Exploring the Endogenous Retroelement Transcriptome for Epithelial Cancer-Specific Targets and Biomarkers

Rachael S. Thompson

University College London

and

The Francis Crick Institute

PhD Supervisor: George Kassiotis

A thesis submitted for the degree of

Doctor of Philosophy

University College London

March 2024

Declaration

I, Rachael S. Thompson, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Retrotransposable elements (RTEs), although fragmented and mutated, contain regulatory elements able to influence the genome and transcriptome. The hypomethylated state of cancer genomes allows for RTE expression, thus revealing cancer-specific effects on the transcriptome. These effects were recently annotated through a genome-guided *de novo* transcriptome assembly, which is explored in this thesis to identify targets and biomarkers in epithelial cancers. The increased search space of the RTE transcriptome was used to identify known and novel breast tumour-specific transcripts, based on expression in The Cancer Genome Atlas (TCGA) and Genotype-Tissue Expression (GTEx) datasets, that may be present in liquid biopsies. These tissue-specific transcripts were searched for in blood derived extracellular RNAseq data, with the aim to separate breast tumour bearing donors and others, both healthy and not. Alternatively, cancer-specific plasma membrane proteins could be used to diagnose patients from tissue biopsies, and be used as therapy targets. The RTE transcriptome encodes predicted transmembrane domain containing open reading frames (ORFs) arising purely from RTEs and from gene-RTE chimeras. Although RTEs are present in the healthy genome, their derived peptides are not always tolerogenic, this combined with being less polymorphic than mutations, make RTEs a potentially rich source of universal cancer-specific antigen. In this work, candidates for *in vitro* stability testing were selected computationally based on their cancer specificity, possible antigenicity, and predicted transmembrane domain position. Beyond diagnosis, markers of patient survival and immunotherapy response are important for patient stratification and further elucidating disease mechanisms. In kidney renal clear cell carcinoma (KIRC), previous work had suggested several RTE loci were associated with response and give rise to tumour-associated antigens. This work explores the RTE expression landscape using an updated loci annotation in the context of KIRC-characteristic pseudohypoxia, as well as surveying the KIRC-specific RTE transcriptome for markers of survival.

Impact Statement

Due to the hypomethylated state of the cancer genome retrotransposable elements (RTEs) are able to influence the transcriptome, altering gene and isoform expression patterns, and contributing peptide sequences. However, the extent of this influence is incompletely understood, alongside the contribution of these sequences to cancer-specific biomarkers and targets. This thesis explores a previously built pan-cancer *de novo* transcriptome assembly uncovering the influence of RTEs on the cancer-specific transcriptome. Here it is shown that presence of RTE-derived sequences in RNA liquid biopsies is influenced by methodological artefacts highlighting the need for consistent quality controls and methodology when collecting and preparing extracellular RNA samples. Although RTE sequences in liquid biopsies were not useful for demarcating patient groups, RTE-derived sequences potentially giving rise to stable transmembrane proteins were identified for use as tissue biomarkers and immunotherapy targets. Alternatively, any non-coding isoforms identified may reduce expression of canonical transmembrane targets highlighting the requirement for further understanding of activation of RTE sequences prior to use of demethylating drugs to treat patients. Further testing of other cancer-specific candidate transcripts identified in this work is required as they have the potential to be stable and localise to the cell surface. Moreover, transcripts are identified which stratify patient response to immunotherapy and patient survival prior to treatment. Additionally, corrections have been made to HERV loci previously associated with response to immunotherapy.

Acknowledgement

I would like to express my sincere gratitude to my supervisor, George Kassiotis, for the insightful feedback and discussion throughout my projects. Your immense knowledge of retrotransposable elements, immunology, and Qlucore has been a huge help for both my thesis and my academic journey. Your encouragement and support have shaped me into a better scientist.

Furthermore, I would like to extend my heartfelt thanks to my fellow lab mates. Laura for training me in both the lab and Italian cuisine, and for making my wet lab life so much easier; Kevin for the academic advice; Nasta for the bioinformatics guidance; Callum for the computing input, pep talks, and encouragement in exploring vegan baking; Tom for teaching me flow cytometry and introducing me to your wonderful poetry; Jane for enlightening me to the layout editor on FlowJo, encouraging me through Western blots, and helping me write my impact statement; Toby for your PCR advice and scarf modelling; Niki for opening up the lab to flexible work hours and support during the valleys of despair; Judith for training me in so many techniques, and for your knowledge of earrings and makeup.

I would like to thank my thesis committee Javier Herrero, Peter Van Loo, Nicholas Luscombe, and Samra Turajlic for all of their advice and guidance. I am also very appreciative of all of the interesting scientific discussions about HERVs in renal cancers with Lewis, Benjy, Ángel, Robert, and Kevin.

I am indebted to the amazing core facilities at the Crick who made my laboratory and computational work so much easier, particularly the glass-wash, media, delivery, cell services, flow cytometry, and scientific computing teams. As well as a big thank you to Patti who saved my thesis when my referencing software stopped working.

Mum, Dad, and Sam thank you for your thoughtfulness in sending me food and chocolate during stressful times providing much-needed comfort and sustenance. Thank you for the stream of pictures of dogs, hens, food, and the garden offering moments of respite from my studies and reminding me of the greenery outside of London. I am grateful for your support throughout my academic life granting me the freedom to choose my own career path. Dobby, Ginny, and Lily, thank you for the endless games and distraction when my studies were tough, I can never repay you in enough cheese and biscuits.

I am forever grateful for the love I have received from my friends. Bethanie you have been a source of unending support in both my work and my life, I'm glad we got to go on so many adventures together and I am looking forward to many more. Ella you have given me so much wonderful encouragement, I'm sorry I couldn't name any proteins after you but I'm glad our friendship of 16 years lasted through it. Trina, thank you for all the pep talks you've sent me and for reminding me that there are many other things outside of my thesis, I'm so happy to have been able to grow up with you. Jenna thank you for the reminders to look after myself and my plants, and for encouraging me to try lots of new recipes to fuel me through my studies. I am also thankful to Terry, Jule, Siobhan, Natasha, and Rebecca for their humour and support.

Finally, thank you to Beryl, Demeter, Edith, Eurydice, Gertrude, Mildred, Orpheus, Perpetua, Stewart, Vera, and Willie for making the air in our flat crisp and constantly bringing joy.

Table of Contents

Table of Figures

List of Tables

Abbreviations

AA	Amino acids
ACC	Adrenocortical carcinoma
ACTB	Beta-actin
ADAR1	Adenosine deaminase RNA specific 1
AKT	AKT serine/threonine kinase
AM	Adrenomedullin peptide
ANOVA	Analysis of variance
APC	Adenomatous polyposis coli
ARNT	Aryl hydrocarbon receptor nuclear translocator
ATF-4	Activating transcription factor 4
ATP	Adenosine triphosphate
BAM	Binary alignment map
BCR	B-cell receptor
BHLHE41	Basic helix-loop-helix family member e41
BLAST	Basic Local Alignment Search Tool
BLAT	BLAST-like alignment tool
BLCA	Bladder Urothelial Carcinoma
BRCA	Breast invasive carcinoma
BRCA1	Breast cancer susceptibility 1 DNA repair associated
CA9	Carbonic anhydrase 9
CALB1	Calbindin 1
CASP8	Caspase 8
CCL28	C-C motif chemokine ligand 28
cDNA	Complementary DNA
CD274	Cluster of differentiation 274
CD5	Cluster of differentiation 5
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
cGAS	Cyclic GMP (guanosine 3',5'-cyclic monophosphate)-AMP (adenosine monophosphate) synthase
CHOL	Cholangiocarcinoma
CI_{95}	95% confidence interval
CMV	Cytomegalovirus
CNTL	Controls
CO_2	Carbon dioxide
COAD	Colon adenocarcinoma
CpG	Cytosine-phosphate-guanine
CRBN	Cereblon
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CSF1R	Colony-stimulating factor 1 receptor

ctDNA	Circulating tumour DNA
Cyr61	Cellular communication network factor 1
DHX9	DExH-box helicase 9
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
DNA	Deoxyribonucleic acid
DNMT	DNA methyltransferases
dsDNA	Double stranded DNA
dsRNA	Double stranded RNA
EGFR	Epidermal growth factor receptor
ELISA	Enzyme-linked immunosorbent assay
ENPP3	Ectonucleotide pyrophosphatase/phosphodiesterase 3
Env	Sequences producing, with splicing, the envelope and accessory proteins of a HERV
ERCC1	Excision repair cross-complementing 1
ERV	Endogenous retrovirus
ESCA	Oesophageal carcinoma
EV	Extracellular vesicle
exRNA	Extracellular RNA
FACS buffer	Flow cytometry staining buffer
FCS	Foetal calf serum
FOXO4	Forkhead Box Protein O4
FPPP	FFPE Pilot Phase II
GABA	Gamma-aminobutyric acid
GABRA3	Gamma-aminobutyric acid A receptor alpha 3 subunit
Gag	Sequence producing the nucleocapsid, capsid, and matrix proteins of a HERV
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase
GBM	Glioblastoma multiforme
GFP	Green fluorescent protein
GRCh37	Genome Reference Consortium human build 37
GRCh38	Genome Reference Consortium human build 38
GTEx	Genotype-Tissue Expression
GTF2F1	General Transcription Factor IIF Subunit 1
HECTD2	HECT (Homologous to the E6-AP Carboxyl Terminus) domain E3 ubiquitin ligase 2
HER2	Human epidermal growth factor receptor 2
HERV	Human endogenous retrovirus
HGNC	Human Genome Organisation Gene Nomenclature Committee
HIF	Hypoxia inducible factor
HLA	Human leukocyte antigen
hnRNPA2B1	Heterogeneous nuclear ribonucleoprotein A2/B1

HNSC	Head and Neck squamous cell carcinoma
HP1	Heterochromatin protein 1
hPSC	Human pluripotent stem cells
HR	Hazard ratio
HRP	Horseradish peroxidase
HUSH	Human silencing hub
IGV	Integrative Genomics Viewer
IMDM	Iscove's Modified Dulbecco's Medium
IRES	Internal ribosome entry site
IRF3	Interferon regulatory factor 3
KAP1	Krüppel-associated box-associated protein 1
KICH	Kidney Chromophobe
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
KLF5	Krüppel-like factor 5
KRAB-ZFP	Krüppel-associated box domain zinc finger protein
LAML	Acute Myeloid Leukaemia
LB media	Luria-Bertani media
LCML	Chronic Myelogenous Leukaemia
LGG	Brain Lower Grade Glioma
LIHC	Liver hepatocellular carcinoma
LINE	Long interspersed nuclear element
lncRNA	Long non-coding RNA
LTR	Long terminal repeat
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MAGEA3	Melanoma antigen gene family member A3
MaLR	Mammalian apparent LTR retrotransposons
MAVS	Mitochondrial antiviral-signalling protein
MDA5	Melanoma differentiation-associated protein 5
MESO	Mesothelioma
MHC-I	Major histocompatibility complex class I
miRNA	MicroRNA
MITF	Microphthalmia-associated transcription factor
MPP8	M-phase phosphoprotein 8
mRNA	Messenger RNA
MSH2	MutS homolog 2
MSH6	MutS homolog 6
MUTYH	MutY DNA glycosylase
NPP1	Nucleotide pyrophosphatase/phosphodiesterase 1
NSCLC	Non-small cell lung cancer
O_2	Oxygen

OCT4	Octamer-binding transcription factor 4
ORF	Open reading frame
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PAMP	Proadrenomedullin N-terminal 20 peptide
PBRM1	Polybromo 1
PBS	Phosphate-buffered saline
PCPG	Pheochromocytoma and Paraganglioma
PD-1	Programmed cell death protein 1
PD-L1	Programmed death-ligand 1
PLA2G4A	phospholipase A2 group IVA
PLD3	Phospholipase D family member 3
Poly(A)	Polyadenylation
PRAD	Prostate adenocarcinoma
Pro and pol	Sequences producing, with splicing, a protease, reverse transcriptase, and integrase proteins of a HERV
PTBP1	Polypyrimidine tract binding protein
Ras	Rat sarcoma virus
RBM4	RNA-binding motif protein 4
READ	Rectum adenocarcinoma
RIG-I	Retinoic acid-inducible gene I
RIPA	Radioimmunoprecipitation assay
RNA	Ribonucleic acid
RPM	Revolutions per minute
RTE	Retrotransposable element
SAM	Sequence Alignment Map
SARC	Sarcoma
SETDB1	Su(var)3-9, enhancer-of-zeste and trithorax domain bifurcated histone lysine methyltransferase 1
SINE	Short interspersed nuclear element
SKCM	Skin Cutaneous Melanoma
SKCM_m	Metastatic skin cutaneous melanoma
S.O.C	Super optimal broth with catabolite repression
SPARC	Secreted protein acidic and cysteine rich
SPEN	Msx2-interacting protein
SRA	Sequence Read Archive
STAD	Stomach adenocarcinoma
STC2	Stanniocalcin 2
STING	Stimulator of interferon genes protein
SVA	SINE-VNTR-Alu
TASOR	Transcription Activation Suppressor

TBK1	TANK (tumor necrosis factor receptor-associated factor family member-associated NF-kappa-B activator)-binding kinase 1
TBS-T	Tris-buffered saline with 0.5% Tween-20
TCGA	The Cancer Genome Atlas
TCR	T-cell receptor
TF	Transcription factor
TGCT	Testicular Germ Cell Tumours
THCA	Thyroid carcinoma
THYM	Thymoma
TLR	Toll-like receptors
TPM	Transcripts per million
Tregs	Regulatory T-cells
TRPV1	Transient receptor potential vanilloid 1
TRPV3	Transient receptor potential vanilloid 3
TSD	Target site duplication
UCEC	Uterine Corpus Endometrial Carcinoma
UCS	Uterine Carcinosarcoma
UCSC	University of California, Santa Cruz
UPF1	UPF1 RNA helicase and ATPase
UTR	Untranslated region
UVM	Uveal Melanoma
VEGFA	Vascular endothelial growth factor A
VHL	Von Hippel–Lindau tumour suppressor
VNTR	Variable number tandem repeat
VSVg	Vesicular stomatitis virus glycoprotein
ZMYND8	MYND (myeloid, Nervy, and DEAF-1) domain containing 8

Chapter 1. General Introduction

The genomes of eukaryotic organisms vary greatly and seemingly randomly in size, without this variation reflecting the known number of genes encoded. With advancements in deoxyribonucleic acid (DNA) sequencing and genome assembly it became clear that the majority of many species' genomes were comprised of repetitive sequence (Hartl, 2000). When the first draft of the human genome was published it was estimated only 5% of sequence was taken up by coding genes, with repetitive sequences, either of short repeated k-mers, much larger duplications, or transposable element-derived sequences representing 45-60% of the genome (Lander et al., 2001). In the telomere-to-telomere assembly of the human genome 53.9% is estimated to be repetitive sequence (Hoyt et al., 2022). Transposable elements are able to move their genetic information around the host genome and are inherited in a Mendelian fashion. They arose in genomes before humans evolved, with some insertions shared with a wide range of species, whilst others are primate- or human-specific. Retrotransposable elements (RTEs) are a form of transposable element and here specifically refer to ribonucleic acid (RNA) transposons which copy and paste themselves increasing the element copy number each time. These include long terminal repeat (LTR)-containing elements such as human endogenous retroviruses (HERVs). As well as non-LTR containing elements such as autonomous long interspersed nuclear elements (LINEs), and non-autonomous short interspersed nuclear elements (SINEs) and SINE-variable number tandem repeat (VNTR)-Alu (SVA) elements. Over evolutionary time these elements have become fixed in the genome and are present in all humans, but are mutated and degraded so that they are mainly no longer able to transpose. Over time some RTEs have been co-opted by the host working to increase the functional diversity of the genome. However, sequences that have not been co-opted yet remain in human DNA can still influence the transcriptome and therefore the biology of healthy and diseased cells.

1.1 Retrotransposable elements

1.1.1 LTR retrotransposons

LTR-containing elements include HERV and mammalian apparent LTR retrotransposon (MaLR) elements which account for approximately 8% of the human genome (Lander et al., 2001). HERVs are believed to have originated exogenously from infection of the germline by ancient retroviruses and have since undergone fixation, but still carry a similar genome to modern exogenous retroviruses. Local homologous recombination of HERVs and MaLRs leaves solo LTRs. The LTRs contain transcriptional regulatory elements and may flank the coding regions gag (producing the nucleocapsid, capsid, and matrix proteins), pro and pol (with splicing producing a protease, reverse transcriptase, and integrase), and env (producing the envelope and accessory proteins) (Figure 1) (Lander et al., 2001; Mao et al., 2021). Accessory proteins produced include Rec (Figure 1) which aids in nuclear export of HERV transcripts (Magin et al., 1999). However, some HERV-K (HML-2) insertions have a deletion of 292 base pairs which removes part of the envelope coding sequence and a splice site for Rec production (Figure 1). Instead the protein Np9 is produced (Lower et al., 1993; Lower et al., 1995). Most HERV loci are not polymorphic with only 20-30 polymorphic HERV-K loci thus far identified (Chen and Li, 2019; Li et al., 2019). Although some loci may carry the same integration, some insertions remain whole whilst others are represented by solo LTRs (Chu et al., 2021). HERV expression is most notable in early human embryos where each group of HERVs is activated in a concerted way, producing both fully HERV-derived RNAs and RNAs spliced between HERV and non-repetitive genome sequences (Goke et al., 2015). The envelope proteins of *HERV-W* and *HERV-FRD* have been co-opted and are also known as syncytin-1 and syncytin-2 respectively. Both are specifically expressed in the placenta and allow cell-cell fusion for formation of the syncytiotrophoblast layer (Blaise et al., 2003; Mi et al., 2000).

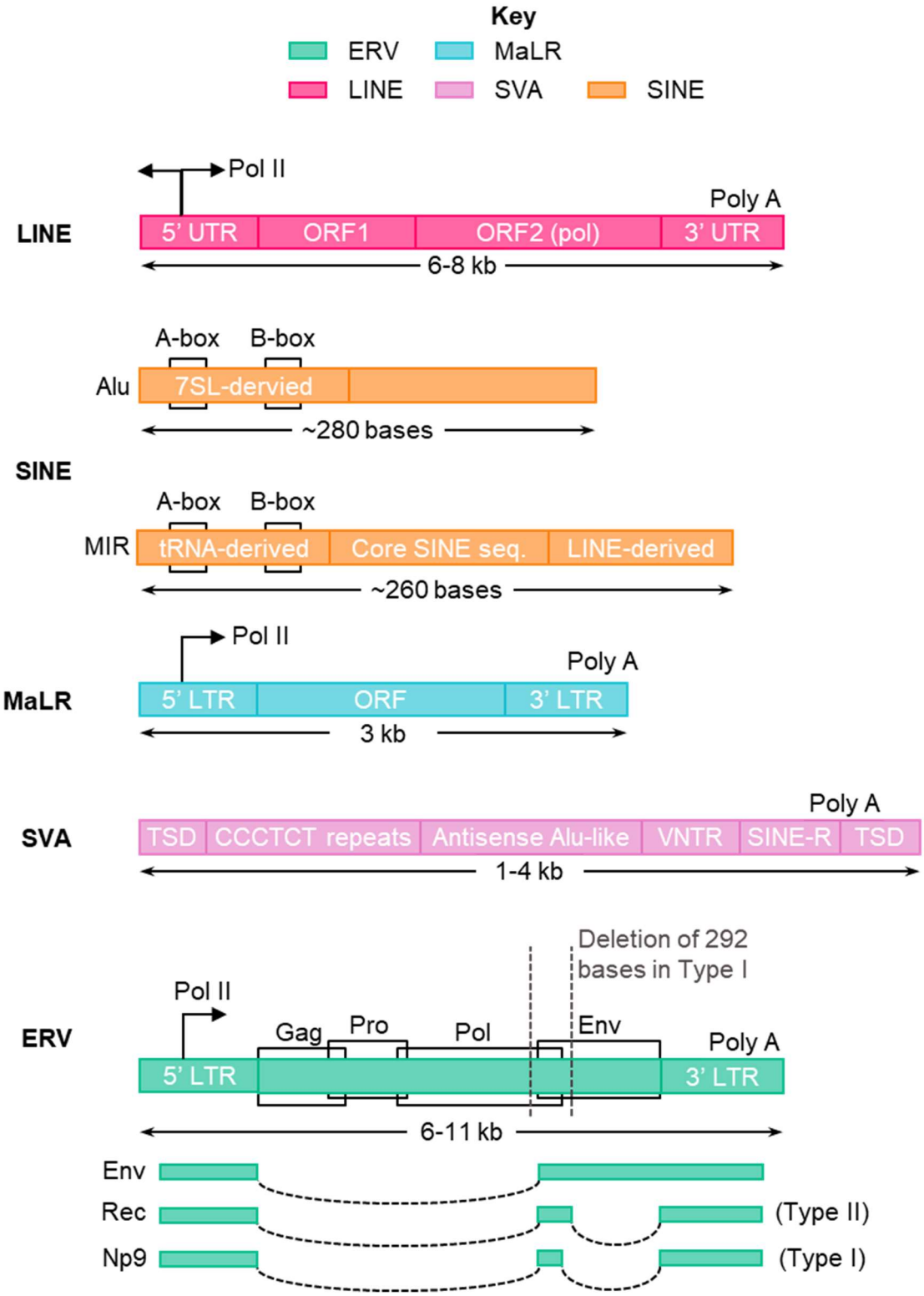

Key
ERV
MaLR
LINE
SVA
SINE
LINE
Pol II
Poly A
5' UTR
ORF1
ORF2 (pol)
3' UTR
6-8 kb
SINE
A-box
B-box
Alu
7SL-dervied
~280 bases
A-box
B-box
MIR
tRNA-derived
Core SINE seq.
LINE-derived
~260 bases
MaLR
Pol II
Poly A
5' LTR
ORF
3' LTR
3 kb
SVA
Poly A
TSD
CCCTCT repeats
Antisense Alu-like
VNTR
SINE-R
TSD
1-4 kb
ERV
Deletion of 292 bases in Type I
Pol II
Gag
Pro
Pol
Env
Poly A
5' LTR
3' LTR
6-11 kb
Env
Rec
(Type II)
Np9
(Type I)

Figure 1: The structures of the main groups of RTEs present in the human genome. For the ERV structure the splicing patterns necessary to produce the envelope and accessory proteins Rec and Np9 in the HERVK group are shown.

1.1.2 LINEs

LINEs represent approximately 20% of the human genome (Lander et al., 2001) with 180 LINE1 polymorphisms on average per person (Bennett et al., 2004). LINEs are autonomous elements responsible for the retrotransposition of not only themselves but SINEs, SVAs, and pseudogenes (Esnault et al., 2000). LINEs contain an RNA-polymerase II promoter in the 5' untranslated region (UTR) able to drive transcription of both the LINE *orf1* and *orf2* (Figure 1) as well as sequences upstream of the LINE element giving rise to transcripts containing LINE sequence alongside sequences of upstream genes (Tubio et al., 2014). The internal LINE transcript is translated to two proteins (ORF1p and ORF2p) which have strong *cis*-preference to bind the exact RNA molecule encoding them (Wei et al., 2001). This specificity is mainly driven through ORF1p (Martin, 2006), although ORF2p may instead bind SINE, SVA, or processed messenger RNA (mRNA). Upon protein binding to the RNA, the complex moves to the nucleus where a new insertion is created. The endonuclease activity of ORF2p puts a nick in one strand of DNA and uses the break to prime reverse transcription of the RNA from the 3' end. Often reverse transcription does not complete so many LINE loci are truncated at the 5' end. New insertions are flanked by small target site duplications (TSDs). There are three groups of LINE elements, LINE1, LINE2, and LINE3 (Lander et al., 2001). The LINE1 group are the only autonomous elements still actively able to retrotranspose sequences (Esnault et al., 2000; Lander et al., 2001).

1.1.3 SINEs

SINEs, non-autonomous elements reliant on LINEs to retrotranspose, make up approximately 13% of the human genome yet encode no proteins (Figure 1) (Lander et al., 2001). The SINE promoter regions are either derived from the signal recognition particle component 7SL or from transfer RNA (tRNA) sequences (Lander et al., 2001). Only the Alu group of SINE elements with approximately 1.2 million copies is active within the human genome (Lander et al., 2001) with approximately 1283 polymorphisms per person (Bennett et al.,

2004). The A-box and B-box within SINEs act as promoters which RNA polymerase III is able to bind (Orioli et al., 2012). Transcription of SINEs also occurs in an RNA polymerase III-independent manner due to insertion into gene introns (Zhang et al., 2021). In some cases, Alu elements form inverted repeats able to influence the degradation and translation of the host gene mRNA (Elbarbary and Maquat, 2017).

1.1.4 SVAs

There are approximately 2700 SVA element insertions in the human genome, with a subset restricted to the human genome (Wang et al., 2005). These are non-autonomous elements, reliant on LINE sequences for retrotransposition (Ostertag et al., 2003). Full length SVAs consist of five domains some with homology to other repeat types and flanked by TSDs (Figure 1). The Alu-like region consists of two antisense Alu elements and the SINE-region appears to be homologous to the 3' end of the HERV-K10 envelope sequence alongside the 3' LTR. There are two regions with repeated sequences, a simple repeat region containing a repeated CCCTCT sequence and a variable number tandem repeat (VNTR) region containing repeats of a 35-50 base pair sequence. Before the final target site duplication there is a putative polyadenylation (poly(A)) signal and poly(A) tail (Wang et al., 2005). There are an estimated 56 polymorphic SVA insertions per person (Bennett et al., 2004) with all insertions potentially able to alter the transcriptome (Quinn and Bubb, 2014). Polymorphic SVA insertions have been shown to cause disease, with insertion into *α-spectrin* causing hereditary elliptocytosis (Ostertag et al., 2003).

1.2 Control of RTE expression

RTE expression must be tightly controlled in order to reduce transposition events that increase genome instability (Cao et al., 2020), to reduce production of inflammatory RTE-derived nucleic acid structures, and to stop deleterious control of gene isoform expression through RTE sequences (Babarinde et al., 2021). Expression is controlled in two main ways, through methylation of element loci and direct protein binding both to RTE DNA and RNA. However, demethylation of RTEs is not sufficient for expression with requirement also for transcription factor (TF) binding to induce transcription (Attig et al., 2019; Kong et al., 2019).

Double-stranded RNA (dsRNA), cytoplasmic double stranded DNA (dsDNA), and DNA-RNA hybrids lead to inflammatory responses within the cell. DsRNA is sensed via the retinoic acid-inducible gene I (RIG-I), melanoma differentiation-associated protein 5 (MDA5), and mitochondrial antiviral-signalling protein (MAVS) dsRNA response pathway (Aktaş et al., 2017; Mehdipour et al., 2020). Cytoplasmic dsDNA and RNA:DNA hybrids can be sensed via the cyclic GMP (guanosine 3',5'-cyclic monophosphate)-AMP (adenosine monophosphate) synthase (cGAS) and stimulator of interferon genes protein (STING) pathway (Sun et al., 2013). Both pathways signal through the TANK (tumour necrosis factor receptor-associated factor family member-associated NF-kappa-B activator)-binding kinase 1 (TBK1) protein which then phosphorylates the interferon regulatory factor 3 (IRF3) allowing dimerization and activation of type I and III interferons (Zhou et al., 2020). These nucleic acid structures are formed most commonly through mis-localised LINE1 reverse transcription in the cytoplasm, however why this occurs outside the nucleus is unknown (De Cecco et al., 2019; Thomas et al., 2017).

Protein binding to RTE sequences suppresses expression at the transcriptional level through recruitment of histone and DNA methyltransferases. Different methylation patterns are enriched across different elements, where for example intermediate age LTRs are associated with histone methylation and young LTRs rich in cytosine-phosphate-guanine (CpG) islands are suppressed with DNA

methylation (Ohtani et al., 2018). In order to inhibit transcription of young LINE1 elements in euchromatic regions of DNA, transcription activation suppressor (TASOR) and M-phase phosphoprotein 8 (MPP8) bind and recruit histone methyltransferases (Liu et al., 2018). Krüppel-associated box domain zinc finger proteins (KRAB-ZFPs) bind a range of RTEs in DNA and repress transcription. KRAB-ZFPs recruit Krüppel-associated box-associated protein 1 (KAP1), and DNA and histone methyltransferases such as SETDB1 (su(var)3-9, enhancer-of-zeste and trithorax domain bifurcated histone lysine methyltransferase 1), DNA methyltransferases (DNMTs), and heterochromatin protein 1 (HP1) (Ivanov et al., 2007; Margolin et al., 1994; Witzgall et al., 1994). This leads to silencing in both early development and in differentiated adult cells (Tie et al., 2018). Knockout of various proteins binding RTEs at the DNA level to recruit DNA and histone methylators leads to increased RTE expression and increased signalling through dsRNA receptors (Aktaş et al., 2017; Mehdipour et al., 2020; Tie et al., 2018).

Protein binding RTE-derived RNA works to sequester sequences in the nucleus and prevent potential translation. DExH-box helicase 9 (DHX9) binds inverted Alu repeats contained within transcripts and destabilises the dsRNA structure formed allowing mRNA processing and translation of the host transcript (Aktaş et al., 2017). Alternatively, adenosine deaminase RNA specific 1 (ADAR1) catalyses adenosine-to-inosine RNA editing of these Alu repeat hairpins (Athanasiadis et al., 2004) which leads to sequestering of the host transcript in the nucleus (Chen et al., 2008). RNA-binding motif protein 4 (RBM4) post-transcriptionally regulates HERV-K and HERV-H group members, by binding the derived RNA and suppressing translation, with loss of RBM4 leading to increased translation of HERV-K envelope protein (Foroushani et al., 2020).

To allow RTE sequence transcription, binding of specific TFs is required. When tumour and adjacent healthy tissue samples were compared the ratio of RTE RNA expression was consistent between tissue pairs suggesting TFs able to bind RTEs are tissue specific (Kong et al., 2019). Furthermore, similar upregulation of RTE-derived transcripts is seen across cancer types derived from similar healthy

tissues (Attig et al., 2019). Some RTEs are enriched for lineage-specific master TF binding sites (Cao et al., 2019). HERV-K LTRs contain sequences recognised by octamer-binding TF 4 (OCT4) which would have allowed fixation in the germline and early embryo (Fuentes et al., 2018). Some LTRs also contain hypoxia inducible factor (HIF) binding sites, which leads to upregulation of the corresponding HERVs as well as neighbouring genes in kidney renal clear cell carcinoma (KIRC) samples due to the uncontrolled activation of HIFs (Siebenthall et al., 2019). Additionally, LINE1 activation in placenta and pluripotent stem cells requires both hypomethylation and binding by the oestrogen receptor (Lanciano et al., 2024).

1.3 Expression of RTEs in healthy tissues

RTE expression has been seen in healthy embryonic and adult tissues. Loci in healthy cells have tissue-specific expression patterns (Chung et al., 2019; Goke et al., 2015; Larouche et al., 2020), and there is expression in medullary thymic epithelial cells suggesting some tolerance for RTE-derived peptides (Larouche et al., 2020). From analysis of the CHM13hTERT foetal human cell line, nascent transcription from the majority of full-length LINE1HS, SVA, and AluY insertions was detected (Hoyt et al., 2022). In peripheral blood mononuclear cells approximately 5.5% of all HERV and MaLR insertions were expressed, with some altering expression under interferon influence (Mommert et al., 2018). Furthermore, analysis of 48 healthy tissues in Genotype-Tissue Expression (GTEx) data showed polymorphic RTE insertions led to changes of isoform proportions and overall expression of neighbouring genes (Cao et al., 2020). Peptides bound to major histocompatibility complex class I (MHC-I) molecules on healthy cells derived from RTE sequences have been detected, suggesting that RTE-derived transcripts are also translated (Larouche et al., 2020).

1.4 RTEs and the non-cancerous transcriptome

RTEs contain transcriptional control elements able to alter expression of other genes and lead to production of RTE-gene chimeric transcripts (Figure 2). RTEs contain elements such as TF binding sites, enhancer and promoter signals, RNA polymerase binding sites, and splice acceptor and donor sites.

Some of these control elements within RTEs have been co-opted for use within the cell. On analysis of RTEs present in the human reference genome, 45.4% of enhancers and 5.1% of promoters were found to overlap an RTE (Simonti et al., 2017). Additionally, enhancer-like RTEs, which are defined as enriched for binding lineage-specific master transcription factors, were able to demarcate cell identity (Cao et al., 2019). Younger and older RTEs were seen to be enriched for different TF binding motifs (Simonti et al., 2017) including enrichment for lineage-specific master transcription factors (Cao et al., 2019; Fuentes et al., 2018; Kazachenka et al., 2023), and multiple RTE-derived promoters were identified near oncogenes (Jang et al., 2019). LINE1-containing transcripts have been shown to regulate T-cell exhaustion and quiescence through silencing of the corresponding protein-coding isoforms until the T-cell is activated (Marasca et al., 2022). Upon T-cell activation general transcription factor IIF subunit 1 (GTF2F1) binds the intronic LINE1 elements leading to splicing out of those introns and production of the respective protein coding isoform (Marasca et al., 2022). In human pluripotent stem cells (hPSCs), 65% of non-coding and 26% of coding transcripts contained RTE-derived sequences (Babarinde et al., 2021). Both HERV and LINE1 sequences were also included in protein producing transcripts (Babarinde et al., 2021). However, RTE-containing transcripts in hPSCs were more often localised to the nucleus instead of the cytoplasm, had lower expression, and had disrupted coding sequences when compared to the non-RTE containing transcripts (Babarinde et al., 2021). Furthermore, a LINE2 acts as a start site for several *OCT4* variants able to produce functional protein (Papamichos, 2021). Additionally, the *cluster of differentiation 274* (*CD274*) gene splices into a *LINE2A* omitting the canonical transmembrane domain and producing a soluble programmed death-ligand 1 (PD-L1) protein able to act as a

receptor antagonist involved in regulation of the immune response (Ng et al., 2019).

RTEs can influence the environmental stimuli that neighbouring genes respond to due to the TF-binding, enhancer, and promoter sites they contain. HERVs specifically have been suggested to influence interferon response pathways as many interferon-responsive enhancers are derived from HERV sequences, with insertion of HERVs near genes leading to interferon-inducible expression (Chuong et al., 2016). Additionally, an SVA between *transient receptor potential vanilloid 1* (*TRPV1*) and *3* (*TRPV3*) is required for co-regulation of the two genes (Price et al., 2021).

Analysis of polymorphic RTE insertions further show the influence these elements can have on the transcriptome. Upon RTE insertion expression of nearby genes was significantly altered due to enhancers losing chromatin accessibility in lymphoblastoid cell lines and induced pluripotent stem cells (Goubert et al., 2020). LINE1 insertions were able to influence the upstream methylation of the insertion site up to 300 bases away (Lanciano et al., 2024). Polymorphic Alu insertions have also been seen to disrupt control elements thus influencing expression of nearby genes (Payer et al., 2021). In reverse, the systematic clustered regularly interspaced short palindromic repeat (CRISPR) knock out targeting of HERV-K (HML-2) LTRs revealed long range effects on gene expression (Fuentes et al., 2018) likely due to the 3-dimensional structure of the genome (Raviram et al., 2018). Although, RTE insertions seen somatically in colorectal cancer samples did not have an effect on neighbouring gene expression (Cajuso et al., 2019).

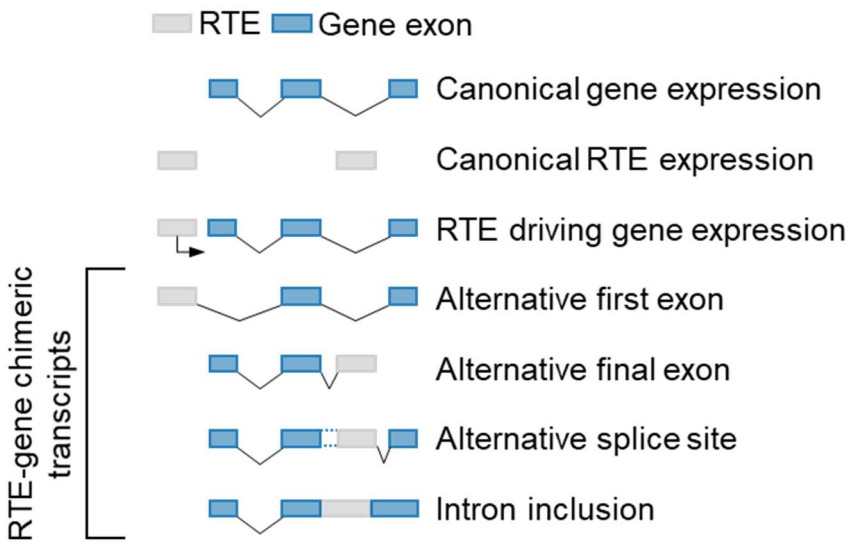


Figure 2: The influence of RTEs on the transcriptome and production of RTE-gene chimeric transcripts.

1.5 Expression of RTEs in cancer

Dysregulation of RTE expression in cancer cells is likely due to the hypomethylated state of the cancer genome. Proximal hypomethylation is associated with RTE expression across cancer types (Kong et al., 2019) although specific TFs are also required for expression leading to cancer-specific expression of certain RTEs (Attig et al., 2019). Epigenetic profiling of KIRC cell lines showed demethylation and subsequent reactivation of HERVs alongside transcription of HERV sequences and transcription of other genes driven through HERV LTRs (Siebenthall et al., 2019). DNA demethylation is enriched in areas near RTEs (Kong et al., 2019) which may be due to downregulation of proteins targeting methylation complexes to the loci. MPP8, part of the human silencing hub (HUSH) complex which targets LINE-1 elements for histone methylation, is downregulated in multiple cancers (Tunbak et al., 2020).

However, understanding of the progression of demethylation in the cancer genome remains incomplete. From analysis of samples of pre-neoplastic monoclonal B-cell lymphocytosis and chronic lymphocytic leukaemia the demethylated state appears early in cancer development and remains stable

throughout disease progression (Kretzmer et al., 2021). Although there were similarities in methylation levels across patients, gene expression varied, implying other levels of control are required for expression (Kretzmer et al., 2021). Furthermore, in colon tissue adjacent to colon tumours there was no significant difference in HERV-H and HERV-K (HML-2) LTR methylation, though HERV-K (HML-2) envelope protein was only detected in tumour tissues whilst the corresponding polymerase was detected more highly in the adjacent healthy tissue (Dolci et al., 2020). However, it is unknown if this hypomethylated state plays a role in causing cancer.

Treatment of tumours with demethylating agents also increases RTE expression. Treatment of cancer cell lines with DNA demethylating agents led to increase in transcripts driven by RTE elements, particularly those induced from LTRs, with novel chimeric transcripts produced (Brocks et al., 2017; Goyal et al., 2023). Peptides derived from these treatment-induced chimeric transcripts were seen in cell lines and in acute myeloid leukaemia patient samples (Goyal et al., 2023). The suggested mode of action of the anti-tumoral effect of demethylating agents is through upregulation of inverted Alu repeat transcription and upregulation of other RTEs producing dsRNA leading to MAVS-dependent immune responses (Mehdipour et al., 2020). With depletion of ADAR1 in cancer cells, prohibiting the sequestering of dsRNA in the nucleus, shown to improve the efficacy of demethylating agents in reducing tumour growth in mice (Mehdipour et al., 2020; Sakurai et al., 2017). However, the expression of RTEs post-treatment in myelodysplastic syndromes and acute myeloid leukaemia did not determine treatment response (Kazachenka et al., 2019).

Expression of RTEs still requires specific TFs, as in healthy tissues, leading to tissue-type specific expression of certain RTEs (Attig et al., 2019; Kong et al., 2019). A *HERV-H* on Xp22.3 is specifically upregulated in colorectal, pancreatic, gastric, and oesophageal cancers (Kazachenka et al., 2023; Wentzensen et al., 2007). A *HERV-K (HML-2)* on 17p13.1 is upregulated in hepatoblastoma when compared to healthy liver (Grabski et al., 2021). Whilst HERV-K gag RNA and

protein are upregulated in prostate cancer samples (Rezaei et al., 2021), and the RNA of HERV-K env, gag, and np9 is also upregulated in BRCA samples (Tavakolian et al., 2019). Regardless of the homology of these HERV sequences, they have expression specific to different tissues and conditions.

Additionally, within every tumour there is a range of DNA mutations and dysregulated pathways, which can also affect RTE locus expression and RTE inclusion within chimeric transcripts. Mutations in genes producing the splicing machinery, as well as other mutation-independent mechanisms lead to abnormal splicing in tumours compared to healthy controls (Dvinge and Bradley, 2015; Frankiw et al., 2019). This abnormal splicing leads to intron retention, where those introns may contain RTEs, with conserved patterns of intron retention across cancers suggesting there is predisposition for inefficient splicing at specific intron boundaries (Dvinge and Bradley, 2015) potentially due to weaker splice signals (Wang et al., 2021). For example, oesophageal adenocarcinoma has especially high rates of intron inclusion across the genome, leading to high rates of RTE inclusion in transcripts (Kazachenka et al., 2023).

1.5.1 Transposition of RTEs in cancer

Upon de-repression of RTEs in cancer some are able to transpose. In colorectal cancer a mean of 25 polymorphic insertions were seen per tumour, and in two cases out of 202 patients there was an insertion into the *adenomatous polyposis coli* gene (*APC*) which would have disrupted the protein-coding sequence and potentially caused the cancer (Cajuso et al., 2019). This had been seen previously in a case of colorectal cancer initiated by a polymorphic *LINE-1* insertion into *APC*. This *LINE-1* was restricted to certain genetic backgrounds with the patient's family having developed a range of epithelial cancers (Scott et al., 2016). An Alu insertion into the *mutY DNA glycosylase* (*MUTYH*) is also associated with increased chance of developing early onset breast and gastric cancers (Zhu et al., 2011). Additionally, an SVA insertion into *caspase 8* (*CASP8*), which led to abnormal splicing, is associated with increased risk of breast cancer

and cutaneous basal-cell carcinoma, but reduced risk of prostate cancer (Stacey et al., 2016). Analysis of 2954 samples from 38 cancer subtypes showed 35% of cancer samples had somatic polymorphic insertions when compared to matched adjacent tissues (Rodriguez-Martin et al., 2020). Though only 66/19166 events hit cancer-associated genes, there were some LINE-1 integrations which led to deletion of segments of chromosomes removing genes associated with tumour suppression (Rodriguez-Martin et al., 2020). Polymorphic insertions can lead to hereditary predisposition to cancers as in the cases of Lynch syndrome caused by insertion of an SVA in *mutS homolog 2* (*MSH2*) or *mutS homolog 6* (*MSH6*) (Yamamoto et al., 2021; Yang et al., 2021a), or an Alu insertion into *MLH1* (Li et al., 2020b; Solassol et al., 2019). Some cases of hereditary breast and ovarian cancer syndrome are caused by Alu insertion into the *breast cancer susceptibility 1 DNA repair associated* (*BRCA1*) gene (Bouras et al., 2021).

1.5.2 Cancer-specific control of the transcriptome

Alongside cancer-specific expression of canonical RTE-derived RNA and protein, there is cancer-specific expression of RTE-driven genes and non-canonical RTE-gene chimeric transcripts (Figure 2). In large scale analyses of tumour transcriptomes, chimeric transcripts overlapping both RTE and known gene sequences have been identified (Attig et al., 2019; Babarinde et al., 2021; Burbage et al., 2023; Goyal et al., 2023; Merlotti et al., 2023; Shah et al., 2023). Promoters in RTEs drive expression of chimeric transcripts, and splice sites allow inclusion within gene transcripts (Attig et al., 2019; Babarinde et al., 2021; Merlotti et al., 2023; Shah et al., 2023), with some genes producing multiple chimeric isoforms (Attig et al., 2019; Shah et al., 2023). Splice donor sites from RTEs were found to be enriched in SINEs, and splice acceptors enriched in HERVs and DNA transposons (Merlotti et al., 2023). Furthermore, exonisation of RTEs was found to be biased towards the beginning of protein coding sequences, and in positions where the protein coding sequence would not be disrupted (Sela et al., 2010).

In some cases, RTE-derived expression drives cancer progression. A *HERV-E* provides a promoter site for a truncated cluster of differentiation 5 (CD5) protein, reducing expression of full-length CD5 which is required to regulate B-cell receptor (BCR) signalling. Increase in the truncated form increases BCR signalling allowing uncontrolled expansion of B lymphocytes potentially leading to chronic lymphocytic leukaemia (Renaudineau et al., 2005). Further alteration of BCR signalling driven by HERVs is seen in anaplastic large cell lymphoma and Hodgkin's lymphoma where an *LTR* activates the *colony-stimulating factor 1 receptor* (*CSF1R*) bypassing BCR signalling and again allowing uncontrolled cell expansion (Lamprecht et al., 2010). A HERV-derived long non-coding RNA (lncRNA), *TROJAN*, drives ubiquitination of the *zinc finger and MYND (myeloid, Nervy, and DEAF-1) domain containing 8* (*ZMYND8*) metastasis-repressing factor leading to progression of triple-negative breast invasive carcinoma (triple-negative BRCA) (Jin et al., 2019). Treatment of mice carrying BRCA tumours with an anti-sense oligonucleotide against *TROJAN* suppressed the tumours and reduced metastases to the liver, bone, and lung (Jin et al., 2019).

In other cases, RTE-derived expression alters cancer biology. Truncated isoforms of *calbindin 1* (*CALB1*) initiated by a *HERV* element promote growth of lung squamous cell carcinoma (LUSC) cell lines both *in vitro* and *in vivo* (Attig et al., 2023). Antisense *HECT (homologous to the E6-AP carboxyl terminus) domain E3 ubiquitin protein ligase 2* (*HECTD2*) transcripts terminating in a *HERV-H* element were associated with better prognosis in uveal melanoma (UVM) and skin cutaneous melanoma (SKCM). Nuclear sequestration of transcripts with inverted Alu repeats due to adenosine-to-inosine editing by ADAR1 also controls translation of the host transcript (Chen et al., 2008). This method of control most commonly occurs in transcripts with inverted Alu repeats in the 3'UTR (Ku et al., 2024). Global shortening of UTRs in tumours may exclude 3'UTR localised inverted Alu repeats from transcription, thus allowing translation of proteins previously downregulated through nuclear sequestration of their transcripts (Ku et al., 2024).

1.5.3 Expression of antigenic proteins

In order to effectively treat cancer patients with immunotherapies there is a need for universal cancer-specific targets, allowing the immune system to differentiate healthy from tumour cells in all patients. Advances in whole genome sequencing have focused the search for tumour specific antigens on mutations of known proteins. However recent immunopeptidomics (sequencing of peptides bound by MHC-I molecules) data has shown peptides derived from presumed non-coding regions are displayed on MHC-I molecules of tumour cells, spurred by the finding that some tumours although responsive to immune checkpoint blockade had a low mutational burden. Immunopeptidomics data and Ribo-Seq (sequencing of transcripts bound by ribosomes) data have shown peptides derived from lncRNAs, pseudogenes, out of frame ORFs of known proteins, ORFs within 5' and 3' UTRs, and proteins derived from presumed non-coding transcripts (Chen et al., 2020; Chong et al., 2020; Laumont et al., 2018; Lu et al., 2019; Ouspenskaia et al., 2020). These peptides have been seen in healthy tissues and malignant samples of glioblastoma, chronic lymphocytic leukaemia, and melanoma, with 50.6% of peptides detected in this study found in two or three of 10 cancer samples, suggesting the translation is not random (Ouspenskaia et al., 2020). Similar results of MHC-I displayed peptides derived from non-canonical coding sequences were also seen in separate samples from melanoma and lung tumours, again with the same peptide detected in multiple samples (Chong et al., 2020). It is possible that these peptides are pervasively produced and are not functional as they show different characteristics to the known human proteome including lower expression levels of both RNA and protein, lower predicted protein stability, higher iso-electric points, and are encoded by fewer exons (Lu et al., 2019). Although functionality may not be relevant for the purposes for antigen targeting, it should be noted that some proteins derived from lncRNAs are stable and specifically localise within the cell (Chen et al., 2020; Lu et al., 2019). These displayed non-canonical peptides are derived from non-mutated yet aberrantly expressed regions, increasing the chance of being universally expressed compared to mutation-dependent antigens.

RTEs are a potential source of cancer-specific non-canonical proteins, with derived peptides previously seen displayed on tumour MHC-I molecules (Kong et al., 2019). RTE expression in cancer samples is associated with immune infiltration in both primary and metastatic samples suggesting these sequences are a source of antigenic proteins (Kong et al., 2019; Topham et al., 2020). RTE-derived antigens include canonical proteins derived from internal coding sequences, such as HERV envelope glycoproteins, and non-canonical proteins derived from RTE-gene chimeric transcripts.

Canonical peptides derived from RTE sequences can act as antigens. Peptides derived from a range of RTE sequences have been detected displayed on MHC-I molecules across cancer types (Kong et al., 2019). From analysis of breast cancer tumours, 192 cancer-specific HERV loci with expression correlated with cytotoxic T-cell signatures have been identified (Bonaventura et al., 2022). Of these, 6 selected candidates coded for protein eliciting high-avidity T-cell responses able to lyse patient-derived organoids. Additionally, 13 of the epitopes identified were coded for by at least 10 individual HERV loci reducing the potential for epitope silencing (Bonaventura et al., 2022). A previous study also identified HERV-K envelope derived peptides were displayed on MHC-I on tumours, with antibodies targeting these peptides reducing tumour mass in mice bearing BRCA tumours (Wang-Johanning et al., 2012). Furthermore, treatment with chimeric antigen receptor T-cells targeted against HERV-K envelope protein reduced growth of mouse BRCA xenograft models and prevented metastasis (Zhou et al., 2015). Peptides derived from HERV-E envelope protein have been detected bound to MHC-I molecules in KIRC patient samples, with *in vitro* recognition by cytotoxic T-cells (Cherkasova et al., 2016). In myeloid malignancies T-cells targeting HERV-derived peptides have also been detected, though response to therapy was not predicted by the expression of targeted HERVs potentially due to the small cohort size (Campbell et al., 2020).

