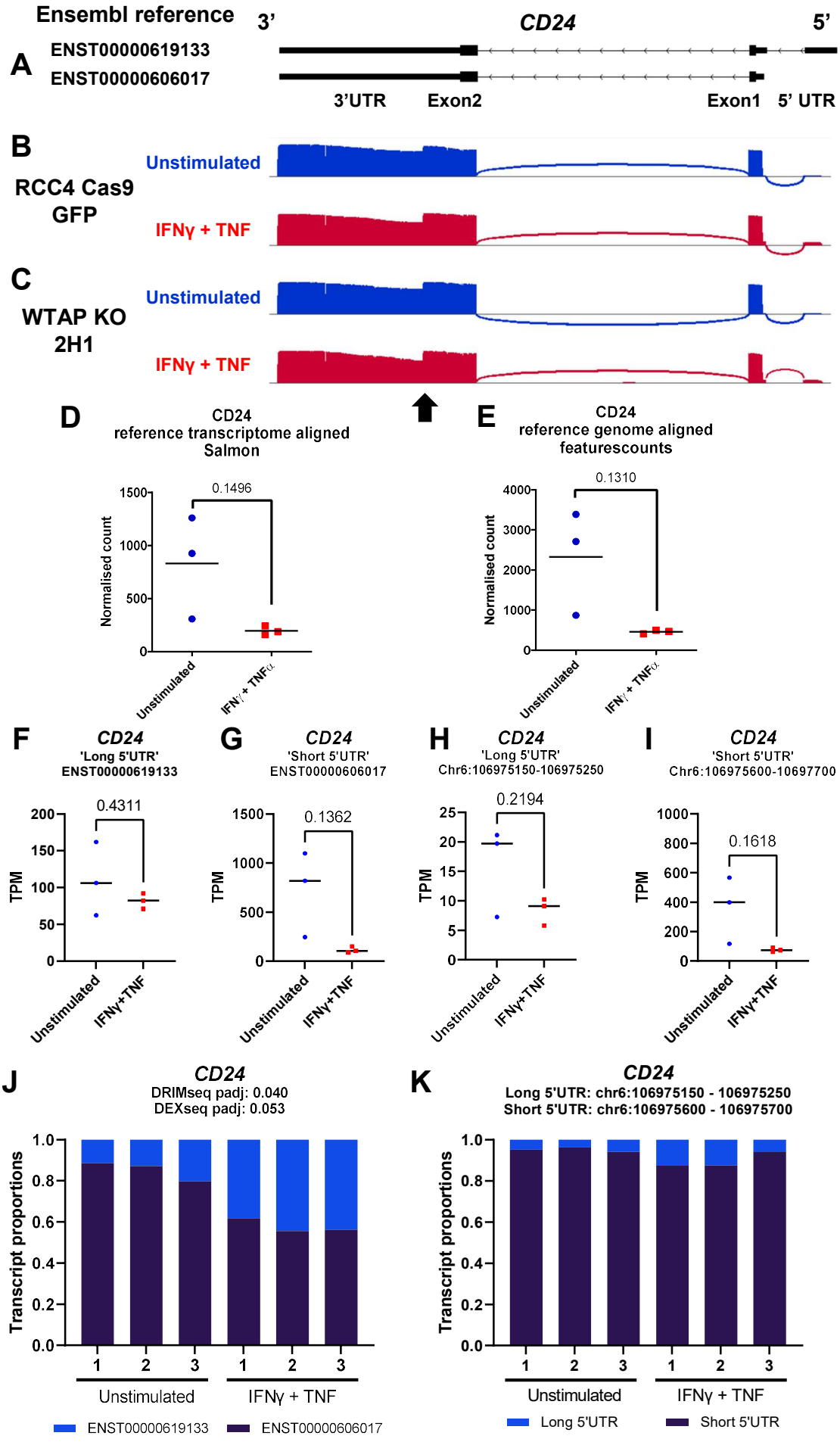

A
Ensembl reference
ENST00000619133
ENST00000606017
3'
CD24
5'
3'UTR
Exon2
Exon1
5' UTR
B
RCC4 Cas9 GFP
Unstimulated
IFNγ + TNF
C
WTAP KO 2H1
Unstimulated
IFNγ + TNF
D
CD24
reference transcriptome aligned
Salmon
Normalised count
0.1496
Unstimulated
IFNγ + TNFα
E
CD24
reference genome aligned
featurescounts
Normalised count
0.1310
Unstimulated
IFNγ + TNFα
F
CD24
'Long 5'UTR'
ENST00000619133
TPM
0.4311
Unstimulated
IFNγ+TNF
G
CD24
'Short 5'UTR'
ENST00000606017
TPM
0.1362
Unstimulated
IFNγ+TNF
H
CD24
'Long 5'UTR'
Chr6:106975150-106975250
TPM
0.2194
Unstimulated
IFNγ+TNF
I
CD24
'Short 5'UTR'
Chr6:106975600-10697700
TPM
0.1618
Unstimulated
IFNγ+TNF
J
CD24
DRIMseq padj: 0.040
DEXseq padj: 0.053
Transcript proportions
Unstimulated
IFNγ + TNF
ENST00000619133
ENST00000606017
K
CD24
Long 5'UTR: chr6:106975150 - 106975250
Short 5'UTR: chr6:106975600 - 106975700
Transcript proportions
Unstimulated
IFNγ + TNF
Long 5'UTR
Short 5'UTR

Figure 5.26: IFNγ and TNF exposure induces DTU of CD24 in RCC4 Cas9 GFP and WTAP KO 2H1 cells

A) Graphical representation of *CD24* transcripts ENST00000619133 and ENST00000606017 from Ensembl reference annotation (GRCh38). **B)** IGV visualisation of combined unstimulated (blue) and IFNγ + TNF treated (red) RCC4 Cas9 GFP reads coverage track in the region of the *CD24* gene. **C)** As in **B**, but for WTAP KO 2H1. Black arrow indicates previously identified novel 3'UTR. **D)** Grouped dot plot showing reference transcriptome aligned DESeq2 normalised *CD24* expression in untreated (blue) and IFNγ + TNF treated (red) RCC4 Cas9 GFP. **E)** as in **D** but with reference genome aligned data. **F)** Grouped dot plot showing reference transcriptome aligned expression (transcripts per million (TPM)) of ENST0000619133 (*CD24* long 5'UTR) in untreated (blue) and IFNγ + TNF treated (red) RCC4 Cas9 GFP. **G)** Grouped dot plot showing reference transcriptome aligned expression (TPM) of ENST0000606017 (*CD24* short 5'UTR). **H)** Grouped dot plot showing reference genome aligned expression (TPM) of *CD24* long 5'UTR transcript (defined by reads that aligned to 6:106,975,150 – 106,975,250) in untreated (blue) and IFNγ + TNF treated (red) RCC4 Cas9 GFP. **I)** Grouped dot plot showing reference genome aligned expression (TPM) of *CD24* short 5'UTR transcript (defined by reads that aligned to 6:106,974,600 – 106,974,700 and not a *CD24* long 5' UTR transcript) in untreated (blue) and IFNγ + TNF treated (red) RCC4 Cas9 GFP. **J)** Stacked bar graphs representing proportions of *CD24* isoforms in reference transcriptome mapped, untreated and IFNγ + TNF treated RCC4 Cas9 GFP cells. DRIMseq and DEXseq p_{adj} values for DTU of *CD24* are indicated in graph, with $p \leq 0.1$ considered significant. **K)** Stacked bar graphs representing proportions of *CD24* isoforms in reference genome mapped, untreated and IFNγ + TNF treated RCC4 Cas9 GFP cells. *CD24* long 5'UTR transcripts were defined by reads that aligned to 6:106,975,150 – 106,975,250. *CD24* short 5'UTR transcripts were defined by reads that aligned to 6:106,974,600 – 106,974,700 and not a *CD24* long 5' UTR transcript. For **D** – **I**, two-tailed unpaired T-tests with Welch's correction were used, with $p \leq 0.05$ considered significant. P values of non-significant results are indicated in graphs. Centre line represents median for each group.

5.3.20 IFNγ and TNF stimulation preferentially upregulates membrane *PD-L1* transcripts

Aside from CD24, IFNγ and TNF stimulation also induced DTU in the isoforms of the immune checkpoint *PD-L1 (CD274)*. In RCC4 Cas9 GFP and WTAP KO 2H1, *PD-L1* transcripts were mapped to either ENST00000381577, which encodes for membrane PD-L1 protein, or ENST00000474218, the surrogate soluble *PD-L1* transcript with 3'UTR that overlaps with soluble *PD-L1* (Figure 5.27A). Since the Ensembl reference transcriptome lacks the Reference genome-aligned IGV coverage tracks of RCC4 Cas9 GFP samples showed an increased representation of exon 5, 6, 7 and 3'UTR coverage to exon 4 in IFNγ and TNF-stimulated samples compared to unstimulated samples (Figure 5.27B). The increased coverage and change in proportions indicated increased levels of membrane *PD-L1* transcripts compared to soluble *PD-L1* transcripts (which lacks exon 5, 6, 7 and the 3'UTR of the membrane *PD-L1* isoform transcripts).

Both reference genome and transcriptome-aligned data showed that IFNγ and TNF stimulation induced significant upregulations of *PD-L1* gene expression levels in RCC4 Cas9 GFP (Figure 5.27C, 5.27 F). In reference transcriptome aligned data, membrane *PD-L1* transcripts showed a 10-fold increase in expression (median TPM: 6.227 vs 72.80) in RCC4 Cas9 GFP after IFNγ and TNF stimulation (Figure 5.27D). In contrast, surrogate soluble *PD-L1* transcript displayed a comparatively modest increase in expression levels (median TPM: 24.29 vs 35.70) (Figure 5.27E). The difference in the magnitude of expression induction between the two transcripts translated into differential transcript proportions at unstimulated and IFNγ and TNF-stimulated RCC4 Cas9 GFP. At unstimulated RCC4 Cas9 GFP, soluble *PD-L1* transcripts outnumbered membrane *PD-L1* transcripts, with an approximate averaged ratio of 6:4 (Figure 5.27I). After 24 hours of IFNγ and TNF stimulation, the ratio reverts to 4:6 on average, with most *PD-L1* transcripts encoding for membrane PD-L1.

A similar trend was observed in reference genome alignment data. Here, membrane *PD-L1* transcripts were defined as transcripts containing sequence overlapping *PD-L1* exon

7 and 3'UTR (chr9: 5,468,000 – 5,470,600). Soluble *PD-L1* transcripts were defined as transcripts with sequence overlapping the soluble *PD-L1* transcript exclusive 3'UTR (chr9: 5,462,830 – 5,463,330). With reference genome-aligned data, both membrane *PD-L1* and soluble *PD-L1* transcripts were significantly upregulated in RCC4 Cas9 GFP after IFNγ and TNF stimulation (Figure 5.27G – H). Membrane *PD-L1* transcripts exhibited a nine-fold increase in expression levels after IFNγ + TNF stimulation (median TPM: 8.41 vs 77.22), whereas 'soluble PD-L1' transcripts showed an approximately two-fold increase in expression (median TPM: 17.85 vs 39.51). Here, reference genome-aligned data also showed that soluble *PD-L1* transcripts were the dominantly expressed transcripts in RCC4 Cas9 GFP with no cytokines exposure, with an approximate ratio of 3:7 between membrane *PD-L1* and soluble *PD-L1* transcripts. After IFNγ and TNF stimulation, the ratio between membrane *PD-L1* and soluble *PD-L1* transcripts changed to approximately 6:4 (Figure 5.27H).

A time-course experiment was conducted with RCC4 cells either stimulated by IFNγ and TNF or unstimulated to validate this finding (n = 3 independent experiments). Using primers which specifically target membrane *PD-L1* and soluble *PD-L1* transcripts, qRT-PCR showed that IFNγ and TNF stimulation induced high levels of fold-induction of membrane *PD-L1* mRNAs than soluble *PD-L1* mRNAs at all time points, compared to unstimulated controls. The differences in expression induction were observed as early as 2 hours post-IFNγ and TNF stimulation (hpt), where membrane *PD-L1* mRNA in stimulated cells showed a 3.8-fold increase in expression, whereas soluble *PD-L1*mRNA showed a 2.5-fold increase compared to unstimulated controls. By 24 hpt, membrane *PD-L1* mRNA showed a 5.8-fold increase in expression levels compared to unstimulated cells, compared to the 2.7-fold increase in soluble *PD-L1* mRNAs. Data here demonstrated that IFNγ and TNF stimulation in RCC4 Cas9 GFP preferentially upregulated the expression of membrane *PD-L1* transcripts

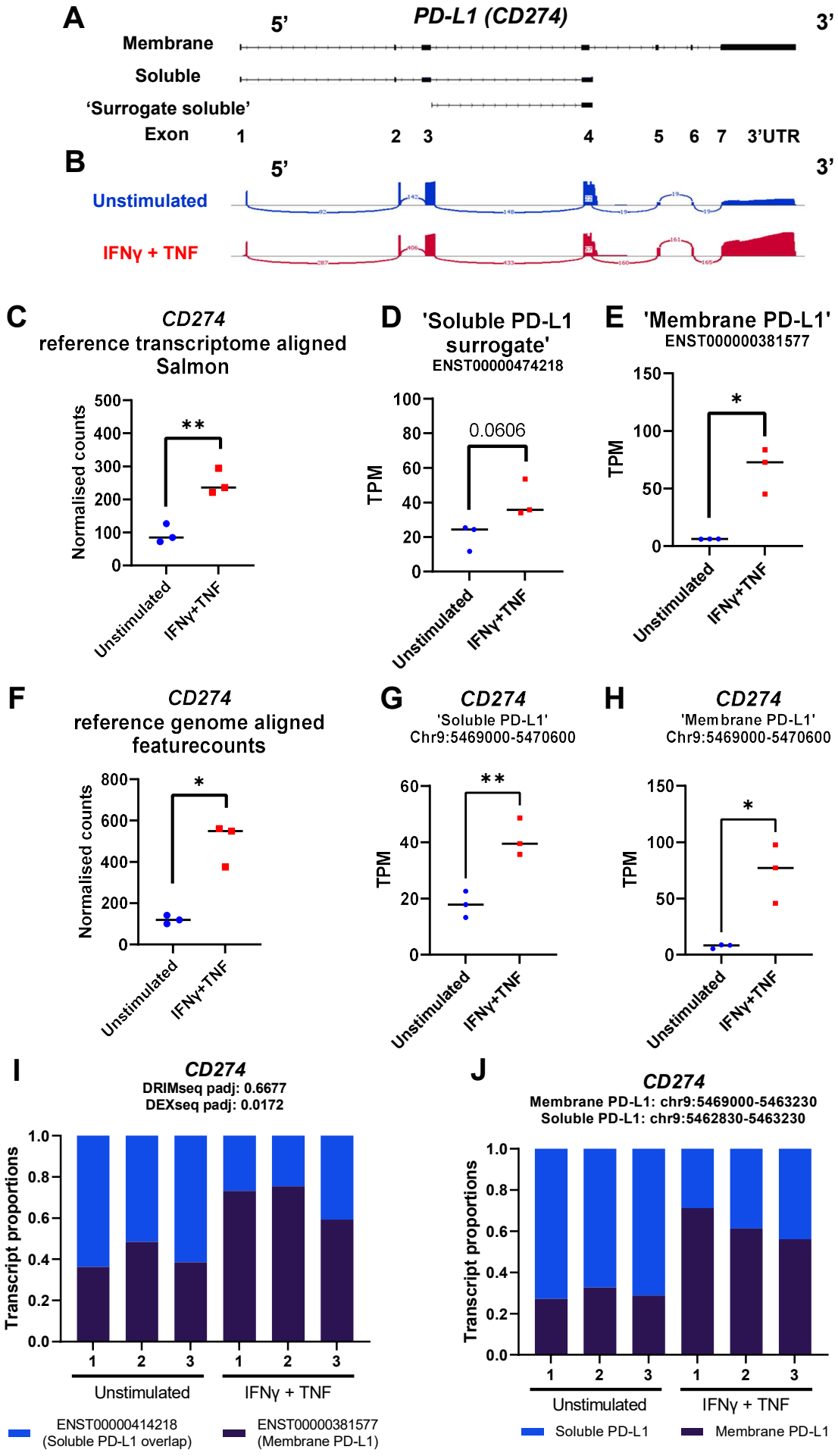

A
PD-L1 (CD274)
5'
3'
Membrane
Soluble
'Surrogate soluble'
Exon
1
2
3
4
5
6
7
3'UTR
B
5'
3'
Unstimulated
IFNγ + TNF
C
CD274
reference transcriptome aligned
Salmon
Normalised counts
**
Unstimulated
IFNγ+TNF
D
'Soluble PD-L1 surrogate'
ENST00000474218
TPM
0.0606
Unstimulated
IFNγ+TNF
E
'Membrane PD-L1'
ENST00000381577
TPM
*
Unstimulated
IFNγ+TNF
F
CD274
reference genome aligned
featurecounts
Normalised counts
*
Unstimulated
IFNγ+TNF
G
CD274
'Soluble PD-L1'
Chr9:5469000-5470600
TPM
**
Unstimulated
IFNγ+TNF
H
CD274
'Membrane PD-L1'
Chr9:5469000-5470600
TPM
*
Unstimulated
IFNγ+TNF
I
CD274
DRIMseq padj: 0.6677
DEXseq padj: 0.0172
Transcript proportions
Unstimulated
IFNγ + TNF
ENST00000414218
(Soluble PD-L1 overlap)
ENST00000381577
(Membrane PD-L1)
J
CD274
Membrane PD-L1: chr9:5469000-5463230
Soluble PD-L1: chr9:5462830-5463230
Transcript proportions
Unstimulated
IFNγ + TNF
Soluble PD-L1
Membrane PD-L1

Figure 5.27: IFNγ and TNF treatment specifically upregulates membrane *PD-L1* transcripts

A) Graphical representation of *PD-L1* (*CD274*) transcript isoforms that encode for membrane PD-L1 (ENST00000381577), soluble PD-L1 (NM_001314029), and the surrogate Ensembl reference transcript for soluble PD-L1 (ENST00000474218). **B)** IGV visualisation of combined unstimulated (blue) and IFNγ + TNF treated (red) RCC4 Cas9 GFP reads coverage track in the region of the *PD-L1* gene. **C)** Grouped dot plot showing reference transcriptome aligned DESeq2 normalised *PD-L1* expression in untreated (blue) and IFNγ + TNF treated (red) RCC4 Cas9 GFP. **D)** Grouped dot plot showing reference transcriptome aligned expression (transcripts per million (TPM)) of surrogate soluble *PD*-L1 transcript (ENST00000474218) in untreated (blue) and IFNγ + TNF treated (red) RCC4 Cas9 GFP. **E)** As in **D**, but for membrane *PD-L1* (ENST00000381577). **F)** As in **C** but fore reference genome aligned gene expression of *PD-L1*. **G)** Grouped dot plot showing reference genome aligned expression (TPM) of soluble *PD-L1* transcripts (defined by reads mapping to chr9: 5,462,830 – 5,463,330 and not a membrane *PD-L1* transcript) in untreated (blue) and IFNγ + TNF treated (red) RCC4 Cas9 GFP. **H)** As in **G**, but for membrane *PD-L1* transcripts (defined by reads mapping to chr9: 5,468,000 – 5,470,600). **I)** Stacked bar graphs representing proportions of *PD-L1* isoforms in reference transcriptome mapped, untreated and IFNγ + TNF treated RCC4 Cas9 GFP cells. DRIMseq and DEXseq p_{adj} values for DTU of *PD-L1* are indicated in graph, with $p \leq 0.1$ considered significant. **J)** Stacked bar graphs representing proportions of *PD-L1* isoforms in reference genome-aligned, untreated and IFNγ + TNF treated RCC4 Cas9 GFP cells. *PD-L1* transcripts that encodes for membrane *PD-L1* were defined by reads mapping to chr9: 5,468,000 – 5,470,600. Soluble *PD-L1* transcripts were defined by reads mapping to chr9: 5,462,830 – 5,463,330 and not a membrane *PD-L1* transcript. For **C – H**, two-tailed unpaired T-tests with Welch's correction were used, with $p \leq 0.05$ considered significant. P values of non-significant results are indicated in graphs. Centre line represents the median for each group.

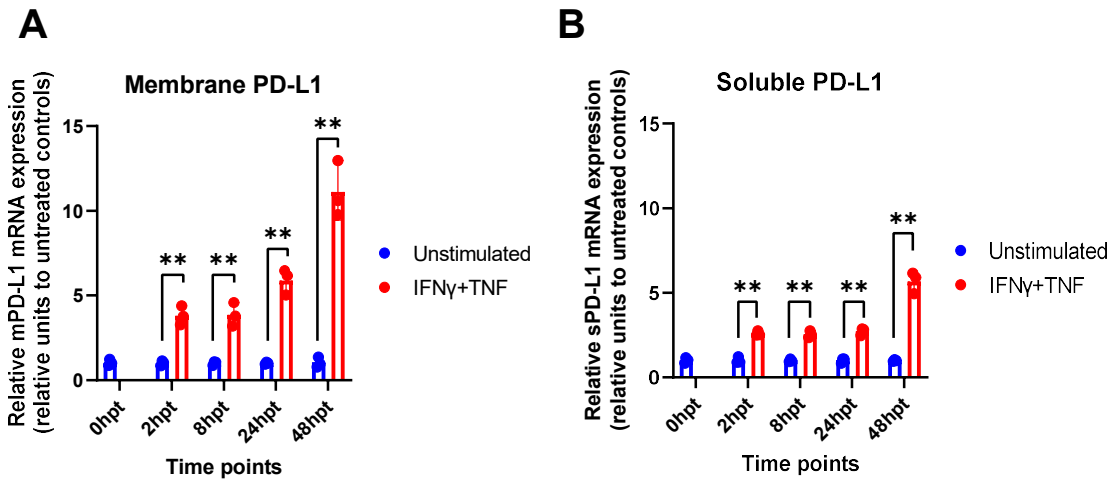


Figure 5.28: qRT-PCR validation of membrane *PD-L1* specific induction by IFNγ and TNF in RCC4

A) Bar graph demonstrating median membrane *PD-L1* mRNA levels measured by qRT-PCR in IFNγ and TNF treated (red) RCC4 cells at 0, 2, 8, 24, 48 hours post treatment (hpt), relative to averaged membrane *PD-L1* mRNA levels in untreated (blue) RCC4 cells at respective treatment time point. Membrane *PD-L1* expression were normalised to *GAPDH* expression. **B)** As in **A**, but for soluble *PD-L1* mRNA levels. Two-tailed unpaired T-tests with Welch's correction were used, with $p \leq 0.05$ considered significant. Asterisks indicate statistical significance levels (** = $p < 0.01$).

5.3.21 Novel soluble *PD-L1* transcripts are expressed in ccRCC tumour cells

Previously, using both DRS and PCS of ccRCC archival tumours, a novel *PD-L1* isoform was discovered to encode soluble *PD-L1* with an extended 3'UTR compared to the reference annotation (see Chapter 4, Figure 4.29). Here, IGV coverage tracks for PCS of ccRCC tumours, as well as RCC4 Cas9 GFP and WTAP KO 2H1, demonstrated the expression of novel isoform in both unstimulated and IFNγ and TNF stimulated conditions (Figure 5.29 B – D). This indicated that the novel soluble *PD-L1* transcript described in Chapter 4 is indeed expressed in ccRCC tumour cells.

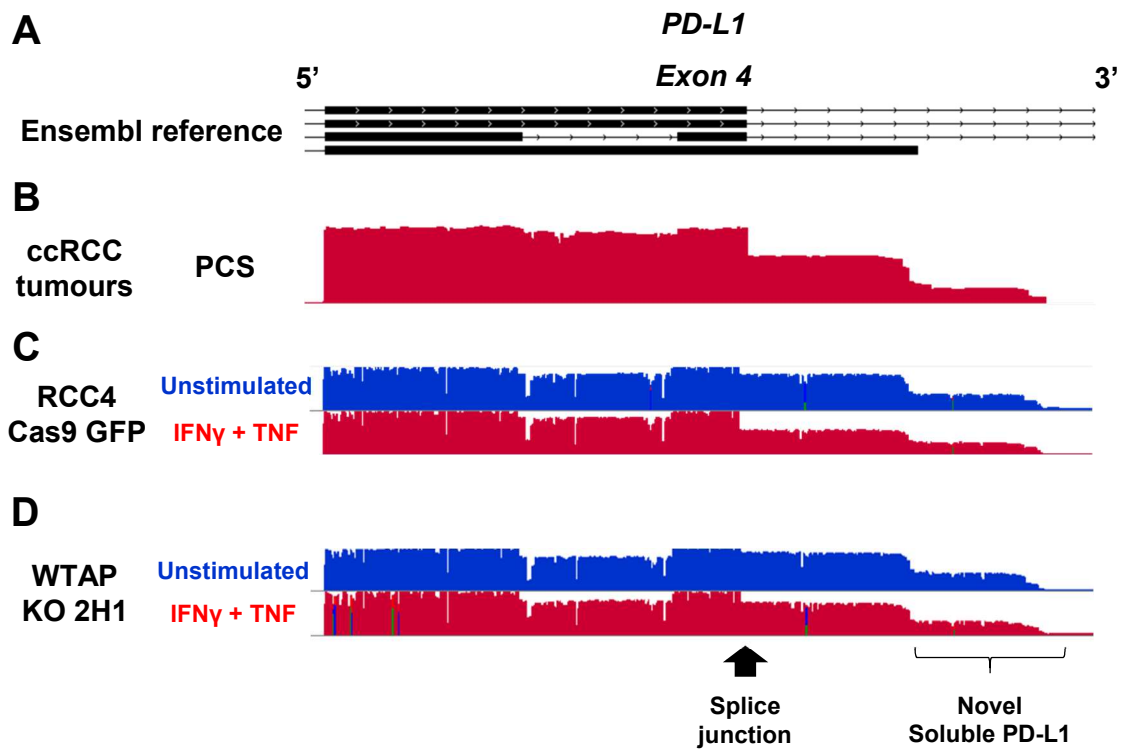


Figure 5.29: Identification of novel soluble *PD-L1* mRNAs in RCC4 Cas9 GFP and WTAP KO 2H1

A) Graphical representation of *PD-L1* transcript isoforms structures near the exon 4 region (hg38 chr9: 5,450,542 – 5,463,350) from Ensembl reference annotation (GRCh38). **B)** IGV visualisation of combined PCS reads coverage track (Red) for all sequenced ccRCC tumours at *PD-L1* locus between hg38 chr9: 5,450,542 – 5,463,350. **C)** IGV visualisation of combined unstimulated (blue) and IFNγ + TNF treated (red) RCC4 Cas9 GFP reads coverage track in the region of chr9: 5,450,542 – 5,463,350. **D)** As in **C**, but for WTAP KO 2H1. Exon 4 splice junction for membrane PD-L1 transcripts and the extended 3'UTR for novel soluble *PD-L1* transcripts are indicated below the reads coverage tracks.

5.3.22 IFNγ and TNF treatment induces global mRNA poly(A) tail lengthening in RCC4 Cas9 GFP cells

After DGE and DTU analyses, the poly(A) tail profiles of unstimulated and IFNγ + TNF treated RCC4 Cas9 GFP cells were analysed using nanopolish. Both mean and median mRNA poly(A) tail lengths were lower in untreated compared to IFNγ + TNF treated RCC4 Cas9 GFP cells (Figure 5.30A). Combining all transcripts from both conditions, mRNA molecules from IFNγ + TNF treated RCC4 Cas9 GFP cells displayed significantly longer poly(A) tail lengths than mRNAs from untreated cells (Mean length: 93.5 nt vs 84.5 nt; Median length: 84.6 nt vs 74.6 nt) (Figure 5.30B).

Next, the expression levels of known nuclear and cytoplasmic poly(A) deadenylases and polymerases were surveyed. No poly(A) deadenylases or polymerases were previously highlighted as a significantly differentially expressed gene ($p_{adj} \leq 0.1$, $|\log_2 FoldChange| > 2$). Amongst the poly(A) polymerases, the expression levels of the cytoplasmic poly(A) polymerase *TENT5A* (Terminal Nucleotidyltransferase 5A) were upregulated in IFNγ + TNF treated RCC4 Cas9 GFP cells ($\log_2 FoldChange = 1.49$, $p_{adj} = 0.0886$) (Figure 5.30C). However, the poly(A) deadenylase subunit *CNOT6L* (CCR4-NOT Transcription Complex Subunit 6 Like) was also found to be upregulated in IFNγ + TNF treated RCC4 Cas9 GFP cells ($\log_2 FoldChange = 1.84$, $p_{adj} = 0.0709$) (Figure 5.30D).

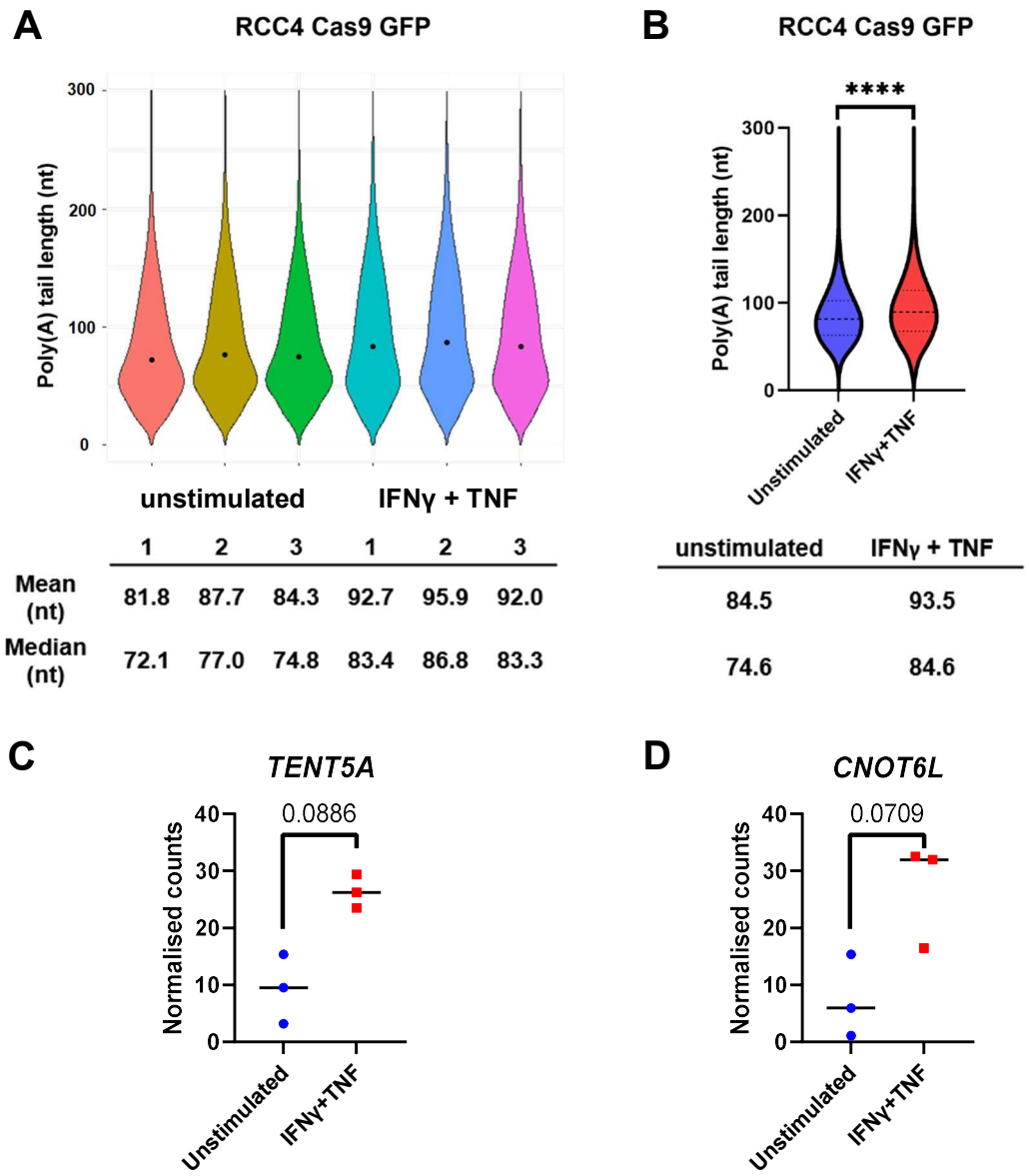


Figure 5.30: Global poly(A) tail length profiling in untreated and IFNγ + TNF treated RCC4 Cas9 GFP cells

A) Violin plots showing poly(A) tail lengths of transcripts from untreated and IFNγ + TNF treated RCC4 Cas9 GFP cells, estimated by nanopolish using DRS data. Dot within violin represents median. **B)** Violin plots showing the combined poly(A) tail length profiles of untreated and IFNγ + TNF treated RCC4 Cas9 GFP cells. Horizontal dotted lines represent inter-quartile range (25%, 50%, 75%). **C)** Grouped dot plot showing reference genome aligned DESeq2 normalised *TENT5A* expression in untreated (blue) and IFNγ + TNF treated (red) RCC4 Cas9 GFP. **D)** Grouped dot plot showing reference transcriptome aligned DESeq2 normalised *CNOT6L* expression in untreated (blue) and IFNγ + TNF treated (red) RCC4 Cas9 GFP. For **B**, nested two-tailed nested t-test was used, with $p \leq 0.05$ considered significant. **** ≤ 0.0001. For **C** – **D**, p_{adj} values were generated by DESeq2 using Wald tests followed by Benjamini-Hochberg corrections. The centre lines represent the median for each group.

5.3.23 Differential poly(A) tail length regulation by IFNγ and TNF in *PD-L1* and *CD24* transcript isoforms

Focussing on the immune checkpoints, the poly(A) tail length profiles of *PD-L1* and *CD24* isoforms in untreated and IFNγ + TNF treated RCC4 Cas9 GFP cells were analysed. The poly(A) tails of *PD-L1* transcripts (including both membrane and soluble PD-L1 isoforms) were significantly longer in IFNγ + TNF treated RCC4 Cas9 GFP cells (median: 120.7 vs 99.9 nt) (Figure 5.31A). Individually, membrane PD-L1 transcripts showed nonsignificant increasing trends in poly(A) tail lengths in IFNγ + TNF treated RCC4 Cas9 GFP cells compared to mRNAs from untreated cells (median: 127.5 vs 108.2 nt) (Figure 5.31B). For soluble PD-L1 transcripts, the poly(A) tails were significantly longer in IFNγ + TNF treated RCC4 Cas9 GFP cells than untreated cells (median: 112.3 vs 93.1 nt) (Figure 5.31C).

Comparing the two PD-L1 isoforms in both untreated and IFNγ + TNF treated RCC4 Cas9 GFP cells, the poly(A) tail lengths of soluble PD-L1 transcripts were observed to be significantly shorter compared to membrane PD-L1 transcripts (median: 126.8 vs 108.7 nt) (Figure 5.31D). Furthermore, when analysed separately at either untreated or IFNγ + TNF treated conditions, the median poly(A) tail lengths of soluble PD-L1 transcripts were both lower than membrane PD-L1 transcripts (93.1 vs 108.2nt & 112.3 vs 127.5nt for unstimulated and IFNγ + TNF treated, respectively), albeit not statistically significant (Figure 5.31E – F).

For *CD24*, both long 3'UTR and short 3'UTR transcripts together and individually, no differences in their poly(A) tail lengths were observed between untreated and IFNγ + TNF treated RCC4 Cas9 GFP (Figure 5.31G – I). Furthermore, when untreated, the poly(A) tail lengths of the long 3'UTR and short 3'UTR *CD24* isoform were not significantly different (median: 102.2 & 99.6 nt) (Figure 5.31K). However, when treated with IFNγ + TNF, the poly(A) tails of long 3'UTR *CD24* transcripts were significantly longer than the short 3'UTR transcripts (median: 100.8 vs 88.92) (Figure 5.31L), which was previously seen in ccRCC archival tumour samples (Figure 4.34E).

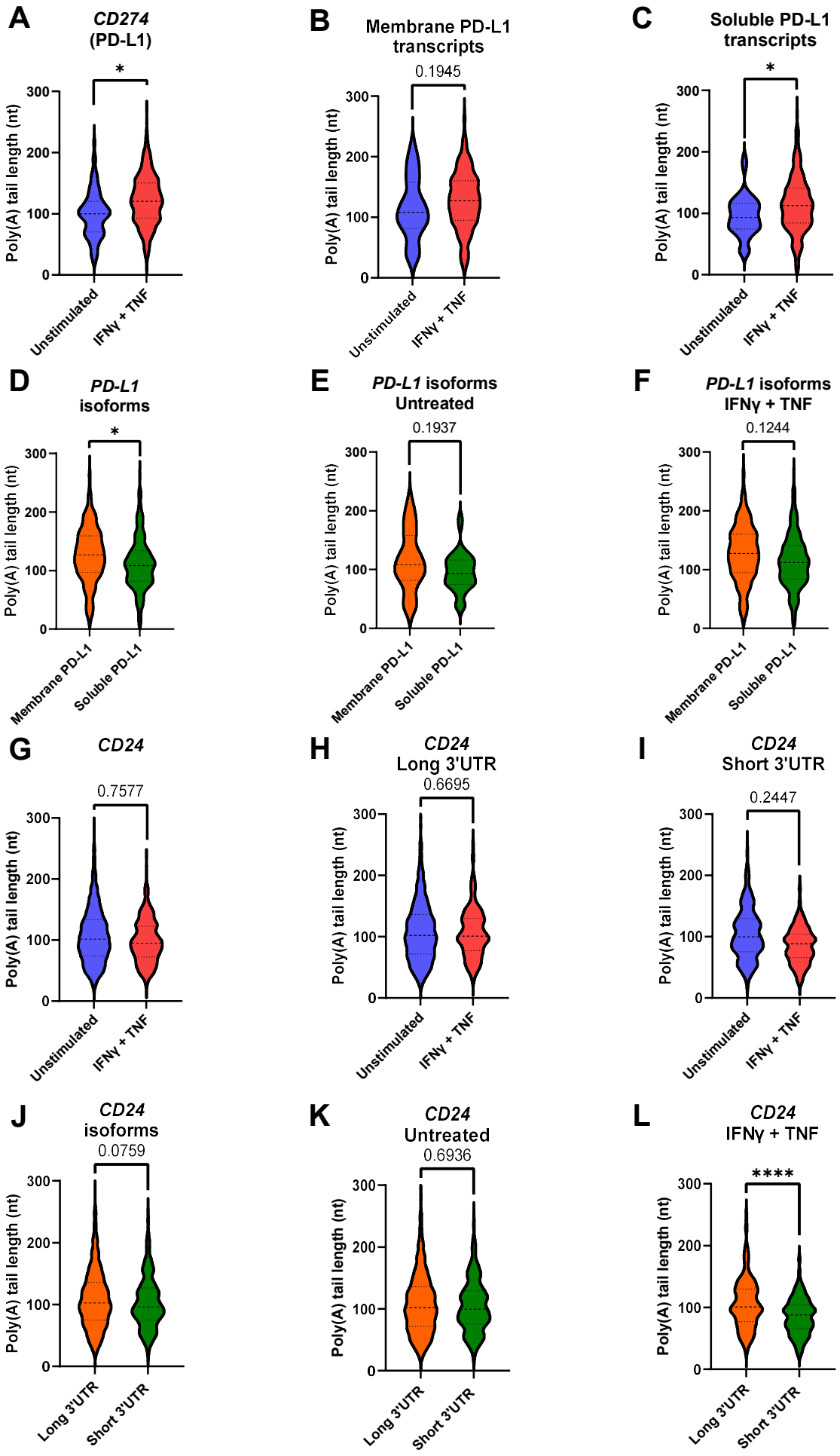

A
CD274
(PD-L1)
*
Poly(A) tail length (nt)
300
200
100
0
Unstimulated
IFNγ + TNF
B
Membrane PD-L1
transcripts
0.1945
C
Soluble PD-L1
transcripts
*
D
PD-L1
isoforms
*
Membrane PD-L1
Soluble PD-L1
E
PD-L1 isoforms
Untreated
0.1937
F
PD-L1 isoforms
IFNγ + TNF
0.1244
G
CD24
0.7577
H
CD24
Long 3'UTR
0.6695
I
CD24
Short 3'UTR
0.2447
J
CD24
isoforms
0.0759
Long 3'UTR
Short 3'UTR
K
CD24
Untreated
0.6936
L
CD24
IFNγ + TNF
****

Figure 5.31: Differential *PD-L1* and *CD24* transcript isoform poly(A) tail length profiles in untreated and IFNγ + TNF treated RCC4 Cas9 GFP cells

A) Violin plots showing poly(A) tail length profiles of all *PD-L1* (*CD274*) transcripts from untreated and IFNγ + TNF treated RCC4 Cas9 GFP cells, estimated by nanopolish using DRS data. **B)** As in **A**, but for membrane *PD-L1* transcripts. **C)** As in **A**, but for soluble *PD-L1* transcripts. **D)** Violin plots showing poly(A) tail length profiles of membrane *PD-L1* and soluble *PD-L1* transcripts in both untreated and IFNγ + TNF treated RCC4 Cas9 GFP cells. **E)** As in **D**, but only with *PD-L1* transcripts from untreated RCC4 Cas9 GFP cells. **F)** As in **D**, but only with *PD-L1* transcripts from IFNγ + TNF treated RCC4 Cas9 GFP cells. **G)** Violin plots showing poly(A) tail length profiles of all *CD24* transcripts from untreated and IFNγ + TNF treated RCC4 Cas9 GFP cells. **H)** As in **G**, but for *CD24* long 3'UTR transcripts. **I)** as in **G**, but for *CD24* short 3'UTR transcripts. **J)** Violin plots showing poly(A) tail length profiles of *CD24* long 3'UTR and short 3'UTR transcripts in both untreated and IFNγ + TNF treated RCC4 Cas9 GFP cells. **K)** As in **J**, but only in *CD24* transcripts from untreated RCC4 Cas9 GFP cells. **L)** As in **J**, but only in *CD24* transcripts from IFNγ + TNF treated RCC4 Cas9 GFP cells. Throughout, horizontal dotted lines represent inter-quartile range (25%, 50%, 75%). Nested two-tailed nested t-tests were used, with $p \leq 0.05$ considered significant. * ≤ 0.05, **** ≤ 0.0001. P values of non-significant results are indicated in graphs.

5.3.24 WTAP KO resulted in decreased mRNA m^{6}A modification in *PD-L1* transcripts

Previously using dot blot analysis, preliminary data showed that global RNA m^{6}A levels were suppressed in WTAP KO 2E6 and 2H1 cells compared to RCC4 Cas9 GFP cells (Figure 5.3I). To see if the m^{6}A levels of *PD*-L1 were specifically downregulated in the WTAP KO cells, the m^{6}A sites of *PD-L1*, as well as *SETD7* transcripts (previously validated to contain WTAP-dependent m^{6}A sites by Schwartz *et al*.), were profiled using the m^{6}A miCLIP database m^{6}A atlas (Schwartz *et al.*, 2014; Tang *et al.*, 2021).

m^{6}A atlas compiled miCLIP databases from 27 different human miCLIP experiments, with a collection of nearly 180,000 unique m^{6}A sites. In *PD-L1*, two high-confidence m^{6}A were identified, one at exon 4 (chr9:5462895) and one at the 3'UTR adjacent to the stop codon (chr9:5467980) (Figure 5.32A). For *SETD7*, 18 m^{6}A sites at the 3'UTR and 1 m^{6}A site at the exon 3 were previously recorded using miCLIP (Figure 5.32B).

MeRIP-qRTPCR is a commonly used m^{6}A quantification method which uses anti-m^{6}A antibodies to pull down fragmented mRNA containing m^{6}A, followed by qRT-PCR (Zeng *et al.*, 2018). m^{6}A site-specific primer sets (red) and negative control primer sets (blue) were designed for *PD-L1* and *SETD7* (Figure 5.32A – B). Amplicon lengths ranged between 100 – 200nt, and both negative primers were at least 350 nucleotides away from the closest reported m^{6}A site. The primer sets' amplification efficiencies (between 90 – 110%) and amplification product specificities (single amplification product as confirmed by melt curve analysis) were validated before use.

RNA fragmentation was optimised to produce an accurate assay for m^{6}A detection using qRT-PCR. Here, total RNA was incubated with Ambion RNA Fragmentation Buffer (Thermo Fisher) for 0, 1, 2, 5, 10, and 15 minutes at 70^{o}C. Fragmentation samples were assessed using Agilent 2100 Bioanalyzer (Figure 5.32C – D). A fragmentation length of 10 minutes was chosen to provide RNA fragment profiles that were both large enough

for PCR amplification and short enough such that m^6A negative primers would not be able to amplify from immunoprecipitated fragments.

MeRIP was performed on RNA purified from IFNγ and TNF-stimulated RCC4 Cas9 GFP and the WTAP KO clonal cell lines (2E6, 2H1, 3A1), followed by qRT-PCR. Transcripts' m^6A percentages were calculated by comparing Ct values of m^6A-immunoprecipitated samples with standard curves generated via titration of corresponding input RNA. MeRIP-qRT-PCR results (n = 2 independent experiments) showed that the rates of m^6A methylation in *PD-L1* and the positive control *SETD7* were substantially higher in RCC4 Cas9 GFP compared to WTAP KO 2E6, 2H1 and 3A1 (Figure 5.33A, C). *PD-L1* negative control primers resulted in broadly lower m^6A % compared to the m^6A-specific primers (Figure 5.33B). *SETD7* negative control primers did not return detectable signals from m^6A immunoprecipitated samples (Figure 5.33D). Previously, *GAPDH was* reported as a viable negative control in MeRIP-qRT-PCR assays (Zeng *et al.*, 2018). Here, low levels of *GAPDH* m^6A modification were found across all four cell lines. Overall, the results showed suppressed mRNA m^6A modification levels in the WTAP KO cell lines compared to parental RCC4 Cas9 GFP cells.

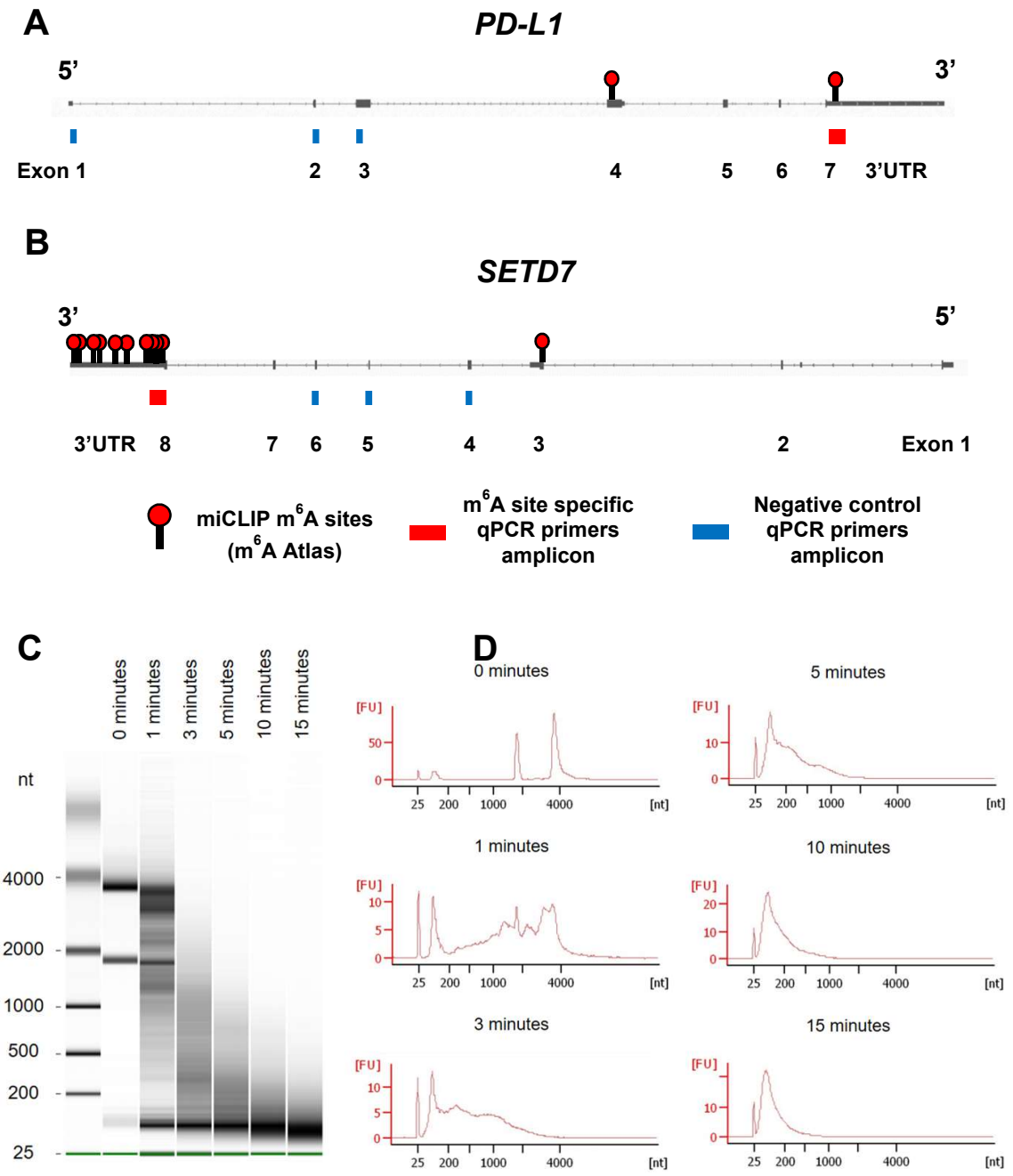


Figure 5.32: m^{6}A-site specific primers design for MeRIP-qRT-PCR experiments

A) Graphical representation of *PD-L1* transcript (ENST00000381577, membrane PD-L1) structure, with published miCLIP studies identified m^{6}A sites from m^{6}A Atlas highlighted above. Coverage of m^{6}A site specific qPCR primers amplicon (red bar) and negative control amplicon (blue bar) are shown below transcript. **B)** As in **A**, but for *SETD7* transcript (ENST00000274031). **C)** Bioanalyzer gel image for the chemically fragmented RNA products, with the corresponding length of fragmentation listed above graph. **D)** Electropherograms showing the RNA fragment size profiles of RNA samples after incubation at 70°C in RNA fragmentation buffer for 0, 1, 3, 5, 10 and 15 minutes. FU represents height threshold.

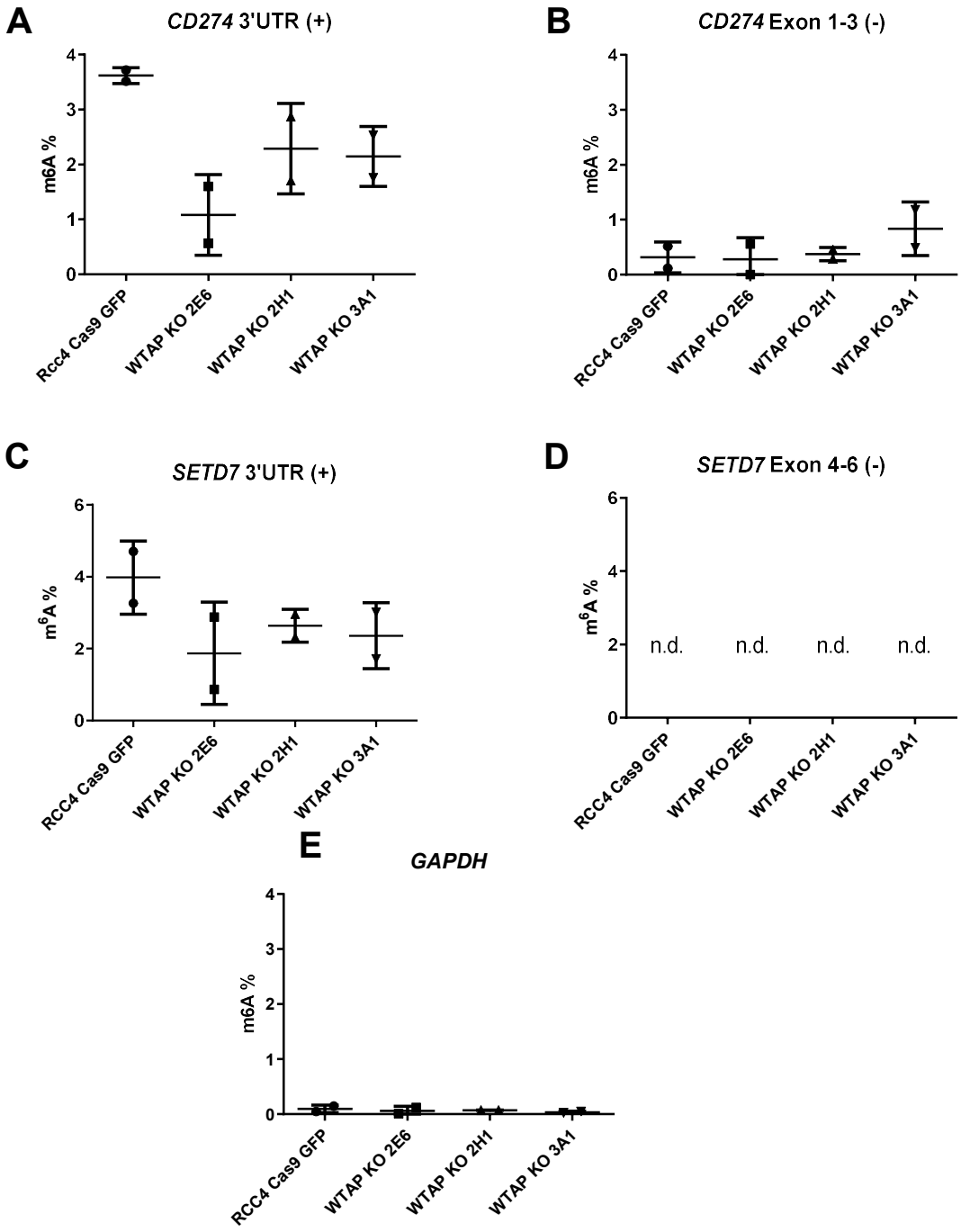


Figure 5.33: 3'UTR of *PD-L1* mRNA contains m^{6}A modifications

MeRIP-qRT-PCR from total RNA extracted from RCC4 Cas9 GFP, WTAP KO 2E6, 2H1 and 3A1 cell lines (n = 2) using primers targeting **A)** *PD-L1* (*CD274*) 3'UTR m^{6}A site, **B)** *PD-L1* (*CD274*) exon 1 – 3 (negative control), **C)** *SETD7* 3'UTR m^{6}A sites, **D)** *SETD7* exon 4 – 6 (negative control), and **E)** *GAPDH* (negative control). m6A % levels were calculated by comparing Ct values of immunoprecipitated samples with a standard curve generated by titrating corresponding input samples. Throughout, centre horizontal lines represent mean values. Error bars represent standard deviation.

5.3.25 *WTAP* KO suppresses membrane PD-L1 protein expression levels

Next, the role of WTAP and m^{6}A modification on the expression levels of membrane PD-L1 was examined. Based on DRS gene expression data, the mRNA levels of *PD-L1* and ENST00000381577 (membrane *PD-L1* mRNA) were not significantly differentially expressed between RCC4 Cas9 GFP and WTAP KO 2H1, both untreated and after 24 hours IFNγ + TNF treatment. Using primers which specifically target membrane *PD-L1* transcripts (exon 6 – 7), qRT-PCR showed no significant differences in expression levels between RCC4 Cas9 GFP and all three WTAP KO clonal cell lines (2E6, 2H1 and 3A1), at both untreated and IFNγ + TNF treated conditions (Figure 5.34A).

Protein expression levels of PD-L1 were examined using western blotting and flow cytometry analysis. For western blot, the anti-PD-L1 antibody E1L3N (cell signaling technology) binds specifically to the cytoplasmic domain of PD-L1 (encoded in exon 6 – 7), which is absent in the soluble PD-L1 proteins (Lawson *et al.*, 2020). Preliminary western blot analysis showed that at both untreated and IFNγ + TNF treated conditions, WTAP KO clonal cell lines expressed substantially lower levels of membrane PD-L1 proteins compared to the parental RCC4 cas9 GFP cell line. In addition, METTL3 did not show compensatory protein expression in the WTAP KO cell lines (Figure 5.34B).

For flow cytometry expression analysis, cell surface expression of PD-L1 was probed by the PE-conjugated (phycoerythrin) antibody clone 29E.2A3, which recognises the extracellular domains of PD-L1 (Haile *et al.*, 2013). PE fluorescent signal and antibody staining specificity were demonstrated in Figure 5.34C by comparing with unstained and isotype-control antibody-stained samples. Results here showed that cell surface expression levels of PD-L1 in RCC4 Cas9 GFP and WTAP KO 2H1 are IFNγ concentration-dependent (Figure 5.34D). Surface PD-L1 expression levels peaked at 1000U/mL for both cell lines, which was the IFNγ concentration used throughout this study. Across all IFNγ concentrations, PD-L1 cell surface protein expression was significantly lower in WTAP KO 2H1 cells than in RCC4 Cas9 GFP. Consistent with

previous western blot analysis results, flow cytometry data showed that cell surface PD-L1 was expressed at significantly lower levels in all three WTAP KO clones, compared to parental RCC4 Cas9 GFP cell line in both untreated and IFNγ + TNF treated conditions (Figure 5.34E – F). Finally, time-course experiments were conducted with IFNγ + TNF treated and untreated control RCC4 Cas9 GFP and WTAP KO 2H1, where cell surface PD-L1 expression levels were assessed by flow cytometry at 8, 24, 48 and 72 hours post-treatment. Data showed that cell surface PD-L1 expression levels were lower in WTAP KO 2H1 than RCC4 Cas9 GFP at all time points in both untreated and IFNγ + TNF treated conditions (Figure 5.34G). Data here demonstrated that CRISPR-Cas9-mediated KO of WTAP resulted in suppressed PD-L1 protein expression whilst having no impact on the mRNA levels.

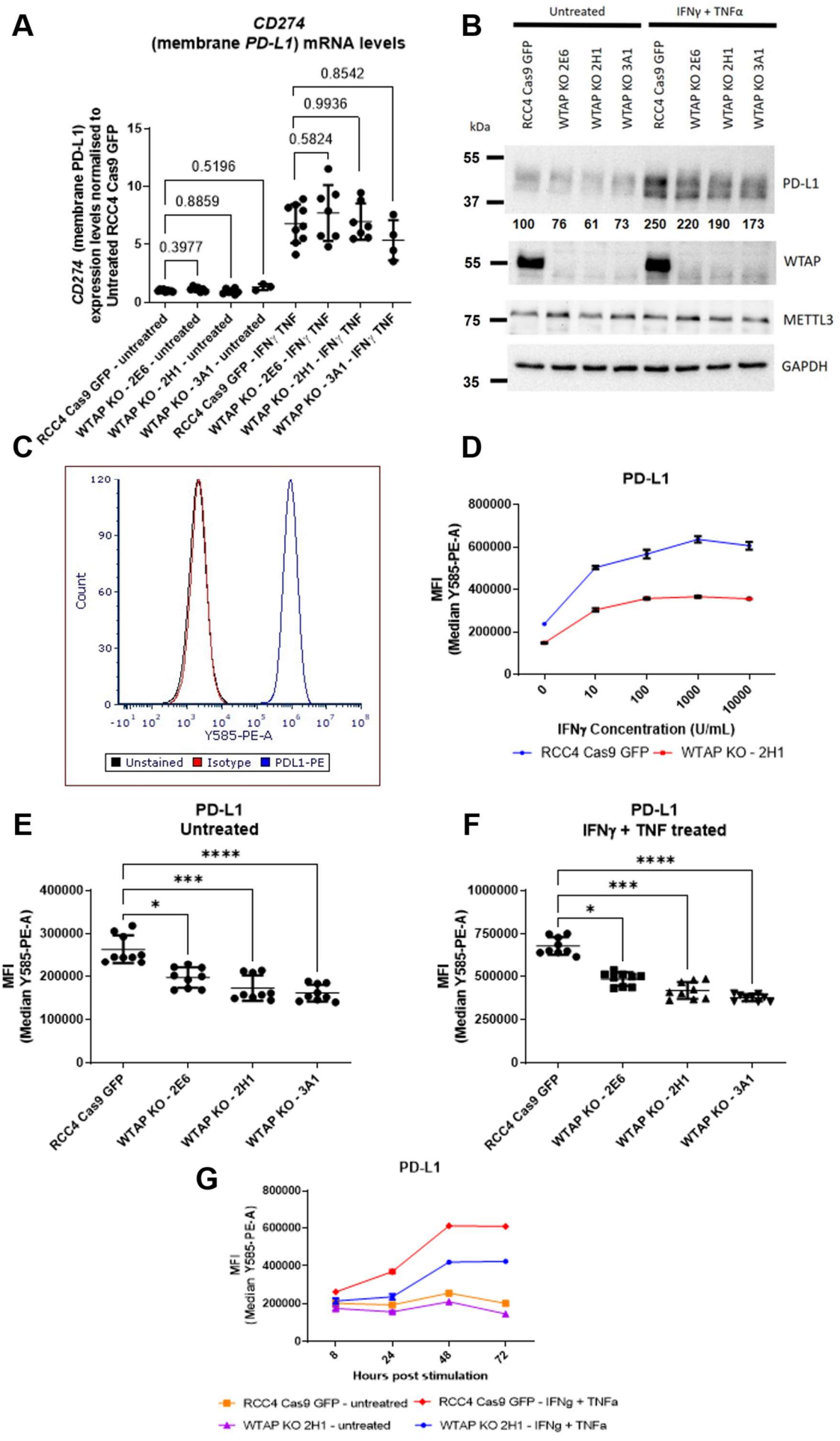

A
CD274
(membrane PD-L1) mRNA levels
CD274 (membrane PD-L1) expression levels normalised to Untreated RCC4 Cas9 GFP
0.5196
0.8859
0.3977
0.8542
0.9936
0.5824
RCC4 Cas9 GFP - untreated
WTAP KO - 2E6 - untreated
WTAP KO - 2H1 - untreated
WTAP KO - 3A1 - untreated
RCC4 Cas9 GFP - IFNγ TNF
WTAP KO - 2E6 - IFNγ TNF
WTAP KO - 2H1 - IFNγ TNF
WTAP KO - 3A1 - IFNγ TNF
B
Untreated
IFNγ + TNFα
RCC4 Cas9 GFP
WTAP KO 2E6
WTAP KO 2H1
WTAP KO 3A1
kDa
PD-L1
WTAP
METTL3
GAPDH
100 76 61 73 250 220 190 173
C
Count
Y585-PE-A
Unstained
Isotype
PDL1-PE
D
PD-L1
MFI (Median Y585-PE-A)
IFNγ Concentration (U/mL)
RCC4 Cas9 GFP
WTAP KO - 2H1
E
PD-L1
Untreated
MFI (Median Y585-PE-A)
F
PD-L1
IFNγ + TNF treated
MFI (Median Y585-PE-A)
G
PD-L1
MFI (Median Y585-PE-A)
Hours post stimulation
RCC4 Cas9 GFP - untreatred
RCC4 Cas9 GFP - IFNg + TNFa
WTAP KO 2H1 - untreated
WTAP KO 2H1 - IFNg + TNFa

Figure 5.34: Characterisation of membrane *PD-L1* mRNA and protein expression levels in RCC4 Cas9 GFP and WTAP KO clonal cell lines

A) *CD274* (membrane PD-L1) mRNA levels measured by qRT-PCR in untreated and IFNγ + TNF treated (24 hours) RCC4 Cas9 GFP, WTAP KO 2E6, 2H1, 3A1 cell lines, relative to averaged *PD-L1* mRNA levels in untreated RCC4 Cas9 GFP. *PD-L1* mRNA levels were normalised to *GAPDH*. **B)** Western blot analysis of PD-L1, WTAP, METTL3 and GAPDH (loading control) in untreated and in IFNγ and TNFα stimulated (24 hours) RCC4 Cas9 GFP, WTAP KO 2E6, WTAP KO 2H1 and WTAP KO 3A1 cell lines (n = 1). PD-L1 expression levels were quantified relative to GAPDH levels using densitometry analysis. **C)** Representative flow cytometry analysis of IFNγ + TNF treated (24 hours) RCC4 Cas9 GFP cells stained for PD-L1 (with PE-conjugated anti-PD-L1 antibody), compared to unstained (black) and PE-conjugated isotype control (red). **D)** Flow cytometry analysis of the expression of cell surface PD-L1 (median fluorescence intensity (MFI), Y585-PE-A) in RCC4 Cas9 GFP (blue) and WTAP KO 2H1 (red) stimulated with increasing amount of IFNγ and without IFNγ for 24 hours (n = 3) . **E)** Flow cytometry analysis of the expression of cell surface PD-L1 (MFI, Y585-PE-A) in unstimulated RCC4 Cas9 GFP, WTAP KO 2E6, 2H1 and 3A1 cells (n = 9) **F)** as in **E**, but for IFNγ + TNF treated (24 hours) cells. **G)** Time course experiment showing kinetics of cell surface PD-L1 expression (MFI, Y585-PE-A) after 8, 24, 48, and 72 hours of IFNγ and TNFα stimulation using flow cytometry analysis (n = 2). For **A**, **E** and **F**, non-parametric Kruskal-Wallis one-way ANOVA tests were performed, with $p \leq 0.05$ considered statistically significant. Asterisks indicate statistical significance levels (**** = $p < 0.0001$, *** = $p < 0.001$, * = $p < 0.05$). p values of nonsignificant results are indicated in graphs. Centre line represents median for each group.

5.3.26 Characterisation of the effects of siRNA-mediated m^{6}A writers depletion on PD-L1 expression in RCC4 cells

Effects of WTAP and mRNA m^{6}A depletion on the protein expression of PD-L1 were further investigated using the transient siRNA-mediated gene depletion approach. Flow cytometry analysis showed that in both untreated and IFNγ + TNF treated (24 hours) conditions, cell surface PD-L1 expression levels were significantly suppressed in siMETTL3 transfected RCC4 cells compared to siNTC controls (n = 6) (Figure 5.35A – B). Conversely, no expression changes were observed in siWTAP transfected cells compared to siNTC controls for both untreated and IFNγ + TNF treated conditions.

Next, using western blot and densitometry analysis, membrane PD-L1 expression was quantified relative to GAPDH expression levels in four independent experiments. Under the untreated condition, no significant differences in membrane PD-L1 expression were observed in siWTAP and siMETTL3 individually transfected RCC4 cells compared to siNTC control cells (Figure 5.35C – D). When both siMETTL3 and siWTAP were simultaneously knocked down when untreated, a significant reduction in membrane PD-L1 was observed compared to siNTC transfected cells (Figure 5.35E). In IFNγ + TNF treated cells, siMETTL3 transfected cells displayed a nonsignificant downregulation trend in cell surface PD-L1 expression compared to siNTC transfected cells (p_{adj} = 0.09, median expression level compared to siNTC: 71.5%). No statistically significant differences were found between all m^{6}A writers' knockdowns and siNTC control RCC4 cells with IFNγ + TNF treatment (Figure 5.35F).

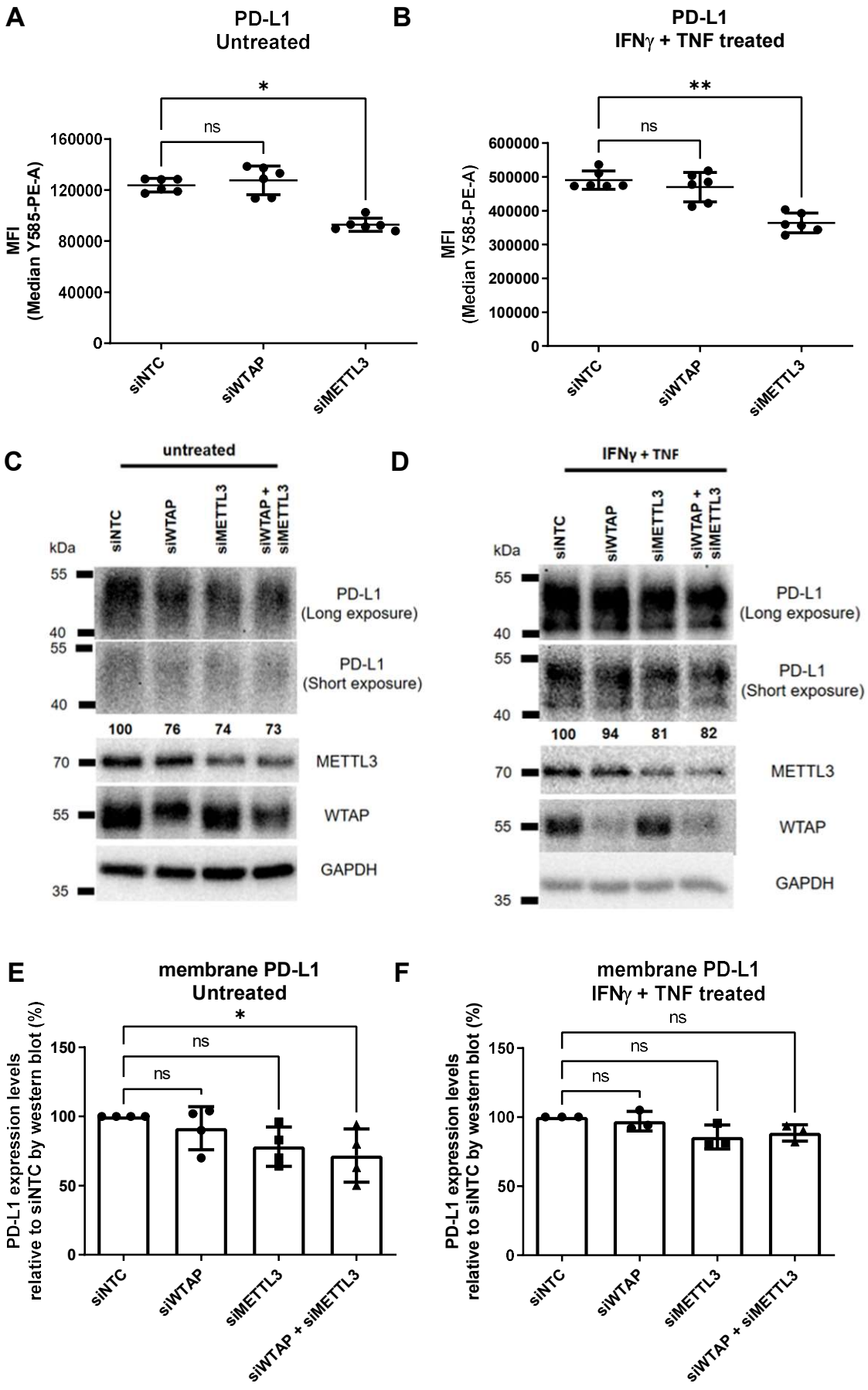

A
PD-L1
Untreated
MFI
(Median Y585-PE-A)
160000
120000
80000
40000
0
*
ns
siNTC
siWTAP
siMETTL3
B
PD-L1
IFNγ + TNF treated
MFI
(Median Y585-PE-A)
600000
500000
400000
300000
200000
100000
0
**
ns
siNTC
siWTAP
siMETTL3
C
untreated
siNTC
siWTAP
siMETTL3
siWTAP + siMETTL3
kDa
55
40
55
40
PD-L1
(Long exposure)
PD-L1
(Short exposure)
100
76
74
73
70
55
35
METTL3
WTAP
GAPDH
D
IFNγ + TNF
siNTC
siWTAP
siMETTL3
siWTAP + siMETTL3
kDa
55
40
55
40
PD-L1
(Long exposure)
PD-L1
(Short exposure)
100
94
81
82
70
55
35
METTL3
WTAP
GAPDH
E
membrane PD-L1
Untreated
PD-L1 expression levels
relative to siNTC by western blot (%)
150
100
50
0
*
ns
ns
siNTC
siWTAP
siMETTL3
siWTAP + siMETTL3
F
membrane PD-L1
IFNγ + TNF treated
PD-L1 expression levels
relative to siNTC by western blot (%)
150
100
50
0
ns
ns
ns
siNTC
siWTAP
siMETTL3
siWTAP + siMETTL3

Figure 5.35: Effects of siRNA-mediated m^{6}A writers depletion on PD-L1 expression in RCC4 cells

A) Flow cytometry analysis of the expression of cell surface PD-L1 (median fluorescence intensity (MFI), Y585-PE-A) in siNTC, siWTAP and siMETTL3 transfected (54 hours), untreated RCC4 cells (n = 6). **B)** Flow cytometry analysis of the expression of cell surface PD-L1 (median fluorescence intensity (MFI), Y585-PE-A) in siNTC, siWTAP and siMETTL3 transfected (54 hours), IFNγ and TNF stimulated (24 hours) RCC4 cells (n = 6). **C)** Representative western blot analysis of PD-L1, WTAP, METTL3 and GAPDH (loading control) in untreated RCC4 cells, transfected with siNTC, siWTAP, siMETTL3, and siWTAP + siMETTL3 (54 hours). PD-L1 expression levels were quantified relative to GAPDH levels using densitometry analysis. **D)** As in **C**, but for IFNγ and TNF stimulated (24 hours) RCC4 cells. **E)** Bar chart showing western blot densitometry analysis of membrane PD-L1 protein expression in untreated sNTC, siWTAP, siMETTL3, and siWTAP + siMETTL3 (54 hours) transfected RCC4 cells, relative to siNTC membrane PD-L1 expression levels (n = 4). **F)** As in **E**, but for IFNγ and TNF stimulated (24 hours) RCC4 cells.

For **A**, **B, E** and **F**, non-parametric Kruskal-Wallis one-way ANOVA tests were performed, with $p \leq 0.05$ considered statistically significant. Asterisks indicate statistical significance levels (**** = $p < 0.0001$, *** = $p < 0.001$, * = $p < 0.05$). Error bars represent standard diviations. For **A – B**, centre lines represent median for each group. For **E – F**, bars represent mean for each group.

5.4 Discussion

5.4.1 Role of m^{6}A in ccRCC tumour cells

One of the main aims of this chapter was to explore the role of mRNA m^{6}A in ccRCC tumour cells. Firstly, TCGA KIRC genomics data analysis demonstrated that CNVs of m^{6}A regulators are widespread in ccRCC tumours. Interestingly, data here showed that CNVs of all three m^{6}A writers (*METTL3*, *METTL14* and *WTAP*) in ccRCC were near-universally deleterious, whilst CNVs of the m^{6}A eraser *FTO* were primarily identified as 'low-level gain' (gene duplications). With the gene expression levels highly correlated with their CNV status, the trend from the data here agree with a previous study which showed suppressed global m^{6}A levels of ccRCC tumours compared to normal tissue (Shen *et al.*, 2022). Survival analysis also showed that gene deletion of *METTL3* alone and in combination with other m^{6}A writers correlated with worse overall survival of ccRCC patients. The data above suggest a potential involvement of aberrations in m^{6}A regulation in ccRCC progression.

There are several limitations to the genomic CNV analysis presented. Firstly, previous work has demonstrated that the prevalence of CNV significantly correlates with the ccRCC tumour stage (Correa *et al.*, 2020). Therefore, it is unclear if the CNVs m^{6}A regulators were 'drivers' or 'passenger' alterations due to the enhanced genome instability in later staged ccRCC. The heterogeneous nature of ccRCC TME also presents challenges in attributing CNVs to tumour cells. There are limited bioinformatics tools which adjust CNV frequencies by tumour purity, but ultimately single cell genome sequencing is required for accurate attribution of CNVs for each cell type in clinical tumour samples (Mahdipour-Shirayeh *et al.*, 2022)

Similar difficulties arise from tumour bulk-RNAseq data analysis, where observed changes in gene expression cannot be easily attributed to a specific cell type within the tumour. To directly ascribe the role of m^{6}A on gene expression profiles of ccRCC tumour cells, CRISPR-Cas9-mediated gene deletion of m^{6}A writers was conducted in the *VHL*-null RCC4 cells. CRISPR-Cas9 system has been used to generate m^{6}A writers (METTL3,

METTL14, WTAP) KO cancer cell line, including in HEK293T and A549 cells (L. Wang *et al.*, 2020; Ge *et al.*, 2021). However, no ccRCC cancer cell line with m^6A writer genes deleted via the CRISPR-Cas9 system has been reported at the time of writing. Several WTAP KO clonal RCC4 cell lines were isolated, with KO validated using western blot. In contrast, the inability to generate viable METTL3 KO clones suggests that the *METTL3* gene may be essential for RCC4 survival.

Previous studies have indicated that the deletion of *WTAP* or *METTL3* genes leads to early embryonic lethality in mice (Fukusumi *et al.*, 2008; Geula *et al.*, 2015). However, several viable *METTL3* CRISPR-Cas9 KO mouse embryonic stem cell lines (mESC) have been described, with global m^6A levels ranging between 0 – 40% (Poh *et al.*, 2022). Using DepMap, a database which compiles identified essential genes from published genome-wide CRISPR-Cas9 and RNAi gene depletion screens, a recent study has shown that *METTL3* and *WTAP* are required for survival in 801 and 836 out of the 1,054 profiled cell lines (Pacini *et al.*, 2021). Amongst the viable cancer cell lines, the impact of METTL3 KO on their proliferation capacities also varied. For example, depletion of *METTL3* via CRISPR-Cas9-mediated gene deletion in the human haepatocellular carcinoma Huh7 cells was found to increase the transcript stability of the tumour suppressor *SOCS2*, leading to suppression in cell proliferation (Mengnuo Chen *et al.*, 2018). In contrast, *METTL3* KO by CRISPR-Cas9-mediated editing in the murine colorectal cancer cell line CT26 and melanoma cell line B16 showed no growth defects, both *in vitro* and *in vivo* (syngeneic model by subcutaneous injection) (L. Wang *et al.*, 2020). Overall, the gene essentiality of *METTL3* and *WTAP* on cell viability and fitness is context-dependent.

Transcriptomic analysis showed that KO of WTAP in RCC4 Cas9 GFP cells did not significantly change the cellular response to IFNγ + TNF treatment. However, dozens of significantly differentially expressed genes between the two cell lines were identified, many of which are associated with the hypoxic response and glycolytic pathways. Strikingly, some of the top DEGs, such as *NDUFA4L2, BNIP3*, *CA9* and *EGLN3*, are

known to be classic gene markers for ccRCC with their expression levels correlating with tumour progression and poorer prognosis(Macher-Goeppinger *et al.*, 2017; Apanovich *et al.*, 2021). The mRNA and protein expression levels of these DEGs were validated by qRT-PCR and western blotting in WTAP KO 2E6 and 2H1 cells but not in WTAP KO 3A1 cells. Subsequent orthogonal validation by siRNA-mediated gene silencing of *WTAP* and *METTL3* did not recapitulate findings from WTAP KO 2H1 cells. Thus, further work is needed to establish the role of m^6A writers and m^6A on the gene expression of ccRCC tumour cells.

Several factors may have contributed to the differential phenotypes between the KO clones. Firstly, given the importance of mRNA m^6A modifications on gene expression regulation and the fact that *WTAP* is essential in the majority of human cell lines that were screened by transcriptome-wide CRISPR-Cas9 deletion assay, the selected WTAP KO clones may have been viable due to compensatory responses. The function of the depleted gene in a gene regulatory network can be compensated by enhanced expression of other genes within the same network. For example, in human melanoma cell line A375, deletion of the β-Actin gene (*ACTB)* induces the expression of the other actin isoform γ-Actin (*ACTG1*), and vice versa (Malek *et al.*, 2020). Here, although WTAP KO did not result in enhanced METTL3 protein expression and no m^6A regulators were shown to be differentially expressed from DRS data, the pro-survival role in RCC4 cells may be compensated by other genes involved in cell proliferation pathways. Hyperactive hypoxic response and glycolytic pathways are crucial for sustaining ccRCC tumour cell growth (Semenza, 2007). The significant suppression of these pathways and the selection of WTAP KO 2E6 and 2H1 clones indicates a potential interplay between WTAP and the ccRCC proliferative pathways. Alternatively, activation of other compensatory pro-survival pathways may have allowed WTAP KO 3A1 to remain proliferative with activated hypoxic and glycolytic pathways. Further characterisation of the transcriptomic profiles from the other WTAP KO clones may provide valuable information on the potential compensatory responses.

In addition to compensatory responses, the phenotypic variations observed can also originate from differential, unforeseen off-target effects between clones. For example, a genome-wide sequencing study has revealed that, on average, 45% of the resulting Cas9 cleavage sites across the genome were not predicted as either target sites or potential off-target sites (Höijer *et al.*, 2020). In addition, although on-target CRISPR/Cas9 gene editing typically causes short genomic deletion (under 50bp), unexpected megabase-scale genome deletions and interchromosomal genomic rearrangement events have also been recently observed (Kosicki *et al.*, 2018; Cullot *et al.*, 2019). Here, *WTAP* was edited with the transfection of a pool of three different guide RNAs. It is possible that different off-target effects were propagated in different WTAP KO lines. Therefore, detailed genomic characterisation of the WTAP KO clones and RCC4 Cas9 GFP cells will be crucial to assess any off-target gene editing or genomic aberrations that could have contributed to the differential phenotypes.

The interclonal variability observed could have also been introduced by the inherent heterogeneity between RCC4 Cas9 GFP cells. In this experiment, RCC4 Cas9 GFP cells were first isolated by FACS, and the non-clonal pool was transfected with guide RNAs, followed by single-cell cloning for the KO clones. A recent study has compared the gene expression profiles of unedited monoclonal cells from a parental mouse kidney epithelial cell line (mIMCD-3) and demonstrated high levels of gene expression variability (783 significantly differentially expressed genes, $|Log_2FoldChange| > 1$, $p_{adj} < 0.01$) between clones. The differential expression profiles were shown to persist after guide RNA transfection and downstream single-cell subcloning. However, the variability in gene expression levels between KO clones was significantly reduced by single-cell sorting before the transfection of guide RNAs (Westermann *et al.*, 2022). Therefore, characterising multiple sets of WTAP KO clones generated from different clonal RCC4 Cas9 GFP could provide data with lower noise and fewer false-positive gene targets.

CRISPR-Cas9 gene knockout and siRNA-mediated gene silencing are two of the most well-used methods for studying loss-of-function phenotypes in cells. However, there are

essential differences between the methods, which may contribute to the differential phenotypes observed between the KO and KD experiments. Firstly, the CRISPR-Cas9 knockout approach represents a complete and permanent loss of gene function. In contrast, siRNA-mediated gene silencing represents a transient, partial loss of gene function (Zimmer *et al.*, 2019). Nevertheless, it is possible that the partial WTAP and METTL3 knockdown still resulted in fully functional m^{6}A writer complexes in RCC4 cells, therefore retaining their biological roles. In addition, the time between guide RNA transfection and transcriptomic characterisation for WTAP KO 2H1 could have allowed compensation to occur, whereas the acute nature of the siRNA approach may not permit RCC4 cells to do so. Therefore, another approach to validate the effects of WTAP KO is by overexpressing the gene in the KO clones and characterising if the rescue of WTAP would reverse the differential gene expression.

Although WTAP KO 2E6, 2H1 and 3A1 cells did not exhibit the same transcriptomic changes, the MeRIP-qRT-PCR assay showed a universal depletion of m^{6}A levels in *SETD7* and soluble PD-L1 mRNA molecules. Preliminary data also showed decreased levels of m^{6}A globally in WTAP KO 2E6 and 2H1 cells. The decreased m^{6}A levels in membrane PD-L1 transcripts were of particular interest since the protein expression levels of membrane PD-L1 were also significantly suppressed, whilst mRNA expression levels were unchanged between WTAP KO clones and RCC4 Cas9 GFP. KD of m^{6}A writers showed modest effects on membrane PD-L1 protein expression. However, it is unclear to what extent KD of m^{6}A writers depletes m^{6}A levels in the transfected RCC4 cells. It is also important to note that in addition to transcriptional regulation by the hypoxic response, growth factors and cytokines, PD-L1 expression is highly regulated post-transcriptionally by miRNAs and post-translationally via protein phosphorylation, ubiquitination and glycosylation (Figure 5.36). Further work is needed to assess if the reduction in membrane PD-L1 protein expression in WTAP KO lines were m^{6}A dependent.

Interestingly, the m^{6}A site here sits exclusively at the membrane PD-L1 transcript and is not shared by soluble PD-L1 transcript. Thus, this may represent a membrane PD-L1 isoform exclusive post-transcriptional regulation. Therefore, it will be beneficial to interrogate the m^{6}A levels of the membrane PD-L1 transcripts from existing DRS results of RCC4 Cas9 GFP and WTAP KO 2H1, as well as between untreated and IFNγ + TNF treated cells using published m^{6}A detection algorithms such as Nanocompore and xPore (Leger *et al.*, 2021; Pratanwanich *et al.*, 2021). Finally, integrating transcriptome-wide m^{6}A mapping with other bioinformatic analyses (DGE, DTU, poly(A) profiling) will provide novel insights into the co-regulation of gene expression by multiple post-transcriptional modifications.

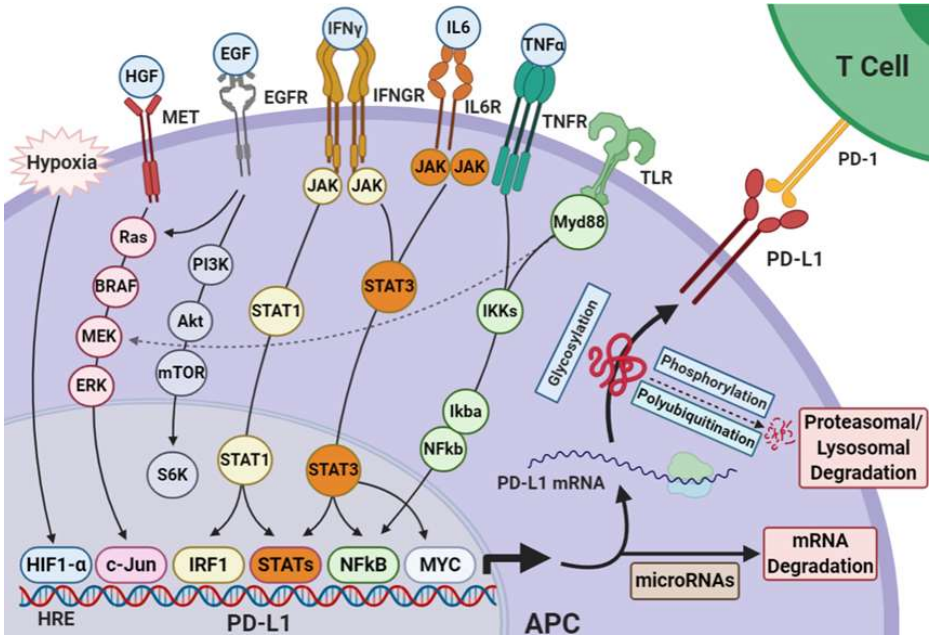


Figure 5.36: PD-L1 expression regulation

PD-L1 expression in regulated at the transcriptional, post-transcriptional and post-translational levels. Interactions between receptors (such as MET, EGFR, IFNGR, IL6R, TNFR, and TLR) and their respective extra cellular ligands trigger cascades of signaling activities. Activation of these pathways activate transcription factors such as IRF1, STAT1, STAT3 and NFkB, resulting in *PD-L1* transcription. Expression of PD-L1 is also tightly regulated post-transcriptionally by microRNAs, either directly by targeting the transcript or indirectly by inhibiting upstream signaling pathways. After translation, post-translational modification can also regulate PD-L1 protein stability and translocation. Phosphorylation and polyubiquitination of the protein results in proteasomal and lysosomal degradation. Glycosylation stabilises PD-L1 proteins, allowing efficient translocation of the protein to the plasma membrane.

5.4.2 Role of IFNγ + TNF in ccRCC tumour cells

Transcriptomic analysis showed dramatic changes when RCC4 Cas9 GFP and WTAP KO 2H1 cells were treated with IFNγ + TNF. PCA showed that 82% of the variations in gene expression could be explained by the cytokine treatment alone, and hundreds of differentially expressed genes were identified. In agreement with the literature, GSEA results showed significantly upregulated antigen presentation pathways in both RCC4 Cas9 GFP and WTAP KO 2H1 cells after IFNγ + TNF treatment. Expression regulation of antigen presentation pathway components by cytokines represents a key mechanism to orchestrate an anti-tumour response. Exposure to IFNγ and TNF induces the expression of MHC class I molecules transcriptionally, facilitating enhanced antigen presentation by the tumour cells. This results in increased antigen recognition by $CD8^+$ T cells and activation of their effector functions and promotes anti-tumour immunity (Figure 5.33) (Wieczorek *et al.*, 2017).

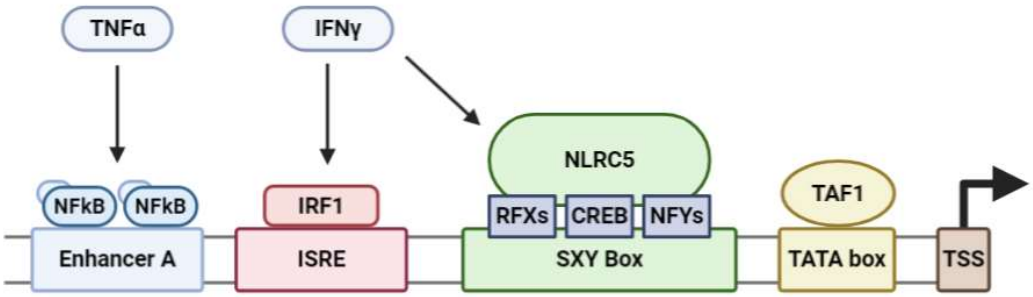


Figure 5.37: IFNγ and TNF induce the expression of MHC Class I molecules

In human, MHC class I molecules consist of two main domains: the heavy α chain encoded by *HLA-A*, *HLA-B* and *HLA-C*, and the beta-2 microglobulin (*B2M*) light chain. For the *HLA*s, their constitutive expressions are mainly driven by the transcription factor TAF1 (TATA box binding protein associated factor 1). Their promoter regions contain an enhancer A where NFkB binding facilitates both constitutive and TNF induced expression. IFNγ can also induce expression via upregulation of IRF1 and NLRC5 (NOD-like receptor family CARD domain containing 5), the key transcriptional regulator for MHC class I. NLRC5 does not bind to the promoter region directly but rather via an enhanceosome protein complex consist of various transcription factors (Regulatory factor X (RFX), cAMP response element binding protein (CREB) and nuclear transcription factor Y (NFY)).

Inhibition of MHC class I molecules expression is a commonly evolved immune evasion strategy by the tumour cell. Here, ccRCC tumour cells were not seen to display an impaired antigen presentation pathway. With the high levels of tumour immune infiltration seen in the ccRCC tumours from the previous chapter, transcriptomic data here suggest that ccRCC tumour cells may have to rely on alternative pathways to evade tumour immunity. In both RCC4 Cas9 GFP and WTAP KO 2H1 cells, IFNγ and TNF treatment increased the expression of co-inhibitory immune checkpoints *IDO1* and *PD-L1*. IFNγ + TNF-induced *IDO1* and *PD-L1* expression were previously described in the literature in other cancer cell types (Robinson *et al.*, 2003; Li *et al.*, 2018). However, DRS results also showed that immune checkpoint genes, such as *CD24* and *PD-L1,* may display DTU before and after IFNγ and TNF treatment.

Transcription factors downstream of IFNγ and TNF activation pathways were indicated to induce alternative splicing and alternative poly(A) site usage in tumour immune genes. For example, the binding of TNF-inducible NFkB to the enhancer sequence of *PTEN* promotes alternative poly(A) site usage and the shortening of *PTEN* mRNA 3'UTR (Kwon *et al.*, 2022). In addition, IRF1, an IFNγ induced transcription factor, was also shown to regulate alternative splicing of the immune checkpoint *CEACAM1* in a hnRNP L- and hnRNP A1-dependent manner (Dery *et al.*, 2014).

Transcript isoform usage is also influenced by differential isoform stability. Both *cis*- and *trans*-regulatory factors contribute to transcript stability. For *cis*-regulatory elements, codon optimality, secondary structures in the 5' and 3' UTRs, microRNA- and RBP-binding sites have all been implicated in dictating the stability of mRNA molecules (Cheng *et al.*, 2017). In addition, the expression levels of *trans*-factors (miRNAs and RBPs) can vary widely depending on cell type and in response to external stress or stimuli (Van Nostrand *et al.*, 2020; Zarnack *et al.*, 2020; Keller *et al.*, 2022). IFNγ and TNF may influence multiple regulatory pathways, resulting in differential transcript usage in genes that regulate tumour immunity.

For *PD-L1*, it was previously shown that IFNγ differentially and independently regulates the transcription of soluble *PD-L1* and membrane *PD-L1* isoforms. For example, in the human B lymphocyte cell line Ramos, IFNγ specifically induced the expression of the soluble *PD-L1* transcripts. In contrast, only the membrane *PD-L1* transcripts were upregulated in the human leukaemia monocytic cell line THP-1 after IFNγ treatment (Ng *et al.*, 2019). Here, the soluble PD-L1 transcripts were found to be the primary transcript isoform at untreated RCC4 Cas9 GFP and WTAP KO 2H1. Once treated with IFNγ and TNF, membrane PD-L1 transcripts were upregulated to a higher level than soluble PD-L1 transcripts, becoming the predominant isoform. Results here correlated with the transcriptomic results from previous archival tumour samples where high levels of membrane *PD-L1* transcripts were found, in conjunction with high levels of immune infiltrates that can express IFNγ and TNF. Future experiments with additional cytokines which induce *PD-L1* expression in tumour cells (such as IL-6, IFNα), both alone and in combinations, would be informative for modelling and predicting the expression patterns of membrane and soluble PD-L1.

The 3'UTR of the membrane *PD-L1* transcript is a crucial determinant of its stability. Within the 3'UTR, various miRNAs, including miR-155 and miR-34a, have been shown to negatively regulate its transcript expression levels (Xi Wang *et al.*, 2015; Yee *et al.*, 2017). In addition, there are multiple AU-rich elements in the 3'UTR where TTP could bind and promote its degradation (Coelho *et al.*, 2017). Lastly, as the MeRIP-qRT-PCR experiment has shown, membrane *PD-L1* transcripts harbour m^6A modification at the 3'UTR. With alternative 3'UTRs, soluble *PD-L1* transcripts are likely to be regulated differently than membrane *PD-L1* transcripts. Further characterisation and comparisons between the soluble *PD-L1* 3'UTRs and membrane *PD-L1* 3'UTR will be valuable for understanding how the isoforms are differentially regulated.

Identifying soluble *PD-L1* expression in ccRCC tumour cells, both with and without exposure to IFNγ and TNF, has important implications. As the ligand for PD-1, tumour cell PD-L1 expression is the most established predictive biomarker for PD-1/PD-L1

blockade treatment. However, for many cancer types, including in ccRCC, PD-L1 expression does not correlate with better treatment outcomes (Ueda *et al.*, 2018). Currently, tumour PD-L1 expression is usually assessed by immunohistochemistry staining. However, most of the antibodies used for these PD-L1 expression assays were found to bind to the cytoplasmic domain of PD-L1, which is absent in the soluble isoforms (Lawson *et al.*, 2020). It is possible that many of the 'PD-L1 negative' tumours expressed high levels of soluble PD-L1 proteins instead of the membrane PD-L1 isoform. Indeed, DRS and PCS data of archival ccRCC tumour samples did show the expression of soluble *PD-L1* transcripts. Together, the results demonstrated that long-read RNA sequencing analysis could complement IHC assays for tumour biomarkers.

In addition to differential gene expression and differential transcript usage, treatment of IFNγ + TNF also resulted in global mRNA poly(A) tail lengthening. A previous study showed that upon heat shock, mRNA poly(A) tails of upregulated genes from yeast were elongated, whereas mRNAs from downregulating genes displayed significantly shorter poly(A) tails (Tudek *et al.*, 2021). Here, volcano plots showed that more genes were upregulated than downregulated upon IFNγ and TNF treatment. A similar trend was also observed for the immune checkpoints, where the poly(A) tail of upregulated PD-L1 transcripts was significantly elongated. In contrast, *CD24* expression levels were, on average, downregulated after IFNγ and TNF treatment, and there was no change in *CD24* mRNA poly(A) tail lengths. Although current literature suggests that the length of the poly(A) tail does not correlate with transcript stability, the length profiles may indicate the transcription rate and gene expression dynamics.

Previously in the archival ccRCC tumour samples, *CD24* transcripts with long 3'UTR were found to have longer poly(A) tails than transcripts with short 3'UTR (Figure 4.34). Interestingly, for RCC4 Cas9 GFP, the discrepancies in poly(A) tail lengths only appeared when cells were treated with IFNγ and TNF but not when untreated. This shows that the differential poly(A) profiles between transcripts could be cytokine-dependent.

The poly(A) tail length is controlled by poly(A) polymerases and deadenylases. After IFNγ + TNF treatment, the expression of the cytoplasmic poly(A) polymerase *Tent5a* was found to be upregulated in RCC4 Cas9 GFP cells. A recent pre-print showed that the gene expression levels of *Tent5a,* along with other immune response-related genes, were upregulated at the tissue of the SARS-CoV-2 mRNA vaccine (Moderna mRNA-1273) injection site. Further characterisation by DRS analysis revealed that TENT5A is responsible for lengthening the poly(A) tails from the upregulated transcripts and re-adenylates mRNA-1273 transcripts in the macrophages. Notably, gene knockout of *Tent5a* in mRNA-1273 injected mice showed significantly lower levels of SARS-CoV-2 Spike antigen (encoded by mRNA-1273) and spike-specific antibody response (Krawczyk *et al.*, 2022). The study focused on the effects on immune cells, but the increased Tent5a expression may be cytokine-dependent. It is also important to note that increased expression levels of the deadenylase *CNOT6L* were also found in RCC4 Cas9 GFP cells treated with IFNγ + TNF. Thus, potential feedback mechanisms may be at play in regulating the poly(A) tail lengths after cytokine stimulation. Further investigation is needed to understand how cytokines regulate the tumour immune transcriptome by modulating poly(A) tail metabolism.

5.4.3 Performance of DRS

The transcriptomic profiles of RCC4 Cas9 GFP and WTAP-KO 2H1 cells, with and without IFNγ + TNF treatments, were profiled using DRS. The average DRS reads here were longer than those obtained from archival tumour samples. This disparity in read lengths resulted from sequencing high-quality, non-degraded RNA, which produced long sequencing reads. Like previous sequencing experiments, a higher number of reads significantly correlated with the number of unique genes identified, showing the significance of sequencing throughput to encapsulate a wide range of transcripts within each sample. However, although a higher number of sequencing reads were generated, a significantly lower number of unique genes were identified in the cell lines compared to the tumour samples, which could reflect the cell-type heterogeneity in ccRCC tumours.

Similar to the results from the previous chapter, reference genome alignment resulted in a wider variety of RNA biotypes being detected than reference transcriptome alignment. Reference genome alignment exclusive genes were mainly lncRNAs, whereas reference transcriptome alignment exclusive genes were primarily protein-coding genes. Overall, gene expression levels provided by reference genome and reference transcriptome alignment data were highly correlated. However, the thousands of genes that were exclusively mapped using reference alignment methods here and in previous sequencing experiments showed the importance of analysing transcriptomic data with multiple read-mapping pipelines.

Focussing on the performance of differential transcript usage analysis, both DRIMseq and DEXseq identified DTU events between the cell lines and between untreated and IFNγ + TNF treatments. Using reference transcriptome-aligned data, DTU analysis revealed that IFNγ + TNF treatment specifically induced membrane PD-L1 transcripts, which was also seen from reference genome-aligned data and subsequently validated using qRT-PCR. Both DRIMseq and DEXseq identified DTU in *CD24* after IFNγ + TNF treatment. Upon IFNγ + TNF treatment, the long 5'UTR *CD24* isoform showed an increased proportion of total *CD24* transcripts compared to the drop in short 5'UTR *CD24*

isoforms. Although DTU can also be seen from reference genome-aligned coverage data, it was apparent that reference transcriptome-aligned data assigned an over-estimated proportion of 5'UTR *CD24* isoforms across the samples.

The DTU analysis (DRIMseq and DEXseq) relies on transcript assignment of raw sequencing reads by minimap2, followed by quantification by Salmon. Minimap2 is currently the most frequently used read-aligner for long-read RNAseq experiments, with a previous benchmarking study showing the strategy used here (minimap2 alignment followed by Salmon quantification) being the most effective method in assigning reads to transcripts (Soneson *et al.*, 2019). Various long-read sequencing optimised aligners (QAlign, desalt, uLTRA) with suggested improvement in alignment accuracies compared to minimap2 have since been published (B. Liu *et al.*, 2019; Joshi *et al.*, 2021; Sahlin and Mäkinen, 2021). Independent benchmarking studies on the performance of sequence aligners will be beneficial for designing DRS data analytic pipeline. Results here also show the importance of using multiple reference mapping methods to validate results.

Finally, transcriptome assembly methods (StringTie2 and FLARE) can profile novel transcript isoforms in the ccRCC tumour cells. The tumour cell origin of novel transcripts can be deduced by comparing the results with previous transcriptome assemblies from the archival nephrectomy samples. It will also be helpful to investigate whether the quality and accuracy of read mapping would improve using transcriptome assemblies rather than reference transcriptome.