Chimeric transcripts derived from RTE and known gene sequences may also produce proteins with novel antigenic peptide either localised to the cell surface

plasma membrane or processed and displayed on MHC-I molecules for detection as non-self. Protein produced by chimeric transcripts has been detected displayed on MHC-I molecules from cancer cell lines and detected in cell lysates (Burbage et al., 2023; Shah et al., 2023). In non-small cell lung cancer (NSCLC) novel splice sites between RTEs and genes are a source of antigen with cytotoxic T-cells detected in patients targeted against the derived peptides (Merlotti et al., 2023). These antigens were shared between patients as the RTEs involved were not polymorphic and therefore provided potential splice junctions in all patients. Non-canonical peptides may also arise from RTE-internal sequences. A peptide derived from a conserved LTR12C region in 88 transcripts which was expressed in a range of cancer cell lines has been seen presented on MHC-I molecules in tumour samples (Goyal et al., 2023).

1.5.4 The effect of RTE expression on cancer biology and patients bearing tumours

Different studies have associated RTE expression with different effects on patient survival, which may be due to overall RTE expression reflecting a different biology from expression of individual loci. Expression of HERVs has been shown to positively correlate with survival of KIRC patients, as well as with response to immunotherapy and cytotoxic T-cell signatures (Panda et al., 2018; Rooney et al., 2015; Smith et al., 2018). However, other evidence suggests that the HERV loci annotation used in these studies was incorrect and HERV expression instead correlates with immune infiltrate and tumour purity (Au et al., 2021). Expression of the HERV-K envelope is required for rat sarcoma virus protein (Ras)-induced tumorigenesis of BRCA cell lines, with knockdown of all HERV-K loci leading to reduction in cell proliferation, invasion, and metastasis (Zhou et al., 2016). On the other hand, HERV-K expression in melanoma cell lines reduced tumorigenesis, as microphthalmia-associated transcription factor (MITF) binds HERV-K LTRs increasing expression of Rec which in turn reduces the invasive phenotypes and epithelial to mesenchymal transition of cells (Singh et al., 2020). Furthermore, in melanoma patients, HERV-K hypomethylation was associated with disease-free survival (Cardelli et al., 2020). HERV expression in UVM predicts the metastatic

potential of tumours, with dysregulation of HERVs on chromosomes 3 and 8 associated with metastatic risk (Bendall et al., 2022). Increased *HERV-H Xp22.3* in oesophageal adenocarcinoma samples predicted better patient survival, alongside increased intronic RTE retention in transcripts also predicting better survival likely due to the concomitant downregulation of the respective functional protein-coding transcripts (Kazachenka et al., 2023). Additionally, in colorectal cancer samples, higher RTE transposition rates were associated with poor patient survival (Cajuso et al., 2019).

1.6 A *de novo* transcriptome assembly

Previously a *de novo* transcriptome assembly was built to reveal effects of RTEs on the cancer transcriptome (Attig et al., 2019), this thesis utilises transcripts identified in this assembly. The pan-cancer *de novo* transcriptome assembly was created using a subset of randomly selected, subtype and sex (where possible) balanced, samples from 31 primary and one metastatic TCGA cancer sub-types (Attig et al., 2019). The method of assembly has been described previously (Attig et al., 2019). Briefly, the RNAseq data for 24 samples per subtype were trimmed of poor-quality regions and adapters using cutadapt (Martin, 2011) before being mapped to the Genome Reference Consortium human build 38 (GRCh38) using STAR (Dobin et al., 2013). Trinity (Grabherr et al., 2011) was then used for genome guided assembly of transcripts, creating one assembly per cancer subtype. The contigs were poly(A)-trimmed using SeqClean and filtered to remove artefactual or low quality contigs using bbduk (Bushnell, 2014). The respective RNAseq data was then remapped to the cancer-specific assembly using Salmon to ensure surviving contigs had even and accurate read coverage, as well as sufficient expression. The assemblies were then merged to find the longest continuous unique contigs in each region using gffread (Trapnell et al., 2010). The assembly was then compared to the GENCODE database (Frankish et al., 2019) to assign strands to transcripts directly overlapping known exons, as well as assigning annotation levels to each transcript. As all transcripts expressed across cancers should have been captured, the assembly contains both cancer-specific and healthy expressed transcripts. Additionally, although the assembly was created to identify RTE-derived transcripts, transcripts not overlapping RTEs will also have been assembled.

The *de novo* transcriptome assembly revealed novel isoforms and expression drivers, some of which have already been described. For example, a novel isoform of *CD274* was identified which produces a soluble form of PD-L1 due to read through into an intronic *LINE2A* element truncating the transcript after exon four (Figure 3a) (Ng et al., 2019). This truncated isoform does not contain the transmembrane domain, and does not function to inhibit T cells, though does act

as an antagonist by binding PD-1 and blocking canonical PD-L1 binding (Ng et al., 2019). Furthermore, three transcripts in the antisense orientation over the *HECTD2* locus were found with some homology to the previously annotated antisense *HECTD2* (*HECTD2-AS*), but with different transcriptional start sites (Figure 3b) (Attig et al., 2019). One of the antisense transcripts was expressed in uveal melanoma and both primary and metastatic skin cutaneous melanoma samples, and was linked to better prognosis in uveal melanoma and primary skin cutaneous melanoma. Another transcript was expressed in bladder urothelial carcinoma and as well as some healthy tissues. Samples with expression of these antisense transcripts had little or no sense expression of canonical *HECTD2* (Attig et al., 2019). Additionally, three novel transcripts producing a truncated CALB1 protein have been identified (Figure 3c) initiated by a *HERVH* with expression driven through the TF Krüppel-like factor 5 (KLF5) (Attig et al., 2023). The protein produced initiates in the third exon of canonical *CALB1* removing the first 51 amino acids (AA), but is still believed to be functional as loss of these transcripts in lung squamous cell carcinoma cell lines reduced growth both *in vitro* and *in vivo* (Attig et al., 2023).

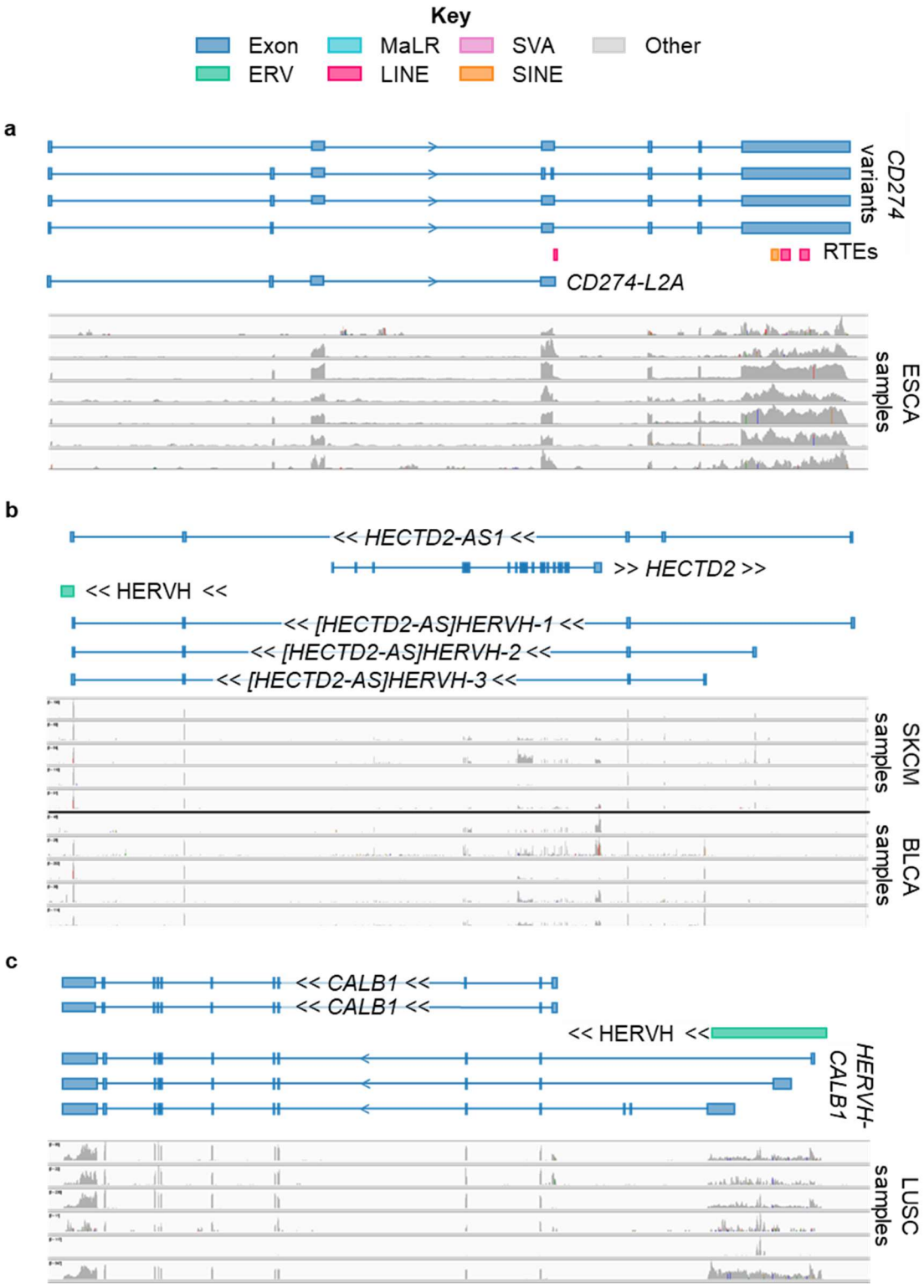

Key
Exon
MaLR
SVA
Other
ERV
LINE
SINE
a
CD274 variants
RTEs
CD274-L2A
ESCA samples
b
<< HECTD2-AS1 <<
>> HECTD2 >>
<< HERVH <<
<< [HECTD2-AS]HERVH-1 <<
<< [HECTD2-AS]HERVH-2 <<
<< [HECTD2-AS]HERVH-3 <<
SKCM samples
BLCA samples
c
<< CALB1 <<
<< CALB1 <<
<< HERVH <<
HERVH-CALB1
LUSC samples

Figure 3: The structure of novel isoforms identified by the *de novo* transcriptome assembly. a. The structure of the *CD274* locus and the novel truncated isoform identified (Ng et al., 2019), alongside BAM files of RNAseq data from oesophageal carcinoma (ESCA) patient samples from TCGA. **b.** The structure of the *HECTD2* locus and the three novel antisense transcripts identified (Attig et al., 2019), alongside BAM files of RNAseq data from skin cutaneous melanoma (SKCM) and bladder urothelial carcinoma (BLCA) patient samples from TCGA. **c.** The structure of the *CALB1* locus and the three novel transcripts identified (Attig et al., 2023), alongside BAM files of RNAseq data from lung squamous cell carcinoma (LUSC) patient samples from TCGA.

Transcripts identified by the *de novo* transcriptome assembly may serve as cancer-specific biomarkers or therapy targets (Figure 4). The *de novo* transcriptome assembly has expanded the transcriptional search space available for identifying cancer-specific isoforms which may be used as biomarkers or as therapy targets, these are explored in Chapter 3: Results 1: Identification of cancer-specific transcripts. The identified cancer-specific transcripts can be used to separate specific cancer types from healthy tissues using tissue RNA profiles and it is possible this expression pattern would be reflected in the blood exRNA profiles of the respective cancer bearing patients. This is explored in the context of liquid biopsies for BRCA in Chapter 4: Results 2: Tumour-specific transcripts in extracellular RNA. Additionally, as the transcripts are so highly cancer specific, peptides derived from these transcripts may be useful for targeted therapies. The *de novo* transcriptome assembly revealed chimeric isoforms with both transmembrane protein coding gene-derived and RTE-derived regions. These peptides, altered from stable canonical peptides, are likely to be antigenic, may be localised to the same place as the canonical protein, and are more likely to be stable than the small potentially transmembrane peptides derived from other fully novel transcripts. The stability and localisation of three candidates are explored in Chapter 5: Results 3: Novel transmembrane domain containing proteins. In a separate approach to identifying antigenic transcripts as well as transcripts influencing patient survival, metastatic KIRC samples pre- and post-immunotherapy treatment were analysed, alongside an extended dataset from TCGA. In Chapter 6: Results 4: Exploration of HERV expression in metastatic KIRC, both the expression of transcripts from the *de novo* transcriptome assembly and expression of individual RTE loci were analysed to identify potential antigen sources and to further understand disease mechanisms through transcripts linked to patient survival. Finally, to further understand disease mechanisms associated with levels of hypoxia and patient survival in KIRC, KIRC-upregulated transcripts assembled in the *de novo* transcriptome are explored in Chapter 7: Results 5: Exploration of transcripts upregulated in KIRC.

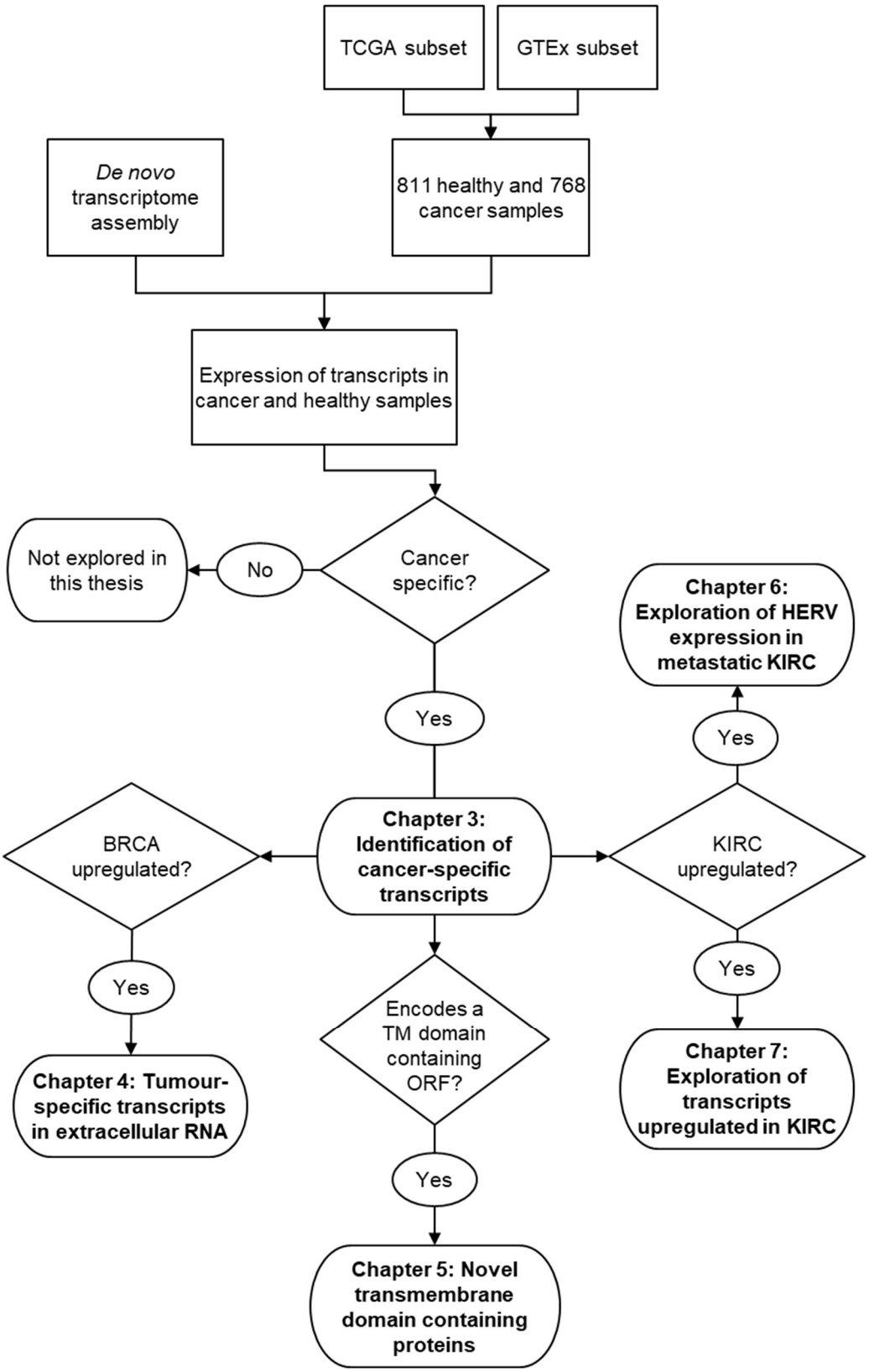


Figure 4: An overview of the structure of this thesis and the exploration of the selected cancer-specific transcripts Results chapters are shown in bold.

1.7 Aims

Previous work has begun to elucidate a role for RTEs in both healthy tissue and cancer biology, with roles in initiation, progression, and immunogenicity of tumours. A few RTE integrations have been co-opted over evolutionary time to be involved in processes including placental formation, regulation of T-cell quiescence, and as a source of enhancer, promoter, and TF-binding site sequences. However, less is known about the remainder of the integrations, most of which are epigenetically silenced in healthy tissues. Due to the hypomethylated state of the cancer genome, as well as mutations in splice machinery and RTE-specific silencing machinery, RTEs become expressed both as fully RTE-derived transcripts and as RTE-gene chimeric transcripts. These transcripts may produce antigenic peptide, control downstream signalling pathways, or influence expression of other isoforms of the same gene. Due to the complexity of the effects of RTEs on the transcriptome, understanding of the roles these repetitive elements play is incomplete. We therefore set out to ask:

1. To what extent do RTEs contribute to the cancer-specific transcriptome?
2. As some of these RTE-overlapping transcripts can differentiate healthy and tumour tissue, are they also released into the blood for use as a liquid biomarker of tumour presence?
3. Can the cancer-specific RTE-overlapping transcripts act as a source of transmembrane antigen for therapy targeting and for use as tissue biomarkers?
4. Can the transcripts be used to stratify patients for immune checkpoint blockade treatment?

Chapter 2. Materials & Methods

2.1 RNAseq data

2.1.1 Original tissue datasets (Attig et al., 2019)

RNAseq data from 768 cancer patients was downloaded from TCGA, with 24 samples for each of the 31 primary and one metastatic cancer types (Table 1). Where possible, healthy tissue-matched controls were downloaded from TCGA and the GTEx consortium, totalling 811 samples (Attig et al., 2019) (Table 1).

2.1.2 Expanded BRCA tissue dataset

BRCA-specific transcript expression was validated in an additional TCGA BRCA dataset of 100 patients totalling 33 basal-like, 25 human epidermal growth factor receptor 2 (HER2)-enriched, 30 luminal A, 26 luminal B, and 28 matched normal tissue samples. The 100 patients were selected randomly ensuring a balance of cancer subtypes.

2.1.3 Extracellular RNA sequencing datasets

Publicly available independent extracellular RNA (exRNA) datasets were downloaded using the Sequence Read Archive (SRA) Toolkit (version 3.0.0). The Melanoma exRNA dataset was downloaded from the authors' online repository (Table 2). In datasets where donors had given multiple samples, samples were pooled for analysis. Variation in methods for plasma or serum collection, presence of an extracellular vesicle (EV) isolation step, RNA isolation, and RNA sequencing was present amongst the studies (Table 2).

Table 1: Overview of RNAseq samples in the original dataset (Attig et al., 2019). Abbreviations of cancer types are shown where used.

Tissue type	Tissue origin	Abbreviation	Number of samples
Healthy	Adipose tissue		24
Cancer	Adrenocortical carcinoma	ACC	24
Cancer	Pheochromocytoma and paraganglioma	PCPG	24
Healthy	Adrenal gland		15
Cancer	Cholangiocarcinoma	CHOL	24
Healthy	Bile duct		9
Cancer	Bladder urothelial carcinoma	BLCA	24
Healthy	Bladder		22
Healthy	Blood		24
Healthy	Blood vessel		36
Cancer	Brain lower grade glioma	LGG	24
Cancer	Glioblastoma multiforme	GBM	24
Healthy	Brain		156
Cancer	Breast invasive carcinoma	BRCA	24
Healthy	Breast		24
Cancer	Cervical squamous cell carcinoma and endocervical adenocarcinoma	CESC	24
Healthy	Cervix		14
Cancer	Colon adenocarcinoma	COAD	24
Cancer	Rectum adenocarcinoma	READ	24
Healthy	Colorectal		42
Cancer	Esophageal carcinoma	ESCA	24
Healthy	Esophagus		47
Cancer	Uveal melanoma	UVM	24
Healthy	Fallopian tube		7
Cancer	Head and neck squamous cell carcinoma	HNSC	24
Healthy	Head and neck		12
Healthy	Heart		24
Cancer	Kidney renal clear cell carcinoma	KIRC	24
Cancer	Kidney renal papillary cell carcinoma	KIRP	24
Healthy	Kidney		36
Cancer	Liver hepatocellular carcinoma	LIHC	24
Healthy	Liver		24
Cancer	Lung adenocarcinoma	LUAD	24
Cancer	Lung squamous cell carcinoma	LUSC	24
Healthy	Lung		36
Cancer	Lymphoid neoplasm diffuse large B-cell lymphoma	DLBC	24
Healthy	Muscle		12

Healthy	Nerve		12
Cancer	Ovarian serous cystadenocarcinoma	OV	24
Healthy	Ovary		12
Cancer	Pancreatic adenocarcinoma	PAAD	24
Healthy	Pancreas		16
Healthy	Pituitary		12
Cancer	Mesothelioma	MESO	24
Cancer	Prostate adenocarcinoma	PRAD	24
Healthy	Prostate		24
Healthy	Salivary gland		12
Cancer	Skin cutaneous melanoma	SKCM	24
Cancer	Metastatic skin cutaneous melanoma	SKCM_m	24
Healthy	Skin		36
Healthy	Small intestine		12
Cancer	Sarcoma	SARC	24
Healthy	Spleen		12
Cancer	Stomach adenocarcinoma	STAD	24
Healthy	Stomach		24
Cancer	Testicular germ cell tumours	TGCT	24
Healthy	Testis		12
Cancer	Thymoma	THYM	24
Healthy	Thymus		2
Cancer	Thyroid carcinoma	THCA	24
Healthy	Thyroid		24
Cancer	Uterine carcinosarcoma	UCS	24
Cancer	Uterine corpus endometrial carcinoma	UCEC	24
Healthy	Uterus		25
Healthy	Vagina		12

Table 2: Description of extracellular RNA sequencing datasets collected for analysis.

Project accession number	Publication	Sample type	Sequencing method	Condition	# donors	# samples
PRJNA543872	(Zhou et al., 2019)	Serum	Small input liquid volume extracellular RNAseq	Healthy	32	64
				Breast cancer	96	192
PRJNA454814	(Max et al., 2018)	Plasma and serum	Illumina	Healthy	13	156
PRJNA290097	(Yuan et al., 2016)	Plasma EVs	Illumina	Healthy	50	100
				Colorectal cancer	100	200
				Pancreatic cancer	6	12
				Prostate cancer	36	72
PRJEB24913	(Buschmann et al., 2018)	Serum EVs	Illumina	Healthy	10	49
				Sepsis/septic shock	9	36
https://github.com/alvinshi20/ExosomeData	(Shi et al., 2020)	Plasma EVs	Illumina	Melanoma	25	25
PRJNA589238	(Wang et al., 2020)	Plasma	Illumina	Healthy	6	6
				Lung cancer	6	6
PRJNA655240	(Sproviero et al., 2021)	Plasma EVs	Illumina	Healthy	12	24
				Alzheimer's disease	12	24
				Fronto-temporal dementia	18	36
				Parkinson's disease	18	36
				Amyotrophic lateral sclerosis	12	24

2.1.4 Metastatic KIRC samples from the ADAPTeR study

Tumour RNAseq data from the ADAPTeR study was analysed as detailed previously (Au et al., 2021). Of the 15 patients enrolled in this phase II, single-arm, open-label clinical trial of anti-programmed cell death protein 1 (PD-1) therapy (nivolumab) in treatment-naïve metastatic KIRC, 14 patients had RNAseq data of acceptable quality. A total of 60 primary tumour samples sequenced from these patients were used in this analysis, with 33 pre-treatment samples and 27 post-treatment samples.

2.1.5 Purified immune cell datasets

RNAseq samples of purified immune cell subsets were downloaded from publicly available sources GSE60424 (Linsley et al., 2014) and E-MTAB-8208 (Kazachenka et al., 2019). Immune cells were sorted from peripheral blood and bone marrow aspirates respectively.

2.1.6 Expanded KIRC tissue dataset

All KIRC samples were downloaded from TCGA alongside any adjacent healthy kidney data, totalling 538 KIRC samples and 72 healthy samples (downloaded August 2021). The corresponding clinical data was also downloaded. Samples were not filtered by the VHL status.

2.1.7 Renal carcinoma cell line dataset

RNAseq data from a renal cell carcinoma cell line RCC4 with the Von Hippel–Lindau tumour suppressor protein (VHL) stably transfected ($RCC4^{VHL+}$) was downloaded from PRJNA494827 (Smythies et al., 2019). RCC4 lines have lost function of VHL which would usually allow response to hypoxia by guiding degradation of constitutively expressed HIFs under normoxic conditions. In order to reintroduce the response to hypoxia (instead of the cell continuously perceiving hypoxia) functional VHL is added back. Three $RCC4^{VHL+}$ samples were cultured

in normoxic conditions, and three samples were cultured in hypoxic conditions of 1% oxygen (O_2) for 24 hours before RNA was extracted.

2.2 RNAseq processing

2.2.1 Annotation of HERV loci

Previous studies had analysed expression of 66 (Mayer et al., 2011; Panda et al., 2018; Rooney et al., 2015) and 3173 (Smith et al., 2018; Vargiu et al., 2016) HERV loci with regards to cytotoxic T-cell signatures and response to immune checkpoint blockade in KIRC. In order to analyse the expression of these HERV loci in the context of GRCh38, the 66 loci (Mayer et al., 2011) sequences had chromosomal coordinates identified using the basic local alignment search tool for nucleotide sequences (BLASTn) where the match with the greatest homology over the greatest length within GRCh38 was taken. For the 3173 HERV loci (Vargiu et al., 2016), where chromosomal coordinates for GRCh37 were given, the Lift Genome Annotations tool from the University of California, Santa Cruz (UCSC) (https://genome.ucsc.edu/cgi-bin/hgLiftOver) was used to convert to GRCh38 coordinates.

Previously, a custom annotation derived from the Dfam 2.0 library (version 150923) for GRCh38 was created (Attig et al., 2017). Here regions of the same provirus had been merged for quantification of whole locus expression. The GRCh38 coordinates for the 66 (Mayer et al., 2011) and 3173 (Vargiu et al., 2016) HERV loci were compared to this custom annotation to find overlapping loci. Loci from the Dfam-derived library had to begin, end, or be fully contained within the previously annotated loci to be considered a match (with a 5 nucleotide buffer added to either end of the locus). For further analysis, the list of LTR-containing elements identified in the Dfam-derived library were used.

2.2.2 Tissue and cell line RNAseq processing

2.2.2.1 Expression of transcripts assembled in the de novo transcriptome

To quantify expression of transcripts in the *de novo* transcriptome assembly Salmon was used as previously described (Attig et al., 2019). GNU parallel (Tange, 2023) was used to submit jobs. Binary alignment map (BAM) files were first converted to fastq format using SAMtools (version 1.3.1 (Danecek et al., 2021)) before Trimmomatic (version 0.36 (Bolger et al., 2014)) was used to remove poor quality sequences and adapters. FastQC (version 0.11.8) was used to check the quality of trimmed fastq files. Salmon (version 0.8.2) was used to quasi-map reads to GRCh38 and to quantify expression of transcripts.

2.2.2.2 Expression of individual RTE loci

Hisat2 (version 2.1.0) was used to align reads to GRCh38 and SAMtools (version 1.3.1,(Danecek et al., 2021)) was used to convert the output to BAM files. The featureCounts function from Subread (Liao et al., 2014) (version 1.5.0, with parameters -p -C -B -f -T 2 --primary), was used to measure expression of individual RTE loci, with multi-mapping reads randomly assigned.

2.2.3 Extracellular RNAseq processing

RNAseq data from each study were uniformly processed. GNU parallel (Tange, 2023) was used to submit jobs in groups, Trimmomatic (version 0.36 (Bolger et al., 2014)) was used to remove adapters and poor quality sequences, as well as removing reads of less than 35 nucleotides. FastQC (version 0.11.8) was used to check the quality of trimmed fastq files. Sed was used to convert fastq files to fasta format, before the BLAST-like alignment tool (BLAT, version 37x1) (Kent, 2002) was used to align the reads to the cancer-specific transcripts, and control sequences with 100% identity, and align reads to the *LINE1HS* and *AluSp* consensus sequences with 90% identity.

```
blat transcriptAndControl.fasta \
trimmed.fasta \
-q=rnax -t=dnax -minMatch=1 \
-minScore=0  tileSize=3 -out=psl -minIdentity=[90|100] \
-dots=100000 \
BLATout.psl
```

Outputs were then filtered to ensure the match length equalled the read length using awk. Prior to analysis of read alignment, donors with fewer than 20000 reads surviving trimming were binned, this removed five healthy donors and five sepsis donors from the study PRJEB24913. Multimapping reads were counted once per patient. To normalise the data, reads aligned were expressed as a percentage of the reads surviving trimming.

2.2.3.1 Analysis of spliced reads

MATLAB (version R2022b, The MathWorks) was used to filter BLAT outputs to identify reads over splice junctions in exRNA data, ensuring reads overlapped at least the splice point and one nucleotide either side. Splice junctions at known exon boundaries in the housekeeping genes *beta-actin* (*ACTB NM_001101.5_1*) and *glyceraldehyde-3-phosphate dehydrogenase* (*GAPDH NM_002046.7_5*) were analysed (2.3.2: Selecting transcripts for use in RNA liquid biopsies).
Reads overlapping splice sites in TCGA BRCA data were counted from BAM files downloaded directly from TCGA. SAMtools (version 1.3.1 (Danecek et al., 2021)) was used to convert the BAM files to the sequence alignment map (SAM) format followed by awk to filter the file for reads aligned to chromosomes 7 and 12 (where control sequences were located) to reduce processing time. BEDOPS (version 2.4.20 (Neph et al., 2012)) was used to convert the filtered SAM file to a BED file. Then the BEDTools (version 2.30.0 (Quinlan and Hall, 2010)) function intersect was used to find reads which overlapped the exon junctions.

2.3 Selection of cancer-specific transcripts

2.3.1 Selecting the cancer-specific transcriptome

Cancer specific transcripts were selected by comparing expression of each transcript in each cancer to the expression in the respective healthy tissue where available, as well as expression in all other healthy tissues individually (Attig et al., 2019). These filters were defined previously (Attig et al., 2019). To be defined as cancer specific each transcript had to have a) expression of less than 10 transcripts per million (TPM) in at least 90% of all healthy samples, b) expression of at least 1 TPM in 25% of the respective patients, c) median cancer expression of at least three times median expression of each healthy tissue, and d) median cancer expression of at least three times the 90th percentile of the respective healthy tissue.

2.3.2 Selecting transcripts for use in RNA liquid biopsies

From the list of cancer-specific transcripts selected above (2.3.1) transcripts with median expression of more than 0.5 TPM in each healthy tissue were removed. Followed by manual inspection of the specificity of transcripts across cancer and healthy samples.

Housekeeping genes *ACTB (NM_001101.5_1*) and *GAPDH (NM_002046.7_5*) were also aligned to. As well as the *Homo sapiens LINE1* (*LINE1HS*) and the *AluSp* consensus sequences.

2.3.3 Selection of transmembrane domain containing candidates

The list of cancer-specific transcripts selected above (2.3.1) were further filtered by removing all transcripts with any median healthy tissue expression greater than 0.5 TPM. The function orf_scanner (Young and Attig, 2019) was used to find all possible open reading frames in both directions containing at least 100 codons (including the stop codon). The ORFs were translated within orf_scanner, and the probability the sequence contained a transmembrane domain was predicted

by TMHMM (version 1.0 (Krogh et al., 2001)). Manual inspection of expression of transcripts predicted to contain at least one open reading frame encoding at least one transmembrane domain was carried out to ensure low levels of expression across healthy tissues. Boxplots of expression across the original TCGA and GTEx dataset (2.1.1) were plotted in MATLAB (version R2022b, The MathWorks) and transcripts with many high-expressing outliers across healthy tissues were removed. The structures of transcripts with highly cancer-specific expression were also manually inspected to ensure confidence in the assembly, RNAseq alignment data was visualised on IGV (Robinson et al., 2011) using BAM files directly downloaded from TCGA which had been aligned to GRCh38. For confidence in the alignment all splice sites and exons had to be well covered in the respective cancer samples. The selected ORFs were then checked for homology with known proteins using BLAST for peptide sequences (BLASTp) filtering for hits with at least 85% homology, and the position and direction of the peptide sequence was checked to ensure alignment with the directionality of splice sites within the transcript.

2.3.4 Selection of transcripts upregulated in KIRC

From the 32264 cancer-specific transcripts described above (2.3.1), kidney cancer (clear cell or papillary) upregulated transcripts were identified by selecting transcripts with mean expression of at least 0.5 TPM in either patient group giving 8135 transcripts. Data from TCGA KIRC and TCGA adjacent healthy kidney were then compared to find transcripts significantly upregulated in KIRC using the Qlucore Omics Explorer (www.qlucore.com) with help from Prof. George Kassiotis (The Francis Crick Institute). Using linear expression values, 1914 of the 8135 transcripts were significantly upregulated in KIRC (fold change ≥ 2, q = 0.05). Using log2 values with a cut-off of 0.1, 3200 of 8135 transcripts were significantly upregulated in KIRC (fold change ≥ 2, q = 0.05). Taking the union of the two lists, 3681 transcripts were selected as overexpressed in KIRC compared to both adjacent kidney tissue and other healthy tissues, though many transcripts may also be expressed in other cancers.

2.4 Statistical analysis and plotting

2.4.1 RNAseq alignment

RNAseq alignment data was visualised on IGV (Robinson et al., 2011) using BAM files directly downloaded from TCGA which had been aligned to GRCh38. Screenshots were taken to be included in figures.

2.4.2 Venn diagrams

Venn diagrams were made using Python (Van et al., 2009) packages matplotlib (Hunter, 2007), numpy (Harris et al., 2020), and matplotlib_venn (Tretyakov, 2024), and were labelled in Microsoft PowerPoint.

2.4.3 Enrichment of repeat types

Enrichment analysis of repeat types was performed using Fisher's Exact test in MATLAB (version R2022b, TheMathWorks), followed by the Bonferroni-Holm method to correct for multiple testing (Groppe, 2010).

2.4.4 Differential expression analysis for HERVs in metastatic KIRC

Differential expression analysis on $\log_2$ transformed data (with an expression cut off of 0.05) and heatmap plotting was performed on Qlucore Omics Explorer (www.qlucore.com) with help from Prof. George Kassiotis (The Francis Crick Institute).

2.4.5 Heatmaps

Heatmaps were plotted using Qlucore Omics Explorer (www.qlucore.com) with help from Prof. George Kassiotis (The Francis Crick Institute). Values were log2 normalised using a cut-off of 0.05, and were further normalised between samples using a z-score. The heatmaps show the relative expression of RTE loci or transcripts. Variables were ordered using the hierarchical clustering function.

2.4.6 Correlation with the hypoxia score

The hypoxia scores for all TCGA samples were kindly sent by Prof. David Mole (Nuffield Department of Medicine, University of Oxford) (Lombardi et al., 2022). The hypoxia scores for KIRC samples were filtered, however mismatches in sample names due to updates by TCGA to the naming system of files meant the mean hypoxia score per patient was used here. Of the 479 KIRC patients with hypoxia scores calculated for tumour samples, 406 had hypoxia scores for one sample, 70 for two samples, and three patients had the mean hypoxia score calculated from three samples.

2.4.7 Survival analysis

Overall survival time was downloaded alongside immune subtype annotations (which were not used) (Thorsson et al., 2018), 515 KIRC patients had available data. Survival analysis was run using R and RStudio (version 2023.12.0 Build 369). The surv_cutpoint function of the survminer package (version 0.4.9) was used to calculate the best expression cut-off for each transcript ensuring a minimum of 20% of the patients per group. Of the 3861 transcripts 17 failed this step. Then the survfit and coxph functions of the survival package (version 3.5-7) were used to fit a Cox proportional hazards regression model to the data and calculate the hazard ratio (HR), 95% confidence interval (CI_{95}) and *p-value*. The survival curve was visualised using the simple survival analysis (Kaplan-Meier) option on GraphPad Prism (version 10.0.2).

For multivariate analysis, association of the transcript expression with other clinical variables potentially affecting survival was tested using analysis of variance (ANOVA) or the students t-test on MATLAB (version R2022b, The MathWorks) with a significance threshold of $p \leq 0.05$. Variables tested were: age at diagnosis, race, gender, prior treatment, prior malignancy, AJCC pathologic stage, and presence of treatment or therapy. Multivariate analysis was run using the coxph function of the survival package (version 3.5-7) incorporating variables associated with the expression of the selected transcript. The function ggforest

of the survminer package (version 0.4.9) was used to visualise the multivariate analysis.

2.4.8 Other plotting

All other data was plotted using MATLAB (version R2022b, The MathWorks) and GraphPad Prism (version 10.0.2). Pearson's correlation analysis was performed in MATLAB, whilst other statistical tests were performed in GraphPad Prism. Figures were prepared using Microsoft PowerPoint.

2.5 Preparation of stably transduced cell lines

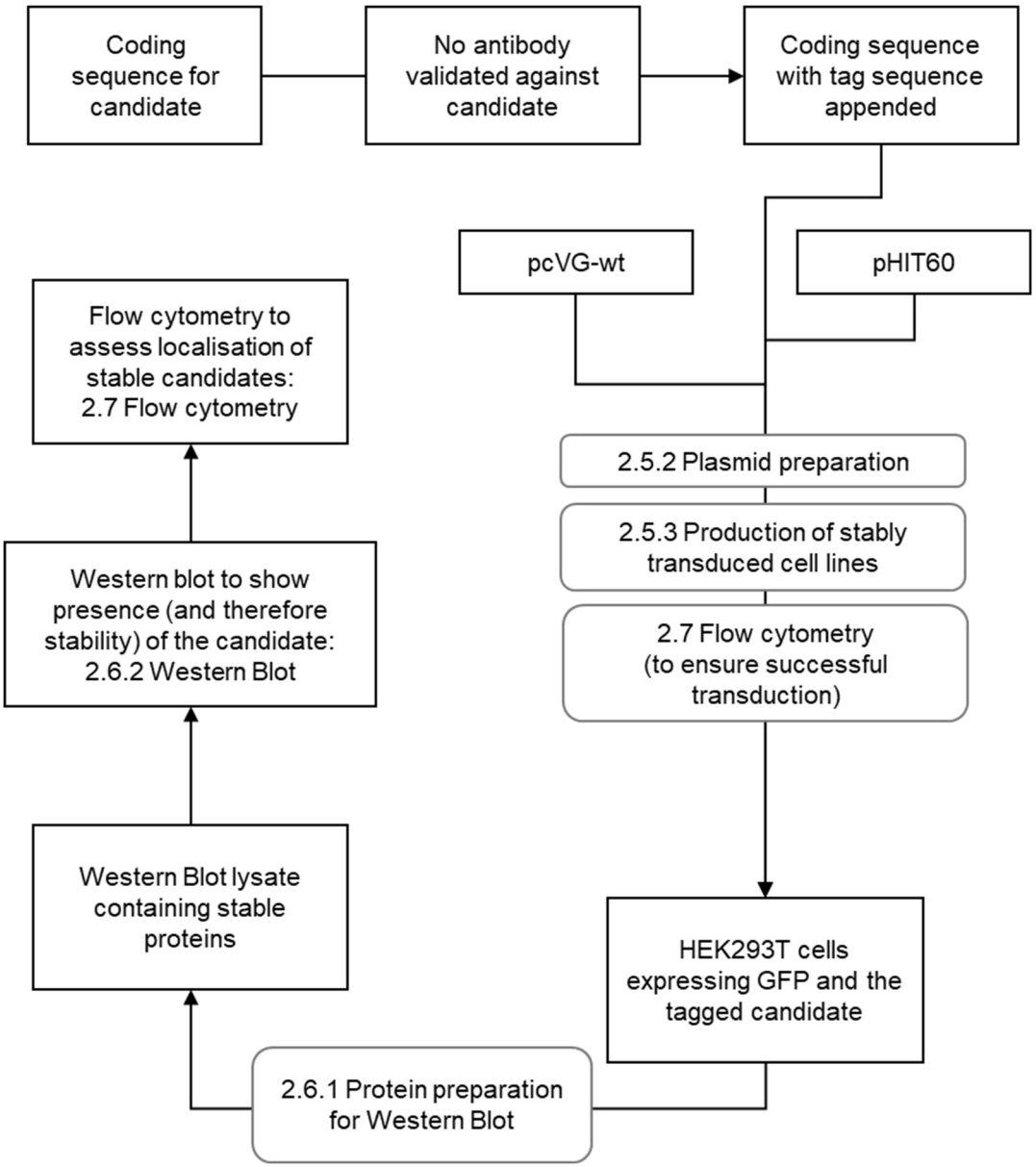


Figure 5: An overview of laboratory work carried out. Methods are referenced in grey boxes. The tag sequence appended to the coding sequence of candidates allowed for specific antibody targeting of the candidate protein without a validated antibody for the protein itself.

2.5.1 Cell culture

HEK293T cells were used to test the stability of various candidate proteins. HEK293T cells are a derivative of the HEK293 immortalised human embryonic kidney cells (Graham et al., 1977). HEK293T cells additionally contain the SV40 large T antigen allowing use of the SV40 promoter (DuBridge et al., 1987).
HEK293T cells were cultured in Iscove's Modified Dulbecco's Medium (IMDM, Sigma-Aldrich, I3390) supplemented with 5% foetal calf serum (FCS, Thermo Fisher Scientific), penicillin-streptomycin solution (Millipore-sigma, P4333-100ML, penicillin at 100 U/mL, streptomycin at 0.1 mg/mL), and L-glutamine (2 mM, Merck, G7513-100ML), and cultured at 95% humidity, 37°C, and 5% carbon dioxide (CO_2). All cells were sourced from and verified as mycoplasma free by the Cell Services facility at The Francis Crick Institute.

2.5.2 Plasmid preparation

Plasmids were amplified using One Shot™ TOP10 competent *Escherichia coli* (Thermo Fisher Scientific, C404003). The bacteria stored at -80°C were thawed on ice, before 0.5 µL of plasmid DNA at 400 ng/µL was added to 25 µL of bacteria cells. The mixture was incubated on ice for 30 minutes, and heat shocked at 42°C for 30 seconds. After a further 5-minute incubation on ice, 350 µL of super optimal broth with catabolite repression (S.O.C) medium (Invitrogen, 15544034) was added, and the solution was shaken at 200 revolutions per minute (RPM) for 1 hour at 37°C. From this solution, 100 µL was then spread onto a pre-warmed ampicillin plate before overnight incubation at 37°C. The next day single colonies were selected and added to 5 mL Luria-Bertani media (LB media, Media Team, The Francis Crick Institute) supplemented with 100 µg/mL ampicillin (Sigma-Aldrich, A5354), this was shaken at 200 RPM for 6 hours at 37°C, and then transferred to 50 mL LB media and placed back into the shaker overnight. Plasmid DNA was then extracted using the Plasmid Plus Maxi Kit (Qiagen, 12963). Aliquots of 50 mL of bacterial culture were spun at 3500 RPM for 20 minutes at 4°C before resuspension in 8 mL of resuspension buffer. To this, 8 mL of lysis buffer was added and the solution inverted until viscous followed by a 3-

minute incubation at room temperature. 8 mL of neutralisation buffer was added and the solution was inverted 6 times before the samples were spun at 3500 RPM for 10 minutes at 4^{o}C. The plasmid containing supernatant was column purified using a vacuum pump, washed, and then eluted using nuclease-free water. The plasmid DNA concentration was measured using a NanoDrop spectrophotometer (Thermo Fisher Scientific).

2.5.3 Production of stably transduced cell lines

Stably transduced cell lines were produced through viral infection and, if required, single cell sorting on green fluorescent protein (GFP) using the MoFLO XDP cell sorter (BD Biosciences, Flow Cytometry Team, The Francis Crick Institute). Virus was generated using HEK293T cells plated at a density of 1.5x10^{6} cells per 60 mm well. Cells were plated the day before transfection in 5 mL of media and allowed to settle overnight. Prior to transfection the media was changed and the transfection solution was added dropwise to the well. The transfection solution contained per 60 mm well, 280 µL serum-free IMDM (Sigma-Aldrich, I3390) and 30 µL GeneJuice (VWR International Ltd., 70967-4) which were mixed, vortexed, and incubated at room temperature for 5 minutes. To this 5 µL of DNA at a concentration of 1 µg/µL was added dropwise before a 15-minute room temperature incubation. The DNA comprised of an equal mix of vesicular stomatitis virus glycoprotein (VSVg) plasmid (pcVG-wt) and pHIT60, both kindly provided by Dr. Jonathan Stoye (The Francis Crick Institute, London, UK), as well as the open reading frames of the sequences of interest cloned into the pRV-IRES-GFP vector (Table 3, Figure 6). Cloning the open reading frames into the vector was carried out Genewiz LLC, and was followed by sequencing to verify the plasmid structure. The open reading frame consisted of the coding sequence of the given target and either a FLAG tag or three HA tags. The supernatant containing the virus was collected three days after transfection and stored at -80^{o}C until use. 2 mL of viral stock with 4 µg/mL of polybrene (Sigma-Aldrich, TR-1003-G) was added to HEK293T cells, plated at a density of 8.5x10^{4} cells per 35 mm well and spun at 1200 RPM for 45 minutes. After three days, in the case of

the ENPP3 and truncated ENPP3 transduced cells, populations were single cell sorted on GFP expression using a BD FACSAria II (BD Biosciences) (Flow Cytometry STP, The Francis Crick Institute).

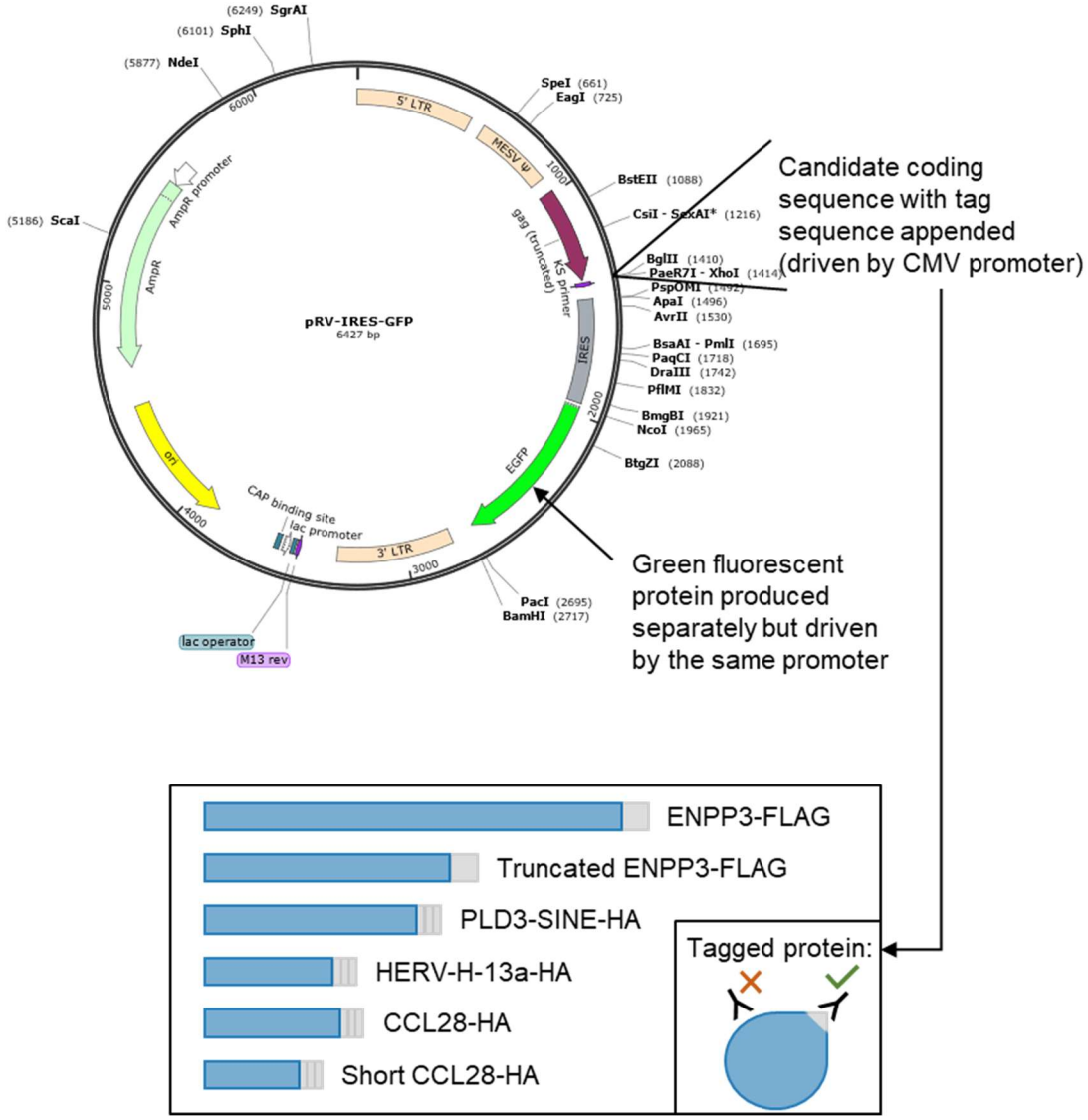


Figure 6: Map of the pRV-IRES-GFP vector. Overviews of the coding sequences and tags are shown below the map. Adding tags to the coding sequences allows the proteins to be labelled using validated antibodies as there was no validated antibody available for the novel proteins. (IRES: internal ribosome entry site; eGFP: green fluorescent protein; CMV: cytomegalovirus)

Table 3: List of plasmids used for production of stably transfected lines

Plasmid	Description	Reasoning
pcVG-wt	Contains VSVg membrane fusion protein	Packaged separately to prevent virus from becoming self-replicating
pHIT60	Contains the gag-pol sequence	
pRV-IRES-GFP	Contains coding sequence of candidate protein, alongside an internal ribosome entry site and the GFP coding sequence.	Allows production of candidate protein with a tag and successfully transduced cells to be sorted by presence of GFP

2.6 Sample preparation and Western Blot

2.6.1 Protein preparation for Western Blot

Cells were washed twice with phosphate-buffered saline (PBS) stored at 4°C before being incubated on ice with radioimmunoprecipitation assay (RIPA, Sigma-Aldrich, R0278-50ML) buffer for 30 minutes to lyse the cells. The mixture was then spun at 14000 RPM for 10 minutes at 4°C. The protein concentration of the lysate was measured using the Pierce™ BCA protein assay kit (Thermo Scientific, 23225). Stock solutions at a protein concentration of 500 µg/mL were made by mixing 100 µL of sample buffer (Laemmli 2x concentrate, Sigma-Aldrich, S3401-10VL), with 100 µg of protein lysate and RIPA buffer to a final volume of 200 µL. Stock solutions were heat denatured at 95°C for 5 minutes before being frozen at -20°C.