5.5 Evaluation of key objectives

- **Generation of CRISPR-Cas9-mediated m^{6}A writers KO ccRCC cell lines**

  WTAP KO clonal RCC4 cell lines were isolated and validated with significantly lower global m^{6}A levels compared to controls. However, m^{6}A methyltransferase *METTL3* KO clones were not viable after passaging, suggesting potential key roles of m^{6}A in regulating the growth and survival of ccRCC tumour cells.

- **Evaluate the role of m^{6}A on ccRCC gene expression profiles**

  Sequencing results showed significant suppression in hypoxic response and glycolytic pathways in the WTAP KO cells compared to parental control cells. This was validated and replicated in 2 of 3 KO clones that were tested. KD of m^{6}A writers via siRNA did not reproduce suppression in key hypoxic/glycolytic gene expression. Detailed genomic characterisation and comparisons between WTAP KO clones and the parental cells are needed to evaluate the role of m^{6}A and WTAP in ccRCC.

- **Characterise the effects of IFNγ & TNF on mRNA transcripts in ccRCC cells**

  The exposure of IFNγ and TNF induces profound differential gene expression and differential transcript isoform usage across the transcriptome, including immune checkpoints, in ccRCC tumour cells. IFNγ and TNF treatment also induces global poly(A) tail lengthening. Notably, poly(A) tail of different immune checkpoint isoforms from the same gene displayed differential responses towards the cytokines treatment. Future work linking mRNA transcripts' poly(A) tail length and transcriptional dynamics will be of great interest.

- **Role of inflammatory cytokines and m^{6}A on the expression of the immune checkpoint PD-L1 in ccRCC**

  IFNγ and TNF induce PD-L1 mRNA and protein expression with a bias towards the membrane isoform. KO of WTAP suppressses membrane PD-L1 protein expression but not at mRNA levels. siRNA-mediated KD of m^{6}A writers displayed nonsignificant downregulation of membrane PD-L1 protein expression. This suggests m^{6}A may have a moderate impact on facilitating membrane PD-L1 mRNA translation.

5.6 Summary

This chapter examined the role of mRNA m^{6}A modification and cytokines (IFNγ and TNF) in regulating ccRCC gene expression. Genomics data analysis showed that m^{6}A writer genes are frequently deleted in ccRCC tumours. Deletion of *METTL3* only and in a combination of *METTL14* and *WTAP* in ccRCC patients correlates with worse overall survival. Clonal WTAP KO lines were generated via CRISPR-Cas9-mediate gene deletion in a Cas9-expressing RCC4 cell line (RCC4 Cas9 GFP). DRS analysis identified suppressed hypoxic response and glycolytic pathways in the WTAP KO 2H1 clone compared to unedited control cells. However, WTAP KO-related DEG was only validated in 2 out of 3 WTAP KO clones. siRNA-mediated gene depletion of *WTAP* and *METTL3* also did not reproduce the differential gene expression observed from DRS data. Whilst MeRIP-qRT-PCR assay showed suppression in m^{6}A levels of membrane PD-L1 transcripts and decreased membrane PD-L1 protein expression levels in WTAP KO clones, subsequent siRNA-mediated gene depletion of *WTAP* and *METTL3* showed modest differences compared to non-targeting controls. Further work is needed to characterise the role of m^{6}A and WTAP in ccRCC.

Exposure to IFNγ and TNF induced profound changes in the gene expression profile of RCC4 Cas9 GFP cells, including upregulation in the antigen presentation pathway. DTU analysis identified isoform-switching events after IFNγ and TNF treatment, including the immune checkpoints *CD24* and *PD-L1*. RCC4 Cas9 GFP predominantly expresses soluble PD-L1 isoforms at basal state, and IFNγ and TNF preferentially upregulate membrane PD-L1 expression. Finally, IFNγ and TNF treatment caused global poly(A) tail lengthening in RCC4 Cas9 GFP cells. mRNAs of different *PD-L1* and *CD24* isoforms displayed different poly(A) tail profiles and responses to IFNγ and TNF treatment. Overall, exposure to IFNγ and TNF remodelled the gene expression, transcript isoform profiles and mRNA poly(A) tail lengths in ccRCC tumour cells. Integration of mRNA m^{6}A analysis from DRS data in the future will provide further details on the co-regulation of gene expression by multiple co- and post-transcriptional regulatory events.

Chapter 6

Discussion

6.1 Summary

This study set out to explore the transcriptomic landscape of ccRCC and investigate how it is regulated by co- / post-transcriptome regulatory events using long-read sequencing technologies. In addition, this study aimed to explore the roles of tumour-infiltrating T cells and cytokines in augmenting the expression, isoform usages and post-transcriptional modification of key cancer immune gene transcripts. To address these aims, archival nephrectomy tissues from ccRCC patients and *in vitro* ccRCC tumour cell lines were sequenced using ONT Direct RNAseq and PCR-cDNAseq. In addition to bioinformatics analysis, the ccRCC gene expression profile and post-transcriptional regulatory events were further characterised using MeRIP-qRT-PCR, western blot and flow cytometry analysis. Here, the list of primary aims and main findings from this study are presented:

Aim i) To explore ccRCC transcriptome by long-read sequencing using archival nephrectomy tissues from non-recurrent/recurrent ccRCC patients.

Main findings:

- ccRCC tumours were successfully sequenced using ONT DRS and PCS
- DRS and PCS libraries can be prepared using total RNA instead of poly(A)$^+$ RNA
- DRS and PCS detected a high number of transcripts from a wide variety of RNA biotypes
- A large proportion of the identified transcripts represent novel isoforms
- Novel isoforms of immune checkpoints (*CTLA4*, *CD24*, *PD-L1* and *IDO1*) were found
- Gene expression levels were highly correlated between DRS and PCS and between the reference genome and reference transcriptome alignment data
- The choice of using reference genome or reference transcriptome alignment influenced the ability to identify different subsets of transcripts

Aim ii) To identify key differential expression genes and transcript isoforms between tumours from non-recurrent/recurrent ccRCC patients.

Main findings:

- ccRCC recurrence-associated differentially expressed genes were identified using DRS and PCS
- GSEA revealed that recurrent ccRCC tumours showed suppressed immune cell activation and antigen presentation pathways
- ccRCC recurrence-associated differential transcript usage events were identified using DRIMseq and DEXseq

Aim iii) To compare the immune landscapes between non-recurrent/recurrent ccRCC tumours via RNAseq immune cell-type deconvolution analysis.

Main findings:

- Significantly lower levels of tumour-infiltrating immune cells were found in recurrent ccRCC tumours compared to non-recurrent tumours
- The proportions of $CD8^+$ T cells were suppressed in recurrent ccRCC tumours compared to non-recurrent tumours
- A subgroup of non-recurrent ccRCC tumours showed significantly high levels of $CD8^+$ T cells and high expression levels of exhausted $CD8^+$ T cell markers

Aim iv) To investigate the roles of pro-inflammatory cytokines (IFNγ & TNF) in shaping the transcriptome of ccRCC tumour cells using DRS

Main findings:

- IFNγ + TNF stimulation of ccRCC tumour cells (RCC4 Cas9 GFP and WTAP KO 2H1) induced hundreds of significantly differentially expressed genes
- GSEA revealed that upregulated antigen presentation pathways are upregulated by IFNγ + TNF stimulation
- IFNγ + TNF treatment in ccRCC tumour cells resulted in differential transcript usage on dozens of genes, including the immune checkpoints *PD-L1* and *CD24*
- IFNγ + TNF stimulation specifically induced membrane PD-L1 transcripts compared to soluble PD-L1 transcripts
- IFNγ + TNF treatment causes global poly(A) tail lengthening in RCC4 Cas9 GFP cells
- The poly(A) tail profiles of different *PD-L1* and *CD24* isoforms showed different responses to IFNγ + TNF

Aim v) To characterise roles of m^6A in transcriptomic regulation in ccRCC tumour cells by applying DRS analysis on CRISPR-Cas9 mediated KO of m^6A writer *WTAP*

Main findings:

- RCC4 Cas9 GFP and WTAP KO 2H1 cells showed similar DEGs after treatment from IFNγ + TNF
- WTAP KO 2H1 cells showed suppressed hypoxic response and glycolytic pathways compared to unedited RCC4 Cas9 GFP cells
- WTAP KO-related DEGs were only validated in 2 out of 3 WTAP KO clones surveyed
- siRNA-mediated gene depletion of *WTAP* and *METTL3* did not reproduce DRS gene expression data from WTAP KO 2H1

6.2 Reflection on the application of long-read RNAseq to cancer biology and RNA research

The recent arrival of long-read RNA sequencing technologies has addressed many inherent technical limitations of the short-read sequencing platform, allowing researchers to study the transcriptome at the single molecule level. Results from this study have showcased many advantages of utilising long-read RNAseq. The ability to characterise full-length transcripts provided increased confidence in both isoform assignments and novel isoform discovery. Moreover, examining native RNA molecules by DRS allows a unique opportunity to integrate information on gene expression, transcript isoform usage and poly(A) tail. Another advantage of the nanopore sequencing approach is the ability to re-evaluate current signal data. With continuous improvements in base-calling models and RNA modification detection algorithms, long-read sequencing technologies have great potential to provide novel biological insights.

Whilst there are many benefits of using long-read RNA sequencing, improvement in the following areas may further facilitate the use of the technology in the broader research community. Firstly, the current DRS and PCS library preparation utilise an adaptor primer with a poly(T) overhang to capture poly(A)$^+$ RNA from the input. Thus, non-polyadenylated transcripts such as histone mRNAs, various lncRNAs (*Neat1* and *Malat1)*, and circular RNAs are not represented in the sequencing results (Yang *et al.*, 2011). Recent studies have applied polymerases *in vitro* to add polyinosine or polyuridine tracks at the 3' end of RNA transcripts, followed by RNA capture using customised adaptors (Drexler *et al.*, 2021; Zhang *et al.*, 2022). Implementing these methods may allow a broader range of transcripts that can be characterised using the long read-sequencing approach.

Since DRS and PCS libraries are typically prepared with 3' end poly(A) tail capturing, transcript coverage is universally 3' end enriched. This is contributed by both RNA degradation and potential mid-sequencing interruption, leading to incomplete RNA reads. In addition, DRS and PCS cannot resolve final nucleotides at the 5' end of the molecule

(Mulroney *et al.*, 2022). Thus, it is not possible for DRS and PCS to differentiate between degraded RNA, interrupted RNA sequencing, or if the 5' end represents the true end of the sequenced RNA molecule at present. To mitigate this issue, recent studies have coupled 5' cap capturing method with poly(A) tail purification for full-length transcripts enrichment (Jiang *et al.*, 2019; Ugolini *et al.*, 2022). Whilst these methods successfully retrieve full-length RNA molecules, the input requirements (1.5 – 6µg of poly(A) enriched RNA before 5' cap capture) used in these studies may not be feasible for clinical tumour samples. The sequencing output of the 5'cap-capture DRS experiments (270,000 – 1,500,000 raw reads) was also relatively low, compared to the DRS performed here in this study. Intriguingly, a recent paper has shown that 5' uncapped, polyadenylated transcripts produced by post-transcriptional mRNA cleavage (by RNA endonucleases) can be stable within cancer cells and translated in a cap-independent manner. Furthermore, the translated products were presented by MHC class I molecules in the tumour cells, which can potentially impact the tumour immune response (Malka *et al.*, 2022). Further development is needed to improve RNA capture methods to ensure the maximum varieties of RNA molecules are represented in long read RNA sequencing experiments.

Finally, this study has demonstrated the power of using cell-type deconvolution methods to investigate tumour heterogeneity. Deconvolution algorithms, such as CIBERSORTx, rely on cell-type specific expression matrices to infer cell-type proportion(Newman *et al.*, 2019). With the development of single-cell RNAseq and spatial transcriptomics using long-read RNAseq technologies, the expression of cell-type specific isoform markers is increasingly recognised (Boileau *et al.*, 2022; Volden and Vollmers, 2022). Therefore, the performance of bulk-RNAseq deconvolution methods is likely to be significantly improved using isoform-level expression data. Moreover, the unprecedented resolution brought by these methods will provide valuable insights into the complex cross-talk within the TME.

6.3 Future work

i) Mapping m^{6}A in ccRCC nephrectomy samples and RCC4 cell line using m^{6}Anet

Since the introduction of DRS in 2018, it has been seen as a promising technology for mRNA m^{6}A detection. Several bioinformatics tools have since been published, but the accuracy and sensitivity of these tools rely on the availability of m^{6}A-null control samples (such as METTL3 knockout cell lines or *in vitro* transcribed RNA). Moreover, the accuracy and resolution of these tools are insufficient for determining the m^{6}A site stoichiometry (m^{6}A vs unmodified A %). Due to the lack of 100% m^{6}A-free control (WTAP-KO cells presented a 30% drop in global m^{6}A levels), this analysis were not performed for the DRS data here. A recently published m^{6}A base-calling algorithm, m^{6}Anet, can detect mRNA m^{6}A levels at the single RNA molecule and single nucleotide resolution (Hendra *et al.* 2022). Importantly, unlike other DRS m^{6}A base-calling algorithms, m^{6}Anet does not require an m^{6}A negative training set. This provides an excellent opportunity to explore the m^{6}A landscape in both ccRCC nephrectomy samples and in RCC4 Cas9 GFP / WTAP KO 2H1 cell lines using existing DRS data.

For the nephrectomy samples, m^{6}anet would enable comparisons of global mRNA m^{6}A modification patterns between recurrent and non-recurrent ccRCC tumours. Previous research showed that global mRNA m^{6}A levels of ccRCC tumours are significantly higher than adjacent normal tissues (Chen *et al.* 2019). Moreover, increased global mRNA m^{6}A levels have been linked with enhanced tumour cell growth and proliferation (Cho *et al.* 2021). Although m^{6}A regulators were not found to be differentially expressed, it will be interesting to see if there are any differences in global mRNA m^{6}A levels between recurrent and non-recurrent ccRCC tumours. m^{6}Anet analysis will also allow investigation into the links between m^{6}A modification rate and isoform expression levels. For RCC4 Cas9 GFP/ WTAP KO 2H1, m^{6}A profiling can provide a direct link between levels of m^{6}A and gene expression, which will help in distinguishing the impact of WTAP KO with potential off-target effects. Overall, this analysis will provide a much deeper understanding of post-transcriptional gene regulation in ccRCC.

ii) Characterisation of novel isoforms in ccRCC nephrectomy and RCC4 cells

Using transcriptome assembly methods (Stringtie2 and FLAIR), this study has discovered tens of thousands of novel transcript isoforms that were not annotated in the Ensembl reference database. The existence of novel soluble PD-L1 isoform was validated using qPCR, but the origin and functionalities of these novel isoforms remain unclear.

To identify the potential tumour cell origin of these novel transcripts, the transcriptome assembly generated from the ccRCC tumours can be used to map the RCC4 DRS data. Comparing the expression of these novel transcripts between untreated and IFNγ and TNF treated RCC4 cells can also shed light on the number of the novel transcripts that may only be induced with the presence of immune cells. Differential expression analysis of the novel transcripts should also be carried out in the ccRCC tumours. This may identify potential novel biomarkers for ccRCC recurrence.

For the potential functionalities of the novel transcripts, their coding potential can be determined via bioinformatics tools (such as CPC2) based on their intrinsic sequence features (Kang *et al*. 2017). Further characterisation of potential protein domains can be executed using protein databases, for example, InterPro (Paysan-Lafosse *et al.* 2023). Reanalysis of publicly available proteomics data with the addition of these potential novel isoforms to the reference proteome can validate the existence of these proteins. For the predicted non-coding RNAs, several computational models that have been published to predict their functions based on their sequence composition, genomic locations and gene co-expression data (Chen *et al.* 2019). Altogether, these methods can provide clues to the potential roles of the novel transcripts, which can be further validated by wet-lab methods such as siRNA-mediated gene KD, CRISPR-Cas9-meadiated gene KO and gene over expression assays.

iii) Validation of sequencing results using additional cohorts of ccRCC patients

Using long-read sequencing technologies, this exploratory study has uncovered differentially expressed genes and important discrepancies in tumour infiltrating immune cell populations between recurrent and non-recurrent ccRCC tumours. With the small sample size (n = 12), further validatory cohorts, preferably from different centres, will be needed to substantiate the findings. Top differentially expressed genes, especially immune cell related genes, can be validated using qPCR assays. Using primers targeting $CD8^+$ T cell specific genes (such as CD8B) can also help validate the sequencing results which showed a significant decrease in $CD8^+$ T cells in recurrent ccRCC tumours. Overall, the usage of a validatory cohort will greatly strengthen the evidence of the results presented here, as well as the reliability of using long-read sequencing on tumour samples.

iv) Usage of long-read sequencing optimised bioinformatics tools

Many of the bioinformatics used in this study, including the gene expression quantification tools featurecounts and Salmon, were initially designed for analysing short-read RNA sequencing data. As a rapidly developing field, there is an increasing number of bioinformatics tools that are being developed specifically for long-read RNA sequencing with reported improvement in performance. Recently, a transcript expression quantifier NanoCount has been developed specifically for long read RNA sequencing data. Benchmarking experiments using known spike-in RNA controls showed a significant improvement in transcript quantification accuracy by NanoCount compared to other tools, including salmon (Gleeson *et al.*, 2022). An improved quantification accuracy will also benefit downstream differential transcript usage analysis. For novel isoform discovery, reference-free transcriptome assembly tools (RATTLE and RNA-Bloom2) specifically designed for long read sequencing data have recently been released (de la Rubia *et al.*, 2022; Nip *et al.*, 2022). Using reference-free transcriptome assemblers may avoid the over-correction problem previously observed using the reference-guided FLAIR and StringTie2.

6.4 Concluding remarks

The advent of long-read RNA sequencing technologies has provided the means to study tumour transcriptomes at an unprecedented resolution. This study shows the feasibility of using ONT DRS and PCS to characterise archival ccRCC tumour samples. Bioinformatics analysis shows that the tumour immune infiltration populations, particularly $CD8^+$ T cells, are significantly suppressed in recurrent ccRCC tumours. Using transcriptome assembly methods, thousands of novel isoforms, including immune checkpoints, are found in ccRCC tumours. DRS showed that the exposure of the cytokines IFNγ and TNF remodels the transcriptomic profiles of ccRCC tumour cells *in vitro* and causes differential transcript usage of the immune checkpoint genes *CD24* and *PD-L1*. Finally, IFNγ and TNF treatment causes global lengthening of mRNA poly(A) tails, but different isoforms of the same gene can display differential responses. This study demonstrates the ability of long-read RNA sequencing to integrate gene expression data with multiple mRNA regulatory events at a single molecule resolution. Future improvements and wider implementation of the technology will contribute to unravelling the complexity of the cancer transcriptome.

Chapter 7

Appendix

	Ensembl ID	Gene Symbol	Biotype	Description	Log_2FoldChange	p_{adj}
1	ENSG00000156395	SORCS3	protein_coding	sortilin related VPS10 domain containing receptor 3	5.9758	2.32E-08
2	ENSG00000181847	TIGIT	protein_coding	T cell immunoreceptor with Ig and ITIM domains	-2.6793	5.25E-04
3	ENSG00000105374	NKG7	protein_coding	natural killer cell granule protein 7	-2.7744	5.25E-04
4	ENSG00000113088	GZMK	protein_coding	granzyme K	-2.9823	8.71E-04
5	ENSG00000172543	CTSW	protein_coding	cathepsin W	-2.7764	1.06E-03
6	ENSG00000182240	BACE2	protein_coding	beta-secretase 2	2.0720	1.22E-03
7	ENSG00000163508	EOMES	protein_coding	eomesodermin	-3.0677	3.54E-03
8	ENSG00000145284	SCD5	protein_coding	stearoyl-CoA desaturase 5	2.6385	3.54E-03
9	ENSG00000277734	TRAC	TR_C_gene	T cell receptor alpha constant	-2.5722	3.54E-03
10	ENSG00000204252	HLA-DOA	protein_coding	major histocompatibility complex, class II, DO alpha	-2.0705	4.06E-03
11	ENSG00000077984	CST7	protein_coding	cystatin F	-3.0239	4.06E-03
12	ENSG00000115009	CCL20	protein_coding	C-C motif chemokine ligand 20	4.1538	4.07E-03
13	ENSG00000153563	CD8A	protein_coding	CD8a molecule	-2.9169	4.07E-03
14	ENSG00000084674	APOB	protein_coding	apolipoprotein B	6.2096	8.01E-03
15	ENSG00000101082	SLA2	protein_coding	Src like adaptor 2	-2.3214	9.37E-03
16	ENSG00000145649	GZMA	protein_coding	granzyme A	-2.1944	1.02E-02
17	ENSG00000089692	LAG3	protein_coding	lymphocyte activating 3	-3.1803	1.02E-02
18	ENSG00000211772	TRBC2	TR_C_gene	T cell receptor beta constant 2	-2.2302	1.21E-02
19	ENSG00000049249	TNFRSF9	protein_coding	TNF receptor superfamily member 9	-2.6865	1.25E-02
20	ENSG00000163564	PYHIN1	protein_coding	pyrin and HIN domain family member 1	-2.7398	1.25E-02
21	ENSG00000027869	SH2D2A	protein_coding	SH2 domain containing 2A	-2.0125	1.63E-02
22	ENSG00000275302	CCL4	protein_coding	C-C motif chemokine ligand 4	-2.0013	1.63E-02
23	ENSG00000172116	CD8B	protein_coding	CD8b molecule	-3.1705	1.67E-02
24	ENSG00000188389	PDCD1	protein_coding	programmed cell death 1	-3.3466	1.67E-02
25	ENSG00000148600	CDHR1	protein_coding	cadherin related family member 1	-5.8014	1.67E-02
26	ENSG00000160654	CD3G	protein_coding	CD3 gamma subunit of T-cell receptor complex	-2.2751	1.71E-02
27	ENSG00000197471	SPN	protein_coding	sialophorin	-2.1205	1.94E-02
28	ENSG00000137825	ITPKA	protein_coding	inositol-trisphosphate 3-kinase A	4.7044	2.35E-02
29	ENSG00000202538	RNU4-2	snRNA	RNA, U4 small nuclear 2	-2.6781	2.57E-02
30	ENSG00000167286	CD3D	protein_coding	CD3 delta subunit of T-cell receptor complex	-2.3153	3.07E-02
31	ENSG00000132855	ANGPTL3	protein_coding	angiopoietin like 3	-4.1111	3.17E-02
32	ENSG00000177822	TENM3-AS1	lncRNA	TENM3 antisense RNA 1	4.9254	3.22E-02
33	ENSG00000186810	CXCR3	protein_coding	C-X-C motif chemokine receptor 3	-2.3150	3.22E-02
34	ENSG00000013588	GPRC5A	protein_coding	G protein-coupled receptor class C group 5 member A	4.3243	3.22E-02
35	ENSG00000079263	SP140	protein_coding	SP140 nuclear body protein	-2.1908	3.35E-02

Table 7.1 DEGs between recurrent and non-recurrent ccRCC tumours profiled by reference genome aligned DRS

	Ensembl ID	Gene Symbol	Biotype	Description	Log_2FoldChange	p_{adj}
36	ENSG00000160185	UBASH3A	protein_coding	ubiquitin associated and SH3 domain containing A	-2.3670	3.75E-02
37	ENSG00000154451	GBP5	protein_coding	guanylate binding protein 5	-2.3455	3.96E-02
38	ENSG00000026751	SLAMF7	protein_coding	SLAM family member 7	-2.0962	4.05E-02
39	ENSG00000167608	TMC4	protein_coding	transmembrane channel like 4	3.4598	4.05E-02
40	ENSG00000151062	CACNA2D4	protein_coding	calcium voltage-gated channel auxiliary subunit alpha2delta 4	2.8263	4.51E-02
41	ENSG00000133477	FAM83F	protein_coding	family with sequence similarity 83 member F	-5.5827	4.51E-02
42	ENSG00000271503	CCL5	protein_coding	C-C motif chemokine ligand 5	-2.0824	4.58E-02
43	ENSG00000270722	RNVU1-31	snRNA	RNA, variant U1 small nuclear 31	-2.2317	4.65E-02
44	ENSG00000147138	GPR174	protein_coding	G protein-coupled receptor 174	-2.3419	4.65E-02
45	ENSG00000232453	LINC02777	lncRNA	long intergenic non-protein coding RNA 2777	3.4527	4.69E-02
46	ENSG00000276649		lncRNA	novel transcript, antisense to SIRPA	2.2227	4.77E-02
47	ENSG00000146678	IGFBP1	protein_coding	insulin like growth factor binding protein 1	8.6205	5.05E-02
48	ENSG00000176919	C8G	protein_coding	complement C8 gamma chain	3.7615	5.28E-02
49	ENSG00000163606	CD200R1	protein_coding	CD200 receptor 1	-2.3814	5.35E-02
50	ENSG00000137078	SIT1	protein_coding	signaling threshold regulating transmembrane adaptor 1	-2.0356	7.44E-02
51	ENSG00000229754	CXCR2P1	transcribed_unprocessed_pseudogene	C-X-C motif chemokine receptor 2 pseudogene 1	-3.8093	8.30E-02
52	ENSG00000089012	SIRPG	protein_coding	signal regulatory protein gamma	-2.2888	8.30E-02
53	ENSG00000166796	LDHC	protein_coding	lactate dehydrogenase C	3.8192	8.34E-02
54	ENSG00000105996	HOXA2	protein_coding	homeobox A2	-3.0764	8.63E-02
55	ENSG00000198846	TOX	protein_coding	thymocyte selection associated high mobility group box	-2.4069	8.63E-02
56	ENSG00000115041	KCNIP3	protein_coding	potassium voltage-gated channel interacting protein 3	2.2159	8.69E-02
57	ENSG00000196878	LAMB3	protein_coding	laminin subunit beta 3	2.6536	8.71E-02
58	ENSG00000004468	CD38	protein_coding	CD38 molecule	-2.3405	8.72E-02
59	ENSG00000111728	ST8SIA1	protein_coding	ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase 1	-2.7527	9.48E-02
60	ENSG00000006740	ARHGAP44	protein_coding	Rho GTPase activating protein 44	3.5627	9.54E-02
61	ENSG00000117560	FASLG	protein_coding	Fas ligand	-2.1567	9.57E-02
62	ENSG00000210151	MT-TS1	Mt_tRNA	mitochondrially encoded tRNA-Ser (UCN) 1	-3.1101	9.68E-02
63	ENSG00000168389	MFSD2A	protein_coding	major facilitator superfamily domain containing 2A	2.5617	9.73E-02
64	ENSG00000229743	LINC01159	lncRNA	long intergenic non-protein coding RNA 1159	2.3433	9.73E-02
65	ENSG00000105427	CNFN	protein_coding	cornifelin	2.3768	9.73E-02
66	ENSG00000225826	LINC00626	lncRNA	long intergenic non-protein coding RNA 626	4.2370	9.84E-02
67	ENSG00000124875	CXCL6	protein_coding	C-X-C motif chemokine ligand 6	4.6556	9.84E-02
68	ENSG00000148702	HABP2	protein_coding	hyaluronan binding protein 2	6.3113	9.84E-02

Table 7.1 (cont.) DEGs between recurrent and non-recurrent ccRCC tumours profiled by reference genome aligned DRS

	Ensembl ID	Gene Symbol	Biotype	Description	Log_2FoldChange	p_{adj}
1	ENSG00000156395	SORCS3	protein_coding	sortilin related VPS10 domain containing receptor 3	6.2563	4.89E-08
2	ENSG00000105374	NKG7	protein_coding	natural killer cell granule protein 7	-2.7936	4.21E-04
3	ENSG00000113088	GZMK	protein_coding	granzyme K	-3.0141	7.38E-04
4	ENSG00000077984	CST7	protein_coding	cystatin F	-3.0711	5.24E-03
5	ENSG00000100385	IL2RB	protein_coding	interleukin 2 receptor subunit beta	-2.0395	5.24E-03
6	ENSG00000145284	SCD5	protein_coding	stearoyl-CoA desaturase 5	2.5949	5.99E-03
7	ENSG00000084674	APOB	protein_coding	apolipoprotein B	6.2777	6.74E-03
8	ENSG00000163564	PYHIN1	protein_coding	pyrin and HIN domain family member 1	-2.8305	6.74E-03
9	ENSG00000277734	TRAC	TR_C_gene	T cell receptor alpha constant	-2.5075	6.74E-03
10	ENSG00000139193	CD27	protein_coding	CD27 molecule	-3.1845	8.06E-03
11	ENSG00000145649	GZMA	protein_coding	granzyme A	-2.1839	1.00E-02
12	ENSG00000182240	BACE2	protein_coding	beta-secretase 2	2.0682	1.25E-02
13	ENSG00000115009	CCL20	protein_coding	C-C motif chemokine ligand 20	4.9642	1.46E-02
14	ENSG00000089692	LAG3	protein_coding	lymphocyte activating 3	-3.2627	1.90E-02
15	ENSG00000147138	GPR174	protein_coding	G protein-coupled receptor 174	-2.4198	2.81E-02
16	ENSG00000013588	GPRC5A	protein_coding	G protein-coupled receptor class C group 5 member A	4.4363	3.82E-02
17	ENSG00000132855	ANGPTL3	protein_coding	angiopoietin like 3	-4.0829	3.88E-02
18	ENSG00000211772	TRBC2	TR_C_gene	T cell receptor beta constant 2	-2.8379	3.88E-02
19	ENSG00000206503	HLA-A	protein_coding	major histocompatibility complex, class I, A	-4.6910	3.97E-02
20	ENSG00000137825	ITPKA	protein_coding	inositol-trisphosphate 3-kinase A	4.5264	4.26E-02
21	ENSG00000154451	GBP5	protein_coding	guanylate binding protein 5	-2.6112	4.26E-02
22	ENSG00000160185	UBASH3A	protein_coding	ubiquitin associated and SH3 domain containing A	-2.2383	4.26E-02
23	ENSG00000160654	CD3G	protein_coding	CD3 gamma subunit of T-cell receptor complex	-2.2020	4.26E-02
24	ENSG00000167286	CD3D	protein_coding	CD3 delta subunit of T-cell receptor complex	-2.3397	4.26E-02
25	ENSG00000172116	CD8B	protein_coding	CD8b molecule	-3.2585	4.26E-02
26	ENSG00000012504	NR1H4	protein_coding	nuclear receptor subfamily 1 group H member 4	2.8814	4.55E-02
27	ENSG00000229754	CXCR2P1	transcribed_unprocessed_pseudogene	C-X-C motif chemokine receptor 2 pseudogene 1	-4.0386	4.55E-02
28	ENSG00000133477	FAM83F	protein_coding	family with sequence similarity 83 member F	-5.7248	5.47E-02
29	ENSG00000204653	ASPDH	protein_coding	aspartate dehydrogenase domain containing	2.7868	5.47E-02
30	ENSG00000276849	TRBC2	TR_C_gene	T cell receptor beta constant 2	-2.5361	5.97E-02
31	ENSG00000198846	TOX	protein_coding	thymocyte selection associated high mobility group box	-2.5599	6.48E-02
32	ENSG00000148600	CDHR1	protein_coding	cadherin related family member 1	-5.7871	6.60E-02
33	ENSG00000004468	CD38	protein_coding	CD38 molecule	-2.4644	8.31E-02
34	ENSG00000186715	MST1L	transcribed_unprocessed_pseudogene	macrophage stimulating 1 like (pseudogene)	3.3649	9.26E-02

Table 7.2 DEGs between recurrent and non-recurrent ccRCC tumours profiled by reference transcriptome aligned DRS

	Ensembl ID	Gene Symbol	Biotype	Description	Log_2FoldChange	p_{adj}
1	ENSG00000229740		lncRNA	novel transcript	22.7129	5.73E-10
2	ENSG00000248515		lncRNA	novel transcript	22.6095	5.73E-10
3	ENSG00000276241		lncRNA	novel transcript, antisense TBC1D3B	-4.8518	1.05E-08
4	ENSG00000172116	CD8B	protein_coding	CD8b molecule	-3.5990	1.01E-03
5	ENSG00000234261		lncRNA	novel transcript	-3.9535	1.01E-03
6	ENSG00000188389	PDCD1	protein_coding	programmed cell death 1	-3.3777	1.78E-03
7	ENSG00000139193	CD27	protein_coding	CD27 molecule	-3.0987	1.78E-03
8	ENSG00000277089	CCL3-AS1	lncRNA	CCL3 antisense RNA 1	-3.5249	1.78E-03
9	ENSG00000163508	EOMES	protein_coding	eomesodermin	-2.8084	2.00E-03
10	ENSG00000163606	CD200R1	protein_coding	CD200 receptor 1	-2.6771	2.00E-03
11	ENSG00000277632	CCL3	protein_coding	C-C motif chemokine ligand 3	-2.2991	2.28E-03
12	ENSG00000105374	NKG7	protein_coding	natural killer cell granule protein 7	-2.8367	2.31E-03
13	ENSG00000147138	GPR174	protein_coding	G protein-coupled receptor 174	-2.4459	2.75E-03
14	ENSG00000133477	FAM83F	protein_coding	family with sequence similarity 83 member F	-5.1938	2.75E-03
15	ENSG00000049249	TNFRSF9	protein_coding	TNF receptor superfamily member 9	-3.1534	2.88E-03
16	ENSG00000177494	ZBED2	protein_coding	zinc finger BED-type containing 2	-3.4242	2.88E-03
17	ENSG00000225826	LINC00626	lncRNA	long intergenic non-protein coding RNA 626	3.2897	3.19E-03
18	ENSG00000153563	CD8A	protein_coding	CD8a molecule	-3.0758	3.19E-03
19	ENSG00000128040	SPINK2	protein_coding	serine peptidase inhibitor Kazal type 2	5.0985	3.19E-03
20	ENSG00000113088	GZMK	protein_coding	granzyme K	-2.9725	3.19E-03
21	ENSG00000261801	LOXL1-AS1	lncRNA	LOXL1 antisense RNA 1	2.7973	3.40E-03
22	ENSG00000197057	DTHD1	protein_coding	death domain containing 1	-3.0433	3.43E-03
23	ENSG00000167608	TMC4	protein_coding	transmembrane channel like 4	3.4241	3.47E-03
24	ENSG00000277734	TRAC	TR_C_gene	T cell receptor alpha constant	-2.2422	3.55E-03
25	ENSG00000275025		lncRNA	novel transcript, antisense to PDK2	2.0826	3.55E-03
26	ENSG00000084674	APOB	protein_coding	apolipoprotein B	6.5111	3.97E-03
27	ENSG00000254126	CD8B2	protein_coding	CD8b2 molecule	-5.2703	3.97E-03
28	ENSG00000146678	IGFBP1	protein_coding	insulin like growth factor binding protein 1	8.0826	3.97E-03
29	ENSG00000176919	C8G	protein_coding	complement C8 gamma chain	3.8171	3.97E-03
30	ENSG00000166796	LDHC	protein_coding	lactate dehydrogenase C	3.9963	3.97E-03
31	ENSG00000089692	LAG3	protein_coding	lymphocyte activating 3	-2.9374	3.97E-03
32	ENSG00000225555		lncRNA	novel transcript	2.7178	3.97E-03
33	ENSG00000234663	LINC01934	lncRNA	long intergenic non-protein coding RNA 1934	-2.9449	4.40E-03
34	ENSG00000172543	CTSW	protein_coding	cathepsin W	-2.6144	4.40E-03
35	ENSG00000117560	FASLG	protein_coding	Fas ligand	-2.9479	4.52E-03

Table 7.3 DEGs between recurrent and non-recurrent ccRCC tumours profiled by reference genome aligned PCS

	Ensembl ID	Gene Symbol	Biotype	Description	Log_2FoldChange	p_{adj}
36	ENSG00000254364		lncRNA	novel transcript	4.7261	4.59E-03
37	ENSG00000077984	CST7	protein_coding	cystatin F	-2.8627	4.59E-03
38	ENSG00000163739	CXCL1	protein_coding	C-X-C motif chemokine ligand 1	4.5808	6.01E-03
39	ENSG00000286039		lncRNA	novel transcript	4.2567	6.09E-03
40	ENSG00000184613	NELL2	protein_coding	neural EGFL like 2	-2.7565	7.55E-03
41	ENSG00000211772	TRBC2	TR_C_gene	T cell receptor beta constant 2	-2.1443	7.75E-03
42	ENSG00000198846	TOX	protein_coding	thymocyte selection associated high mobility group box	-2.6027	7.75E-03
43	ENSG00000278030	TRBV7-9	TR_V_gene	T cell receptor beta variable 7-9	-3.5023	8.11E-03
44	ENSG00000154027	AK5	protein_coding	adenylate kinase 5	4.0734	8.89E-03
45	ENSG00000232453	LINC02777	lncRNA	long intergenic non-protein coding RNA 2777	3.5451	8.89E-03
46	ENSG00000145284	SCD5	protein_coding	stearoyl-CoA desaturase 5	2.0719	8.89E-03
47	ENSG00000105996	HOXA2	protein_coding	homeobox A2	-2.7624	8.89E-03
48	ENSG00000013588	GPRC5A	protein_coding	G protein-coupled receptor class C group 5 member A	4.1916	8.89E-03
49	ENSG00000015413	DPEP1	protein_coding	dipeptidase 1	-5.1038	8.89E-03
50	ENSG00000271503	CCL5	protein_coding	C-C motif chemokine ligand 5	-2.5762	8.89E-03
51	ENSG00000185433	LINC00158	lncRNA	long intergenic non-protein coding RNA 158	-2.8242	8.89E-03
52	ENSG00000182240	BACE2	protein_coding	beta-secretase 2	2.1052	8.89E-03
53	ENSG00000186810	CXCR3	protein_coding	C-X-C motif chemokine receptor 3	-2.4552	1.00E-02
54	ENSG00000183395	PMCH	protein_coding	pro-melanin concentrating hormone	-2.7213	1.00E-02
55	ENSG00000229754	CXCR2P1	transcribed_unprocessed_pseudogene	C-X-C motif chemokine receptor 2 pseudogene 1	-3.9142	1.05E-02
56	ENSG00000180644	PRF1	protein_coding	perforin 1	-2.0877	1.05E-02
57	ENSG00000173626	TRAPPC3L	protein_coding	trafficking protein particle complex subunit 3L	-2.8232	1.11E-02
58	ENSG00000199377	RNU5F-1	snRNA	RNA, U5F small nuclear 1	-2.0625	1.20E-02
59	ENSG00000109255	NMU	protein_coding	neuromedin U	5.3832	1.24E-02
60	ENSG00000107593	PKD2L1	protein_coding	polycystin 2 like 1, transient receptor potential cation channel	-2.3370	1.31E-02
61	ENSG00000135480	KRT7	protein_coding	keratin 7	2.8181	1.31E-02
62	ENSG00000183570	PCBP3	protein_coding	poly(rC) binding protein 3	6.3074	1.31E-02
63	ENSG00000225079	FTH1P22	processed_pseudogene	ferritin heavy chain 1 pseudogene 22	-3.0507	1.35E-02
64	ENSG00000148600	CDHR1	protein_coding	cadherin related family member 1	-4.9266	1.38E-02
65	ENSG00000181847	TIGIT	protein_coding	T cell immunoreceptor with Ig and ITIM domains	-2.2153	1.39E-02
66	ENSG00000137078	SIT1	protein_coding	signaling threshold regulating transmembrane adaptor 1	-2.4182	1.39E-02
67	ENSG00000185668	POU3F1	protein_coding	POU class 3 homeobox 1	-3.5502	1.43E-02
68	ENSG00000286084		lncRNA	novel transcript	-2.6396	1.49E-02
69	ENSG00000271005	CTHRC1P1	processed_pseudogene	collagen triple helix repeat containing 1 pseudogene 1	-5.0930	1.49E-02
70	ENSG00000167286	CD3D	protein_coding	CD3 delta subunit of T-cell receptor complex	-2.4641	1.49E-02

Table 7.3 (cont.) **DEGs between recurrent and non-recurrent ccRCC tumours profiled by reference genome aligned PCS**

	Ensembl ID	Gene Symbol	Biotype	Description	Log_2FoldChange	p_{adj}
71	ENSG00000152207	CYSLTR2	protein_coding	cysteinyl leukotriene receptor 2	-3.1342	1.49E-02
72	ENSG00000137825	ITPKA	protein_coding	inositol-trisphosphate 3-kinase A	3.8131	1.49E-02
73	ENSG00000275302	CCL4	protein_coding	C-C motif chemokine ligand 4	-2.0511	1.49E-02
74	ENSG00000145649	GZMA	protein_coding	granzyme A	-2.0014	1.51E-02
75	ENSG00000160654	CD3G	protein_coding	CD3 gamma subunit of T-cell receptor complex	-2.1556	1.51E-02
76	ENSG00000006740	ARHGAP44	protein_coding	Rho GTPase activating protein 44	3.2994	1.52E-02
77	ENSG00000235304	LINC01281	lncRNA	long intergenic non-protein coding RNA 1281	-4.2835	1.55E-02
78	ENSG00000156395	SORCS3	protein_coding	sortilin related VPS10 domain containing receptor 3	5.9958	1.59E-02
79	ENSG00000266088		lncRNA	novel transcript	-2.5765	1.59E-02
80	ENSG00000254427	LINC02696	lncRNA	long intergenic non-protein coding RNA 2696	3.0344	1.61E-02
81	ENSG00000247699	FABP6-AS1	lncRNA	FABP6 antisense RNA 1	3.2545	1.62E-02
82	ENSG00000116299	ELAPOR1	protein_coding	endosome-lysosome associated apoptosis and autophagy regulator 1	-2.4591	1.85E-02
83	ENSG00000115009	CCL20	protein_coding	C-C motif chemokine ligand 20	3.8192	1.85E-02
84	ENSG00000287347		lncRNA	novel transcript	-5.3471	1.85E-02
85	ENSG00000164112	SMIM43	protein_coding	small integral membrane protein 43	-3.3484	1.93E-02
86	ENSG00000105427	CNFN	protein_coding	cornifelin	2.4190	1.94E-02
87	ENSG00000163884	KLF15	protein_coding	Kruppel like factor 15	2.1314	1.95E-02
88	ENSG00000130173	ANGPTL8	protein_coding	angiopoietin like 8	6.8663	1.97E-02
89	ENSG00000265185	SNORD3B-1	snoRNA	small nucleolar RNA, C/D box 3B-1	-3.5068	2.04E-02
90	ENSG00000004468	CD38	protein_coding	CD38 molecule	-2.2787	2.12E-02
91	ENSG00000151062	CACNA2D4	protein_coding	calcium voltage-gated channel auxiliary subunit alpha2delta 4	3.2298	2.12E-02
92	ENSG00000152192	POU4F1	protein_coding	POU class 4 homeobox 1	-5.2023	2.12E-02
93	ENSG00000154451	GBP5	protein_coding	guanylate binding protein 5	-2.2586	2.12E-02
94	ENSG00000163519	TRAT1	protein_coding	T cell receptor associated transmembrane adaptor 1	-2.2153	2.12E-02
95	ENSG00000198851	CD3E	protein_coding	CD3 epsilon subunit of T-cell receptor complex	-2.2247	2.15E-02
96	ENSG00000248243	LINC02014	lncRNA	long intergenic non-protein coding RNA 2014	4.7178	2.16E-02
97	ENSG00000177822	TENM3-AS1	lncRNA	TENM3 antisense RNA 1	3.7944	2.31E-02
98	ENSG00000250770		transcribed_unprocessed_pseudogene	tetraspanin 11 (TSPAN11) pseudogene	4.4273	2.31E-02
99	ENSG00000224167	LINC01357	lncRNA	long intergenic non-protein coding RNA 1357	-2.2965	2.34E-02
100	ENSG00000101425	BPI	protein_coding	bactericidal permeability increasing protein	3.7255	2.34E-02
101	ENSG00000163564	PYHIN1	protein_coding	pyrin and HIN domain family member 1	-2.0215	2.49E-02
102	ENSG00000235688	SNTG2-AS1	lncRNA	SNTG2 antisense RNA 1	6.5182	2.60E-02
103	ENSG00000089012	SIRPG	protein_coding	signal regulatory protein gamma	-2.4870	2.60E-02
104	ENSG00000257048	LINC02417	lncRNA	long intergenic non-protein coding RNA 2417	5.4794	2.62E-02
105	ENSG00000274767		lncRNA	novel transcript, antisense CCL3L3	-2.6285	2.62E-02

Table 7.3 (cont.) DEGs between recurrent and non-recurrent ccRCC tumours profiled by reference genome aligned PCS

	Ensembl ID	Gene Symbol	Biotype	Description	Log_2FoldChange	p_{adj}
106	ENSG00000130540	SULT4A1	protein_coding	sulfotransferase family 4A member 1	5.7841	2.62E-02
107	ENSG00000171658	NMRAL2P	transcribed_unprocessed_pseudogene	NmrA like redox sensor 2, pseudogene	3.8867	2.63E-02
108	ENSG00000026751	SLAMF7	protein_coding	SLAM family member 7	-2.0634	2.78E-02
109	ENSG00000286481		lncRNA	novel transcript	2.6107	2.78E-02
110	ENSG00000238171		lncRNA	novel transcript	-4.0056	2.78E-02
111	ENSG00000272282	LINC02084	lncRNA	long intergenic non-protein coding RNA 2084	-2.2935	2.78E-02
112	ENSG00000253537	PCDHGA7	protein_coding	protocadherin gamma subfamily A, 7	2.9443	2.78E-02
113	ENSG00000211734	TRBV5-1	TR_V_gene	T cell receptor beta variable 5-1	-2.6537	2.78E-02
114	ENSG00000245522	LINC02709	lncRNA	long intergenic non-protein coding RNA 2709	2.5021	2.78E-02
115	ENSG00000237702	TRBV3-1	TR_V_gene	T cell receptor beta variable 3-1	-3.1253	2.80E-02
116	ENSG00000228509		lncRNA	novel transcript, antisense to NAB1	-2.3995	2.90E-02
117	ENSG00000211810	TRAV29DV5	TR_V_gene	T cell receptor alpha variable 29/delta variable 5	-3.5356	2.98E-02
118	ENSG00000161267	BDH1	protein_coding	3-hydroxybutyrate dehydrogenase 1	2.8923	3.00E-02
119	ENSG00000115041	KCNIP3	protein_coding	potassium voltage-gated channel interacting protein 3	2.4485	3.04E-02
120	ENSG00000004799	PDK4	protein_coding	pyruvate dehydrogenase kinase 4	2.0135	3.13E-02
121	ENSG00000211727	TRBV7-6	TR_V_gene	T cell receptor beta variable 7-6	-3.2373	3.13E-02
122	ENSG00000211777	TRAV3	TR_V_gene	T cell receptor alpha variable 3	-4.0043	3.28E-02
123	ENSG00000112137	PHACTR1	protein_coding	phosphatase and actin regulator 1	2.5826	3.31E-02
124	ENSG00000135069	PSAT1	protein_coding	phosphoserine aminotransferase 1	3.1226	3.31E-02
125	ENSG00000104953	TLE6	protein_coding	TLE family member 6, subcortical maternal complex member	2.9030	3.31E-02
126	ENSG00000273102		lncRNA	novel transcript	3.3979	3.32E-02
127	ENSG00000081051	AFP	protein_coding	alpha fetoprotein	2.8469	3.35E-02
128	ENSG00000230020	NHS-AS1	lncRNA	NHS antisense RNA 1	3.7894	3.35E-02
129	ENSG00000238120	LINC01589	lncRNA	long intergenic non-protein coding RNA 1589	3.7506	3.48E-02
130	ENSG00000187288	CIDEC	protein_coding	cell death inducing DFFA like effector c	6.2014	3.55E-02
131	ENSG00000185652	NTF3	protein_coding	neurotrophin 3	3.3306	3.80E-02
132	ENSG00000259511	UBE2Q2L	transcribed_unprocessed_pseudogene	ubiquitin conjugating enzyme E2 Q2 like	2.3873	4.06E-02
133	ENSG00000232518		lncRNA	novel transcript	-3.5361	4.06E-02
134	ENSG00000236751	LINC01186	lncRNA	long intergenic non-protein coding RNA 1186	2.0144	4.06E-02
135	ENSG00000234789		lncRNA	novel transcript	5.3551	4.06E-02
136	ENSG00000262074	SNORD3B-2	snoRNA	small nucleolar RNA, C/D box 3B-2	-2.1544	4.06E-02
137	ENSG00000261012		lncRNA	novel transcript	3.6010	4.08E-02
138	ENSG00000257067	LINC02703	lncRNA	long intergenic non-protein coding RNA 2703	3.7568	4.16E-02
139	ENSG00000184828	ZBTB7C	protein_coding	zinc finger and BTB domain containing 7C	3.5471	4.23E-02
140	ENSG00000232862		transcribed_processed_pseudogene	implantation-associated protein (DKFZp564K142) pseudogene	-4.4846	4.39E-02

Table 7.3 (cont.) **DEGs between recurrent and non-recurrent ccRCC tumours profiled by reference genome aligned PCS**

	Ensembl ID	Gene Symbol	Biotype	Description	Log_2FoldChange	p_{adj}
141	ENSG00000283063	TRBV6-2	TR_V_gene	T cell receptor beta variable 6-2	-5.2354	4.43E-02
142	ENSG00000257924	LINC02416	lncRNA	long intergenic non-protein coding RNA 2416	-4.5207	4.43E-02
143	ENSG00000196878	LAMB3	protein_coding	laminin subunit beta 3	2.6480	4.47E-02
144	ENSG00000074211	PPP2R2C	protein_coding	protein phosphatase 2 regulatory subunit Bgamma	5.8644	4.47E-02
145	ENSG00000211788	TRAV13-1	TR_V_gene	T cell receptor alpha variable 13-1	-2.0060	4.47E-02
146	ENSG00000086696	HSD17B2	protein_coding	hydroxysteroid 17-beta dehydrogenase 2	2.2091	4.47E-02
147	ENSG00000185168	LINC00482	lncRNA	long intergenic non-protein coding RNA 482	2.9538	4.50E-02
148	ENSG00000202538	RNU4-2	snRNA	RNA, U4 small nuclear 2	-2.5267	4.52E-02
149	ENSG00000225720		transcribed_unprocessed_pseudogene	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3 (APOBEC3) family pseudogene	-2.6595	4.79E-02
150	ENSG00000155719	OTOA	protein_coding	otoancorin	-2.0076	4.81E-02
151	ENSG00000198910	L1CAM	protein_coding	L1 cell adhesion molecule	-2.0026	4.84E-02
152	ENSG00000167165	UGT1A6	protein_coding	UDP glucuronosyltransferase family 1 member A6	2.0185	5.00E-02
153	ENSG00000186493	C5orf38	lncRNA	chromosome 5 open reading frame 38	3.2715	5.00E-02
154	ENSG00000198785	GRIN3A	protein_coding	glutamate ionotropic receptor NMDA type subunit 3A	-3.6114	5.00E-02
155	ENSG00000126562	WNK4	protein_coding	WNK lysine deficient protein kinase 4	3.1989	5.00E-02
156	ENSG00000286587		lncRNA	novel transcript	5.0318	5.00E-02
157	ENSG00000073792	IGF2BP2	protein_coding	insulin like growth factor 2 mRNA binding protein 2	3.7695	5.00E-02
158	ENSG00000074966	TXK	protein_coding	TXK tyrosine kinase	-2.0631	5.00E-02
159	ENSG00000249574		lncRNA	novel transcript	3.4475	5.00E-02
160	ENSG00000147570	DNAJC5B	protein_coding	DnaJ heat shock protein family (Hsp40) member C5 beta	-2.5597	5.00E-02
161	ENSG00000156127	BATF	protein_coding	basic leucine zipper ATF-like transcription factor	-2.0360	5.00E-02
162	ENSG00000118271	TTR	protein_coding	transthyretin	5.7834	5.00E-02
163	ENSG00000287566		lncRNA	novel transcript	5.9837	5.07E-02
164	ENSG00000278558	TMEM191B	protein_coding	transmembrane protein 191B	-3.9272	5.10E-02
165	ENSG00000258346		lncRNA	novel transcript	3.7233	5.17E-02
166	ENSG00000186897	C1QL4	protein_coding	complement C1q like 4	-3.3886	5.17E-02
167	ENSG00000237484	LINC01684	lncRNA	long intergenic non-protein coding RNA 1684	-2.2699	5.17E-02
168	ENSG00000128918	ALDH1A2	protein_coding	aldehyde dehydrogenase 1 family member A2	2.0315	5.27E-02
169	ENSG00000213279		lncRNA	novel transcript	3.2438	5.60E-02
170	ENSG00000148702	HABP2	protein_coding	hyaluronan binding protein 2	5.7425	5.67E-02
171	ENSG00000073150	PANX2	protein_coding	pannexin 2	3.7966	5.67E-02
172	ENSG00000247193		lncRNA	novel transcript, antisense to CENTD1	-2.4103	5.76E-02
173	ENSG00000243537	RPL32P20	processed_pseudogene	ribosomal protein L32 pseudogene 20	2.8043	5.76E-02
174	ENSG00000255968		lncRNA	novel transcript to ITPR2	3.4090	5.76E-02
175	ENSG00000149633	KIAA1755	protein_coding	KIAA1755	3.0108	5.76E-02

Table 7.3 (cont.) DEGs between recurrent and non-recurrent ccRCC tumours profiled by reference genome aligned PCS

	Ensembl ID	Gene Symbol	Biotype	Description	Log_2FoldChange	p_{adj}
176	ENSG00000116032	GRIN3B	protein_coding	glutamate ionotropic receptor NMDA type subunit 3B	-5.0178	5.76E-02
177	ENSG00000099994	SUSD2	protein_coding	sushi domain containing 2	2.1444	5.76E-02
178	ENSG00000103175	WFDC1	protein_coding	WAP four-disulfide core domain 1	2.0374	5.78E-02
179	ENSG00000231948	HS1BP3-IT1	lncRNA	HS1BP3 intronic transcript 1	3.5020	5.81E-02
180	ENSG00000229743	LINC01159	lncRNA	long intergenic non-protein coding RNA 1159	2.5661	5.84E-02
181	ENSG00000165899	OTOGL	protein_coding	otogelin like	2.4621	5.86E-02
182	ENSG00000185304	RGPD2	protein_coding	RANBP2 like and GRIP domain containing 2	-2.8852	5.95E-02
183	ENSG00000272382		lncRNA	novel transcript	-2.2010	5.95E-02
184	ENSG00000088002	SULT2B1	protein_coding	sulfotransferase family 2B member 1	2.8261	5.95E-02
185	ENSG00000258875		lncRNA	novel transcript, antisense to GPR68	-2.3426	6.18E-02
186	ENSG00000226912	ISCA2P1	processed_pseudogene	iron-sulfur cluster assembly 2 pseudogene 1	5.4023	6.38E-02
187	ENSG00000111728	ST8SIA1	protein_coding	ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase 1	-2.4060	6.43E-02
188	ENSG00000180432	CYP8B1	protein_coding	cytochrome P450 family 8 subfamily B member 1	2.8881	6.48E-02
189	ENSG00000164604	GPR85	protein_coding	G protein-coupled receptor 85	-2.0485	6.63E-02
190	ENSG00000130829	DUSP9	protein_coding	dual specificity phosphatase 9	3.7215	6.63E-02
191	ENSG00000286374		lncRNA	novel transcript	5.5858	7.02E-02
192	ENSG00000180535	BHLHA15	protein_coding	basic helix-loop-helix family member a15	2.0207	7.20E-02
193	ENSG00000211710	TRBV4-1	TR_V_gene	T cell receptor beta variable 4-1	-2.3399	7.20E-02
194	ENSG00000172216	CEBPB	protein_coding	CCAAT enhancer binding protein beta	2.0711	7.90E-02
195	ENSG00000124507	PACSIN1	protein_coding	protein kinase C and casein kinase substrate in neurons 1	-2.3491	8.07E-02
196	ENSG00000222042		lncRNA	novel transcript	-4.0179	8.10E-02
197	ENSG00000154277	UCHL1	protein_coding	ubiquitin C-terminal hydrolase L1	2.4328	8.14E-02
198	ENSG00000250620	LINC02515	lncRNA	long intergenic non-protein coding RNA 2515	-3.7541	8.14E-02
199	ENSG00000255689		lncRNA	novel transcript	3.8510	8.14E-02
200	ENSG00000256714		lncRNA	novel transcript	-4.3038	8.14E-02
201	ENSG00000101255	TRIB3	protein_coding	tribbles pseudokinase 3	2.1360	8.19E-02
202	ENSG00000149571	KIRREL3	protein_coding	kirre like nephrin family adhesion molecule 3	5.2139	8.35E-02
203	ENSG00000197826	CFAP299	protein_coding	cilia and flagella associated protein 299	2.4698	8.39E-02
204	ENSG00000138109	CYP2C9	protein_coding	cytochrome P450 family 2 subfamily C member 9	3.3377	8.39E-02
205	ENSG00000128965	CHAC1	protein_coding	ChaC glutathione specific gamma-glutamylcyclotransferase 1	2.4764	8.39E-02
206	ENSG00000156414	TDRD9	protein_coding	tudor domain containing 9	3.5328	8.59E-02
207	ENSG00000266521		lncRNA	novel transcript	-3.2824	8.84E-02
208	ENSG00000205424	PCBP3-AS1	lncRNA	PCBP3 antisense RNA 1	6.3351	8.84E-02
209	ENSG00000204385	SLC44A4	protein_coding	solute carrier family 44 member 4	2.5042	9.19E-02
210	ENSG00000255202		lncRNA	novel transcript	3.3771	9.38E-02

Table 7.3 (cont.) **DEGs between recurrent and non-recurrent ccRCC tumours profiled by reference genome aligned PCS**

	Ensembl ID	Gene Symbol	Biotype	Description	Log_2FoldChange	p_{adj}
211	ENSG00000169876	MUC17	protein_coding	mucin 17, cell surface associated	9.6685	9.41E-02
212	ENSG00000272620	AFAP1-AS1	lncRNA	AFAP1 antisense RNA 1	5.5809	9.58E-02
213	ENSG00000125878	TCF15	protein_coding	transcription factor 15	2.8215	9.60E-02
214	ENSG00000238290	ERRFI1-DT	lncRNA	ERRFI1 divergent transcript	2.4445	9.64E-02
215	ENSG00000132855	ANGPTL3	protein_coding	angiopoietin like 3	-3.4697	9.78E-02
216	ENSG00000240045	STRIT1	protein_coding	small transmembrane regulator of ion transport 1	3.4174	9.78E-02
217	ENSG00000112333	NR2E1	protein_coding	nuclear receptor subfamily 2 group E member 1	-4.2088	9.78E-02
218	ENSG00000132702	HAPLN2	protein_coding	hyaluronan and proteoglycan link protein 2	2.8090	9.99E-02
219	ENSG00000128564	VGF	protein_coding	VGF nerve growth factor inducible	-5.7079	9.99E-02

Table 7.3 (cont.) DEGs between recurrent and non-recurrent ccRCC tumours profiled by reference genome aligned PCS

	Ensembl ID	Gene Symbol	Biotype	Description	Log_2FoldChange	p_{adj}
1	ENSG00000169876	MUC17	protein_coding	mucin 17, cell surface associated	24.1211	1.59E-11
2	ENSG00000172116	CD8B	protein_coding	CD8b molecule	-3.5878	1.66E-04
3	ENSG00000133477	FAM83F	protein_coding	family with sequence similarity 83 member F	-5.5527	3.65E-04
4	ENSG00000229295	HLA-DPB1	protein_coding	major histocompatibility complex, class II, DP beta 1	10.1926	3.65E-04
5	ENSG00000254126	CD8B2	protein_coding	CD8b2 molecule	-5.4280	3.65E-04
6	ENSG00000146678	IGFBP1	protein_coding	insulin like growth factor binding protein 1	8.2423	9.28E-04
7	ENSG00000211727	TRBV7-6	TR_V_gene	T cell receptor beta variable 7-6	-4.4967	9.28E-04
8	ENSG00000147570	DNAJC5B	protein_coding	DnaJ heat shock protein family (Hsp40) member C5 beta	-2.7302	1.24E-03
9	ENSG00000105374	NKG7	protein_coding	natural killer cell granule protein 7	-2.8250	1.27E-03
10	ENSG00000113088	GZMK	protein_coding	granzyme K	-2.8493	1.50E-03
11	ENSG00000211772	TRBC2	TR_C_gene	T cell receptor beta constant 2	-2.5194	1.72E-03
12	ENSG00000177494	ZBED2	protein_coding	zinc finger BED-type containing 2	-3.3809	1.83E-03
13	ENSG00000013588	GPRC5A	protein_coding	G protein-coupled receptor class C group 5 member A	4.1241	2.13E-03
14	ENSG00000117560	FASLG	protein_coding	Fas ligand	-2.8989	2.13E-03
15	ENSG00000139193	CD27	protein_coding	CD27 molecule	-3.0984	2.13E-03
16	ENSG00000077984	CST7	protein_coding	cystatin F	-2.7447	2.33E-03
17	ENSG00000084674	APOB	protein_coding	apolipoprotein B	6.7454	2.33E-03
18	ENSG00000176919	C8G	protein_coding	complement C8 gamma chain	3.8977	2.33E-03
19	ENSG00000277734	TRAC	TR_C_gene	T cell receptor alpha constant	-2.5570	2.33E-03
20	ENSG00000183395	PMCH	protein_coding	pro-melanin concentrating hormone	-3.0383	2.36E-03
21	ENSG00000225079	FTH1P22	processed_pseudogene	ferritin heavy chain 1 pseudogene 22	-3.7702	4.16E-03
22	ENSG00000228405	RNF5	protein_coding	ring finger protein 5	-6.1846	5.79E-03
23	ENSG00000128040	SPINK2	protein_coding	serine peptidase inhibitor Kazal type 2	4.8044	6.23E-03
24	ENSG00000089692	LAG3	protein_coding	lymphocyte activating 3	-2.7375	7.75E-03
25	ENSG00000167183	PRR15L	protein_coding	proline rich 15 like	4.6225	8.00E-03
26	ENSG00000172543	CTSW	protein_coding	cathepsin W	-2.6239	8.15E-03
27	ENSG00000163687	DNASE1L3	protein_coding	deoxyribonuclease 1 like 3	-2.1899	8.23E-03
28	ENSG00000089012	SIRPG	protein_coding	signal regulatory protein gamma	-2.8162	8.30E-03
29	ENSG00000182240	BACE2	protein_coding	beta-secretase 2	2.0508	8.30E-03
30	ENSG00000229754	CXCR2P1	transcribed_unprocessed_pseudogene	C-X-C motif chemokine receptor 2 pseudogene 1	-3.8223	1.07E-02
31	ENSG00000197057	DTHD1	protein_coding	death domain containing 1	-2.8671	1.07E-02
32	ENSG00000198846	TOX	protein_coding	thymocyte selection associated high mobility group box	-2.4993	1.07E-02
33	ENSG00000241622	RARRES2P1	processed_pseudogene	retinoic acid receptor responder 2 pseudogene 1	5.7042	1.07E-02
34	ENSG00000186715	MST1L	transcribed_unprocessed_pseudogene	macrophage stimulating 1 like (pseudogene)	2.7616	1.12E-02
35	ENSG00000135480	KRT7	protein_coding	keratin 7	2.7580	1.12E-02

Table 7.4 DEGs between recurrent and non-recurrent ccRCC tumours profiled by reference transcriptome aligned PCS

	Ensembl ID	Gene Symbol	Biotype	Description	Log_2FoldChange	p_{adj}
36	ENSG00000274233	CCL5	protein_coding	C-C motif chemokine ligand 5	-2.6845	1.12E-02
37	ENSG00000137825	ITPKA	protein_coding	inositol-trisphosphate 3-kinase A	3.7599	1.15E-02
38	ENSG00000206478	IER3	protein_coding	immediate early response 3	-4.1458	1.20E-02
39	ENSG00000167286	CD3D	protein_coding	CD3 delta subunit of T-cell receptor complex	-2.4056	1.23E-02
40	ENSG00000281981	TRBC1	TR_C_gene	T cell receptor beta constant 1	-5.3044	1.33E-02
41	ENSG00000160654	CD3G	protein_coding	CD3 gamma subunit of T-cell receptor complex	-2.0218	1.33E-02
42	ENSG00000163519	TRAT1	protein_coding	T cell receptor associated transmembrane adaptor 1	-2.3521	1.34E-02
43	ENSG00000185652	NTF3	protein_coding	neurotrophin 3	4.0574	1.34E-02
44	ENSG00000145284	SCD5	protein_coding	stearoyl-CoA desaturase 5	2.1153	1.38E-02
45	ENSG00000004468	CD38	protein_coding	CD38 molecule	-2.1687	1.40E-02
46	ENSG00000137078	SIT1	protein_coding	signaling threshold regulating transmembrane adaptor 1	-2.2844	1.49E-02
47	ENSG00000027869	SH2D2A	protein_coding	SH2 domain containing 2A	-2.1886	1.70E-02
48	ENSG00000115009	CCL20	protein_coding	C-C motif chemokine ligand 20	3.6735	1.82E-02
49	ENSG00000260762	ACSM5P1	unprocessed_pseudogene	acyl-CoA synthetase medium chain family member 5 pseudogene 1	2.5610	1.83E-02
50	ENSG00000152192	POU4F1	protein_coding	POU class 4 homeobox 1	-5.2332	1.83E-02
51	ENSG00000135069	PSAT1	protein_coding	phosphoserine aminotransferase 1	3.1583	1.90E-02
52	ENSG00000002746	HECW1	protein_coding	HECT, C2 and WW domain containing E3 ubiquitin protein ligase 1	4.5899	2.00E-02
53	ENSG00000049249	TNFRSF9	protein_coding	TNF receptor superfamily member 9	-3.0868	2.17E-02
54	ENSG00000156395	SORCS3	protein_coding	sortilin related VPS10 domain containing receptor 3	6.1354	2.17E-02
55	ENSG00000276849	TRBC2	TR_C_gene	T cell receptor beta constant 2	-2.2272	2.17E-02
56	ENSG00000228913	UBD	protein_coding	ubiquitin D	-5.6581	2.22E-02
57	ENSG00000273722	TMC4	protein_coding	transmembrane channel like 4	4.8340	2.22E-02
58	ENSG00000171658	NMRAL2P	transcribed_unprocessed_pseudogene	NmrA like redox sensor 2, pseudogene	4.2132	2.22E-02
59	ENSG00000185668	POU3F1	protein_coding	POU class 3 homeobox 1	-3.5037	2.22E-02
60	ENSG00000107593	PKD2L1	protein_coding	polycystin 2 like 1, transient receptor potential cation channel	-2.2624	2.25E-02
61	ENSG00000243537	RPL32P20	processed_pseudogene	ribosomal protein L32 pseudogene 20	3.1624	2.25E-02
62	ENSG00000257017	HP	protein_coding	haptoglobin	8.1861	2.37E-02
63	ENSG00000213402	PTPRCAP	protein_coding	protein tyrosine phosphatase receptor type C associated protein	-2.2772	2.75E-02
64	ENSG00000173626	TRAPPC3L	protein_coding	trafficking protein particle complex subunit 3L	-2.1712	2.76E-02
65	ENSG00000206502	POLR1H	protein_coding	RNA polymerase I subunit H	-5.9549	2.91E-02
66	ENSG00000258732		unprocessed_pseudogene	arginine-glutamic acid dipeptide (RE) repeats (RERE) pseudogene	-4.7174	2.91E-02
67	ENSG00000105427	CNFN	protein_coding	cornifelin	2.3934	2.93E-02
68	ENSG00000163606	CD200R1	protein_coding	CD200 receptor 1	-2.9314	3.06E-02
69	ENSG00000234906	APOC2	protein_coding	apolipoprotein C2	-2.1741	3.35E-02
70	ENSG00000198851	CD3E	protein_coding	CD3 epsilon subunit of T-cell receptor complex	-2.3320	4.72E-02

Table 7.4 (cont.) **DEGs between recurrent and non-recurrent ccRCC tumours profiled by reference transcriptome aligned PCS**

	Ensembl ID	Gene Symbol	Biotype	Description	Log_2FoldChange	p_{adj}
71	ENSG00000214142	RPL7P60	processed_pseudogene	ribosomal protein L7 pseudogene 60	4.9688	4.74E-02
72	ENSG00000163739	CXCL1	protein_coding	C-X-C motif chemokine ligand 1	4.5743	4.93E-02
73	ENSG00000198648	STK39	protein_coding	serine/threonine kinase 39	4.7001	4.93E-02
74	ENSG00000253537	PCDHGA7	protein_coding	protocadherin gamma subfamily A, 7	2.7579	4.93E-02
75	ENSG00000078804	TP53INP2	protein_coding	tumor protein p53 inducible nuclear protein 2	2.0736	4.97E-02
76	ENSG00000151062	CACNA2D4	protein_coding	calcium voltage-gated channel auxiliary subunit alpha2delta 4	4.3792	5.10E-02
77	ENSG00000275824	CCL4	protein_coding	C-C motif chemokine ligand 4	-2.2611	5.18E-02
78	ENSG00000157890	MEGF11	protein_coding	multiple EGF like domains 11	2.9365	5.21E-02
79	ENSG00000146122	DAAM2	protein_coding	dishevelled associated activator of morphogenesis 2	2.6795	5.34E-02
80	ENSG00000169436	COL22A1	protein_coding	collagen type XXII alpha 1 chain	3.8243	5.34E-02
81	ENSG00000183918	SH2D1A	protein_coding	SH2 domain containing 1A	-2.3412	5.34E-02
82	ENSG00000004799	PDK4	protein_coding	pyruvate dehydrogenase kinase 4	2.1269	5.48E-02
83	ENSG00000180638	SLC47A2	protein_coding	solute carrier family 47 member 2	3.3974	5.48E-02
84	ENSG00000163884	KLF15	protein_coding	Kruppel like factor 15	2.1953	5.81E-02
85	ENSG00000180432	CYP8B1	protein_coding	cytochrome P450 family 8 subfamily B member 1	2.9827	5.87E-02
86	ENSG00000154451	GBP5	protein_coding	guanylate binding protein 5	-2.0564	5.95E-02
87	ENSG00000266964	FXYD1	protein_coding	FXYD domain containing ion transport regulator 1	-3.2452	5.95E-02
88	ENSG00000271503	CCL5	protein_coding	C-C motif chemokine ligand 5	-2.0502	5.95E-02
89	ENSG00000126562	WNK4	protein_coding	WNK lysine deficient protein kinase 4	3.1346	6.00E-02
90	ENSG00000166796	LDHC	protein_coding	lactate dehydrogenase C	3.5565	6.03E-02
91	ENSG00000271005	CTHRC1P1	processed_pseudogene	collagen triple helix repeat containing 1 pseudogene 1	-4.6246	6.03E-02
92	ENSG00000213413	PVRIG	protein_coding	PVR related immunoglobulin domain containing	-2.3762	6.36E-02
93	ENSG00000111537	IFNG	protein_coding	interferon gamma	-2.8621	6.45E-02
94	ENSG00000167701	GPT	protein_coding	glutamic--pyruvic transaminase	2.2185	6.64E-02
95	ENSG00000184363	PKP3	protein_coding	plakophilin 3	7.0527	6.64E-02
96	ENSG00000227191	TRGC2	TR_C_gene	T cell receptor gamma constant 2	-2.1249	6.67E-02
97	ENSG00000232862		transcribed_processed_pseudogene	implantation-associated protein (DKFZp564K142) pseudogene	-4.4493	6.67E-02
98	ENSG00000172216	CEBPB	protein_coding	CCAAT enhancer binding protein beta	2.3011	6.76E-02
99	ENSG00000186810	CXCR3	protein_coding	C-X-C motif chemokine receptor 3	-2.5261	6.76E-02
100	ENSG00000224916	APOC4-APOC2	protein_coding	APOC4-APOC2 readthrough (NMD candidate)	-2.3638	6.76E-02
101	ENSG00000230034	PSMB8	protein_coding	proteasome 20S subunit beta 8	3.1796	6.76E-02
102	ENSG00000118271	TTR	protein_coding	transthyretin	5.8045	6.82E-02
103	ENSG00000147724	FAM135B	protein_coding	family with sequence similarity 135 member B	2.3538	6.82E-02
104	ENSG00000103175	WFDC1	protein_coding	WAP four-disulfide core domain 1	2.0445	6.90E-02

Table 7.4 (cont.) **DEGs between recurrent and non-recurrent ccRCC tumours profiled by reference transcriptome aligned PCS**

	Ensembl ID	Gene Symbol	Biotype	Description	Log_2FoldChange	p_{adj}
105	ENSG00000125878	TCF15	protein_coding	transcription factor 15	3.0471	6.90E-02
106	ENSG00000284953	CACNA2D4	protein_coding	calcium voltage-gated channel auxiliary subunit alpha2delta 4	4.1189	6.90E-02
107	ENSG00000241360	PDXP	protein_coding	pyridoxal phosphatase	2.4229	6.91E-02
108	ENSG00000137403	HLA-F	protein_coding	major histocompatibility complex, class I, F	-2.4068	7.33E-02
109	ENSG00000166407	LMO1	protein_coding	LIM domain only 1	5.4658	7.33E-02
110	ENSG00000197272	IL27	protein_coding	interleukin 27	-2.7301	7.86E-02
111	ENSG00000240045	STRIT1	protein_coding	small transmembrane regulator of ion transport 1	3.5329	8.34E-02
112	ENSG00000107984	DKK1	protein_coding	dickkopf WNT signaling pathway inhibitor 1	-2.6575	8.40E-02
113	ENSG00000164604	GPR85	protein_coding	G protein-coupled receptor 85	-2.4707	8.52E-02
114	ENSG00000203805	PLPP4	protein_coding	phospholipid phosphatase 4	3.6354	8.52E-02
115	ENSG00000143851	PTPN7	protein_coding	protein tyrosine phosphatase non-receptor type 7	-2.2140	8.71E-02
116	ENSG00000255641		protein_coding	novel protein	-3.4445	8.71E-02
117	ENSG00000223793	HLA-DQA2	protein_coding	major histocompatibility complex, class II, DQ alpha 2	-3.6939	8.78E-02
118	ENSG00000276977	PDCD1	protein_coding	programmed cell death 1	-4.3754	8.78E-02
119	ENSG00000197238	H4C11	protein_coding	H4 clustered histone 11	2.2068	8.90E-02
120	ENSG00000124615	MOCS1	protein_coding	molybdenum cofactor synthesis 1	2.3892	9.42E-02
121	ENSG00000073792	IGF2BP2	protein_coding	insulin like growth factor 2 mRNA binding protein 2	4.1634	9.43E-02
122	ENSG00000147168	IL2RG	protein_coding	interleukin 2 receptor subunit gamma	-2.0680	9.43E-02
123	ENSG00000161643	SIGLEC16	protein_coding	sialic acid binding Ig like lectin 16	-2.0569	9.43E-02
124	ENSG00000206501	PPP1R11	protein_coding	protein phosphatase 1 regulatory inhibitor subunit 11	-5.1174	9.43E-02
125	ENSG00000128965	CHAC1	protein_coding	ChaC glutathione specific gamma-glutamylcyclotransferase 1	2.3234	9.85E-02
126	ENSG00000182866	LCK	protein_coding	LCK proto-oncogene, Src family tyrosine kinase	-2.0564	9.94E-02

Table 7.4 (cont.) **DEGs between recurrent and non-recurrent ccRCC tumours profiled by reference transcriptome aligned PCS**

	ID	Description	setSize	NES	p_{adj}
1	GO:0002449	lymphocyte mediated immunity	350	-2.1911	2.00E-13
2	GO:0042110	T cell activation	500	-2.0090	2.00E-13
3	GO:0051249	regulation of lymphocyte activation	494	-2.0255	2.58E-13
4	GO:0002443	leukocyte mediated immunity	427	-2.0954	3.49E-13
5	GO:0002697	regulation of immune effector process	342	-2.0731	3.10E-11
6	GO:0050863	regulation of T cell activation	356	-2.0671	3.10E-11
7	GO:0002460	adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	359	-2.0646	3.10E-11
8	GO:1902105	regulation of leukocyte differentiation	265	-2.1908	6.51E-11
9	GO:0002696	positive regulation of leukocyte activation	388	-1.9821	2.75E-10
10	GO:0045619	regulation of lymphocyte differentiation	177	-2.3333	2.88E-10
11	GO:0050867	positive regulation of cell activation	399	-1.9834	3.05E-10
12	GO:0001906	cell killing	171	-2.3477	3.98E-10
13	GO:0051251	positive regulation of lymphocyte activation	351	-1.9880	3.98E-10
14	GO:0031341	regulation of cell killing	117	-2.4359	4.60E-10
15	GO:0030217	T cell differentiation	253	-2.1396	7.18E-10
16	GO:0002683	negative regulation of immune system process	386	-1.9797	8.62E-10
17	GO:0045580	regulation of T cell differentiation	152	-2.3240	9.95E-10
18	GO:0001909	leukocyte mediated cytotoxicity	143	-2.3520	1.24E-09
19	GO:1903706	regulation of hemopoiesis	355	-1.9576	1.51E-09
20	GO:0019882	antigen processing and presentation	186	-2.2464	1.54E-09
21	GO:0031343	positive regulation of cell killing	84	-2.5472	1.70E-09
22	GO:0048002	antigen processing and presentation of peptide antigen	137	-2.3244	2.12E-09
23	GO:0001910	regulation of leukocyte mediated cytotoxicity	104	-2.4440	2.44E-09
24	GO:0002521	leukocyte differentiation	492	-1.8251	3.74E-09
25	GO:0002366	leukocyte activation involved in immune response	263	-2.0699	4.12E-09
26	GO:0046631	alpha-beta T cell activation	166	-2.2541	7.28E-09
27	GO:0001912	positive regulation of leukocyte mediated cytotoxicity	78	-2.4340	1.08E-08
28	GO:0030098	lymphocyte differentiation	345	-1.9403	1.08E-08
29	GO:0002263	cell activation involved in immune response	267	-2.0449	1.23E-08
30	GO:0002285	lymphocyte activation involved in immune response	192	-2.1196	1.40E-08
31	GO:1903131	mononuclear cell differentiation	397	-1.8550	1.70E-08
32	GO:0007159	leukocyte cell-cell adhesion	384	-1.8759	1.78E-08
33	GO:0002703	regulation of leukocyte mediated immunity	243	-2.0620	2.03E-08
34	GO:0019884	antigen processing and presentation of exogenous antigen	106	-2.3482	2.43E-08
35	GO:0002699	positive regulation of immune effector process	245	-2.0518	2.44E-08
36	GO:0050870	positive regulation of T cell activation	256	-2.0056	2.86E-08
37	GO:1903037	regulation of leukocyte cell-cell adhesion	350	-1.8763	3.12E-08
38	GO:0022407	regulation of cell-cell adhesion	446	-1.7904	3.12E-08
39	GO:0001913	T cell mediated cytotoxicity	78	-2.3900	3.31E-08
40	GO:0002478	antigen processing and presentation of exogenous peptide antigen	96	-2.3864	3.31E-08
41	GO:0002706	regulation of lymphocyte mediated immunity	189	-2.1470	3.31E-08
42	GO:0002705	positive regulation of leukocyte mediated immunity	154	-2.2230	3.94E-08
43	GO:0046634	regulation of alpha-beta T cell activation	119	-2.2737	4.43E-08
44	GO:0002456	T cell mediated immunity	138	-2.2486	4.78E-08
45	GO:0022409	positive regulation of cell-cell adhesion	311	-1.8983	7.61E-08
46	GO:0002228	natural killer cell mediated immunity	86	-2.3639	9.88E-08
47	GO:1903039	positive regulation of leukocyte cell-cell adhesion	273	-1.9312	1.45E-07
48	GO:0044283	small molecule biosynthetic process	483	1.7601	1.77E-07
49	GO:0019883	antigen processing and presentation of endogenous antigen	69	-2.3588	2.03E-07
50	GO:0002440	production of molecular mediator of immune response	276	-1.8992	2.09E-07

Table 7.5 Top 50 GO BP terms (by p_{adj}) between non-recurrent and recurrent ccRCC profiled by DRS

	ID	Description	setSize	NES	p_{adj}
1	GO:0042110	T cell activation	482	-1.9925	8.32E-10
2	GO:0030217	T cell differentiation	260	-2.1837	7.71E-08
3	GO:1903037	regulation of leukocyte cell-cell adhesion	326	-1.9785	8.85E-07
4	GO:1903706	regulation of hemopoiesis	360	-1.9341	1.00E-06
5	GO:0050863	regulation of T cell activation	329	-1.9703	1.29E-06
6	GO:0007159	leukocyte cell-cell adhesion	363	-1.8957	1.65E-06
7	GO:1902105	regulation of leukocyte differentiation	273	-2.0045	2.72E-06
8	GO:0030098	lymphocyte differentiation	366	-1.8771	9.34E-06
9	GO:1903131	mononuclear cell differentiation	416	-1.7956	1.11E-05
10	GO:0002697	regulation of immune effector process	340	-1.8271	1.90E-05
11	GO:0050870	positive regulation of T cell activation	217	-1.9974	2.24E-05
12	GO:0002683	negative regulation of immune system process	399	-1.7733	2.24E-05
13	GO:0022407	regulation of cell-cell adhesion	435	-1.7230	2.24E-05
14	GO:0002263	cell activation involved in immune response	273	-1.9151	2.65E-05
15	GO:0002366	leukocyte activation involved in immune response	269	-1.9070	2.73E-05
16	GO:0045619	regulation of lymphocyte differentiation	177	-2.0701	3.57E-05
17	GO:1903039	positive regulation of leukocyte cell-cell adhesion	238	-1.9474	3.57E-05
18	GO:0002703	regulation of leukocyte mediated immunity	230	-1.9255	4.15E-05
19	GO:0046632	alpha-beta T cell differentiation	110	-2.0971	4.96E-05
20	GO:0045088	regulation of innate immune response	221	-1.9689	9.84E-05
21	GO:0045580	regulation of T cell differentiation	148	-2.0390	2.00E-04
22	GO:0001906	cell killing	163	-1.9785	2.04E-04
23	GO:0002228	natural killer cell mediated immunity	73	-2.2593	2.31E-04
24	GO:0046631	alpha-beta T cell activation	156	-1.9644	2.47E-04
25	GO:0002706	regulation of lymphocyte mediated immunity	170	-1.9526	3.49E-04
26	GO:0045089	positive regulation of innate immune response	135	-1.9526	3.49E-04
27	GO:0034341	response to interferon-gamma	137	-2.0120	3.50E-04
28	GO:0071346	cellular response to interferon-gamma	116	-2.0060	3.89E-04
29	GO:0002699	positive regulation of immune effector process	240	-2.1323	3.96E-04
30	GO:0002831	regulation of response to biotic stimulus	336	-1.8312	3.96E-04
31	GO:0032609	interferon-gamma production	113	-1.7204	3.96E-04
32	GO:0032649	regulation of interferon-gamma production	113	-2.0368	6.58E-04
33	GO:0002285	lymphocyte activation involved in immune response	188	-2.0368	6.58E-04
34	GO:0043367	CD4-positive, alpha-beta T cell differentiation	81	-1.8826	7.07E-04
35	GO:0002819	regulation of adaptive immune response	182	-2.1359	8.55E-04
36	GO:0001909	leukocyte mediated cytotoxicity	126	-1.8502	9.66E-04
37	GO:0002833	positive regulation of response to biotic stimulus	171	-1.9662	1.03E-03
38	GO:0022409	positive regulation of cell-cell adhesion	284	-1.8587	1.14E-03
39	GO:0042267	natural killer cell mediated cytotoxicity	70	-1.6972	1.14E-03
40	GO:0002504	antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	34	-2.2011	1.57E-03
41	GO:0032729	positive regulation of interferon-gamma production	74	-2.1515	1.86E-03
42	GO:0002685	regulation of leukocyte migration	211	-2.1192	2.34E-03
43	GO:0050900	leukocyte migration	371	-1.8049	2.34E-03
44	GO:0002705	positive regulation of leukocyte mediated immunity	136	-1.6232	2.47E-03
45	GO:0002495	antigen processing and presentation of peptide antigen via MHC class II	32	-1.9379	2.65E-03
46	GO:0002286	T cell activation involved in immune response	103	-2.1670	2.89E-03
47	GO:1902107	positive regulation of leukocyte differentiation	153	-1.9900	2.89E-03
48	GO:1903708	positive regulation of hemopoiesis	153	-1.8933	2.91E-03
49	GO:0002443	leukocyte mediated immunity	430	-1.8933	2.91E-03
50	GO:0031349	positive regulation of defense response	272	-1.5410	2.91E-03

Table 7.6 Top 50 GO BP terms (by p_{adj}) between non-recurrent and recurrent ccRCC profiled by PCS

	ID	Description	setSize	NES	p_{adj}
1	GO:0042605	peptide antigen binding	95	-2.4366	4.99E-08
2	GO:0140375	immune receptor activity	156	-2.0128	1.19E-05
3	GO:0003823	antigen binding	158	-2.0132	2.47E-05
4	GO:0032395	MHC class II receptor activity	42	-2.3815	3.28E-05
5	GO:0042277	peptide binding	323	-1.7263	3.28E-05
6	GO:0046977	TAP binding	41	-2.3136	8.19E-05
7	GO:0023023	MHC protein complex binding	77	-2.1407	8.19E-05
8	GO:0023026	MHC class II protein complex binding	67	-2.2004	1.06E-04
9	GO:0005344	oxygen carrier activity	13	-2.2572	4.50E-04
10	GO:0031720	haptoglobin binding	9	-2.1294	1.19E-03
11	GO:0042608	T cell receptor binding	24	-2.1747	2.43E-03
12	GO:0042288	MHC class I protein binding	32	-2.1244	2.98E-03
13	GO:0033218	amide binding	398	-1.4831	3.46E-03
14	GO:0005201	extracellular matrix structural constituent	131	1.7627	5.73E-03
15	GO:0042287	MHC protein binding	48	-2.0106	1.40E-02
16	GO:0016684	oxidoreductase activity, acting on peroxide as acceptor	47	-1.9456	1.59E-02
17	GO:0004601	peroxidase activity	44	-1.9615	2.13E-02
18	GO:0045296	cadherin binding	328	1.5370	2.47E-02
19	GO:0019825	oxygen binding	28	-1.9963	2.77E-02
20	GO:0046978	TAP1 binding	20	-2.0394	2.84E-02
21	GO:0008514	organic anion transmembrane transporter activity	143	1.6826	2.84E-02
22	GO:0010997	anaphase-promoting complex binding	7	-1.9366	3.19E-02
23	GO:0005244	voltage-gated ion channel activity	142	-1.5978	3.58E-02
24	GO:0022832	voltage-gated channel activity	142	-1.5978	3.58E-02
25	GO:0019842	vitamin binding	123	1.6360	3.92E-02
26	GO:0016176	superoxide-generating NADPH oxidase activator activity	7	-1.9081	4.93E-02

Table 7.7 GO MF terms (by p_{adj}) between non-recurrent and recurrent ccRCC profiled by DRS

	ID	Description	setSize	NES	p_{adj}
1	GO:0042605	peptide antigen binding	36	-2.4855	9.19E-05
2	GO:0042287	MHC protein binding	41	-2.5160	2.11E-04
3	GO:0023023	MHC protein complex binding	34	-2.2228	2.74E-03
4	GO:0042277	peptide binding	305	-1.5508	4.13E-02
5	GO:0004867	serine-type endopeptidase inhibitor activity	82	1.8922	4.45E-02

Table 7.8 GO MF terms (by p_{adj}) between non-recurrent and recurrent ccRCC profiled by PCS

	ID	Description	setSize	NES	p_{adj}
1	GO:0042611	MHC protein complex	92	-2.5320	1.58E-09
2	GO:0071556	integral component of lumenal side of endoplasmic reticulum membrane	91	-2.3857	4.94E-08
3	GO:0098553	lumenal side of endoplasmic reticulum membrane	91	-2.3857	4.94E-08
4	GO:0030666	endocytic vesicle membrane	241	-2.0384	4.94E-08
5	GO:0098576	lumenal side of membrane	97	-2.3391	1.15E-07
6	GO:0042101	T cell receptor complex	52	-2.4954	4.02E-07
7	GO:0009897	external side of plasma membrane	342	-1.7966	1.64E-06
8	GO:0012507	ER to Golgi transport vesicle membrane	119	-2.1609	2.46E-06
9	GO:0030139	endocytic vesicle	368	-1.7322	2.49E-06
10	GO:0042613	MHC class II protein complex	58	-2.2426	3.90E-05
11	GO:0030134	COPII-coated ER to Golgi transport vesicle	149	-1.9006	5.86E-05
12	GO:0045335	phagocytic vesicle	163	-1.8338	1.03E-04
13	GO:0030670	phagocytic vesicle membrane	109	-1.9644	1.42E-04
14	GO:0098802	plasma membrane signaling receptor complex	185	-1.7583	2.98E-04
15	GO:0030176	integral component of endoplasmic reticulum membrane	234	-1.7334	3.44E-04
16	GO:0042612	MHC class I protein complex	36	-2.1741	4.42E-04
17	GO:0001772	immunological synapse	51	-2.0646	6.21E-04
18	GO:0030658	transport vesicle membrane	230	-1.6862	8.01E-04
19	GO:0031227	intrinsic component of endoplasmic reticulum membrane	242	-1.6797	8.31E-04
20	GO:0030662	coated vesicle membrane	223	-1.6733	8.31E-04
21	GO:0030669	clathrin-coated endocytic vesicle membrane	99	-1.8631	8.73E-04
22	GO:0045334	clathrin-coated endocytic vesicle	112	-1.7896	1.21E-03
23	GO:0031901	early endosome membrane	194	-1.6088	2.59E-03
24	GO:0031012	extracellular matrix	440	1.5308	2.59E-03
25	GO:0005811	lipid droplet	87	1.8104	2.62E-03
26	GO:0030312	external encapsulating structure	441	1.5306	2.62E-03
27	GO:0062023	collagen-containing extracellular matrix	343	1.5420	3.41E-03
28	GO:0045177	apical part of cell	358	1.5460	3.68E-03
29	GO:0032588	trans-Golgi network membrane	123	-1.6851	3.75E-03
30	GO:0005833	hemoglobin complex	11	-2.0238	4.65E-03
31	GO:0009925	basal plasma membrane	209	1.6205	5.05E-03
32	GO:0005759	mitochondrial matrix	452	1.4839	6.27E-03
33	GO:0045178	basal part of cell	224	1.5774	7.79E-03
34	GO:0016323	basolateral plasma membrane	188	1.5929	8.66E-03
35	GO:0016324	apical plasma membrane	301	1.5002	1.52E-02
36	GO:0005796	Golgi lumen	67	1.7710	1.58E-02
37	GO:0005743	mitochondrial inner membrane	459	1.4092	1.77E-02
38	GO:0030665	clathrin-coated vesicle membrane	134	-1.5676	1.87E-02
39	GO:0034703	cation channel complex	157	-1.4807	2.25E-02
40	GO:0098686	hippocampal mossy fiber to CA3 synapse	28	1.8078	2.52E-02
41	GO:0000786	nucleosome	98	-1.6543	2.52E-02
42	GO:0042824	MHC class I peptide loading complex	25	-1.8469	2.79E-02
43	GO:0042105	alpha-beta T cell receptor complex	8	-1.9218	2.83E-02
44	GO:0043235	receptor complex	357	-1.3779	2.93E-02
45	GO:0098862	cluster of actin-based cell projections	139	1.5779	3.37E-02
46	GO:0034702	ion channel complex	208	-1.4585	3.37E-02
47	GO:0030135	coated vesicle	325	-1.3661	3.72E-02
48	GO:0005769	early endosome	390	-1.3047	4.53E-02
49	GO:0031838	haptoglobin-hemoglobin complex	10	-1.8223	4.93E-02

Table 7.9 GO CC terms (by p_{adj}) between non-recurrent and recurrent ccRCC profiled by DRS

	ID	Description	setSize	NES	p_{adj}
1	GO:0042101	T cell receptor complex	122	-3.0701	6.51E-20
2	GO:0098802	plasma membrane signaling receptor complex	282	-2.4006	5.46E-13
3	GO:0043235	receptor complex	495	-1.9655	2.35E-10
4	GO:0042613	MHC class II protein complex	15	-2.1022	1.45E-02
5	GO:0042105	alpha-beta T cell receptor complex	9	-1.9819	1.94E-02
6	GO:0042611	MHC protein complex	22	-2.2719	3.01E-02

Table 7.10 GO CC terms (by p_{adj}) between non-recurrent and recurrent ccRCC profiled by PCS

	ID	Description	setSize	NES	p_{adj}
1	hsa01240	Biosynthesis of cofactors	136	1.7148	2.21E-02
2	hsa04613	Neutrophil extracellular trap formation	159	-1.6566	2.21E-02
3	hsa01230	Biosynthesis of amino acids	64	1.8270	3.70E-02
4	hsa05340	Primary immunodeficiency	34	-1.9802	4.08E-02
5	hsa04975	Fat digestion and absorption	28	1.8364	4.08E-02
6	hsa00982	Drug metabolism - cytochrome P450	50	1.8116	4.08E-02
7	hsa05120	Epithelial cell signaling in Helicobacter pylori infection	65	1.7067	4.08E-02
8	hsa04080	Neuroactive ligand-receptor interaction	218	-1.5111	4.08E-02
9	hsa04520	Adherens junction	65	1.6868	4.97E-02

Table 7.11 GSEA of KEGG pathways between non-recurrent and recurrent ccRCC profiled by DRS

	ID	Description	setSize	NES	p_{adj}
1	hsa04612	Antigen processing and presentation	68	-2.6057	1.55E-08
2	hsa05332	Graft-versus-host disease	36	-2.4157	1.81E-05
3	hsa05330	Allograft rejection	35	-2.3835	1.98E-05
4	hsa05140	Leishmaniasis	73	-2.2503	4.65E-05
5	hsa04940	Type I diabetes mellitus	39	-2.1922	1.31E-03
6	hsa04660	T cell receptor signaling pathway	101	-2.0282	1.94E-03
7	hsa04650	Natural killer cell mediated cytotoxicity	109	-2.0457	2.19E-03
8	hsa04975	Fat digestion and absorption	38	1.9722	2.19E-03
9	hsa05320	Autoimmune thyroid disease	38	-2.1722	2.36E-03
10	hsa04640	Hematopoietic cell lineage	95	-2.0003	2.37E-03
11	hsa04658	Th1 and Th2 cell differentiation	87	-1.9877	2.79E-03
12	hsa05416	Viral myocarditis	57	-2.0729	3.90E-03
13	hsa00350	Tyrosine metabolism	34	1.9273	4.11E-03
14	hsa05235	PD-L1 expression and PD-1 checkpoint pathway in cancer	86	-1.9440	6.21E-03
15	hsa04659	Th17 cell differentiation	104	-1.8430	1.13E-02
16	hsa05340	Primary immunodeficiency	38	-2.0275	1.19E-02
17	hsa05168	Herpes simplex virus 1 infection	484	-1.4016	1.19E-02
18	hsa05321	Inflammatory bowel disease	62	-1.9985	1.31E-02
19	hsa05144	Malaria	48	-1.9444	1.31E-02
20	hsa05164	Influenza A	153	-1.6620	1.31E-02
21	hsa00515	Mannose type O-glycan biosynthesis	23	-2.0300	1.52E-02
22	hsa04613	Neutrophil extracellular trap formation	175	-1.6171	1.56E-02
23	hsa05310	Asthma	27	-1.9512	2.73E-02
24	hsa05152	Tuberculosis	163	-1.5826	2.73E-02
25	hsa04514	Cell adhesion molecules	150	-1.5978	3.54E-02
26	hsa00590	Arachidonic acid metabolism	58	1.6825	4.43E-02

Table 7.12 GSEA of KEGG pathways between non-recurrent and recurrent ccRCC profiled by PCS

	Ensembl ID	Gene Symbol	Description	DRIMseq p_{adj}	DEXseq p_{adj}
1	ENSG00000176731	RBIS	Ribosomal biogenesis factor	3.97E-02	1

Table 7.13 DTU genes between recurrent and non-recurrent ccRCC tumours by DRIMseq and DEXseq profiled by DRS

	Ensembl ID	Gene Symbol	Description	DRIMseq p_{adj}	DEXseq p_{adj}
1	ENSG00000187118	CMC1	C-X9-C motif containing 1	4.06E-14	7.22E-05
2	ENSG00000197530	MIB2	MIB E3 ubiquitin protein ligase 2	2.90E-05	0.002468232
3	ENSG00000160408	ST6GALNAC6	ST6 N-acetylgalactosaminide alpha-2,6-sialyltransferase 6	0.032775212	0.064348518
4	ENSG00000183690	EFHC2	EF-hand domain containing 2	0.023421849	1
5	ENSG00000136754	ABI1	ABL interactor 1	0.023421849	1
6	ENSG00000118257	NRP2	Neuropilin 2	0.023421849	1

Table 7.14 DTU genes between recurrent and non-recurrent ccRCC tumours by DRIMseq and DEXseq profiled by PCS

	Ensembl ID	Gene Symbol	Biotype	Description	Log_2FoldChange	p_{adj}
1	ENSG00000140105	WARS1	protein_coding	tryptophanyl-tRNA synthetase 1	4.9355	9.46E-114
2	ENSG00000112096	SOD2	protein_coding	superoxide dismutase 2	4.5742	8.25E-99
3	ENSG00000108691	CCL2	protein_coding	C-C motif chemokine ligand 2	9.6200	4.52E-90
4	ENSG00000090339	ICAM1	protein_coding	intercellular adhesion molecule 1	5.3560	3.47E-85
5	ENSG00000125730	C3	protein_coding	complement C3	5.1080	8.12E-79
6	ENSG00000156587	UBE2L6	protein_coding	ubiquitin conjugating enzyme E2 L6	4.1409	2.20E-73
7	ENSG00000125347	IRF1	protein_coding	interferon regulatory factor 1	5.1981	1.75E-72
8	ENSG00000168394	TAP1	protein_coding	transporter 1, ATP binding cassette subfamily B member	4.4816	1.90E-72
9	ENSG00000240065	PSMB9	protein_coding	proteasome 20S subunit beta 9	4.7561	9.94E-69
10	ENSG00000137496	IL18BP	protein_coding	interleukin 18 binding protein	7.2752	4.31E-62
11	ENSG00000234745	HLA-B	protein_coding	major histocompatibility complex, class I, B	3.4388	2.03E-58
12	ENSG00000117228	GBP1	protein_coding	guanylate binding protein 1	12.5927	2.56E-57
13	ENSG00000111331	OAS3	protein_coding	2'-5'-oligoadenylate synthetase 3	3.9894	3.27E-50
14	ENSG00000068079	IFI35	protein_coding	interferon induced protein 35	4.5009	1.24E-47
15	ENSG00000006210	CX3CL1	protein_coding	C-X3-C motif chemokine ligand 1	4.1378	2.12E-46
16	ENSG00000184371	CSF1	protein_coding	colony stimulating factor 1	3.4870	1.24E-45
17	ENSG00000184557	SOCS3	protein_coding	suppressor of cytokine signaling 3	4.1865	2.53E-45
18	ENSG00000119917	IFIT3	protein_coding	interferon induced protein with tetratricopeptide repeats 3	6.2392	1.34E-44
19	ENSG00000002549	LAP3	protein_coding	leucine aminopeptidase 3	3.8431	2.55E-43
20	ENSG00000089127	OAS1	protein_coding	2'-5'-oligoadenylate synthetase 1	7.6480	6.50E-43
21	ENSG00000185338	SOCS1	protein_coding	suppressor of cytokine signaling 1	5.4393	7.40E-42
22	ENSG00000172183	ISG20	protein_coding	interferon stimulated exonuclease gene 20	7.5528	1.10E-37
23	ENSG00000133321	PLAAT4	protein_coding	phospholipase A and acyltransferase 4	6.5817	7.62E-37
24	ENSG00000100342	APOL1	protein_coding	apolipoprotein L1	4.0876	2.81E-35
25	ENSG00000164442	CITED2	protein_coding	Cbp/p300 interacting transactivator with Glu/Asp rich carboxy-terminal domain 2	-2.9314	3.56E-35
26	ENSG00000196954	CASP4	protein_coding	caspase 4	3.7612	5.45E-35
27	ENSG00000221963	APOL6	protein_coding	apolipoprotein L6	4.5524	3.15E-34
28	ENSG00000130176	CNN1	protein_coding	calponin 1	-3.1941	3.47E-32
29	ENSG00000138496	PARP9	protein_coding	poly(ADP-ribose) polymerase family member 9	4.1654	5.01E-31
30	ENSG00000188313	PLSCR1	protein_coding	phospholipid scramblase 1	3.3081	5.88E-31
31	ENSG00000113263	ITK	protein_coding	IL2 inducible T cell kinase	6.8133	6.71E-31
32	ENSG00000198959	TGM2	protein_coding	transglutaminase 2	3.1985	1.81E-29
33	ENSG00000196776	CD47	protein_coding	CD47 molecule	2.9358	1.41E-28
34	ENSG00000204642	HLA-F	protein_coding	major histocompatibility complex, class I, F	4.2530	1.41E-28
35	ENSG00000205336	ADGRG1	protein_coding	adhesion G protein-coupled receptor G1	-3.9306	2.31E-28

Table 7.15 Top 100 DEGs (by p_{adj}) between untreated and IFNγ + TNF treated RCC4 Cas9 GFP by reference genome aligned DRS