2.6.2 Western Blot

Sample stock solutions were thawed on ice before being boiled at 95°C for 5 minutes. 10 µg of protein per sample was loaded into a 4–20% Mini-PROTEAN® TGX™ precast polyacrylamide gel (Bio-Rad, 4561094) alongside a protein ladder (PageRuler Plus Prestained Protein Ladder, 10 kDa to 250 kDa, ThermoFisher, 26619). The gel electrophoresis was run in a Mini-PROTEAN® Tetra Vertical Electrophoresis Cell (Bio-Rad) filled with protein running buffer (Media Team, The

Francis Crick Institute) at 180 V for 40 minutes. Samples were transferred to a 0.2 µm nitrocellulose membrane (Trans-Blot Turbo Mini 0.2 µm Nitrocellulose Transfer Pack, Bio-Rad, 1704158) using the Trans-Blot Turbo dry transfer system (Bio-Rad, 1704150) turbo setting for mini TGX gels before blocking with 5% skimmed milk (Marvel) in Tris-buffered saline with 0.5% Tween-20 (TBS-T, Media Team, The Francis Crick Institute) for 1 hour. Membranes were stained overnight at 4°C with the primary antibody (Table 4) diluted in 5% skimmed milk in TBS-T. Membranes were washed for 15 minutes four times in TBS-T at room temperature before the horseradish peroxidase (HRP)-conjugated secondary antibody (Table 4) was added, diluted in 3% skimmed milk in TBS-T and incubated at room temperature for 1 hour. Membranes were then washed for 10 minutes three times in TBS-T and visualised by enhanced chemiluminescence using Clarity™ Western ECL Substrate (Bio-Rad, 1705060) on a ChemiDoc XRS+ (Bio-Rad). Adobe Illustrator (v27.5, 64-bit) and Microsoft PowerPoint were used for labelling Western blot images.

2.7 Flow cytometry

2.7.1 Sample preparation

Extracellular and intracellular staining was done in parallel in v-bottomed 96-well plates. $5x10^5$ cells per well suspended in media were added before the plate was centrifuged at 1200 RPM for 5 minutes at 4°C. Media was flicked off and 100 µL of either flow cytometry staining buffer (FACS buffer, PBS, 2% FCS, 0.1% Azide) for extracellular staining or fixation buffer (3:1 fixation diluent (Invitrogen, 00-5223-56) and fixation concentrate (Invitrogen, 00-5123-43)) for intracellular staining was added to each well, followed by a 20-minute incubation at 4°C. After incubation an additional 100 µL of either FACS buffer for extracellular staining or permeabilization buffer (Invitrogen, 00-8333-56) for intracellular staining was added to each well and the plate was centrifuged at 1200 RPM for 5 minutes at 4°C. Buffers were flicked off and antibodies (Table 5) diluted in 200 µL of FACS or permeabilization buffer were added to each well for extracellular and

intracellular staining respectively. The plate was incubated at 4°C for 30 minutes, before washing the samples twice with 200 µL of either FACS or permeabilization buffer and centrifuging at 1200 RPM at 4°C for 5 minutes each. The secondary antibody (Table 5) was then added, again diluted in 200 µL of either FACS buffer for extracellular or permeabilization buffer for intracellular staining. The plate was incubated at 4°C for 30 minutes, before samples were washed twice with 200 µL of FACS or permeabilization buffer and centrifuged at 1200 RPM at 4°C for 5 minutes each. All samples were then resuspended in 100 µL FACS buffer and filtered through a 100 µm mesh sieve.

2.7.2 Sample and data analysis

Samples were analysed on an LRS Fortessa (BD Biosciences) and output data were analysed using BD FACSDiva v8.0 and FlowJo v10.8.1 (BD Life Sciences).

Table 4: List of Western blot antibodies.

Target	Clone	Species	Conjugation	Dilution	Source
FLAG	M2	Mouse		1:1000	Sigma-Aldrich (F1804)
Mouse IgG	Polyclonal	Rabbit	HRP	1:2000	Abcam (ab6728)
HA	3F10	Rat		1:200	Roche (ROAHAHA)
Rat IgG	Polyclonal	Goat	HRP	1:1000	Cell Signalling (7077S)
Vinculin	Polyclonal	Rabbit		1:1000	Cell Signalling (4650S)
β-actin	AC-15	Mouse	HRP	1:1000	Abcam (ab49900)

Table 5: List of flow cytometry antibodies.

Target	Clone	Species	Conjugation	Dilution	Source
HA	3F10	Rat		1:50	Roche (ROAHAHA)
Rat IgG	Polyclonal	Goat	Alexa Fluor® 594	1:1000	Abcam (ab150160)

Chapter 3. Results 1: Identification of cancer-specific transcripts

3.1 Aims

From the previously assembled *de novo* transcriptome cancer-specific transcripts overlapping LTRs have been characterised in order to further understanding of the role of LTR elements in cancer biology (Attig et al., 2019). To create a more comprehensive list of transcripts overlapping other RTEs, or no RTE, all potential cancer-specific transcripts were identified from those assembled in the *de novo* transcriptome (2.3.1: Selecting the cancer-specific transcriptome; Figure 7). This list of 32264 transcripts (Figure 7) will be the focus of work presented in this thesis (Figure 4, Figure 7).

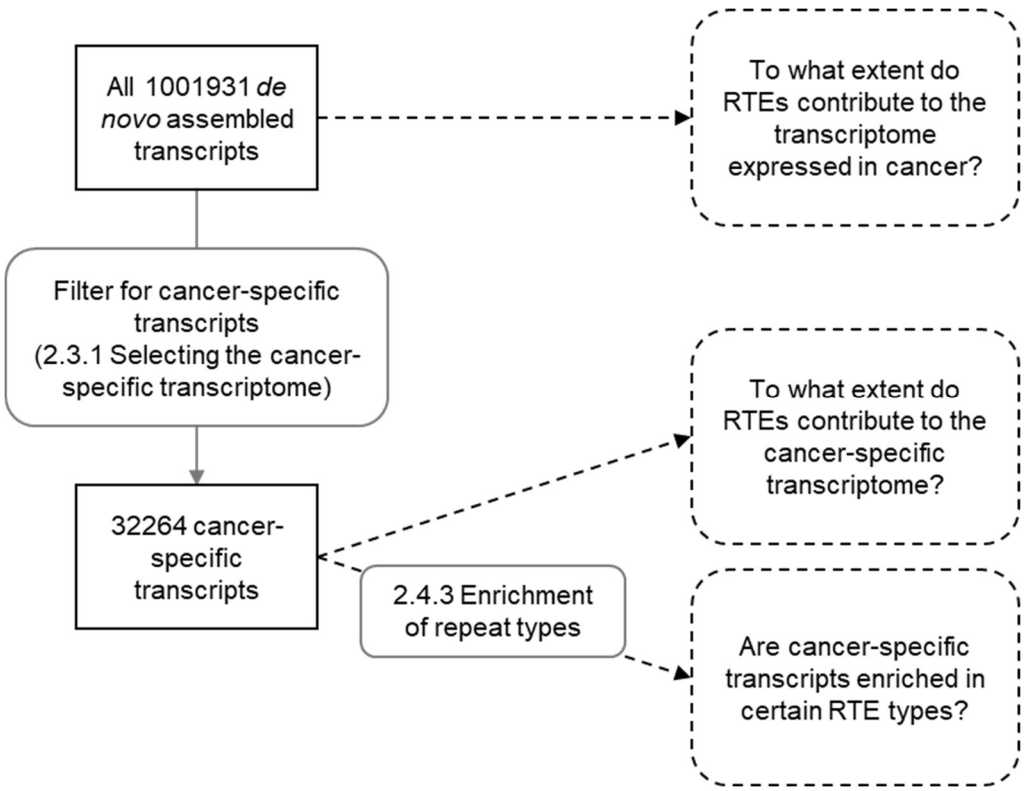


Figure 7: Aims for Results 1: Identification of cancer-specific transcripts. Aims are shown in dashed boxes and methods are referenced in grey boxes.

3.2 Introduction

RTEs contain control elements able to influence the transcriptome (1.4: RTEs and the non-cancerous transcriptome, 1.5.2: Cancer-specific control of the transcriptome). In cancer, the hypomethylated genome unleashes the potential effects of RTEs on the transcriptome (1.2: Control of RTE expression, 1.5: Expression of RTEs in cancer) producing cancer-specific effects not seen in healthy tissues. A *de novo* transcriptome has previously been built (Attig et al., 2019) (1.6: A *de novo* transcriptome assembly) revealing the effects of RTEs on the cancer transcriptome.

3.3 Results

3.3.1 Sequences represented by the *de novo* transcriptome assembly

In order to understand the extent to which RTEs contribute to the transcriptome expressed in cancer, annotation of the *de novo* transcriptome assembly was analysed. The *de novo* transcriptome assembly (Attig et al., 2019) represents transcripts of both known and novel sequence, overlapping both gene and RTE sequences (Figure 8). The *de novo* transcriptome assembly contains 1001931 transcripts, 75.6% (757566/1001931) of which overlap at least one RTE (Figure 8a). Of the 757566 transcripts that overlap RTE sequences, 98.0% (742748/757566) overlap at least one SINE, LINE, or HERV, though they may also overlap other repeat types (Figure 8a). At the time of annotation, 8.23% (82462/1001931) of the transcripts did not overlap a known gene or repeat (Figure 8a). The transcripts assembled were mainly multiexonic, though 23.05% were monoexonic (Figure 8b).

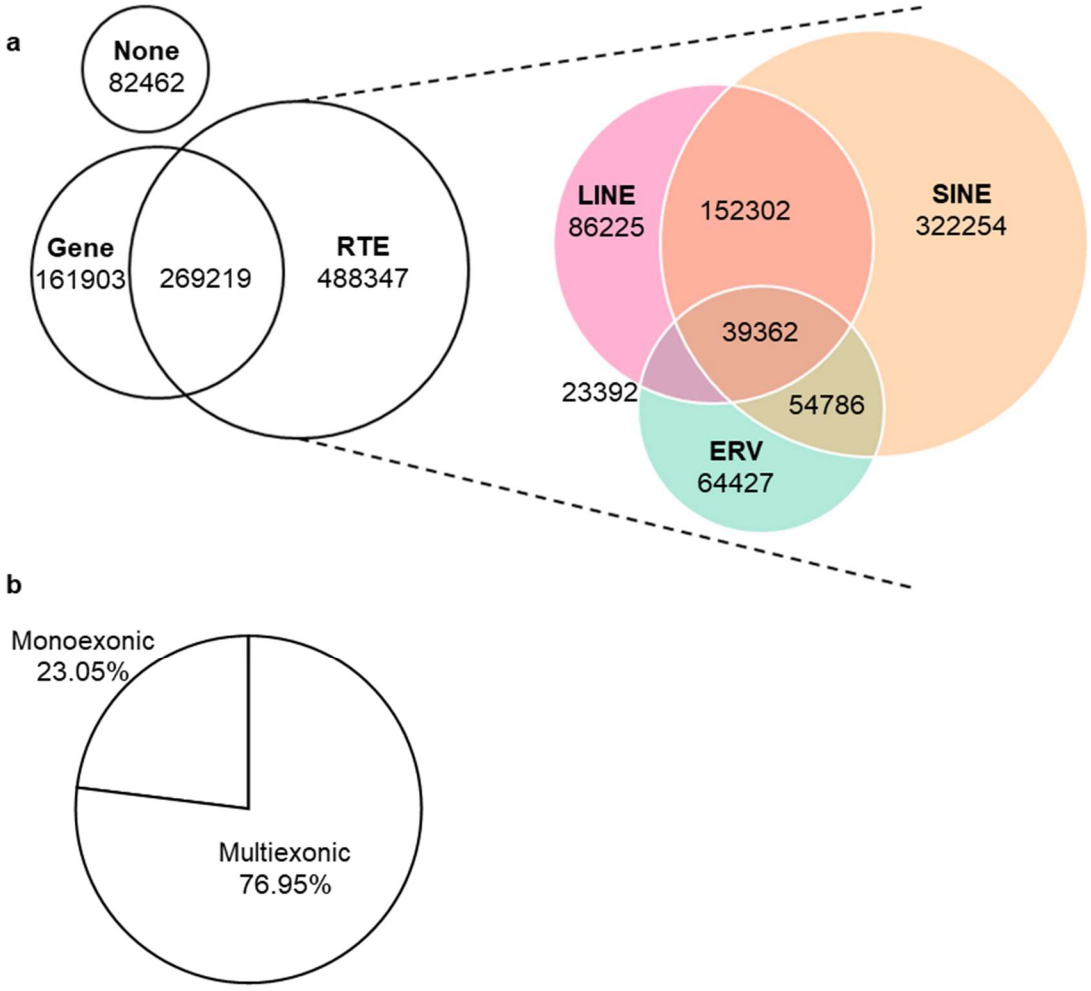


Figure 8: Overview of the sequences represented by the *de novo* transcriptome assembly. a. The total number of *de novo* assembled transcripts overlapping gene and RTE sequences, alongside the numbers of transcripts overlapping the three most well-represented RTE groups. The number of transcripts overlapping neither known gene or RTE sequences are shown under "none". **b.** The proportion of monoexonic and multiexonic transcripts assembled.

3.3.2 Sequences represented by the cancer-specific transcriptome

To understand the contribution of RTEs to the cancer-specific transcriptome, cancer-specific transcripts were identified (2.3.1: Selecting the cancer-specific transcriptome). The cancer-specific transcriptome also represents known and novel transcripts, containing gene and RTE sequences (Figure 9a). From the 1001931 transcripts assembled, 32264 were selected as cancer-specific, some expressed in one cancer and others across cancer types. These transcripts represent a range of known and novel sequences, the majority of which overlap at least one RTE (Figure 9a). Of the 32264 transcripts, 1245 do not overlap a known gene or repeat at the time of annotation. Only 1.51% of all transcripts overlapping neither a gene or repeat (1245/82462) were selected as cancer-specific, compared to 3.37% of transcripts overlapping either a gene, repeat, or both being selected as cancer-specific (31019/919469). This increased selection of transcripts overlapping some known sequence may be because fully novel transcripts have less expression in cancer, or because these transcripts are less likely to be assembled correctly. Of the 32264 transcripts 23217 overlap at least one RTE, with 95.90% (22265/23217) of these overlapping at least one SINE, LINE, or HERV though they may also overlap other repeat types (Figure 9a). Of the 32264 transcripts 17648 overlap known genes, representing sequences from 11115 unique genes. The majority of the transcripts are multiexonic (Figure 9b). When compared to the entire *de novo* transcriptome assembly (2.4.3: Enrichment of repeat types), the cancer-specific transcripts are enriched in specific subtypes of RTEs. Although most of these are HERVs, there was also enrichment for SVA, MaLR, SINE, and LINE subtypes (Figure 9c).

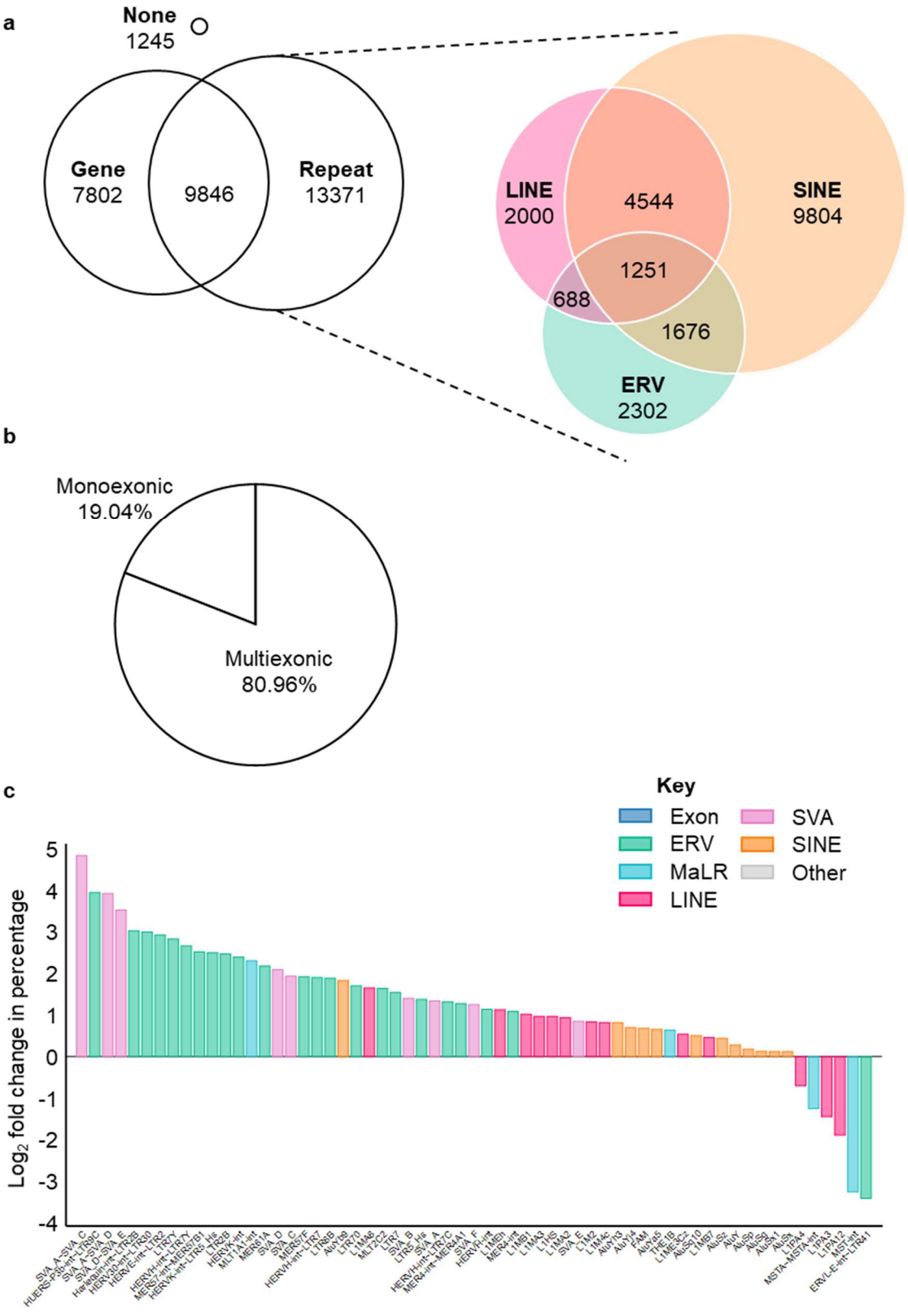

a
None
1245
Gene
7802
9846
Repeat
13371
LINE
2000
4544
SINE
9804
1251
688
1676
ERV
2302
b
Monoexonic
19.04%
Multiexonic
80.96%
c
Key
Exon
ERV
MaLR
LINE
SVA
SINE
Other
Log2 fold change in percentage
5
4
3
2
1
0
-1
-2
-3
-4

Figure 9: Overview of the sequences represented by the 32264 cancer-specific transcripts. a. The total number of cancer-specific transcripts overlapping gene and RTE sequences, alongside the numbers of transcripts overlapping the three most well-represented RTE groups. The number of transcripts overlapping neither known gene or RTE sequences are shown under “none”. **b.** The proportion of monoexonic and multiexonic transcripts selected as cancer-specific. **c.** RTE subtypes with significant enrichment when compared to the whole *de novo* transcriptome assembly (as described in 2.4.3: Enrichment of repeat types).

3.4 Discussion

Other work has also identified RTE-containing transcripts, in both cancer and human pluripotent stem cells. Using a combination of long and short read sequencing, transcripts expressed in human pluripotent stem cells were identified, with 65% of non-coding and 26% of coding transcripts overlapping at least one RTE (Babarinde et al., 2021). These RTE-containing transcripts were generally expressed at lower levels than non-RTE containing isoforms over the same gene. The incorporation of RTEs into novel isoforms overlapping known genes was likely to disrupt open reading frames in coding sequences, which in some instances led to HERV and LINE1 derived peptides (Babarinde et al., 2021). Additionally, to identify potential antigenic cancer-specific transcripts driven by RTEs, a transcriptome assembly was built using pan-cancer samples from TCGA (Shah et al., 2023). In agreement with the transcriptome assembly analysed in this thesis, some genes had multiple isoforms overlapping RTEs, with enrichment for overlapping ERV and SVA RTEs (Shah et al., 2023). Peptide production from specific candidates was also confirmed through whole cell mass spectrometry and immunopeptidomics (Shah et al., 2023).

Isoforms specific to cancer may contain canonically intronic regions. Increased intron retention is seen across cancer cell lines and patient samples, and may be due to mutations in splice machinery. Though abnormal splicing in samples has also been seen in the absence of splicing machinery mutants (Dvinge and Bradley, 2015). Some introns, termed “exitrons”, are included due to weaker splice signals at their boundaries, often do not contain stop codons, and are in frame with other gene exons. Exitron-containing isoforms of a given gene have been shown to be expressed at higher levels than isoforms with intron retention, and in some cases may be the dominant isoform. This additional inclusion of sequence may still disrupt canonical open reading frames though, as in the cases of the forkhead box protein O4 (FOXO4) and Msx2-interacting protein (SPEN) tumour suppressors which were found to be truncated in tumour samples due to exitron inclusion (Wang et al., 2021).

Expression of some transcripts may be artefactual. Monoexonic transcripts identified here, and in other transcriptome assemblies, may be derived from real monoexonic transcripts, or may be due to the aligner being unable to assemble the contig fully due to gaps in the sequence created by either damage during RNA preparation or differences in the reference and sample genomes (Babarinde et al., 2021). It is possible the monoexonic transcript may also be a spliced-out intron sequenced prior to degradation (Babarinde et al., 2021). Furthermore, artefacts from the reverse transcription step of library preparation for sequencing can introduce false introns, or “falsitrons” (Schulz et al., 2021). These falsitrons are created by the reverse transcriptase skipping regions of sequence contained within hairpin loops, these loops are often formed by inverted Alu repeats. Additionally, as the *de novo* transcriptome assembly analysed here was created in an un-stranded way, some transcripts represent artefactual merging of almost overlapping genes running in opposite directions. Due to these caveats, candidate transcripts were manually inspected before further analysis.

Quantifying the expression of RTEs is complex due to their repetitive nature and presence in introns. Approximately one third of human protein coding transcripts and three quarters of non-coding transcripts contain an exon derived from an RTE (Lanciano and Cristofari, 2020). Although poly(A) filtered RNAseq datasets reduce the measurement of intronic and pervasive transcription of RTEs (Lanciano and Cristofari, 2020) it is still difficult to ensure that transcription seen from an individual RTE locus is due to activation of that locus. Due to the repetitive nature of RTEs long reads are required to accurately map to a specific locus. Increasing the read length from 50 to 100 base pairs increases the amount of the genome which can be uniquely mapped to from 68% to 88% (Lanciano and Cristofari, 2020). However, some sequences overlapping Alu, LINE1, and tandem repeat regions cannot be uniquely mapped to even with synthetic read lengths of 1000 base pairs (Li et al., 2014). For the analysis in this thesis Salmon (Patro et al., 2017) has been used to quantify the expression of the *de novo* assembled transcripts. Salmon assigns multimapping reads based on a probabilistic model after aligning all uniquely mapping reads. This can lead to an

underestimation of transcripts with smaller proportions of unique areas (Storvall et al., 2013), but is still believed to be the most accurate method of quantifying repetitive elements (Lanciano and Cristofari, 2020).

Polymorphisms in the sample and reference genome leads to misalignment of reads, which would be especially problematic for polymorphic RTE insertions. RTE insertions differing from the current reference genome have been seen in healthy (Cao et al., 2020) and cancer patient samples (Bennett et al., 2004; Rodriguez-Martin et al., 2020). These polymorphic insertions led to changes in isoform proportions and overall expression of nearby genes (Cao et al., 2020) which may produce unusual gene expression patterns when RNAseq samples are aligned to the reference genome and polymorphic insertions are ignored. Integrations of LINE1s have also been seen to delete parts of chromosomes, knocking out tumour suppressor genes, and again producing unusual gene expression patterns when polymorphisms are not considered during RNAseq alignment (Rodriguez-Martin et al., 2020).

3.5 Conclusion

Overall, the *de novo* transcriptome previously assembled revealed the effects of RTE reactivation on the cancer transcriptome (Attig et al., 2019). The 32264 cancer-specific transcripts identified will serve as the basis for this thesis. Manual inspection of transcripts will be carried out where feasible due to the caveats of the assembly and of quantifying RTE expression.

Chapter 4. Results 2: Tumour-specific transcripts in extracellular RNA

4.1 Aims

Some RTE-overlapping transcripts are able to demark healthy and tumour tissues, as well as differentiate between tumour types (Attig et al., 2019). It is possible these transcripts are released into patient blood, with detection potentially useful as a biomarker of disease presence. In order to validate the use of these transcripts as a blood-based biomarker, the presence of BRCA tumour-specific transcripts was quantified in blood of healthy donors and patients bearing BRCA tumours. BRCA was selected due to the availability of data. In order to ensure other diseases would not confound the specificity of this detection, BRCA tumour-specific transcripts were also quantified in the blood of independent healthy donors, patients bearing other tumour types, and patients with non-cancerous diseases. Furthermore, samples were also analysed to assess whether the liquid biopsy RNAseq data quality confounded the detection of tumour-specific transcripts (Figure 10).

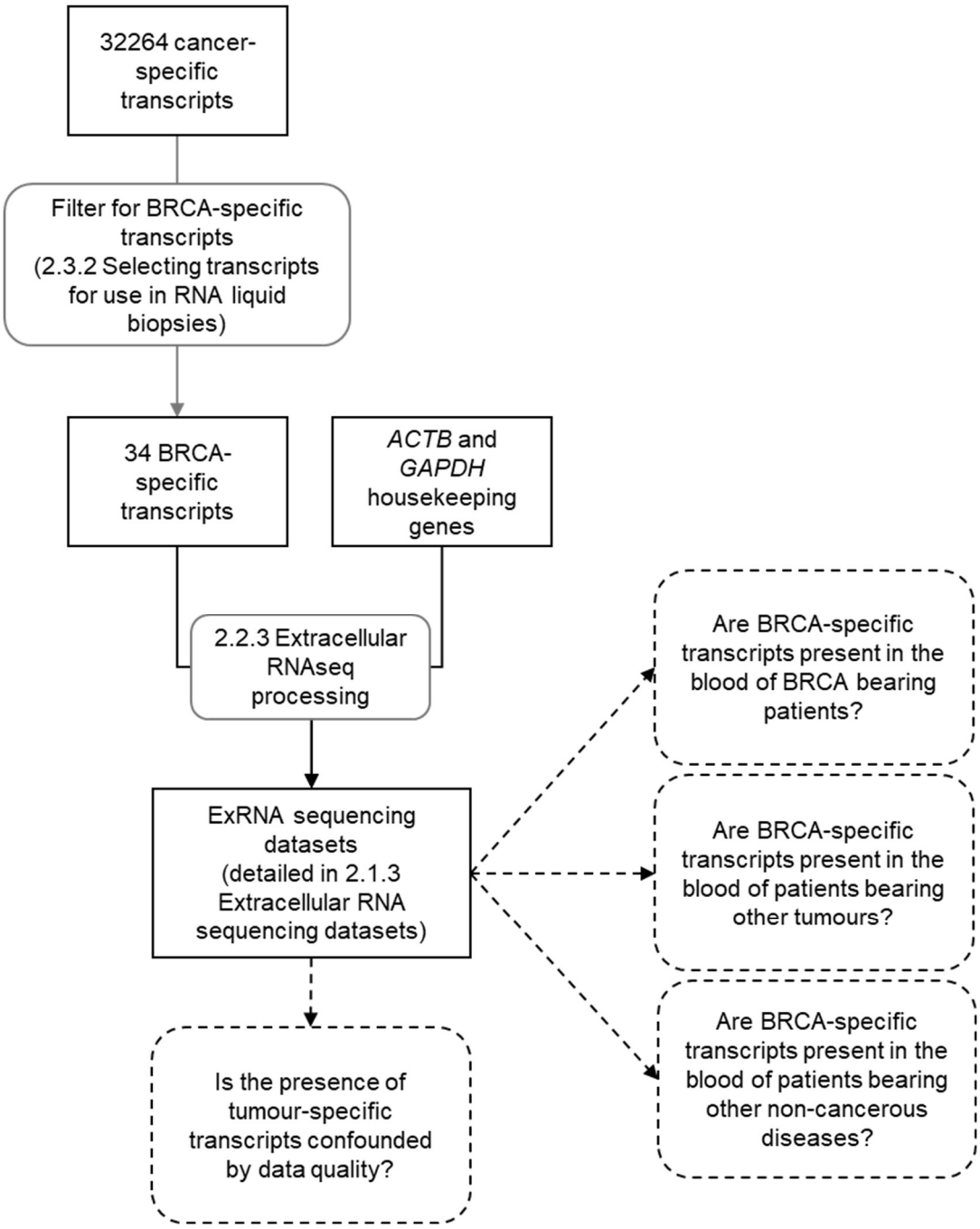


Figure 10: Aims for Results 2: Tumour-specific transcripts in extracellular RNA. Aims are shown in dashed boxes and methods are referenced in grey boxes. (BRCA: breast invasive carcinoma; exRNA: extracellular RNA).

4.2 Introduction

RNA is present both freely in blood and contained within extracellular vesicles. A range of exRNA species have been detected in blood including microRNAs (miRNAs), piwi-interacting RNAs, lncRNA, pseudogenes, LINEs, SINEs, HERVs, and gene-derived mRNA (Freedman et al., 2016; Larson et al., 2021; Nikitina et al., 2016; Qin et al., 2016; Savelyeva et al., 2017; Skog et al., 2008; Wang-Johanning et al., 2013; Yuan et al., 2016). The majority of sequences detected in exRNA samples are ribosomal and mitochondrial RNA, and only 2% of reads map to mRNA, which is consistent with cellular levels (Larson et al., 2021). The mechanism behind export of RNA from the cell is unknown, but it is suggested to be controlled as the whole transcriptome is not represented in pooled exRNA samples (Freedman et al., 2016; Groot and Lee, 2020; Hinger et al., 2018; Zhou et al., 2019). RNA-binding proteins such as the heterogeneous nuclear ribonucleoprotein A2/B1 (hnRNPA2B1) and Argonaute-2, along with membrane proteins involved in EV biogenesis (caveolin-1 and neural sphingomyelinase 2) have been shown to influence sorting of miRNAs into EVs (Groot and Lee, 2020). ExRNA may also exist outside of EVs (Murillo et al., 2019), therefore an EV isolation step in sample preparation will influence which RNAs are detected in sequencing. The function of exRNA is unknown, though multiple studies have suggested that exRNA packaged into EVs could be used for intercellular communication. This could allow transfer of drug resistance, and encourage malignant growth of nearby healthy cells as reviewed by O'Neill and colleagues (O'Neill et al.). It should be noted that the evidence for this communication is based on *in vitro* systems and the way in which EVs target specific cells *in vivo* is unknown.

Liquid biopsies would theoretically allow sampling of oligonucleotides the whole tumour has contributed to, at multiple time points, in a less invasive manner than tumour biopsies. This could be used understand the expression profile of the tumour, or to detect the presence of a tumour, with the potential for large-scale screening. Many previous studies have focused on circulating tumour DNA (ctDNA) due to its stability and representation of tumour mutational burden

(Happel et al., 2020). But exRNA is an enticing alternative, not only reflecting mutations (Shi et al., 2020; Skog et al., 2008), but splicing, RNA-editing (Giraldez et al., 2018), and transcript expression changes (Murillo et al., 2019; Shi et al., 2020; Siravegna et al., 2017). Paired RNAseq from tumour and blood EVs shows strong correlation between expression levels of transcripts (average R^2 = 0.82) (Shi et al., 2020). Further to this, unlike ctDNA which requires cell death for release, production of exRNA occurs normally in healthy cells (Happel et al., 2020). Although the source of each exRNA molecule is unknown, it could be inferred if the source transcript is specific to a certain cell type, RNA-carrier, or disease state (Murillo et al., 2019; Shi et al., 2020).

Previous studies have shown the exRNA species found in cancer patient's serum and plasma samples differentiate them from healthy individuals (Mithraprabhu et al., 2019; Mugoni et al., 2022; Skog et al., 2008; Tian et al., 2020; Wang et al., 2020; Yuan et al., 2016; Zhou et al., 2019). From a finger prick of serum Zhou and colleagues (Zhou et al., 2019) found differences in the exRNA profiles of 32 healthy individuals and 96 BRCA patients. Using defined sets of genes the majority of individuals could be correctly classified suggesting exRNA could be used as a diagnostic tool (Zhou et al., 2019). *Excision repair cross-complementing 1* (*ERCC1*) transcripts bound to EVs also differentiates BRCA bearing individuals from healthy (Keup et al., 2021). Skog and colleagues (Skog et al., 2008) were able to detect lowering levels of *epidermal growth factor receptor* (*EGFR*) mutations in exRNA after surgical resection of glioblastoma which tracked with tumour burden. Differences in exRNA profiles from individuals with benign oesophagitis and early stage oesophageal squamous cell carcinoma have also been seen (Tian et al., 2020). Additionally, exosomal *LINC02418* expression levels have been shown to distinguish colorectal cancer patients from healthy (Zhou et al., 2019), and 640 genes were shown to be differentially expressed between patients with and without non-small cell lung cancer (Wang et al., 2020). Further to diagnosing patients, work has begun on stratifying patients into treatment groups based on exRNA profiles. Mithraprabhu and colleagues (Mithraprabhu et al., 2019) found high *cereblon* (*CRBN*) expression

coupled with low *secreted protein acidic and cysteine rich* (*SPARC*) expression before treatment associated with shorter overall survival in relapsed and refractory multiple myeloma patients (Mithraprabhu et al., 2019). Furthermore, profiling extracellular vesicles of melanoma patients during immune checkpoint inhibitor treatment revealed differentially expressed genes between those responding and not responding to treatment (Shi et al., 2020). Differences in chronological age and sex have also been found to influence exRNA profiles (Max et al., 2018; Yuan et al., 2016; Zhou et al., 2019).

The majority of previous studies have removed RTEs from consideration. This was partially due to the difficulty mapping short reads to repetitive sequences, and partially to remove intronic repeat reads from confounding expression measurements of genes of interest (Giraldez et al., 2019). Though some targeted studies have been performed, Balaj and colleagues (Balaj et al., 2011) used microarrays to measure RTE expression in tumour cell line conditioned media, finding enrichment in LINE1 and HERV RNA compared to healthy controls. Previous findings also showed HERV-K (HML-2) mRNA in the serum of patients with early stage breast cancer which was not present in healthy individuals (Wang-Johanning et al., 2013). More recently, HERV envelope RNA was found in lung cancer patient blood, distinguishing them from healthy samples (Zare et al., 2018). Furthermore, pancreatic cancer patient blood was found to be enriched in expression, though not diversity, of Alu sequences (Reggiardo et al., 2023).

4.3 Results

4.3.1 Differences in data sources

Independent sources of data were collected for uniform analysis (2.1.3: Extracellular RNA sequencing datasets, 2.2.3: Extracellular RNAseq processing). Between conditions in different datasets, there were large differences in the read length and number of reads surviving trimming (Figure 11). Read lengths ranged between a median of 36.9 nucleotides (PRJEB24913 Healthy) to a median of 131 (PRJNA589238 Lung cancer) nucleotides (Figure 11a). Shorter reads may have been more likely to align non-specifically than longer reads, but this could not be corrected for. There was also large variation in the number of reads surviving trimming both between and within conditions (Figure 11b). Mean read length and read survival after trimming were significantly positively correlated (Pearson's correlation coefficient, r = 0.49, p = 9.60 x 10^{-29}), this was expected as after trimming shorter reads were more likely to fall below the 35-nucleotide threshold. Additionally, there was a significant positive correlation between the number of reads surviving trimming and the raw number of reads aligned to both *ACTB* (Pearson's correlation coefficient, r = 0.60, p = 1.11 x 10^{-44}, Figure 12a) and *GAPDH* (Pearson's correlation coefficient, r = 0.59, p = 2.30 x 10^{-43}, Figure 12b). This suggests the exRNA samples do not represent a saturated pool of reads, and increasing exRNA sample size would increase detection of other transcripts. The large differences in the number of reads surviving trimming was partially corrected for by expressing aligned reads as a percentage of the reads surviving trimming.

4.3.2 Alignment to control sequences

Alignment to control sequences showed variation between conditions (Figure 13a-b, 2.2.3: Extracellular RNAseq processing)**.** Samples from all conditions in the studies PRJNA290097 and PRJNA454814 had poor alignment to *ACTB* and *GAPDH*. Alignment to controls was higher in more recent studies (Table 2) suggesting this data may be of better quality.

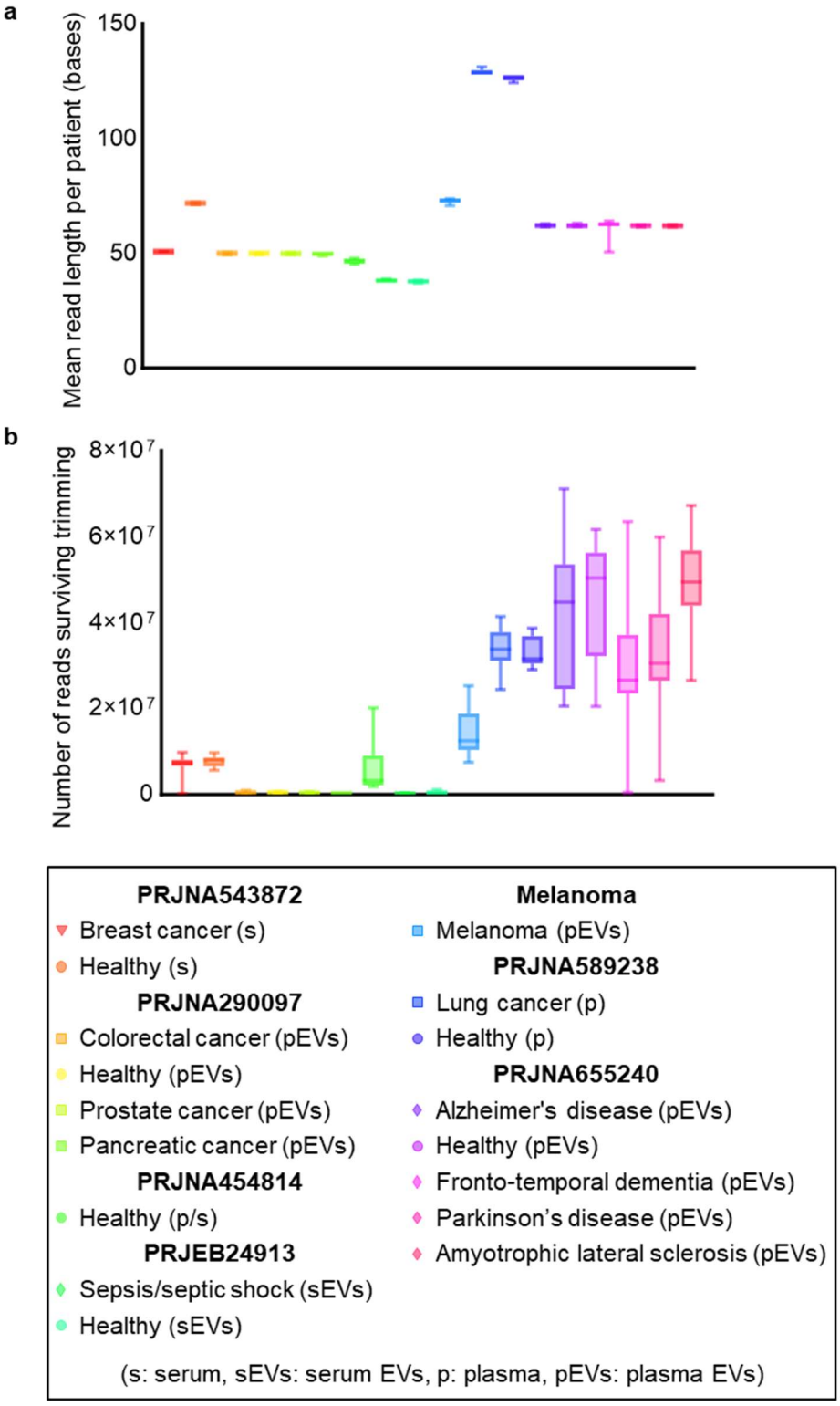


Figure 11: Differences in exRNA data sources (The legend for all graphs is at the bottom of the figure). **a.** Mean read length per patient grouped by condition. **b.** Number of reads surviving trimming per patient grouped by condition.

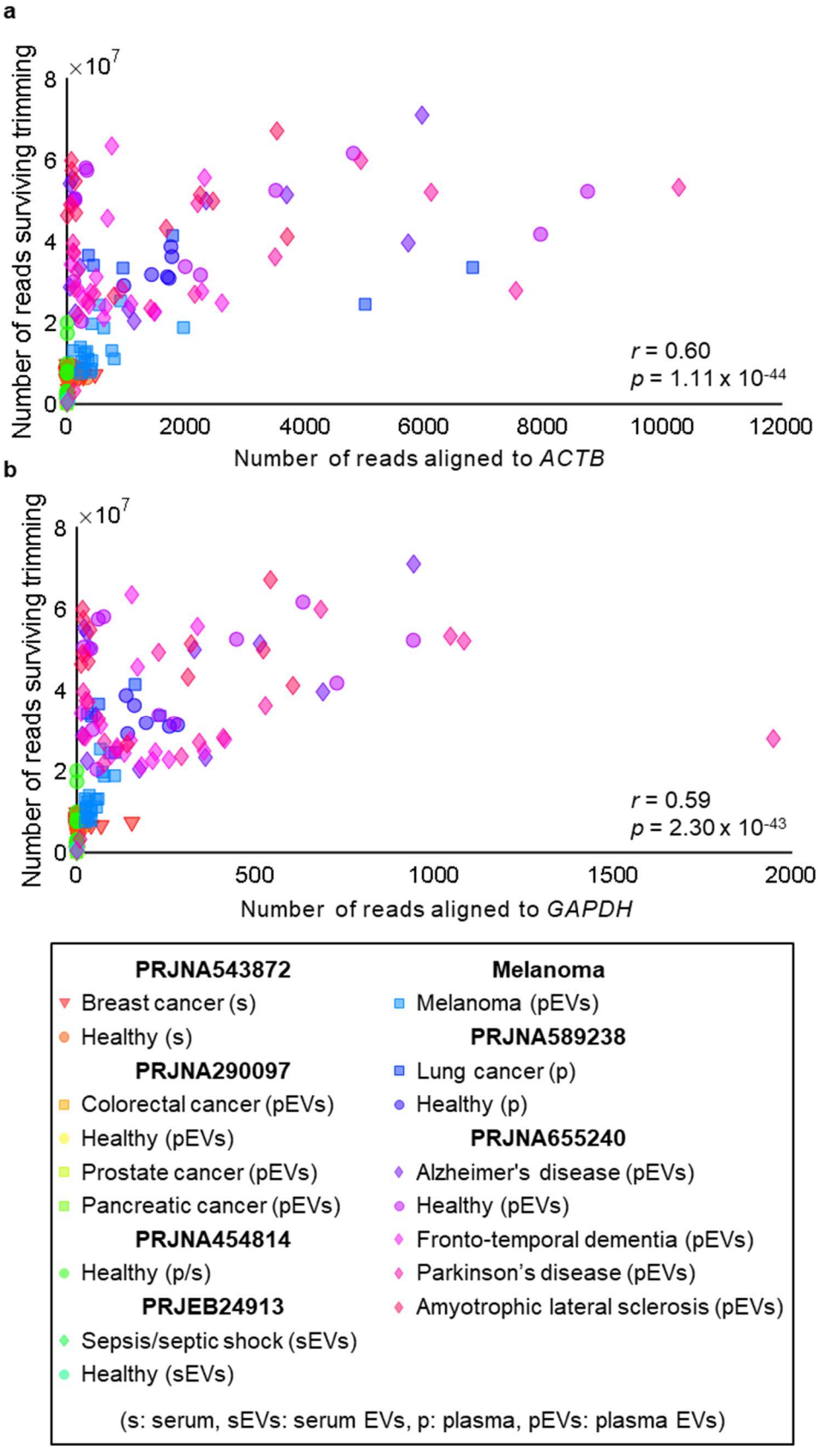

a
Number of reads surviving trimming
×10^7
8
6
4
2
0
0
2000
4000
6000
8000
10000
12000
Number of reads aligned to ACTB
r = 0.60
p = 1.11 x 10^-44
b
Number of reads surviving trimming
×10^7
8
6
4
2
0
0
500
1000
1500
2000
Number of reads aligned to GAPDH
r = 0.59
p = 2.30 x 10^-43
PRJNA543872
Breast cancer (s)
Healthy (s)
PRJNA290097
Colorectal cancer (pEVs)
Healthy (pEVs)
Prostate cancer (pEVs)
Pancreatic cancer (pEVs)
PRJNA454814
Healthy (p/s)
PRJEB24913
Sepsis/septic shock (sEVs)
Healthy (sEVs)
Melanoma
Melanoma (pEVs)
PRJNA589238
Lung cancer (p)
Healthy (p)
PRJNA655240
Alzheimer's disease (pEVs)
Healthy (pEVs)
Fronto-temporal dementia (pEVs)
Parkinson's disease (pEVs)
Amyotrophic lateral sclerosis (pEVs)
(s: serum, sEVs: serum EVs, p: plasma, pEVs: plasma EVs)

Figure 12: Alignment to control sequences normalised to the number of reads surviving trimming (The legend for all graphs is at the bottom of the figure). **a.** The relationship between the number of reads surviving trimming and the raw number of reads aligned to *ACTB* and **b.** *GAPDH* per patient.

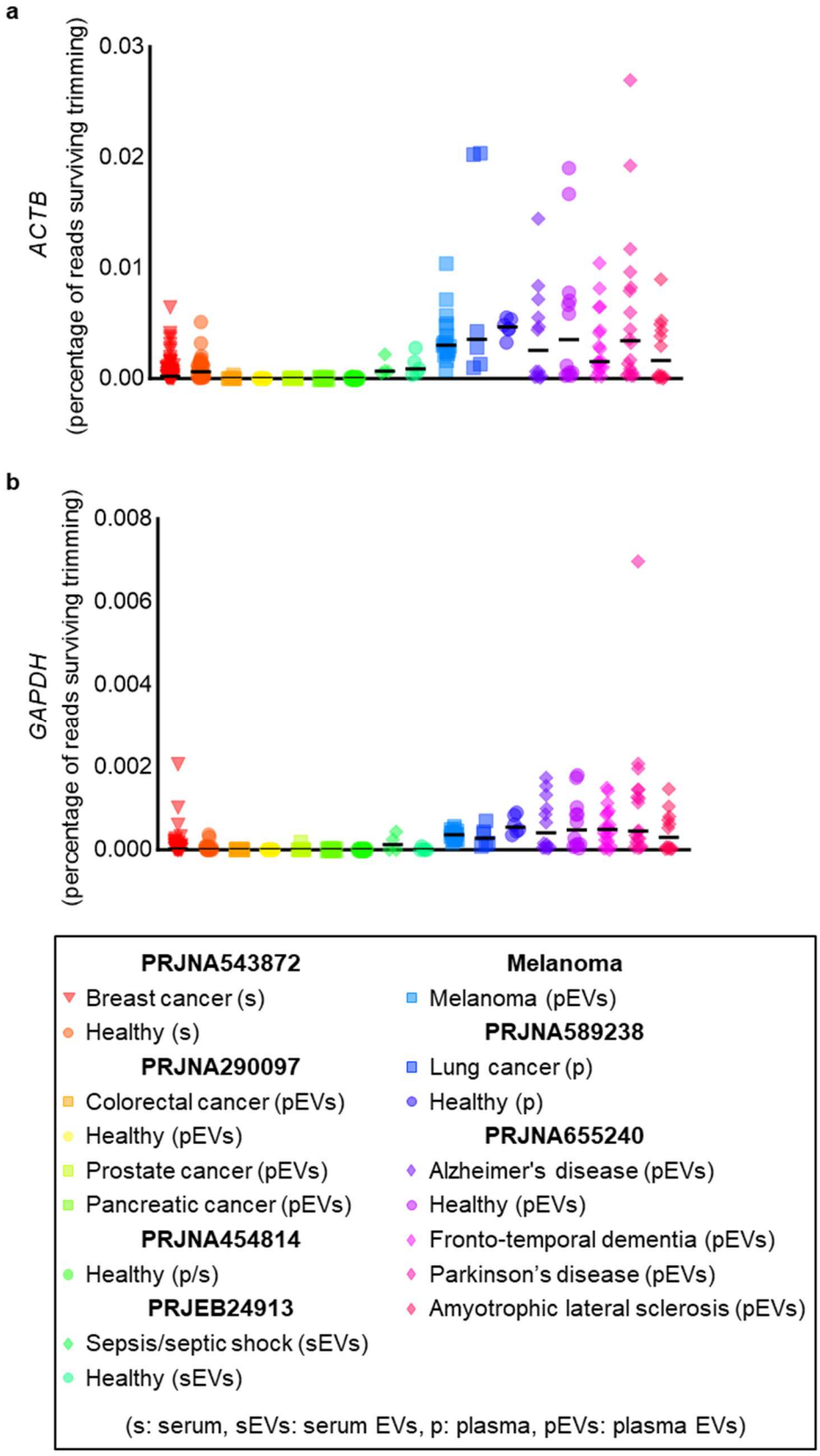

a
ACTB
(percentage of reads surviving trimming)
0.03
0.02
0.01
0.00
b
GAPDH
(percentage of reads surviving trimming)
0.008
0.006
0.004
0.002
0.000
PRJNA543872
Breast cancer (s)
Healthy (s)
PRJNA290097
Colorectal cancer (pEVs)
Healthy (pEVs)
Prostate cancer (pEVs)
Pancreatic cancer (pEVs)
PRJNA454814
Healthy (p/s)
PRJEB24913
Sepsis/septic shock (sEVs)
Healthy (sEVs)
Melanoma
Melanoma (pEVs)
PRJNA589238
Lung cancer (p)
Healthy (p)
PRJNA655240
Alzheimer's disease (pEVs)
Healthy (pEVs)
Fronto-temporal dementia (pEVs)
Parkinson's disease (pEVs)
Amyotrophic lateral sclerosis (pEVs)
(s: serum, sEVs: serum EVs, p: plasma, pEVs: plasma EVs)

Figure 13: Alignment to control sequences in each condition (The legend for all graphs is at the bottom of the figure). **a.** The percentage of reads surviving trimming aligned to *ACTB* and **b.** *GAPDH* per patient grouped by condition.

4.3.3 Selection of breast cancer specific transcripts

In total, 34 BRCA specific transcripts were selected (2.3.2: Selecting transcripts for use in RNA liquid biopsies). From hierarchical clustering of transcript expression across TCGA cancer, TCGA adjacent healthy, and GTEx healthy tissues, the transcripts form two main clusters (Figure 14a, 2.4.5: Heatmaps). One cluster of 18 transcripts appear to be breast tissue specific with some low-level expression in healthy breast tissue and higher expression in BRCA samples. The other 16 transcripts were very lowly expressed in healthy tissue, but are expressed at higher levels across different cancer types. In order to validate the transcript expression in a larger cohort of samples an additional 100 TCGA BRCA samples were analysed (2.1.2: Expanded BRCA tissue dataset, 2.2.2.1: Expression of transcripts assembled in the de novo transcriptome). Hierarchical clustering of transcript expression in these additional patients again showed two clusters of transcript expression (Figure 14b, 2.4.5: Heatmaps). One cluster of 12 transcripts was expressed across all four subtypes and their represented stages whilst maintaining low expression in adjacent healthy breast tissue. The other cluster of 22 transcripts were only sporadically expressed in basal-like tumours, but were expressed across HER2-enriched, luminal A, and luminal B tumours, with some sporadic expression also in adjacent healthy tissue.

The selected transcripts represented both known gene and RTE sequences. Transcripts overlapped exons of 24 genes, with some genes represented more than once (Table 6). Transcripts also overlapped a range of DNA, LINE, SINE, and HERV elements (Table 7). As some sequences were represented multiple times, only unique reads were counted per patient. Of the transcript sequences represented, 14 overlap sequences already known, 12 overlap sequences part of which are known and part of which are novel, and eight overlap fully novel transcript sequences. Examples of known and novel transcript structures are shown in Figure 15.

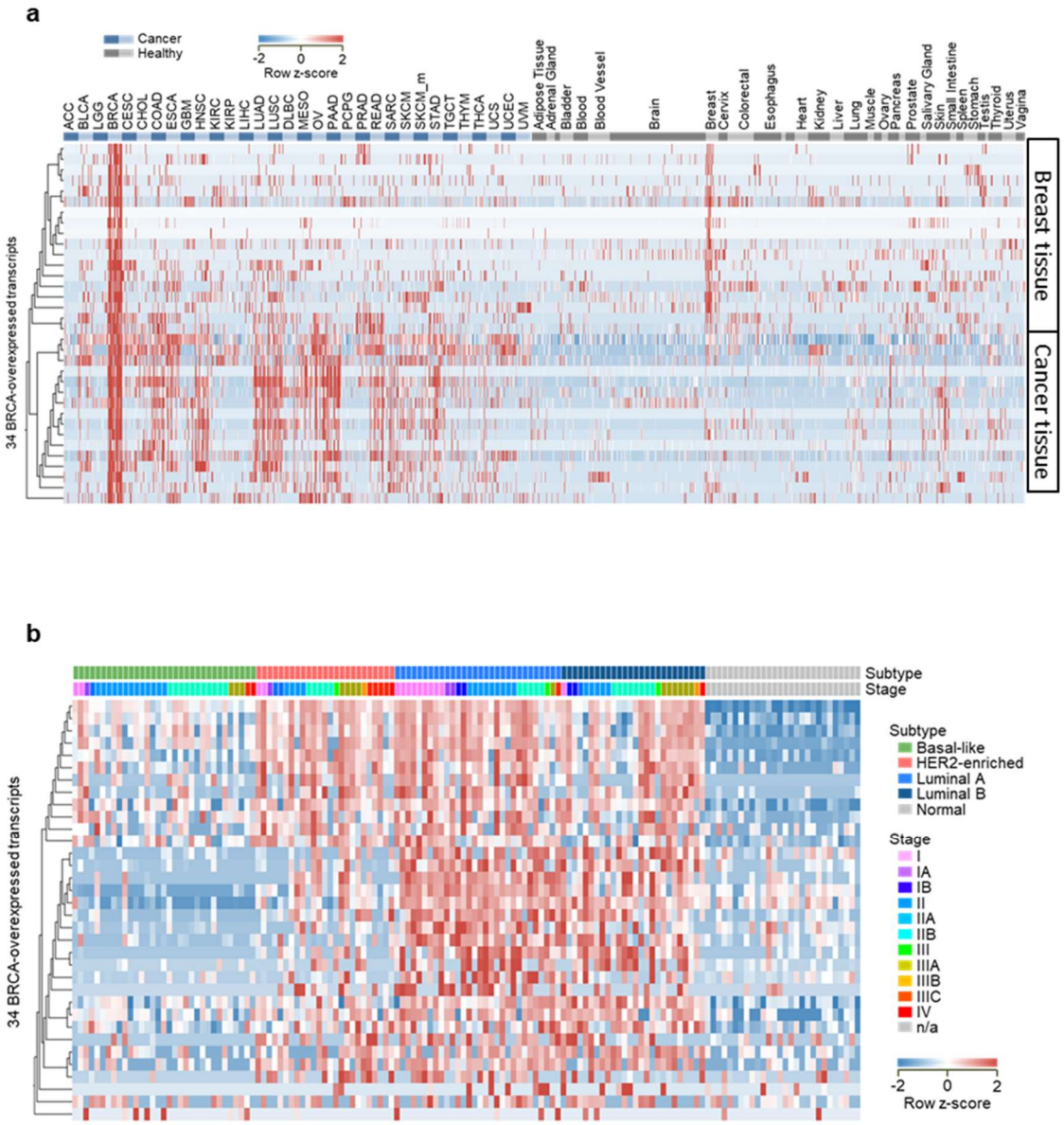


Figure 14: Expression of selected breast cancer specific transcripts. a. Expression of the 34 transcripts across TCGA cancer, TCGA adjacent healthy, and GTEx healthy tissues, ordered by hierarchical clustering. Transcripts split into two main clusters, one with expression across healthy and cancerous breast tissue, and a second with expression across cancers but little in healthy. **b.** Expression of the 34 transcripts across the additional 100 TCGA BRCA patients, ordered by hierarchical clustering.

Table 6: List of genes represented by the 34 BRCA specific transcripts.

Gene represented	Number of transcripts	Cancer tissue or breast tissue cluster	Previous associations with breast cancer
ABCC11	1	Breast tissue	(Honorat et al., 2008)
ANKRD30A	1	Breast tissue	(Mathe et al., 2015)
C1QTNF6	1		-
COL10A1	3	Cancer tissue	(Zhou et al., 2022)
COL11A1	1	Cancer tissue	(Shi et al., 2022)
DLG5	1	Cancer tissue	(Liu et al., 2019)
ELP2	1	Breast tissue	-
FHAD1	1	Breast tissue	-
FSIP1	2	Breast/cancer tissue	(Li et al., 2020a)
H2AC18	1	Cancer tissue	-
H2AC19	1	Cancer tissue	-
IQCK	1	Cancer tissue	-
KNOP1	1	Cancer tissue	(Li et al., 2023)
LINC00504	1	Breast tissue	(Feng et al., 2021)
LINC02544	1	Cancer tissue	(Guo et al., 2020)
LRRC15	3	Cancer tissue	(Yang et al., 2021b)
MMP11	1	Cancer tissue	(Kim et al., 2021)
MMP13	1	Cancer tissue	(Li et al., 2022)
MS4A7	1	Breast tissue	(Wu et al., 2023)
NT5DC1	1		(Jia et al., 2023)
POTEJ	1	Cancer tissue	-
SLC39A6	1	Breast tissue	(de Nonneville et al., 2023)
SRMS	2	Breast tissue	(Limsakul et al., 2023)
TESMIN	1	Cancer tissue	-
none	8	Breast tissue	

Table 7: List of RTEs represented by the 34 BRCA specific transcripts

RTE type	Number of transcripts
DNA	6
HERV	5
LINE	11
SINE	13
none	13

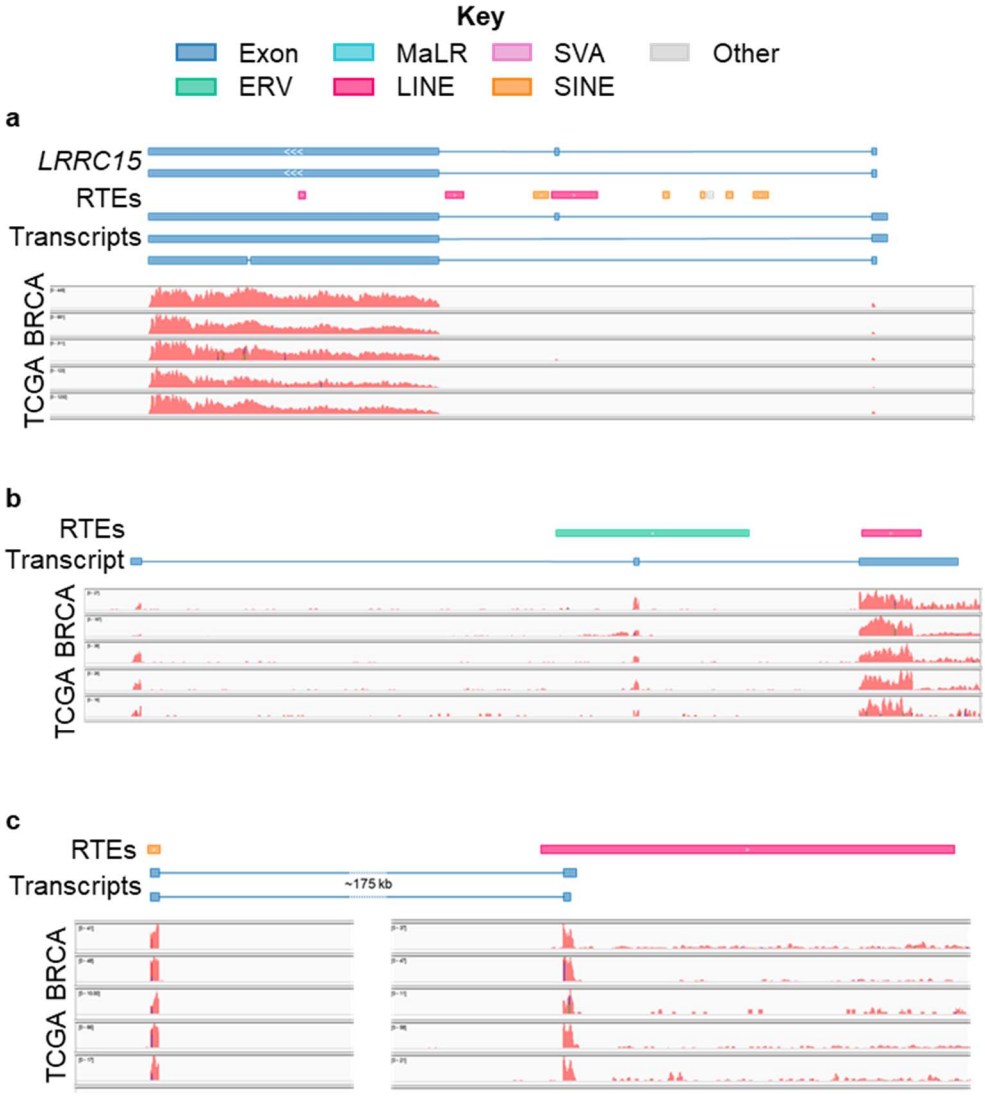


Figure 15: Structures of six of the selected transcripts alongside BAM files from five TCGA BRCA samples. **a.** The structure and expression of three transcripts overlapping a known sequence from the *LRRC15* gene. **b.** The structure and expression of a novel transcript overlapping RTE sequences and a region overlapping no known gene or RTE. **c.** The structure and expression of two novel transcripts spliced between two RTE elements.

4.3.4 Alignment to breast cancer specific transcripts

ExRNA sequencing data from 96 BRCA bearing donors and 32 healthy donors (PRJNA543872) were aligned to the 34 BRCA specific transcripts (2.2.3: Extracellular RNAseq processing). Although there was little change in the number of transcripts aligned to, reads aligned were more prevalent in BRCA bearing donors compared to healthy donors (Figure 16a-b). There was a small increase in the number of different transcripts aligned to in samples from BRCA bearing donors compared to healthy donors (Mann-Whitney, $p = 6.90 \times 10^{-4}$, BRCA: n = 96, median = 19 transcripts, healthy: n = 32, median = 17 transcripts, Figure 16a). However, BRCA bearing donors had a marked increase in the percentage of reads surviving trimming aligned to the 34 BRCA specific transcripts (Mann-Whitney, $p = 1.35 \times 10^{-30}$, BRCA: n = 96, median = 0.02498% reads surviving trimming, healthy: n = 32, median = 0.005691% reads surviving trimming, Figure 16b). This suggested that using a cut off of 0.01238% reads surviving trimming aligned to the 34 BRCA specific transcripts the BRCA bearing donors could be separated from healthy donors.

It is possible other cancers and patient comorbidities may impact the sensitivity and specificity of the test. In order to test whether samples from patients with other diseases would contain sequences which would confound expression of the selected transcripts, exRNA sequencing data from independent studies (Table 2, 2.2.3: Extracellular RNAseq processing) was aligned to the 34 BRCA specific transcripts. A total of 317 additional donors were used across six studies, including 86 healthy donors (13 sampled longitudinally (Max et al., 2018)), 167 donors bearing cancerous diseases, and 64 bearing non-cancerous diseases. Again, there were some similarities in the number of transcripts aligned to (Figure 16a), but read prevalence was much higher in BRCA bearing donors compared to all other donors (Figure 16b). Using the same cut-off defined previously (0.01238% reads surviving trimming) it appeared that BRCA bearing donors could be separated from all other samples (Figure 17).

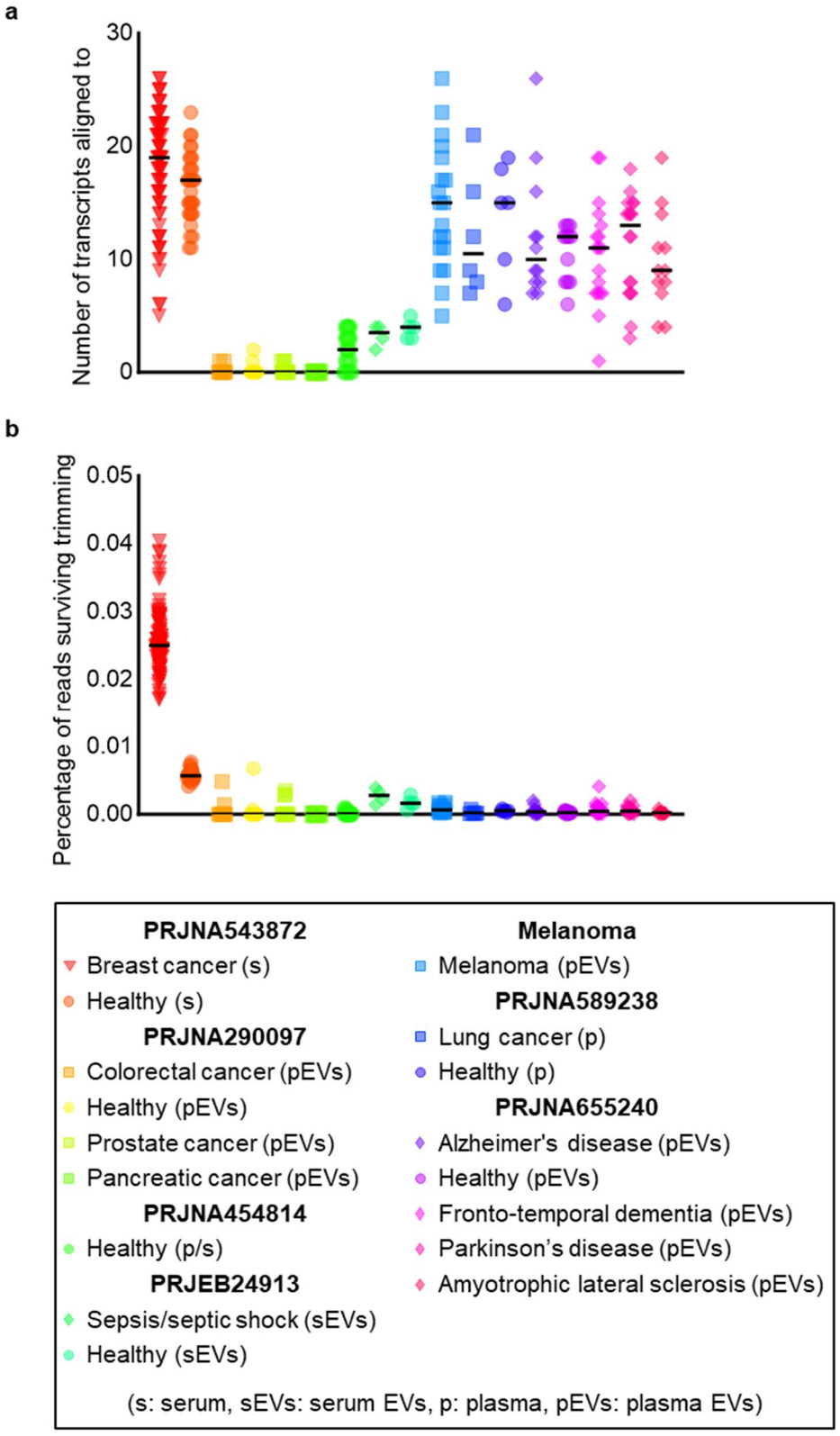

a
Number of transcripts aligned to
30
20
10
0
b
Percentage of reads surviving trimming
0.05
0.04
0.03
0.02
0.01
0.00
PRJNA543872
Breast cancer (s)
Healthy (s)
PRJNA290097
Colorectal cancer (pEVs)
Healthy (pEVs)
Prostate cancer (pEVs)
Pancreatic cancer (pEVs)
PRJNA454814
Healthy (p/s)
PRJEB24913
Sepsis/septic shock (sEVs)
Healthy (sEVs)
Melanoma
Melanoma (pEVs)
PRJNA589238
Lung cancer (p)
Healthy (p)
PRJNA655240
Alzheimer's disease (pEVs)
Healthy (pEVs)
Fronto-temporal dementia (pEVs)
Parkinson's disease (pEVs)
Amyotrophic lateral sclerosis (pEVs)
(s: serum, sEVs: serum EVs, p: plasma, pEVs: plasma EVs)

Figure 16: Alignment to the 34-breast cancer specific transcripts per condition (The legend for all graphs is at the bottom of the figure). **a.** Number of transcripts aligned to (of the 34) per patient in each condition across all datasets. **b.** The percentage of unique reads surviving trimming aligned to all 34 BRCA specific transcripts per patient in each condition across all datasets.

However, there were some anomalies in the data which required further exploration. Firstly, some transcripts were substantially more highly aligned to than others, and secondly, the healthy donors from PRJNA543872 were easily distinguished from other healthy donors in the independent studies (Figure 17).

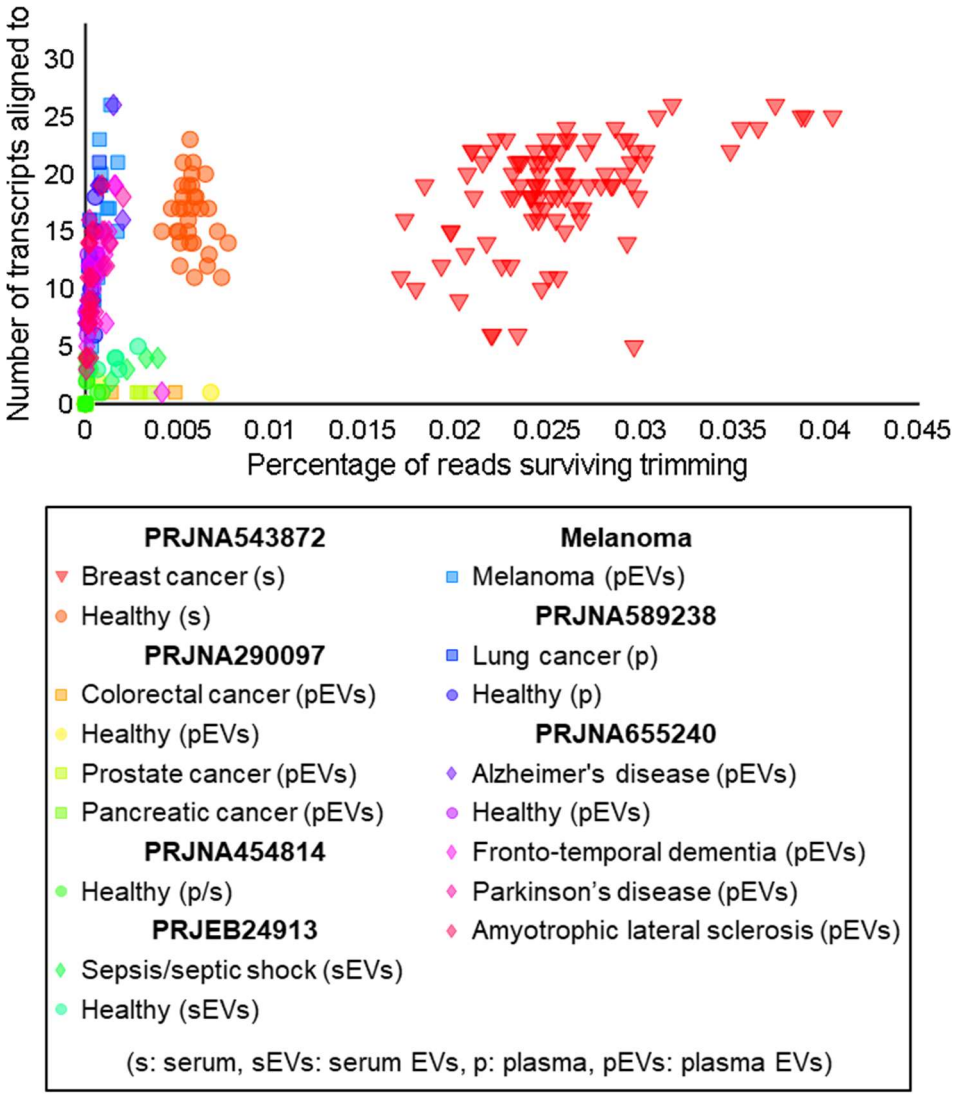


Figure 17: Alignment to the 34-breast cancer specific transcripts per patient. Number of transcripts aligned to per patient against the percentage of unique reads surviving trimming aligned to those transcripts per patient in each condition across all datasets.

4.3.5 Questions about data quality

In an attempt to validate the sequencing had been performed on RNA instead of DNA, the number of reads which overlapped splice sites in the control sequences were analysed (Figure 18, 2.2.3.1: Analysis of spliced reads)**.** To lend context to the results, this splicing was compared to 24 TCGA BRCA tissue samples (2.1.1: Original tissue datasets (Attig et al., 2019)). The amount of splicing over each control was expressed as a percentage of the total number of reads overlapping that control sequence. Many donors had zero spliced reads over the controls, not because there was no splicing, but because no reads were detected anywhere along the control sequences. Here it is impossible to tell if RNA or DNA was sequenced using spliced reads on *ACTB* and *GAPDH* alone. In cases were 100% of reads over the control were spliced, only one or two reads were aligned to the control and all were aligned over splice junctions, indicating RNA was present but at low levels.

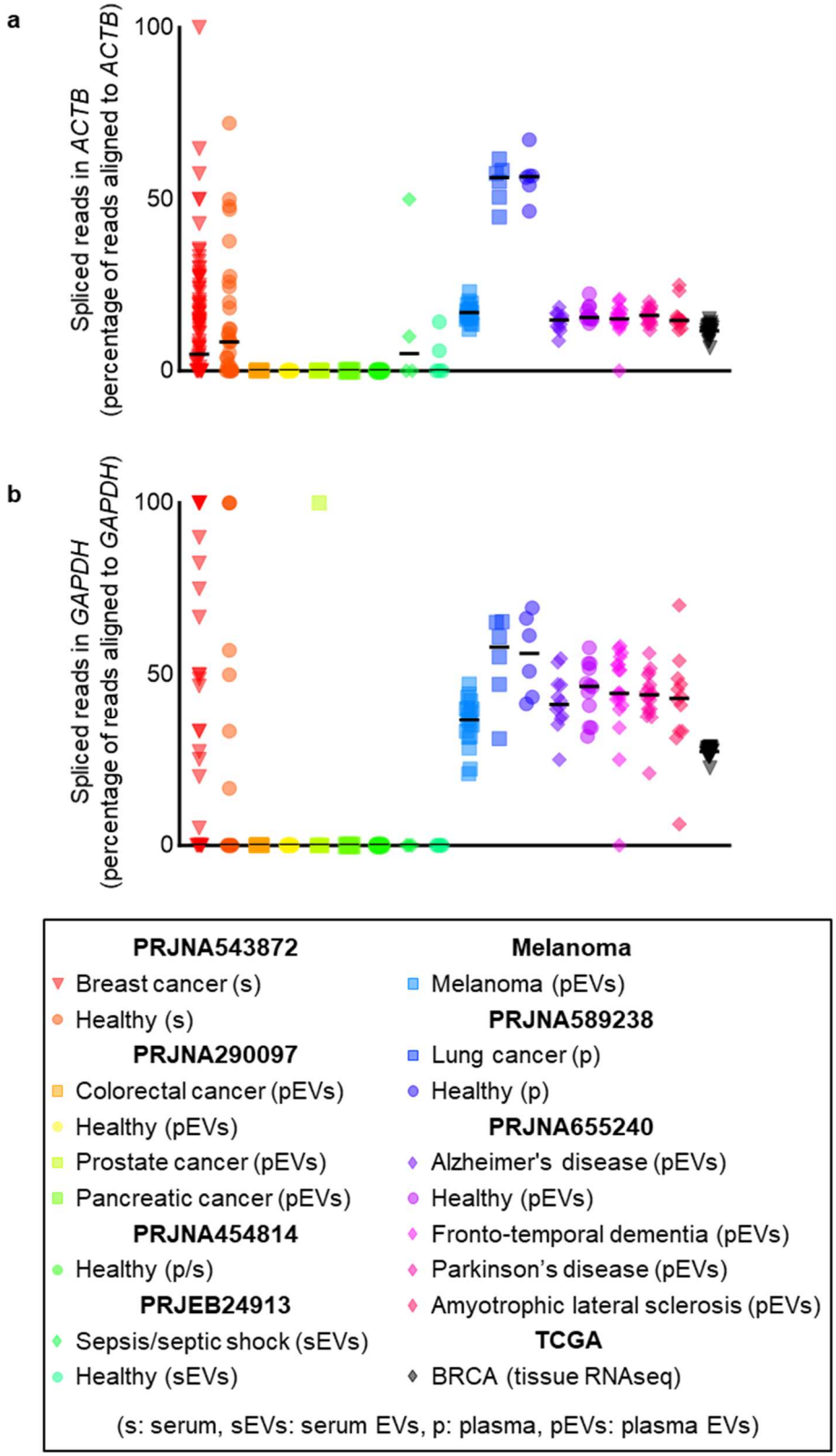

a
Spliced reads in ACTB
(percentage of reads aligned to ACTB)
100
50
0
b
Spliced reads in GAPDH
(percentage of reads aligned to GAPDH)
100
50
0
PRJNA543872
Breast cancer (s)
Healthy (s)
PRJNA290097
Colorectal cancer (pEVs)
Healthy (pEVs)
Prostate cancer (pEVs)
Pancreatic cancer (pEVs)
PRJNA454814
Healthy (p/s)
PRJEB24913
Sepsis/septic shock (sEVs)
Healthy (sEVs)
Melanoma
Melanoma (pEVs)
PRJNA589238
Lung cancer (p)
Healthy (p)
PRJNA655240
Alzheimer's disease (pEVs)
Healthy (pEVs)
Fronto-temporal dementia (pEVs)
Parkinson's disease (pEVs)
Amyotrophic lateral sclerosis (pEVs)
TCGA
BRCA (tissue RNAseq)
(s: serum, sEVs: serum EVs, p: plasma, pEVs: plasma EVs)

Figure 18: Reads overlapping splice junctions in control sequences.(The legend for all graphs is at the bottom of the figure). **a.** Reads overlapping splice junctions in *ACTB* and **b.** *GAPDH* as a percentage of the total reads aligned to *ACTB* and *GAPDH* respectively per patient in each condition across datasets.

There were also discrepancies in the read alignment patterns across breast cancer specific transcripts. Instead of equal coverage across the whole transcript some transcripts contained large read peaks overlapping RTEs (Figure 19). These peaks overlapped *LINE1HS* and *AluSp* sequences, and regions were well conserved when compared to the respective consensus sequences. There are many RTE copies in DNA, and not all are expressed in RNA, in order to explore whether this upregulation in RTE-derived sequences was BRCA-specific or an artefact of the study, the RNAseq data was re-aligned to the *LINE1HS* and *AluSp* consensus sequences. In order to capture reads across all the *LINE1HS* and *AluSp* family members, BLAT was run using a minimum identity of 90%, whilst ensuring the match and read length were the same (2.2.3: Extracellular RNAseq processing). This showed a study-specific upregulation of both RTEs in PRJNA543872 when compared to the other independent datasets (Figure 20).

Artefactual differences in PRJNA543872 could also be seen when comparing healthy data across studies. The healthy data from PRJNA543872 had increased alignment to the *LINE1HS* (Figure 20a) and *AluSp* (Figure 20b) consensus sequences, alongside increased alignment to BRCA-specific transcripts (Figure 17) when compared to healthy donors from other studies. PRJNA543872 healthy donors could be clearly distinguished from other healthy donors using the percentage of reads surviving trimming aligned to BRCA-specific transcripts (Figure 17), these reads come from alignment with *LINE1HS* and *AluSp* (Figure 19).

Overall, some transcripts have especially high expression in BRCA patient exRNA samples. The read alignment to these transcripts is not evenly spread across the transcript length but is instead concentrated to *LINE1HS* and *AluSp* overlapping regions (Figure 19). The detection of *LINE1HS* and *AluSp* is increased in both healthy and BRCA-bearing donors of the PRJNA543872 study, this is not seen in other studies, and thus is a study-specific artefact (Figure 20). In an attempt to correct for the artefactual upregulation of *LINE1HS* and *AluSp*, transcripts containing large read peaks over these RTEs were removed. This

reduced the magnitude of variation between healthy datasets, but also removed the main source of variation between all conditions and BRCA bearing donors were no longer distinguishable from other donors (Figure 21).

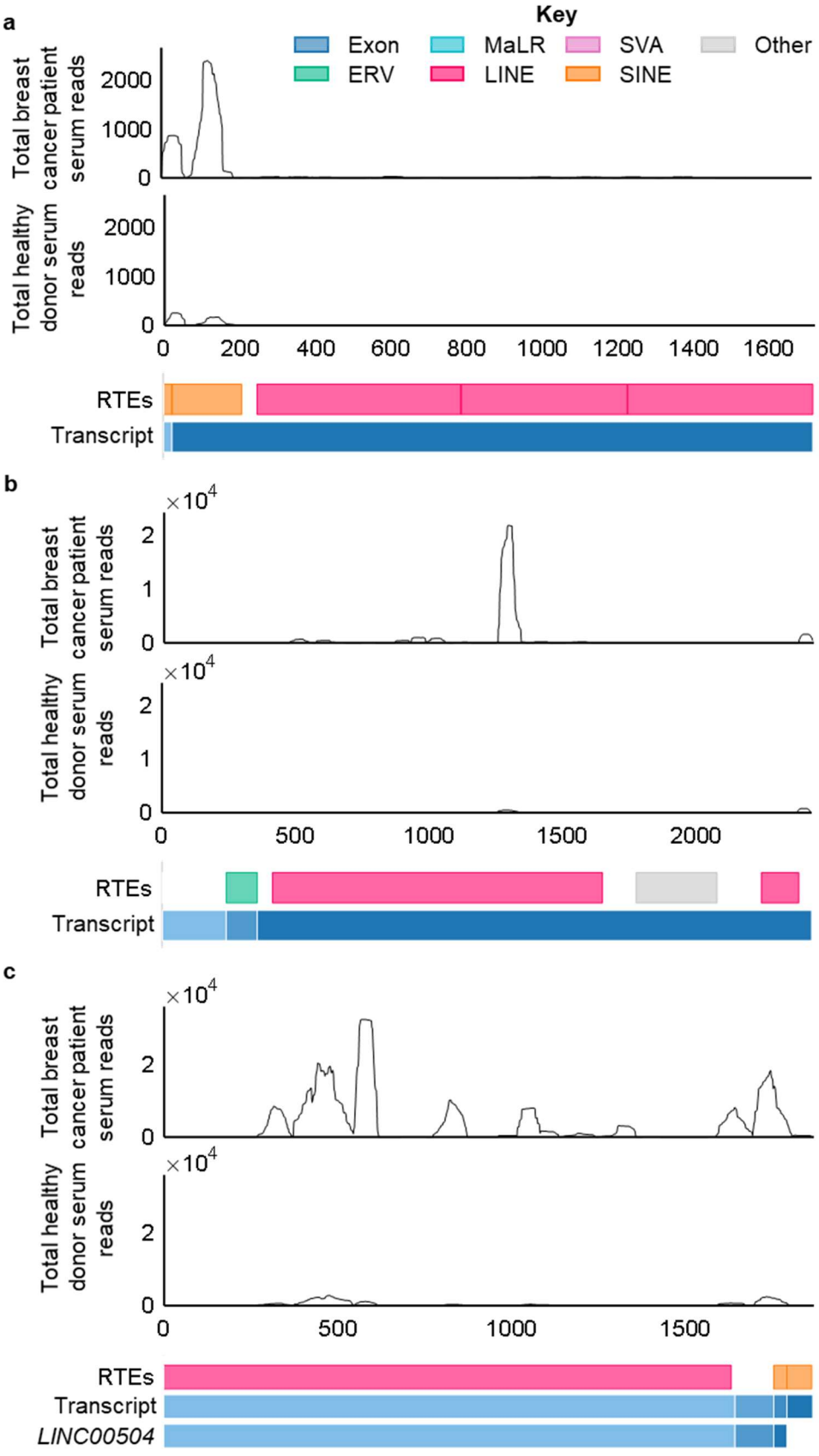

a
Key
Exon
ERV
MaLR
LINE
SVA
SINE
Other
Total breast cancer patient serum reads
Total healthy donor serum reads
2000
1000
0
0
200
400
600
800
1000
1200
1400
1600
RTEs
Transcript
b
×10^4
2
1
0
500
1000
1500
2000
c
LINC00504

Figure 19: Structures of the three BRCA specific transcripts containing large peaks of read alignment over RTE elements (intronic regions of the transcripts are not shown). Read counts shown are pooled from all patients. Different exons are shown in different shades of blue, gene exons overlapped by transcripts are not necessarily complete.

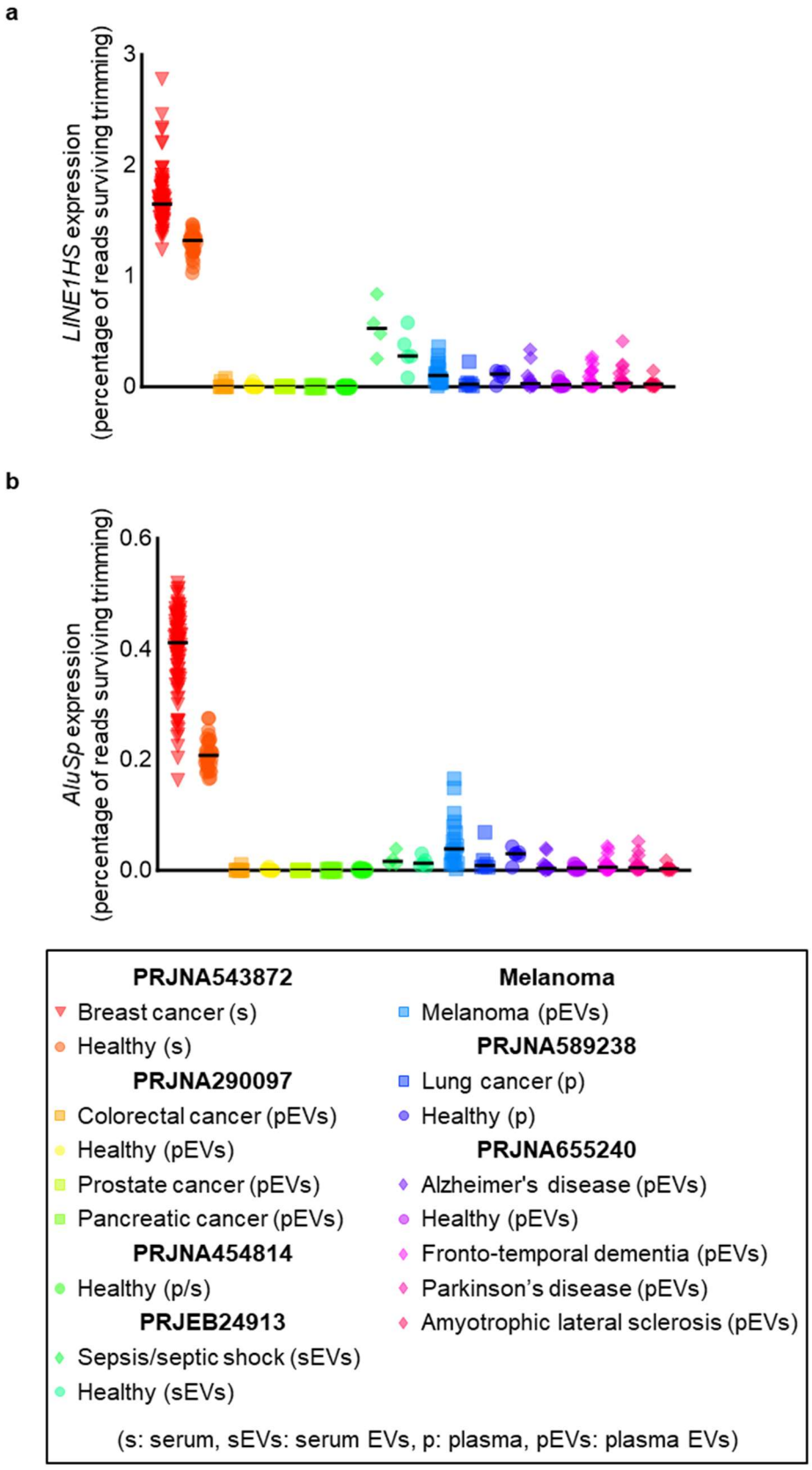

a
LINE1HS expression
(percentage of reads surviving trimming)
3
2
1
0
b
AluSp expression
(percentage of reads surviving trimming)
0.6
0.4
0.2
0.0
PRJNA543872
Breast cancer (s)
Healthy (s)
PRJNA290097
Colorectal cancer (pEVs)
Healthy (pEVs)
Prostate cancer (pEVs)
Pancreatic cancer (pEVs)
PRJNA454814
Healthy (p/s)
PRJEB24913
Sepsis/septic shock (sEVs)
Healthy (sEVs)
Melanoma
Melanoma (pEVs)
PRJNA589238
Lung cancer (p)
Healthy (p)
PRJNA655240
Alzheimer's disease (pEVs)
Healthy (pEVs)
Fronto-temporal dementia (pEVs)
Parkinson's disease (pEVs)
Amyotrophic lateral sclerosis (pEVs)
(s: serum, sEVs: serum EVs, p: plasma, pEVs: plasma EVs)

Figure 20: Expression of *LINE1HS* (a) and *AluSp* (b) per patient in each condition allowing for 90% match identity. (The legend for all graphs is at the bottom of the figure). There is a study specific increase of both repeat types in PRJNA543872.

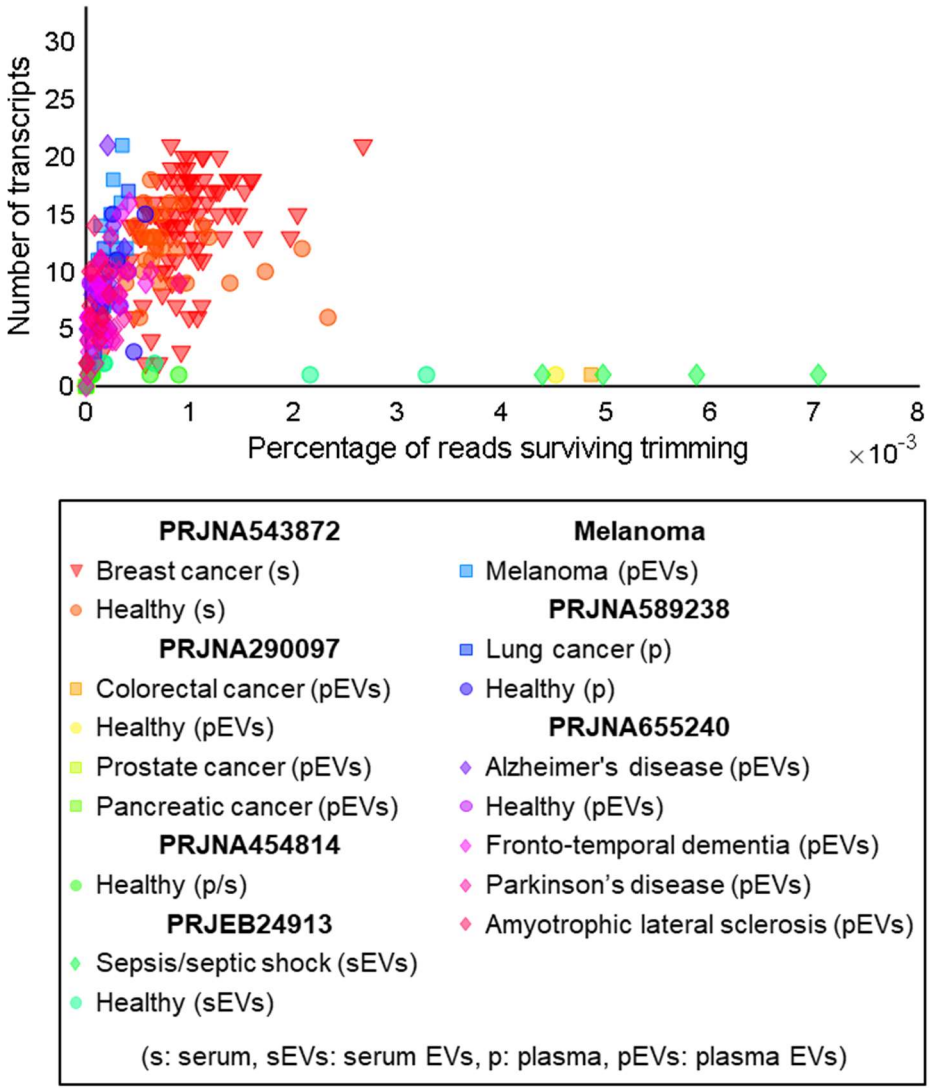


Figure 21: ExRNA expression per condition after removal of the three transcripts containing large peaks over RTEs. Number of transcripts aligned to per patient against the percentage of unique reads surviving trimming aligned to those transcripts per patient in each condition

4.3.6 Attempts in other cancer types

Following the same transcript selection method as used previously (2.3.2: Selecting transcripts for use in RNA liquid biopsies), SKCM specific transcripts were selected and reads from the Melanoma study were aligned using BLAT (2.2.3: Extracellular RNAseq processing). Although there was good alignment to several transcripts, the study did not contain any healthy donors to compare alignment levels to (Table 2). Alignment of an independent dataset showed melanoma bearing donors to have lower levels of SKCM specific transcripts compared to healthy donors.

In an attempt to reduce non-specific read alignment from homologous sequences, and given the lung cancer dataset (PRJNA589238) had such long reads (Figure 11a), Salmon was used for alignment as described for tissue data (2.2.2.1: Expression of transcripts assembled in the de novo transcriptome). Lung cancer specific transcripts were selected with the same method as used previously (2.3.2: Selecting transcripts for use in RNA liquid biopsies). However, both lung cancer patient and healthy donor samples were equally lacking in transcript presence.

4.4 Discussion

BRCA bearing donors could not be distinguished from others using the 34 BRCA specific transcripts from the *de novo* transcriptome assembly. Although TCGA BRCA patient samples had high and specific expression of the 34 BRCA specific transcripts, many of these transcripts were poorly represented in BRCA bearing donor exRNA samples. This may be because the 96 BRCA-bearing patients did not have tumours expressing the selected transcripts. The transcript expression was validated from the original pool of 24 patients with a larger group of 100. However, as these 100 patients represent under one tenth of the total TCGA BRCA patient pool (100/1095) it is possible that selection of a different 100 patients would have shown no expression of transcripts in some tumours. There may also be biases in RNA packaging for extracellular release (Freedman et al., 2016; Groot and Lee, 2020; Hinger et al., 2018; Zhou et al., 2019), or lack of stability of the transcripts in the blood, although there is disagreement on this area. Some suggest that even in the presence of RNase in the blood exRNA species can be stable through protection in EVs (Freedman et al., 2016). Others suggest there is a higher frequency of RNA modification of RNA in EVs compare to intracellular RNA (Hinger et al., 2018).

Transcripts that contained conserved RTE regions were detected at higher levels in BRCA bearing donor exRNA samples. Read coverage across the highly expressed transcripts was not constant, with specific peaks over regions of *LINE1HS* and *AluSp* RTEs. The increased detection of *LINE1HS* and *AluSp* was not BRCA donor sample specific but study specific. This may have been due to any number of differences between datasets collected in the methods of collection, preparation, and sequencing of the exRNA samples. The increased detection of *LINE1HS* and *AluSp* sequences may also be due to DNA contamination, but as there were so few reads aligned to *ACTB* and *GAPDH* sequences in samples from the breast cancer study the frequency of spliced reads could not be usefully analysed. Comparison of RTEs in cell free DNA and in cellular DNA showed Alu was detected at higher proportions and LINE1 at

lower proportions outside the cell, but both were still detected at very high levels (Gezer et al., 2022).

Alignment to repetitive sequences using short reads requires a balance of specificity and sensitivity. In order to increase alignment specificity, BLAT was used, an alignment software specialised for shorter reads with few gaps or mismatches, alongside a minimum alignment identity of 100%. Although this meant that reads from mutated transcripts would be binned, as well as any reads altered whilst in the blood, it increased the likelihood reads would map to their true origin. But as the database BLAT was given only contained the 34 BRCA specific transcripts, it was blind to the rest of the human transcriptome, so with that data alone it was impossible to tell if the reads could have also been aligned elsewhere. In order to improve alignment specificity a more stringent filter on read length could have been applied, but this would have further restricted the size of the exRNA samples. The correlation of *ACTB* and *GAPDH* read alignment with read length suggested that reduced exRNA sample sizes would reduce the likelihood of detecting any given sequence, thus reducing the sensitivity of the test. It has previously been shown that the ability to distinguish sepsis donors from healthy donors is dependent on the miRNA yield (Buschmann et al., 2018). However, increasing the acceptable read length may have improved the specificity of read alignment. When aligning to the Genome Reference Consortium human build 37 (GRCh37) with a read length of 35 nucleotides used as the minimum here, there was a 16% probability the read would map to multiple coordinates (Li et al., 2014). As the transcriptome covers a smaller sequence space than the genome, the probability of multimapping is likely to be lower in this case. But the probability of multimapping could perhaps act as a guide to help set the minimum read length required for confident read alignment to any given sequence of interest.

There is a lack of systematic studies to show the impact of technical methods on the exRNA species detected in samples. This lack of understanding limits the use of exRNA as a biomarker as there may be unknown impacts from sample storage,

EV purification, adapter ligation, and sequencing techniques (Everaert et al., 2019; Giraldez et al., 2018; Qin et al., 2016; Yuan et al., 2016). Study of different exRNA isolation, library preparation, and sequencing methods have shown these methods define expression differences of samples in clustering analysis (Murillo et al., 2019). Furthermore, biases from library preparation altered the exRNA species detected, with more recently developed methods detecting more long intergenic non-coding RNA than older ones (Murillo et al., 2019). Additionally, differences have been seen when sequencing the same synthetic miRNA pools in different laboratories (Giraldez et al., 2018). These synthetic exRNA profile results clustered by extraction and library preparation protocols. Although the relative expression values within samples were similar, comparison between samples using read counts was inaccurate (Giraldez et al., 2018). This variation has also been explored in patient samples, with differentially expressed miRNAs between healthy and sepsis donors being dependent on the EV isolation method used (Buschmann et al., 2018).

Due to the lack of accepted quality controls in the field, there is no way to define the best protocols for processing patient samples. Although various methods are impacting the exRNA species detected in each sample, which method reflects the true pool of exRNA species is unknown. Additionally, with exRNA samples originating from such small input volumes, heterogeneity in detected species between samples from the same individual is likely to be high (Everaert et al., 2019), further reducing the ability to identify the best protocols for patient samples. Though one longitudinal study has shown some stability of exRNA profiles of healthy donors over two months (Max et al., 2018). This stability would not necessarily be reflected in a disease state such as cancer, especially when combined with anti-cancer treatments which can also affect the function of healthy cells.