	Ensembl ID	Gene Symbol	Biotype	Description	Log_2FoldChange	p_{adj}
36	ENSG00000019582	CD74	protein_coding	CD74 molecule	2.2918	2.84E-28
37	ENSG00000103196	CRISPLD2	protein_coding	cysteine rich secretory protein LCCL domain containing 2	-3.0673	1.02E-27
38	ENSG00000130813	SHFL	protein_coding	shiftless antiviral inhibitor of ribosomal frameshifting	3.8414	1.02E-27
39	ENSG00000135148	TRAFD1	protein_coding	TRAF-type zinc finger domain containing 1	2.8213	2.90E-27
40	ENSG00000204264	PSMB8	protein_coding	proteasome 20S subunit beta 8	3.0719	3.02E-27
41	ENSG00000197142	ACSL5	protein_coding	acyl-CoA synthetase long chain family member 5	3.1399	3.94E-27
42	ENSG00000142089	IFITM3	protein_coding	interferon induced transmembrane protein 3	2.7929	5.83E-27
43	ENSG00000141682	PMAIP1	protein_coding	phorbol-12-myristate-13-acetate-induced protein 1	3.5367	7.21E-27
44	ENSG00000140464	PML	protein_coding	PML nuclear body scaffold	2.4005	2.80E-26
45	ENSG00000131203	IDO1	protein_coding	indoleamine 2,3-dioxygenase 1	13.2973	8.51E-26
46	ENSG00000117226	GBP3	protein_coding	guanylate binding protein 3	5.0641	2.40E-25
47	ENSG00000130589	HELZ2	protein_coding	helicase with zinc finger 2	3.7990	5.51E-25
48	ENSG00000204267	TAP2	protein_coding	transporter 2, ATP binding cassette subfamily B member	3.2237	6.22E-24
49	ENSG00000169245	CXCL10	protein_coding	C-X-C motif chemokine ligand 10	13.0779	1.74E-23
50	ENSG00000173821	RNF213	protein_coding	ring finger protein 213	2.7511	1.83E-23
51	ENSG00000204592	HLA-E	protein_coding	major histocompatibility complex, class I, E	2.7068	2.39E-23
52	ENSG00000136883	KIF12	protein_coding	kinesin family member 12	3.6726	4.79E-23
53	ENSG00000116514	RNF19B	protein_coding	ring finger protein 19B	2.6088	9.07E-23
54	ENSG00000197646	PDCD1LG2	protein_coding	programmed cell death 1 ligand 2	3.6637	2.64E-22
55	ENSG00000185507	IRF7	protein_coding	interferon regulatory factor 7	3.0573	5.38E-22
56	ENSG00000139112	GABARAPL1	protein_coding	GABA type A receptor associated protein like 1	2.8478	1.14E-21
57	ENSG00000134470	IL15RA	protein_coding	interleukin 15 receptor subunit alpha	3.1341	1.80E-21
58	ENSG00000128335	APOL2	protein_coding	apolipoprotein L2	3.3095	2.20E-21
59	ENSG00000163131	CTSS	protein_coding	cathepsin S	10.5452	2.79E-21
60	ENSG00000132109	TRIM21	protein_coding	tripartite motif containing 21	2.3720	3.18E-21
61	ENSG00000118503	TNFAIP3	protein_coding	TNF alpha induced protein 3	3.0443	3.74E-21
62	ENSG00000100336	APOL4	protein_coding	apolipoprotein L4	11.9835	4.69E-21
63	ENSG00000059378	PARP12	protein_coding	poly(ADP-ribose) polymerase family member 12	2.6921	4.79E-21
64	ENSG00000103257	SLC7A5	protein_coding	solute carrier family 7 member 5	-2.3883	1.37E-20
65	ENSG00000040608	RTN4R	protein_coding	reticulon 4 receptor	-2.6034	2.00E-20
66	ENSG00000056558	TRAF1	protein_coding	TNF receptor associated factor 1	3.5549	5.59E-20
67	ENSG00000135899	SP110	protein_coding	SP110 nuclear body protein	2.8155	1.14E-19
68	ENSG00000162654	GBP4	protein_coding	guanylate binding protein 4	11.5964	1.77E-19
69	ENSG00000163840	DTX3L	protein_coding	deltex E3 ubiquitin ligase 3L	3.1508	3.80E-19
70	ENSG00000185745	IFIT1	protein_coding	interferon induced protein with tetratricopeptide repeats 1	4.4510	7.43E-19

Table 7.15 (cont.) Top 100 DEGs (by p_{adj}) between untreated and IFNγ + TNF treated RCC4 Cas9 GFP by reference genome aligned DRS

	Ensembl ID	Gene Symbol	Biotype	Description	Log_2FoldChange	p_{adj}
71	ENSG00000173193	PARP14	protein_coding	poly(ADP-ribose) polymerase family member 14	4.9333	2.04E-18
72	ENSG00000101347	SAMHD1	protein_coding	SAM and HD domain containing deoxynucleoside triphosphate triphosphohydrolase 1	3.3214	2.15E-18
73	ENSG00000146674	IGFBP3	protein_coding	insulin like growth factor binding protein 3	-2.3081	2.37E-18
74	ENSG00000173432	SAA1	protein_coding	serum amyloid A1	5.5625	3.51E-18
75	ENSG00000140853	NLRC5	protein_coding	NLR family CARD domain containing 5	2.7971	3.62E-18
76	ENSG00000123609	NMI	protein_coding	N-myc and STAT interactor	2.9745	7.73E-18
77	ENSG00000187608	ISG15	protein_coding	ISG15 ubiquitin like modifier	4.0058	8.29E-18
78	ENSG00000163735	CXCL5	protein_coding	C-X-C motif chemokine ligand 5	2.3437	8.90E-18
79	ENSG00000226025		transcribed_unprocessed_pseudogene	Charcot-Leyden crystal protein pseudogene	11.0172	1.03E-17
80	ENSG00000116337	AMPD2	protein_coding	adenosine monophosphate deaminase 2	-2.1931	1.07E-17
81	ENSG00000100065	CARD10	protein_coding	caspase recruitment domain family member 10	-2.1201	1.14E-17
82	ENSG00000170581	STAT2	protein_coding	signal transducer and activator of transcription 2	2.6599	1.62E-17
83	ENSG00000169252	ADRB2	protein_coding	adrenoceptor beta 2	-3.0931	1.71E-17
84	ENSG00000159403	C1R	protein_coding	complement C1r	7.2260	1.85E-17
85	ENSG00000168404	MLKL	protein_coding	mixed lineage kinase domain like pseudokinase	2.7130	1.86E-17
86	ENSG00000081041	CXCL2	protein_coding	C-X-C motif chemokine ligand 2	5.2950	3.12E-17
87	ENSG00000182326	C1S	protein_coding	complement C1s	10.8689	3.22E-17
88	ENSG00000158714	SLAMF8	protein_coding	SLAM family member 8	10.8472	4.28E-17
89	ENSG00000166710	B2M	protein_coding	beta-2-microglobulin	3.2177	9.48E-17
90	ENSG00000126878	AIF1L	protein_coding	allograft inflammatory factor 1 like	-2.1694	1.59E-16
91	ENSG00000023445	BIRC3	protein_coding	baculoviral IAP repeat containing 3	4.1381	1.69E-16
92	ENSG00000154451	GBP5	protein_coding	guanylate binding protein 5	10.6363	1.83E-16
93	ENSG00000163565	IFI16	protein_coding	interferon gamma inducible protein 16	4.3515	4.28E-16
94	ENSG00000149131	SERPING1	protein_coding	serpin family G member 1	9.1194	6.42E-16
95	ENSG00000170989	S1PR1	protein_coding	sphingosine-1-phosphate receptor 1	-4.2881	6.60E-16
96	ENSG00000091409	ITGA6	protein_coding	integrin subunit alpha 6	3.8946	6.60E-16
97	ENSG00000115267	IFIH1	protein_coding	interferon induced with helicase C domain 1	4.6565	7.66E-16
98	ENSG00000162645	GBP2	protein_coding	guanylate binding protein 2	9.2395	9.19E-16
99	ENSG00000139289	PHLDA1	protein_coding	pleckstrin homology like domain family A member 1	2.1874	1.07E-15
100	ENSG00000256128	LINC00944	lncRNA	long intergenic non-protein coding RNA 944	2.9553	1.19E-15

Table 7.15 (cont.) Top 100 DEGs (by p_{adj}) between untreated and IFNγ + TNF treated RCC4 Cas9 GFP by reference genome aligned DRS

	Ensembl ID	Gene Symbol	Biotype	Description	Log_2FoldChange	p_{adj}
1	ENSG00000108691	CCL2	protein_coding	C-C motif chemokine ligand 2	9.5526	3.59E-83
2	ENSG00000090339	ICAM1	protein_coding	intercellular adhesion molecule 1	5.3366	1.28E-69
3	ENSG00000140105	WARS1	protein_coding	tryptophanyl-tRNA synthetase 1	4.6384	1.28E-69
4	ENSG00000291237	SOD2	protein_coding	superoxide dismutase 2	4.7208	1.57E-69
5	ENSG00000125347	IRF1	protein_coding	interferon regulatory factor 1	5.3658	2.46E-69
6	ENSG00000125730	C3	protein_coding	complement C3	5.0671	1.82E-64
7	ENSG00000156587	UBE2L6	protein_coding	ubiquitin conjugating enzyme E2 L6	4.0482	3.57E-50
8	ENSG00000228964	HLA-B	protein_coding	major histocompatibility complex, class I, B	3.7614	3.91E-45
9	ENSG00000240065	PSMB9	protein_coding	proteasome 20S subunit beta 9	4.6691	7.40E-39
10	ENSG00000184557	SOCS3	protein_coding	suppressor of cytokine signaling 3	4.1862	2.75E-36
11	ENSG00000234745	HLA-B	protein_coding	major histocompatibility complex, class I, B	3.0775	9.35E-35
12	ENSG00000184371	CSF1	protein_coding	colony stimulating factor 1	3.4814	1.77E-34
13	ENSG00000185338	SOCS1	protein_coding	suppressor of cytokine signaling 1	5.3499	1.87E-34
14	ENSG00000126709	IFI6	protein_coding	interferon alpha inducible protein 6	4.6347	1.03E-33
15	ENSG00000205220	PSMB10	protein_coding	proteasome 20S subunit beta 10	3.4862	8.71E-33
16	ENSG00000164442	CITED2	protein_coding	Cbp/p300 interacting transactivator with Glu/Asp rich carboxy-terminal domain 2	-2.9385	1.48E-30
17	ENSG00000133321	PLAAT4	protein_coding	phospholipase A and acyltransferase 4	6.4103	3.73E-29
18	ENSG00000068079	IFI35	protein_coding	interferon induced protein 35	4.0680	5.80E-29
19	ENSG00000237988	OR2I1P	protein_coding	olfactory receptor family 2 subfamily I member 1 pseudogene	13.7220	2.93E-27
20	ENSG00000196954	CASP4	protein_coding	caspase 4	3.7227	1.12E-26
21	ENSG00000198959	TGM2	protein_coding	transglutaminase 2	3.1545	5.81E-26
22	ENSG00000130589	HELZ2	protein_coding	helicase with zinc finger 2	3.8064	7.96E-26
23	ENSG00000126246	IGFLR1	protein_coding	IGF like family receptor 1	3.3509	1.08E-24
24	ENSG00000206468	UBD	protein_coding	ubiquitin D	13.3142	1.23E-24
25	ENSG00000141682	PMAIP1	protein_coding	phorbol-12-myristate-13-acetate-induced protein 1	3.5309	1.80E-24
26	ENSG00000196776	CD47	protein_coding	CD47 molecule	2.9252	2.00E-23
27	ENSG00000117228	GBP1	protein_coding	guanylate binding protein 1	12.5278	1.48E-22
28	ENSG00000221963	APOL6	protein_coding	apolipoprotein L6	3.3590	1.51E-22
29	ENSG00000130176	CNN1	protein_coding	calponin 1	-3.1944	5.73E-22
30	ENSG00000103196	CRISPLD2	protein_coding	cysteine rich secretory protein LCCL domain containing 2	-3.1483	6.73E-22
31	ENSG00000185885	IFITM1	protein_coding	interferon induced transmembrane protein 1	7.8895	5.29E-21
32	ENSG00000187608	ISG15	protein_coding	ISG15 ubiquitin like modifier	3.9035	6.54E-21
33	ENSG00000131203	IDO1	protein_coding	indoleamine 2,3-dioxygenase 1	11.8615	3.40E-20
34	ENSG00000135148	TRAFD1	protein_coding	TRAF-type zinc finger domain containing 1	2.8934	3.40E-20
35	ENSG00000188313	PLSCR1	protein_coding	phospholipid scramblase 1	3.2514	3.63E-20

Table 7.16 Top 100 DEGs (by p_{adj}) between untreated and IFNγ + TNF treated RCC4 Cas9 GFP by reference transcriptome aligned DRS

	Ensembl ID	Gene Symbol	Biotype	Description	Log_2FoldChange	p_{adj}
36	ENSG00000113263	ITK	protein_coding	IL2 inducible T cell kinase	6.3396	3.50E-19
37	ENSG00000197646	PDCD1LG2	protein_coding	programmed cell death 1 ligand 2	3.6951	7.07E-19
38	ENSG00000169245	CXCL10	protein_coding	C-X-C motif chemokine ligand 10	13.0577	2.10E-18
39	ENSG00000089127	OAS1	protein_coding	2'-5'-oligoadenylate synthetase 1	7.2958	2.13E-18
40	ENSG00000142089	IFITM3	protein_coding	interferon induced transmembrane protein 3	2.6005	4.06E-18
41	ENSG00000197142	ACSL5	protein_coding	acyl-CoA synthetase long chain family member 5	3.1455	6.07E-18
42	ENSG00000103257	SLC7A5	protein_coding	solute carrier family 7 member 5	-2.3681	8.11E-18
43	ENSG00000146859	TMEM140	protein_coding	transmembrane protein 140	3.0521	1.27E-17
44	ENSG00000137496	IL18BP	protein_coding	interleukin 18 binding protein	5.3866	2.18E-17
45	ENSG00000006210	CX3CL1	protein_coding	C-X3-C motif chemokine ligand 1	3.8640	2.35E-17
46	ENSG00000166710	B2M	protein_coding	beta-2-microglobulin	3.2167	4.73E-17
47	ENSG00000123609	NMI	protein_coding	N-myc and STAT interactor	2.9231	1.53E-16
48	ENSG00000130813	SHFL	protein_coding	shiftless antiviral inhibitor of ribosomal frameshifting	4.1300	3.00E-16
49	ENSG00000019582	CD74	protein_coding	CD74 molecule	2.3961	4.16E-16
50	ENSG00000081041	CXCL2	protein_coding	C-X-C motif chemokine ligand 2	5.2961	4.75E-16
51	ENSG00000163840	DTX3L	protein_coding	deltex E3 ubiquitin ligase 3L	3.1550	4.75E-16
52	ENSG00000182326	C1S	protein_coding	complement C1s	10.6121	5.18E-16
53	ENSG00000162654	GBP4	protein_coding	guanylate binding protein 4	10.5683	8.89E-16
54	ENSG00000136883	KIF12	protein_coding	kinesin family member 12	3.9827	9.91E-16
55	ENSG00000226025		transcribed_unprocessed_pseudogene	Charcot-Leyden crystal protein pseudogene	10.4833	1.10E-15
56	ENSG00000169252	ADRB2	protein_coding	adrenoceptor beta 2	-3.1320	1.29E-15
57	ENSG00000158714	SLAMF8	protein_coding	SLAM family member 8	10.4749	1.31E-15
58	ENSG00000184979	USP18	protein_coding	ubiquitin specific peptidase 18	5.4764	2.40E-15
59	ENSG00000059378	PARP12	protein_coding	poly(ADP-ribose) polymerase family member 12	2.7713	2.62E-15
60	ENSG00000168404	MLKL	protein_coding	mixed lineage kinase domain like pseudokinase	2.6749	5.26E-15
61	ENSG00000169248	CXCL11	protein_coding	C-X-C motif chemokine ligand 11	10.4657	6.21E-15
62	ENSG00000132109	TRIM21	protein_coding	tripartite motif containing 21	2.3476	6.33E-15
63	ENSG00000163131	CTSS	protein_coding	cathepsin S	10.2003	7.11E-15
64	ENSG00000170989	S1PR1	protein_coding	sphingosine-1-phosphate receptor 1	-4.3608	7.15E-15
65	ENSG00000040608	RTN4R	protein_coding	reticulon 4 receptor	-2.5476	1.68E-14
66	ENSG00000108700	CCL8	protein_coding	C-C motif chemokine ligand 8	10.2044	2.34E-14
67	ENSG00000119917	IFIT3	protein_coding	interferon induced protein with tetratricopeptide repeats 3	6.5352	2.34E-14
68	ENSG00000140464	PML	protein_coding	PML nuclear body scaffold	2.1612	4.26E-14
69	ENSG00000146674	IGFBP3	protein_coding	insulin like growth factor binding protein 3	-2.2958	5.92E-14
70	ENSG00000241935	HOGA1	protein_coding	4-hydroxy-2-oxoglutarate aldolase 1	-2.7646	9.26E-14

Table 7.16 (cont.)Top 100 DEGs (by p_{adj}) between untreated and IFNγ+TNF treated RCC4 Cas9 GFP by reference transcriptome aligned DRS

	Ensembl ID	Gene Symbol	Biotype	Description	Log_2FoldChange	p_{adj}
71	ENSG00000163735	CXCL5	protein_coding	C-X-C motif chemokine ligand 5	2.3469	9.38E-14
72	ENSG00000134470	IL15RA	protein_coding	interleukin 15 receptor subunit alpha	3.0705	1.03E-13
73	ENSG00000168062	BATF2	protein_coding	basic leucine zipper ATF-like transcription factor 2	9.7654	1.04E-13
74	ENSG00000116514	RNF19B	protein_coding	ring finger protein 19B	2.6129	1.46E-13
75	ENSG00000168899	VAMP5	protein_coding	vesicle associated membrane protein 5	2.1629	2.81E-13
76	ENSG00000115415	STAT1	protein_coding	signal transducer and activator of transcription 1	3.1337	3.57E-13
77	ENSG00000130303	BST2	protein_coding	bone marrow stromal cell antigen 2	2.1787	3.57E-13
78	ENSG00000132274	TRIM22	protein_coding	tripartite motif containing 22	8.3786	7.59E-13
79	ENSG00000154451	GBP5	protein_coding	guanylate binding protein 5	9.4070	1.14E-12
80	ENSG00000104951	IL4I1	protein_coding	interleukin 4 induced 1	3.2651	2.89E-12
81	ENSG00000139112	GABARAPL1	protein_coding	GABA type A receptor associated protein like 1	2.8129	2.90E-12
82	ENSG00000114315	HES1	protein_coding	hes family bHLH transcription factor 1	-3.5846	3.24E-12
83	ENSG00000273686	B2M	protein_coding	beta-2-microglobulin	3.9808	3.41E-12
84	ENSG00000138496	PARP9	protein_coding	poly(ADP-ribose) polymerase family member 9	3.2781	5.40E-12
85	ENSG00000008517	IL32	protein_coding	interleukin 32	2.3654	5.47E-12
86	ENSG00000039523	RIPOR1	protein_coding	RHO family interacting cell polarization regulator 1	-2.3228	6.89E-12
87	ENSG00000101384	JAG1	protein_coding	jagged canonical Notch ligand 1	-2.6099	9.82E-12
88	ENSG00000135899	SP110	protein_coding	SP110 nuclear body protein	3.2846	1.12E-11
89	ENSG00000117226	GBP3	protein_coding	guanylate binding protein 3	5.7010	1.33E-11
90	ENSG00000149131	SERPING1	protein_coding	serpin family G member 1	9.0687	1.70E-11
91	ENSG00000171223	JUNB	protein_coding	JunB proto-oncogene, AP-1 transcription factor subunit	2.1007	1.94E-11
92	ENSG00000112149	CD83	protein_coding	CD83 molecule	2.4764	3.69E-11
93	ENSG00000188820	CALHM6	protein_coding	calcium homeostasis modulator family member 6	2.0264	4.32E-11
94	ENSG00000111331	OAS3	protein_coding	2'-5'-oligoadenylate synthetase 3	2.9470	8.73E-11
95	ENSG00000138755	CXCL9	protein_coding	C-X-C motif chemokine ligand 9	9.0710	1.12E-10
96	ENSG00000168685	IL7R	protein_coding	interleukin 7 receptor	3.2159	1.13E-10
97	ENSG00000111335	OAS2	protein_coding	2'-5'-oligoadenylate synthetase 2	8.7788	1.13E-10
98	ENSG00000183018	SPNS2	protein_coding	sphingolipid transporter 2	-2.1691	1.60E-10
99	ENSG00000050344	NFE2L3	protein_coding	NFE2 like bZIP transcription factor 3	3.0303	1.67E-10
100	ENSG00000152518	ZFP36L2	protein_coding	ZFP36 ring finger protein like 2	2.5334	2.10E-10

Table 7.16 (cont.)Top 100 DEGs (by p_{adj}) between untreated and IFNγ+TNF treated RCC4 Cas9 GFP by reference transcriptome aligned DRS

	Ensembl ID	Gene Symbol	Biotype	Description	Log_2FoldChange	p_{adj}
1	ENSG00000117228	GBP1	protein_coding	guanylate binding protein 1	10.3937	7.66E-131
2	ENSG00000090339	ICAM1	protein_coding	intercellular adhesion molecule 1	5.5519	3.13E-115
3	ENSG00000108691	CCL2	protein_coding	C-C motif chemokine ligand 2	9.6430	6.76E-111
4	ENSG00000140105	WARS1	protein_coding	tryptophanyl-tRNA synthetase 1	5.0225	2.80E-104
5	ENSG00000125347	IRF1	protein_coding	interferon regulatory factor 1	5.2562	3.24E-102
6	ENSG00000240065	PSMB9	protein_coding	proteasome 20S subunit beta 9	5.0124	1.93E-89
7	ENSG00000112096	SOD2	protein_coding	superoxide dismutase 2	4.3175	8.62E-88
8	ENSG00000168394	TAP1	protein_coding	transporter 1, ATP binding cassette subfamily B member	4.4960	2.13E-75
9	ENSG00000182326	C1S	protein_coding	complement C1s	7.1305	2.41E-66
10	ENSG00000234745	HLA-B	protein_coding	major histocompatibility complex, class I, B	3.7402	1.84E-64
11	ENSG00000089127	OAS1	protein_coding	2'-5'-oligoadenylate synthetase 1	5.0887	2.70E-64
12	ENSG00000119917	IFIT3	protein_coding	interferon induced protein with tetratricopeptide repeats 3	6.2490	5.08E-60
13	ENSG00000006210	CX3CL1	protein_coding	C-X3-C motif chemokine ligand 1	4.7430	1.41E-57
14	ENSG00000125730	C3	protein_coding	complement C3	4.0302	2.00E-53
15	ENSG00000184371	CSF1	protein_coding	colony stimulating factor 1	3.4205	1.17E-50
16	ENSG00000068079	IFI35	protein_coding	interferon induced protein 35	3.9297	9.37E-47
17	ENSG00000156587	UBE2L6	protein_coding	ubiquitin conjugating enzyme E2 L6	3.6274	1.13E-46
18	ENSG00000133321	PLAAT4	protein_coding	phospholipase A and acyltransferase 4	5.8804	4.10E-46
19	ENSG00000002549	LAP3	protein_coding	leucine aminopeptidase 3	3.2801	3.00E-44
20	ENSG00000185338	SOCS1	protein_coding	suppressor of cytokine signaling 1	4.8565	5.92E-41
21	ENSG00000137496	IL18BP	protein_coding	interleukin 18 binding protein	8.8305	1.36E-40
22	ENSG00000159403	C1R	protein_coding	complement C1r	5.6130	2.31E-40
23	ENSG00000118503	TNFAIP3	protein_coding	TNF alpha induced protein 3	4.0870	2.77E-39
24	ENSG00000111331	OAS3	protein_coding	2'-5'-oligoadenylate synthetase 3	3.4977	8.21E-39
25	ENSG00000184557	SOCS3	protein_coding	suppressor of cytokine signaling 3	3.8583	8.45E-38
26	ENSG00000056558	TRAF1	protein_coding	TNF receptor associated factor 1	4.1042	2.45E-37
27	ENSG00000205336	ADGRG1	protein_coding	adhesion G protein-coupled receptor G1	-3.4566	8.98E-36
28	ENSG00000188313	PLSCR1	protein_coding	phospholipid scramblase 1	3.0225	1.01E-33
29	ENSG00000221963	APOL6	protein_coding	apolipoprotein L6	4.5032	1.01E-33
30	ENSG00000130589	HELZ2	protein_coding	helicase with zinc finger 2	3.5422	2.46E-33
31	ENSG00000138496	PARP9	protein_coding	poly(ADP-ribose) polymerase family member 9	3.8480	3.58E-32
32	ENSG00000169245	CXCL10	protein_coding	C-X-C motif chemokine ligand 10	14.3000	6.18E-30
33	ENSG00000135148	TRAFD1	protein_coding	TRAF-type zinc finger domain containing 1	2.9590	2.54E-29
34	ENSG00000131203	IDO1	protein_coding	indoleamine 2,3-dioxygenase 1	13.9260	1.45E-28
35	ENSG00000164442	CITED2	protein_coding	Cbp/p300 interacting transactivator with Glu/Asp rich carboxy-terminal domain 2	-2.5011	1.93E-27

Table 7.17 Top 100 DEGs (by p_{adj}) between untreated and IFNγ + TNF treated WTAP KO 2H1 by reference genome aligned DRS

	Ensembl ID	Gene Symbol	Biotype	Description	Log_2FoldChange	p_{adj}
36	ENSG00000113263	ITK	protein_coding	IL2 inducible T cell kinase	5.7545	3.18E-27
37	ENSG00000173432	SAA1	protein_coding	serum amyloid A1	5.3321	3.18E-27
38	ENSG00000168062	BATF2	protein_coding	basic leucine zipper ATF-like transcription factor 2	8.6265	1.03E-26
39	ENSG00000149131	SERPING1	protein_coding	serpin family G member 1	7.6699	1.86E-25
40	ENSG00000196954	CASP4	protein_coding	caspase 4	2.7088	1.50E-24
41	ENSG00000172183	ISG20	protein_coding	interferon stimulated exonuclease gene 20	8.2853	2.01E-24
42	ENSG00000142089	IFITM3	protein_coding	interferon induced transmembrane protein 3	2.7175	2.64E-24
43	ENSG00000204642	HLA-F	protein_coding	major histocompatibility complex, class I, F	4.2269	5.96E-24
44	ENSG00000196776	CD47	protein_coding	CD47 molecule	2.5198	8.31E-24
45	ENSG00000162772	ATF3	protein_coding	activating transcription factor 3	2.8962	1.10E-23
46	ENSG00000204592	HLA-E	protein_coding	major histocompatibility complex, class I, E	2.6372	1.63E-23
47	ENSG00000163840	DTX3L	protein_coding	deltex E3 ubiquitin ligase 3L	3.1435	1.43E-22
48	ENSG00000008517	IL32	protein_coding	interleukin 32	2.6166	1.82E-22
49	ENSG00000132109	TRIM21	protein_coding	tripartite motif containing 21	2.6789	1.94E-22
50	ENSG00000204264	PSMB8	protein_coding	proteasome 20S subunit beta 8	2.8500	3.52E-22
51	ENSG00000112149	CD83	protein_coding	CD83 molecule	3.0314	4.10E-22
52	ENSG00000140853	NLRC5	protein_coding	NLR family CARD domain containing 5	3.0574	4.70E-22
53	ENSG00000101347	SAMHD1	protein_coding	SAM and HD domain containing deoxynucleoside triphosphate triphosphohydrolase 1	3.1483	7.35E-22
54	ENSG00000130813	SHFL	protein_coding	shiftless antiviral inhibitor of ribosomal frameshifting	3.0528	1.57E-21
55	ENSG00000173821	RNF213	protein_coding	ring finger protein 213	2.5898	2.05E-21
56	ENSG00000100336	APOL4	protein_coding	apolipoprotein L4	11.9727	7.46E-21
57	ENSG00000128335	APOL2	protein_coding	apolipoprotein L2	3.3841	1.12E-20
58	ENSG00000136883	KIF12	protein_coding	kinesin family member 12	3.4638	3.42E-20
59	ENSG00000146859	TMEM140	protein_coding	transmembrane protein 140	5.6078	6.39E-20
60	ENSG00000170581	STAT2	protein_coding	signal transducer and activator of transcription 2	2.6308	6.80E-20
61	ENSG00000162654	GBP4	protein_coding	guanylate binding protein 4	11.5638	1.69E-19
62	ENSG00000166710	B2M	protein_coding	beta-2-microglobulin	2.5760	2.42E-19
63	ENSG00000204525	HLA-C	protein_coding	major histocompatibility complex, class I, C	2.0493	2.77E-19
64	ENSG00000134326	CMPK2	protein_coding	cytidine/uridine monophosphate kinase 2	6.8006	3.29E-19
65	ENSG00000117226	GBP3	protein_coding	guanylate binding protein 3	4.2850	3.85E-19
66	ENSG00000187608	ISG15	protein_coding	ISG15 ubiquitin like modifier	3.5452	1.04E-18
67	ENSG00000171223	JUNB	protein_coding	JunB proto-oncogene, AP-1 transcription factor subunit	2.3362	1.13E-18
68	ENSG00000139192	TAPBPL	protein_coding	TAP binding protein like	2.8530	1.19E-18
69	ENSG00000163131	CTSS	protein_coding	cathepsin S	11.1939	2.41E-18
70	ENSG00000141682	PMAIP1	protein_coding	phorbol-12-myristate-13-acetate-induced protein 1	2.7738	3.39E-18

Table 7.17 (cont.) Top 100 DEGs (by p_{adj}) between untreated and IFNγ+TNF treated WTAP KO 2H1 by reference genome aligned DRS

	Ensembl ID	Gene Symbol	Biotype	Description	Log_2FoldChange	p_{adj}
71	ENSG00000204267	TAP2	protein_coding	transporter 2, ATP binding cassette subfamily B member	2.8382	3.58E-18
72	ENSG00000139112	GABARAPL1	protein_coding	GABA type A receptor associated protein like 1	2.3640	4.20E-18
73	ENSG00000158714	SLAMF8	protein_coding	SLAM family member 8	11.1708	4.35E-18
74	ENSG00000226025	0	transcribed_unprocessed_pseudogene	Charcot-Leyden crystal protein pseudogene	11.1412	5.12E-18
75	ENSG00000173193	PARP14	protein_coding	poly(ADP-ribose) polymerase family member 14	3.8317	7.72E-18
76	ENSG00000059378	PARP12	protein_coding	poly(ADP-ribose) polymerase family member 12	2.6128	1.75E-17
77	ENSG00000197142	ACSL5	protein_coding	acyl-CoA synthetase long chain family member 5	2.4193	1.75E-17
78	ENSG00000140464	PML	protein_coding	PML nuclear body scaffold	2.3665	2.02E-17
79	ENSG00000185745	IFIT1	protein_coding	interferon induced protein with tetratricopeptide repeats 1	3.5250	3.61E-17
80	ENSG00000116514	RNF19B	protein_coding	ring finger protein 19B	2.3217	4.25E-17
81	ENSG00000121858	TNFSF10	protein_coding	TNF superfamily member 10	2.8496	4.74E-17
82	ENSG00000154451	GBP5	protein_coding	guanylate binding protein 5	10.8860	4.76E-17
83	ENSG00000188015	S100A3	protein_coding	S100 calcium binding protein A3	2.6567	6.46E-17
84	ENSG00000138755	CXCL9	protein_coding	C-X-C motif chemokine ligand 9	10.7879	6.58E-17
85	ENSG00000172216	CEBPB	protein_coding	CCAAT enhancer binding protein beta	2.2619	1.15E-16
86	ENSG00000128284	APOL3	protein_coding	apolipoprotein L3	7.8386	2.09E-16
87	ENSG00000185507	IRF7	protein_coding	interferon regulatory factor 7	2.6767	2.30E-16
88	ENSG00000168404	MLKL	protein_coding	mixed lineage kinase domain like pseudokinase	2.2498	2.54E-16
89	ENSG00000103196	CRISPLD2	protein_coding	cysteine rich secretory protein LCCL domain containing 2	-2.3792	3.74E-16
90	ENSG00000023445	BIRC3	protein_coding	baculoviral IAP repeat containing 3	3.6142	3.75E-16
91	ENSG00000134470	IL15RA	protein_coding	interleukin 15 receptor subunit alpha	3.3137	4.15E-16
92	ENSG00000108700	CCL8	protein_coding	C-C motif chemokine ligand 8	10.5583	4.93E-16
93	ENSG00000123609	NMI	protein_coding	N-myc and STAT interactor	2.5034	8.99E-16
94	ENSG00000225492	GBP1P1	transcribed_unprocessed_pseudogene	guanylate binding protein 1 pseudogene 1	10.4345	9.98E-16
95	ENSG00000184979	USP18	protein_coding	ubiquitin specific peptidase 18	5.1197	1.28E-15
96	ENSG00000115415	STAT1	protein_coding	signal transducer and activator of transcription 1	2.5164	1.42E-15
97	ENSG00000111335	OAS2	protein_coding	2'-5'-oligoadenylate synthetase 2	10.2747	2.30E-15
98	ENSG00000198959	TGM2	protein_coding	transglutaminase 2	2.4372	3.63E-15
99	ENSG00000246985	SOCS2-AS1	lncRNA	SOCS2 antisense RNA 1	-2.6091	3.81E-15
100	ENSG00000135899	SP110	protein_coding	SP110 nuclear body protein	2.4241	4.74E-15

Table 7.17 (cont.) Top 100 DEGs (by p_{adj}) between untreated and IFNγ+TNF treated WTAP KO 2H1 by reference genome aligned DRS

	Ensembl ID	Gene Symbol	Biotype	Description	Log_2FoldChange	p_{adj}
1	ENSG00000090339	ICAM1	protein_coding	intercellular adhesion molecule 1	5.5790	1.82E-103
2	ENSG00000108691	CCL2	protein_coding	C-C motif chemokine ligand 2	9.6923	1.82E-103
3	ENSG00000125347	IRF1	protein_coding	interferon regulatory factor 1	5.3375	1.92E-98
4	ENSG00000291237	SOD2	protein_coding	superoxide dismutase 2	4.4111	2.48E-85
5	ENSG00000140105	WARS1	protein_coding	tryptophanyl-tRNA synthetase 1	4.7849	1.81E-83
6	ENSG00000228964	HLA-B	protein_coding	major histocompatibility complex, class I, B	4.1856	9.25E-64
7	ENSG00000125730	C3	protein_coding	complement C3	4.0530	9.79E-55
8	ENSG00000184371	CSF1	protein_coding	colony stimulating factor 1	3.4411	1.71E-52
9	ENSG00000156587	UBE2L6	protein_coding	ubiquitin conjugating enzyme E2 L6	3.5298	1.31E-49
10	ENSG00000240065	PSMB9	protein_coding	proteasome 20S subunit beta 9	5.0740	1.22E-45
11	ENSG00000089127	OAS1	protein_coding	2'-5'-oligoadenylate synthetase 1	4.6015	7.57E-38
12	ENSG00000126709	IFI6	protein_coding	interferon alpha inducible protein 6	4.6587	7.91E-38
13	ENSG00000185338	SOCS1	protein_coding	suppressor of cytokine signaling 1	4.9718	1.30E-37
14	ENSG00000184557	SOCS3	protein_coding	suppressor of cytokine signaling 3	3.9065	4.07E-37
15	ENSG00000234745	HLA-B	protein_coding	major histocompatibility complex, class I, B	3.1331	7.19E-37
16	ENSG00000133321	PLAAT4	protein_coding	phospholipase A and acyltransferase 4	5.8205	8.52E-37
17	ENSG00000205220	PSMB10	protein_coding	proteasome 20S subunit beta 10	3.6227	1.05E-34
18	ENSG00000182326	C1S	protein_coding	complement C1s	6.8814	1.43E-31
19	ENSG00000068079	IFI35	protein_coding	interferon induced protein 35	3.6221	3.81E-31
20	ENSG00000169245	CXCL10	protein_coding	C-X-C motif chemokine ligand 10	14.2993	7.28E-30
21	ENSG00000130589	HELZ2	protein_coding	helicase with zinc finger 2	3.4888	2.08E-28
22	ENSG00000126246	IGFLR1	protein_coding	IGF like family receptor 1	4.0383	1.59E-27
23	ENSG00000164442	CITED2	protein_coding	Cbp/p300 interacting transactivator with Glu/Asp rich carboxy-terminal domain 2	-2.5120	4.29E-27
24	ENSG00000188313	PLSCR1	protein_coding	phospholipid scramblase 1	3.0290	1.37E-25
25	ENSG00000117228	GBP1	protein_coding	guanylate binding protein 1	11.2666	2.11E-25
26	ENSG00000006210	CX3CL1	protein_coding	C-X3-C motif chemokine ligand 1	4.2675	1.90E-24
27	ENSG00000131203	IDO1	protein_coding	indoleamine 2,3-dioxygenase 1	12.4859	1.11E-22
28	ENSG00000206468	UBD	protein_coding	ubiquitin D	12.4734	1.45E-22
29	ENSG00000187608	ISG15	protein_coding	ISG15 ubiquitin like modifier	3.5801	7.58E-22
30	ENSG00000132109	TRIM21	protein_coding	tripartite motif containing 21	2.7008	4.04E-21
31	ENSG00000163840	DTX3L	protein_coding	deltex E3 ubiquitin ligase 3L	3.1337	1.19E-20
32	ENSG00000171223	JUNB	protein_coding	JunB proto-oncogene, AP-1 transcription factor subunit	2.3624	1.23E-20
33	ENSG00000168062	BATF2	protein_coding	basic leucine zipper ATF-like transcription factor 2	8.5723	6.90E-20
34	ENSG00000142089	IFITM3	protein_coding	interferon induced transmembrane protein 3	2.5470	8.46E-20
35	ENSG00000166710	B2M	protein_coding	beta-2-microglobulin	2.5578	1.99E-19

Table 7.18 Top 100 DEGs (by p_{adj}) between untreated and IFNγ + TNF treated WTAP KO 2H1 by reference transcriptome aligned DRS

	Ensembl ID	Gene Symbol	Biotype	Description	Log_2FoldChange	p_{adj}
36	ENSG00000196954	CASP4	protein_coding	caspase 4	2.6452	2.97E-19
37	ENSG00000237988	OR2I1P	protein_coding	olfactory receptor family 2 subfamily I member 1 pseudogene	11.4534	3.54E-19
38	ENSG00000008517	IL32	protein_coding	interleukin 32	2.5808	1.66E-18
39	ENSG00000185885	IFITM1	protein_coding	interferon induced transmembrane protein 1	8.1700	3.02E-18
40	ENSG00000172216	CEBPB	protein_coding	CCAAT enhancer binding protein beta	2.2901	3.78E-18
41	ENSG00000118503	TNFAIP3	protein_coding	TNF alpha induced protein 3	3.8458	4.30E-18
42	ENSG00000196776	CD47	protein_coding	CD47 molecule	2.5008	6.29E-18
43	ENSG00000112149	CD83	protein_coding	CD83 molecule	2.9708	9.92E-18
44	ENSG00000184979	USP18	protein_coding	ubiquitin specific peptidase 18	5.1305	6.83E-17
45	ENSG00000188015	S100A3	protein_coding	S100 calcium binding protein A3	2.6500	6.83E-17
46	ENSG00000138755	CXCL9	protein_coding	C-X-C motif chemokine ligand 9	10.7910	1.02E-16
47	ENSG00000226025		transcribed_unprocessed_pseudogene	Charcot-Leyden crystal protein pseudogene	10.6728	2.91E-16
48	ENSG00000158714	SLAMF8	protein_coding	SLAM family member 8	10.6818	2.99E-16
49	ENSG00000135148	TRAFD1	protein_coding	TRAF-type zinc finger domain containing 1	2.7189	3.27E-16
50	ENSG00000141682	PMAIP1	protein_coding	phorbol-12-myristate-13-acetate-induced protein 1	2.7475	3.35E-16
51	ENSG00000162654	GBP4	protein_coding	guanylate binding protein 4	10.5450	3.41E-16
52	ENSG00000154451	GBP5	protein_coding	guanylate binding protein 5	10.6639	4.04E-16
53	ENSG00000108700	CCL8	protein_coding	C-C motif chemokine ligand 8	10.5704	7.40E-16
54	ENSG00000123609	NMI	protein_coding	N-myc and STAT interactor	2.4861	2.82E-15
55	ENSG00000221963	APOL6	protein_coding	apolipoprotein L6	3.1610	6.04E-15
56	ENSG00000169248	CXCL11	protein_coding	C-X-C motif chemokine ligand 11	10.1549	9.79E-15
57	ENSG00000139192	TAPBPL	protein_coding	TAP binding protein like	2.6276	1.32E-14
58	ENSG00000168404	MLKL	protein_coding	mixed lineage kinase domain like pseudokinase	2.2257	1.86E-14
59	ENSG00000163131	CTSS	protein_coding	cathepsin S	9.8738	4.71E-14
60	ENSG00000137965	IFI44	protein_coding	interferon induced protein 44	4.4708	1.37E-13
61	ENSG00000146859	TMEM140	protein_coding	transmembrane protein 140	2.6497	1.41E-13
62	ENSG00000134470	IL15RA	protein_coding	interleukin 15 receptor subunit alpha	3.4939	1.77E-13
63	ENSG00000137496	IL18BP	protein_coding	interleukin 18 binding protein	7.1623	2.00E-13
64	ENSG00000059378	PARP12	protein_coding	poly(ADP-ribose) polymerase family member 12	2.6195	2.52E-13
65	ENSG00000103196	CRISPLD2	protein_coding	cysteine rich secretory protein LCCL domain containing 2	-2.4172	2.76E-13
66	ENSG00000198959	TGM2	protein_coding	transglutaminase 2	2.3333	2.76E-13
67	ENSG00000197646	PDCD1LG2	protein_coding	programmed cell death 1 ligand 2	2.9439	2.93E-13
68	ENSG00000121858	TNFSF10	protein_coding	TNF superfamily member 10	2.8335	3.41E-13
69	ENSG00000113263	ITK	protein_coding	IL2 inducible T cell kinase	4.9805	4.09E-13
70	ENSG00000140464	PML	protein_coding	PML nuclear body scaffold	2.2560	6.09E-13

Table 7.18 (cont.) Top 100 DEGs (by p_{adj}) between untreated and IFNγ + TNF treated WTAP KO 2H1 by reference transcriptome aligned DRS

	Ensembl ID	Gene Symbol	Biotype	Description	Log_2FoldChange	p_{adj}
71	ENSG00000136883	KIF12	protein_coding	kinesin family member 12	3.1670	8.16E-13
72	ENSG00000188820	CALHM6	protein_coding	calcium homeostasis modulator family member 6	2.1099	8.21E-13
73	ENSG00000241935	HOGA1	protein_coding	4-hydroxy-2-oxoglutarate aldolase 1	-2.5611	1.61E-12
74	ENSG00000120833	SOCS2	protein_coding	suppressor of cytokine signaling 2	-2.5652	4.54E-12
75	ENSG00000111331	OAS3	protein_coding	2'-5'-oligoadenylate synthetase 3	2.8382	5.50E-12
76	ENSG00000081041	CXCL2	protein_coding	C-X-C motif chemokine ligand 2	4.4784	6.29E-12
77	ENSG00000104951	IL4I1	protein_coding	interleukin 4 induced 1	2.5093	1.11E-11
78	ENSG00000132274	TRIM22	protein_coding	tripartite motif containing 22	8.9911	1.58E-11
79	ENSG00000163735	CXCL5	protein_coding	C-X-C motif chemokine ligand 5	2.0346	2.02E-11
80	ENSG00000128203	ASPHD2	protein_coding	aspartate beta-hydroxylase domain containing 2	3.0668	2.27E-11
81	ENSG00000100065	CARD10	protein_coding	caspase recruitment domain family member 10	-2.9599	3.45E-11
82	ENSG00000140853	NLRC5	protein_coding	NLR family CARD domain containing 5	2.8205	3.87E-11
83	ENSG00000170989	S1PR1	protein_coding	sphingosine-1-phosphate receptor 1	-3.0169	3.93E-11
84	ENSG00000138496	PARP9	protein_coding	poly(ADP-ribose) polymerase family member 9	3.2279	4.62E-11
85	ENSG00000152518	ZFP36L2	protein_coding	ZFP36 ring finger protein like 2	2.4398	4.85E-11
86	ENSG00000139112	GABARAPL1	protein_coding	GABA type A receptor associated protein like 1	2.4458	2.35E-10
87	ENSG00000162645	GBP2	protein_coding	guanylate binding protein 2	8.7050	2.43E-10
88	ENSG00000114315	HES1	protein_coding	hes family bHLH transcription factor 1	-3.3386	2.82E-10
89	ENSG00000169429	CXCL8	protein_coding	C-X-C motif chemokine ligand 8	2.0981	3.13E-10
90	ENSG00000115415	STAT1	protein_coding	signal transducer and activator of transcription 1	2.6075	3.69E-10
91	ENSG00000290525	GBP1P1	lncRNA	guanylate binding protein 1 pseudogene 1	8.5727	5.09E-10
92	ENSG00000130813	SHFL	protein_coding	shiftless antiviral inhibitor of ribosomal frameshifting	2.6614	7.16E-10
93	ENSG00000123685	BATF3	protein_coding	basic leucine zipper ATF-like transcription factor 3	2.5972	7.19E-10
94	ENSG00000108771	DHX58	protein_coding	DExH-box helicase 58	3.9128	7.57E-10
95	ENSG00000134215	VAV3	protein_coding	vav guanine nucleotide exchange factor 3	-4.1575	1.33E-09
96	ENSG00000132530	XAF1	protein_coding	XIAP associated factor 1	8.2943	3.41E-09
97	ENSG00000111335	OAS2	protein_coding	2'-5'-oligoadenylate synthetase 2	8.1050	3.69E-09
98	ENSG00000116514	RNF19B	protein_coding	ring finger protein 19B	2.1793	3.85E-09
99	ENSG00000101384	JAG1	protein_coding	jagged canonical Notch ligand 1	-2.1988	4.38E-09
100	ENSG00000134326	CMPK2	protein_coding	cytidine/uridine monophosphate kinase 2	8.0790	4.47E-09

Table 7.18 (cont.) Top 100 DEGs (by p_{adj}) between untreated and IFNγ + TNF treated WTAP KO 2H1 by reference transcriptome aligned DRS

	Ensembl ID	Gene Symbol	Biotype	Description	Log_2FoldChange	p_{adj}
1	ENSG00000185633	NDUFA4L2	protein_coding	NDUFA4 mitochondrial complex associated like 2	-9.4549	1.15E-53
2	ENSG00000107159	CA9	protein_coding	carbonic anhydrase 9	-3.9148	8.32E-35
3	ENSG00000129521	EGLN3	protein_coding	egl-9 family hypoxia inducible factor 3	-6.6649	1.24E-18
4	ENSG00000176171	BNIP3	protein_coding	BCL2 interacting protein 3	-2.1640	3.78E-17
5	ENSG00000104419	NDRG1	protein_coding	N-myc downstream regulated 1	-2.8417	4.43E-16
6	ENSG00000168209	DDIT4	protein_coding	DNA damage inducible transcript 4	-2.7900	2.07E-13
7	ENSG00000109107	ALDOC	protein_coding	aldolase, fructose-bisphosphate C	-3.3031	1.57E-11
8	ENSG00000114268	PFKFB4	protein_coding	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 4	-2.6906	3.71E-10
9	ENSG00000247095	MIR210HG	lncRNA	MIR210 host gene	-2.9386	1.23E-06
10	ENSG00000123146	ADGRE5	protein_coding	adhesion G protein-coupled receptor E5	-7.8071	1.23E-06
11	ENSG00000169903	TM4SF4	protein_coding	transmembrane 4 L six family member 4	2.6865	1.34E-06
12	ENSG00000006606	CCL26	protein_coding	C-C motif chemokine ligand 26	2.8217	1.41E-06
13	ENSG00000145934	TENM2	protein_coding	teneurin transmembrane protein 2	4.2436	3.33E-06
14	ENSG00000129757	CDKN1C	protein_coding	cyclin dependent kinase inhibitor 1C	2.1747	3.58E-06
15	ENSG00000164237	CMBL	protein_coding	carboxymethylenebutenolidase homolog	-7.5557	4.00E-06
16	ENSG00000124785	NRN1	protein_coding	neuritin 1	-6.3510	7.32E-06
17	ENSG00000101210	EEF1A2	protein_coding	eukaryotic translation elongation factor 1 alpha 2	-3.7639	7.32E-06
18	ENSG00000184441		lncRNA	novel transcript, antisense to C21orf2	-2.4704	8.95E-06
19	ENSG00000186352	ANKRD37	protein_coding	ankyrin repeat domain 37	-3.2692	2.66E-05
20	ENSG00000242221	PSG2	protein_coding	pregnancy specific beta-1-glycoprotein 2	-2.7742	8.63E-05
21	ENSG00000100342	APOL1	protein_coding	apolipoprotein L1	-4.6286	1.60E-04
22	ENSG00000287453		lncRNA	novel transcript	6.8444	2.10E-04
23	ENSG00000112655	PTK7	protein_coding	protein tyrosine kinase 7 (inactive)	-2.2750	3.91E-04
24	ENSG00000186472	PCLO	protein_coding	piccolo presynaptic cytomatrix protein	6.9226	4.21E-04
25	ENSG00000120594	PLXDC2	protein_coding	plexin domain containing 2	3.3161	1.44E-03
26	ENSG00000235385	LINC02154	lncRNA	long intergenic non-protein coding RNA 2154	3.2272	2.76E-03
27	ENSG00000073849	ST6GAL1	protein_coding	ST6 beta-galactoside alpha-2,6-sialyltransferase 1	-6.1666	4.16E-03
28	ENSG00000125841	NRSN2	protein_coding	neurensin 2	-2.4948	5.15E-03
29	ENSG00000162069	BICDL2	protein_coding	BICD family like cargo adaptor 2	-5.7698	5.63E-03
30	ENSG00000197632	SERPINB2	protein_coding	serpin family B member 2	-6.3611	6.25E-03
31	ENSG00000256292	LINC02376	lncRNA	long intergenic non-protein coding RNA 2376	-2.7829	8.16E-03
32	ENSG00000157680	DGKI	protein_coding	diacylglycerol kinase iota	2.2009	8.82E-03
33	ENSG00000091622	PITPNM3	protein_coding	PITPNM family member 3	-2.2346	1.10E-02
34	ENSG00000234147		lncRNA	novel transcript	5.9214	1.24E-02
35	ENSG00000230432		lncRNA	novel transcript	-5.4813	1.35E-02

Table 7.19 DEGs between unstimulated RCC4 Cas9 GFP and WTAP KO 2H1 profiled by reference genome alignment