4.5 Conclusion

Overall, the selected 34 BRCA specific transcripts (which could distinguish breast cancer patients using RNA expression from tumour biopsies) could not be used to distinguish BRCA bearing donors from others using serum RNA, partially due to artefacts of the methods used in the independent datasets, and partially due to the poor representation of most of the 34 transcripts in the BRCA bearing donor samples. Further work needs to be done to define agreed upon quality controls for exRNA datasets, and agreed upon protocols to allow comparison between studies if exRNA species are to become a reliable biomarker for disease.

Chapter 5. Results 3: Novel transmembrane domain containing proteins

5.1 Aims

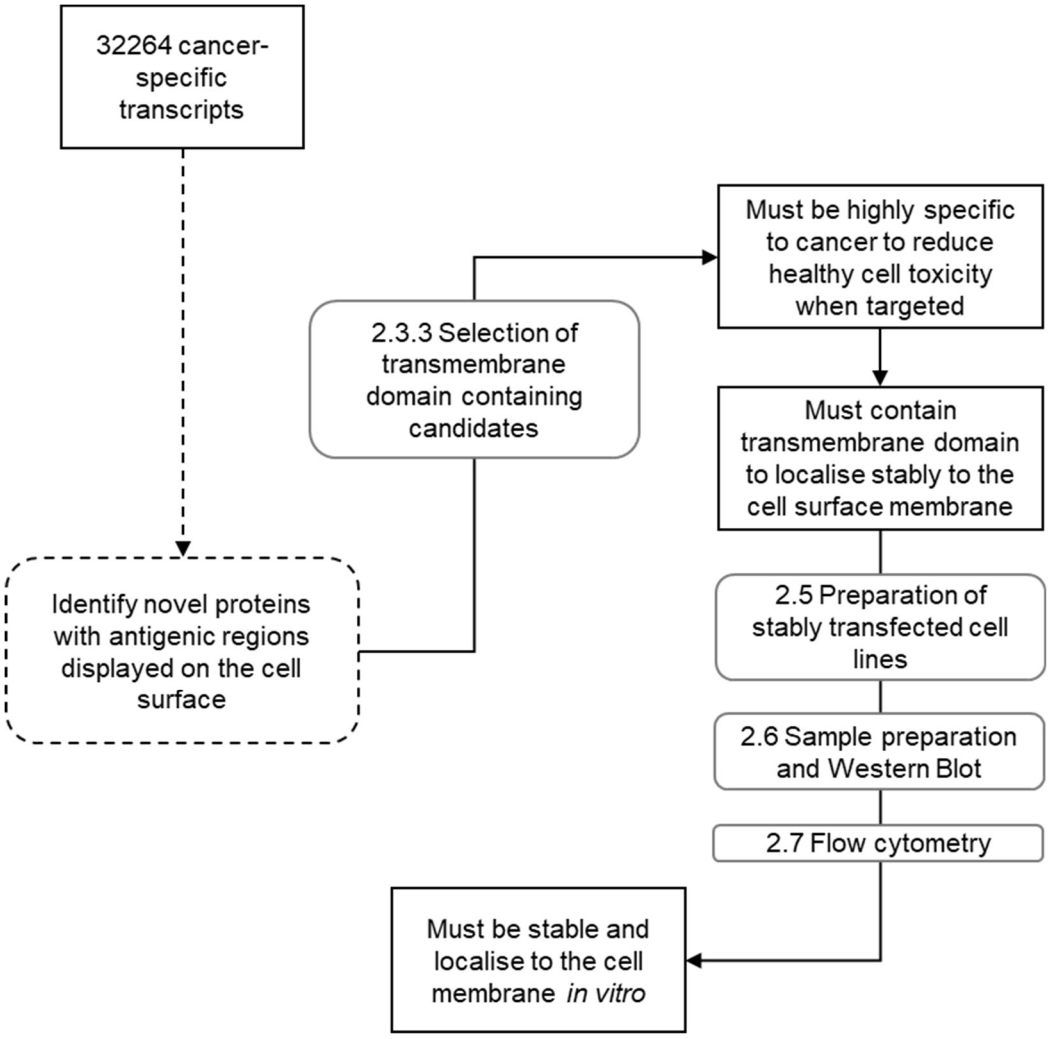


Figure 22: Aims for Results 3: Novel transmembrane domain containing proteins. Aims are shown in dashed boxes and methods are referenced in grey boxes.

5.2 Introduction

Novel transcripts identified by the *de novo* transcriptome assembly may give rise to antigenic proteins localised to the cell surface plasma membrane (1.5.3: Expression of antigenic proteins). Proteins localised to the cell surface are available for B cell or antibody recognition without the requirements of processing and display on MHC molecules, thus increasing the likelihood of being shared across many tumours. The *de novo* transcriptome assembly identified both sequences chimeric with canonical transmembrane domain containing proteins, and fully novel sequences which may also code for transmembrane domain proteins. Here a list of potential highly cancer-specific candidates was selected, and three candidates were tested *in vitro* for protein stability and localisation (Figure 22).

5.3 Results

5.3.1 Selection of candidate transcripts

From the original cancer-specific list of 32264 transcripts, 313 were selected after identifying highly cancer-specific transcripts containing at least one open reading frame coding for at least one transmembrane domain (Figure 22, Figure 23a, 2.3.3: Selection of transmembrane domain containing candidates). The selected transcripts were upregulated in a range of cancer types, with some transcripts upregulated in multiple cancers (Figure 23b). Most transcripts were multiexonic, though some were monoexonic (Figure 24a), several of which overlapped HERV elements. The open reading frames within the transcripts ranged in length from 100 to 3312 AA (Figure 24c), with the much larger open reading frames generally overlapping known gene exons.

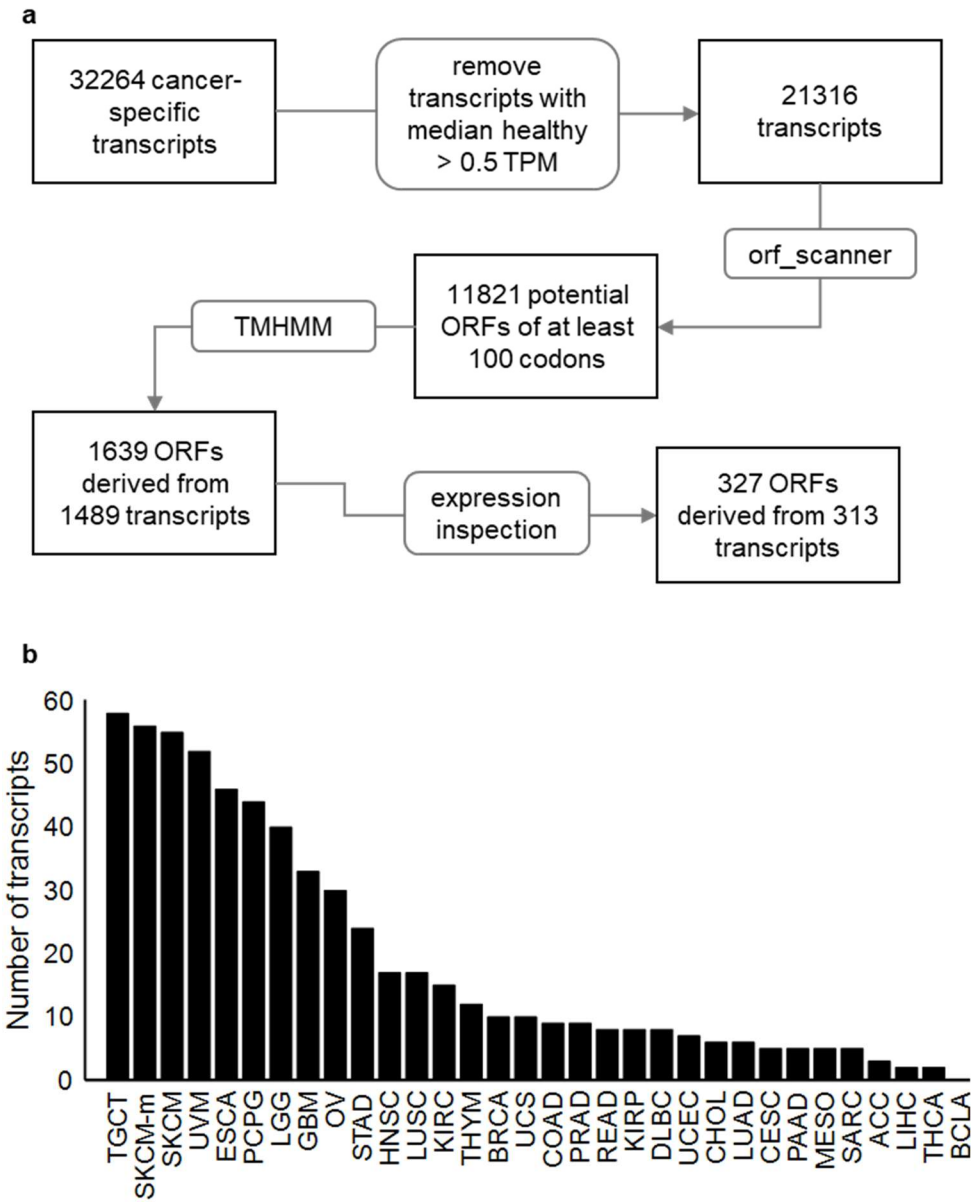


Figure 23: Selection of the candidate cancer-specific transmembrane-domain coding transcripts. a. The filtering steps used to select the candidate transcripts alongside the number of transcripts kept at each stage. Manual inspection of the raw expression values of each transcript in every tissue type was performed to ensure healthy tissues had few outlying values (expression inspection). **b.** The cancers that the 313 selected transcripts were upregulated in, some transcripts are counted multiple times if upregulated in multiple cancer types.

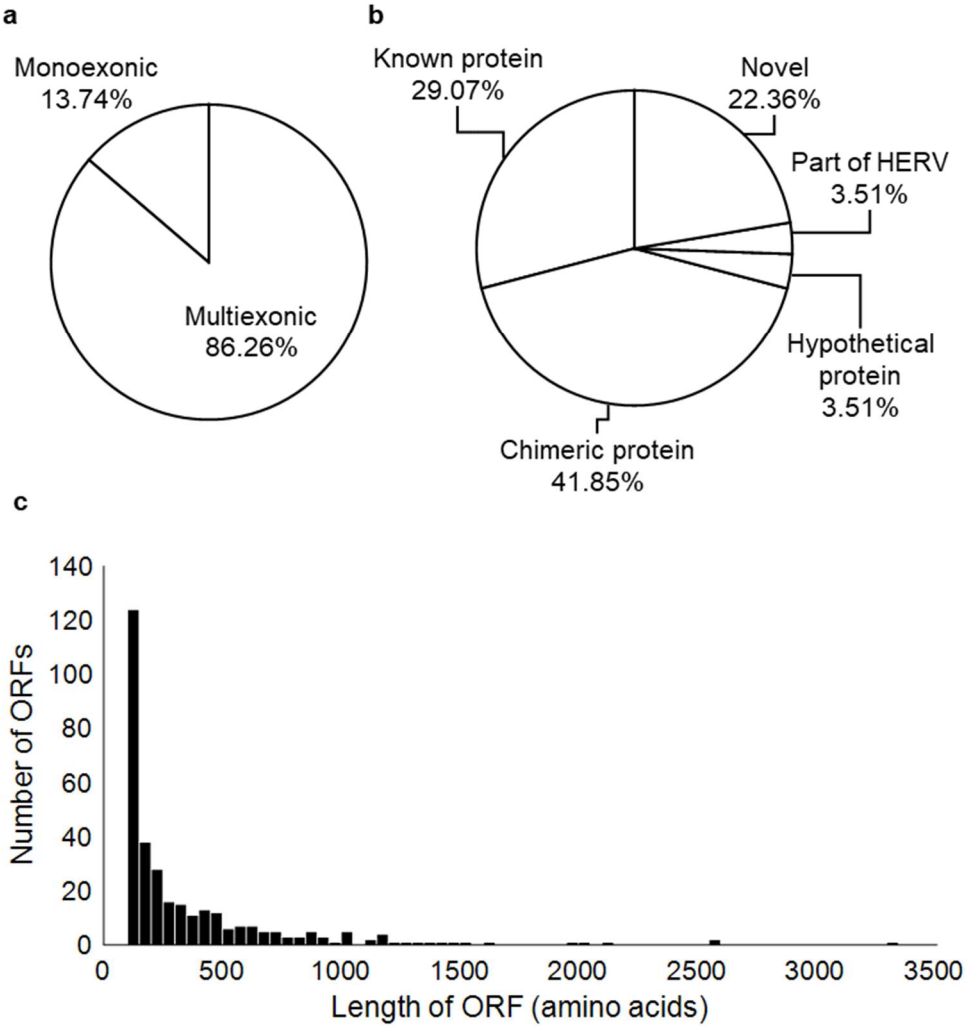


Figure 24: Structure of the candidate cancer-specific transmembrane-domain coding transcripts. a. The proportions of monoexonic and multiexonic transcripts. **b.** The proportion of already annotated sequence represented by the candidate transcripts. **c.** A histogram of the length of each of the 327 ORFs derived from the 313 candidate transcripts (using a bin width of 50 AA).

The selected transcripts represented both known and novel sequences (Figure 24b, 2.3.3: Selection of transmembrane domain containing candidates). Of the 313 transcripts, 131/313 partially overlapped a known sequence with additional sequence derived from either RTEs, intergenic, or intragenic regions, with a range of effects on the canonical sequence including truncating the peptide and addition of RTE-derived peptide to the ORF. A small number of the selected transcripts (10/313) directly overlapped a HERV, and the ORF predicted to contain a transmembrane domain had homology with HERV Env proteins. Furthermore, 70/313 transcripts were comprised of completely novel sequence, not overlapping an exon of a known or hypothetical gene, and not overlapping any known open reading frame of a HERV element. Alongside 11/313 transcripts representing, either partially or fully, hypothetical gene sequences. Additionally, 91 were transcripts which represented, partially or fully, known gene sequences with no unique regions. In some cases these were lncRNAs with predicted ORFs, in others known genes coding for transmembrane domain proteins which were ectopically expressed in cancer. The presence of known transmembrane domain containing proteins in the candidate list lent confidence to the transmembrane domain predictions.

For example, a novel transcript isoform of *gamma-aminobutyric acid A receptor alpha 3 subunit* (*GABRA3*) was identified (Figure 25a), which codes for the same transmembrane domain containing ORF (Figure 25b) as the canonical isoform. Canonical *GABRA3* is expressed in brain tissue, brain lower grade glioma, and glioblastoma multiforme samples (Figure 26). Alongside other subunits, the protein forms a gamma-aminobutyric acid (GABA)-responsive chloride channel. GABRA3 RNA and protein expression have been associated with increased metastasis and cell proliferation alongside poor patient survival in breast cancer (Gumireddy et al., 2016), hepatocellular carcinoma (Liu et al., 2008), lung cancer (Liu et al., 2016; Liu et al., 2009; Loriot et al., 2014), pancreatic cancer (Long et al., 2017), and melanoma (Loriot et al., 2014). The *GABRA3* transcript, alongside coding for the GABRA3 protein, also harbours miRNA sequences *miR-105* and *miR-767* in the first intron (Loriot et al., 2014). *MiR-105* promotes metastasis

through weakening vascular endothelial barriers (Zhou et al., 2014). *MiR-767* inhibits TET2 which regulates DNA methylation levels and increased expression of *miR-767* is associated with increased cell proliferation and invasiveness (Jia et al., 2020; Zhou et al., 2014).

The isoform identified by the *de novo* transcriptome assembly appears to use a separate promoter to the canonical contained within a *LINE1* and splices into a second novel exon within an *AluJb* before splicing into the second canonical exon of *GABRA3* (Figure 25a). The novel isoform is predicted to lead to ectopic expression of the same protein in LUSC, testicular germ cell tumours (TGCT), and both primary and metastatic SKCM samples (Figure 26).

Another cancer-specific isoform of GABRA3 has previously been annotated (Loriot et al., 2014). From 5' RACE data in melanoma cell lines an isoform driven through the bidirectional promoter of the *melanoma antigen gene family member A3* (*MAGEA3*) was identified. This isoform skipped the first exon of GABRA3, splicing instead into the second exon, maintaining the canonical ORF and the miRNA sequences harboured in the first intron. The exons spliced into upstream of GABRA3 code for an upstream ORF likely to elicit transcript degradation through nonsense-mediated decay, thus no GABRA3 protein was detected in melanoma lines expressing this transcript (Loriot et al., 2014). This may be the true source of the transcript identified by the de novo transcriptome assembly, though RTE sequences are still used to form the transcript the expression may not be driven through these elements. There must however be a source of the ectopic GABRA3 protein which may not be produced from the upstream ORF-containing *MAGEA3*-associated transcript.

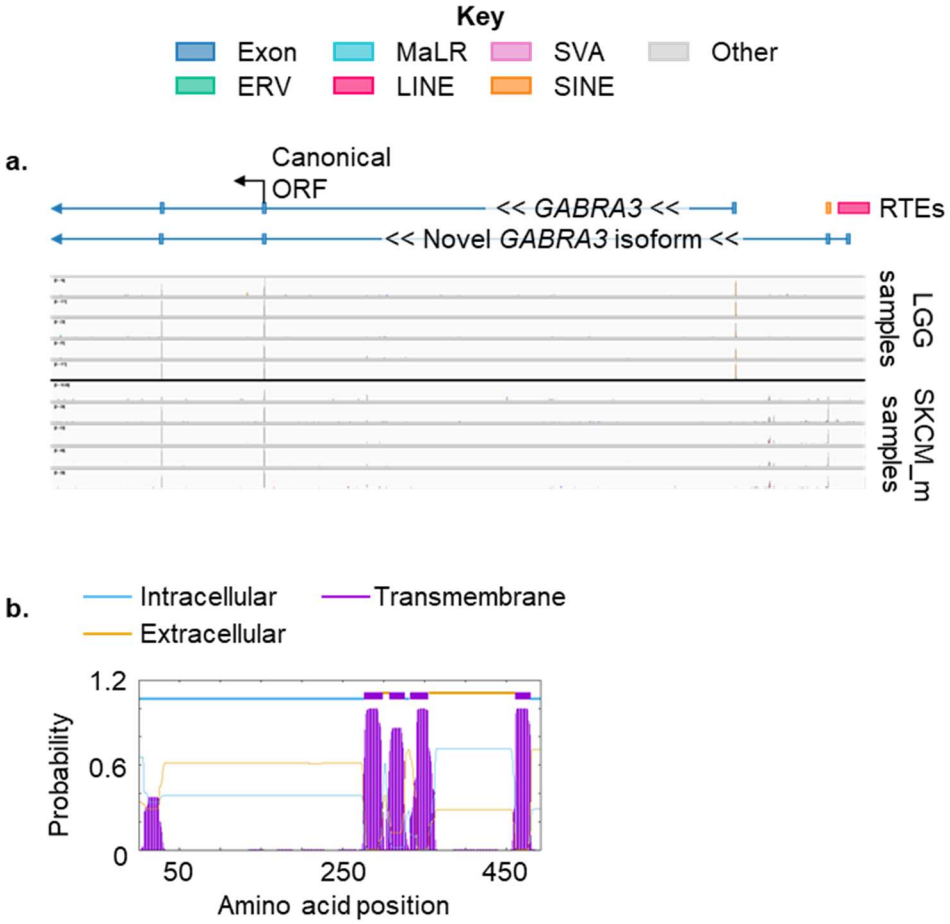


Figure 25: The structure of a novel *GABRA3* isoform identified by the *de novo* transcriptome assembly. a. The structure of the *GABRA3* locus and the novel isoform identified, alongside BAM files of RNAseq data from brain lower grade glioma (LGG) and metastatic skin cutaneous melanoma (SKCM_m) patient samples from TCGA. **b.** The TMHMM output showing the position of transmembrane domains within GABRA3.

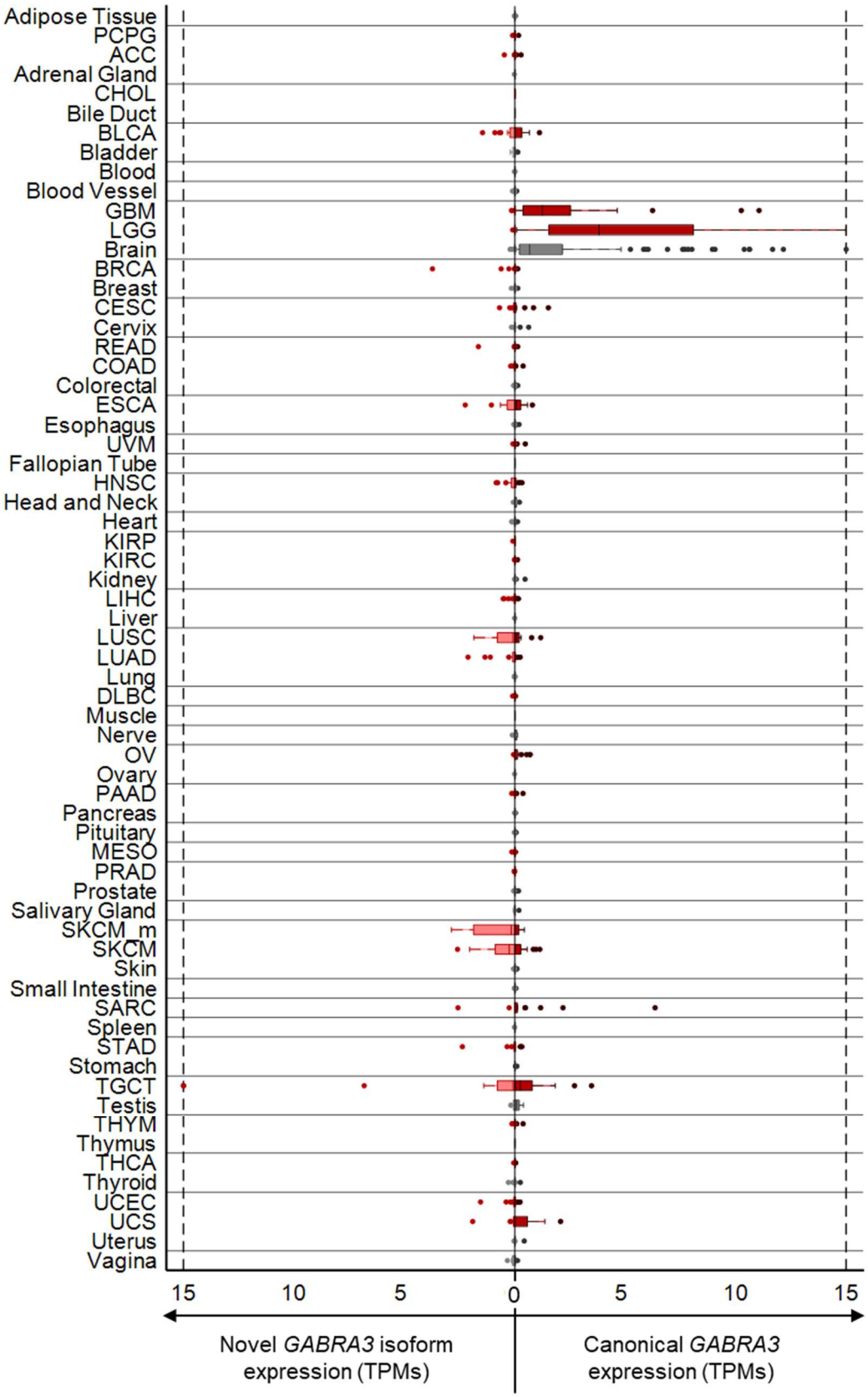

Adipose Tissue
PCPG
ACC
Adrenal Gland
CHOL
Bile Duct
BLCA
Bladder
Blood
Blood Vessel
GBM
LGG
Brain
BRCA
Breast
CESC
Cervix
READ
COAD
Colorectal
ESCA
Esophagus
UVM
Fallopian Tube
HNSC
Head and Neck
Heart
KIRP
KIRC
Kidney
LIHC
Liver
LUSC
LUAD
Lung
DLBC
Muscle
Nerve
OV
Ovary
PAAD
Pancreas
Pituitary
MESO
PRAD
Prostate
Salivary Gland
SKCM_m
SKCM
Skin
Small Intestine
SARC
Spleen
STAD
Stomach
TGCT
Testis
THYM
Thymus
THCA
Thyroid
UCEC
UCS
Uterus
Vagina
15
10
5
0
5
10
15
Novel GABRA3 isoform expression (TPMs)
Canonical GABRA3 expression (TPMs)

Figure 26: The expression of a novel *GABRA3* isoform identified by the *de novo* transcriptome assembly alongside the expression of canonical *GABRA3.* Expression values per patient are shown in a mirrored boxplot for TCGA and GTEx healthy (grey), and TCGA cancer (red) samples.Data are clipped to a maximum of 15 TPM.

Three candidate transcripts were selected for *in vitro* validation of protein stability and localisation (2.5: Preparation of stably transduced cell lines, 2.6: Sample preparation and Western Blot, 2.7: Flow cytometry). As highest confidence was placed in stability of transmembrane domains from genes known to code for transmembrane domain proteins and sequences with highest antigenicity were likely to be derived from RTEs, two candidates selected were truncations of known proteins with addition of RTE-derived peptides. A third candidate, tested for stability by Dr Jane Loong, was derived from a transcript overlapping a *HERV-H* but with no homology to known HERV proteins.

5.3.2 A novel truncated isoform of *ENPP3*

A novel isoform of the *ectonucleotide pyrophosphatase/phosphodiesterase 3* (*ENPP3*) was identified, with specific expression in KIRC (Figure 27a and Figure 28). The canonical form of *ENPP3* is expressed on a range of epithelial and mucosal cells, as well as on mast cells and basophils (Bühring et al., 2004). Canonical ENPP3 cleaves extracellular ATP (Tsai et al., 2015) and cGAMP (Mardjuki et al., 2024), and intracellular UDP-GlcNAc (Korekane et al., 2013). Cleavage of ATP regulates basophil and mast cell responses (Tsai et al., 2015). Alteration of ATP concentration in tumours may influence Treg function, as ATP has been shown to inhibit Tregs (Schenk et al., 2011), thus increased ENPP3 expression may reduce inflammation. Some cancer cell lines continuously release cGAMP due to mis-segregation of DNA (Carozza et al., 2020; Mackenzie et al., 2017), if this also occurs in tumours ENPP3 may further reduce inflammation through cleavage of cGAMP therefore reducing STING activation (Mardjuki et al., 2024; Wang et al., 2023). An *LTR* element upstream of *ENPP3* has been shown to control expression of the gene in a HIF-dependent manner (Siebenthall et al., 2019). *ENPP3* is expressed in healthy kidney tissue and is further upregulated in KIRC (Doñate et al., 2016; Thompson et al., 2018) (Figure 28). Due to the upregulation in KIRC, an antibody drug conjugate against ENPP3 has been tested in Phase I clinical trials, but although some patients responded there were dose-limiting reversible effects to patient corneas (Thompson et al., 2018). The novel *ENPP3* isoform identified codes for a truncated ENPP3 protein

(Figure 27b). While the canonical protein is 875 AA with only the first 22 AA localised intracellularly and 829 AA localised extracellularly, the truncated form is 492 AA long with the first 471 AA identical to the canonical and the final 21 AA donated by a *LINE2* along with a stop codon (Figure 27a). These final 21 AA may be displayed outside the cell, though were predicted by TMHMM to form a second transmembrane helix (Figure 27b). Additionally, the truncation of the protein would display peptide sequences and structures that although are contained within the canonical form would not be available for antibody or BCR binding of the canonical form. The truncated isoform had lesser expression in KIRC compared to the canonical isoform, but the expression was more cancer-specific and still very high (Figure 28). Considering isoform expression in TCGA and GTEx across adjacent healthy and healthy tissue types, 49% (396/811) of samples have canonical *ENPP3* expression over 0.5 TPM, whereas only 3.7% (30/811) express the truncated isoform above this threshold, reducing the likelihood of on-target toxic effects. However, Western blot of cell clone lysates transduced with the FLAG-tagged coding sequence for either the canonical or truncated *ENPP3* isoforms (see Supplementary figure 1 for transduction efficiency of clones, 2.5: Preparation of stably transduced cell lines, 2.6: Sample preparation and Western Blot) showed the protein derived from the truncated isoform was not detectable in HEK293T cells (Figure 27c) and was therefore unlikely to be stable.

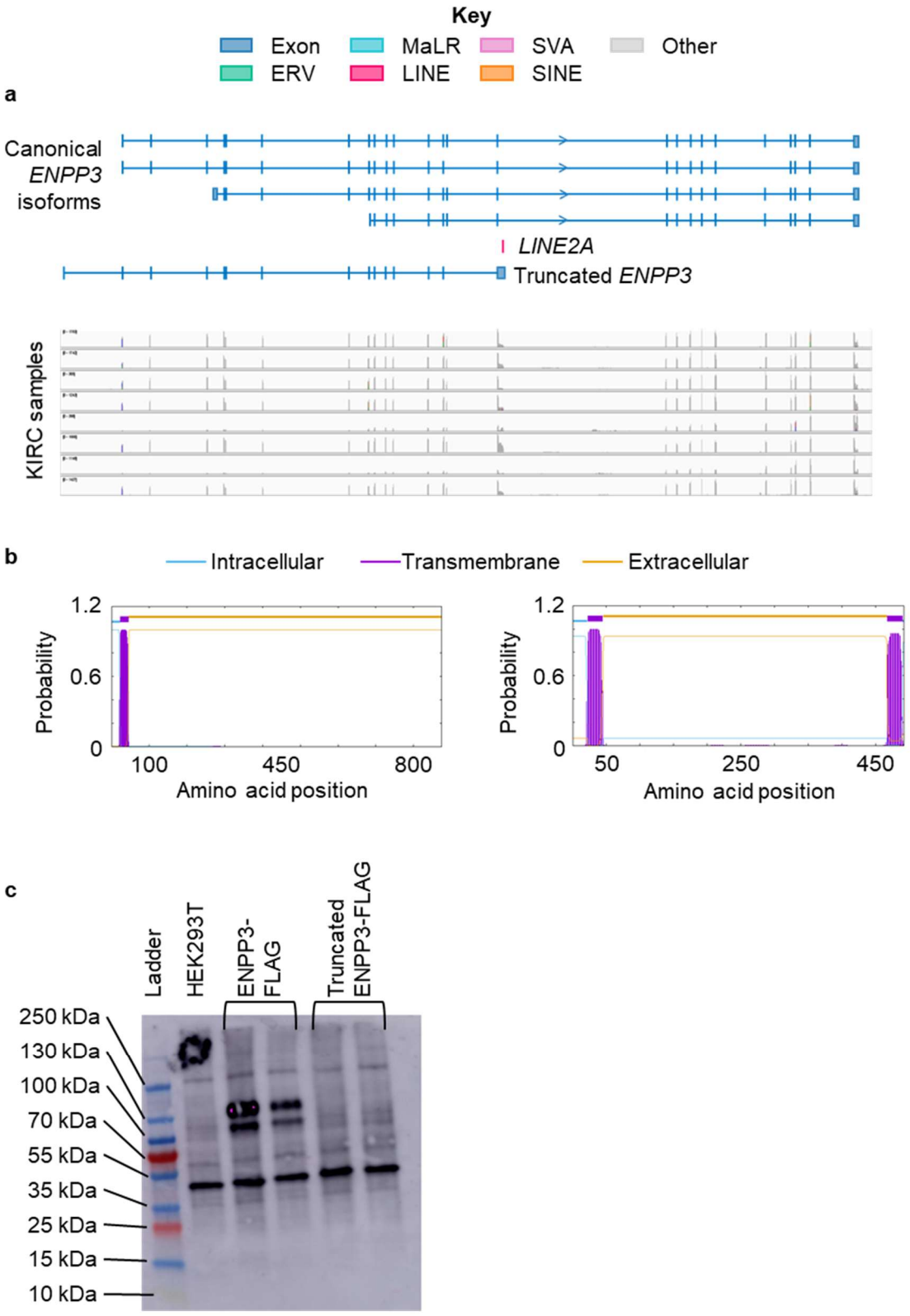

Key
Exon
MaLR
SVA
Other
ERV
LINE
SINE
a
Canonical ENPP3 isoforms
LINE2A
Truncated ENPP3
KIRC samples
b
Intracellular
Transmembrane
Extracellular
Probability
1.2
0.6
0
100
450
800
Amino acid position
50
250
450
c
Ladder
HEK293T
ENPP3-FLAG
Truncated ENPP3-FLAG
250 kDa
130 kDa
100 kDa
70 kDa
55 kDa
35 kDa
25 kDa
15 kDa
10 kDa

Figure 27: The structure and stability of a novel *ENPP3* isoform identified by the *de novo* transcriptome assembly. a. The structure of the *ENPP3* locus and the novel isoform identified, alongside BAM files of RNAseq data from kidney renal clear cell carcinoma (KIRC) patient samples from TCGA. **b.** The TMHMM output showing the position of transmembrane domains within ENPP3 (left) and the potential peptide of truncated ENPP3 (right). **c.** A Western blot (with help from Dr Laura Doglio) showing the stability of canonical ENPP3-FLAG (with an estimated molecular weight of 100 kDa), however the truncated ENPP3-FLAG (with an estimated molecular weight of 54 kDa) could not be detected.

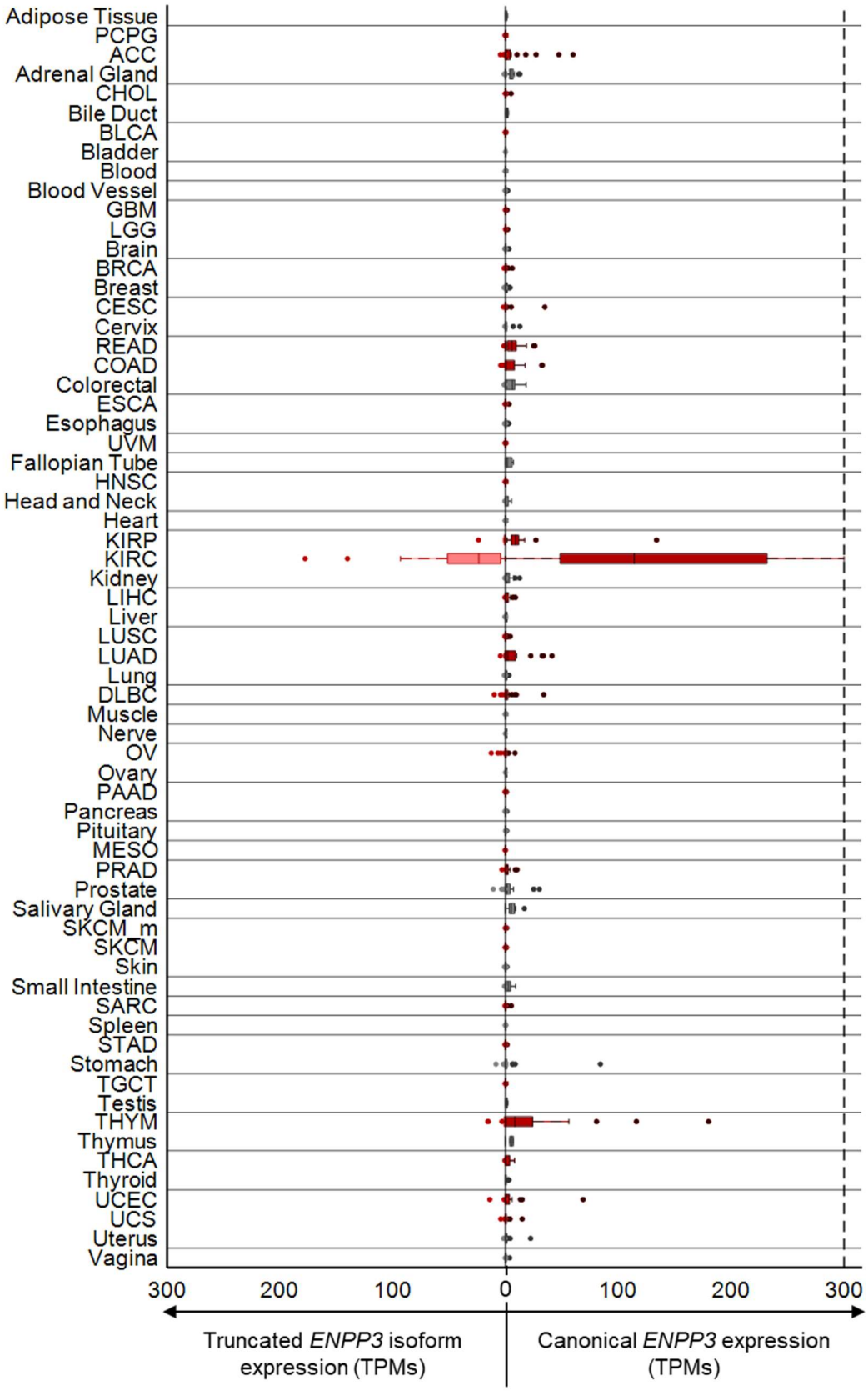

Adipose Tissue
PCPG
ACC
Adrenal Gland
CHOL
Bile Duct
BLCA
Bladder
Blood
Blood Vessel
GBM
LGG
Brain
BRCA
Breast
CESC
Cervix
READ
COAD
Colorectal
ESCA
Esophagus
UVM
Fallopian Tube
HNSC
Head and Neck
Heart
KIRP
KIRC
Kidney
LIHC
Liver
LUSC
LUAD
Lung
DLBC
Muscle
Nerve
OV
Ovary
PAAD
Pancreas
Pituitary
MESO
PRAD
Prostate
Salivary Gland
SKCM_m
SKCM
Skin
Small Intestine
SARC
Spleen
STAD
Stomach
TGCT
Testis
THYM
Thymus
THCA
Thyroid
UCEC
UCS
Uterus
Vagina
300
200
100
0
100
200
300
Truncated ENPP3 isoform expression (TPMs)
Canonical ENPP3 expression (TPMs)

Figure 28: The expression of a novel *ENPP3* isoform identified by the *de novo* transcriptome assembly alongside the expression of canonical *ENPP3.* Expression values per patient are shown in a mirrored boxplot for TCGA and GTEx healthy (grey), and TCGA cancer (red) samples. Data are clipped to a maximum of 300 TPM.

A novel truncated isoform of *PLD3*

A novel isoform of *phospholipase D family member 3* (*PLD3*) was also identified, with specific expression in uterine carcinosarcoma (UCS) (Figure 29a, Figure 30). The canonical form of PLD3 is a lysosomal protein which is localised to the endoplasmic reticulum, Golgi apparatus, and early endosomal membranes (Gonzalez et al., 2018). In lysosomes the protein is cleaved near the transmembrane domain releasing the large catalytic domain into the lysosome (Gonzalez et al., 2018). The catalytic domain cleaves single stranded RNA and DNA, preventing accumulation which would lead to continuous activation of toll-like receptors (TLRs) TLR9 and TLR7, and potential autoimmunity (Gavin et al., 2021; Gavin et al., 2018). For trafficking into endosomes, PLD3 must be trafficked first to the cell membrane (Gonzalez et al., 2018).

The novel *PLD3* isoform identified codes for a truncated PLD3 protein (Figure 29a). The canonical *PLD3* isoform codes for a 490 AA protein, while the truncated isoform codes for a 263 AA protein with the first 227 AA identical to the canonical and 36 amino acids donated by an *AluJr*. Only the first 37 AA of both proteins would be localised intracellularly if localised to the cell surface membrane, with the novel 36 AA donated by the *AluJr* available for antibody binding extracellularly (Figure 29b), as well as any peptides revealed by the protein truncation.
Western blot of cell population lysates transduced with the HA-tagged coding sequence for the truncated *PLD3* isoform showed the protein derived from this sequence was stable in HEK293T cells (Figure 31a) (see Figure 31b for transduction efficiency of the cell population, 2.5: Preparation of stably transduced cell lines, 2.6: Sample preparation and Western Blot, 2.7: Flow cytometry). This was also shown through flow cytometry, where intracellular staining of the transduced cell population was seen, however extracellular staining showed no detectable surface protein in HEK293T cells (Figure 31b). This protein may localise to intracellular membranes instead along with the canonical PLD3.

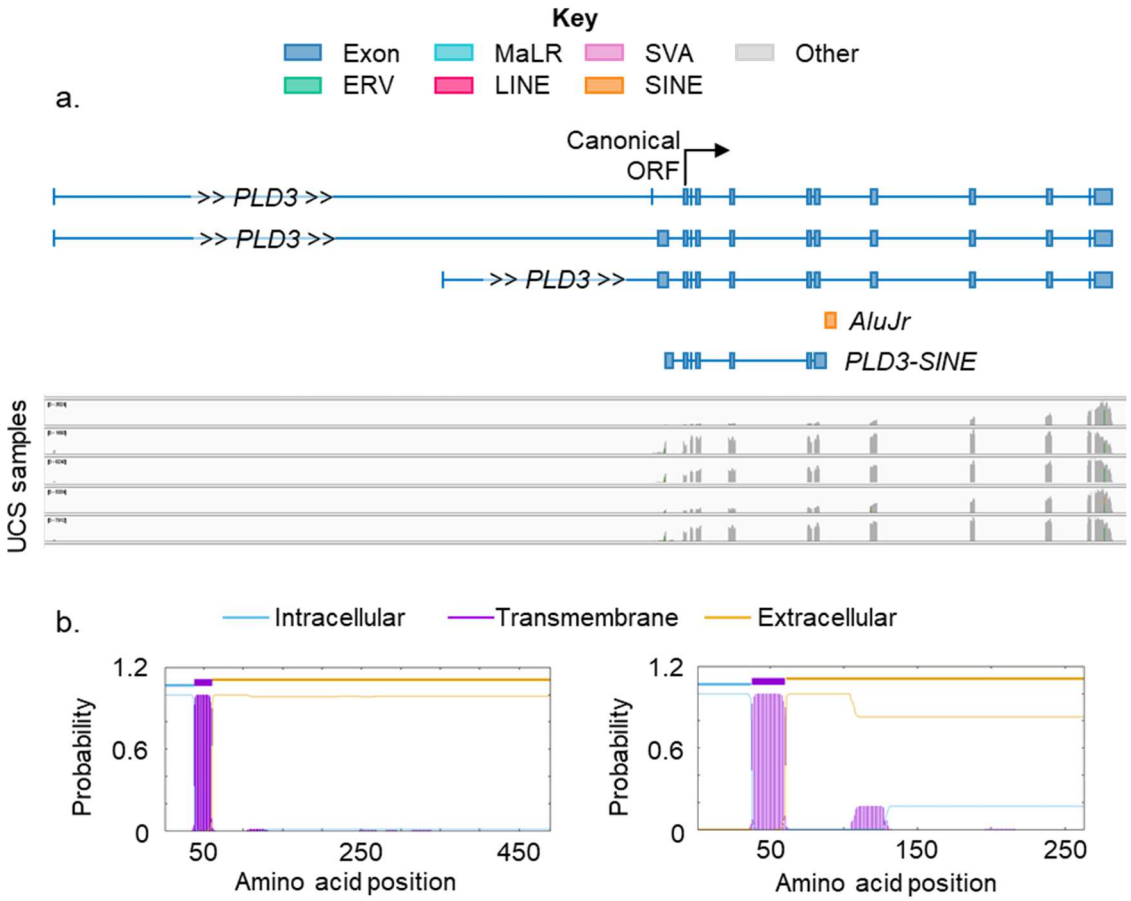


Figure 29: The structure of a novel *PLD3* isoform identified by the *de novo* transcriptome assembly. **a.** The structure of the *PLD3* locus and the novel isoform identified, alongside BAM files of RNAseq data from uterine carcinosarcoma patient samples from TCGA. **b.** The TMHMM output showing the position of transmembrane domains within PLD3 (left) and the potential peptide of truncated PLD3 (right).

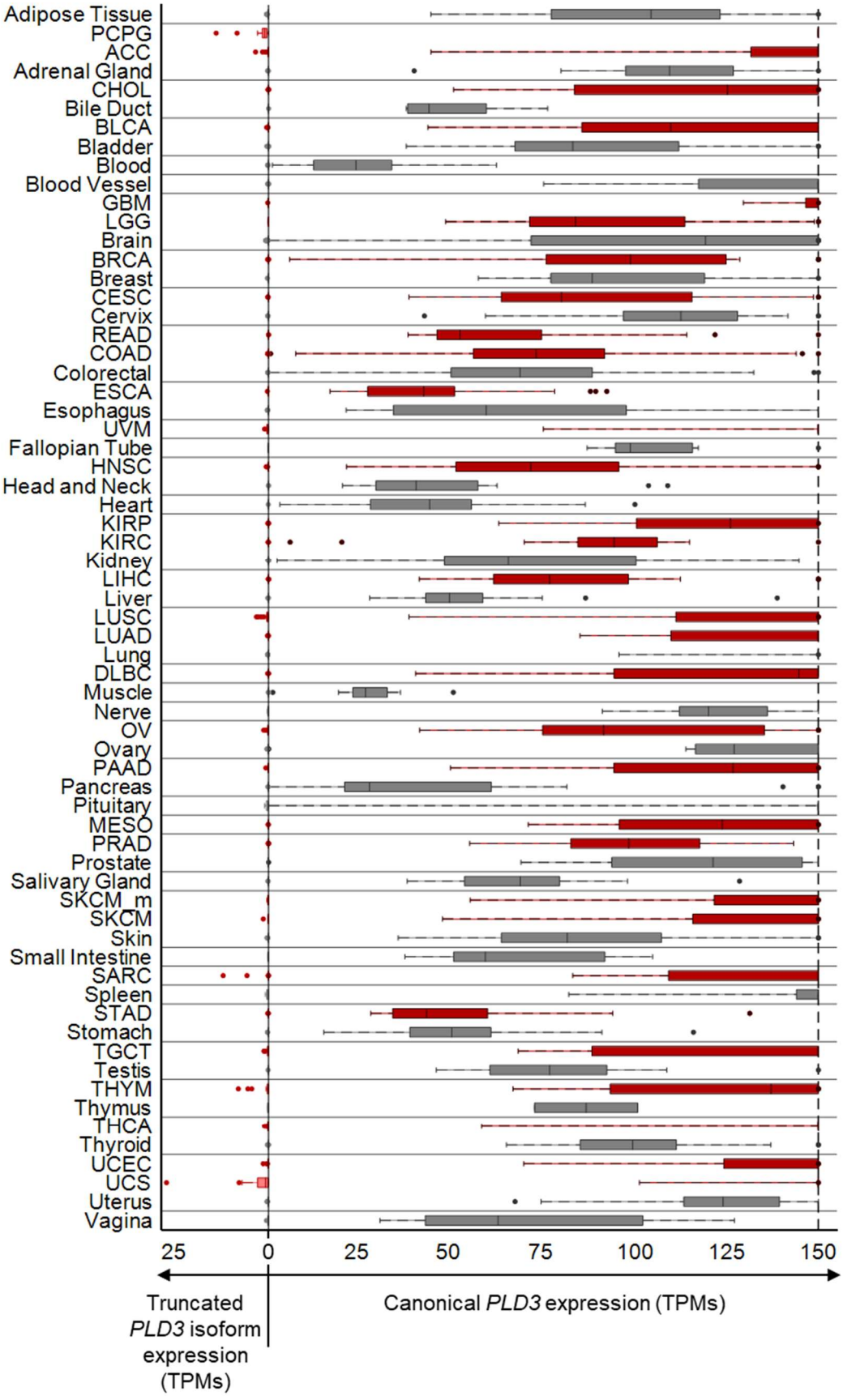

Adipose Tissue
PCPG
ACC
Adrenal Gland
CHOL
Bile Duct
BLCA
Bladder
Blood
Blood Vessel
GBM
LGG
Brain
BRCA
Breast
CESC
Cervix
READ
COAD
Colorectal
ESCA
Esophagus
UVM
Fallopian Tube
HNSC
Head and Neck
Heart
KIRP
KIRC
Kidney
LIHC
Liver
LUSC
LUAD
Lung
DLBC
Muscle
Nerve
OV
Ovary
PAAD
Pancreas
Pituitary
MESO
PRAD
Prostate
Salivary Gland
SKCM_m
SKCM
Skin
Small Intestine
SARC
Spleen
STAD
Stomach
TGCT
Testis
THYM
Thymus
THCA
Thyroid
UCEC
UCS
Uterus
Vagina
25
0
25
50
75
100
125
150
Truncated PLD3 isoform expression (TPMs)
Canonical PLD3 expression (TPMs)

Figure 30: The expression of a novel *PLD3* isoform identified by the *de novo* transcriptome assembly alongside the expression of canonical *PLD3*. Expression values per patient are shown in a mirrored boxplot for TCGA and GTEx healthy (grey), and TCGA cancer (red) samples. Data are clipped to a maximum of 150 TPM.

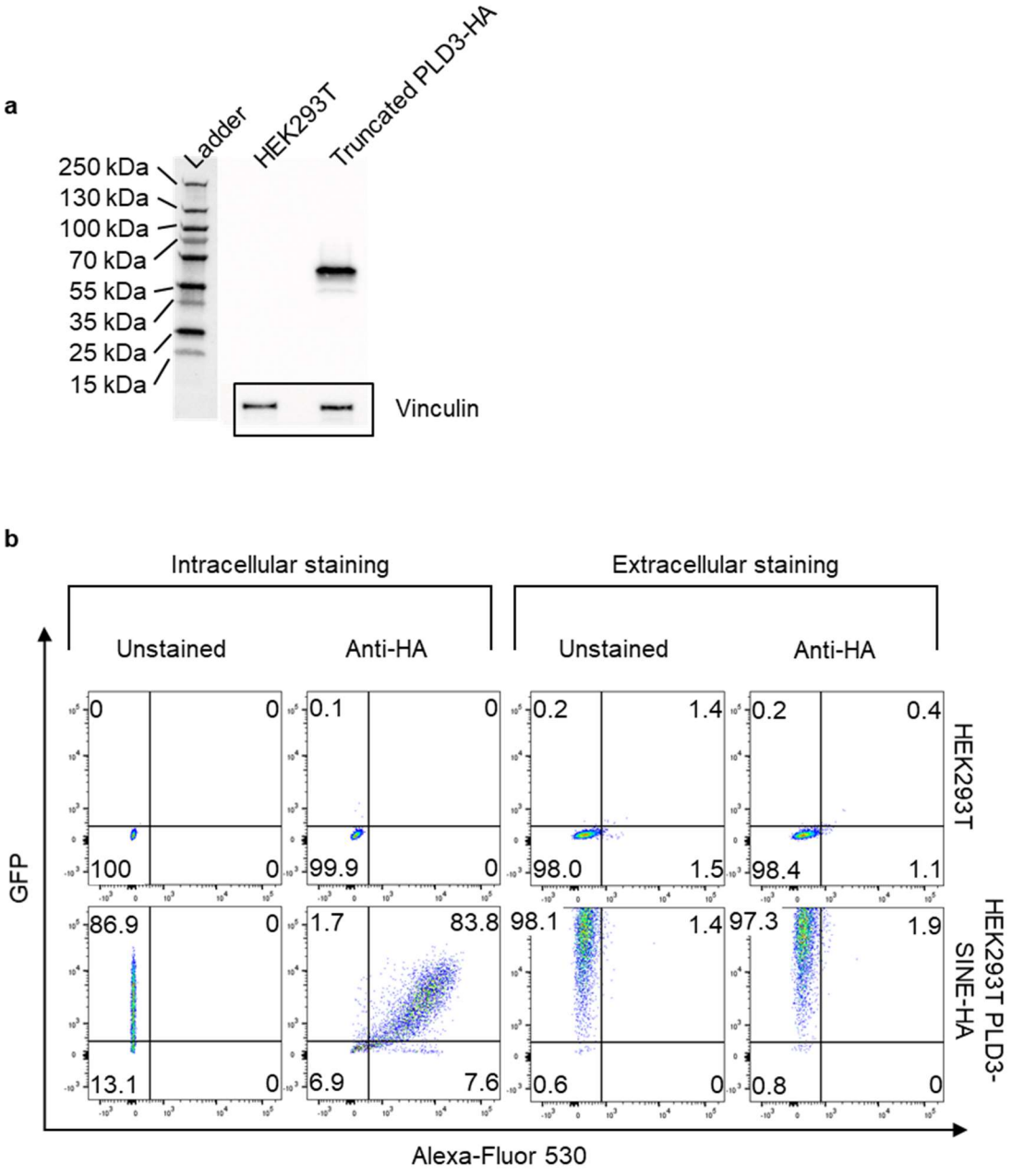

a
Ladder
HEK293T
Truncated PLD3-HA
250 kDa
130 kDa
100 kDa
70 kDa
55 kDa
35 kDa
25 kDa
15 kDa
Vinculin
b
Intracellular staining
Extracellular staining
Unstained
Anti-HA
Unstained
Anti-HA
0
0
0.1
0
0.2
1.4
0.2
0.4
100
0
99.9
0
98.0
1.5
98.4
1.1
86.9
0
1.7
83.8
98.1
1.4
97.3
1.9
13.1
0
6.9
7.6
0.6
0
0.8
0
GFP
HEK293T
HEK293T PLD3-SINE-HA
Alexa-Fluor 530

Figure 31: The stability and localisation of the truncated PLD3 protein produced by the novel *PLD3* isoform identified by the *de novo* transcriptome assembly. a. A Western blot showing the stability of the truncated PLD3-HA (PLD3-SINE-HA) protein (with an estimated molecular weight of 34 kDa). **b.** Flow cytometry showing extracellular staining (left) and intracellular staining (right) of parental HEK293T cells and transduced HEK293T cells. Successfully transduced cells are marked with GFP (Figure 6), and the anti-HA antibody was visualised with a secondary antibody conjugated to Alexa-Fluor 530.

5.3.4 A HERVH-derived transcript

Additionally, a fully novel transcript overlapping a *HERV-H* on chromosome 13 was identified, with specific expression in oesophageal carcinoma (ESCA), head and neck squamous cell carcinoma (HNSC), and stomach adenocarcinoma (STAD) (Figure 32a, Figure 33). This transcript does not overlap a canonical open reading frame within the *HERV-H*, and the derived protein has no homology to any known protein (BLASTp, with minimum homology of 85%). Within this novel 2626 nucleotide transcript a 116 AA protein was identified which was predicted to contain a transmembrane domain, with the first 53 AA predicted to be displayed outside the cell if the protein localised to the surface plasma membrane (Figure 32b)

Western blot of cell population lysates transduced with the HA-tagged coding sequence for the open reading frame, with cells prepared and samples analysed by Dr Jane Loong (2.5: Preparation of stably transduced cell lines, 2.6: Sample preparation and Western Blot), showed the protein derived from this sequence was not detectable in HEK293T cells (Figure 32c) and therefore unlikely to be stable.

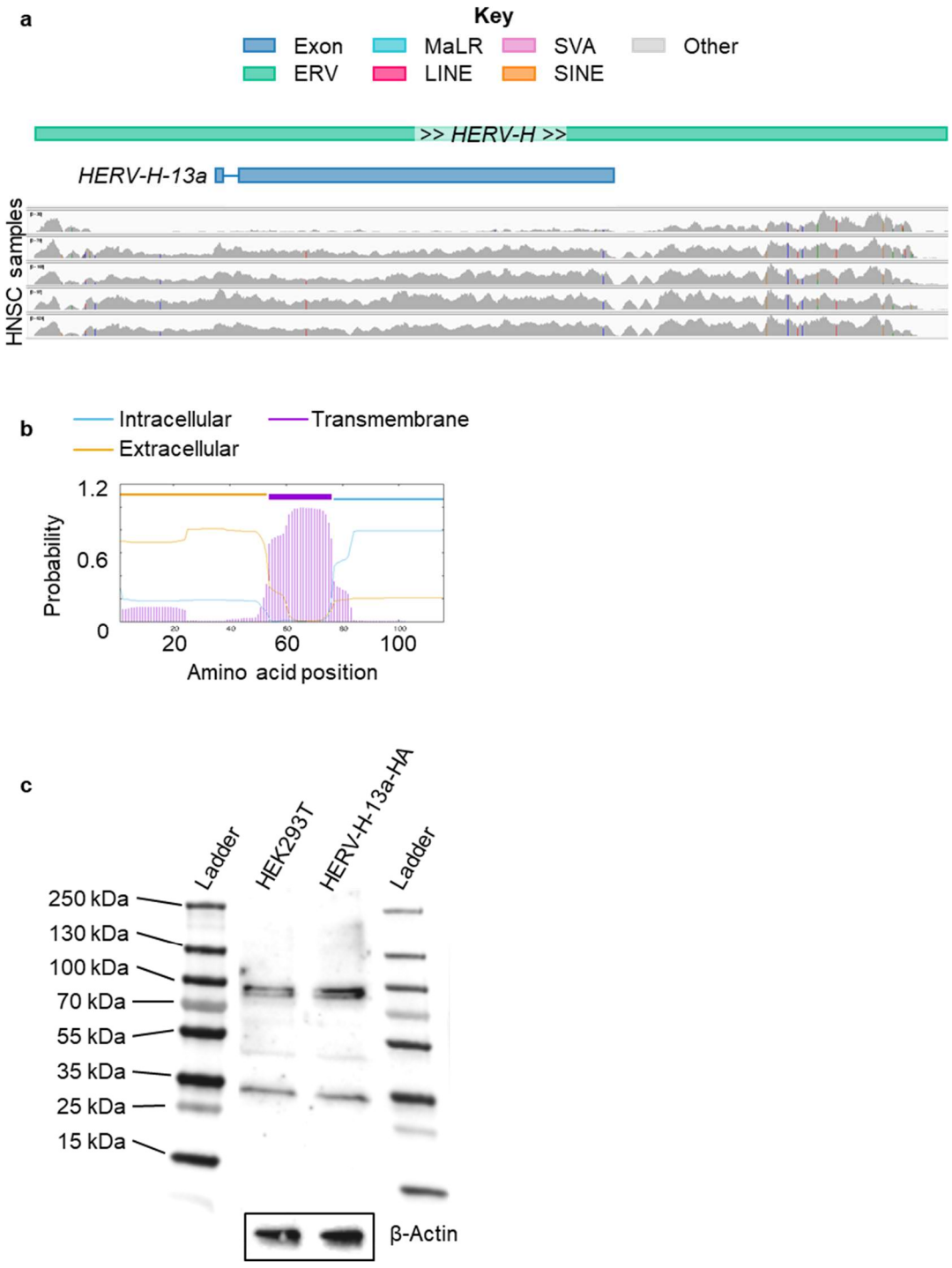


Figure 32: The structure and stability of a novel HERV-H-derived transcript identified by the *de novo* transcriptome assembly. **a.** The structure of the HERV-H locus and the novel transcript identified, alongside BAM files of RNAseq data from head and neck squamous cell carcinoma (HNSC) patient samples from TCGA. **b.** The TMHMM output showing the position of the transmembrane domain within the derived protein. **c.** A Western blot (Dr Jane Loong) showing the lack of stability of the HA-tagged HERV-H-derived protein (with an estimated molecular weight of 16 kDa).

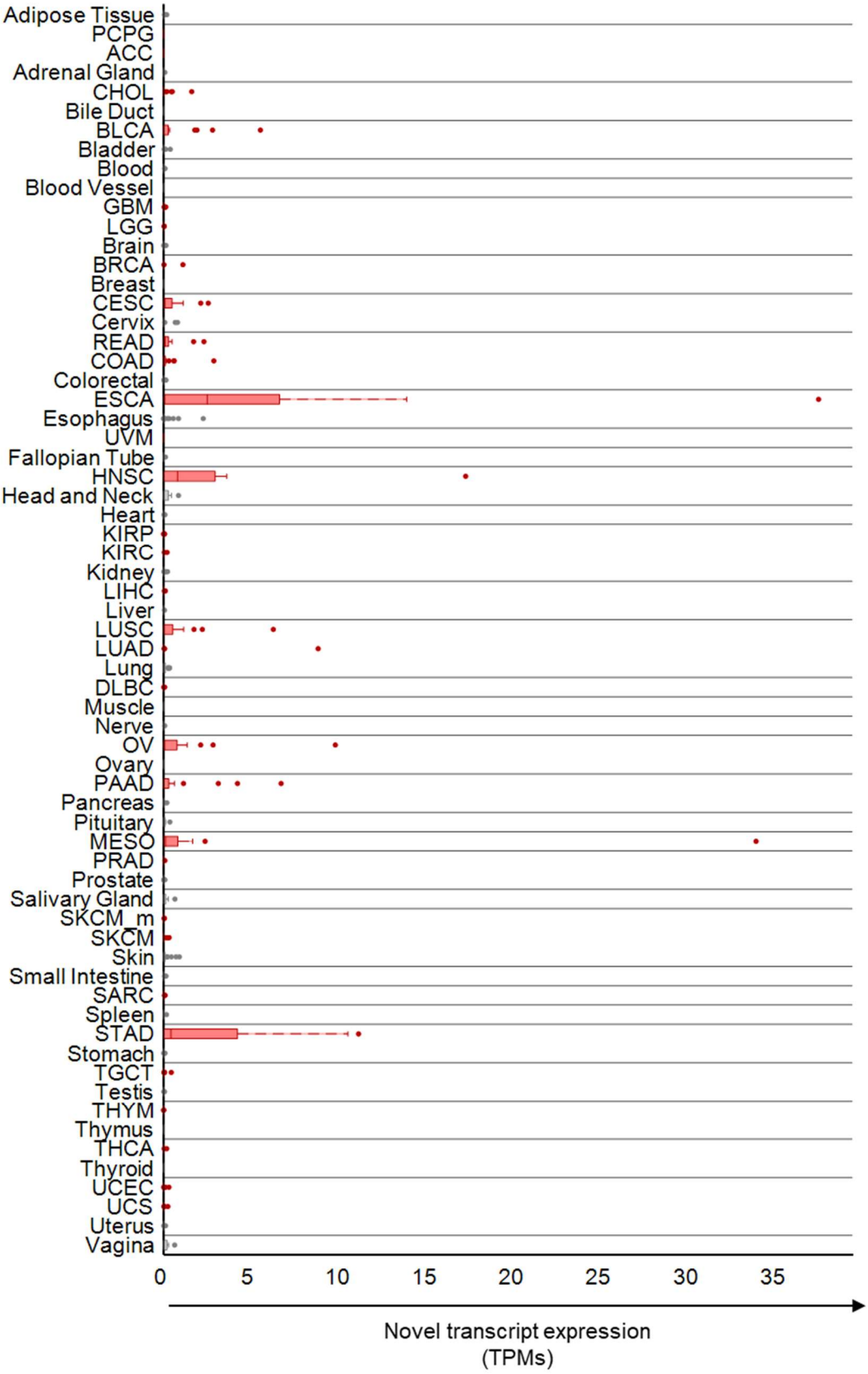

Adipose Tissue
PCPG
ACC
Adrenal Gland
CHOL
Bile Duct
BLCA
Bladder
Blood
Blood Vessel
GBM
LGG
Brain
BRCA
Breast
CESC
Cervix
READ
COAD
Colorectal
ESCA
Esophagus
UVM
Fallopian Tube
HNSC
Head and Neck
Heart
KIRP
KIRC
Kidney
LIHC
Liver
LUSC
LUAD
Lung
DLBC
Muscle
Nerve
OV
Ovary
PAAD
Pancreas
Pituitary
MESO
PRAD
Prostate
Salivary Gland
SKCM_m
SKCM
Skin
Small Intestine
SARC
Spleen
STAD
Stomach
TGCT
Testis
THYM
Thymus
THCA
Thyroid
UCEC
UCS
Uterus
Vagina
0
5
10
15
20
25
30
35
Novel transcript expression
(TPMs)

Figure 33: The expression of a novel HERV-H-derived transcript identified by the *de novo* transcriptome assembly. Expression values shown for TCGA and GTEx healthy (grey), and TCGA cancer (red) samples.

5.4 Discussion

It is possible that transmembrane domain containing proteins with peptide donated by RTEs can be stable. Although the truncated ENPP3 isoform and the HERV-H-derived protein were not stable when transduced into HEK293T cells, the truncated PLD3 was. It is possible the RNA of the two candidates that were not stable at the protein level has some function, however for the aims of this project stable protein was required. The instability of truncated ENPP3 aligns with previous studies of mouse nucleotide pyrophosphatase/phosphodiesterase 1 (NPP1), a paralogue of ENPP3, where the orthologous domains removed in this truncation were shown to be required for stability and localisation of NPP1 (Gijsbers et al., 2003). Although the truncated PLD3 was stable the protein did not localise in detectable levels to the cell surface plasma membrane. Some surface localisation of the canonical form had been seen previously, though this was rare (Gonzalez et al., 2018). The truncation of PLD3 may still have implications for the cell as the domains removed are required for catalysing cleavage of single stranded RNA and DNA (Gavin et al., 2018). Knock-out of PLD3 in mice led to development of fatal liver inflammation, which was rescued when all TLRs required for activation of the immune response against accumulation of single stranded RNA and DNA in the cell were concomitantly knocked out (Gavin et al., 2021; Gavin et al., 2018). However, the expression of the truncated PLD3 is much lower than the canonical form (Figure 30), and it is unknown to what extent the truncated form would allow for accumulation of oligonucleotides and the following inflammation. Other work has also shown that peptides containing RTE-derived sequences can produce stable protein through mass spectrometry analysis of cell lysates and *in vitro* testing (Burbage et al., 2023; Merlotti et al., 2023; Ng et al., 2019; Shah et al., 2023). Two additional transmembrane domain proteins with RTE sequences appended have also been shown to be stable and localise to the cell surface membrane, with the potentially-antigenic RTE sequences displayed outside the cell. L1PA2_GABRG2 and THE1C_TMEM260 with 64 and 4 AA derived from RTEs added to the start of the canonical proteins respectively, were both shown to be stable in Western blot analysis of cell membranes and through immunofluorescence staining of natively

expressing cells (Shah et al., 2023). Neither of these transcripts were assembled in this *de novo* transcriptome assembly.

The ORFs derived from some candidate transcripts may not be translated. Many of the candidate ORFs selected here are very short compared to the majority of canonical proteins. Previous work surveying ORFs in prokaryotic and eukaryotic genomes showed the average protein length in eukaryotes to be 472 AA, with a lack of proteins below 100 AA and a preference for those over 250 AA long (Tiessen et al., 2012). The majority of ORFs selected for in this analysis are under 250 AA (190/327), including the unstable HERVH-derived protein. Although there are examples of extremely small proteins active once complexes are formed (Tiessen et al., 2012), small ORFs are poorly conserved compared to the longer canonical ORFs (Couso and Patraquim, 2017). Though on the other hand, a long ORF derived from the truncated ENPP3 transcript was not stable. Additionally, analysis of the known human proteome has shown correlation between transcript length and protein size (Lopes et al., 2021). This is somewhat reflected in the selected candidates, however there are some candidates with very large transcripts containing very small ORFs, these are perhaps much less likely to be translated, such as the HERVH-derived transcript where a 116 AA protein was derived from a 2626 nucleotide transcript.

Accurate prediction of translatability of ORFs would make candidate selection more efficient. In order to prioritise candidates for *in vitro* stability testing a range of filters can be used including selecting for peptides with high predicted stability, peptides coded for by larger numbers of exons, or by the longest ORF within the transcript. But previous work on non-canonical peptide expression derived from lncRNAs has shown the peptides produced have different characteristics to the current known human proteome (Chen et al., 2020; Lu et al., 2019). The lncRNA-derived peptides were shorter, coded for by fewer exons, and were predicted to have lower stability and iso-electric points (Lu et al., 2019). Thus, to create an algorithm able to predict translatability of non-canonical ORFs based on knowledge of canonical ORFs may mean many real and stable proteins would

not pass the filtering. Additionally, some peptides which are not stable but are highly homologous to canonical proteins, such as the truncated ENPP3, may wrongly pass filtering. Peptides could also be searched for in mass spectrometry data, but including all possible ORFs would drastically increase the false discovery rate, added to difficulties distinguishing the canonical protein from isoforms with small unique regions of peptide derived from RTEs.

If the proteins are targetable, expression may be selected against. Complete regression of tumours requires continuous expression of the antigenic protein being targeted in all cancer cells. This situation is most likely to arise if the protein being targeted is required for cell survival within the tumour. However, only 4.1% of genes encoding cell surface proteins have been shown to be necessary for survival compared to 14% of genes encoding proteins localised elsewhere (Hu et al., 2021). Added to this, in selecting cancer-specific transcripts, it is likely all transcripts required for cell survival are removed as expression would also be needed in healthy cells. Alternatively, there may also be some proteins the cell is unable to downregulate. The reactivation of RTEs in cancer appears to be due to a combination of genome hypomethylation and TF availability. How much control the short-term evolution seen in tumours has over which specific RTE loci are active and are therefore able to influence the proteome is unknown. Further studies into the timeline of RTE activation in cancer development, and mechanisms involved in re-silencing specific loci are needed.

5.5 Conclusion

In summary, novel transcripts identified by the *de novo* transcriptome assembly have the potential to code for stable transmembrane domain containing proteins with RTE-derived peptides. Although, neither the truncated ENPP3 isoform or HERV-H-derived protein were stable, the truncated PLD3 with 36 AA donated by an *AluJr* element was stable. The localisation of the truncated PLD3 was not to the cell surface membrane, but the stability of this protein suggests the potential for stability of the other cancer-specific candidates expressed across patients

identified here which are yet to be tested. Those with highest potential to be stable are likely to be those with homology to known transmembrane domain proteins. Furthermore, this analysis has revealed a potential mechanism behind the upregulation of an ectopically expressed known transmembrane protein in cancer GABRA3, where the novel cancer-specific transcript isoform produces the same protein as the canonical brain-specific isoform, with pro-tumorigenic effects.

Chapter 6. Results 4: Exploration of HERV expression in metastatic KIRC

6.1 Aims

Sequences produced by RTEs may be antigenic driving an immune response against tumours (1.5.3: Expression of antigenic proteins). Certain HERV loci have been associated with spontaneous regression, response to immune checkpoint blockade, and cytotoxic T-cell signatures in KIRC (Panda et al., 2018; Rooney et al., 2015; Smith et al., 2018; Takahashi et al., 2008). In order to assess whether HERV expression associated with anti-PD-1 treatment of metastatic KIRC patients in the ADAPTeR study (2.1.4: Metastatic KIRC samples from the ADAPTeR study), expression of loci was analysed (Figure 34, 2.2.2.2: Expression of individual RTE loci, 2.4.4: Differential expression analysis for HERVs in metastatic KIRC). As loci previously associated with response were not mapped to GRCh38 these loci were mapped and matched to a custom Dfam library (Attig et al., 2017) (2.2.1: Annotation of HERV loci). Expression of these and other HERV loci from the Dfam library was then analysed in the RNAseq dataset from the ADAPTeR study, consisting of 60 tumour biopsies from 14 patients (Figure 34, 2.1.4: Metastatic KIRC samples from the ADAPTeR study). Additionally, HERV-derived transcript expression (Attig et al., 2019) was also analysed for association with response to anti-PD-1 therapy (2.2.2.1: Expression of transcripts assembled in the *de novo* transcriptome, 2.4.4: Differential expression analysis for HERVs in metastatic KIRC). Additionally, to ensure associations with response were not confounded by immune cell infiltrate, expression of HERV loci and derived transcripts was assessed in purified immune cell datasets (2.1.5: Purified immune cell datasets).

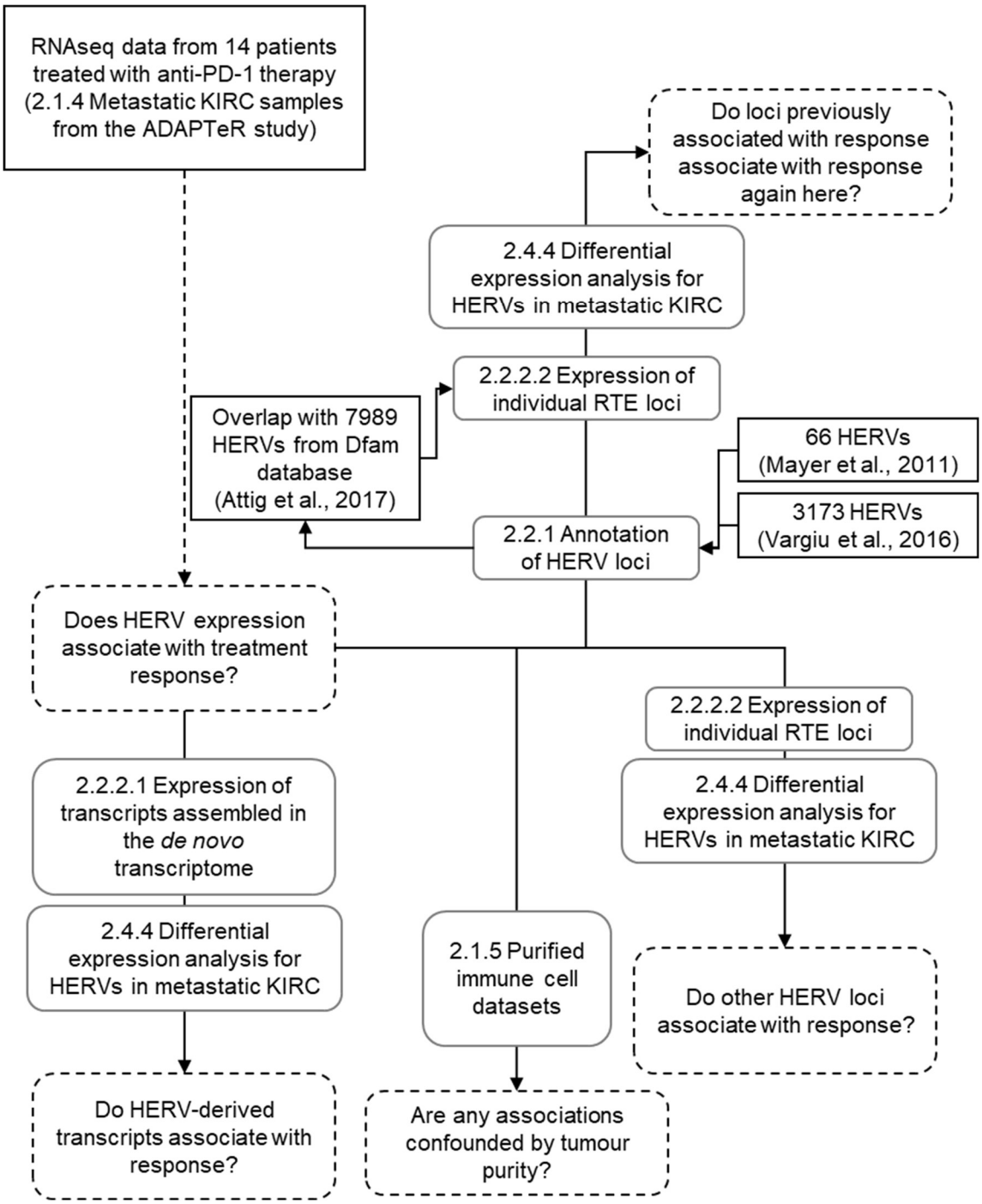


Figure 34: Aims for Results 4: Exploration of HERV expression in metastatic KIRC. Aims are shown in dashed boxes and methods are referenced in grey boxes.

6.2 Introduction

KIRC has been seen to spontaneously regress and respond to immune checkpoint blockade, but the antigen source is unknown. KIRC is the most common histological subtype of kidney cancer (Ricketts et al., 2018) and is highly immune infiltrated (Ricketts et al., 2018; Rooney et al., 2015; Thorsson et al., 2018). Spontaneous regression has been seen in rare cases (Cole and Everson, 1956; Janiszewska et al., 2013; Snow and Schellhammer, 1982), alongside response to checkpoint inhibitors seen even in metastatic disease (Albiges et al., 2019; Au et al., 2021; Motzer et al., 2015; Xu et al., 2020). But the source of antigen driving the immune response to the tumour is unknown. Expressed non-synonymous single nucleotide variants, tumour mutational burden, and frameshift insertions or deletions do not correlate with response to immunotherapy in KIRC (Au et al., 2021; Braun et al., 2020; McDermott et al., 2018; Motzer et al., 2020; Turajlic et al., 2017).

Previously, expression of HERVs in KIRC has been associated with response to immune checkpoint blockade and cytotoxic T-cell tumour infiltrate levels. In a study of patients with metastatic KIRC treated with hematopoietic stem cell transplantation, one patient had regression leading to survival of four years post-treatment. The donor T-cells in this patient responded to a tumour-restricted antigen derived from a *HERV-E* sequence (*ERVE-4* or *HERV 2256*) (Takahashi et al., 2008). Furthermore, in a pan-cancer analysis, *ERV3-2* expression has been associated with response to immune checkpoint blockade in 11 cancer types, with increased *ERV3-2* expression associated with response to anti-PD-1 therapy in primary KIRC (Panda et al., 2018). Expression of HERV-E, HERV-H, and HERV-K group members have been associated with increased cytotoxic T-cell signatures, both across various tumour types and in primary KIRC (Rooney et al., 2015). *HERV 4700* has also been shown to associate with immunotherapy response in KIRC, with Ribo-Seq data suggesting this locus may be translated (Smith et al., 2018).

To allow quantification of whole HERV loci potentially able to produce antigen, two lists of complete HERV proviruses have been previously compiled. These loci have been used to correlate HERV expression with immunotherapy response and cytotoxic T-cell signatures. In 2011 a list of potentially transcribed HERV loci was compiled to create a uniform nomenclature of separate transcriptional units for inclusion within the standardised nomenclature of human genes defined by the Human Genome Organisation Gene Nomenclature Committee (HGNC) (Mayer et al., 2011). Inclusion on this list required the HERV locus to have an mRNA sequence published in a public database, with the complementary DNA (cDNA) sequence uniquely mappable to the human reference genome (GRCh37), and with sequence representative of viral genes (as opposed to solo LTR elements) (Mayer et al., 2011). Fusions of HERVs and HERV insertions within other HERVs would not be represented by this list of loci, nor would any HERV locus overlapping another known transcript, such as those from known genes, as they would no longer be defined as a separate transcriptional unit in this case. From these criteria, a limited list of 66 loci were identified from the RepBase (Bao et al., 2015) database of repeat loci, with all names beginning “ERV” such as *ERV3-2* (Mayer et al., 2011). A second larger list was compiled in 2016, again focused on potentially transcribed loci and excluding solo LTRs. This list included HERVs inserted within other HERVs and loci where recombination had occurred which were merged into single HERV locus annotations. The list of 3173 potentially transcribed HERV loci was expected to represent around one quarter of total HERV sequences within the human genome, with the naming convention for the list beginning “HERV” followed by a unique number, such as *HERV 4700* (Vargiu et al., 2016). These lists were used in previous studies analysing HERV expression in KIRC and the correlation with immune checkpoint blockade response, and the list of 3173 is the basis of the tool hervQuant (Smith et al., 2018). The list of 66 loci from Mayer and colleagues (2011) was used in the study of response to immune checkpoint blockade by Rooney and colleagues (2015) and in the study of correlation with cytotoxic T-cell signatures by Panda and colleagues (2018). The list of 3173 loci from Vargiu and colleagues (2016) was

used by Smith and colleagues (2018) to correlate HERV expression with immunotherapy response in KIRC and other cancers.

In order to identify correlates of immune checkpoint blockade response, data from a clinical trial of anti-PD-1 therapy in treatment-naïve metastatic KIRC patients was analysed. This data is from the ADAPTeR clinical trial (NCT02446860) consisted of 15 patients with metastatic KIRC treated with an anti-PD-1 antibody (nivolumab) (Au et al., 2021). To assign therapy responsiveness to these patients, those with partial response or stable disease of at least 6 months were selected as responders regardless of the overall change in tumour size whilst on treatment. Therefore, 5 patients were selected as responders, and 10 as non-responders. Patients underwent multi-region tumour biopsies of both primary and metastatic sites, with a total of 115 biopsies taken over four time points: at baseline, week 9 after treatment start, nephrectomy if performed, and at disease progression. In alignment with other studies, analysis of genomic features such as tumour mutational burden, expressed non-synonymous single nucleotide variants, and frameshift insertions or deletions did not correlate with response to therapy. However, T-cell receptor (TCR) sequencing did reveal immunotherapy response was associated with maintenance of pre-treatment expanded TCR clones, suggesting an ongoing antigen response boosted by anti-PD-1 therapy. Analysis of T-cells in responding patients showed upregulation of granzymes B and K required for T-cell cytotoxicity.

6.3 Results

6.3.1 Analysis of previously annotated HERV loci

In order to analyse the expression of HERV loci previously associated with response to immunotherapy and cytotoxic T-cell signatures in KIRC the loci had to be mapped to GRCh38. Mapping the previously annotated HERV loci to GRCh38 revealed errors in the older assemblies (2.2.1: Annotation of HERV loci). Of the 66 loci identified by Mayer and colleagues (2011), 47 were also identified by Vargiu and colleagues (2016). The previously annotated lists in total corresponded to 7989 unique LTR element annotations from the Dfam-derived library, with many previously annotated single loci overlapping multiple Dfam-derived elements (Figure 35a). From the Mayer list the majority of annotations (49/66) only overlapped a single locus from the Dfam-derived library, one locus had no match identified, and one locus overlapped four LTR elements (Figure 35a). The list from Vargiu and colleagues (2016) was more poorly annotated when compared to the Dfam-derived library, with only 22% (701/3173) of loci corresponding to a single element. For 12% (381/3173) of loci no match could be found in the Dfam library, though perhaps matching would have been more efficient if the buffer allowed for locus overlap had been increased. The majority of the loci annotated previously (2091/3173, 66%) matched more than one LTR element from the Dfam-derived library, with one locus matching 18 separate elements (Figure 35a). Upon further inspection of locus alignment in GRCh38 there was also overlap with other RTE types, including SINEs and LINEs, and overlap with canonical gene exons.

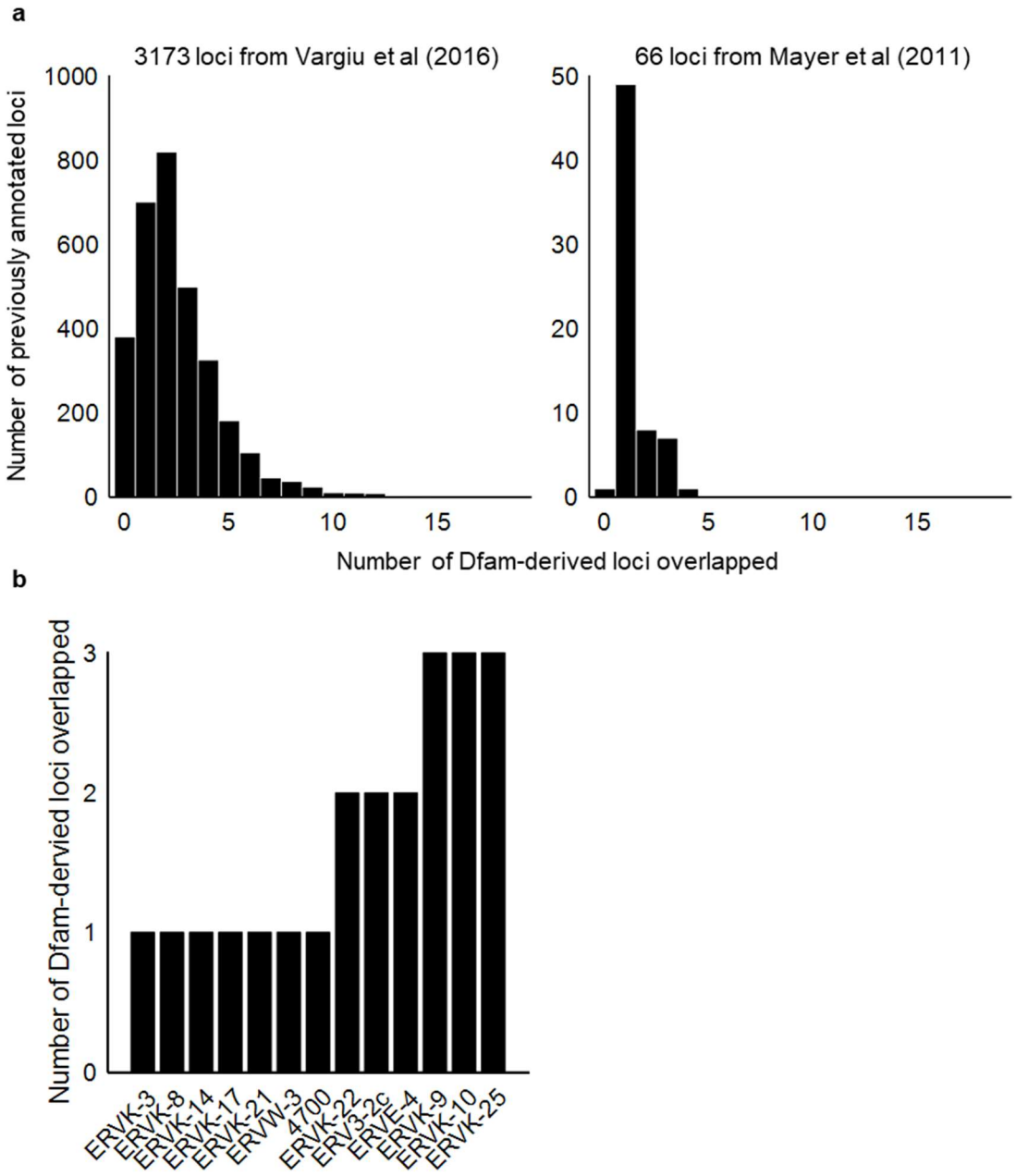


Figure 35: The number of Dfam-derived LTR-containing loci the previously annotated HERV lists overlapped. a. The number of Dfam-derived loci overlapped by the 3173 loci from Vargiu and colleagues (2016) (left) and by the 66 loci from Mayer and colleagues (2011) (right). **b.** The number of Dfam-derived loci overlapped by HERVs previously associated with response to immune checkpoint blockade or associated with cytotoxic T-cell signatures.

Expression of previously identified HERVs did not associate with response to anti-PD-1 therapy (Figure 36a). When responders and non-responders were compared (2.4.4: Differential expression analysis for HERVs in metastatic KIRC), expression of none of the HERVs previously associated with response correlated with response to anti-PD-1 therapy, or were significantly differentially expressed between responders and non-responders. This may be due to the lack of statistical power with such a small sample size. Although, several of the loci were more highly expressed, though not significantly, in non-responders opposing completely previous observations (Figure 36a).

This lack of association with response may also be due to poor assembly of previously annotated loci. In this analysis, Dfam-derived loci overlapping the previously associated loci were used (2.2.1: Annotation of HERV loci). This was because many of the loci previously associated with response were incorrectly annotated (Figure 35a). If the incorrectly annotated loci positions had been used instead it is possible a correlation with response may have been seen. Of the HERV loci previously associated with either cytotoxic T-cell signatures or immune checkpoint blockade response, seven matched a single Dfam-derived locus, though some of these were poorly annotated (Figure 35b). The *ERVK-3* locus (Mayer et al., 2011) previously correlated with cytotoxic T-cell signatures in primary KIRC (Rooney et al., 2015) overlapped one Dfam-derived locus, however the overlap was incomplete, with *ERVK-3* only overlapping a small portion of the HERV-K annotated in RepeatMasker and the Dfam-derived library (Attig et al., 2017) (Figure 37). Additionally, the locus of *ERV3-2*, previously shown to correlate with immune checkpoint blockade response in 11 cancer types including primary KIRC (Panda et al., 2018), overlapped two HERV loci incompletely (Figure 38b). The corresponding locus (*HERV 2637*) from the list published by Vargiu and colleagues (2016) not only overlapped multiple LTR-containing loci, but also a LINE element, and part of the final exon of *WDR91* (Figure 38b). It was possible these errors appeared when the coordinates were lifted from GRCh37 to GRCh38, but looking at the original GRCh37 coordinates, the *HERV 2637* locus did originally overlap non-LTR elements and the final exon

of *WDR91* (Figure 38a). It is possible that expression of genes such as *WDR91* may associate with patient response to immunotherapy, which may have driven any association seen previously.

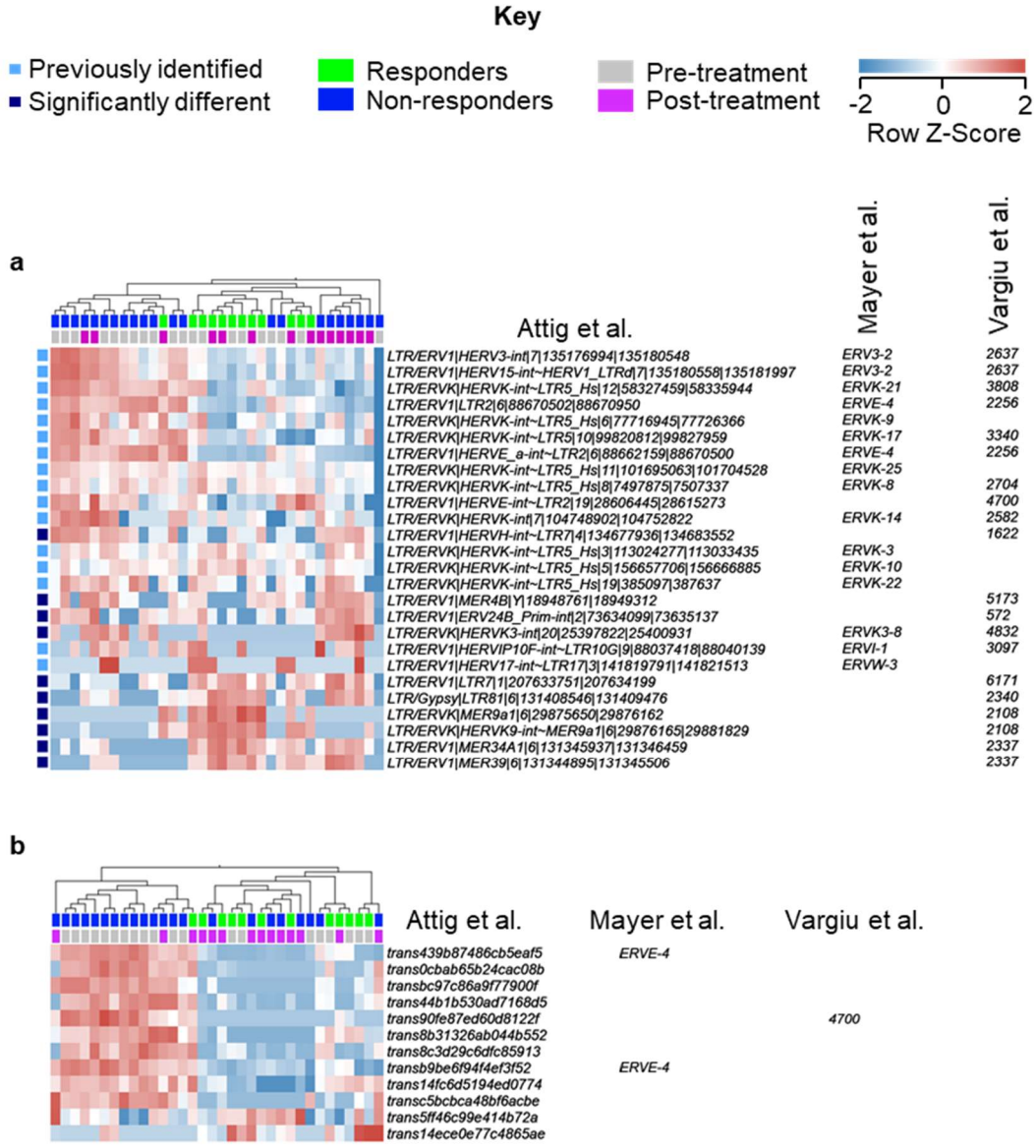


Figure 36: The relative expression of HERVs and LTR-overlapping transcripts in pre-treatment and week 9 samples from patients treated with anti-PD-1 therapy. Figure adapted from Au et al (2021). **a.** Relative expression of HERV loci significantly differentially expressed between responders and non-responders, as well as those previously associated with immune checkpoint blockade response and cytotoxic T-cell signatures. **b.** Relative expression of LTR-overlapping transcripts significantly differentially expressed between responders and non-responders, which includes transcripts overlapping the previously identified *ERVE-4* (Smith et al., 2019; Takahashi et al., 2008) and *HERV 4700* (Smith et al., 2018).

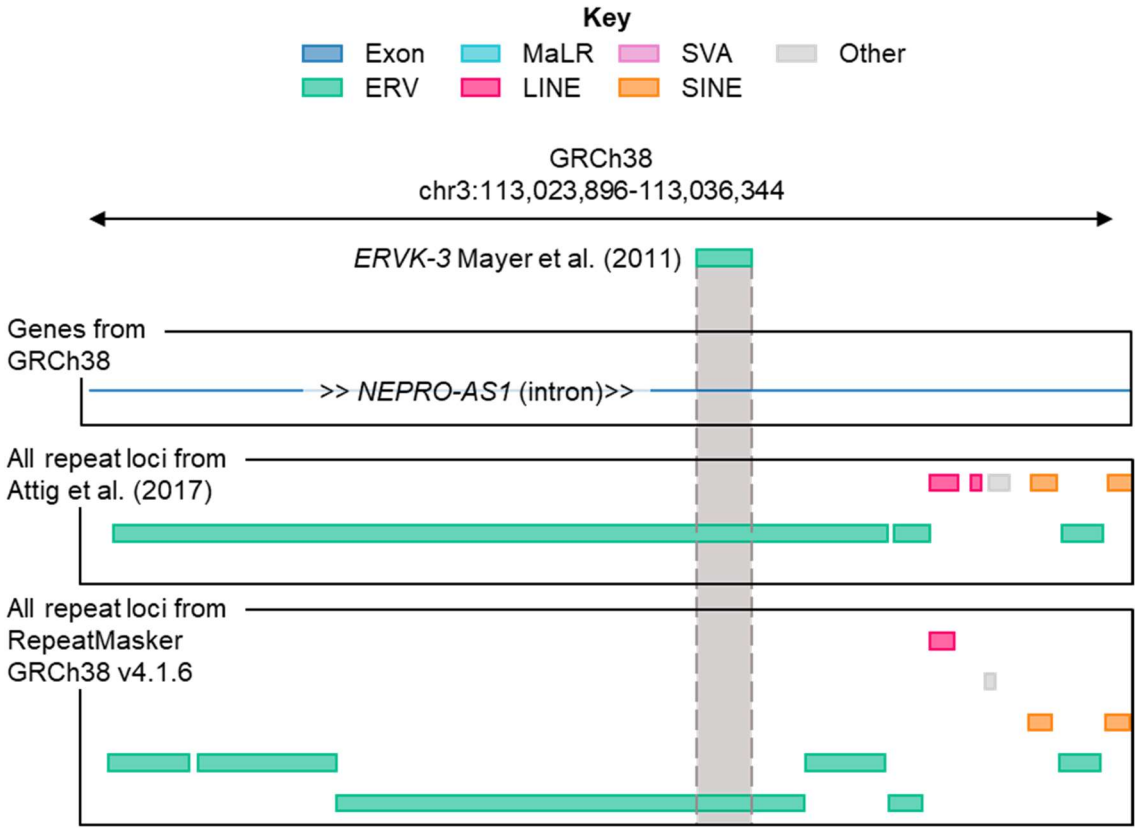


Figure 37: The structure of *ERVK-3* in GRCh38. *ERVK-3* overlaps an intron of *NEPRO-AS1*. In the Dfam derived database of repeat loci (Attig et al., 2017) the *ERVK-3* overlaps a portion of a HERV, in the most recent update of RepeatMasker (May 2024) the *ERVK-3* also overlaps a portion of a HERV. The annotation from 2016 is therefore incomplete.

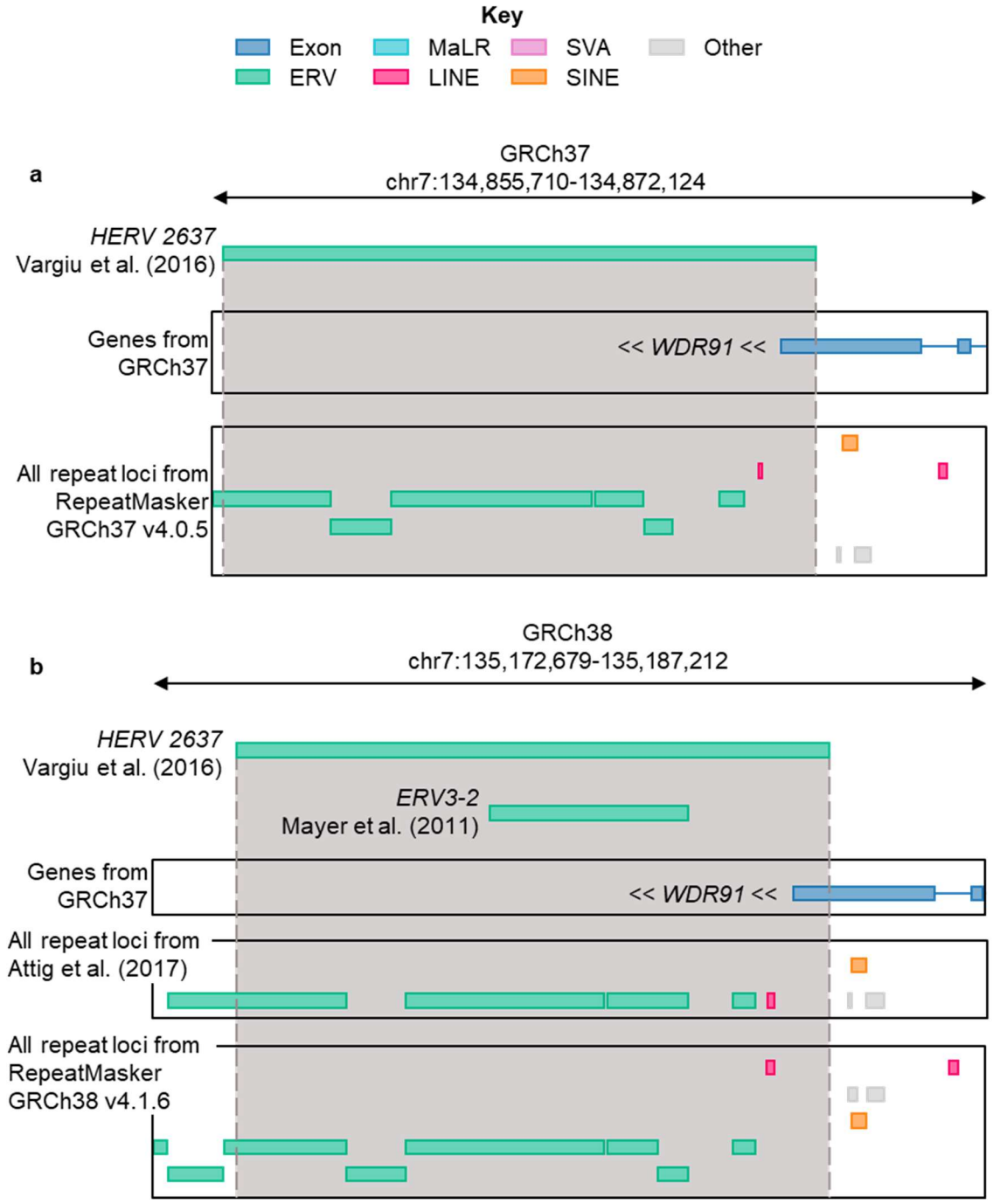


Figure 38: The structure of *ERV3-2* and the corresponding *HERV 2637*. Loci positions shown in both **a.** GRCh37 and **b.** GRCh38. *HERV 2637* overlaps part of a gene exon, as well as multiple HERV loci and a LINE locus in both the Dfam derived library of repeat loci (Attig et al., 2017) and in the RepeatMasker database. This is true for both GRCh38 and GRCh37. The corresponding *ERV3-2* overlaps multiple HERV loci in both the Dfam derived library of repeat loci (Attig et al., 2017) and in the RepeatMasker database.

Additionally, many of the previously identified HERV loci were expressed in purified immune cell samples (Figure 39, 2.1.5: Purified immune cell datasets). The *ERV3-2* locus had especially strong expression across immune cell subsets in both GSE60424 (Linsley et al., 2014) and E-MTAB-8208 (Kazachenka et al., 2019). This may explain the correlation with immune checkpoint blockade response seen across cancers (Panda et al., 2018), as expression of *ERV3-2* pre-treatment reflects immune infiltrate pre-treatment. On the other hand, the *ERVE-4* integration which has been shown to produce antigen leading to tumour regression (Takahashi et al., 2008) had lesser and more sporadic expression in immune cells (Figure 39), suggesting it is more specific to KIRC tumour cells.