	Ensembl ID	Gene Symbol	Biotype	Description	Log_2FoldChange	p_{adj}
36	ENSG00000164849	GPR146	protein_coding	G protein-coupled receptor 146	-2.2227	1.38E-02
37	ENSG00000089127	OAS1	protein_coding	2'-5'-oligoadenylate synthetase 1	2.8945	1.38E-02
38	ENSG00000258667	HIF1A-AS3	lncRNA	HIF1A antisense RNA 3	-5.4958	1.47E-02
39	ENSG00000143847	PPFIA4	protein_coding	PTPRF interacting protein alpha 4	-4.8083	1.52E-02
40	ENSG00000128512	DOCK4	protein_coding	dedicator of cytokinesis 4	2.5549	1.89E-02
41	ENSG00000155893	PXYLP1	protein_coding	2-phosphoxylose phosphatase 1	2.0415	3.93E-02
42	ENSG00000286162		lncRNA	novel transcript	-3.0663	4.13E-02
43	ENSG00000197358	BNIP3P1	transcribed_processed_pseudogene	BCL2 interacting protein 3 pseudogene 1	-2.1060	5.00E-02
44	ENSG00000147676	MAL2	protein_coding	mal, T cell differentiation protein 2	2.3687	6.26E-02
45	ENSG00000246228	CASC8	lncRNA	cancer susceptibility 8	2.3550	6.58E-02
46	ENSG00000196611	MMP1	protein_coding	matrix metallopeptidase 1	3.8388	6.65E-02
47	ENSG00000275216		lncRNA	novel transcript	5.3890	6.65E-02
48	ENSG00000243944		lncRNA	novel transcript, antisense to PFN2	5.3797	6.87E-02

Table 7.19 (cont.) DEGs between unstimulated RCC4 Cas9 GFP and WTAP KO 2H1 profiled by reference genome alignment

	Ensembl ID	Gene Symbol	Biotype	Description	Log_2FoldChange	p_{adj}
1	ENSG00000107159	CA9	protein_coding	carbonic anhydrase 9	-3.8500	3.29E-29
2	ENSG00000185633	NDUFA4L2	protein_coding	NDUFA4 mitochondrial complex associated like 2	-6.2287	1.69E-20
3	ENSG00000129521	EGLN3	protein_coding	egl-9 family hypoxia inducible factor 3	-6.5782	2.40E-17
4	ENSG00000176171	BNIP3	protein_coding	BCL2 interacting protein 3	-2.0811	1.64E-11
5	ENSG00000114268	PFKFB4	protein_coding	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 4	-2.6523	1.52E-08
6	ENSG00000168209	DDIT4	protein_coding	DNA damage inducible transcript 4	-2.8242	7.57E-07
7	ENSG00000169903	TM4SF4	protein_coding	transmembrane 4 L six family member 4	2.7267	2.65E-06
8	ENSG00000145934	TENM2	protein_coding	teneurin transmembrane protein 2	4.3793	1.25E-05
9	ENSG00000164237	CMBL	protein_coding	carboxymethylenebutenolidase homolog	-7.4756	1.27E-05
10	ENSG00000104419	NDRG1	protein_coding	N-myc downstream regulated 1	-2.6208	2.08E-04
11	ENSG00000109107	ALDOC	protein_coding	aldolase, fructose-bisphosphate C	-3.1072	2.08E-04
12	ENSG00000197632	SERPINB2	protein_coding	serpin family B member 2	-7.2573	2.19E-04
13	ENSG00000124785	NRN1	protein_coding	neuritin 1	-6.5465	3.75E-04
14	ENSG00000151364	KCTD14	protein_coding	potassium channel tetramerization domain containing 14	2.6647	1.44E-03
15	ENSG00000186352	ANKRD37	protein_coding	ankyrin repeat domain 37	-2.5005	1.51E-03
16	ENSG00000023608	SNAPC1	protein_coding	small nuclear RNA activating complex polypeptide 1	2.8078	2.07E-03
17	ENSG00000186472	PCLO	protein_coding	piccolo presynaptic cytomatrix protein	6.7931	2.61E-03
18	ENSG00000242221	PSG2	protein_coding	pregnancy specific beta-1-glycoprotein 2	-2.7170	2.61E-03
19	ENSG00000143590	EFNA3	protein_coding	ephrin A3	-2.4337	4.90E-03
20	ENSG00000120594	PLXDC2	protein_coding	plexin domain containing 2	3.2277	1.76E-02
21	ENSG00000054356	PTPRN	protein_coding	protein tyrosine phosphatase receptor type N	-4.5761	5.61E-02
22	ENSG00000157680	DGKI	protein_coding	diacylglycerol kinase iota	2.1016	5.61E-02
23	ENSG00000147676	MAL2	protein_coding	mal, T cell differentiation protein 2	2.4844	6.16E-02
24	ENSG00000257302	FAHD2P1	processed_pseudogene	fumarylacetoacetate hydrolase domain containing 2 pseudogene 1	5.5561	6.71E-02
25	ENSG00000290907		lncRNA	novel transcript	2.3475	8.88E-02
26	ENSG00000164849	GPR146	protein_coding	G protein-coupled receptor 146	-2.1718	9.77E-02

Table 7.20 DEGs between unstimulated RCC4 Cas9 GFP and WTAP KO 2H1 profiled by reference transcriptome alignment

	Ensembl ID	Gene Symbol	Biotype	Description	Log_2FoldChange	p_{adj}
1	ENSG00000185633	NDUFA4L2	protein_coding	NDUFA4 mitochondrial complex associated like 2	-9.4352	2.48E-58
2	ENSG00000176171	BNIP3	protein_coding	BCL2 interacting protein 3	-2.3804	2.80E-50
3	ENSG00000107159	CA9	protein_coding	carbonic anhydrase 9	-5.5873	1.16E-49
4	ENSG00000104419	NDRG1	protein_coding	N-myc downstream regulated 1	-3.1401	2.06E-41
5	ENSG00000129521	EGLN3	protein_coding	egl-9 family hypoxia inducible factor 3	-6.9169	3.85E-28
6	ENSG00000100342	APOL1	protein_coding	apolipoprotein L1	-2.5643	7.13E-23
7	ENSG00000168209	DDIT4	protein_coding	DNA damage inducible transcript 4	-2.4659	1.86E-19
8	ENSG00000113739	STC2	protein_coding	stanniocalcin 2	-2.0671	4.72E-19
9	ENSG00000181458	TMEM45A	protein_coding	transmembrane protein 45A	-2.2411	5.48E-16
10	ENSG00000114268	PFKFB4	protein_coding	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 4	-2.8761	1.66E-13
11	ENSG00000114023	FAM162A	protein_coding	family with sequence similarity 162 member A	-2.4742	1.92E-13
12	ENSG00000167772	ANGPTL4	protein_coding	angiopoietin like 4	-2.4112	1.92E-13
13	ENSG00000100867	DHRS2	protein_coding	dehydrogenase/reductase 2	2.7247	1.55E-11
14	ENSG00000180730	SHISA2	protein_coding	shisa family member 2	-2.3236	1.94E-09
15	ENSG00000109107	ALDOC	protein_coding	aldolase, fructose-bisphosphate C	-2.8935	2.10E-09
16	ENSG00000242221	PSG2	protein_coding	pregnancy specific beta-1-glycoprotein 2	-3.2933	1.11E-08
17	ENSG00000123146	ADGRE5	protein_coding	adhesion G protein-coupled receptor E5	-7.8088	1.58E-08
18	ENSG00000256292	LINC02376	lncRNA	long intergenic non-protein coding RNA 2376	-2.4770	1.63E-08
19	ENSG00000247095	MIR210HG	lncRNA	MIR210 host gene	-4.0633	2.92E-08
20	ENSG00000186352	ANKRD37	protein_coding	ankyrin repeat domain 37	-3.5256	3.23E-08
21	ENSG00000178776	C5orf46	protein_coding	chromosome 5 open reading frame 46	2.1485	9.09E-08
22	ENSG00000169903	TM4SF4	protein_coding	transmembrane 4 L six family member 4	3.5213	2.17E-07
23	ENSG00000112655	PTK7	protein_coding	protein tyrosine kinase 7 (inactive)	-3.2531	9.30E-07
24	ENSG00000186472	PCLO	protein_coding	piccolo presynaptic cytomatrix protein	4.5957	1.77E-06
25	ENSG00000124785	NRN1	protein_coding	neuritin 1	-6.7569	5.13E-06
26	ENSG00000006606	CCL26	protein_coding	C-C motif chemokine ligand 26	2.5026	1.10E-05
27	ENSG00000265096	C1QTNF1-AS1	lncRNA	C1QTNF1 antisense RNA 1	2.8938	1.54E-05
28	ENSG00000145934	TENM2	protein_coding	teneurin transmembrane protein 2	6.0027	1.86E-05
29	ENSG00000157111	TMEM171	protein_coding	transmembrane protein 171	6.7420	7.79E-05
30	ENSG00000287453	0	lncRNA	novel transcript	6.6680	1.60E-04
31	ENSG00000196611	MMP1	protein_coding	matrix metallopeptidase 1	2.3623	1.99E-04
32	ENSG00000164237	CMBL	protein_coding	carboxymethylenebutenolidase homolog	-6.1555	5.78E-04
33	ENSG00000258667	HIF1A-AS3	lncRNA	HIF1A antisense RNA 3	-5.9960	1.13E-03
34	ENSG00000197632	SERPINB2	protein_coding	serpin family B member 2	-6.9934	1.17E-03
35	ENSG00000000971	CFH	protein_coding	complement factor H	2.0633	1.49E-03

Table 7.21 DEGs between IFNγ+TNF treated RCC4 Cas9 GFP and WTAP KO 2H1 profiled by reference genome alignment

	Ensembl ID	Gene Symbol	Biotype	Description	Log_2FoldChange	p_{adj}
36	ENSG00000228742	LINC02577	lncRNA	long intergenic non-protein coding RNA 2577	2.6975	2.47E-03
37	ENSG00000073849	ST6GAL1	protein_coding	ST6 beta-galactoside alpha-2,6-sialyltransferase 1	-5.7071	3.79E-03
38	ENSG00000162069	BICDL2	protein_coding	BICD family like cargo adaptor 2	-5.3070	4.09E-03
39	ENSG00000143847	PPFIA4	protein_coding	PTPRF interacting protein alpha 4	-5.2993	4.15E-03
40	ENSG00000074047	GLI2	protein_coding	GLI family zinc finger 2	4.2384	4.78E-03
41	ENSG00000248323	LUCAT1	lncRNA	lung cancer associated transcript 1	-3.2517	5.12E-03
42	ENSG00000154274	C4orf19	protein_coding	chromosome 4 open reading frame 19	5.4732	5.42E-03
43	ENSG00000155629	PIK3AP1	protein_coding	phosphoinositide-3-kinase adaptor protein 1	2.0153	6.03E-03
44	ENSG00000137673	MMP7	protein_coding	matrix metallopeptidase 7	2.8668	6.23E-03
45	ENSG00000117480	FAAH	protein_coding	fatty acid amide hydrolase	5.7601	6.39E-03
46	ENSG00000197614	MFAP5	protein_coding	microfibril associated protein 5	3.2087	8.79E-03
47	ENSG00000244383	FAM3D-AS1	lncRNA	FAM3D antisense RNA 1	-4.3722	1.96E-02
48	ENSG00000234147	0	lncRNA	novel transcript	4.9914	2.69E-02
49	ENSG00000272870	SAP30-DT	lncRNA	SAP30 divergent transcript	-2.6687	2.70E-02
50	ENSG00000018236	CNTN1	protein_coding	contactin 1	3.5031	2.86E-02
51	ENSG00000157680	DGKI	protein_coding	diacylglycerol kinase iota	2.7553	3.13E-02
52	ENSG00000189252	SPANXN3	protein_coding	SPANX family member N3	4.8453	3.70E-02
53	ENSG00000129757	CDKN1C	protein_coding	cyclin dependent kinase inhibitor 1C	2.1386	3.73E-02
54	ENSG00000174837	ADGRE1	protein_coding	adhesion G protein-coupled receptor E1	3.7113	4.50E-02
55	ENSG00000162894	FCMR	protein_coding	Fc mu receptor	2.7852	4.76E-02
56	ENSG00000161267	BDH1	protein_coding	3-hydroxybutyrate dehydrogenase 1	-2.1327	4.84E-02
57	ENSG00000120594	PLXDC2	protein_coding	plexin domain containing 2	4.7479	4.87E-02
58	ENSG00000171951	SCG2	protein_coding	secretogranin II	-2.8942	4.90E-02
59	ENSG00000140511	HAPLN3	protein_coding	hyaluronan and proteoglycan link protein 3	3.6874	4.90E-02
60	ENSG00000184254	ALDH1A3	protein_coding	aldehyde dehydrogenase 1 family member A3	2.0103	5.11E-02
61	ENSG00000204711	#N/A	protein_coding	cilia and flagella associated protein 95	5.1328	5.18E-02
62	ENSG00000005249	PRKAR2B	protein_coding	protein kinase cAMP-dependent type II regulatory subunit beta	2.5621	5.27E-02
63	ENSG00000235385	LINC02154	lncRNA	long intergenic non-protein coding RNA 2154	2.2539	6.32E-02
64	ENSG00000164920	OSR2	protein_coding	odd-skipped related transcription factor 2	3.9045	8.93E-02
65	ENSG00000113504	SLC12A7	protein_coding	solute carrier family 12 member 7	-2.8965	9.84E-02

Table 7.21 (cont.) DEGs between IFNγ+TNF treated RCC4 Cas9 GFP and WTAP KO 2H1 profiled by reference genome alignment

	Ensembl ID	Gene Symbol	Biotype	Description	Log_2FoldChange	p_{adj}
1	ENSG00000107159	CA9	protein_coding	carbonic anhydrase 9	-5.5520	2.30E-45
2	ENSG00000176171	BNIP3	protein_coding	BCL2 interacting protein 3	-2.2918	2.79E-30
3	ENSG00000129521	EGLN3	protein_coding	egl-9 family hypoxia inducible factor 3	-6.7761	1.63E-25
4	ENSG00000185633	NDUFA4L2	protein_coding	NDUFA4 mitochondrial complex associated like 2	-5.8081	3.20E-22
5	ENSG00000181458	TMEM45A	protein_coding	transmembrane protein 45A	-2.1781	7.39E-11
6	ENSG00000114268	PFKFB4	protein_coding	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 4	-2.8086	1.21E-10
7	ENSG00000197632	SERPINB2	protein_coding	serpin family B member 2	-9.0007	1.21E-10
8	ENSG00000114023	FAM162A	protein_coding	family with sequence similarity 162 member A	-2.2913	6.89E-09
9	ENSG00000167772	ANGPTL4	protein_coding	angiopoietin like 4	-2.2329	1.18E-08
10	ENSG00000178776	C5orf46	protein_coding	chromosome 5 open reading frame 46	2.2643	4.20E-08
11	ENSG00000169903	TM4SF4	protein_coding	transmembrane 4 L six family member 4	3.7211	1.65E-07
12	ENSG00000168209	DDIT4	protein_coding	DNA damage inducible transcript 4	-2.6444	2.76E-07
13	ENSG00000180730	SHISA2	protein_coding	shisa family member 2	-2.2616	3.76E-07
14	ENSG00000186352	ANKRD37	protein_coding	ankyrin repeat domain 37	-2.8231	6.46E-07
15	ENSG00000242221	PSG2	protein_coding	pregnancy specific beta-1-glycoprotein 2	-3.0313	1.10E-06
16	ENSG00000104419	NDRG1	protein_coding	N-myc downstream regulated 1	-2.6604	1.67E-06
17	ENSG00000186472	PCLO	protein_coding	piccolo presynaptic cytomatrix protein	4.5066	7.18E-06
18	ENSG00000123146	ADGRE5	protein_coding	adhesion G protein-coupled receptor E5	-6.6208	7.47E-05
19	ENSG00000145934	TENM2	protein_coding	teneurin transmembrane protein 2	5.6313	1.95E-04
20	ENSG00000124785	NRN1	protein_coding	neuritin 1	-6.0479	2.07E-04
21	ENSG00000196611	MMP1	protein_coding	matrix metallopeptidase 1	2.4244	2.74E-04
22	ENSG00000109107	ALDOC	protein_coding	aldolase, fructose-bisphosphate C	-3.6411	2.94E-04
23	ENSG00000100867	DHRS2	protein_coding	dehydrogenase/reductase 2	2.7197	6.11E-04
24	ENSG00000290907		lncRNA	novel transcript	2.9720	7.14E-04
25	ENSG00000164237	CMBL	protein_coding	carboxymethylenebutenolidase homolog	-6.0692	1.20E-03
26	ENSG00000137673	MMP7	protein_coding	matrix metallopeptidase 7	2.9201	7.61E-03
27	ENSG00000157111	TMEM171	protein_coding	transmembrane protein 171	5.7660	1.04E-02
28	ENSG00000117480	FAAH	protein_coding	fatty acid amide hydrolase	5.7287	1.10E-02
29	ENSG00000155629	PIK3AP1	protein_coding	phosphoinositide-3-kinase adaptor protein 1	2.1210	1.23E-02
30	ENSG00000197614	MFAP5	protein_coding	microfibril associated protein 5	4.0368	1.46E-02
31	ENSG00000112655	PTK7	protein_coding	protein tyrosine kinase 7 (inactive)	-5.0037	1.54E-02
32	ENSG00000182118	FAM89A	protein_coding	family with sequence similarity 89 member A	2.8492	1.90E-02
33	ENSG00000074047	GLI2	protein_coding	GLI family zinc finger 2	3.9876	2.05E-02
34	ENSG00000204389	HSPA1A	protein_coding	heat shock protein family A (Hsp70) member 1A	-2.2103	4.00E-02
35	ENSG00000189252	SPANXN3	protein_coding	SPANX family member N3	4.8887	4.75E-02

Table 7.22 DEGs between IFNγ + TNF treated RCC4 Cas9 GFP and WTAP KO 2H1 profiled by reference transcriptome alignment

	Ensembl ID	Gene Symbol	Biotype	Description	Log_2FoldChange	p_{adj}
36	ENSG00000154274	C4orf19	protein_coding	chromosome 4 open reading frame 19	4.8342	5.70E-02
37	ENSG00000204310	AGPAT1	protein_coding	1-acylglycerol-3-phosphate O-acyltransferase 1	-2.0749	5.90E-02
38	ENSG00000157680	DGKI	protein_coding	diacylglycerol kinase iota	2.6943	6.56E-02
39	ENSG00000291174		lncRNA	novel transcript	-2.4503	6.68E-02
40	ENSG00000120594	PLXDC2	protein_coding	plexin domain containing 2	4.6605	9.34E-02
41	ENSG00000285839		protein_coding	novel transcript	3.2566	9.34E-02

Table 7.22 (cont.) DEGs between IFNγ + TNF treated RCC4 Cas9 GFP and WTAP KO 2H1 profiled by reference transcriptome alignment

	ID	Description	setSize	NES	p_{adj}
1	GO:0009615	response to virus	314	2.8256	3.95E-33
2	GO:0034341	response to interferon-gamma	104	3.0175	1.67E-26
3	GO:0051607	defense response to virus	239	2.7844	1.67E-26
4	GO:0140546	defense response to symbiont	239	2.7844	1.67E-26
5	GO:0002831	regulation of response to biotic stimulus	285	2.6038	1.46E-21
6	GO:0031347	regulation of defense response	436	2.3980	2.35E-20
7	GO:0045088	regulation of innate immune response	191	2.6958	1.29E-19
8	GO:0019221	cytokine-mediated signaling pathway	326	2.4943	1.29E-19
9	GO:0009617	response to bacterium	384	2.3838	4.70E-19
10	GO:0002250	adaptive immune response	302	2.4748	1.59E-18
11	GO:0071346	cellular response to interferon-gamma	85	2.8303	3.67E-18
12	GO:0006954	inflammatory response	482	2.2542	1.12E-17
13	GO:0050778	positive regulation of immune response	361	2.3405	1.94E-17
14	GO:0001819	positive regulation of cytokine production	323	2.3682	3.68E-16
15	GO:0050777	negative regulation of immune response	135	2.6382	2.91E-15
16	GO:0002460	adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	214	2.4619	9.90E-15
17	GO:0016032	viral process	364	2.2603	1.09E-14
18	GO:0002683	negative regulation of immune system process	286	2.3567	5.12E-14
19	GO:0002252	immune effector process	392	2.1961	1.39E-13
20	GO:0019058	viral life cycle	273	2.3314	1.83E-13
21	GO:1903900	regulation of viral life cycle	120	2.6191	2.58E-13
22	GO:0050792	regulation of viral process	138	2.5715	2.75E-13
23	GO:0045824	negative regulation of innate immune response	75	2.6941	4.16E-13
24	GO:0048525	negative regulation of viral process	82	2.6423	9.47E-13
25	GO:0019882	antigen processing and presentation	115	2.5527	1.39E-12
26	GO:0019079	viral genome replication	122	2.5725	1.42E-12
27	GO:0002443	leukocyte mediated immunity	264	2.2655	5.30E-12
28	GO:0032103	positive regulation of response to external stimulus	293	2.2247	8.88E-12
29	GO:0002521	leukocyte differentiation	384	2.1220	8.88E-12
30	GO:0048002	antigen processing and presentation of peptide antigen	78	2.6294	2.92E-11
31	GO:0034612	response to tumor necrosis factor	182	2.3706	3.73E-11
32	GO:1903131	mononuclear cell differentiation	301	2.1951	3.97E-11
33	GO:0002449	lymphocyte mediated immunity	210	2.3157	4.15E-11
34	GO:0002832	negative regulation of response to biotic stimulus	107	2.5019	6.93E-11
35	GO:0002764	immune response-regulating signaling pathway	298	2.1734	8.24E-11
36	GO:0002819	regulation of adaptive immune response	147	2.4003	8.78E-11
37	GO:0031348	negative regulation of defense response	179	2.3638	8.78E-11
38	GO:0045069	regulation of viral genome replication	78	2.5934	1.17E-10
39	GO:0045071	negative regulation of viral genome replication	51	2.6166	2.76E-10
40	GO:0002697	regulation of immune effector process	232	2.2293	4.11E-10
41	GO:0002221	pattern recognition receptor signaling pathway	141	2.3507	8.69E-10
42	GO:0034340	response to type I interferon	63	2.5814	8.96E-10
43	GO:0032496	response to lipopolysaccharide	211	2.2168	1.06E-09
44	GO:0042110	T cell activation	342	2.0589	1.41E-09
45	GO:0002237	response to molecule of bacterial origin	224	2.2034	1.74E-09
46	GO:0051249	regulation of lymphocyte activation	322	2.0710	2.42E-09
47	GO:0002822	regulation of adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	134	2.3411	2.59E-09
48	GO:0098586	cellular response to virus	73	2.5556	2.91E-09
49	GO:0071357	cellular response to type I interferon	58	2.5202	3.68E-09
50	GO:0031349	positive regulation of defense response	192	2.2317	6.48E-09

Table 7.23 Top 50 GO BP terms (by p_{adj}) between unstimulated and IFNγ + TNF treated RCC4 Cas9 GFP

	ID	Description	setSize	NES	p_{adj}
1	GO:0005126	cytokine receptor binding	152	-1.9925	3.90E-09
2	GO:0042379	chemokine receptor binding	33	-2.1837	2.25E-08
3	GO:0042605	peptide antigen binding	48	-1.9785	1.93E-07
4	GO:0008009	chemokine activity	24	-1.9341	3.78E-07
5	GO:0003823	antigen binding	63	-1.9703	5.08E-07
6	GO:0046977	TAP binding	35	-1.8957	2.85E-06
7	GO:0005125	cytokine activity	110	-2.0045	1.24E-05
8	GO:0003725	double-stranded RNA binding	67	-1.8771	5.72E-05
9	GO:0048020	CCR chemokine receptor binding	16	-1.7956	6.48E-05
10	GO:0008233	peptidase activity	388	-1.8271	6.48E-04
11	GO:0001664	G protein-coupled receptor binding	163	-1.9974	8.49E-04
12	GO:0023026	MHC class II protein complex binding	18	-1.7733	1.45E-03
13	GO:0045236	CXCR chemokine receptor binding	12	-1.7230	1.45E-03
14	GO:0140098	catalytic activity, acting on RNA	382	-1.9151	1.66E-03
15	GO:0023023	MHC protein complex binding	24	-1.9070	3.83E-03
16	GO:0140375	immune receptor activity	75	-2.0701	3.83E-03
17	GO:0004298	threonine-type endopeptidase activity	22	-1.9474	4.42E-03
18	GO:0016887	ATP hydrolysis activity	291	-1.9255	4.42E-03
19	GO:0140097	catalytic activity, acting on DNA	204	-2.0971	5.28E-03
20	GO:0004175	endopeptidase activity	257	-1.9689	6.34E-03
21	GO:0004386	helicase activity	158	-2.0390	7.41E-03
22	GO:0042277	peptide binding	229	-1.9785	7.41E-03
23	GO:0022836	gated channel activity	158	-2.2593	1.20E-02
24	GO:0004896	cytokine receptor activity	53	-1.9644	1.44E-02
25	GO:0030545	signaling receptor regulator activity	257	-1.9526	1.44E-02
26	GO:0016796	exonuclease activity, active with either ribo- or deoxyribonucleic acids and producing 5'-phosphomonoesters	57	1.8742	1.44E-02
27	GO:0008234	cysteine-type peptidase activity	134	1.7979	1.44E-02
28	GO:0004532	exoribonuclease activity	40	2.0196	1.62E-02
29	GO:0031730	CCR5 chemokine receptor binding	5	1.8163	1.62E-02
30	GO:0030546	signaling receptor activator activity	241	1.6163	1.62E-02
31	GO:0016896	exoribonuclease activity, producing 5'-phosphomonoesters	38	1.9576	1.68E-02
32	GO:0004197	cysteine-type endopeptidase activity	84	1.8081	1.78E-02
33	GO:0004540	ribonuclease activity	103	1.7902	1.84E-02
34	GO:0008094	ATP-dependent activity, acting on DNA	100	1.8072	1.89E-02
35	GO:0140657	ATP-dependent activity	453	1.4952	1.89E-02
36	GO:0030881	beta-2-microglobulin binding	19	1.9619	2.07E-02
37	GO:0035325	Toll-like receptor binding	5	1.7977	2.21E-02
38	GO:0005525	GTP binding	324	1.5501	2.21E-02
39	GO:0048018	receptor ligand activity	235	1.6155	2.21E-02
40	GO:0019001	guanyl nucleotide binding	336	1.5054	2.45E-02
41	GO:0032561	guanyl ribonucleotide binding	336	1.5054	2.45E-02
42	GO:0046978	TAP1 binding	15	1.9282	2.77E-02
43	GO:0032395	MHC class II receptor activity	6	1.8153	3.00E-02
44	GO:0019955	cytokine binding	79	1.8027	3.00E-02
45	GO:0071889	14-3-3 protein binding	35	1.9090	3.09E-02
46	GO:0015075	ion transmembrane transporter activity	499	-1.3339	3.18E-02
47	GO:0005244	voltage-gated ion channel activity	97	-1.6824	3.63E-02
48	GO:0022832	voltage-gated channel activity	97	-1.6824	3.63E-02
49	GO:0070003	threonine-type peptidase activity	28	1.9573	4.28E-02
50	GO:0036312	phosphatidylinositol 3-kinase regulatory subunit binding	7	1.8052	4.28E-02

Table 7.24 Top 50 GO MF terms (by p_{adj}) between unstimulated and IFNγ + TNF treated RCC4 Cas9 GFP

	ID	Description	setSize	NES	p_{adj}
1	GO:0009615	response to virus	314	2.5910	4.46E-28
2	GO:0034341	response to interferon-gamma	102	2.8452	3.63E-26
3	GO:0009617	response to bacterium	386	2.4387	1.58E-24
4	GO:0051607	defense response to virus	239	2.5915	5.06E-24
5	GO:0140546	defense response to symbiont	239	2.5915	5.06E-24
6	GO:0031347	regulation of defense response	437	2.3262	2.39E-22
7	GO:0002831	regulation of response to biotic stimulus	282	2.4838	3.35E-22
8	GO:0002250	adaptive immune response	303	2.4504	5.43E-22
9	GO:0006954	inflammatory response	489	2.2537	9.47E-21
10	GO:0001819	positive regulation of cytokine production	332	2.3658	3.32E-20
11	GO:0045088	regulation of innate immune response	188	2.5544	7.21E-20
12	GO:0050778	positive regulation of immune response	356	2.3273	1.33E-19
13	GO:0019221	cytokine-mediated signaling pathway	339	2.3384	4.15E-19
14	GO:0071346	cellular response to interferon-gamma	82	2.6621	5.64E-17
15	GO:0002460	adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	220	2.4132	1.45E-16
16	GO:0002449	lymphocyte mediated immunity	210	2.4149	2.49E-16
17	GO:0002683	negative regulation of immune system process	294	2.2870	3.24E-16
18	GO:0002237	response to molecule of bacterial origin	230	2.3135	9.40E-15
19	GO:0002819	regulation of adaptive immune response	154	2.4055	3.29E-14
20	GO:0002252	immune effector process	399	2.1326	4.06E-14
21	GO:0002443	leukocyte mediated immunity	264	2.2623	5.17E-14
22	GO:0019058	viral life cycle	276	2.2327	6.23E-14
23	GO:0045824	negative regulation of innate immune response	73	2.5666	1.93E-13
24	GO:0019079	viral genome replication	120	2.4659	2.76E-13
25	GO:0002709	regulation of T cell mediated immunity	77	2.5310	2.95E-13
26	GO:0032496	response to lipopolysaccharide	217	2.2868	2.95E-13
27	GO:0050777	negative regulation of immune response	139	2.4134	3.02E-13
28	GO:0016032	viral process	367	2.1061	5.04E-13
29	GO:0032103	positive regulation of response to external stimulus	295	2.1841	6.81E-13
30	GO:0002706	regulation of lymphocyte mediated immunity	132	2.3810	4.29E-12
31	GO:0002822	regulation of adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	141	2.3748	4.59E-12
32	GO:0002456	T cell mediated immunity	95	2.4387	4.59E-12
33	GO:0050792	regulation of viral process	139	2.3619	4.59E-12
34	GO:0019882	antigen processing and presentation	107	2.4177	5.31E-12
35	GO:1903900	regulation of viral life cycle	121	2.4029	5.31E-12
36	GO:0002821	positive regulation of adaptive immune response	99	2.3973	1.54E-11
37	GO:0002703	regulation of leukocyte mediated immunity	175	2.2733	1.58E-11
38	GO:0002711	positive regulation of T cell mediated immunity	61	2.5113	1.76E-11
39	GO:0002697	regulation of immune effector process	244	2.1771	2.57E-11
40	GO:0030595	leukocyte chemotaxis	145	2.3139	3.23E-11
41	GO:0031341	regulation of cell killing	75	2.4427	4.15E-11
42	GO:0002832	negative regulation of response to biotic stimulus	103	2.4187	4.89E-11
43	GO:0001906	cell killing	117	2.3856	5.38E-11
44	GO:0002764	immune response-regulating signaling pathway	293	2.0978	5.74E-11
45	GO:0002824	positive regulation of adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	95	2.3802	8.30E-11
46	GO:1990868	response to chemokine	53	2.5103	9.81E-11
47	GO:1990869	cellular response to chemokine	53	2.5103	9.81E-11
48	GO:0097529	myeloid leukocyte migration	139	2.2963	1.07E-10
49	GO:0048002	antigen processing and presentation of peptide antigen	74	2.4582	1.53E-10
50	GO:0071219	cellular response to molecule of bacterial origin	147	2.2710	1.55E-10

Table 7.25 Top 50 GO BP terms (by p_{adj}) between unstimulated and IFNγ + TNF treated WTAP KO 2H1

	ID	Description	setSize	NES	p_{adj}
1	GO:0005126	cytokine receptor binding	160	2.3697	3.29E-11
2	GO:0042379	chemokine receptor binding	32	2.4657	3.66E-10
3	GO:0008009	chemokine activity	23	2.4099	8.46E-10
4	GO:0005125	cytokine activity	114	2.2678	2.42E-08
5	GO:0003823	antigen binding	61	2.3599	6.15E-08
6	GO:0042605	peptide antigen binding	44	2.3530	2.12E-07
7	GO:0046977	TAP binding	32	2.3272	3.60E-07
8	GO:0001664	G protein-coupled receptor binding	165	2.0502	7.94E-06
9	GO:0042277	peptide binding	224	1.9280	7.94E-06
10	GO:0048020	CCR chemokine receptor binding	15	2.1695	5.35E-05
11	GO:0030545	signaling receptor regulator activity	265	1.8143	1.06E-04
12	GO:0030546	signaling receptor activator activity	249	1.8073	1.09E-04
13	GO:0030881	beta-2-microglobulin binding	19	2.1416	1.77E-04
14	GO:0048018	receptor ligand activity	243	1.8085	1.78E-04
15	GO:0045236	CXCR chemokine receptor binding	13	2.0817	9.12E-04
16	GO:0033218	amide binding	293	1.7175	9.69E-04
17	GO:0003725	double-stranded RNA binding	67	1.9734	1.26E-03
18	GO:0008308	voltage-gated anion channel activity	13	-2.0257	8.29E-03
19	GO:0046873	metal ion transmembrane transporter activity	242	-1.5965	8.44E-03
20	GO:0070097	delta-catenin binding	6	-1.8107	1.83E-02
21	GO:0019955	cytokine binding	80	1.8135	2.49E-02
22	GO:0005216	ion channel activity	246	-1.4962	2.49E-02
23	GO:0140375	immune receptor activity	76	1.8091	2.77E-02
24	GO:0042610	CD8 receptor binding	8	1.8529	2.79E-02
25	GO:0031730	CCR5 chemokine receptor binding	5	1.7432	2.83E-02
26	GO:1990404	NAD+-protein ADP-ribosyltransferase activity	17	1.9232	2.88E-02
27	GO:0015318	inorganic molecular entity transmembrane transporter activity	431	-1.3702	2.88E-02
28	GO:0004888	transmembrane signaling receptor activity	416	1.4544	2.95E-02
29	GO:0015267	channel activity	265	-1.4643	3.19E-02
30	GO:0022803	passive transmembrane transporter activity	265	-1.4643	3.19E-02
31	GO:0022836	gated channel activity	174	-1.5679	3.41E-02
32	GO:0046979	TAP2 binding	11	1.8836	3.79E-02
33	GO:0002162	dystroglycan binding	9	1.8412	3.89E-02
34	GO:0046978	TAP1 binding	14	1.8960	4.25E-02
35	GO:0005007	fibroblast growth factor-activated receptor activity	5	-1.7188	4.25E-02

Table 7.26 Top GO MF terms (by p_{adj}) between unstimulated and IFNγ+TNF treated WTAP KO 2H1

	Ensembl ID	Gene Symbol	Description	DRIMseq p_{adj}	DEXseq p_{adj}
1	ENSG00000162522	KIAA1522	KIAA1522	2.37E-02	2.01E-02
2	ENSG00000162437	RAVER2	ribonucleoprotein, PTB binding 2	7.82E-02	9.65E-02
3	ENSG00000171848	RRM2	ribonucleotide reductase regulatory subunit M2	5.47E-05	8.90E-01
4	ENSG00000168036	CTNNB1	catenin beta 1	1.77E-01	1
5	ENSG00000164054	SHISA5	shisa family member 5	7.87E-02	9.70E-02
6	ENSG00000041880	PARP3	poly(ADP-ribose) polymerase family member 3	7.50E-02	9.53E-02
7	ENSG00000188313	PLSCR1	phospholipid scramblase 1	2.72E-02	5.81E-02
8	ENSG00000065882	TBC1D1	TBC1 domain family member 1	3.97E-02	2.42E-01
9	ENSG00000138757	G3BP2	G3BP stress granule assembly factor 2	1	1
10	ENSG00000049167	ERCC8	ERCC excision repair 8, CSA ubiquitin ligase complex subunit	1	7.97E-02
11	ENSG00000214944	ARHGEF28	Rho guanine nucleotide exchange factor 28	5.37E-04	4.71E-01
12	ENSG00000153113	CAST	calpastatin	6.55E-02	1
13	ENSG00000145730	PAM	peptidylglycine alpha-amidating monooxygenase	2.04E-02	1
14	ENSG00000204267	TAP2	transporter 2, ATP binding cassette subfamily B member	3.36E-02	1
15	ENSG00000204264	PSMB8	proteasome 20S subunit beta 8	7.37E-02	9.53E-02
16	ENSG00000137216	TMEM63B	transmembrane protein 63B	2.04E-02	2.22E-02
17	ENSG00000272398	CD24	CD24 molecule	3.97E-02	5.28E-02
18	ENSG00000010818	HIVEP2	HIVEP zinc finger 2	4.14E-02	5.65E-02
19	ENSG00000196262	PPIA	peptidylprolyl isomerase A	2.10E-05	3.00E-05
20	ENSG00000105825	TFPI2	tissue factor pathway inhibitor 2	3.36E-02	4.29E-02
21	ENSG00000122783	CYREN	cell cycle regulator of NHEJ	1.41E-05	5.05E-01
22	ENSG00000153707	PTPRD	protein tyrosine phosphatase receptor type D	7.50E-02	9.53E-02
23	ENSG00000119139	TJP2	tight junction protein 2	1	7.85E-02
24	ENSG00000187764	SEMA4D	semaphorin 4D	1	1
25	ENSG00000150093	ITGB1	integrin subunit beta 1	6.42E-02	1
26	ENSG00000266412	NCOA4	nuclear receptor coactivator 4	6.32E-02	8.03E-02
27	ENSG00000107796	ACTA2	actin alpha 2, smooth muscle	7.50E-02	9.53E-02
28	ENSG00000214413	BBIP1	BBSome interacting protein 1	4.14E-02	1
29	ENSG00000165868	HSPA12A	heat shock protein family A (Hsp70) member 12A	1	1
30	ENSG00000142089	IFITM3	interferon induced transmembrane protein 3	1.08E-03	1.66E-03
31	ENSG00000149091	DGKZ	diacylglycerol kinase zeta	1	1
32	ENSG00000173039	RELA	RELA proto-oncogene, NF-kB subunit	7.82E-02	9.65E-02
33	ENSG00000137497	NUMA1	nuclear mitotic apparatus protein 1	6.32E-02	1
34	ENSG00000159063	ALG8	ALG8 alpha-1,3-glucosyltransferase	4.95E-07	8.03E-02
35	ENSG00000257261	SLC38A4-AS1	SLC38A4 antisense RNA 1	2.34E-02	2.38E-02

Table 7.27 DTU genes between unstimulated and IFNγ+TNF treated RCC4 Cas9 GFP by DRIMseq and DEXseq

	Ensembl ID	Gene Symbol	Description	DRIMseq p_{adj}	DEXseq p_{adj}
36	ENSG00000123415	SMUG1	single-strand-selective monofunctional uracil-DNA glycosylase 1	1	1
37	ENSG00000140105	WARS1	tryptophanyl-tRNA synthetase 1	1.41E-05	3.89E-06
38	ENSG00000140464	PML	PML nuclear body scaffold	1	1
39	ENSG00000140612	SEC11A	SEC11 homolog A, signal peptidase complex subunit	1.61E-05	1
40	ENSG00000198736	MSRB1	methionine sulfoxide reductase B1	4.88E-02	6.45E-02
41	ENSG00000008517	IL32	interleukin 32	3.36E-02	1
42	ENSG00000085644	ZNF213	zinc finger protein 213	3.36E-02	4.29E-02
43	ENSG00000179044	EXOC3L1	exocyst complex component 3 like 1	1.42E-02	6.45E-02
44	ENSG00000205220	PSMB10	proteasome 20S subunit beta 10	6.55E-02	8.66E-02
45	ENSG00000131473	ACLY	ATP citrate lyase	4.88E-02	1
46	ENSG00000169727	GPS1	G protein pathway suppressor 1	3.59E-02	8.89E-03
47	ENSG00000214049	UCA1	urothelial cancer associated 1	1	1
48	ENSG00000042753	AP2S1	adaptor related protein complex 2 subunit sigma 1	1.60E-03	2.37E-03
49	ENSG00000125826	RBCK1	RANBP2-type and C3HC4-type zinc finger containing 1	5.99E-12	1
50	ENSG00000171552	BCL2L1	BCL2 like 1	9.96E-02	5.28E-02
51	ENSG00000198959	TGM2	transglutaminase 2	1.42E-02	1.67E-02
52	ENSG00000236830	CBR3-AS1	CBR3 antisense RNA 1	4.88E-02	6.45E-02
53	ENSG00000102100	SLC35A2	solute carrier family 35 member A2	1	1
54	ENSG00000115762	PLEKHB2	pleckstrin homology domain containing B2	9.96E-02	1.23E-01
55	ENSG00000066379	POLR1H	RNA polymerase I subunit H	9.96E-02	1.20E-01
56	ENSG00000105854	PON2	paraoxonase 2	8.73E-02	1.08E-01
57	ENSG00000165282	PIGO	phosphatidylinositol glycan anchor biosynthesis class O	3.36E-02	1.31E-01
58	ENSG00000095794	CREM	cAMP responsive element modulator	9.96E-02	1
59	ENSG00000007168	PAFAH1B1	platelet activating factor acetylhydrolase 1b regulatory subunit 1	9.10E-02	1.09E-01
60	ENSG00000172315	TP53RK	TP53 regulating kinase	9.00E-02	1.09E-01
61	ENSG00000156261	CCT8	chaperonin containing TCP1 subunit 8	9.00E-02	1.09E-01

Table 7.27 (cont.) DTU genes between unstimulated and IFNγ+TNF treated RCC4 Cas9 GFP by DRIMseq and DEXseq

	Ensembl ID	Gene Symbol	Description	DRIMseq p_{adj}	DEXseq p_{adj}
1	ENSG00000143436	MRPL9	mitochondrial ribosomal protein L9	3.60E-01	2.23E-03
2	ENSG00000143545	RAB13	RAB13, member RAS oncogene family	1	8.29E-02
3	ENSG00000234741	GAS5	growth arrest specific 5	1	9.01E-02
4	ENSG00000079785	DDX1	DEAD-box helicase 1	1	7.34E-02
5	ENSG00000119801	YPEL5	yippee like 5	1	1.15E-10
6	ENSG00000123609	NMI	N-myc and STAT interactor	1	1.12E-07
7	ENSG00000003402	CFLAR	CASP8 and FADD like apoptosis regulator	3.97E-01	2.87E-03
8	ENSG00000123992	DNPEP	aspartyl aminopeptidase	1	8.95E-02
9	ENSG00000123933	MXD4	MAX dimerization protein 4	1	1.64E-02
10	ENSG00000204387	SNHG32	small nucleolar RNA host gene 32	7.46E-01	4.63E-02
11	ENSG00000204264	PSMB8	proteasome 20S subunit beta 8	7.54E-01	5.87E-04
12	ENSG00000272398	CD24	CD24 molecule	7.54E-01	2.67E-03
13	ENSG00000196262	PPIA	peptidylprolyl isomerase A	6.96E-01	8.22E-02
14	ENSG00000105825	TFPI2	tissue factor pathway inhibitor 2	4.05E-01	1.54E-03
15	ENSG00000004864	SLC25A13	solute carrier family 25 member 13	1	6.35E-02
16	ENSG00000257923	CUX1	cut like homeobox 1	1	3.21E-06
17	ENSG00000122783	CYREN	cell cycle regulator of NHEJ	1	3.84E-04
18	ENSG00000197694	SPTAN1	spectrin alpha, non-erythrocytic 1	1	2.74E-03
19	ENSG00000151892	GFRA1	GDNF family receptor alpha 1	1	6.29E-04
20	ENSG00000284969		novel protein	1	1.27E-02
21	ENSG00000173039	RELA	RELA proto-oncogene, NF-kB subunit	5.22E-02	6.63E-04
22	ENSG00000171067	C11orf24	chromosome 11 open reading frame 24	1	7.22E-02
23	ENSG00000109861	CTSC	cathepsin C	1	7.71E-03
24	ENSG00000089157	RPLP0	ribosomal protein lateral stalk subunit P0	1	1.48E-04
25	ENSG00000092010	PSME1	proteasome activator subunit 1	1	6.82E-07
26	ENSG00000140105	WARS1	tryptophanyl-tRNA synthetase 1	3.61E-02	1.41E-10
27	ENSG00000078304	PPP2R5C	protein phosphatase 2 regulatory subunit B'gamma	1	7.41E-02
28	ENSG00000140464	PML	PML nuclear body scaffold	1	5.93E-02
29	ENSG00000008517	IL32	interleukin 32	1	2.77E-07
30	ENSG00000102897	LYRM1	LYR motif containing 1	1	6.59E-02
31	ENSG00000149932	TMEM219	transmembrane protein 219	1	4.18E-04
32	ENSG00000153774	CFDP1	craniofacial development protein 1	3.60E-01	1.96E-02
33	ENSG00000167508	MVD	mevalonate diphosphate decarboxylase	1	9.96E-02
34	ENSG00000013306	SLC25A39	solute carrier family 25 member 39	1	8.73E-02
35	ENSG00000108946	PRKAR1A	protein kinase cAMP-dependent type I regulatory subunit alpha	1	7.57E-03

Table 7.28 DTU genes between unstimulated and IFNγ+TNF treated WTAP KO 2H1 by DRIMseq and DEXseq

	Ensembl ID	Gene Symbol	Description	DRIMseq p_{adj}	DEXseq p_{adj}
36	ENSG00000119559	C19orf25	chromosome 19 open reading frame 25	1	1.75E-02
37	ENSG00000042753	AP2S1	adaptor related protein complex 2 subunit sigma 1	9.91E-01	6.18E-04
38	ENSG00000198131	ZNF544	zinc finger protein 544	1	4.27E-03
39	ENSG00000125826	RBCK1	RANBP2-type and C3HC4-type zinc finger containing 1	1	0.00E+00
40	ENSG00000132635	PCED1A	PC-esterase domain containing 1A	1	8.95E-02
41	ENSG00000171552	BCL2L1	BCL2 like 1	1	1.08E-02
42	ENSG00000101000	PROCR	protein C receptor	4.47E-01	4.52E-04
43	ENSG00000198959	TGM2	transglutaminase 2	1	1.17E-02
44	ENSG00000228109	MELTF-AS1	MELTF antisense RNA 1	1	1
45	ENSG00000039523	RIPOR1	RHO family interacting cell polarization regulator 1	5.22E-02	1

Table 7.28 (cont.) DTU genes between unstimulated and IFNγ+TNF treated WTAP KO 2H1 by DRIMseq and DEXseq

	Ensembl ID	Gene Symbol	Description	DRIMseq p_{adj}	DEXseq p_{adj}
1	ENSG00000162437	RAVER2	ribonucleoprotein, PTB binding 2	0.0581	1.0000
2	ENSG00000162664	ZNF326	zinc finger protein 326	0.0581	1.0000
3	ENSG00000124813	RUNX2	RUNX family transcription factor 2	0.0000	0.0000
4	ENSG00000196275	GTF2IRD2	GTF2I repeat domain containing 2	0.0581	1.0000
5	ENSG00000150093	ITGB1	integrin subunit beta 1	0.0980	1.0000
6	ENSG00000121236	TRIM6	tripartite motif containing 6	0.0581	1.0000
7	ENSG00000149925	ALDOA	aldolase, fructose-bisphosphate A	0.0004	0.0000
8	ENSG00000108963	DPH1	diphthamide biosynthesis 1	0.0980	1.0000
9	ENSG00000214049	UCA1	urothelial cancer associated 1	0.0667	1.0000
10	ENSG00000078699	CBFA2T2	CBFA2/RUNX1 partner transcriptional co-repressor 2	0.0845	1.0000
11	ENSG00000160226	CFAP410	cilia and flagella associated protein 410	0.0026	0.0000
12	ENSG00000100280	AP1B1	adaptor related protein complex 1 subunit beta 1	0.0581	1.0000
13	ENSG00000236200	KDM4A-AS1	KDM4A antisense RNA 1	1.0000	0.0373
14	ENSG00000154511	DIPK1A	divergent protein kinase domain 1A	1.0000	0.0320
15	ENSG00000134215	VAV3	vav guanine nucleotide exchange factor 3	1.0000	0.0123
16	ENSG00000121940	CLCC1	chloride channel CLIC like 1	1.0000	0.0000
17	ENSG00000143774	GUK1	guanylate kinase 1	0.9289	0.0072
18	ENSG00000151779	NBAS	NBAS subunit of NRZ tethering complex	1.0000	0.0160
19	ENSG00000138095	LRPPRC	leucine rich pentatricopeptide repeat containing	1.0000	0.0066
20	ENSG00000153250	RBMS1	RNA binding motif single stranded interacting protein 1	0.6999	0.0089
21	ENSG00000003402	CFLAR	CASP8 and FADD like apoptosis regulator	0.2812	0.0139
22	ENSG00000003393	ALS2	alsin Rho guanine nucleotide exchange factor ALS2	1.0000	0.0002
23	ENSG00000144591	GMPPA	GDP-mannose pyrophosphorylase A	0.7253	0.0266
24	ENSG00000185049	NELFA	negative elongation factor complex member A	1.0000	0.0004
25	ENSG00000048342	CC2D2A	coiled-coil and C2 domain containing 2A	1.0000	0.0111
26	ENSG00000214944	ARHGEF28	Rho guanine nucleotide exchange factor 28	1.0000	0.0000
27	ENSG00000235142	LINC02532	long intergenic non-protein coding RNA 2532	1.0000	0.0267
28	ENSG00000146833	TRIM4	tripartite motif containing 4	1.0000	0.0046
29	ENSG00000090263	MRPS33	mitochondrial ribosomal protein S33	0.1001	0.0230
30	ENSG00000105993	DNAJB6	DnaJ heat shock protein family (Hsp40) member B6	0.2256	0.0121
31	ENSG00000156170	NDUFAF6	NADH:ubiquinone oxidoreductase complex assembly factor 6	0.8615	0.0009
32	ENSG00000179832	MROH1	maestro heat like repeat family member 1	1.0000	0.0000
33	ENSG00000164967	RPP25L	ribonuclease P/MRP subunit p25 like	0.1753	0.0010
34	ENSG00000160404	TOR2A	torsin family 2 member A	1.0000	0.0315
35	ENSG00000165698	SPACA9	sperm acrosome associated 9	1.0000	0.0000

Table 7.29 DTU genes between unstimulated RCC4 Cas9 GFP and WTAP KO 2H1 by DRIMseq and DEXseq