6.3.2 Analysis of HERV loci annotated in the Dfam-derived library

Expression of other HERV loci identified in the Dfam-derived library (Attig et al., 2017) are associated with response to anti-PD-1 therapy (Figure 36a). There were 10 HERVs derived from 8 distinct loci which distinguished responders from non-responders in this cohort (2.4.4: Differential expression analysis for HERVs in metastatic KIRC). The HERVs were significantly differentially expressed (absolute fold change ≥2, $q \leq 0.05$) and were mainly expressed in responders both pre- and post- treatment, and non-responders post-treatment (Figure 36a). Some of these loci also had expression in immune cell subsets (Figure 39), and their increased expression in responders may be due to increased immune infiltrate in these tumours at baseline (Au et al., 2021). Increased expression of these HERV loci in post-treatment non-responder samples may indicate increased immune infiltrate in these tumours due to the anti-PD-1 therapy.

6.3.3 Analysis of HERV-overlapping transcripts assembled in the *de novo* transcriptome

Expression of HERV-overlapping transcripts from the *de novo* transcriptome assembly was associated with response to anti-PD-1 therapy (Figure 36b, 2.2.2.1: Expression of transcripts assembled in the de novo transcriptome, 2.4.4: Differential expression analysis for HERVs in metastatic KIRC). Previously, 570

LTR-overlapping transcripts from the *de novo* transcriptome assembly have been identified to be upregulated in a cancer-specific way in primary KIRC (Attig et al., 2019). In differential expression analysis, 12 transcripts derived from 9 HERV loci were significantly differentially expressed (absolute fold change ≥2, q ≤ 0.05) (Figure 36b). These transcripts overlapped HERV loci also overlapping *ERVE-4* and *HERV 4700* which have previously been identified to correlate with immunotherapy response (Panda et al., 2018; Smith et al., 2019; Takahashi et al., 2008). However, here expression was greatest in non-responders pre-treatment, whereas previously increased expression was seen in patients responding to immune checkpoint blockade. This increase in non-responders may be due to the transcripts identified correlating with tumour purity (Figure 40), which would be expected as the transcripts were selected as KIRC-specific. Increased transcript expression indicated increased tumour purity and therefore lack of immune infiltrate pre-treatment.

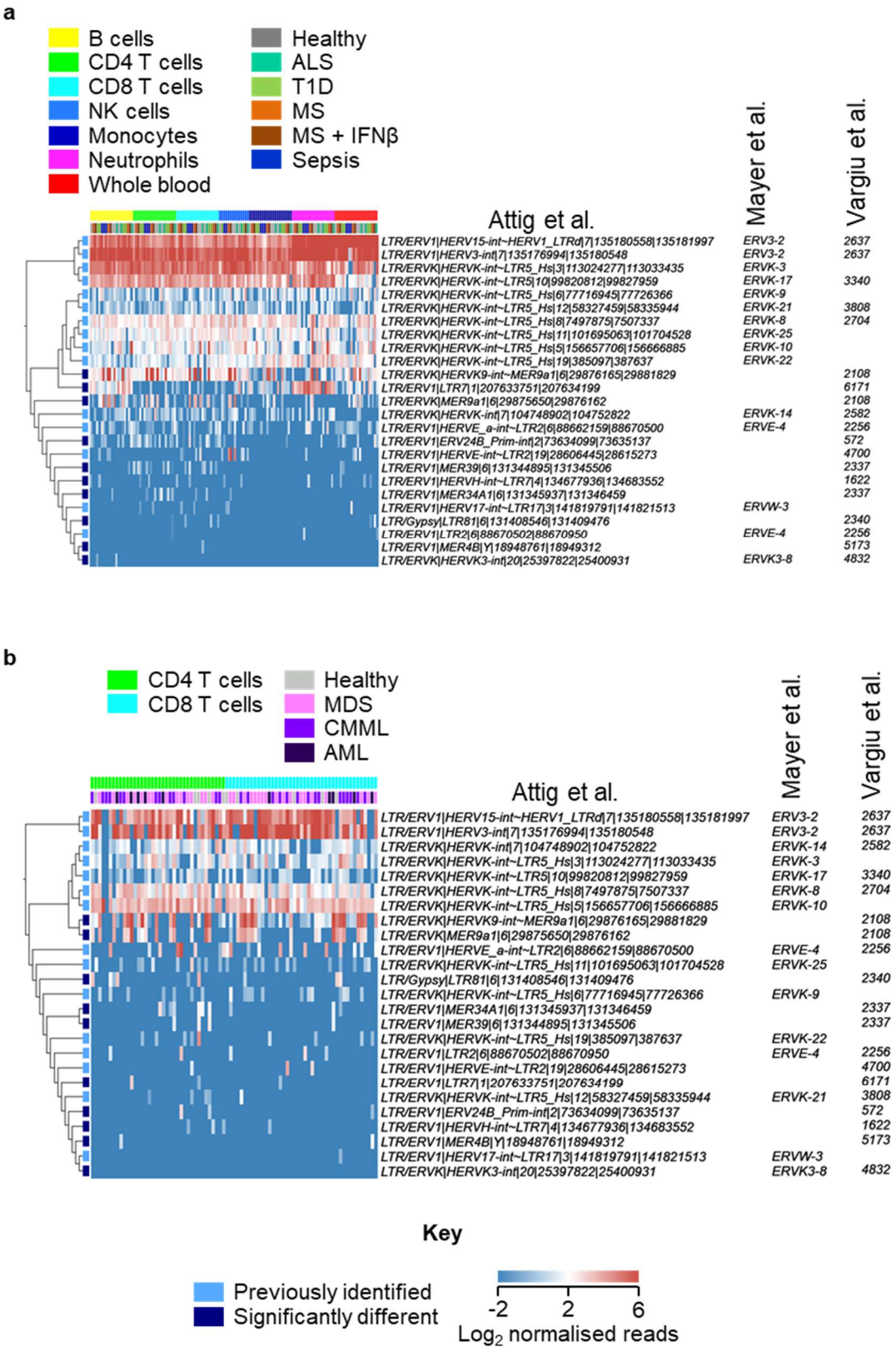

a
B cells
CD4 T cells
CD8 T cells
NK cells
Monocytes
Neutrophils
Whole blood
Healthy
ALS
T1D
MS
MS + IFNβ
Sepsis
Attig et al.
Mayer et al.
Vargiu et al.
LTR/ERV1|HERV15-int~HERV1_LTRd|7|135180558|135181997 ERV3-2 2637
LTR/ERV1|HERV3-int|7|135176994|135180548 ERV3-2 2637
LTR/ERVK|HERVK-int~LTR5_Hs|3|113024277|113033435 ERVK-3
LTR/ERVK|HERVK-int~LTR5|10|99820812|99827959 ERVK-17 3340
LTR/ERVK|HERVK-int~LTR5_Hs|6|77716945|77726366 ERVK-9
LTR/ERVK|HERVK-int~LTR5_Hs|12|58327459|58335944 ERVK-21 3808
LTR/ERVK|HERVK-int~LTR5_Hs|8|7497875|7507337 ERVK-8 2704
LTR/ERVK|HERVK-int~LTR5_Hs|11|101695063|101704528 ERVK-25
LTR/ERVK|HERVK-int~LTR5_Hs|5|156657706|156666885 ERVK-10
LTR/ERVK|HERVK-int~LTR5_Hs|19|385097|387637 ERVK-22
LTR/ERVK|HERVK9-int~MER9a1|6|29876165|29881829 2108
LTR/ERV1|LTR7|1|207633751|207634199 6171
LTR/ERVK|MER9a1|6|29875650|29876162 2108
LTR/ERVK|HERVK-int|7|104748902|104752822 ERVK-14 2582
LTR/ERV1|HERVE_a-int~LTR2|6|88662159|88670500 ERVE-4 2256
LTR/ERV1|ERV24B_Prim-int|2|73634099|73635137 572
LTR/ERV1|HERVE-int~LTR2|19|28606445|28615273 4700
LTR/ERV1|MER39|6|131344895|131345506 2337
LTR/ERV1|HERVH-int~LTR7|4|134677936|134683552 1622
LTR/ERV1|MER34A1|6|131345937|131346459 2337
LTR/ERV1|HERV17-int~LTR17|3|141819791|141821513 ERVW-3
LTR/Gypsy|LTR81|6|131408546|131409476 2340
LTR/ERV1|LTR2|6|88670502|88670950 ERVE-4 2256
LTR/ERV1|MER4B|Y|18948761|18949312 5173
LTR/ERVK|HERVK3-int|20|25397822|25400931 ERVK3-8 4832
b
CD4 T cells
CD8 T cells
Healthy
MDS
CMML
AML
Attig et al.
Mayer et al.
Vargiu et al.
LTR/ERV1|HERV15-int~HERV1_LTRd|7|135180558|135181997 ERV3-2 2637
LTR/ERV1|HERV3-int|7|135176994|135180548 ERV3-2 2637
LTR/ERVK|HERVK-int|7|104748902|104752822 ERVK-14 2582
LTR/ERVK|HERVK-int~LTR5_Hs|3|113024277|113033435 ERVK-3
LTR/ERVK|HERVK-int~LTR5|10|99820812|99827959 ERVK-17 3340
LTR/ERVK|HERVK-int~LTR5_Hs|8|7497875|7507337 ERVK-8 2704
LTR/ERVK|HERVK-int~LTR5_Hs|5|156657706|156666885 ERVK-10
LTR/ERVK|HERVK9-int~MER9a1|6|29876165|29881829 2108
LTR/ERVK|MER9a1|6|29875650|29876162 2108
LTR/ERV1|HERVE_a-int~LTR2|6|88662159|88670500 ERVE-4 2256
LTR/ERVK|HERVK-int~LTR5_Hs|11|101695063|101704528 ERVK-25
LTR/Gypsy|LTR81|6|131408546|131409476 2340
LTR/ERVK|HERVK-int~LTR5_Hs|6|77716945|77726366 ERVK-9
LTR/ERV1|MER34A1|6|131345937|131346459 2337
LTR/ERV1|MER39|6|131344895|131345506 2337
LTR/ERVK|HERVK-int~LTR5_Hs|19|385097|387637 ERVK-22
LTR/ERV1|LTR2|6|88670502|88670950 ERVE-4 2256
LTR/ERV1|HERVE-int~LTR2|19|28606445|28615273 4700
LTR/ERV1|LTR7|1|207633751|207634199 6171
LTR/ERVK|HERVK-int~LTR5_Hs|12|58327459|58335944 ERVK-21 3808
LTR/ERV1|ERV24B_Prim-int|2|73634099|73635137 572
LTR/ERV1|HERVH-int~LTR7|4|134677936|134683552 1622
LTR/ERV1|MER4B|Y|18948761|18949312 5173
LTR/ERV1|HERV17-int~LTR17|3|141819791|141821513 ERVW-3
LTR/ERVK|HERVK3-int|20|25397822|25400931 ERVK3-8 4832
Key
Previously identified
Significantly different
-2
2
6
Log2 normalised reads

Figure 39: The log_2 normalised expression of HERVs previously associated with response to immune checkpoint blockade, cytotoxic T-cell signatures, and correlated with response to anti-PD-1 therapy in this cohort in purified immune cell subsets. Figure adapted from Au et al (2021). **a.** GSE60424 (Linsley et al., 2014) and **b.** E-MTAB-8208 (Kazachenka et al., 2019).

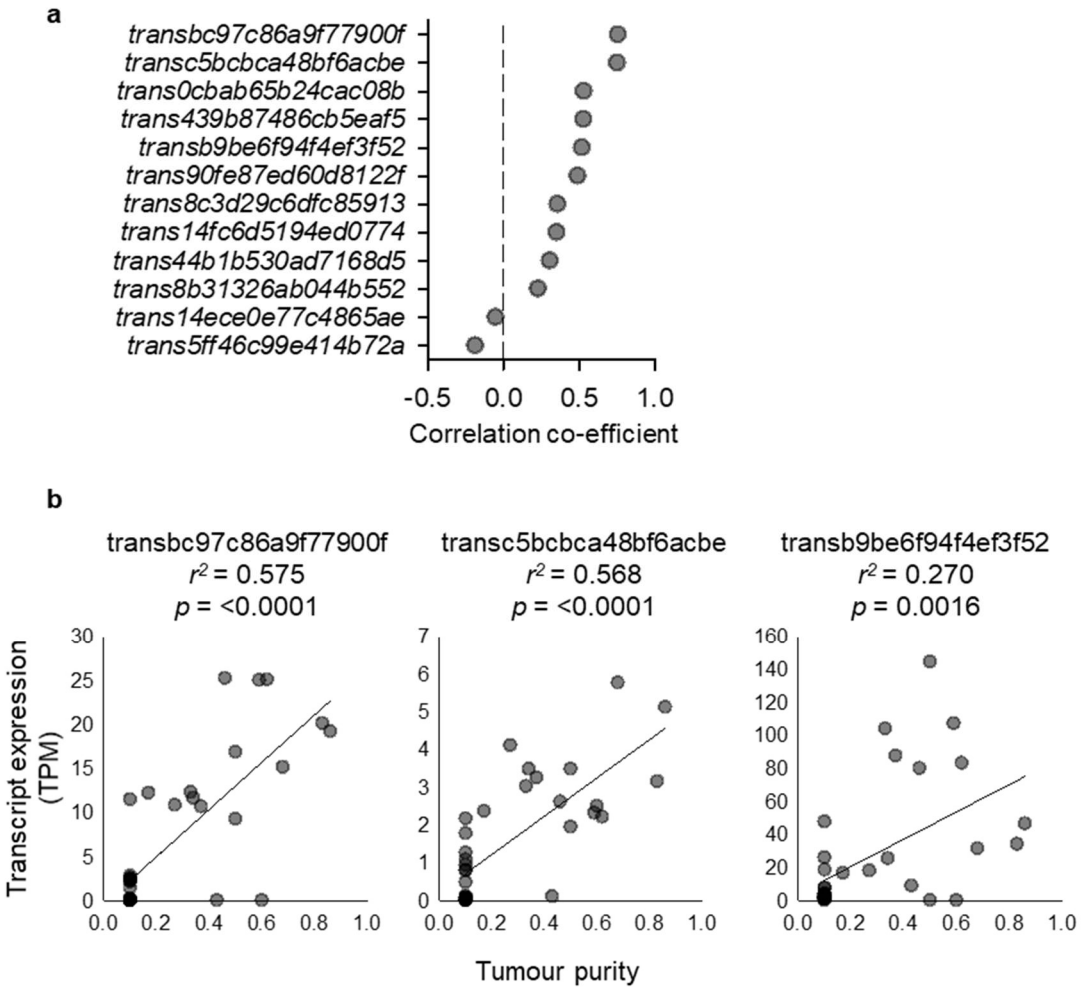


Figure 40: The correlation of LTR-overlapping transcripts associated with response to anti-PD-1 therapy with tumour purity. Figure adapted from Au et al (2021). **a.** The correlation of all transcripts associated with response with tumour purity. **b.** Examples of specific transcripts and the relationship between expression and tumour purity.

6.4 Discussion

In this patient cohort, HERV expression was associated with tumour purity and type of immune infiltrate. Using updated HERV loci annotation from the Dfam-derived library (Attig et al., 2017) and the *de novo* transcriptome assembly (Attig et al., 2019), HERV expression correlated with tumour purity in this dataset of 60 tumour biopsies from 14 anti-PD-1 treated metastatic KIRC patients. Few HERV loci were significantly differentially expressed between responders and non-responders, possibly due to the small sample size, with increased expression in responders pre- and post- treatment and non-responders post-treatment. Several of these loci had expression in purified immune cell subsets, potentially explaining their upregulation in responding tumours as those biopsies had increased tumour infiltrate (Au et al., 2021). Additionally, few HERV-overlapping transcripts were significantly differentially expressed, with increased expression here in non-responders post-treatment. This opposes the expression pattern seen for the HERV loci but is explained by the transcripts analysed being selected for as KIRC-specific. Most transcripts that associated with response also positively correlated with tumour purity. Thus, as the immune-expressed HERV loci increase in expression, the tumour-specific HERV-overlapping transcripts decrease in expression. Here HERV expression reflects tumour purity and immune infiltrate levels, although provision of antigen is not ruled out, it is not well supported in this dataset. The only previously associated HERV loci with overlapping transcripts expressed at significantly different levels in responders and non-responders were *ERVE-4* and *HERV 4700*, both of which have been suggested to provide antigen (Smith et al., 2018; Takahashi et al., 2008). However, although neither are expressed in purified immune cell subsets, both are unexpectedly upregulated in non-responders pre-treatment. Previous work has shown that peptides derived from the *ERVE-4* locus are human leukocyte antigen (HLA)-A*02 and A*11 restricted (Smith et al., 2018), thus correlation with response to immune checkpoint blockade is likely to only be seen in cohorts of patients with these HLA haplotypes.

Previously annotated HERV loci did not associate with response to anti-PD-1 therapy in this cohort, which may be due to poor annotation of the loci coordinates. None of the individual HERV loci previously identified to correlate with immune checkpoint blockade response or cytotoxic T-cell signatures in KIRC were found to correlate with response to anti-PD-1 therapy in this cohort. Furthermore, these HERV loci were non-significantly upregulated in non-responders pre-treatment completely opposing previous results. Many of these previously identified loci, including *ERV3-2* associated with immune checkpoint blockade response in 11 cancer types (Panda et al., 2018), were expressed in purified immune cell subsets, potentially explaining the association previously seen. Added to this, the annotations of the previously identified potentially-translated HERV-loci (Mayer et al., 2011; Vargiu et al., 2016) were very poor when compared to the more recent Dfam-derived loci (Attig et al., 2017). Many of the loci overlapped more than one HERV locus as annotated in the Dfam-derived library, with one locus overlapping 18 separate loci. Some annotated loci also incompletely overlapped HERVs. These issues with previous annotations would have led to different expression levels being associated with the HERV compared to using the HERV coordinates from the Dfam-derived library, explaining why here using the more recent library the same associations are not seen.

HERVs are likely to be upregulated in this cohort as passengers of uncontrolled HIF1α and HIF2α activity. *Polybromo 1* (*PBRM1*) and *VHL* mutation were common in this cohort, with 62% and 77% of patients presenting respectively with these clonal and sub-clonal alterations, which is typical of KIRC samples (Au et al., 2021). VHL works to continuously ubiquitinate HIF1α and HIF2α (hypoxia inducible factors, HIFs) under normoxic conditions. Under hypoxia, or when VHL is mutated, HIF1α and HIF2α are no longer continuously degraded and are able to translocate to the nucleus and act as TFs. It has been shown HIFs are able to bind LTR elements, leading to transcription of HERV loci as well as controlling transcription of other genes (Siebenthall et al., 2019). Furthermore, mutation of the chromatin remodelling protein PBRM1 has been associated with increased HERV expression also in a HIF-dependent manner. If antigen was to be derived

from HERV loci or HERV-overlapping transcripts, the amount of time it would take for the tumour to evolve away from expressing any given HERV locus is unknown, given the expression is due to such an uncontrolled upregulation of HIF activity. Added to this, due to the repetitive nature of HERV elements, it is possible that the same antigen may be derived from multiple loci, further increasing the promise of HERVs as a source of therapeutic targets.

6.5 Conclusion

Overall, in this cohort expression of HERV loci and HERV-overlapping transcripts reflected immune infiltrate and tumour purity. Here significantly differentially expressed HERV loci were upregulated in responders and were expressed across immune cell subsets. Whereas LTR-overlapping KIRC-specific transcripts significantly differentially expressed between responders and non-responders were upregulated in non-responders pre-treatment. These associations indirectly correlated, through levels of immune infiltrate and tumour purity, the HERV expression with response to anti-PD-1 therapy in 14 patients with metastatic KIRC. Further associations and potential antigen sources may be revealed using larger and HLA haplotyped patient cohorts. This analysis does not agree with what has been previously published, where increased HERV expression was seen in tumour samples of patients responding to immune checkpoint blockade and in tumours with increased cytotoxic T-cell signatures (Panda et al., 2018; Rooney et al., 2015; Smith et al., 2018). The reason for differing conclusions is likely to be in the annotated loci used in this analysis, where a more recent list of loci derived from the Dfam database (Attig et al., 2017) were used instead of two previously compiled lists (Mayer et al., 2011; Vargiu et al., 2016). These previously complied HERV loci were poorly annotated when compared to the Dfam-derived library, with fragmented and incomplete elements, as well as elements merging multiple HERV loci alongside non-LTR elements such as SINEs and LINEs, and known gene exons.

Chapter 7. Results 5: Exploration of transcripts upregulated in KIRC

7.1 Aims

To further understand the complex association between hypoxia and patient survival, transcripts upregulated under hypoxia were analysed. As KIRC samples undergo the permanent activation of HIFs this cancer type was used as a model of continuous hypoxia. Thus, transcripts identified by the *de novo* transcriptome assembly upregulated in KIRC were selected for further exploration (Figure 41). Analysis of a transcriptome inclusive of RTE sequences is especially interesting as they contain elements HIFs may bind only under the hypomethylated state of the cancer genome (D'Anna et al., 2020; Siebenthall et al., 2019).

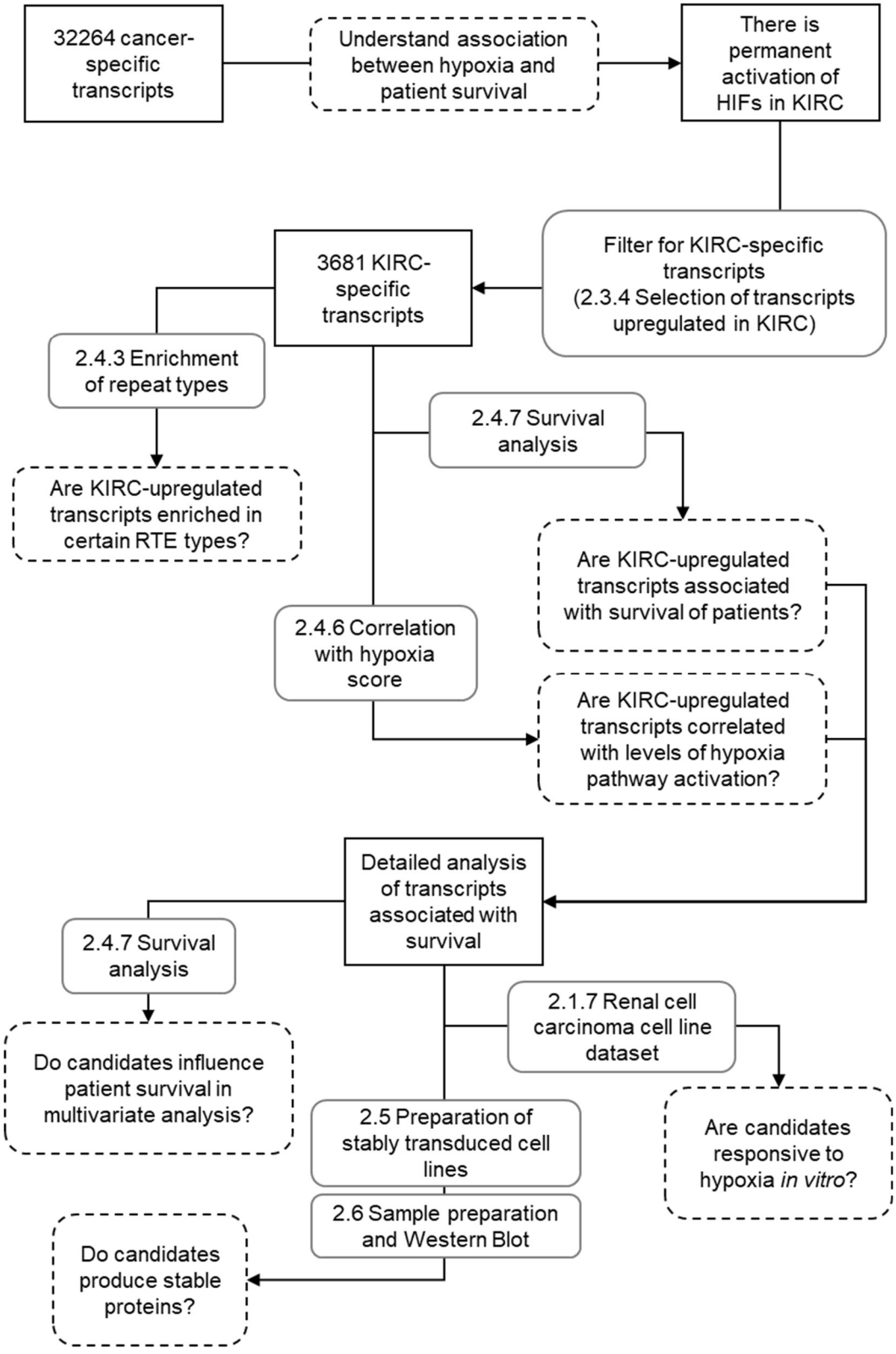


Figure 41: Aims for Results 5: Exploration transcripts upregulated in KIRC. Aims are shown in dashed boxes and methods are referenced in grey boxes. (HIFs: hypoxia inducible factors; KIRC: kidney renal clear cell carcinoma)

7.2 Introduction

The majority of KIRC tumours undergo permanent activation of the hypoxia response pathway, regardless of the presence of a hypoxic environment. In healthy cells, a ubiquitination complex including VHL constitutively ubiquitinates HIF1α and HIF2α, leading to proteasome-dependent degradation of these TFs. Under hypoxia, VHL is no longer able to bind HIF proteins efficiently (Maxwell et al., 1999), allowing HIF1α and HIF2α to form heterodimers with HIF1β (also known as the aryl hydrocarbon receptor nuclear translocator (ARNT)) and bind hypoxia response elements in the genome, inducing hypoxia-responsive genes. However, in KIRC the function of both alleles of the *VHL* gene is lost in 50-80% of tumours through mutation, deletion, or hypermethylation (Foster et al., 1994; Gnarra et al., 1994; Herman et al., 1994), leading to activation of HIF1α and HIF2α targets regardless of the oxygen concentration within the cell.

Hypoxia influences the expression of transcripts by altering gene expression and splicing patterns, as well as inhibiting nonsense-mediated decay. Gene expression under hypoxia is regulated in a HIF-dependent manner, with HIFs directly altering expression of genes involved in angiogenesis, apoptosis, cell motility, cell proliferation, and metabolism (Samanta et al., 2017). Genes upregulated include *stanniocalcin 2* (*STC2*) which promotes proliferation under hypoxic conditions (Law and Wong, 2010), *vascular endothelial growth factor A* (*VEGFA*) which regulates angiogenesis (Siebenthall et al., 2019) and leads to the characteristic vascularisation signatures seen in KIRC (Ricketts et al., 2018), and *carbonic anhydrase 9* (*CA9*) which is involved in ion transport (Sena et al., 2014a; Siebenthall et al., 2019). Additionally, HIFs control expression of other TFs such as *OCT4* involved in maintenance of pluripotency (Nichols et al., 1998) and *basic helix-loop-helix family member e41* (*BHLHE41*) involved in regulation of invasive tumour phenotypes (Montagner et al., 2012).

Furthermore, hypoxia influences splicing patterns leading to preferential expression of specific isoforms. Hypoxia has been shown to regulate the splicing patterns of both HIF and non-HIF targets (Sena et al., 2014a), influencing which

functional isoform is produced (Hirschfeld et al., 2009; Sena et al., 2014a; Sena et al., 2014b) or promoting non-coding isoform production (Memon et al., 2016). The *cellular communication network factor 1* (*CYR61*) gene locus can produce an isoform with intron three retention giving rise to a truncated CYR61 protein, or an isoform with intron three splicing giving rise to the active proangiogenic protein. Under hypoxia in breast tumours, production of the active form was favoured (Memon et al., 2016). *Adrenomedullin* produces two isoforms, one including intron three giving rise to the proadrenomedullin N-terminal 20 peptide (PAMP), and one splicing intron three giving rise to both PAMP and adrenomedullin (AM) proangiogenic peptides. Under hypoxia, in both cancer and normal cell lines, the production of the PAMP/AM isoform is favoured by 62:1 compared to under normoxia where it is favoured at 12:1 (Sena et al., 2014b). In TCGA colon adenocarcinoma samples, DNA damage repair genes, and RNA splicing genes switch to the production of non-coding isoforms in a hypoxia-dependent manner (Memon et al., 2016).

Hypoxia has additionally been shown to inhibit nonsense-mediate decay, further regulating expression of specific isoforms. Under normoxia UPF1 RNA helicase and ATPase (UPF1) binds transcripts targeted for nonsense-mediated decay, and localises to processing bodies where the transcript is degraded. However, under hypoxic conditions, UPF1 aberrantly localises to stress granules, prohibiting degradation of the transcript and potentially allowing translation of open reading frames. The activating transcription factor 4 (ATF-4), a protein involved in protection from cellular stress, contains two upstream ORFs before the open reading frame for ATF-4 and the first exon junction. Thus, under normoxia, the *ATF-4* transcript is degraded via nonsense-mediated decay. However, under hypoxia the transcript is stabilised in stress granules and the ATF-4 protein is produced (Gardner, 2008).

Quantifying levels of hypoxia is complex due to the constitutive transcription, translation, and degradation of HIFs. As the correlation coefficient of HIF1α mRNA and protein is estimated to be 0.02 (Shenoy, 2020), either protein

concentration or target gene expression must be quantified. Here, a pan-cancer hypoxia score calculated by Lombardi and colleagues has been used to quantify HIF activation (Lombardi et al., 2022). This score combines expression of 48 HIF target genes regulated by hypoxia in six cell lines, including the renal cancer cell line RCC4 with VHL stably transfected, as measured by ChIPseq and RNAseq data. All 48 genes were bound by HIF1α and HIF2α increasing the robustness of the signature. The expression of each gene was measured across TCGA cancer subtypes, expression of the genes was quantile normalised to ensure highly expressed genes did not skew the score, and total normalised expression of the 48 genes in each sample was calculated. As expected the hypoxia scores of KIRC samples were significantly higher than those of adjacent healthy kidney samples, and KIRC samples had some of the highest hypoxia scores across all cancer types (Lombardi et al., 2022).

The impact of this permanent activation of the hypoxia response pathway on patient survival is debated in the literature. HIF1α protein staining has been associated with better patient survival (Lidgren et al., 2005), and ChIPseq data has shown HIF1α-bound genes are generally associated with better patient survival (Salama et al., 2015). HIF1α has also been shown to slow tumour growth when re-expressed in cell lines which have lost the functional version of the gene (Biswas et al., 2010) and stabilisation of HIF1α in the 786-O KIRC cell line also stably transfected with VHL slowed tumour growth *in vivo* (Maranchie et al., 2002). Although nuclear HIF1α protein staining has been associated with worse patient survival (Fan et al., 2015). On the other hand, HIF2α has been shown to increase xenograft growth when over-expressed (Biswas et al., 2010). Additionally, cytoplasmic staining of HIF2α has been associated with worse patient survival (Fan et al., 2015), alongside ChIPseq data which has shown HIF2α-bound genes are generally associated with worse patient survival (Salama et al., 2015). It should be noted, any association with patient survival may also be impacted by other variables such as low immune infiltrate which when paired with upregulation of the hypoxia response pathways was associated with poor patient survival (Bai et al., 2022).

7.3 Results

7.3.1 Transcripts upregulated in KIRC

In order to identify transcripts upregulated under the activation of the hypoxia pathway, those upregulated in KIRC were identified (2.3.4: Selection of transcripts upregulated in KIRC). Transcripts upregulated in KIRC represent both gene and RTE sequences. Of the 3681 KIRC-specific transcripts selected, 1065 transcripts overlapped known genes and no RTEs, 1153 only overlapped RTEs, and 1280 overlapped both known genes and RTEs, with 183 transcripts overlapping neither (Figure 42a). The most frequently overlapped RTE type were SINEs, but both LINE and LTR elements were also represented (Figure 42a). When compared to the whole *de novo* transcriptome assembly (2.4.3: Enrichment of repeat types) there was enrichment of LTR elements, as well as a LINE1 group and several Alu groups (Figure 42b). The selected transcripts were mainly multiexonic (Figure 42c). The transcripts represented 1742 genes with most genes only overlapped by one transcript (Figure 42d).

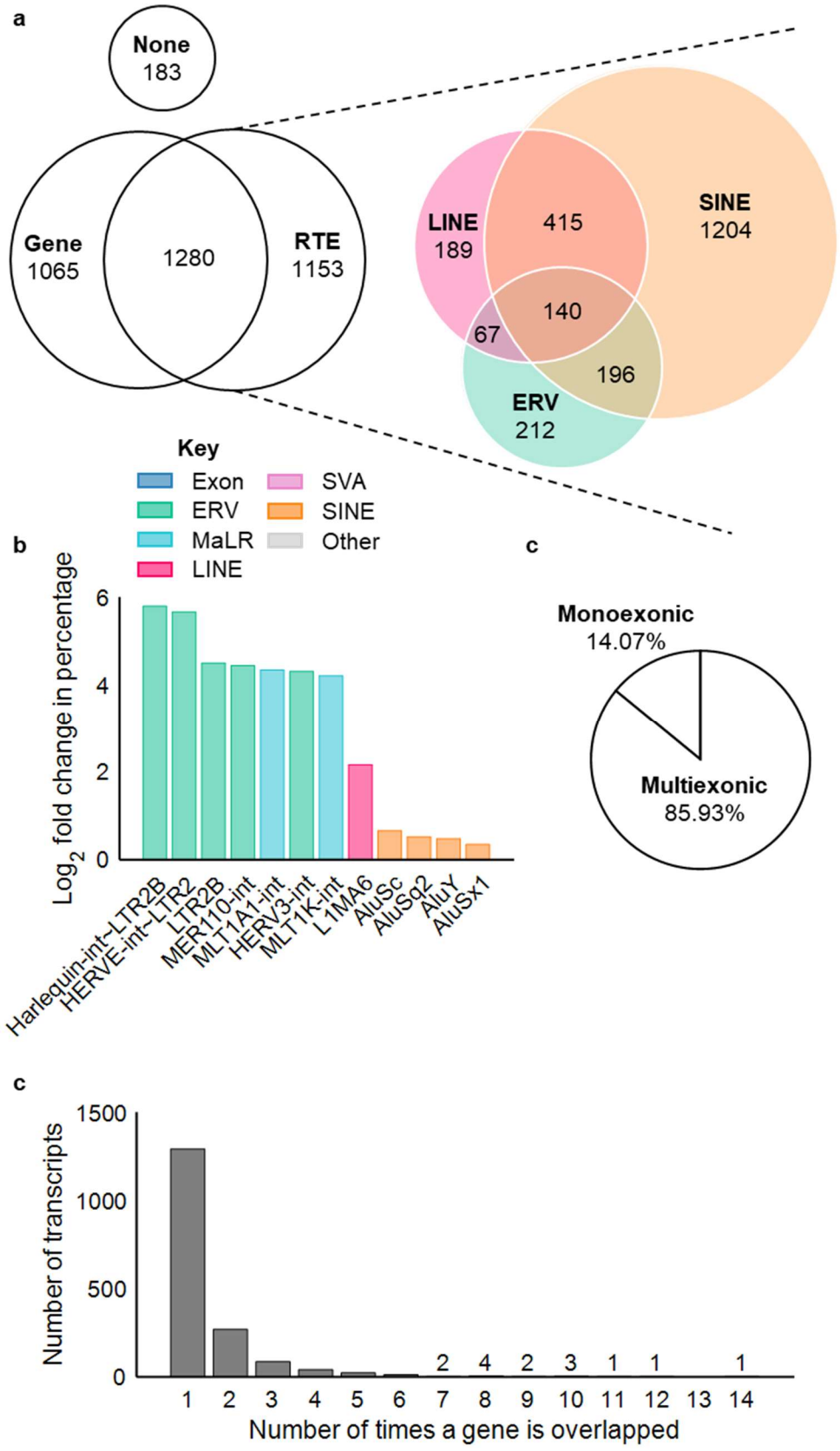

a
None
183
Gene
1065
1280
RTE
1153
LINE
189
415
SINE
1204
140
67
196
ERV
212
Key
Exon
ERV
MaLR
LINE
SVA
SINE
Other
b
Log2 fold change in percentage
6
4
2
0
Harlequin-int~LTR2B
HERVE-int~LTR2
LTR2B
MER110-int
MLT1A1-int
HERV3-int
MLT1K-int
L1MA6
AluSc
AluSq2
AluY
AluSx1
c
Monoexonic
14.07%
Multiexonic
85.93%
c
Number of transcripts
1500
1000
500
0
2 4 2 3 1 1 1
1 2 3 4 5 6 7 8 9 10 11 12 13 14
Number of times a gene is overlapped

Figure 42: Overview of the sequences represented by the 3681 KIRC-specific transcripts. a. The total number of cancer-specific transcripts overlapping gene and RTE sequences, alongside the numbers of transcripts overlapping the three most well-represented RTE groups. **b.** RTE subtypes with significant enrichment in the 3681 transcripts when compared to the whole *de novo* transcriptome assembly. **c.** The proportion of the 3681 transcripts that are either monoexonic or multiexonic. **d.** The number of times each of the unique 1742 genes represented are overlapped by a transcript.

7.3.2 Association of transcripts with hypoxia

Considering the typical lack of VHL in KIRC and therefore continuous activation of the hypoxia response pathway, the correlation with the mean hypoxia score for each sample (Lombardi et al., 2022) was calculated for each transcript (2.4.6: Correlation with the hypoxia score). Of the 3681, 1516 transcripts significantly correlated with the mean hypoxia score in TCGA KIRC samples (Pearson's correlation coefficient, $p \leq 0.05$), though correlations for most transcripts were not strong with only 240 transcripts correlating with $|r| \geq 0.2$ (Figure 43a).

7.3.3 Association of transcripts with survival

Given that these upregulated transcripts may confer a survival advantage to the tumour cells, univariate survival analysis was performed (2.4.7: Survival analysis). Of the 3681 transcripts 3664 had sufficient expression variation for the Cox proportional hazards regression model to be fitted. Of these 3664 transcripts, 2413 significantly associated with survival (HR ≤ 0.667 or ≥ 1.5, $p \leq 0.05$, Figure 43b), with more stringent filtering 934/3664 significantly associated with survival (Cox univariate test, HR ≤ 0.5 or ≥ 2, $p \leq 0.05$, Figure 43b). The association with survival of some of these transcripts is explored here.

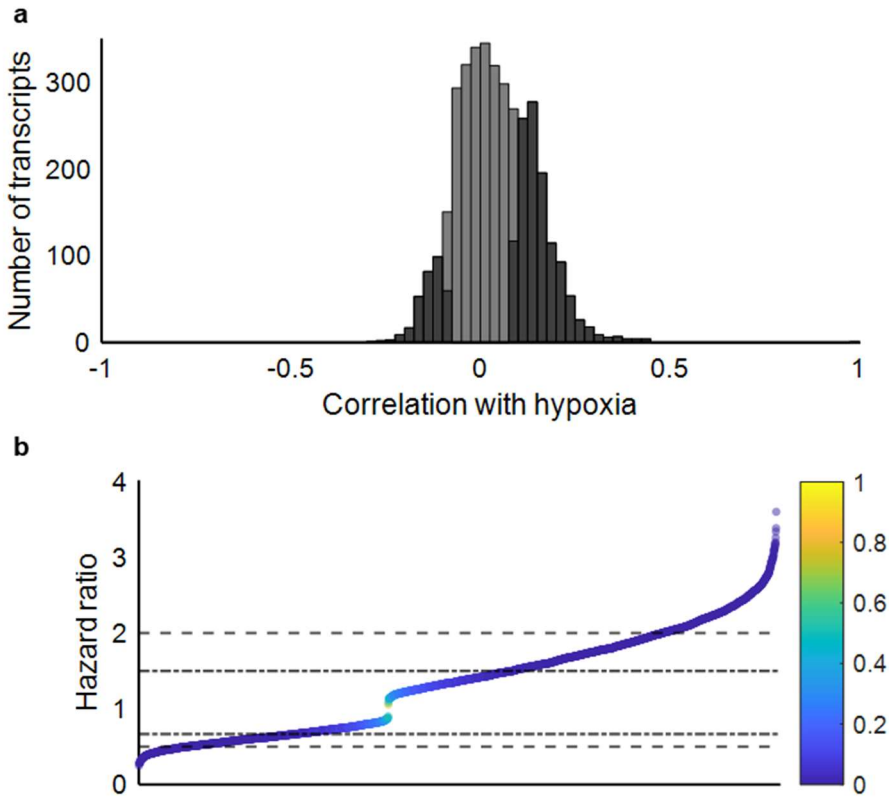


Figure 43: The association of the 3681 KIRC-specific transcripts with hypoxia and survival. **a.** The Pearson's correlation coefficient of the 3681 transcripts with the mean hypoxia score (Lombardi et al., 2022) for that patient. Bars in darker grey show the number of transcripts with significant correlation ($p \leq 0.05$). **b.** The hazard ratio derived from the univariate Cox proportional hazards regression model fitted for each of the 3664/3681 transcripts. Each point is coloured by the *p-value*, and the two pairs of hazard ratio thresholds are shown as dashed lines. Here a value greater than one implies increased expression of the transcript is associated with poorer patient survival.

7.3.3.1 A CCL28 isoform associated with better patient survival

The canonical *C-C motif chemokine ligand 28* (*CCL28*) isoform is induced by hypoxia in a HIF1α-dependent manner and recruits immunosuppressive regulatory T-cells (Tregs) to tumours. The canonical CCL28 protein, as detected by enzyme-linked immunosorbent assays (ELISA), has been shown to increase in the supernatant of ovarian (Facciabene et al., 2011), liver (Ren et al., 2016), and lung cancer lines (Huang et al., 2016; Liu and Wei, 2021) under hypoxia, with this increase ablated upon knock down of HIF1α. Increased CCL28 expression in these cancers increased recruitment of Tregs to tumours (Facciabene et al., 2011; Liu and Wei, 2021; Ren et al., 2016). However, contrary to this immunosuppressive role of canonical CCL28, the transcript detected as upregulated in KIRC (Figure 44a) had expression associated with better patient survival (Figure 45a). This transcript overlapped an isoform of *CCL28* already annotated (*NM_001301875.2*) predicted to code for an 80 AA protein which is 47 AA shorter than the canonical protein. The short isoform utilises a separate first exon, but shares the same second and third exons with the canonical transcript (Figure 44a). From inspection of BAM files (Figure 44a) and analysis of expression data (Figure 44b) it became clear the short isoform is dominant in kidney tissue, with little expression of the canonical isoform in either adjacent healthy kidney or KIRC samples. While 66.7% of adjacent healthy kidney samples (48/72) and 82.9% of KIRC samples (446/538) have short isoform expression of at least 1 TPM, only 1/72 adjacent healthy kidney and 5/538 KIRC patient samples have expression of the canonical isoform above the same threshold.

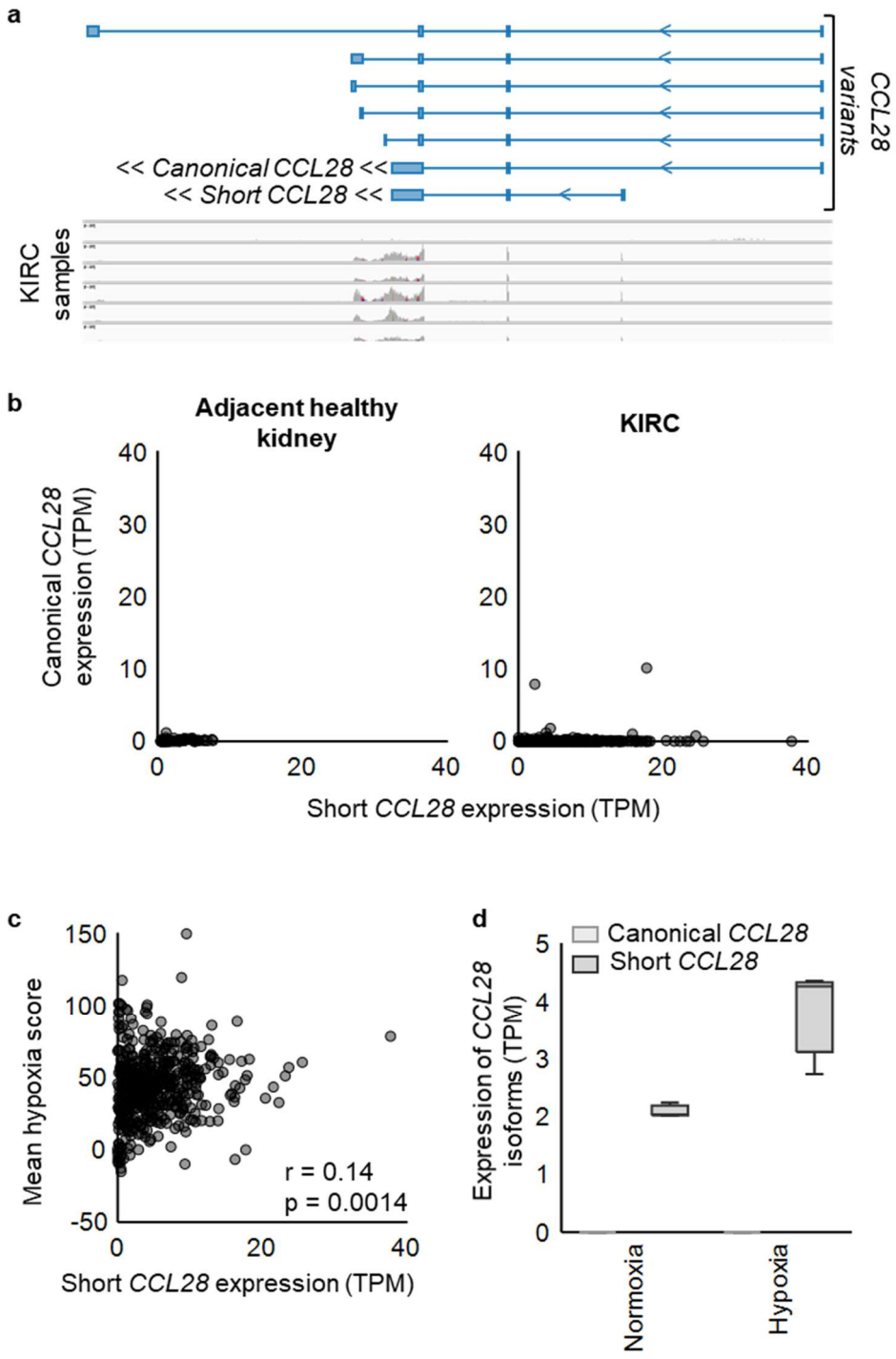


Figure 44: The structure of the *CCL28* locus and expression of the canonical and short isoforms. a. The structure of the *CCL28* locus alongside BAM files of RNAseq data from KIRC patient samples from TCGA. **b.** The expression of the canonical and short isoforms in 72 adjacent healthy kidney samples from TCGA (left) and 538 KIRC samples from TCGA (right). **c.** The association between expression of the short *CCL28* isoform with the mean hypoxia score (Lombardi et al., 2022) per sample in TCGA KIRC patients. **d.** The expression of canonical and short form *CCL28* in RCC4^{VHL+} cells (Smythies et al., 2019) under normoxic and hypoxic conditions.

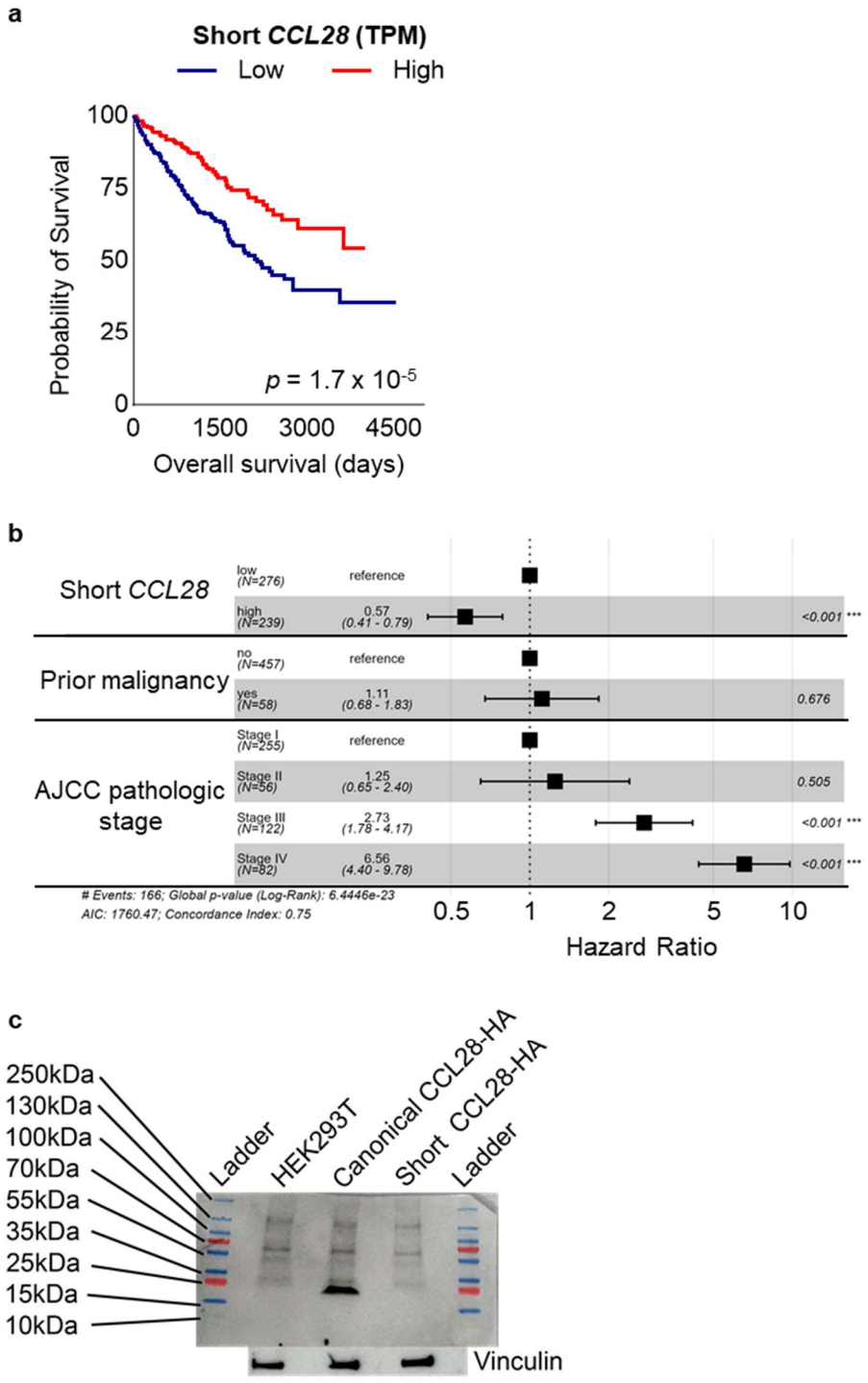


Figure 45: Survival analysis and protein stability of the short *CCL28* isoform. a. Univariate survival analysis for the expression of the short *CCL28* isoform in 515 KIRC patients, using a threshold of 4.36 TPM with 276 patients categorised as having low expression and 239 as high. **b.** Multivariate analysis of the effect of expression of short *CCL28* on patient survival. **c.** A Western blot showing the stability of canonical CCL28-HA (with an estimated molecular weight of 17.6 kDa), however the short CCL28-HA (with an estimated molecular weight of 12.6 kDa) could not be detected.