	Ensembl ID	Gene Symbol	Description	DRIMseq p_{adj}	DEXseq p_{adj}
36	ENSG00000185904	LINC00839	long intergenic non-protein coding RNA 839	0.9372	0.0314
37	ENSG00000183020	AP2A2	adaptor related protein complex 2 subunit alpha 2	1.0000	0.0015
38	ENSG00000006118	TMEM132A	transmembrane protein 132A	0.2784	0.0005
39	ENSG00000168000	BSCL2	BSCL2 lipid droplet biogenesis associated, seipin	0.3876	0.0000
40	ENSG00000177103	DSCAML1	DS cell adhesion molecule like 1	1.0000	0.0000
41	ENSG00000139174	PRICKLE1	prickle planar cell polarity protein 1	1.0000	0.0174
42	ENSG00000111331	OAS3	2'-5'-oligoadenylate synthetase 3	0.1761	0.0000
43	ENSG00000089157	RPLP0	ribosomal protein lateral stalk subunit P0	1.0000	0.0004
44	ENSG00000090615	GOLGA3	golgin A3	1.0000	0.0000
45	ENSG00000204977	TRIM13	tripartite motif containing 13	1.0000	0.0000
46	ENSG00000139926	FRMD6	FERM domain containing 6	0.2510	0.0316
47	ENSG00000140104	CLBA1	clathrin binding box of aftiphilin containing 1	1.0000	0.0046
48	ENSG00000173548	SNX33	sorting nexin 33	1.0000	0.0000
49	ENSG00000167005	NUDT21	nudix hydrolase 21	1.0000	0.0005
50	ENSG00000090863	GLG1	golgi glycoprotein 1	1.0000	0.0234
51	ENSG00000178773	CPNE7	copine 7	0.0684	0.0004
52	ENSG00000262165	C17orf114	chromosome 17 open reading frame 114	1.0000	0.0153
53	ENSG00000108511	HOXB6	homeobox B6	1.0000	0.0039
54	ENSG00000169727	GPS1	G protein pathway suppressor 1	0.8639	0.0000
55	ENSG00000105298	CACTIN	cactin, spliceosome C complex subunit	0.1330	0.0101
56	ENSG00000131351	HAUS8	HAUS augmin like complex subunit 8	0.2536	0.0308
57	ENSG00000121289	CEP89	centrosomal protein 89	0.1692	0.0000
58	ENSG00000123815	COQ8B	coenzyme Q8B	0.3651	0.0000
59	ENSG00000283103		novel transcript, antisense to ERVK3-1	0.9553	0.0102
60	ENSG00000101247	NDUFAF5	NADH:ubiquinone oxidoreductase complex assembly factor 5	0.9716	0.0095
61	ENSG00000125846	ZNF133	zinc finger protein 133	1.0000	0.0004
62	ENSG00000088970	KIZ	kizuna centrosomal protein	1.0000	0.0027
63	ENSG00000242372	EIF6	eukaryotic translation initiation factor 6	0.6786	0.0012
64	ENSG00000185658	BRWD1	bromodomain and WD repeat domain containing 1	1.0000	0.0003
65	ENSG00000079974	RABL2B	RAB, member of RAS oncogene family like 2B	0.7700	0.0001
66	ENSG00000147099	HDAC8	histone deacetylase 8	0.9021	0.0012
67	ENSG00000157600	TMEM164	transmembrane protein 164	1.0000	0.0002
68	ENSG00000203950	RTL8A	retrotransposon Gag like 8A	0.8386	0.0000
69	ENSG00000182195	LDOC1	LDOC1 regulator of NFKB signaling	0.3918	0.0112

Table 7.29 (cont.) DTU genes between unstimulated RCC4 Cas9 GFP and WTAP KO 2H1 by DRIMseq and DEXseq

	Ensembl ID	Gene Symbol	Description	DRIMseq p_{adj}	DEXseq p_{adj}
1	ENSG00000157911	PEX10	peroxisomal biogenesis factor 10	1.0000	0.0002
2	ENSG00000155363	MOV10	Mov10 RISC complex RNA helicase	1.0000	0.0395
3	ENSG00000143774	GUK1	guanylate kinase 1	0.9289	0.0009
4	ENSG00000143633	C1orf131	chromosome 1 open reading frame 131	1.0000	0.0002
5	ENSG00000145191	EIF2B5	eukaryotic translation initiation factor 2B subunit epsilon	1.0000	0.0175
6	ENSG00000131188	PRR7	proline rich 7, synaptic	0.6677	0.0914
7	ENSG00000234127	TRIM26	tripartite motif containing 26	1.0000	0.0763
8	ENSG00000124813	RUNX2	RUNX family transcription factor 2	1.0000	0.0000
9	ENSG00000112592	TBP	TATA-box binding protein	1.0000	0.0000
10	ENSG00000165046	LETM2	leucine zipper and EF-hand containing transmembrane protein 2	1.0000	0.0102
11	ENSG00000077782	FGFR1	fibroblast growth factor receptor 1	1.0000	0.0611
12	ENSG00000120217	CD274	CD274 molecule	0.6677	0.0172
13	ENSG00000168000	BSCL2	BSCL2 lipid droplet biogenesis associated, seipin	1.0000	0.0189
14	ENSG00000137731	FXYD2	FXYD domain containing ion transport regulator 2	0.9919	0.0875
15	ENSG00000196465	MYL6B	myosin light chain 6B	1.0000	0.0105
16	ENSG00000165948	IFI27L1	interferon alpha inducible protein 27 like 1	1.0000	0.0327
17	ENSG00000149925	ALDOA	aldolase, fructose-bisphosphate A	0.0079	0.0000
18	ENSG00000153774	CFDP1	craniofacial development protein 1	0.6247	0.0370
19	ENSG00000132522	GPS2	G protein pathway suppressor 2	1.0000	0.0322
20	ENSG00000184009	ACTG1	actin gamma 1	1.0000	0.0449
21	ENSG00000125730	C3	complement C3	1.0000	0.0020
22	ENSG00000214049	UCA1	urothelial cancer associated 1	1.0000	0.0300
23	ENSG00000153879	CEBPG	CCAAT enhancer binding protein gamma	1.0000	0.0287
24	ENSG00000088832	FKBP1A	FKBP prolyl isomerase 1A	1.0000	0.0190
25	ENSG00000160226	CFAP410	cilia and flagella associated protein 410	0.1065	0.0224
26	ENSG00000100336	APOL4	apolipoprotein L4	1.0000	0.0000
27	ENSG00000203950	RTL8A	retrotransposon Gag like 8A	0.8386	0.0003
28	ENSG00000214717	ZBED1	zinc finger BED-type containing 1	1.0000	0.0076
29	ENSG00000204438	GPANK1	G-patch domain and ankyrin repeats 1	0.1065	1.0000
30	ENSG00000166788	SAAL1	serum amyloid A like 1	0.0935	1.0000
31	ENSG00000120860	WASHC3	WASH complex subunit 3	0.0185	1.0000
32	ENSG00000039523	RIPOR1	RHO family interacting cell polarization regulator 1	0.0935	1.0000

Table 7.30 DTU genes between IFNγ + TNF treated RCC4 Cas9 GFP and WTAP KO 2H1 by DRIMseq and DEXseq

List of abbreviations

4E-BP	eIF4E binding protein
AKT	Protein kinase B
ALYREF	Aly/REF export factor
APA	Alternative cleavage and polyadenylation
APC	Antigen presenting cell
ApoB	Apolipoprotein B
ARE	AU-rich elements
ATP	Adenosine triphosphate
AUBP	AU-rich element binding protein
BAP1	BRCA1 associated protein 1
BCL2	B cell lymphoma 2
CAF	Cancer associated fibroblast
CAGE	Cap analysis gene expression
Cas9	CRISPR-associated protein 9
ccRCC	Clear cell renal cell carcinoma
CCL	Chemokine ligand 5
CDE	CAF-derived-exosomes
CDK	Cyclin dependent kinase
cDNA	Complementary DNA
CFI	Cleavage factor I
chRCC	Chromophobe renal cell carcinoma
CMML	Chronic myelomonocytic leukaemia
c-MYC	MYC proto-oncogene
CNV	Copy number variations
CPE	Cytoplasmic polyadenylation element

CPSF	Cleavage and polyadenylation specificity factor
CPT1A	Carnitine palmitoyltransferase 1A
CRISPR	clustered regularly interspaced short palindromic repeats
CSTF	Cleavage stimulation factor
CT	Computerised tomography scan
CTCF	CCCTC-binding factor
CTD	Carboxy terminal domain
CTLA4	Cytotoxic T-lymphocyte-associated protein 4
CTLs	Cytotoxic T lymphocytes
DC	Dendritic cell
DEG	Differentially expressed genes
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleoside triphosphate
DRS	Direct RNAseq
DTT	Dithiothreitol
DTU	Differential transcript usage
ECM	Extracellular matrix
EGF	Endothelial growth factor
ERK	Extracellular signal regulated kinase
FACS	Fluorescence-activated cell sorting
FBS	Foetal bovine serum
FDA	U.S. food and drug administration
FFPE	Formalin-fixed paraffin-embedded
GAP	GTPase activating protein
GFP	Green fluorescent protein
GTEx	Genotype-tissue expression

GLUT1	Glucose transporter 1
GO	Gene ontology
GO BP	Gene ontology biological processes
GO CC	Gene ontology cellular component
GO MF	Gene ontology molecular function
GSEA	Gene set enrichment analysis
HIF	Hypoxia inducible Factor
HLA	Human leukocyte antigen
hnRNPA1	Heterogenous nuclear ribonucleoprotein A1
HRE	Hypoxia response elements
ICI	Immune checkpoint inhibitor
IDO1	Indoleamine 2,3-dioxygenase
IFNγ	Interferon gamma
IGF	Insulin-like growth factor
IGV	Integrated genomics viewer
IHC	Immunohistochemical analysis
IL-10	Interleukin 10
ILK	Integrin-linked kinase
IRF1	Interferon regulatory factor 1
ISRE	Interferon stimulated response element
JAK	Janus kinase
KEGG	Kyoto encyclopedia of genes and genome
KIRC	Kidney renal clear cell carcinoma
LAG-3	Lymphocyte-activation gene 3
LDHA	Lactate dehydrogenase A
LOH	Loss of heterozygosity

MAPK	Mitogen activated protein kinase
MCT1	Monocarboxylate transporter 1
METTL3	Methyltransferase 3
METTL14	Methyltransferase 14
MHC	Major histocabability complex
MRI	Magnetic resonance imaging
mRNA	Messenger RNA
mRNP	messenger ribonucleoprotein
mTOR	Mechanistic target of rapamycin
mTORC1	mTOR complex 1
mTORC2	mTOR complex 2
ncRNA	non-coding RNA
NGS	Next generation sequencing
NLRC5	NLR family CARD domain containing 5
NMD	Nonsense-mediated decay
NSCLC	Non-small cell lung cancer
ONT	Oxford Nanopore Technologies
ORR	Objective response rate
PABPC1	Poly(A) binding cytoplasmic protein 1
PABPN1	Poly(A) binding nuclear protein 1
PacBio	Pacific Biosciences
PAP	Poly(A) polymerase
PARN	Poly(A)-specific ribonuclease
PAS	Polyadenylation signal
PBRM1	Polybromo 1
PBS	Phosphate-buffered saline

PCA	Principal component analysis
PCR	Polymerase chain reaction
PCS	PCR-cDNAseq
PD-1	Protein cell death protein 1
PDGF	platelet derived growth factor
PDK1	Phosphoinositide 3-kinase-1
PD-L1	Program death ligand 1
PH	Pleckstrin-homology
PHD	Prolyl hydroxylase
PI3K	Phosphatidylinositol-3-kinase
PIP2	Phosphatidylinositol-4,5-biphosphate
PIP3	Phosphatidylinositol-3,4,5-trisphosphate
poly(A)	Polyadenylation
pRCC	Papillary renal cell carcinoma
pre-mRNA	Precursor mRNA
p-TEFb	Positive transcription elongation factors b
qRT-PCR	quantitative reverse transcription PCR
RBP	RNA binding protein
RCC	Renal cell carcinoma
RIN	RNA Integrity Numbers
RNA	Ribonucleic acid
RNAP II	RNA polymerase II
rRNA	Ribosomal RNA
RT	Reverse transcription
RTK	Receptor tyrosine kinase
S6K1	p70S6 kinase 1

scRNAseq	Single-cell RNAseq
Ser	Serine
SETD2	SET domain containing 2
SHH	Sonic hedgehog
siRNA	Small interfering RNA
snRNA	Small nuclear RNA
snRNP	Small nuclear ribonucleoprotein
SR	Serine and arginine rich
SREBP	Sterol responsive element binding protein
SRSF1	Serine and arginine rich splicing factor 1
STAT1	Signal transducer and activator of transcription 1
TAM	Tumoure-associated macrophages
TAP	Transporter associated with antigen processing
TCA	Tricarboxylic acid
TCC	Transitional cell cancer
TCGA	The cancer genome atlas
TCGA	The cancer genome atlas
TCR	T cell receptor
TGF-β	Transforming growth factor beta
Th	T helper cells
Thr	Threonine
TIGIT	T cell immunoreceptor with Ig and ITIM domains
TIM-3	T cell immunoglobulin and mucin domain containing-3
TKI	Tyrosine kinase inhibitors
TLS	Tertiary lymphoid structures
TME	Tumour microenvironment

TNF	Tumour necrosis factor
TOX	Thymocyte selection associated high mobility group box
tracrRNA	transactivating CRISPR RNA
T_{reg}	Regulatory T cells
TTP	Tristetraprolin
U2AF1	U2 small nuclear RNA auxiliary factor 1
UTR	Untranslated region
VEGF	Vascular endothelial growth factor
VHL	von Hippel-Lindau
WTAP	Wilms tumor 1 associated proteins
YTHDF	YTH N6-methyladenosine RNA binding protein

References

Abascal, F. *et al.* (2020) 'Expanded encyclopaedias of DNA elements in the human and mouse genomes', *Nature*, 583(7818), pp. 699 - 710.

Agrawal, V. *et al.* (2019) 'Computer-aided discovery of small molecule inhibitors of thymocyte selection-associated high mobility group box protein (TOX) as potential therapeutics for cutaneous T-cell lymphomas', *Molecules*, 24(19), p. 3459.

Aird, D. *et al.* (2011) 'Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries', *Genome Biology*, 12(2), p. R18.

Aken, B.L. *et al.* (2016) 'The Ensembl gene annotation system', *Database (Oxford)*, 2016, p. baw093.

Albanese, A. *et al.* (2021) 'The role of hypoxia-inducible factor post-translational modifications in regulating its localisation, stability, and activity', *International Journal of Molecular Sciences*, 22(1), pp. 1–18.

Anders, S., Reyes, A. and Huber, W. (2012) 'Detecting differential usage of exons from RNA-seq data', *Genome Research*, 22(10), p. 2008.

Anderson, N.M. and Simon, M.C. (2020) 'The tumor microenvironment', *Current Biology*, 30(16), pp. R921–R925.

Angelini, C., Canditiis, D.D. and Feis, I.D. (2014) 'Computational approaches for isoform detection and estimation: Good and bad news', *BMC Bioinformatics*, 15(1), p.135.

Anreiter, I. *et al.* (2021) 'New Twists in Detecting mRNA Modification Dynamics', *Trends in Biotechnology*, 39(1), pp. 72–89.

Apanovich, N. *et al.* (2021) 'The Choice of Candidates in Survival Markers Based on Coordinated Gene Expression in Renal Cancer', *Frontiers in Oncology* , 11, p. 615787

Aran, D., Hu, Z. and Butte, A.J. (2017) 'xCell: Digitally portraying the tissue cellular heterogeneity landscape', *Genome Biology*, 18(1), p. 220.

Aran, D., Sirota, M. and Butte, A.J. (2015) 'Systematic pan-cancer analysis of tumour purity', *Nature Communications*, 6, p. 8971.

Atkins, M.B. and Tannir, N.M. (2018) 'Current and emerging therapies for first-line treatment of metastatic clear cell renal cell carcinoma', *Cancer Treatment Reviews*, 7(39), pp. 127–137.

Atzpodien, J. *et al.* (2002) ‘Combination chemotherapy with or without s.c. IL-2 and IFN-α: Results of a prospectively randomized trial of the cooperative advanced malignant melanoma chemoimmunotherapy group (ACIMM)’, *British Journal of Cancer*, 86(2), pp. 179 - 184.

Bakhtyar, N. *et al.* (2013) ‘Clear cell renal cell carcinoma induces fibroblast-mediated production of stromal periostin’, *European Journal of Cancer*, 49(16), pp. 3537–3546.

Bao, S. *et al.* (2021) ‘TGF-β1 Induces Immune Escape by Enhancing PD-1 and CTLA-4 Expression on T Lymphocytes in Hepatocellular Carcinoma’, *Frontiers in Oncology*, 11, p. 694145.

Barbieri, I. *et al.* (2017) ‘Promoter-bound METTL3 maintains myeloid leukaemia by m6A-dependent translation control’, *Nature*, 552(7683), pp. 126–131.

Barkal, A.A. *et al.* (2019) ‘CD24 signalling through macrophage Siglec-10 is a target for cancer immunotherapy’, *Nature*, 572(7769), pp. 392–396.

Barnet, M.B. *et al.* (2018) ‘Understanding Immune Tolerance of Cancer: Re-Purposing Insights from Fetal Allografts and Microbes’, *BioEssays*, 40(8), p. e1800050

Becker, C. *et al.* (2010) ‘mRNA and microRNA quality control for RT-qPCR analysis’, *Methods*, 50(4), pp. 237 - 243.

Behbahani, T.E. *et al.* (2011) ‘Tyrosine kinase expression profile in clear cell renal cell carcinoma’, *World Journal of Urology 2011 30:4*, 30(4), pp. 559–565.

Benita, Y. *et al.* (2009) ‘An integrative genomics approach identifies Hypoxia Inducible Factor-1 (HIF-1)-target genes that form the core response to hypoxia’, *Nucleic Acids Research*, 37(14), pp. 4587 - 4602.

Benjamini, Y. and Speed, T.P. (2012) ‘Summarizing and correcting the GC content bias in high-throughput sequencing’, *Nucleic Acids Research*, 40(10), p. e72.

Van den Berge, K. *et al.* (2017) ‘stageR: A general stage-wise method for controlling the gene-level false discovery rate in differential expression and differential transcript usage’, *Genome Biology*, 18(1), pp. 1–14.

Bhat, P. *et al.* (2017) ‘Interferon-γ derived from cytotoxic lymphocytes directly enhances their motility and cytotoxicity’, *Cell Death and Disease*, 8(6), p. e2836.

Bogard, N. *et al.* (2019) 'A Deep Neural Network for Predicting and Engineering Alternative Polyadenylation', *Cell*, 178(1), pp. 91 - 106.

Boguslawska, J. *et al.* (2019) 'TGF-β and microRNA interplay in genitourinary cancers', *Cells*, 3(12), p. 1619.

Boileau, E. *et al.* (2022) 'Full-Length Spatial Transcriptomics Reveals the Unexplored Isoform Diversity of the Myocardium Post-MI', *Frontiers in Genetics*, 13(July), pp. 1–11.

Bond, K.H. *et al.* (2021) 'The extracellular matrix environment of clear cell renal cell carcinoma determines cancer associated fibroblast growth', *Cancers*, 13(23), p. 5873.

Bossel Ben-Moshe, N. *et al.* (2018) 'mRNA-seq whole transcriptome profiling of fresh frozen versus archived fixed tissues', *BMC Genomics*, 19(1), p. 419.

Bourens, M. and Barrientos, A. (2017) ' A CMC1□knockout reveals translation□independent control of human mitochondrial complex IV biogenesis ', *EMBO reports*, 18(3), pp. 477 - 494.

Braun, D.A. *et al.* (2020) 'Interplay of somatic alterations and immune infiltration modulates response to PD-1 blockade in advanced clear cell renal cell carcinoma', *Nature Medicine*, 26(6), pp. 909-918.

Braun, D.A. *et al.* (2021) 'Progressive immune dysfunction with advancing disease stage in renal cell carcinoma', *Cancer Cell*, 39(5), pp. 632-648e8.

Brennan, T. *et al.* (2018) 'Generation of Luciferase-expressing Tumor Cell Lines', *Bio-Protocol*, 8(8), p.e2817.

Budimir, N. *et al.* (2022) 'Reversing T-cell Exhaustion in Cancer: Lessons Learned from PD-1/PD-L1 Immune Checkpoint Blockade', *Cancer Immunology Research*, 10(2), pp. 146-153.

Buen Abad Najar, C.F., Yosef, N. and Lareau, L.F. (2020) 'Coverage-dependent bias creates the appearance of binary splicing in single cells.', *eLife*, 9, pp. 1–23.

Buschauer, R. *et al.* (2020) 'The Ccr4-Not complex monitors the translating ribosome for codon optimality', *Science*, 368(6488), p. eaay6912.

Bushmanova, E. *et al.* (2019) 'RnaSPAdes: A de novo transcriptome assembler and its application to RNA-Seq data', *GigaScience*, 8(9), p. giz100.

Büttner, F.A. *et al.* (2022) 'A novel molecular signature identifies mixed subtypes in renal cell carcinoma with poor prognosis and independent response to immunotherapy', *Genome medicine*, 14(1), p. 105.

Caizzi, L. *et al.* (2021) 'Efficient RNA polymerase II pause release requires U2 snRNP function', *Molecular Cell*, 81(9), pp. 1920-1934e9.

de Campos, E.C.R. *et al.* (2013) 'Analysis of PTEN gene by fluorescent in situ hybridization in renal cell carcinoma', *Rev. Col. Bras. Cir.*, 40(6), pp. 471–475.

Capogrosso, P. *et al.* (2016) 'Follow-up After Treatment for Renal Cell Carcinoma: The Evidence Beyond the Guidelines', *European Urology Focus*., 1(3), pp. 272-281.

Carrillo Oesterreich, F. *et al.* (2016) 'Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II', *Cell*, 165(2), pp. 372-381.

Cartolano, M. *et al.* (2016) 'cDNA library enrichment of full length transcripts for SMRT long read sequencing', *PLoS ONE*, 11(6), p. e0157779.

Castro, F. *et al.* (2018) 'Interferon-gamma at the crossroads of tumor immune surveillance or evasion', *Frontiers in Immunology*., 9(837), p. 847.

Cerami, E. *et al.* (2012) 'The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data', *Cancer Discovery*, 2(5), pp. 401-404.

Chang, A. *et al.* (2022) 'Proteogenomic and clinical implications of unique recurrent splice variants in clear cell renal cell carcinoma.', *Journal of Clinical Oncology*, 40(6), pp. 354-362.

Chang, H. *et al.* (2014) 'TAIL-seq: Genome-wide determination of poly(A) tail length and 3' end modifications', *Molecular Cell*, 53(6), pp. 1044-1052.

Chang, J.H. and Tong, L. (2012) 'Mitochondrial poly(A) polymerase and polyadenylation', *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, 1819(9-10), pp. 992–997.

de Charette, M., Marabelle, A. and Houot, R. (2016) 'Turning tumour cells into antigen presenting cells: The next step to improve cancer immunotherapy?', *European Journal of Cancer*., 68, pp 134-147.

Chathoth, K.T. *et al.* (2014) 'A Splicing-Dependent Transcriptional Checkpoint Associated with Prespliceosome Formation', *Molecular Cell*, 53(5), pp. 779-790.

Chen, Meng *et al.* (2018) '3′ UTR lengthening as a novel mechanism in regulating cellular senescence', *Genome Research*, 28(3), pp. 285-294.

Chen, Mengnuo *et al.* (2018) 'RNA N6-methyladenosine methyltransferase-like 3 promotes liver cancer progression through YTHDF2-dependent posttranscriptional silencing of SOCS2', *Hepatology*, 67(6), pp. 2254-2270.

Chen, S. *et al.* (2021) ‘A Novel m6A Gene Signature Associated With Regulatory Immune Function for Prognosis Prediction in Clear-Cell Renal Cell Carcinoma’, *Frontiers in Cell and Developmental Biology*, 8, p. 616972.

Chen, W. and Moore, M.J. (2015) ‘Spliceosomes’, *Current Biology*, 25(5), pp. R181-R183

Chen, X.-Y., Zhang, J. and Zhu, J.-S. (2019) ‘The role of m6A RNA methylation in human cancer’, *Molecular Cancer*, 18(1), p. 103.

Chen, X. *et al.* (2019) 'Computational models for lncRNA function prediction and functional similarity calculation', *Briefings in Functional Genomics*, 18(1), pp. 58-82.

Cheng, J. *et al.* (2017) ‘Cis-regulatory elements explain most of the mRNA stability variation across genes in yeast’, *RNA*, 23(11), pp. 1648 - 1659.

Cho, S. *et al.* (2011) ‘Interaction between the RNA binding domains of Ser-Arg splicing factor 1 and U1-70K snRNP protein determines early spliceosome assembly’, *Proceedings of the National Academy of Sciences of the United States of America*, 108(20), pp. 8233-8238.

Cho, S. *et al.* (2021) 'mTORC1 promotes cell growth via m^6A-dependent mRNA degradation', *Molecular Cell*, 81, pp. 2064 - 2075

Clark, D.J. *et al.* (2019) ‘Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma Resource Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma’, *Cell*, 179, pp. 964-983.

Coelho, M.A. *et al.* (2017) ‘Oncogenic RAS Signaling Promotes Tumor Immunoresistance by Stabilizing PD-L1 mRNA’, *Immunity*, 47(6), pp. 1083-1099.

Cohen, N. *et al.* (2017) ‘Fibroblasts drive an immunosuppressive and growth-promoting microenvironment in breast cancer via secretion of Chitinase 3-like 1’, *Oncogene*, 36(31), pp. 4457-4468.

Collart, M.A. (2016) ‘The Ccr4-Not complex is a key regulator of eukaryotic gene expression’, *Wiley Interdisciplinary Reviews: RNA*, pp. 438–454.

Conesa, A. *et al.* (2016) ‘A survey of best practices for RNA-seq data analysis’, *Genome Biology*, 17(1), p. 13.

Corbett, A.H. (2018) ‘Post-transcriptional regulation of gene expression and human disease’, *Current Opinion in Cell Biology*, pp. 96–104.

Corchete, L.A. *et al.* (2020) 'Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis', *Scientific Reports*, 10(1), p. 19737.

Correa, A.F. *et al.* (2020) 'Overall tumor genomic instability: an important predictor of recurrence-free survival in patients with localized clear cell renal cell carcinoma', *Cancer Biology and Therapy*, 21(5), pp. 424-431.

Costa, A. *et al.* (2018) 'Fibroblast Heterogeneity and Immunosuppressive Environment in Human Breast Cancer', *Cancer Cell*, 33(3), pp. 463-479.

De Coster, W. *et al.* (2018) 'NanoPack: visualizing and processing long-read sequencing data', *Bioinformatics*, 34(15), pp. 2666–2669.

Courtney, K.D. *et al.* (2018) 'Isotope Tracing of Human Clear Cell Renal Cell Carcinomas Demonstrates Suppressed Glucose Oxidation In Vivo', *Cell Metabolism*, 28(5), pp. 793-800.

Creighton, C.J. *et al.* (2013) 'Comprehensive molecular characterization of clear cell renal cell carcinoma', *Nature 2013 499:7456*, 499(7456), pp. 43–49.

Crick, F. (1970) 'Central dogma of molecular biology', *Nature*, 227(5258), pp. 561-563.

Cui, X. *et al.* (2016) 'A novel algorithm for calling mRNA m6A peaks by modeling biological variances in MeRIP-seq data', *Bioinformatics*, 32(12), pp. i378–i385.

Cui, X. *et al.* (2018) 'MeTDiff: A Novel Differential RNA Methylation Analysis for MeRIP-Seq Data', *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(2), pp. 526–534.

Cullot, G. *et al.* (2019) 'CRISPR-Cas9 genome editing induces megabase-scale chromosomal truncations', *Nature Communications*, 10(1), p. 1136.

Cunningham, F. *et al.* (2022) 'Ensembl 2022', *Nucleic Acids Research*, 50(D1), pp. D988-D995.

Dabestani, S. *et al.* (2019) 'Long-term Outcomes of Follow-up for Initially Localised Clear Cell Renal Cell Carcinoma: RECUR Database Analysis', *European Urology Focus*, 5(5), pp. 857-867.

Dagogo-Jack, I. and Shaw, A.T. (2018) 'Tumour heterogeneity and resistance to cancer therapies', *Nature Reviews Clinical Oncology*, 15(2), pp. 81–94.

Damjanovic, S.S. *et al.* (2016) 'Tuberous sclerosis complex protein 1 expression is affected by VHL Gene alterations and HIF-1α production in sporadic clear-cell renal cell carcinoma', *Experimental and Molecular Pathology*, 101(3), pp. 323–331.

Dana, A. and Tuller, T. (2012) 'Determinants of Translation Elongation Speed and Ribosomal Profiling Biases in Mouse Embryonic Stem Cells', *PLoS Computational Biology*, 8(11), p. e1002755.

David, C.J. *et al.* (2010) 'HnRNP proteins controlled by c-Myc deregulate pyruvate kinase mRNA splicing in cancer', *Nature*, 463(7279), pp. 364-368.

Davidson, L., Muniz, L. and West, S. (2014) '3' end formation of pre-mRNA and phosphorylation of Ser2 on the RNA polymerase II CTD are reciprocally coupled in human cells', *Genes and Development*, 28(4), pp. 342-356.

Davies, J., Denyer, T. and Hadfield, J. (2016) 'Bioanalyzer chips can be used interchangeably for many analyses of DNA or RNA', *BioTechniques*, 60(4), pp. 197-199.

Day, D.S. *et al.* (2016) 'Comprehensive analysis of promoter-proximal RNA polymerase II pausing across mammalian cell types', *Genome Biology*, 17(1), pp. 1-17.

Dery, K.J. *et al.* (2014) 'IRF-1 regulates alternative mRNA splicing of carcinoembryonic antigen-related cell adhesion molecule 1 (CEACAM1) in breast epithelial cells generating an immunoreceptor tyrosine-based inhibition motif (ITIM) containing isoform', *Molecular Cancer*, 13(1), pp. 1-20.

Deschamps-Francoeur, G., Simoneau, J. and Scott, M.S. (2020) 'Handling multi-mapped reads in RNA-seq', *Computational and Structural Biotechnology Journal*, 18, pp. 1569–1576.

Desjardins, P. and Conklin, D. (2010) 'NanoDrop microvolume quantitation of nucleic acids', *Journal of Visualized Experiments*, 45(2526), p. e2565.

Dhatchinamoorthy, K., Colbert, J.D. and Rock, K.L. (2021) 'Cancer Immune Evasion Through Loss of MHC Class I Antigen Presentation', *Frontiers in Immunology*, 12, p. 636568.

Djebali, S. *et al.* (2012) 'Landscape of transcription in human cells', *Nature*, 489(7414), pp. 101–108.

Dominissini, D. *et al.* (2012) 'Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq', *Nature*, 485(7397), pp. 201–206.

Dominissini, D. *et al.* (2013) 'Transcriptome-wide mapping of N6-methyladenosine by m6A-seq based on immunocapturing and massively parallel sequencing', *Nature Protocols*, 8(1), pp. 176–189.

Downes, N.L. *et al.* (2018) 'Differential but Complementary HIF1α and HIF2α Transcriptional Regulation', *Molecular Therapy*, 26(7), pp. 1735–1745.

Drexler, H.L. *et al.* (2021) 'Revealing nascent RNA processing dynamics with nano-COP', *Nature Protocols*, 16(3), pp. 1343-1375.

Du, W. *et al.* (2017) 'HIF drives lipid deposition and cancer in ccRCC via repression of fatty acid metabolism', *Nature Communications*, 8(1), pp. 786-799.

Eisen, T.J. *et al.* (2020) 'The Dynamics of Cytoplasmic mRNA Metabolism', *Molecular Cell*, 77(4).

Elkon, R., Ugalde, A.P. and Agami, R. (2013) 'Alternative cleavage and polyadenylation: Extent, regulation and function', *Nature Reviews Genetics*, 14(7), pp. 496–506.

Elyada, E. *et al.* (2019) 'Cross-species single-cell analysis of pancreatic ductal adenocarcinoma reveals antigen-presenting cancer-associated fibroblasts', *Cancer Discovery*, 9(8), pp. 1102–1123.

Escors, D. *et al.* (2018) 'The intracellular signalosome of PD-L1 in cancer cells', *Signal Transduction and Targeted Therapy*, 3(1), p. 26.

Escudier, B. *et al.* (2019) 'Renal cell carcinoma: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up†', *Annals of Oncology*, 30(5), pp. 706–720.

Fabian, M.R. *et al.* (2013) 'Structural basis for the recruitment of the human CCR4-NOT deadenylase complex by tristetraprolin', *Nature Structural and Molecular Biology*, 20(6), pp. 735-739.

Fontes-Sousa, M. *et al.* (2022) 'Reviewing Treatment Options for Advanced Renal Cell Carcinoma: Is There Still a Place for Tyrosine Kinase Inhibitor (TKI) Monotherapy?', *Advances in Therapy*, 39(3), pp. 1107–1125.

Foster, D.S. *et al.* (2018) 'The evolving relationship of wound healing and tumor stroma', *JCI Insight*, 3(18).

Frankish, A. *et al.* (2021) 'GENCODE 2021', *Nucleic Acids Research*, 49(D1), pp. D916-

D923.

Fridman, W.H. *et al.* (2017) 'The immune contexture in cancer prognosis and treatment', *Nature Reviews Clinical Oncology*, 14(12), pp. 717–734.

Fu, S. *et al.* (2018) 'IDP-denovo: De novo transcriptome assembly and isoform annotation by hybrid sequencing', *Bioinformatics*, 32(13), pp. 2168–2176.

Fukaya, T., Lim, B. and Levine, M. (2017) 'Rapid Rates of Pol II Elongation in the Drosophila Embryo', *Current Biology*, 27(9), pp. 1387-1391.

Fukusumi, Y., Naruse, C. and Asano, M. (2008) 'Wtap is required for differentiation of endoderm and mesoderm in the mouse embryo', *Developmental Dynamics*, 237(3), pp. 618-629.

Gao, R. *et al.* (2021) 'm6A Modification: A Double-Edged Sword in Tumor Development.', *Frontiers in oncology*, 11, p. 679367.

Garalde, D.R. *et al.* (2018) 'Highly parallel direct RN A sequencing on an array of nanopores', *Nature Methods*, 15(3), pp. 201–206.

Garcia-Elias, A. *et al.* (2017) 'Defining quantification methods and optimizing protocols for microarray hybridization of circulating microRNAs', *Scientific Reports*, 7(1), pp. 1-14.

Ge, Y. *et al.* (2021) 'Degradation of WTAP blocks antiviral responses by reducing the m 6 A levels of IRF3 and IFNAR1 mRNA', *EMBO reports*, 22(11), p. e52101.

Gehring, N.H. and Roignant, J.Y. (2021) 'Anything but Ordinary – Emerging Splicing Mechanisms in Eukaryotic Gene Regulation', *Trends in Genetics*, 37(4), pp. 355–372.

Geula, S. *et al.* (2015) 'm6A mRNA methylation facilitates resolution of naïve pluripotency toward differentiation', *Science*, 347(6225), pp. 1002–1006.

Ghatalia, P. *et al.* (2019) 'Prognostic impact of immune gene expression signature and tumor infiltrating immune cells in localized clear cell renal cell carcinoma', *Journal for ImmunoTherapy of Cancer*, 7(1), pp. 1-12.

Giraldo, N.A. *et al.* (2017) 'Tumor-infiltrating and peripheral blood T-cell immunophenotypes predict early relapse in localized clear cell renal cell carcinoma', *Clinical Cancer Research*, 23(15), pp. 4416-4428.

Giraldo, N.A. *et al.* (2019) 'The clinical role of the TME in solid cancer', *British Journal of Cancer*, 120(1), pp. 45–53.

Gleeson, J. *et al.* (2022) 'Accurate expression quantification from nanopore direct RNA sequencing with NanoCount', *Nucleic Acids Research*, 50(4), pp. e19-e19.

Glinos, D.A. *et al.* (2022) 'Transcriptome variation in human tissues revealed by long-read sequencing', *Nature*, 608(7922), pp. 353–359.

Goering, R. *et al.* (2021) 'LABRAT reveals association of alternative polyadenylation with transcript localization, RNA binding protein expression, transcription speed, and cancer survival', *BMC Genomics*, 22(1), p.476.

Goodwin, S., McPherson, J.D. and McCombie, W.R. (2016) 'Coming of age: ten years of next-generation sequencing technologies', *Nature Reviews Genetics*, 17(6), pp. 333–351.

Gossage, L., Eisen, T. and Maher, E.R. (2014) 'VHL, the story of a tumour suppressor gene', *Nature Reviews Cancer*, 15(1), pp. 55–64.

Gracia Villacampa, E. *et al.* (2021) 'Genome-wide spatial expression profiling in formalin-fixed tissues', *Cell Genomics*, 1(3), p. 100065.

Gruber, A.J. *et al.* (2016) 'A comprehensive analysis of 3' end sequencing data sets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation', *Genome Research*, 26(8). pp. 1145-1159.

Gruber, A.R. *et al.* (2014) 'Global 3' UTR shortening has a limited effect on protein abundance in proliferating T cells', *Nature communications*, 5(1), p. 5465.

Grünberger, F., Ferreira-Cerca, S. and Grohmann, D. (2022) 'Nanopore sequencing of RNA and cDNA molecules in Escherichia coli', *RNA*, 28(3), pp.400-417.

Gu, B., Eick, D. and Bensaude, O. (2013) 'CTD serine-2 plays a critical role in splicing and termination factor recruitment to RNA polymerase II in vivo', *Nucleic Acids Research*, 41(3), pp. 1591-1603.

Guo, H. *et al.* (2015) 'The PI3K/AKT Pathway and Renal Cell Carcinoma', *Journal of Genetics and Genomics*, 42(7), pp. 343–353.

Guo, J. *et al.* (2016) 'pVHL suppresses kinase activity of Akt in a proline-hydroxylation–dependentmanner', *Science*, 353(6302), p. 929.

Haas, B.J. *et al.* (2013) ‘De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis’, *Nature Protocols*, 8(8), pp. 1494-1512.

Haile, S.T. *et al.* (2013) ‘Soluble CD80 Restores T Cell Activation and Overcomes Tumor Cell Programmed Death Ligand 1–Mediated Immune Suppression’, *The Journal of Immunology*, 191(5), pp. 2829-2836.

Hammers, H.J. *et al.* (2017) ‘Safety and efficacy of nivolumab in combination with ipilimumab in metastatic renal cell carcinoma: The checkmate 016 study’, *Journal of Clinical Oncology*, 35(34), p. 3851.

Han, B. *et al.* (2021) ‘The clinical implication of soluble PD-L1 (sPD-L1) in patients with breast cancer and its biological function in regulating the function of T lymphocyte’, *Cancer Immunology, Immunotherapy*, 70(10), pp. 2893–2909.

Hanahan, D. and Weinberg, R.A. (2011) ‘Hallmarks of cancer: The next generation’, *Cell*, 144(5), pp. 646–674.

Heard, J.J. *et al.* (2018) ‘An oncogenic mutant of RHEB, RHEB Y35N, exhibits an altered interaction with BRAF resulting in cancer transformation’, *BMC Cancer*, 18(1), pp. 1–11.

Helm, M. and Motorin, Y. (2017) ‘Detecting RNA modifications in the epitranscriptome: Predict and validate’, *Nature Reviews Genetics*, 18(5), pp. 275–291.

Hendra, C. *et al.* (2022) 'Detection of m6A from direct RNA sequencing using a multiple instance learning framework', *Nature Methods*, 19, pp. 1590-1598.

Herbert, Z.T. *et al.* (2018) ‘Cross-site comparison of ribosomal depletion kits for Illumina RNAseq library construction’, *BMC Genomics*, 19, pp. 1-10.

Hess, M.E. *et al.* (2013) ‘The fat mass and obesity associated gene (Fto) regulates activity of the dopaminergic midbrain circuitry.’, *Nature neuroscience*, 16(8), pp. 1042–8.

Hoefflin, R. *et al.* (2020) ‘HIF-1α and HIF-2α differently regulate tumour development and inflammation of clear cell renal cell carcinoma in mice’, *Nature Communications*, 11(1), p. 4111.

Höijer, I. *et al.* (2020) ‘Amplification-free long-read sequencing reveals unforeseen CRISPR-Cas9 off-target activity’, *Genome Biology*, 21(1), pp. 1-19.

Holmqvist, I. *et al.* (2021) ‘FLAME: long-read bioinformatics tool for comprehensive

spliceome characterization', *RNA*, 27(10), pp. 1127-1139.

Horiuchi, K. *et al.* (2006) 'Wilms' tumor 1-associating protein regulates G2/M transition through stabilization of cyclin A2 mRNA', *Proceedings of the National Academy of Sciences of the United States of America*, 103(46), pp. 17278–17283.

Hornyák, L. *et al.* (2018) 'The Role of Indoleamine-2,3-Dioxygenase in Cancer Development, Diagnostics, and Therapy.', *Frontiers in immunology*, 9, p. 151.

Howard, J.M. *et al.* (2018) 'HNRNPA1 promotes recognition of splice site decoys by U2AF2 in vivo', *Genome Research*, 28(5), pp. 689-698.

Hsin, J.P. and Manley, J.L. (2012) 'The RNA polymerase II CTD coordinates transcription and RNA processing', *Genes and Development*, 26(19), pp. 2119–2137.

Hsu, P.J. *et al.* (2017) 'Ythdc2 is an N6 -methyladenosine binding protein that regulates mammalian spermatogenesis', *Cell Research*, 27(9), pp. 1115–1127.

Hu, J. *et al.* (2020) 'Single-Cell Transcriptome Analysis Reveals Intratumoral Heterogeneity in ccRCC, which Results in Different Clinical Outcomes', *Molecular Therapy*, 28(7), pp. 1658-1672.

Huang, D. *et al.* (2010) 'Sunitinib acts primarily on tumor endothelium rather than tumor cells to inhibit the growth of renal cell carcinoma', *Cancer Research*, 70(3), pp. 1053-1062.

Huang, H. *et al.* (2018) 'Recognition of RNA N 6 -methyladenosine by IGF2BP proteins enhances mRNA stability and translation', *Nature Cell Biology*, 20(3), pp. 285–295.

Huang, J.J. and Hsieh, J.J. (2020) 'The Therapeutic Landscape of Renal Cell Carcinoma: From the Dark Age to the Golden Age', *Seminars in Nephrology*, 70(3), pp. 28–41.

Hudes, G.R. (2009) 'Targeting mTOR in renal cell carcinoma', *Cancer*, 115(S10), pp. 2313–2320.

Ibrahim, F. *et al.* (2021) 'TERA-Seq: True end-to-end sequencing of native RNA molecules for transcriptome characterization', *Nucleic Acids Research*, 49(20), pp. e115-e115.

Im, S.J. *et al.* (2016) 'Defining CD8+ T cells that provide the proliferative burst after PD-1 therapy', *Nature*, 537(7620), pp.417-421.

Ivan, M. *et al.* (2001) 'HIFα targeted for VHL-mediated destruction by proline hydroxylation: Implications for O2 sensing', *Science*, 292(5516), pp. 464–468.

Jain, M. *et al.* (2022) ‘Advances in nanopore direct RNA sequencing’, *Nature Methods*, 19(10), pp. 1160–1164.

Janssen, M.W.W. *et al.* (2018) ‘Survival outcomes in patients with large (7cm) clear cell renal cell carcinomas treated with nephron-sparing surgery versus radical nephrectomy: Results of a multicenter cohort with long-term follow-up’, *PLoS ONE*, 13(5), p. e0196427.

Jenjaroenpun, P. *et al.* (2021) ‘Decoding the epitranscriptional landscape from native RNA sequences’, *Nucleic acids research*, 49(2), p. e7.

Jennings, L.J. *et al.* (2017) ‘Guidelines for Validation of Next-Generation Sequencing–Based Oncology Panels’, *The Journal of Molecular Diagnostics*, 19(3), pp. 341–365.

Jia, G. *et al.* (2011) ‘N6-Methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO’, *Nature Chemical Biology*, 7(12), pp. 885–887.

Jia, Y.Y., Yu, Y. and Li, H.J. (2021) ‘POSTN promotes proliferation and epithelial mesenchymal transition in renal cell carcinoma through ILK/AKT/mTOR pathway’, *Journal of Cancer*, 12(14), pp. 4183–4195.

Jiang, F. *et al.* (2019) ‘Long-read direct RNA sequencing by 5’-Cap capturing reveals the impact of Piwi on the widespread exonization of transposable elements in locusts’, *RNA Biology*, 16(7), pp. 950-959.

Jiang, J. *et al.* (2017) ‘Polypyrimidine Tract-Binding Protein 1 promotes proliferation, migration and invasion in clear-cell renal cell carcinoma by regulating alternative splicing of PKM’, *American Journal of Cancer Research*, 7(2), p. 245.

Jonasch, E., Walker, C.L. and Rathmell, W.K. (2020) ‘Clear cell renal cell carcinoma ontogeny and mechanisms of lethality’, *Nature Reviews Nephrology*, 17(4), pp. 245–261.

De Jong, V.M. *et al.* (2016) ‘Variation in the CTLA4 3′UTR has phenotypic consequences for autoreactive T cells and associates with genetic risk for type 1 diabetes’, *Genes and Immunity*, 17(1), pp. 75–78.

Jonkers, I., Kwak, H. and Lis, J.T. (2014) ‘Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons’, *eLife*, 2014(3), p. e02407.

Josephs, S.F. *et al.* (2018) ‘Unleashing endogenous TNF-alpha as a cancer immunotherapeutic.’, *Journal of translational medicine*, 16(1), p. 242.

Joshi, D. *et al.* (2021) ‘QAlign: Aligning nanopore reads accurately using current-level modeling’, *Bioinformatics*, 37(5), p. 625-633.

Juneja, V.R. *et al.* (2017) ‘PD-L1 on tumor cells is sufficient for immune evasion in immunogenic tumors and inhibits CD8 T cell cytotoxicity’, *Journal of Experimental Medicine*, 214(4), pp. 895-904.

Kadomoto, S., Izumi, K., Hiratsuka, K., *et al.* (2020) ‘Tumor-associated macrophages induce migration of renal cell carcinoma cells via activation of the CCL20-CCR6 axis’, *Cancers*, 12(1), p.89.

Kadomoto, S., Izumi, K. and Mizokami, A. (2020) ‘The CCL20-CCR6 axis in cancer progression’, *International Journal of Molecular Sciences*, 21(15), pp. 1–18.

Kahles, A. *et al.* (2018) ‘Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients’, *Cancer Cell*, 34(2), pp. 211-224.

Kalbasi, A. and Ribas, A. (2020) ‘Tumour-intrinsic resistance to immune checkpoint blockade’, *Nature Reviews Immunology*, 20(1), pp. 25–39.

Kalia, V. and Sarkar, S. (2018) ‘Regulation of Effector and Memory CD8 T Cell Differentiation by IL-2—A Balancing Act’, *Frontiers in Immunology*, p. 2987.

Kang, Y. *et al.* (2017) 'CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features', *Nucleic Acids Research*, 45(W1), pp. W12-W16.

Kap, M. *et al.* (2014) ‘Fit for purpose frozen tissue collections by RNA integrity number-based quality control assurance at the Erasmus MC tissue bank’, *Biopreservation and Biobanking*, 12(2), pp. 81-90.

Kartikasari, A.E.R. *et al.* (2021) ‘Tumor-Induced Inflammatory Cytokines and the Emerging Diagnostic Devices for Cancer Detection and Prognosis.’, *Frontiers in oncology*, 11, p. 692142.

Kataoka, K. *et al.* (2016) ‘Aberrant PD-L1 expression through 3′-UTR disruption in multiple cancers’, *Nature*, 534(7607), pp. 402–406.

Keller, A. *et al.* (2022) ‘MiRNATissueAtlas2: An update to the human miRNA tissue atlas’, *Nucleic Acids Research*, 50(D1), pp. D211-D221.

Khan, O. *et al.* (2019) ‘TOX transcriptionally and epigenetically programs CD8+ T cell exhaustion’, *Nature*, 571(7764), pp. 211-218.

Kim, K. *et al.* (2020) 'Single-cell transcriptome analysis reveals TOX as a promoting factor for T cell exhaustion and a predictor for anti-PD-1 responses in human cancer', *Genome Medicine*, 12(1), pp. 1-16.

Kim, M.H. *et al.* (2021) 'PD-1 expression and its correlation with prognosis in clear cell renal cell carcinoma', *In Vivo*, 35(3), pp. 1549-1553.

Kim, S.H. *et al.* (2021) 'A Real-World, Population-Based Retrospective Analysis of Therapeutic Survival for Recurrent Localized Renal Cell Carcinoma After Nephrectomy', *Frontiers in Oncology*, 11, p. 693831.

Knuckles, P. *et al.* (2018) 'Zc3h13/Flacc is required for adenosine methylation by bridging the mRNA-binding factor Rbm15/Spenito to the m6A machinery component Wtap/Fl(2)d.', *Genes & development*, 32(5–6), pp. 415–429.

Korotkevich, G. and Sukhov, V. (2016) 'Fast gene set enrichment analysis', *bioRxiv*, p.060012

Kosicki, M., Tomberg, K. and Bradley, A. (2018) 'Repair of double-strand breaks induced by CRISPR–Cas9 leads to large deletions and complex rearrangements', *Nature Biotechnology*, 36(8), pp. 765-771.

Kovacs, G. *et al.* (1997) 'The Heidelberg classification of renal cell tumours', *The Journal of Pathology*, 183(2), pp. 131–133.

Kovaka, S. *et al.* (2019) 'Transcriptome assembly from long-read RNA-seq alignments with StringTie2', *Genome Biology*, 20(1), pp. 1-3.

Krause, M. *et al.* (2019) 'TailFindR: Alignment-free poly(A) length measurement for Oxford Nanopore RNA and DNA sequencing', *RNA*, 25(10), pp.1229-1241.

Krawczyk, P.S. *et al.* (2022) 'SARS-CoV-2 mRNA vaccine is re-adenylated in vivo, enhancing antigen production and immune response', *bioRxiv*, pp.2022-12

Krzywinski, M., Altman, N. and Blainey, P. (2014) 'Points of Significance: Nested designs', *Nature Methods*, 11(10), p. 879.

Kühn, U. *et al.* (2009) 'Poly(A) tail length is controlled by the nuclear Poly(A)-binding protein regulating the interaction between Poly(A) polymerase and the cleavage and polyadenylation specificity factor', *Journal of Biological Chemistry*, 284(34), p. 22803.

Kumar, B. V., Connors, T.J. and Farber, D.L. (2018) 'Human T Cell Development,

Localization, and Function throughout Life', *Immunity*, 48(2), pp. 202–213.

Kwak, Y. *et al.* (2022) 'Dynamic and widespread control of poly(A) tail length during macrophage activation', *Rna*, 28(7), pp. 947–971.

Kwon, B. *et al.* (2022) 'Enhancers regulate 3' end processing activity to control expression of alternative 3'UTR isoforms', *Nature communications*, 13(1), p. 2709.

de la Rubia, I. *et al.* (2022) 'RATTLE: reference-free reconstruction and quantification of transcriptomes from Nanopore sequencing', *Genome Biology*, 23(1), pp. 1–21.

Laha, D. *et al.* (2021) 'The Role of Tumor Necrosis Factor in Manipulating the Immunological Response of Tumor Microenvironment.', *Frontiers in immunology*, 12, p. 656908.

Lalmahomed, Z.S. *et al.* (2017) 'Multicenter fresh frozen tissue sampling in colorectal cancer: does the quality meet the standards for state of the art biomarker research?', *Cell and Tissue Banking*, 18(3), pp. 425-431.

Lasman, L. *et al.* (2020) 'Context-dependent functional compensation between Ythdf m 6 A reader proteins', *Genes & Development*, 34(19–20), pp. 1373–1391.

Lawrence, M. *et al.* (2013) 'Software for Computing and Annotating Genomic Ranges', *PLOS Computational Biology*, 9(8), p. e1003118.

Lawson, N.L. *et al.* (2020) 'Mapping the binding sites of antibodies utilized in programmed cell death ligand-1 predictive immunohistochemical assays for use with immuno-oncology therapies', *Modern Pathology*, 33(4), pp. 518-530.

Lee, A.Y.S. *et al.* (2013) 'CC Chemokine Ligand 20 and Its Cognate Receptor CCR6 in Mucosal T Cell Immunology and Inflammatory Bowel Disease: Odd Couple or Axis of Evil?', *Frontiers in immunology*, 4, p. 194.

Leger, A. *et al.* (2021) 'RNA modifications detection by comparative Nanopore direct RNA sequencing.', *Nature communications*, 12(1), p. 7198.

Leibovich, B.C. *et al.* (2003) 'Prediction of progression after radical nephrectomy for patients with clear cell renal cell carcinoma', *Cancer*, 97(7), pp. 1663-1671.

Lence, T. *et al.* (2016) 'm6A modulates neuronal functions and sex determination in Drosophila.', *Nature*, 540(7632), pp. 242–247.

Lenz, M. *et al.* (2016) 'Principal components analysis and the reported low intrinsic dimensionality of gene expression microarray data', *Scientific Reports*, 6(1), pp. 1-11.

Lesbirel, S. *et al.* (2018) 'The m6A-methylase complex recruits TREX and regulates mRNA export.', *Scientific reports*, 8(1), p. 13827.

Leung, M.K.K., Delong, A. and Frey, B.J. (2018) 'Inference of the human polyadenylation code', *Bioinformatics*, 34(17), pp. 2889-2898.

Leung, S.K. *et al.* (2021) 'Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing', *Cell Reports*, 37(7), p. 110022.

Li, F. *et al.* (2021) 'The association between CD8+ tumor-infiltrating lymphocytes and the clinical outcome of cancer immunotherapy: A systematic review and meta-analysis.', *EClinicalMedicine*, 41, p. 101134.

Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25(16), pp. 2078–2079.

Li, H. (2018) 'Minimap2: pairwise alignment for nucleotide sequences', *Bioinformatics*, 34(18), pp. 3094–3100.

Li, N. *et al.* (2018) 'Cross-talk between TNF-α and IFN-γ signaling in induction of B7-H1 expression in hepatocellular carcinoma cells', *Cancer Immunology, Immunotherapy*, 67(2), pp. 271–283.

Li, X. *et al.* (2015) 'Lowering the quantification limit of the QubitTM RNA HS Assay using RNA spike-in', *BMC Molecular Biology*, 16(1), pp. 1-7.

Li, X. *et al.* (2019) 'Infiltration of CD8 + T cells into tumor cell clusters in triple-negative breast cancer', *Proceedings of the National Academy of Sciences of the United States of America*, 116(9), pp. 3678-3687.

Li, X. *et al.* (2021) 'Discovery of functional alternatively spliced PKM transcripts in human cancers', *Cancers*, 13(2), p. 348.

Li, X., Xiong, X. and Yi, C. (2017) 'Epitranscriptome sequencing technologies: decoding RNA modifications', *Nature Methods*, 14(1), pp. 23–31.

Lianoglou, S. *et al.* (2013) 'Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression', *Genes and Development*, 27(21), pp. 2380-2396.

Liao, Y., Smyth, G.K. and Shi, W. (2014) 'featureCounts: an efficient general purpose program for assigning sequence reads to genomic features', *Bioinformatics*, 30(7), pp. 923–

930.

Libreros, S., Garcia-Areas, R. and Iragavarapu-Charyulu, V. (2013) 'CHI3L1 plays a role in cancer through enhanced production of pro-inflammatory/pro-tumorigenic and angiogenic factors', *Immunologic Research*, 57, pp. 99-105.

Lima, S.A. *et al.* (2017) 'Short poly(A) tails are a conserved feature of highly expressed genes', *Nature Structural and Molecular Biology*, 24(12), pp. 1057-1063.

Lin, S. *et al.* (2016) 'The m 6 A Methyltransferase METTL3 Promotes Translation in Human Cancer Cells', *Molecular Cell*, 62(3), pp. 335–345.

Linder, B. *et al.* (2015) 'Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome', *Nature Methods*, 12(8), pp. 767–772.

Liu, B. *et al.* (2019) 'DeSALT: Fast and accurate long transcriptomic read alignment with De Bruijn graph-based index', *Genome Biology*, 20(1), pp.1-14.

Liu, H. *et al.* (2019) 'Accurate detection of m6A RNA modifications in native RNA sequences', *Nature Communications*, 10(1), p. 4079.

Liu, J. *et al.* (2014) 'A METTL3–METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation', *Nature Chemical Biology*, 10(2), pp. 93–95.

Liu, M., Li, S. and Li, M.O. (2018) 'TGF-β Control of Adaptive Immune Tolerance: A Break From Treg Cells', *BioEssays*, 40(11), p. e1800063.

Liu, N., Dai, Q., Zheng, G., He, C., Parisien, M. and Pan, T., 2015. N 6-methyladenosine-dependent RNA structural switches regulate RNA–protein interactions. *Nature*, *518*(7540), pp.560-564.

Liu, T. *et al.* (2019) 'Cancer-associated fibroblasts: An emerging target of anti-cancer immunotherapy', *Journal of Hematology and Oncology*, 12(1), pp. 1–15.

Liu, W. *et al.* (2020) 'TNF-α increases breast cancer stem-like cells through up-regulating TAZ expression via the non-canonical NF-κB pathway', *Scientific Reports*, 10(1), p. 1804.

Liu, Y. *et al.* (2019) 'Poly(A) inclusive RNA isoform sequencing (PAIso−seq) reveals wide-spread non-adenosine residues within RNA poly(A) tails', *Nature Communications*, 10(1), pp. 1–13.

Liu, Y. *et al.* (2021) 'Comprehensive analysis of mRNA poly(A) tail reveals complex and conserved regulation', *bioRxiv,* pp.2021-08

Liu, Y. *et al.* (2022) 'Quality control recommendations for RNASeq using FFPE samples based on pre-sequencing lab metrics and post-sequencing bioinformatics metrics', *BMC Medical Genomics*, 15(1), pp. 1–12.

Lopes, I. *et al.* (2021) 'Gene Size Matters: An Analysis of Gene Length in the Human Genome', *Frontiers in Genetics*, 12, p. 559988.

Love, M.I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*, 15(12), pp. 1–21.

Love, M.I., Soneson, C. and Patro, R. (2018) 'Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification', *F1000Research*, 7, p. 952.

Lu, X. *et al.* (2016) 'Multiple P-TEFbs cooperatively regulate the release of promoter-proximally paused RNA polymerase II', *Nucleic Acids Research*, 44(14), pp. 6853-6867.

Lucarelli, G. *et al.* (2019) 'Metabolomic insights into pathophysiological mechanisms and biomarker discovery in clear cell renal cell carcinoma', *Expert Review of Molecular Diagnostics*, 19(5), pp. 397–407.

Macher-Goeppinger, S. *et al.* (2017) 'Expression and Functional Characterization of the BNIP3 Protein in Renal Cell Carcinomas', *Translational Oncology*, 10(6), pp.869-975.

Mahdipour-Shirayeh, A. *et al.* (2022) 'SciCNV: High-throughput paired profiling of transcriptomes and DNA copy number variations at single-cell resolution', *Briefings in Bioinformatics*, 23(1), p. bbab413.

Maher, E.A. *et al.* (2012) 'Metabolism of [U-13C]glucose in Human Brain Tumors In Vivo', *NMR in biomedicine*, 25(11), p. 1234.

Malek, N. *et al.* (2020) 'Knockout of ACTB and ACTG1 with CRISPR/Cas9(D10A) technique shows that non-muscle β and γ actin are not equal in relation to human melanoma cells' motility and focal adhesion formation', *International Journal of Molecular Sciences*, 21(8), p.2746.

Malka, Y. *et al.* (2022) 'Alternative cleavage and polyadenylation generates downstream uncapped RNA isoforms with translation potential', *Molecular cell*, 82(20), pp. 3840–3855.

Malquori, L., Carsetti, L. and Ruberti, G. (2008) 'The 3′ UTR of the human CTLA4 mRNA can regulate mRNA stability and translational efficiency', *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, 1779(1), pp. 60–65.

Mao, X. *et al.* (2021) 'Crosstalk between cancer-associated fibroblasts and immune cells in the tumor microenvironment: new findings and future perspectives.', *Molecular cancer*, 20(1), p. 131.

Marczyk, M. *et al.* (2019) 'The impact of RNA extraction method on accurate RNA sequencing from formalin-fixed paraffin-embedded tissues', *BMC Cancer*, 19(1), pp. 1-12.

Martin, M. (2011) 'Cutadapt removes adapter sequences from high-throughput sequencing reads', *EMBnet.journal*, 17(1), pp. 10-12.

Masamha, C.P. *et al.* (2014) 'CFIm25 links alternative polyadenylation to glioblastoma tumour suppression', *Nature*, 510(7505), pp. 412-416.

Masoud, G.N. and Li, W. (2015) 'HIF-1α pathway: role, regulation and intervention for cancer therapy', *Acta Pharmaceutica Sinica. B*, 5(5), p. 378.

Matsushita, H. *et al.* (2016) 'Neoantigen load, antigen presentation machinery, and immune signatures determine prognosis in clear cell renal cell carcinoma', *Cancer Immunology Research*, 4(5), pp. 463-471.

Mauer, J. and Jaffrey, S.R. (2018) 'FTO, m^6A_m, and the hypothesis of reversible epitranscriptomic mRNA modifications', *FEBS Letters*, 592(12), pp. 2012–2022.

Mayr, C. (2019) 'What are 3′ utrs doing?', *Cold Spring Harbor Perspectives in Biology*, 11(10), p.a034728.

Mayr, C. and Bartel, D.P. (2009) 'Widespread Shortening of 3′UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells', *Cell*, 138(4), pp. 673-684.

McCaw, T.R. *et al.* (2019) 'The expression of MHC class II molecules on murine breast tumors delays T-cell exhaustion, expands the T-cell repertoire, and slows tumor growth', *Cancer Immunology, Immunotherapy*, 68(2), pp.175-188.

Meyer, K.D. *et al.* (2012) 'Comprehensive Analysis of mRNA Methylation Reveals Enrichment in 3′ UTRs and near Stop Codons', *Cell*, 149(7), pp. 1635–1646.

Millis, S.Z. *et al.* (2016) 'Landscape of Phosphatidylinositol-3-Kinase Pathway Alterations Across 19 784 Diverse Solid Tumors', *JAMA Oncology*, 2(12), pp. 1565–1573.

Mills, C. *et al.* (2020) 'PEREGRINE: A genome-wide prediction of enhancer to gene relationships supported by experimental evidence.', *PloS one*, 15(12), p. e0243791.

Mitchell, T.J. *et al.* (2018) 'Timing the Landmark Events in the Evolution of Clear Cell Renal

Cell Cancer: TRACERx Renal', *Cell*, 173(3), pp. 611-623.

Mizuno, R. *et al.* (2019) 'PD-1 Primarily Targets TCR Signal in the Inhibition of Functional T Cell Activation', *Frontiers in Immunology*, 10, p. 630.

Moch, H. *et al.* (2022) 'The 2022 World Health Organization Classification of Tumours of the Urinary System and Male Genital Organs—Part A: Renal, Penile, and Testicular Tumours', *European Urology*, 82(5), pp. 458–468.

Monteran, L. and Erez, N. (2019) 'The Dark Side of Fibroblasts: Cancer-Associated Fibroblasts as Mediators of Immunosuppression in the Tumor Microenvironment.', *Frontiers in immunology*, 10, p. 1835.

Mørch, A.M. *et al.* (2020) 'Coreceptors and TCR Signaling - the Strong and the Weak of It.', *Frontiers in cell and developmental biology*, 8, p. 597627.

Morra, L. and Moch, H. (2011) 'Periostin expression and epithelial-mesenchymal transition in cancer: a review and an update', *Virchows Archiv*, 459(5), p. 465.

Mulroney, L. *et al.* (2022) 'Identification of high-confidence human poly(A) RNA isoform scaffolds using nanopore sequencing', *RNA*, 28(2), pp. 162-176.

Murakami, S. and Jaffrey, S.R. (2022) 'Hidden codes in mRNA: Control of gene expression by m6A', *Molecular Cell*, 82(12), pp. 2236–2251.

Murray, E.L. and Schoenberg, D.R. (2008) 'Assays for determining poly(A) tail length and the polarity of mRNA decay in mammalian cells.', *Methods in enzymology*, 448, pp. 483–504.

Nachtergaele, S. and He, C. (2017) 'The emerging biology of RNA post-transcriptional modifications.', *RNA biology*, 14(2), pp. 156–163.

Nanavaty, V. *et al.* (2020) 'DNA Methylation Regulates Alternative Polyadenylation via CTCF and the Cohesin Complex', *Molecular Cell*, 78(4), pp. 752-764.

Négrier, S. *et al.* (2002) 'Prognostic factors of survival and rapid progression in 782 patients with metastatic renal carcinomas treated by cytokines: A report from the Groupe Français d'Immunothérapie', *Annals of Oncology*, 13(9), pp. 1460-1468.

Neil, C.R. and Fairbrother, W.G. (2019) 'Intronic RNA: Ad"junk" mediator of post-transcriptional gene regulation', *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1862(11–12), p. 194439.

Neugebauer, K.M. (2019) ‘Nascent RNA and the coordination of splicing with transcription’, *Cold Spring Harbor Perspectives in Biology*, 11(8), p.a032227.

Neve, J. *et al.* (2017) ‘Cleavage and polyadenylation: Ending the message expands gene regulation’, *RNA Biology*, 14(7), pp. 865–890.

Newman, A.M. *et al.* (2019) ‘Determining cell type abundance and expression from bulk tissues with digital cytometry’, *Nature Biotechnology*, 37(7), pp.773-782.

Ng, K.W. *et al.* (2019) ‘Soluble PD-L1 generated by endogenous retroelement exaptation is a receptor antagonist’, *eLife*, 8, p. e50256.

Nicholson-Shaw, A.L. *et al.* (2022) ‘Nuclear and cytoplasmic poly(A) binding proteins (PABPs) favor distinct transcripts and isoforms’, *Nucleic Acids Research*, 50(8), pp. 4685–4702.

Nilsson, H. *et al.* (2020) ‘Features of increased malignancy in eosinophilic clear cell renal cell carcinoma’, *The Journal of Pathology*, 252(4), p. 384.

Nip, K.M. *et al.* (2022) ‘Reference-free assembly of long-read transcriptome sequencing data with RNA-Bloom2’, *bioRxiv*, pp. 2022-08.

Noe Gonzalez, M. *et al.* (2018) ‘CTD-dependent and -independent mechanisms govern co-transcriptional capping of Pol II transcripts’, *Nature Communications*, 9(1), p.3392.

Noguchi, S. *et al.* (2017) ‘FANTOM5 CAGE profiles of human and mouse samples’, *Scientific Data*, 4(170112), pp. 1-10.

Van Nostrand, E.L. *et al.* (2020) ‘A large-scale binding and functional map of human RNA-binding proteins’, *Nature*, 583(7818), pp.711-719.

Nourse, J., Spada, S. and Danckwardt, S. (2020) ‘Emerging roles of RNA 3′-end cleavage and polyadenylation in pathogenesis, diagnosis and therapy of human disorders’, *Biomolecules*, 10(6), pp. 1–43.

O’Leary, N.A. *et al.* (2016) ‘Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation’, *Nucleic Acids Research*, 44(D1), pp. D733–D745.

Ogorodnikov, A. *et al.* (2018) ‘Transcriptome 3′end organization by PCF11 links alternative polyadenylation to formation and neuronal differentiation of neuroblastoma’, *Nature Communications*, 9(1), p. 5331.

Oh, S.A. and Li, M.O. (2013) 'TGF-β: Guardian of T Cell Function', *The Journal of Immunology*, 191(8), pp. 3973-3979.

Oka, M. *et al.* (2021) 'Aberrant splicing isoforms detected by full-length transcriptome sequencing as transcripts of potential neoantigens in non-small cell lung cancer', *Genome Biology*, 22(1), pp. 1-30.

Okeyo-Owuor, T. *et al.* (2015) 'U2AF1 mutations alter sequence specificity of pre-mRNA binding and splicing', *Leukemia*, 29(4), pp. 909-917.

Oldham, K.A. *et al.* (2012) 'T lymphocyte recruitment into renal cell carcinoma tissue: A role for chemokine receptors CXCR3, CXCR6, CCR5, and CCR6', *European Urology*, 61(2), pp. 385-394.

Owen, K.L., Brockwell, N.K. and Parker, B.S. (2019) 'Jak-stat signaling: A double-edged sword of immune regulation and cancer progression', *Cancers*, 11(12), p. 2002.

Pacini, C. *et al.* (2021) 'Integrated cross-study datasets of genetic dependencies in cancer', *Nature Communications*, 12(1), p. 1661.

Padala, S.A. *et al.* (2020) 'Epidemiology of renal cell carcinoma', *World Journal of Oncology*, 11(3), pp. 79-87.

Pal, K. *et al.* (2019) 'Synchronous inhibition of mTOR and VEGF/NRP1 axis impedes tumor growth and metastasis in renal cancer', *npj Precision Oncology*, 3(1), pp. 1–11.

Pacheco-Fiallos, B. *et al.* (2023) 'mRNA recognition and packaging by the human transcription–export complex'. *Nature*, pp.1-8.

Papandreou, I. *et al.* (2006) 'HIF-1 mediates adaptation to hypoxia by actively downregulating mitochondrial oxygen consumption', *Cell Metabolism*, 3(3), pp. 187–197.

Pardo, O.E. and Seckl, M.J. (2013) 'S6K2: The Neglected S6 Kinase Family Member', *Frontiers in Oncology*, 3, p. 191.

Paris, J. *et al.* (2019) 'Targeting the RNA m6A Reader YTHDF2 Selectively Compromises Cancer Stem Cells in Acute Myeloid Leukemia', *Cell Stem Cell*, 25(1), pp. 137-148.

Park, H.J. *et al.* (2018) '3′ UTR shortening represses tumor-suppressor genes in trans by disrupting ceRNA crosstalk', *Nature Genetics*, 50(6), pp. 783-789.

Parker, M.T. *et al.* (2020) 'Nanopore direct RNA sequencing maps the complexity of arabidopsis mRNA processing and m6A modification', *eLife*, 9, p.e49658.

Passmore, L.A. and Coller, J. (2022) ‘Roles of mRNA poly(A) tails in regulation of eukaryotic gene expression’, *Nature Reviews Molecular Cell Biology*, 23(2), pp. 93–106.

Patil, D.P., Chen, C.K., Pickering, B.F., *et al.* (2016) ‘M6 A RNA methylation promotes XIST-mediated transcriptional repression’, *Nature*, 537(7620), pp. 369–373.

Patil, D.P., Chen, C.K., Pickering, brian F., *et al.* (2016) ‘m6A RNA methylation promotes XIST-mediated transcriptional repression’, *Nature*, 537(7620), pp. 369–373.

Patro, R. *et al.* (2017b) ‘Salmon provides fast and bias-aware quantification of transcript expression’, *Nature Methods*, 14(4), pp. 417–419.

Paysan-Lafosse, T. *et al.* (2023) 'InterPro in 2022', *Nucleic Acids Research*, 51(D1), pp. D418-D427.

Peña-Llopis, S. *et al.* (2012) ‘BAP1 loss defines a new class of renal cell carcinoma’, *Nature Genetics 2012 44:7*, 44(7), pp. 751–759.

Peng, Y.L. *et al.* (2022) ‘Single-cell transcriptomics reveals a low CD8+ T cell infiltrating state mediated by fibroblasts in recurrent renal cell carcinoma’, *Journal for ImmunoTherapy of Cancer*, 10(2).

Pennock, N.D. *et al.* (2019) ‘RNA-seq from archival FFPE breast cancer samples: Molecular pathway fidelity and novel discovery’, *BMC Medical Genomics*, 12(1), pp. 1-18.

Pertea, G. and Pertea, M. (2020) ‘GFF Utilities: GffRead and GffCompare’, *F1000Research*, 9, p. 304.

Pickup, M.W., Mouw, J.K. and Weaver, V.M. (2014) ‘The extracellular matrix modulates the hallmarks of cancer’, *EMBO reports*, 15(12), pp. 1243-1253.

Ping, X.L. *et al.* (2014) ‘Mammalian WTAP is a regulatory subunit of the RNA N6-methyladenosine methyltransferase’, *Cell Research*, 24(2), pp. 177–189.

Piovesan, A. *et al.* (2019) ‘Human protein-coding genes and gene feature statistics in 2019’, *BMC Research Notes*, 12(1), pp. 1-5.

Poh, H.X. *et al.* (2022) ‘Alternative splicing of METTL3 explains apparently METTL3-independentm6A modifications in mRNA’, *PLoS Biology*, 20(7), pp. 1–25.

Polenkowski, M. *et al.* (2023) 'THOC5 complexes with DDX5, DDX17, and CDK12 to regulate R loop structures and transcription elongation rate'. iScience, 26(1), p. 105784

Pratanwanich, P.N. *et al.* (2021) 'Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore', *Nature Biotechnology*, 39(11), pp.1394-1402.

Price, A.M. *et al.* (2020) 'Direct RNA sequencing reveals m6A modifications on adenovirus RNA are necessary for efficient splicing', *Nature Communications*, 11(1), pp. 1–17.

Pugacheva, E.M. *et al.* (2020) 'CTCF mediates chromatin looping via N-terminal domain-dependent cohesin retention', *Proceedings of the National Academy of Sciences of the United States of America*, 117(4), pp. 2020-2031.

Pühringer, T., Hohmann, U., Fin, L., Pacheco-Fiallos, B., Schellhaas, U., Brennecke, J. and Plaschka, C., 2020. Structure of the human core transcription-export complex reveals a hub for multivalent interactions. *Elife*, *9*, p.e61503.

Puzanov, G.A. (2022) 'Identification of key genes of the ccRCC subtype with poor prognosis', *Scientific Reports*, 12(1), pp. 1–10.

Qi, X. *et al.* (2021) 'The Uniqueness of Clear Cell Renal Cell Carcinoma: Summary of the Process and Abnormality of Glucose Metabolism and Lipid Metabolism in ccRCC', *Frontiers in Oncology*, 11, p. 727778.

Qiu, Q.C. *et al.* (2018) 'CHI3L1 promotes tumor progression by activating TGF-β signaling pathway in hepatocellular carcinoma', *Scientific Reports*, 8(1), p. 15029.

Racle, J. *et al.* (2017) 'Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data', *eLife*, 6, p. e26476.

Radaeva, M. *et al.* (2021) 'Drugging the "undruggable". Therapeutic targeting of protein–DNA interactions with the use of computer-aided drug discovery methods', *Drug Discovery Today*, 26(11), pp. 2660–2679.

Raplee, I.D., Evsikov, A. V. and De Evsikova, C.M. (2019) 'Aligning the aligners: Comparison of rna sequencing data alignment and gene expression quantification tools for clinical breast cancer research', *Journal of Personalized Medicine*, 9(2), p.18.

Raskov, H. *et al.* (2021) 'Cytotoxic CD8+ T cells in cancer and cancer immunotherapy', *British Journal of Cancer*, 124(2), pp. 359–367.

Rassy, E., Flippot, R. and Albiges, L. (2020) 'Tyrosine kinase inhibitors and immunotherapy combinations in renal cell carcinoma', *Therapeutic Advances in Medical Oncology*, 12, p. 1758835920907504.

Ren, L. *et al.* (2019) 'Apolipoproteins and cancer', *Cancer Medicine*, 8(16), pp. 7032–7043.

René, C., Lozano, C. and Eliaou, J.-F. (2016) 'Expression of classical HLA class I molecules: regulation and clinical impacts', *HLA*, 87(5), pp. 338–349.

Richards, K.E. *et al.* (2016) 'Cancer-associated fibroblast exosomes regulate survival and proliferation of pancreatic cancer cells', *Oncogene 2017 36:13*, 36(13), pp. 1770–1778.

Rini, B. *et al.* (2015) 'A 16-gene assay to predict recurrence after surgery in localised renal cell carcinoma: Development and validation studies', *The Lancet Oncology*, 16(6), pp. 676-685.

Rissland, O.S. *et al.* (2017) 'The influence of microRNAs and poly(A) tail length on endogenous mRNA-protein complexes', *Genome Biology*, 18(1), pp. 1-18.

Ritchie, A.W.W., Griffiths, G. and Parmar, M. (1999) 'Interferon-α and survival in metastatic renal carcinoma: Early results of a randomised controlled trial', *Lancet*, 353(9146), pp.14-17.

Robert, C. (2020) 'A decade of immune-checkpoint inhibitors in cancer therapy', *Nature Communications*, 11(1), p. 3801.

Robinson, C.M., Shirey, K.A. and Carlin, J.M. (2003) 'Synergistic transcriptional activation of indoleamine dioxygenase by IFN-γ and tumor necrosis factor-α', *Journal of Interferon and Cytokine Research*, 23(8), pp. 413–421.

Robinson, M.D. and Nowicka, M. (2016) 'DRIMSeq: A Dirichlet-multinomial framework for multivariate count outcomes in genomics', *F1000Research*, 5, p. 1356.

Roundtree, I.A. *et al.* (2017) 'YTHDC1 mediates nuclear export of N6-methyladenosine methylated mRNAs.', *eLife*, 6, p. e31311.

Ruf, M., Moch, H. and Schraml, P. (2016) 'PD-L1 expression is regulated by hypoxia inducible factor in clear cell renal cell carcinoma', *International Journal of Cancer*, 139(2), pp. 396–403.

Ruterbusch, M. *et al.* (2020) 'In Vivo CD4+ T Cell Differentiation and Function: Revisiting the Th1/Th2 Paradigm', *Annual Review of Immunology*, 38, pp. 705–725.

Saba, J. *et al.* (2019) 'The elemental mechanism of transcriptional pausing', *eLife*, 8, p. e40981.

Sahlin, K. and Mäkinen, V. (2021) 'Accurate spliced alignment of long RNA sequencing

reads', *Bioinformatics*, 37(24), pp. 4643-4651.

Salceda, S. and Caro, J. (1997) 'Hypoxia-inducible Factor 1α (HIF-1α) Protein Is Rapidly Degraded by the Ubiquitin-Proteasome System under Normoxic Conditions', *Journal of Biological Chemistry*, 272(36), pp. 22642–22647.

Sallés, F.J. *et al.* (1994) 'Coordinate initiation of Drosophila development by regulated polyadenylation of maternal messenger RNAs', *Science*, 266(5193), pp. 1996-1999.

Sarantopoulou, D. *et al.* (2019) 'Comparative evaluation of RNA-Seq library preparation methods for strand-specificity and low input', *Scientific Reports*, 9(1), p. 13477.

Sautès-Fridman, C. *et al.* (2019) 'Tertiary lymphoid structures in the era of cancer immunotherapy', *Nature Reviews Cancer*, 19(6), pp. 307-325.

Schlautmann, L.P. and Gehring, N.H., 2020. A day in the life of the exon junction complex. *Biomolecules*, *10*(6), p.866.

Schmid, T.A. and Gore, M.E. (2016) 'Sunitinib in the treatment of metastatic renal cell carcinoma', *Therapeutic Advances in Urology*, 8(6), pp. 348–371.

Schödel, J. *et al.* (2016) 'Hypoxia, Hypoxia-inducible Transcription Factors, and Renal Cancer', *European urology*, 69(4), p. 646.

Schöller, E. *et al.* (2018) 'Interactions, localization, and phosphorylation of the m6A generating METTL3–METTL14–WTAP complex', *RNA*, 24(4), pp. 499–512.

Schulz, M.H. *et al.* (2012) 'Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels', *Bioinformatics*, 28(8), pp. 1086-1092.

Schwartz, S. *et al.* (2014) 'Perturbation of m6A Writers Reveals Two Distinct Classes of mRNA Methylation at Internal and 5′ Sites', *Cell Reports*, 8(1), pp. 284–296.

Seiler, M. *et al.* (2018) 'Somatic Mutational Landscape of Splicing Factor Genes and Their Functional Consequences across 33 Cancer Types', *Cell Reports*, 23(1), pp. 282-296.

Sekar, R.R. *et al.* (2016) 'Major histocompatibility complex i upregulation in clear cell renal cell carcinoma is associated with increased survival', *Asian Journal of Urology*, 3(2), pp. 75-81.

Seliger, B. *et al.* (2003) 'Characterization of human lymphocyte antigen class I antigen-processing machinery defects in renal cell carcinoma lesions with special emphasis on transporter-associated with antigen-processing down-regulation', *Clinical Cancer Research*,

9(5), pp. 1721-1727.

Semenza, G.L. (2007) 'HIF-1 mediates the Warburg effect in clear cell renal carcinoma', *Journal of Bioenergetics and Biomembranes*, 39(3), pp. 231–234.

Seo, H. *et al.* (2019) 'TOX and TOX2 transcription factors cooperate with NR4A transcription factors to impose CD8+ T cell exhaustion', *Proceedings of the National Academy of Sciences of the United States of America*, 116(25), pp. 12410-12415.

Sethumadhavan, S. *et al.* (2017) 'Hypoxia and hypoxia-inducible factor (HIF) downregulate antigen-presenting MHC class I molecules limiting tumor cell recognition by T cells', *PLoS ONE*, 12(11), p. e0187314.

Shatsky, I.N. *et al.* (2018) 'Cap-Independent Translation: What's in a Name?', *Trends in Biochemical Sciences*, 43(11), pp. 882–895.

Sheets, M.D., Wu, M. and Wickens, M. (1995) 'Polyadenylation of c-mos mRNA as a control point in Xenopus meiotic maturation', *Nature*, 374(6522), pp. 511-516.

Shen, C. *et al.* (2020) 'RNA Demethylase ALKBH5 Selectively Promotes Tumorigenesis and Cancer Stem Cell Self-Renewal in Acute Myeloid Leukemia', *Cell Stem Cell*, 27(1), pp. 64-80.

Shen, D. *et al.* (2022) 'METTL14-mediated Lnc-LSG1 m6A modification inhibits clear cell renal cell carcinoma metastasis via regulating ESRP2 ubiquitination', *Molecular Therapy - Nucleic Acids*, 27, pp. 547-561.

Sheng, I.Y. and Ornstein, M.C. (2020) 'Ipilimumab and nivolumab as first-line treatment of patients with renal cell carcinoma: The evidence to date', *Cancer Management and Research*, 12, pp. 4871–4881.

Sheng, Q. *et al.* (2017) 'Multi-perspective quality control of Illumina RNA sequencing data analysis', *Briefings in Functional Genomics*, 16(4), pp. 194-204.

Shi, H. *et al.* (2017) 'YTHDF3 facilitates translation and decay of N 6-methyladenosine-modified RNA', *Cell Research*, 27(3), pp. 315–328.

Shi, H. *et al.* (2021) 'Bias in RNA-seq Library Preparation: Current Challenges and Solutions', *BioMed Research International*. Edited by N. Raju, 2021, p. 6647597.

Shimoda, M. *et al.* (2014) 'Loss of the Timp gene family is sufficient for the acquisition of the CAF-like cell state', *Nature Cell Biology 2014 16:9*, 16(9), pp. 889–901.

Shirai, Y.T. *et al.* (2014) 'Multifunctional roles of the mammalian CCR4-NOT complex in physiological phenomena', *Frontiers in Genetics*, 5, p. 286.

Signoretti, S. *et al.* (2018) 'Renal cell carcinoma in the era of precision medicine: From molecular pathology to tissue-based biomarkers', *Journal of Clinical Oncology*, 36(36), p. 3553

Son, K. *et al.* (2018) 'A simple guideline to assess the characteristics of RNA-Seq Data', *BioMed Research International*, 2018.

Soneson, C. *et al.* (2019) 'A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes.', *Nature communications*, 10(1), p. 3359.

Soneson, C., Love, M.I. and Robinson, M.D. (2016) 'Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences', *F1000Research*, 4, p. 1521.

Song, M. *et al.* (2019) 'Low-dose IFNg induces tumor cell stemness in tumor microenvironment of non–small cell lung cancer', *Cancer Research*, 79(14), pp. 3737–3748.

Srivastava, A. *et al.* (2020) 'Alignment and mapping methodology influence transcript abundance estimation', *Genome Biology*, 21(1), pp. 1-29.

Stark, R., Grzelak, M. and Hadfield, J. (2019) 'RNA sequencing: the teenage years', *Nature Reviews Genetics 2019 20:11*, 20(11), pp. 631–656.

Su, S., Akbarinejad, S. and Shahriyari, L. (2021) 'Immune classification of clear cell renal cell carcinoma', *Scientific Reports*, 11(1), p. 4338.

Subtelny, A.O. *et al.* (2014) 'Poly(A)-tail profiling reveals an embryonic switch in translational control', *Nature*, 508(1), pp. 66-71.

Sun, A. *et al.* (2017) 'Phosphorylation of Ser6 in hnRNPA1 by S6K2 regulates glucose metabolism and cell growth in colorectal cancer', *Oncology Letters*, 14(6), pp. 6323-7331.

Sung, H. *et al.* (2021) 'Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries', *CA: A Cancer Journal for Clinicians*, 71(3), pp. 209–249.

Tabana, Y. *et al.* (2021) 'Reversing T-cell exhaustion in immunotherapy: a review on current approaches and limitations', *Expert Opinion on Therapeutic Targets*, 25(5), pp. 347-363.

Takahashi, H. *et al.* (2012) '5′ end-centered expression profiling using cap-analysis gene

expression and next-generation sequencing', *Nature Protocols*, 7(3), pp. 542-561.

Tang, A.D. *et al.* (2020) 'Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns', *Nature Communications*, 11(1), pp. 1–12.

Tang, C. *et al.* (2017) 'ALKBH5-dependent m6A demethylation controls splicing and stability of long 3'-UTR mRNAs in male germ cells', *Proceedings of the National Academy of Sciences of the United States of America*, 115(2), pp. E325–E333.

Tang, Y. *et al.* (2021) 'M6A-Atlas: A comprehensive knowledgebase for unraveling the N6-methyladenosine (m6A) epitranscriptome', *Nucleic Acids Research*, 49(D1), pp. D134–D143.

Tang, Y. *et al.* (2022) 'PBRM1 deficiency oncogenic addiction is associated with activated AKT–mTOR signalling and aerobic glycolysis in clear cell renal cell carcinoma cells', *Journal of Cellular and Molecular Medicine*, 26(14), pp. 3837–3849.

Tay, R.E., Richardson, E.K. and Toh, H.C. (2021) 'Revisiting the role of CD4+ T cells in cancer immunotherapy—new insights into old paradigms', *Cancer Gene Therapy*, 28(1-2), pp. 5–17.

Terzo, E.A. *et al.* (2019) 'SETD2 loss sensitizes cells to PI3Kβ and AKT inhibition', *Oncotarget*, 10(6), p. 647.

Tian, B. and Graber, J.H. (2012) 'Signals for pre-mRNA cleavage and polyadenylation', *Wiley Interdisciplinary Reviews: RNA*, 3(3), pp. 385–396.

Tian, Z.-H. *et al.* (2019) 'Systematic identification of key genes and pathways in clear cell renal cell carcinoma on bioinformatics analysis', *Annals of Translational Medicine*, 7(5).

Trapnell, C. *et al.* (2012) 'Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks', *Nature Protocols*, 7(3), pp. 562-578.

Tudek, A. *et al.* (2021) 'Global view on the metabolism of RNA poly(A) tails in yeast Saccharomyces cerevisiae', *Nature Communications*, 12(1), p. 4951.

Ueda, K. *et al.* (2018) 'Prognostic value of PD-1 and PD-L1 expression in patients with metastatic clear cell renal cell carcinoma', *Urologic Oncology: Seminars and Original Investigations*, 36(11), pp.499-e9.

Ugolini, C. *et al.* (2022) 'Nanopore ReCappable sequencing maps SARS-CoV-2 5′ capping

sites and provides new insights into the structure of sgRNAs', *Nucleic Acids Research*, 50(6), pp. 3475 - 3489.

Vajavaara, H. *et al.* (2021) 'Soluble PD-1 but Not PD-L1 Levels Predict Poor Outcome in Patients with High-Risk Diffuse Large B-Cell Lymphoma', *Cancers*, 13(3), p. 398.

Vasudev, N.S. *et al.* (2020) 'UK Multicenter Prospective Evaluation of the Leibovich Score in Localized Renal Cell Carcinoma: Performance has Altered Over Time', *Urology*, 136, pp.162 - 168.

Velagapudi, S. *et al.* (2018) 'Scavenger receptor BI promotes cytoplasmic accumulation of lipoproteins in clear-cell renal cell carcinoma', *Journal of Lipid Research*, 59(11), pp. 2188 - 2201.

Viscardi, M.J. and Arribere, J.A. (2022) 'Poly(a) selection introduces bias and undue noise in direct RNA-sequencing', *BMC Genomics*, 23(1), pp. 1–10.

Volden, R. and Vollmers, C. (2022) 'Single-cell isoform analysis in human immune cells', *Genome Biology*, 23(1), p. 47.

Wachutka, L. *et al.* (2019) 'Global donor and acceptor splicing site kinetics in human cells.', *eLife*, 8, p. e45056.

Waldman, A.D., Fritz, J.M. and Lenardo, M.J. (2020) 'A guide to cancer immunotherapy: from T cell basic science to clinical practice', *Nature Reviews Immunology*, 20(11), pp. 651–668.

Wang, H. *et al.* (2012) 'Widespread plasticity in CTCF occupancy linked to DNA methylation', *Genome Research*, 22(9), pp. 1680 - 1688.

Wang, H. *et al.* (2015) 'Relevance and Therapeutic Possibility of PTEN-Long in Renal Cell Carcinoma', *PLoS ONE*, 10(2), p. e114250.

Wang, H. *et al.* (2021) 'Identification of a Novel Stem Cell Subtype for Clear Cell Renal Cell Carcinoma Based on Stem Cell Gene Profiling', *Frontiers in Oncology*, 11, p. 4891.

Wang, J. *et al.* (2020) 'Leukemogenic Chromatin Alterations Promote AML Leukemia Stem Cells via a KDM4C-ALKBH5-AXL Signaling Axis', *Cell Stem Cell*, 27(1), pp. 81-97.

Wang, L. *et al.* (2013) 'IFN-γ and TNF-α synergistically induce mesenchymal stem cell impairment and tumorigenesis via NFκB signaling', *Stem Cells*, 31(7), pp. 1383-1395.

Wang, L. *et al.* (2020) 'm6A RNA methyltransferases METTL3/14 regulate immune

responses to anti□PD□1 therapy', *The EMBO Journal*, 39(20), p. e104514.

Wang, P., Doxtader, K.A. and Nam, Y. (2016) 'Structural Basis for Cooperative Function of Mettl3 and Mettl14 Methyltransferases', *Molecular Cell*, 63(2), pp. 306–317.

Wang, Q., Liu, F. and Liu, L. (2017) 'Prognostic significance of PD-L1 in solid tumor: An updated meta-analysis', *Medicine (United States)*, p. 6369.

Wang, W. *et al.* (2015) 'Effect of platelet-derived growth factor-B on renal cell carcinoma growth and progression', *Urologic oncology*, 33(4), pp. 168 -168.

Wang, Xiao *et al.* (2015) 'N(6)-methyladenosine Modulates Messenger RNA Translation Efficiency.', *Cell*, 161(6), pp. 1388–99.

Wang, Xi *et al.* (2015) 'Tumor suppressor miR-34a targets PD-L1 and functions as a potential immunotherapeutic target in acute myeloid leukemia', *Cellular Signalling*, 27(3), pp. 443 - 452.

WANG, Y. *et al.* (2015) 'Mechanism of alternative splicing and its regulation', *Biomedical Reports*, 3(2), pp. 152-158.

Wang, Z., Gerstein, M. and Snyder, M. (2009) 'RNA-Seq: A revolutionary tool for transcriptomics', *Nature Reviews Genetics*, pp. 57–63.

Warburg, O., Wind, F. and Negelein, E. (1927) 'THE METABOLISM OF TUMORS IN THE BODY', *The Journal of General Physiology*, 8(6), p. 519.

Wei, J. *et al.* (2018) 'Differential m 6 A, m 6 A m , and m 1 A Demethylation Mediated by FTO in the Cell Nucleus and Cytoplasm', *Molecular Cell*, 71, pp. 973-985.

Weinstein, J.N. *et al.* (2013) 'The cancer genome atlas pan-cancer analysis project', *Nature Genetics*, 45(10), pp. 1113 - 1120.

Wells, J.P., White, J. and Stirling, P.C., 2019. R loops and their composite cancer connections. *Trends in Cancer*, *5*(10), pp.619-631.

Wen, J. *et al.* (2018) 'Zc3h13 Regulates Nuclear RNA m6A Methylation and Mouse Embryonic Stem Cell Self-Renewal.', *Molecular cell*, 69(6), pp. 1028-1038

Westermann, L. *et al.* (2022) 'Wildtype heterogeneity contributes to clonal variability in genome edited cells', *Scientific Reports*, 12(1), pp. 1–13.

Wherry, E.J. and Kurachi, M. (2015) 'Molecular and cellular insights into T cell exhaustion',

Nature Reviews Immunology., 15(8), pp. 486-499.

Wickramasinghe, V.O. and Laskey, R.A. (2015) 'Control of mammalian gene expression by selective mRNA export', *Nature reviews Molecular cell biology*, 16(7), pp. 431-442

Wieczorek, M. *et al.* (2017) 'Major histocompatibility complex (MHC) class I and MHC class II proteins: Conformational plasticity in antigen presentation', *Frontiers in Immunology*, 8, p. 292.

Wolf, J. and Passmore, L.A. (2014) 'mRNA deadenylation by Pan2-Pan3', *Biochemical Society Transactions*, 42(1), pp. 184-187.

Workman, R.E. *et al.* (2019) 'Nanopore native RNA sequencing of a human poly(A) transcriptome', *Nature Methods*, 16(12), pp. 1297–1305.

Wu, L. *et al.* (2019) 'Changes of N6-methyladenosine modulators promote breast cancer progression', *BMC Cancer*, 19(1), p. 326.

Wu, T. *et al.* (2021) 'clusterProfiler 4.0: A universal enrichment tool for interpreting omics data', *The Innovation*, 2(3), p. 100141.

Wulf, M.G. *et al.* (2019) 'Non-templated addition and template switching by Moloney murine leukemia virus (MMLV)-based reverse transcriptases co-occur and compete with each other', *Journal of Biological Chemistry*, 294(48), pp. 18220 - 18231.

Xia, A. *et al.* (2019) 'T Cell Dysfunction in Cancer Immunity and Immunotherapy', *Frontiers in immunology*, 10, p. 1719.

Xia, Z. *et al.* (2014) 'Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'2-UTR landscape across seven tumour types', *Nature Communications*, 5, p. 5274.

Xiang, K. and Bartel, D.P. (2021) 'The molecular basis of coupling between poly(A)-tail length and translational efficiency', *eLife*, 10, p. e66493.

Xiao, Y. *et al.* (2020) 'The m6A RNA demethylase FTO is a HIF-independent synthetic lethal partner with the VHL tumor suppressor', *Proceedings of the National Academy of Sciences of the United States of America*, 117(35), pp. 21441 - 21449.

Xie, H. *et al.* (2021) 'Glycogen metabolism is dispensable for tumour progression in clear cell renal cell carcinoma', *Nature Metabolism 2021 3:3*, 3(3), pp. 327–336.

Xue, C., Zhao, Y., Li, G. and Li, L., 2021. Multi-Omic Analyses of the m5C Regulator ALYREF Reveal Its Essential Roles in Hepatocellular Carcinoma. *Frontiers in Oncology*, *11*,

p.633415.

Yang, L. *et al.* (2011) ‘Genomewide characterization of non-polyadenylated RNAs’, *Genome Biology*, 12(2), p. 16.

Yang, S.W. *et al.* (2020) ‘A Cancer-Specific Ubiquitin Ligase Drives mRNA Alternative Polyadenylation by Ubiquitinating the mRNA 3′ End Processing Complex’, *Molecular Cell*, 77(6), pp. 1206-1221.

Yang, S.W. *et al.* (2020) ‘Structural basis for substrate recognition and chemical inhibition of oncogenic MAGE ubiquitin ligases’, *Nature Communications 2020 11:1*, 11(1), pp. 1–14.

Yankova, E. *et al.* (2021) ‘Small-molecule inhibition of METTL3 as a strategy against myeloid leukaemia’, *Nature*, 593(7860), pp. 597–601.

Yee, D. *et al.* (2017) ‘MicroRNA-155 induction via TNF-α and IFN-γ suppresses expression of programmed death ligand-1 (PD-L1) in human primary cells.’, *The Journal of biological chemistry*, 292(50), pp. 20683–20693.

Yoshihara, K. *et al.* (2013) ‘Inferring tumour purity and stromal and immune cell admixture from expression data.’, *Nature communications*, 4, p. 2612.

Yu, Y. *et al.* (2013) ‘Cancer-associated fibroblasts induce epithelial–mesenchymal transition of breast cancer cells through paracrine TGF-β signalling’, *British Journal of Cancer 2014 110:3*, 110(3), pp. 724–732.

Yuan, F. *et al.* (2021) ‘Alternative polyadenylation of mRNA and its role in cancer’, *Genes and Diseases*, 8(1), pp. 61–72.

Yue, Y. *et al.* (2018) ‘VIRMA mediates preferential m6A mRNA methylation in 3′UTR and near stop codon and associates with alternative polyadenylation’, *Cell Discovery*, 4(1), p. 10.

Zaccara, S. and Jaffrey, S.R. (2020) ‘A Unified Model for the Function of YTHDF Proteins in Regulating m6A-Modified mRNA’, *Cell*, 181(7), pp. 1582-1595.

Zarnack, K. *et al.* (2020) ‘Dynamic mRNP remodeling in response to internal and external stimuli’, *Biomolecules*, 10(9), pp. 1–32.

Zeng, Y. *et al.* (2018) ‘Refined RIP-seq protocol for epitranscriptome analysis with low input materials’, *PLOS Biology*, 16(9), p. e2006092.

Zhan, X. *et al.* (2020) ‘A pan-kidney cancer study identifies subtype specific perturbations on pathways with potential drivers in renal cell carcinoma’, *BMC Medical Genomics*, 13(11),

pp. 1–15.

Zhang, C., Fu, J. and Zhou, Y. (2019) ‘A Review in Research Progress Concerning m6A Methylation and Immunoregulation’, *Frontiers in immunology*, 10, p. 922.

Zhang, F. *et al.* (2016) ‘TGF-β induces M2-like macrophage polarization via SNAILmediated suppression of a pro-inflammatory phenotype’, *Oncotarget*, 7(32), pp.52294-52306.

Zhang, Q. *et al.* (2019) ‘Activation and function of receptor tyrosine kinases in human clear cell renal cell carcinomas’, *BMC Cancer*, 19(1), p. 1044.

Zhang, S. *et al.* (2019) ‘Systemic Interferon-g Increases MHC Class I Expression and T-cell Infiltration in Cold Tumors: Results of a phase 0 clinical trial’, *Cancer Immunology Research*, 7(8), pp. 1237 - 1243.

Zhang, S. *et al.* (2021) ‘Structure of a transcribing RNA polymerase II-U1 snRNP complex’, *Science*, 371(6526), pp. 305-309.

Zhang, W. *et al.* (2022) ‘Novel Method of Full-Length RNA-seq That Expands the Identification of Non-Polyadenylated RNAs Using Nanopore Sequencing’, *Analytical Chemistry*, 94(36), pp. 12342–12351.

Zhang, Y. *et al.* (2008) ‘Model-based analysis of ChIP-Seq (MACS)’, *Genome Biology*, 9(9), pp. 1–9.

Zhang, Y. *et al.* (2021) ‘Alternative splicing and cancer: a systematic review’, *Signal Transduction and Targeted Therapy*, 6(1), p. 78.

Zhang, Y. and Hamada, M. (2020) ‘MoAIMS: Efficient software for detection of enriched regions of MeRIP-Seq’, *BMC Bioinformatics*, 21(1), pp. 1–12.

Zhao, G. *et al.* (2020) ‘Exosomal Sonic Hedgehog derived from cancer-associated fibroblasts promotes proliferation and migration of esophageal squamous cell carcinoma’, *Cancer Medicine*, 9(7), pp. 2500–2513.

Zhao, H. *et al.* (2016) ‘Tumor microenvironment derived exosomes pleiotropically modulate cancer cell metabolism.’, *eLife*, 5, e10250.

Zhao, T. *et al.* (2020) ‘Chitinase-3 like-protein-1 function and its role in diseases’, *Signal Transduction and Targeted Therapy*., 5(1), p.201.

Zhao, W. *et al.* (2014) ‘Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling’, *BMC Genomics*, 15(1).

Zheng, D. *et al.* (2018) ‘Cellular stress alters 3′UTR landscape through alternative polyadenylation and isoform-specific degradation’, *Nature Communications*, 9(1), p. 2268.

Zheng, K. *et al.* (2013) ‘Co-transcriptional formation of DNA:RNA hybrid G-quadruplex and potential function as constitutional cis element for transcription control.’, *Nucleic acids research*, 41(10), pp. 5533–41.

Zheng, L. *et al.* (2021) ‘Pan-cancer single-cell landscape of tumor-infiltrating T cells’, *Science*, 374(6574), p. abe6474.

Zhou, X. *et al.* (2019) ‘Splicing factor SRSF1 promotes gliomagenesis via oncogenic splice-switching of MYO1B’, *Journal of Clinical Investigation*, 129(2), pp. 676 - 693.

Zhuang, J. *et al.* (2015) ‘TGFβ1 secreted by cancer-associated fibroblasts induces epithelial-mesenchymal transition of bladder cancer cells through lncRNA-ZEB2NAT’, *Scientific Reports 2015*, 5(1), pp. 1–13.

Zimmer, A.M. *et al.* (2019) ‘Loss-of-function approaches in comparative physiology: Is there a future for knockdown experiments in the era of genome editing?’, *Journal of Experimental Biology*, 222(7), p. jeb175737.

Zingone, A. *et al.* (2021) ‘A comprehensive map of alternative polyadenylation in African American and European American lung cancer patients’, *Nature Communications*, 12(1), p. 5605.