This short *CCL28* isoform is also inducible under hypoxia. Although there is weak correlation with the sample hypoxia score (Lombardi et al., 2022) (Pearson's correlation coefficient, $r = 0.14$, $p = 0.0014$, Figure 44c) in a renal cell carcinoma cell line $RCC4^{VHL+}$ (2.1.7: Renal carcinoma cell line dataset) the short isoform expression is increased under hypoxic conditions (Figure 44d). In these cells the canonical isoform is not expressed, reflecting what is seen in the KIRC patient samples.

Cox univariate survival analysis showed the short form of *CCL28* was associated with survival of KIRC patients (2.4.7: Survival analysis). Here increased expression of the short form associated with better patient survival (HR = 0.49, CI_{95} = 0.36 – 0.67, $p = 1.7 \times 10^{-5}$, Figure 45a). Analysis of clinical variables potentially effecting survival showed short form *CCL28* expression associated with pathologic stage (ANOVA, $p \leq 0.05$) and prior malignancy (t-test, $p \leq 0.5$). Multivariate survival analysis showed increased *CCL28* short isoform expression significantly reduced the hazard ratio independently of these other variables (HR = 0.57, CI_{95} = 0.41 – 0.79, $p = 6.97 \times 10^{-4}$, Figure 45b).

To characterise the potential stability of the protein encoded by this short *CCL28* isoform, HEK293T cells were transduced with plasmid expressing either the canonical *CCL28* or the short *CCL28* (2.5: Preparation of stably transduced cell lines). The sequences included a HA tag as the commercial antibody against the canonical CCL28 may not have bound the short form (see Supplementary figure 2 for the transduction efficiency of the cell populations). Western blot of lysates from these cells showed only the canonical isoform was stable (Figure 45c, 2.6: Sample preparation and Western Blot) implying the expression of this transcript must be a marker of some other process.

7.3.3.2 A truncated ENPP3 isoform reducing the survival advantage of canonical ENPP3

Of the genes overlapped by the 3681 KIRC-specific transcripts, *ENPP3* was overlapped 14 times. *ENPP3* has been introduced previously in Chapter 5 (5.3.2: A novel truncated isoform of *ENPP3*). The canonical ENPP3 protein is upregulated in KIRC (Doñate et al., 2016; Thompson et al., 2018) potentially due to a HIF-inducible LTR upstream of the gene (Siebenthall et al., 2019). One of the transcripts identified in this analysis is the truncated *ENPP3* transcript described in Chapter 5 (5.3.2: A novel truncated isoform of *ENPP3*) as possibly coding for an antigenic truncated ENPP3 protein, however the protein was not stable.

There was strong upregulation of both the canonical and truncated *ENPP3* isoforms at the RNA level between adjacent healthy kidney and KIRC samples from TCGA (Figure 46a). Although the *ENPP3* locus has been shown to be hypoxia-inducible (Siebenthall et al., 2019), there was weak correlation of the mean hypoxia score (Lombardi et al., 2022) with both the canonical *ENPP3* isoform (Pearson's correlation coefficient, $r = 0.16$, $p = 5.2 \times 10^{-4}$, Figure 46b) and the truncated isoform (Pearson's correlation coefficient, $r = 0.20$, $p = 8.6 \times 10^{-6}$, Figure 46b). Interestingly, analysis of RNAseq data from a renal cell carcinoma cell line $RCC4^{VHL+}$ (2.1.7: Renal carcinoma cell line dataset) showed only the truncated isoform expression increased under hypoxic conditions (Figure 46c), suggesting this isoform is under separate control to the canonical even though they share many of the same exons.

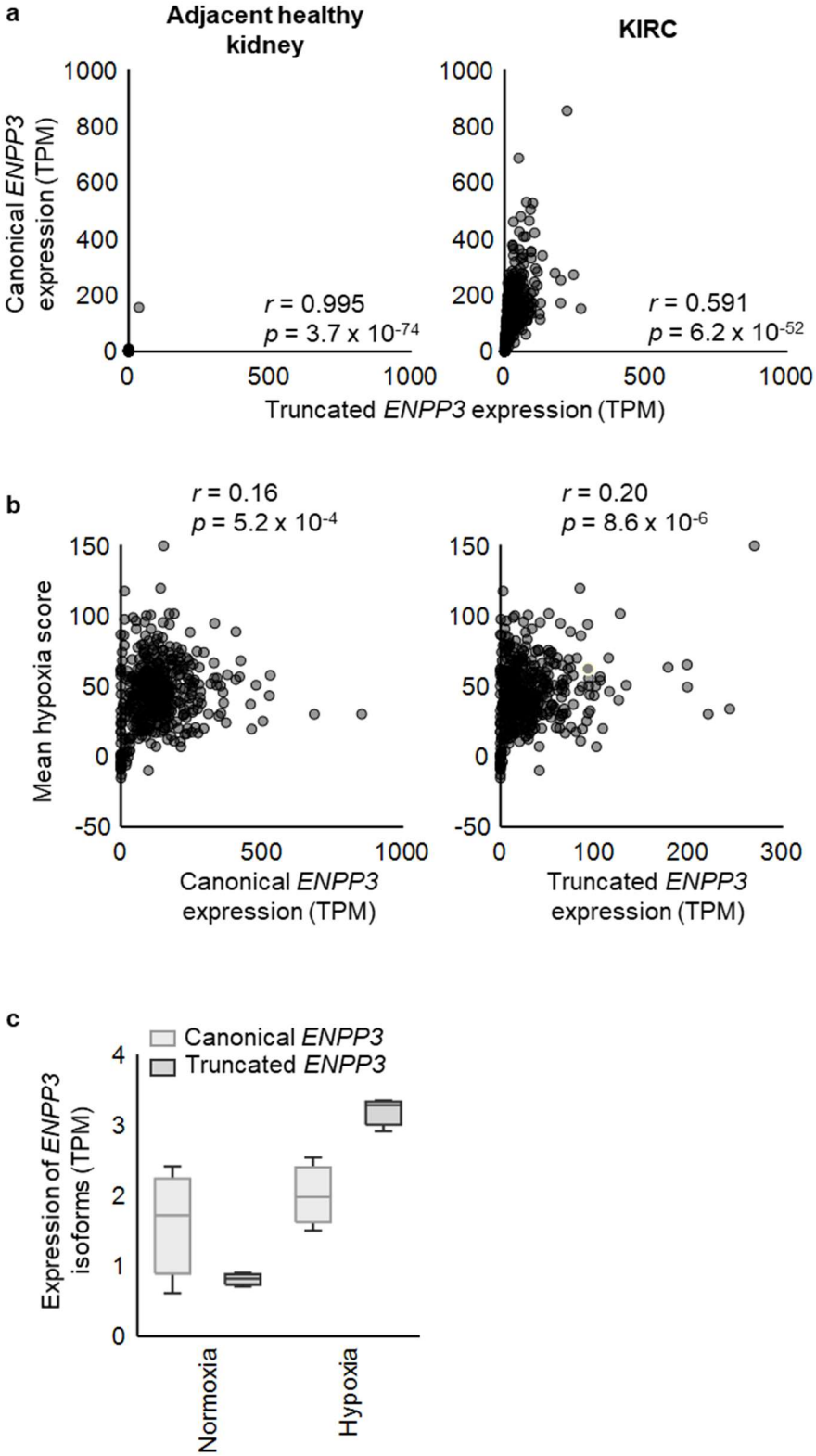

a
Adjacent healthy kidney
KIRC
Canonical ENPP3 expression (TPM)
Truncated ENPP3 expression (TPM)
r = 0.995
p = 3.7 x 10-74
r = 0.591
p = 6.2 x 10-52
b
r = 0.16
p = 5.2 x 10-4
r = 0.20
p = 8.6 x 10-6
Mean hypoxia score
Canonical ENPP3 expression (TPM)
Truncated ENPP3 expression (TPM)
c
Canonical ENPP3
Truncated ENPP3
Expression of ENPP3 isoforms (TPM)
Normoxia
Hypoxia

Figure 46: The expression of the canonical and truncated isoforms of *ENPP3*. a. The expression of the canonical and truncated isoforms in 72 adjacent healthy kidney samples from TCGA (left) and 538 KIRC samples from TCGA (right). **b.** The correlation of the canonical *ENPP3* isoform (left) and the truncated *ENPP3* isoform (right) with the mean hypoxia score (Lombardi et al., 2022) per sample in TCGA KIRC samples. **c.** The expression of canonical and truncated *ENPP3* in RCC4^{VHL+} cells (Smythies et al., 2019) under normoxic and hypoxic conditions.

Both the canonical and truncated *ENPP3* isoforms were associated with survival (2.4.7: Survival analysis). Canonical *ENPP3* expression is associated with better survival in KIRC patients (HR = 0.46, 0.34 – 0.62, CI_{95} = 0.34 – 0.62, p = 8.56 x 10^{-7}, Figure 47a). The truncated ENPP3 transcript previously described in Chapter 5 (5.3.2: A novel truncated isoform of *ENPP3*), which did not produce stable protein, also associated with better patient survival (HR = 0.63, CI_{95} = 0.47 – 0.86, p = 0.0048, Figure 47b). This was expected as the expression of the truncated isoform is highly correlated with the expression of the canonical isoform in both adjacent healthy kidney (Pearson's correlation coefficient, r = 0.995, p = 3.7 x 10^{-74}, Figure 46a) and in KIRC samples (Pearson's correlation coefficient, r = 0.591, p = 6.2 x 10^{-52}, Figure 46a). In order to see if the truncated isoform had an effect on survival independently of the locus expression, the ratio of truncated to canonical *ENPP3* expression was used. This ratio had the opposing survival association to either isoform alone (HR = 1.39, CI_{95} = 0.98 – 1.96, p = 0.045, Figure 47c, 2.4.7: Survival analysis) using a threshold of 0.26 to split the patients. This suggests that regardless of the overall amount of locus expression, if the truncated isoform expression is more than 20.6% of the canonical isoform then too much of the canonical form is replaced by the non-protein producing isoform and the survival advantage of the canonical ENPP3 is lost. The ratio of the truncated and canonical *ENPP3* isoforms did not associate with any other clinical variable tested so multivariate survival analysis was not run.

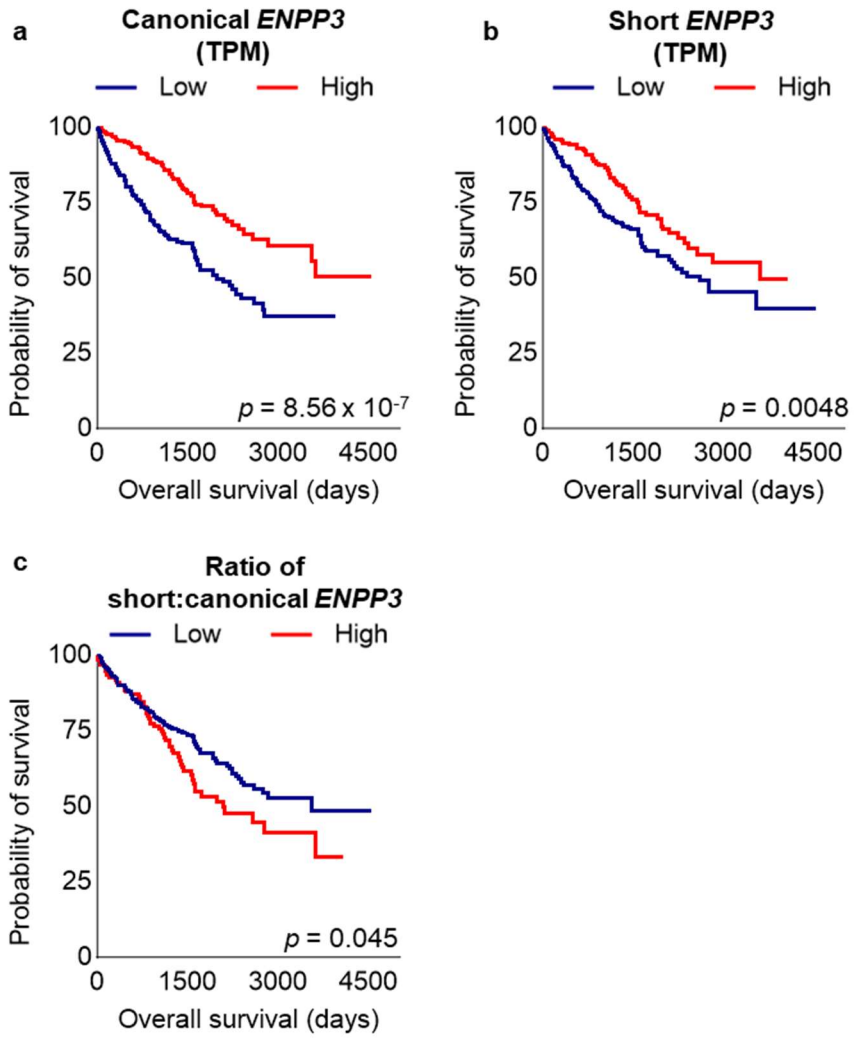


Figure 47: Survival analysis for the canonical and truncated isoforms of *ENPP3* and their ratio. a. Univariate survival analysis for the expression of canonical *ENPP3* in 515 KIRC patients using a threshold of 107 TPM with 241 patients defined as having low expression and 274 as high. **b.** Univariate survival analysis for the expression of truncated *ENPP3* in 515 KIRC patients, with a threshold of 23.5 TPM defining 297 patients as low expressers and 218 as high. **c.** Univariate survival analysis for the ratio of expression of the truncated to the canonical isoforms of *ENPP3* using a threshold of 0.26 defining 380 patients as having a low ratio and 135 as having a high ratio.

7.4 Discussion

Non-coding isoforms induced under hypoxia associate with survival. The majority of KIRC tumours undergo continuous activation of the hypoxia response pathway through constitutive HIF activation due to the loss of VHL. To explore the complex effect tumour hypoxia has on survival of KIRC patients, transcripts assembled in the *de novo* transcriptome were analysed. Of the transcripts assembled, 3681 were upregulated in KIRC compared to healthy tissues though may also be expressed in other cancer types. The majority of these transcripts (2413/3681) were associated with the overall survival of patients. This included a short isoform of *CCL28* and the truncated *ENPP3* isoform which was characterised in Chapter 5 (5.3.2: A novel truncated isoform of *ENPP3*).

The expression of a short *CCL28* isoform was associated with better overall survival of patients despite canonical CCL28 recruiting immunosuppressive Tregs under hypoxic conditions in other cancer types. The difference in the expected result was explained by adjacent healthy kidney and KIRC samples rarely expressing canonical *CCL28*, instead the short isoform which does not give rise to a stable protein is dominant. The canonical isoform is induced under hypoxia in a HIF1α-dependent manner, but in RCC4^{VHL+} cells only the short isoform was expressed and induced further under hypoxia. Thus, here the better overall survival associated with the short isoform of *CCL28* is not due to loss of the canonical form in some patients but is instead a marker of some other process associated with hypoxia. This cell type-specific expression of isoforms both driven by HIFs may be due to methylation of the relevant promoters for each isoform, methylation of CpG nucleotides within the consensus binding sequences for HIFs sterically hinders HIF binding (D'Anna et al., 2020).

Additionally, expression of a truncated *ENPP3* isoform was associated with better overall survival in patients. This truncated isoform does not produce stable protein thus the association with survival was likely due to the high correlation of the truncated *ENPP3* with canonical *ENPP3*, the expression of which was also highly associated with better patient survival. The ratio of truncated to canonical

ENPP3 was calculated to understand the association of the truncated isoform with survival independently of locus expression. Increase in the ratio of truncated to canonical isoforms was associated with poorer patient survival. This implies the canonical isoform is beneficial for patient survival and increases in the proportion of the truncated form (which does not produce stable protein) is therefore detrimental to overall survival. Why the canonical ENPP3 protein is beneficial to patient survival is yet unknown. Previous work has shown the *ENPP3* locus is upregulated under hypoxia due to an upstream HIF-responsive *LTR*. The RCC4^{VHL+} cell line data showed that only the truncated isoform was upregulated under hypoxia, suggesting hypoxia additionally controls usage of the *LINE2A* element the truncated form terminates in.

There is lack of strong correlation between the selected transcripts and the pan-cancer hypoxia score even when cell line data shows hypoxia has an effect on expression. The 3681 transcripts were selected from the original list of 32264 cancer-specific transcripts which were filtered for expression in at least 25% of the sample of a given tumour type. This consistent transcript expression was further selected for by choosing transcripts significantly upregulated in the KIRC sample population when compared to adjacent healthy kidney. This consistent expression of transcripts is not reflected in the range of hypoxia scores calculated for KIRC samples (Lombardi et al., 2022), reducing the ability to detect hypoxia-induced transcripts in the list of 3681. The transcript expression could be correlated with pan-cancer hypoxia scores to allow for a larger range of transcript expression and hypoxia pathway induction, but this would mask any KIRC-specific regulation. To identify a broader set of hypoxia-induced transcripts the pan-cancer scores and cancer-specific transcript list could be correlated. Furthermore, other factors such as tumour purity, the level of hypoxia required for transcript induction, and non-overlapping roles of HIF1α and HIF2α may also have influenced how well correlated the transcript expression and hypoxia score were.

It is possible many of the transcripts upregulated in KIRC only appear to be upregulated due to being stabilised by the inhibition of nonsense-mediated decay. The quantity of a given transcript in a cell is a balance between the production (transcription) and the stability (through nonsense-mediated decay or other pathways), thus to increase the perceived expression of a transcript either the transcription must increase or the decay decrease. Inhibiting nonsense-mediated decay leads to accumulation of transcripts in stress granules (Gardner, 2008) increasing the expression measured through RNAseq. As KIRC undergoes continuous hypoxia there may be a large number of transcripts selected as upregulated in KIRC as they are no longer being degraded. Some of the transcripts here positively associated with survival may be a marker of cellular stress and chronic accumulation of transcripts. In order to understand which transcripts are induced under hypoxia in KIRC, global run-on sequencing (Core et al., 2008) datasets surveying nascent transcription could be used.

7.5 Conclusion

Due to the constitutive activation of the hypoxia response pathways in the majority of KIRC tumours, the transcripts upregulated in this tumour type were analysed to further understand the association of hypoxia in tumours and patient survival. Hypoxia influences the expression of a range of transcripts which may in turn influence patient survival in opposing ways. The selection of transcripts upregulated under hypoxic conditions is dependent upon the methylation landscape of the cell as HIFs are unable to bind methylated regions (D'Anna et al., 2020). A short non-coding *CCL28* isoform is upregulated in a hypoxia dependent manner and is associated with better patient survival, though what this isoform is a marker of and why this isoform is chosen over the canonical in kidney tissues is unknown. With the hypomethylated state of RTEs in cancer, HIFs may be able to influence the expression of more transcripts than in healthy tissues through binding of RTE sequences. *ENPP3* is downstream of an LTR able to be bound by HIFs allowing upregulation of the gene locus under hypoxia (Siebenthall et al., 2019). But the benefit to patients conferred by the canonical

protein expression is limited by expression of a truncated non-protein coding isoform terminating in a *LINE2A* which is induced under hypoxia differently to the canonical isoform.

Chapter 8. General Discussion

8.1 Summary of findings

Analysis of transcripts identified in the previously assembled pan-cancer *de novo* transcriptome has revealed effects of RTEs on cancer-promoting and cancer-repressing genes (Attig et al., 2019) and uncovered a greater search space for the identification of cancer-specific targets and biomarkers. Exploring the 32264 cancer-specific transcripts of the 1001931 assembled transcripts showed that 95.90% overlapped at least one RTE with enrichment of SVA and HERV elements. Although these transcripts could be used to distinguish various tumour types from others and from healthy tissues using RNAseq from solid biopsies, this was not the case with liquid biopsies. There was limited detection of most of the selected BRCA-specific transcripts in patient blood samples, and sequences that were detected appeared to be an artefact of the methods used in one study. However, RTE-derived transcripts may act as a source of transmembrane and other antigens for use as biomarkers or as therapy targets. Of the 32264 cancer-specific transcripts 313 contained at least one ORF predicted to code for at least one transmembrane domain. Of the three candidates tested for stability and localisation, a truncated ENPP3 with sequence donated by a *LINE2A*, a truncated PLD3 with sequence donated by an *AluJr*, and a non-canonical *HERV-H* ORF, only the truncated PLD3 was stable. Although the protein did not localise to the surface membrane at detectable levels, this still indicates potential for other transmembrane proteins containing RTE-derived sequences to be stable. Furthermore, the expression specifically of HERV loci and derived transcripts in metastatic KIRC patients was able to distinguish responders from non-responders treated with anti-PD-1 therapy. This ability to distinguish response was most likely due to the HERV loci expression correlating with immune infiltrate, and the HERV-derived transcript expression correlating with tumour purity. Additionally, transcripts upregulated in primary KIRC and associated with patient survival were also explored to gain a better understanding of disease mechanisms under continuous hypoxia responses. The truncated *ENPP3*

isoform was shown to be induced under hypoxia and reduced the survival advantage conferred to patients by the canonical *ENPP3*. An isoform of *CCL28* was also characterised which had opposing survival associations to the immunosuppressive Treg-recruiting canonical isoform expressed in other cancers. This *CCL28* isoform was also induced under hypoxia and did not give rise to stable protein.

8.2 The extent to which RTEs contribute to the cancer-specific transcriptome

RTEs contribute to both the healthy and cancer transcriptomes, providing control elements like TF binding sites, splice sites, enhancer and promoter regions, and poly(A) tails, as well as contributing peptide sequences to both RTE-only and RTE-chimeric transcript ORFs. In the cancer transcriptome these effects may be more pronounced due to the hypomethylated state of the genome allowing for activation of RTE sequences. Cancer-specific transcripts identified from a previously assembled pan-cancer *de novo* transcriptome (Attig et al., 2019) revealed the broad effects of RTEs across cancer transcriptomes. This transcriptome assembly has identified novel cancer-specific transcripts which, as the selection criteria ensured expression across at least 25% of patients per cancer type, may act as a source of RNA and potentially protein biomarkers and treatment targets. The transcripts assembled also offer an insight into the biology of tumours, such as explaining a potential source of the ectopic expression of a brain-specific gene *GABRA3* in TCGT, LUSC, SKCM, and SKCM_m, and explaining why *CCL28* gene expression is associated with better patient survival in KIRC where a short isoform was dominant over the canonical. The cancer-specific transcripts were mainly enriched in groups of SVA and HERV elements, and there was also enrichment for specific LINE, SINE, and MaLR groups. A few RTE groups were seen at lower than expected levels in cancer, but it would perhaps be more interesting to analyse transcripts specific to healthy tissues to see if these are enriched in other RTE types which are downregulated in cancer. However, the *de novo* transcriptome assembly was created using a subset of cancer samples from TCGA, so healthy-specific transcripts are likely to be rare.

To allow for this analysis, a transcriptome derived of samples from both healthy and cancer disease states must be assembled. A larger number of samples per tissue type also need to be used as pan-cancer assembly created since using a larger dataset from TCGA identified transcripts not assembled here (Shah et al., 2023). Additionally, other conditions may induce RTE expression such as ageing and infection. As cancer specificity here was defined by comparing to healthy tissues of varying ages and infection histories this reduced the likelihood of other conditions being the cause of activation. Furthermore, although the contribution of RTEs overlapping transcripts has been surveyed here, this data does not give insight into why these specific RTEs are upregulated or the effect other loci may be having on the expression of neighbouring transcripts whose sequence they do not donate to (the enhancer activity of RTEs). In order to understand the global context of transcript regulation, regulatory networks need to be built integrating RNAseq for transcript expression, ChIP-seq for DNA binding protein patterns, 3D genomic mapping data to understand DNA-DNA interactions, and DNA sequencing analysis to profile polymorphic insertions which may further be influencing the transcriptome. Long-read RNAseq data may also be incorporated to improve confidence in the source of RTE transcripts.

8.3 The use of RTE-derived sequences in liquid biopsies

As the cancer-specific transcripts identified by the *de novo* transcriptome assembly were able to distinguish healthy and cancerous tissues in solid biopsies, it is possible RNA released from these tissues into the blood would allow exRNA species to define patients bearing certain tumours. Of the 34 BRCA specific transcripts selected, only three were well detected in BRCA patient blood samples. All three transcripts had uneven read coverage with large peaks over *LINE1HS* and *AluSp* overlapping regions. This upregulation was not BRCA patient-specific but specific to the published dataset, with upregulation seen in the healthy samples of this dataset but not in the healthy samples of other independently published datasets. As there is no accepted quality control for exRNA samples it was impossible to tell if the increased detection of RTE

sequences was reflective of better or poorer sample quality. In an attempt to understand the sample quality reads were aligned to *ACTB* and *GAPDH*, and read alignment as well as reads aligned over splice junctions was analysed. More recent studies had greater alignment of reads and spliced reads over these control sequences, and had more similar patterns of splicing to those seen in tissue data, suggesting these patient exRNA profiles were of better quality than earlier data with limited read alignment to controls and limited numbers of spliced reads. This approach may be a valuable measure of DNA contamination and depth of sequencing in exRNA datasets. In order to conclude the usefulness of RTE-derived cancer-specific transcripts as a liquid biomarker, data quality measures need to be agreed upon and datasets with longer read lengths need to be analysed to ensure specific mapping to the transcripts selected. Systematic analysis of variables affecting RTE detection in blood is also required to ensure the detection is due to RNA presence and not DNA contamination.

8.4 RTE-derived transcripts as a source of transmembrane antigen

Highly cancer-specific proteins localised to the cell membrane can be used to distinguish malignant cells from healthy, and can therefore be used as biomarkers and as antigens for targeted therapy. Of the three candidates tested identified from the *de novo* transcriptome assembly, only a truncated PLD3 with sequence donated from an *AluJr* element was stable in HEK293T cells, however the protein did not localise to the surface membrane at detectable levels. The study detecting canonical PLD3 on the surface of cells noted the rarity of the localisation (Gonzalez et al., 2018), a more sensitive approach may have detected localisation of truncated PLD3 to surface but whether this would be enough to elicit an anti-tumour immune response is unknown. Alternatively, the truncated protein may be processed and displayed via MHC-I molecules on the cell surface, with the 36 AA derived from the *AluJr* potentially giving rise to a set of antigenic peptides. Another candidate tested, the truncated *ENPP3* terminating in a *LINE2A*, was not stable and further analysis showed that increased ratios of truncated to canonical *ENPP3* was detrimental to patient survival. Canonical

ENPP3 is used as a transmembrane target of antibody drug conjugates (Thompson et al., 2018) and here activation of a *LINE2A* reduced expression of the canonical. The stability of the truncated PLD3 does suggest other transmembrane protein candidates containing protein sequence derived from RTEs may be stable, but prioritising candidates for *in vitro* testing is difficult without tools to accurately predict stability. Peptidomics and Ribo-seq data could be mined, but the ability of these datasets to differentiate isoforms of genes is limited, and the false discovery rate would drastically increase if all possible RTE-overlapping ORFs were added to the libraries searched. RTE-derived peptides are an attractive source of antigen as they are more likely to be shared between different patients than antigens which require sequence mutation. However, only in rare cases is the tumour dependent on RTE expression meaning it is possible for the tumour to escape a therapy targeted to RTE-derived peptides. As RTE expression is generally dysregulated in cancer due to hypomethylation of the genome, it is unknown how rapidly tumour evolution would be able to silence expression of a given locus. In xenograft mouse models HERV-K and HERV-E envelope protein targeted cytotoxic T-cells were able to reduce tumour growth without the tumour escaping (Cherkasova et al., 2016; Wang-Johanning et al., 2012; Zhou et al., 2016). But these xenograft tumour models are grown over a very short period of time compared to the amount of time tumours remain and are treated for in humans and thus do not reflect the evolutionary potential of a human malignancy.

8.5 RTE expression in stratifying patients for immune checkpoint blockade treatment

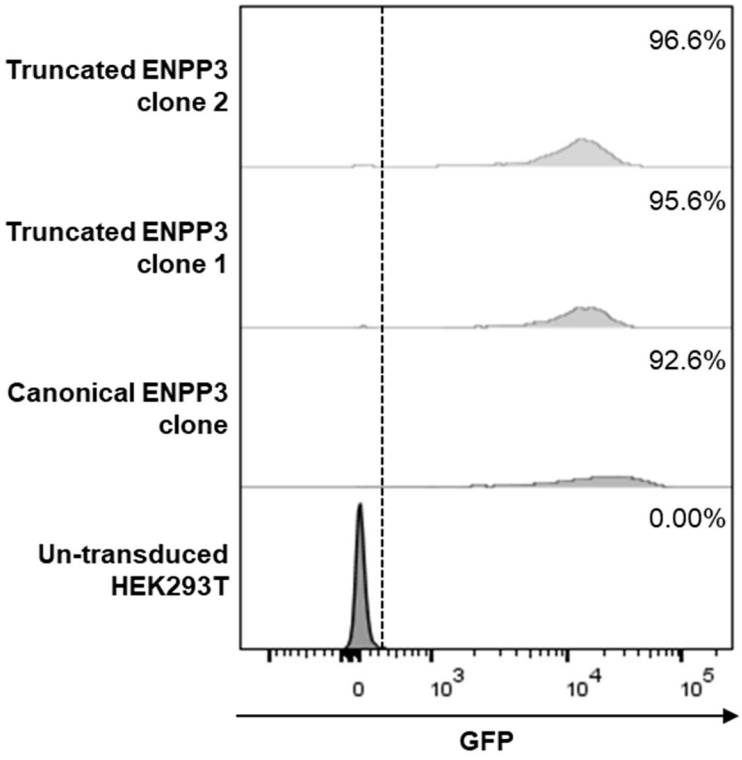
Previous work analysing RNAseq data from KIRC patients has shown HERV expression is increased in patients responding to immune checkpoint blockade therapy and HERV expression is positively correlated with cytotoxic T-cell signatures. In order to analyse expression of previously identified HERV loci (Mayer et al., 2011; Vargiu et al., 2016) in the context of GRCh38, the corresponding coordinates were found and compared to an updated annotation of HERVs from a Dfam-derived library (Attig et al., 2017). This comparison

revealed the errors in previous loci annotations, with some fragmented, incomplete, or over-extended to include other non-LTR RTEs and even parts of gene exons. According to PubMed, 40 studies have cited the list published by Mayer and colleagues (Mayer et al., 2011) and 96 studies have cited the list published by Vargiu and colleagues which is used by the hervQuant tool (Vargiu et al., 2016). This ongoing analysis leads to incorrect associations with HERV expression and hides potentially correct associations. Analysis of the updated annotation of HERV loci for GRCh38 and RNAseq from a small cohort of metastatic KIRC patients treated with anti-PD-1 therapy showed that HERV loci differentially expressed between responding and non-responding patients was correlated with immune infiltrate. These HERV loci were more highly expressed in responders, but on further inspection of purified immune cell subsets, it was found that the HERV expression was increased in these samples due to expression in immune cells. HERV-derived KIRC-specific transcript expression, which represented a different sequence space to the loci alone, was also analysed. The significantly differentially expressed transcripts were upregulated in non-responders, unlike the loci. The transcripts positively correlated with tumour purity which was higher in non-responders likely due to the lack of immune infiltrate (with the immune cells in turn expressing the identified HERV loci). Although HERV expression was differentially expressed between responders and non-responders the association with response was indirect, through tumour purity and immune infiltrate. In order to further understand the contribution of RTEs to patient responses to anti-PD-1 therapy, larger cohorts of patients need to be analysed with all RTE loci considered instead of just HERVs. Additionally, to reveal any immune response against RTE-derived antigens expression of the identified list of 313 candidate transcripts potentially coding for transmembrane domain containing proteins could be compared in responders and non-responders. As well as the expression of a larger list of KIRC-specific RTE-overlapping transcripts in HLA-typed individuals to explore the contribution of MHC-I display of predicted peptides from these sequences to patient responses to immunotherapy.

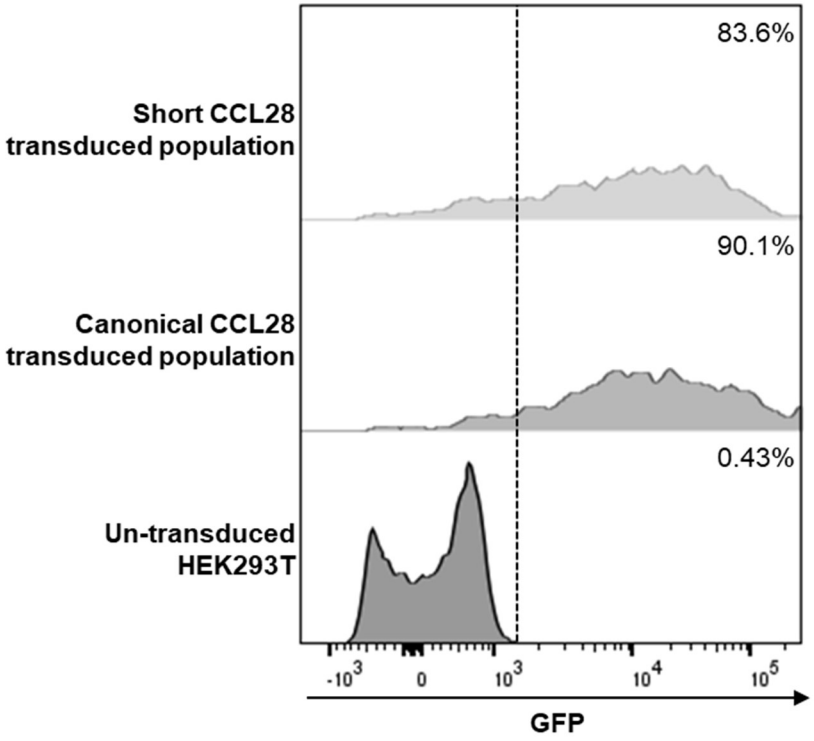
8.6 Conclusions

Overall, the upregulation of RTEs in cancer has complex implications for tumour progression and treatment. RTE-derived transcripts may provide antigen and allow for targeting of tumour cells, but these transcripts may also drive initiation and progression of cancer. Additionally, due to the repetitive and mutated nature of these elements annotating their coordinates has been problematic, further clouding understanding of RTE biology. RTEs also influence the transcriptome of healthy cells, both in co-opted functions and in some cases leading to the development of disease. With advancements in epigenetic drugs to treat cancers which hypomethylate the genome and are currently delivered in a systemic manner, further understanding of the effects of upregulated RTE-derived transcripts is required to ensure treatment is beneficial to patients.

Chapter 9. Appendix



Supplementary figure 1: Transduction of HEK293T cells with coding sequences of the ENPP3 isoforms. GFP expression as measured by flow cytometry of untransduced HEK293T cells and HEK293T cells transduced with constructs containing the canonical and truncated ENPP3-FLAG coding sequences. GFP presence indicates successful transduction (Figure 6) and the percentage of the population which was GFP positive is shown. Transduced populations were grown from single cells sorted on high GFP expression.



Supplementary figure 2: Transduction of HEK293T cells with the coding sequence of short CCL28. GFP expression as measured by flow cytometry of untransduced HEK293T cells and HEK293T cells transduced with a construct containing the short CCL28-HA coding sequence. GFP presence indicates successful transduction (Figure 6) and the percentage of the population which was GFP positive is shown.

Reference List

Aktaş, T., I.A. Ilik, D. Maticzka, V. Bhardwaj, C. Pessoa Rodrigues, G. Mittler, T. Manke, R. Backofen, and A. Akhtar. 2017. DHX9 suppresses RNA processing defects originating from the Alu invasion of the human genome. *Nature*. 544:115-119.

Albiges, L., T. Powles, M. Staehler, K. Bensalah, R.H. Giles, M. Hora, M.A. Kuczyk, T.B. Lam, B. Ljungberg, L. Marconi, A.S. Merseburger, A. Volpe, Y. Abu-Ghanem, S. Dabestani, S. Fernandez-Pello, F. Hofmann, T. Kuusk, R. Tahbaz, and A. Bex. 2019. Updated European Association of Urology Guidelines on Renal Cell Carcinoma: Immune Checkpoint Inhibition Is the New Backbone in First-line Treatment of Metastatic Clear-cell Renal Cell Carcinoma. *Eur. Urol.* 76:151-156.

Athanasiadis, A., A. Rich, and S. Maas. 2004. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS Biol.* 2:e391.

Attig, J., J. Pape, L. Doglio, A. Kazachenka, E. Ottina, G.R. Young, K.S.S. Enfield, I.V. Aramburu, K.W. Ng, N. Faulkner, W. Bolland, V. Papayannopoulos, C. Swanton, and G. Kassiotis. 2023. Human endogenous retrovirus onco-exaptation counters cancer cell senescence through calbindin. *The Journal of clinical investigation*. 133.

Attig, J., G.R. Young, L. Hosie, D. Perkins, V. Encheva-Yokoya, J.P. Stoye, A.P. Snijders, N. Ternette, and G. Kassiotis. 2019. LTR retroelement expansion of the human cancer transcriptome and immunopeptidome revealed by de novo transcript assembly. *Genome Res.*

Attig, J., G.R. Young, J.P. Stoye, and G. Kassiotis. 2017. Physiological and pathological transcriptional activation of endogenous retroelements assessed by RNA-sequencing of B lymphocytes. *Frontiers in Microbiology*. 8.

Au, L., E. Hatipoglu, M. Robert de Massy, K. Litchfield, G. Beattie, A. Rowan, D. Schnidrig, R. Thompson, F. Byrne, S. Horswell, N. Fotiadis, S. Hazell, D. Nicol, S.T.C. Shepherd, A. Fendler, R. Mason, L. Del Rosario, K. Edmonds, K. Lingard, S. Sarker, M. Mangwende, E. Carlyle, J. Attig, K. Joshi, I. Uddin, P.D. Becker, M.W. Sunderland, A. Akarca, I. Puccio, W.W. Yang, T. Lund, K. Dhillon, M.D. Vasquez, E. Ghorani, H. Xu, C. Spencer, J.I. Lopez, A. Green, U. Mahadeva, E. Borg, M. Mitchison, D.A. Moore, I. Proctor, M. Falzon, L. Pickering, A.J.S. Furness, J.L. Reading, R. Salgado, T. Marafioti, M. Jamal-Hanjani, P. Consortium, G. Kassiotis, B. Chain, J. Larkin, C. Swanton, S.A. Quezada, S. Turajlic, and T.R.R. Consortium. 2021. Determinants of anti-PD-1 response and resistance in clear cell renal cell carcinoma. *Cancer Cell*. 39:1497-1518 e1411.

Babarinde, I.A., G. Ma, Y. Li, B. Deng, Z. Luo, H. Liu, M.M. Abdul, C. Ward, M. Chen, X. Fu, L. Shi, M. Duttlinger, J. He, L. Sun, W. Li, Q. Zhuang, G. Tong, J. Frampton, J.B. Cazier, J. Chen, R. Jauch, M.A. Esteban, and A.P. Hutchins. 2021. Transposable element sequence fragments incorporated into coding and noncoding transcripts modulate the transcriptome of human pluripotent stem cells. *Nucleic Acids Res.* 49:9132-9153.

Bai, S., L. Chen, Y. Yan, X. Wang, A. Jiang, R. Li, H. Kang, Z. Feng, G. Li, W. Ma, J. Zhang, and J. Ren. 2022. Identification of Hypoxia–Immune-Related Gene Signatures and Construction of a Prognostic Model in Kidney Renal Clear Cell Carcinoma. *Frontiers in Cell and Developmental Biology*. 9:1-1.

Balaj, L., R. Lessard, L. Dai, Y.J. Cho, S.L. Pomeroy, X.O. Breakefield, and J. Skog. 2011. Tumour microvesicles contain retrotransposon elements and amplified oncogene sequences. *Nat. Commun.* 2:180-180.

Bao, W., K.K. Kojima, and O. Kohany. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA*. 6:11.

Bendall, M.L., J.H. Francis, A.N. Shoushtari, and D.F. Nixon. 2022. Specific human endogenous retroviruses predict metastatic potential in uveal melanoma. *JCI Insight*. 7.

Bennett, E.A., L.E. Coleman, C. Tsui, W.S. Pittard, and S.E. Devine. 2004. Natural genetic variation caused by transposable elements in humans. *Genetics*. 168:933-951.

Biswas, S., H. Troy, R. Leek, Y.L. Chung, J.L. Li, R.R. Raval, H. Turley, K. Gatter, F. Pezzella, J.R. Griffiths, M. Stubbs, and A.L. Harris. 2010. Effects of HIF-1alpha and HIF2alpha on Growth and Metabolism of Clear-Cell Renal Cell Carcinoma 786-0 Xenografts. *J Oncol*. 2010:757908.

Blaise, S., N. de Parseval, L. Benit, and T. Heidmann. 2003. Genomewide screening for fusogenic human endogenous retrovirus envelopes identifies syncytin 2, a gene conserved on primate evolution. *Proc Natl Acad Sci U S A*. 100:13013-13018.

Bolger, A.M., M. Lohse, and B. Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 30:2114-2120.

Bonaventura, P., V. Alcazer, V. Mutez, L. Tonon, J. Martin, N. Chuvin, E. Michel, R.E. Boulos, Y. Estornes, J. Valladeau-Guilemond, A. Viari, Q. Wang, C. Caux, and S. Depil. 2022. Identification of shared tumor epitopes from endogenous retroviruses inducing high-avidity cytotoxic T cells for cancer immunotherapy. *In* Science Advances. Vol. 8.

Bouras, A., M. Leone, V. Bonadona, M. Lebrun, A. Calender, and N. Boutry-Kryza. 2021. Identification and Characterization of New Alu Element Insertion in the BRCA1 Exon 14 Associated with Hereditary Breast and Ovarian Cancer. *Genes (Basel)*. 12.

Braun, D.A., Y. Hou, Z. Bakouny, M. Ficial, M. Sant' Angelo, J. Forman, P. Ross-Macdonald, A.C. Berger, O.A. Jegede, L. Elagina, J. Steinharter, M. Sun, M. Wind-Rotolo, J.C. Pignon, A.D. Cherniack, L. Lichtenstein, D. Neuberg, P. Catalano, G.J. Freeman, A.H. Sharpe, D.F. McDermott, E.M. Van Allen, S. Signoretti, C.J. Wu, S.A. Shukla, and T.K. Choueiri. 2020. Interplay of somatic alterations and immune infiltration modulates response to PD-1 blockade in advanced clear cell renal cell carcinoma. *Nat. Med.* 26:909-918.

Brocks, D., C.R. Schmidt, M. Daskalakis, H.S. Jang, N.M. Shah, D. Li, J. Li, B. Zhang, Y. Hou, S. Laudato, D.B. Lipka, J. Schott, H. Bierhoff, Y. Assenov, M. Helf, A. Ressnerova, M.S. Islam, A.M. Lindroth, S. Haas, M. Essers, C.D. Imbusch, B. Brors, I. Oehme, O. Witt, M. Lübbert, J.P. Mallm, K. Rippe, R. Will, D. Weichenhan, G. Stoecklin, C. Gerhäuser, C.C. Oakes, T. Wang, and C. Plass. 2017. DNMT and HDAC inhibitors induce cryptic transcription start sites encoded in long terminal repeats. *Nat. Genet.* 49:1052-1060.

Bühring, H.J., A. Streble, and P. Valent. 2004. The basophil-specific ectoenzyme E-NPP3 (CD203c) as a marker for cell activation and allergy diagnosis. *Int Arch Allergy Immunol*. 133:317-329.

Burbage, M., A. Rocañín-Arjó, B. Baudon, Y.A. Arribas, A. Merlotti, D.C. Rookhuizen, S. Heurtebise-Chrétien, M. Ye, A. Houy, N. Burgdorf, G. Suarez, M. Gros, B. Sadacca, M. Carrascal, A. Garmilla, M. Bohec, S. Baulande, B. Lombard, D. Loew, J.J. Waterfall, M.H. Stern, C. Goudot, and S. Amigorena. 2023. Epigenetically controlled tumor antigens derived from splice junctions between exons and transposable elements. *Science immunology*. 8.

Buschmann, D., B. Kirchner, S. Hermann, M. Märte, C. Wurmser, F. Brandes, S. Kotschote, M. Bonin, O.K. Steinlein, M.W. Pfaffl, G. Schelling, and M. Reithmair. 2018. Evaluation of serum extracellular vesicle isolation methods for profiling miRNAs by next-generation sequencing. *Journal of Extracellular Vesicles*. 7.

Bushnell, B. 2014. BBMap: A Fast, Accurate, Splice-Aware Aligner.

Cajuso, T., P. Sulo, T. Tanskanen, R. Katainen, A. Taira, U.A. Hänninen, J. Kondelin, L. Forsström, N. Välimäki, M. Aavikko, E. Kaasinen, A. Ristimäki, S. Koskensalo, A. Lepistö, L. Renkonen-Sinisalo, T. Seppälä, T. Kuopio, J. Böhm, J.P. Mecklin, O. Kilpivaara, E. Pitkänen, K. Palin, and L.A. Aaltonen. 2019. Retrotransposon insertions can initiate colorectal cancer and are associated with poor survival. *Nat. Commun.* 10.

Campbell, P.J., G. Getz, J.O. Korbel, J.M. Stuart, J.L. Jennings, L.D. Stein, M.D. Perry, H.K. Nahal-Bose, B.F.F. Ouellette, C.H. Li, E. Rheinbay, G.P. Nielsen, D.C. Sgroi, C.L. Wu, W.C. Faquin, V. Deshpande, P.C. Boutros, A.J. Lazar, K.A. Hoadley, D.N. Louis, L.J. Dursi, C.K. Yung, M.H. Bailey, G. Saksena, K.M. Raine, I. Buchhalter, K. Kleinheinz, M. Schlesner, J. Zhang, W. Wang, D.A. Wheeler, L. Ding, J.T. Simpson, B.D. O'Connor, S. Yakneen, K. Ellrott, N. Miyoshi, A.P. Butler, R. Royo, S.I. Shorser, M. Vazquez, T. Rausch, G. Tiao, S.M. Waszak, B. Rodriguez-Martin, S. Shringarpure, D.Y. Wu, G.M. Demidov, O. Delaneau, S. Hayashi, S. Imoto, N. Habermann, A.V. Segre, E. Garrison, A. Cafferkey, E.G. Alvarez, J.M. Heredia-Genestar, F. Muyas, O. Drechsel, A.L. Bruzos, J. Temes, J. Zamora, A. Baez-Ortega, H.L. Kim, R.J. Mashl, K. Ye, A. DiBiase, K.I. Huang, I. Letunic, M.D. McLellan, S.J. Newhouse, T. Shmaya, S. Kumar, D.C. Wedge, M.H. Wright, V.D. Yellapantula, M. Gerstein, E. Khurana, T. Marques-Bonet, A. Navarro, C.D. Bustamante, R. Siebert, H. Nakagawa, D.F. Easton, S. Ossowski, J.M.C. Tubio, F.M. De La Vega, X. Estivill, D. Yuen, G.L. Mihaiescu, L. Omberg, V. Ferretti, R. Sabarinathan, O. Pich, A.

Gonzalez-Perez, A. Taylor-Weiner, M.W. Fittall, J. Demeulemeester, M. Tarabichi, N.D. Roberts, et al. 2020. Pan-cancer analysis of whole genomes. *Nature*. 578:82-93.

Cao, X., Y. Zhang, L.M. Payer, H. Lords, J.P. Steranka, K.H. Burns, and J. Xing. 2020. Polymorphic mobile element insertions contribute to gene expression and alternative splicing in human tissues. *Genome Biol.* 21.

Cao, Y., G. Chen, G. Wu, X. Zhang, J. McDermott, X. Chen, C. Xu, Q. Jiang, Z. Chen, Y. Zeng, D. Ai, Y. Huang, and J.D.J. Han. 2019. Widespread roles of enhancer-like transposable elements in cell identity and long-range genomic interactions. *Genome Res.* 29:40-52.

Cardelli, M., R.v. Doorn, L. Larcher, M.D. Donato, F. Piacenza, E. Pierpaoli, R. Giacconi, M. Malavolta, S. Rachakonda, N.A. Gruis, A. Molven, P.A. Andresen, D. Pjanova, J.J. van den Oord, M. Provinciali, E. Nagore, and R. Kumar. 2020. Association of HERV-K and LINE-1 hypomethylation with reduced disease-free survival in melanoma patients. *Epigenomics*.

Carozza, J.A., V. Bohnert, K.C. Nguyen, G. Skariah, K.E. Shaw, J.A. Brown, M. Rafat, R. von Eyben, E.E. Graves, J.S. Glenn, M. Smith, and L. Li. 2020. Extracellular cGAMP is a cancer cell-produced immunotransmitter involved in radiation-induced anti-cancer immunity. *Nat Cancer.* 1:184-196.

Chen, J., A.-D. Brunner, J.Z. Cogan, J.K. Nunez, A.P. Fields, B. Adamson, D.N. Itzhak, J.Y. Li, M. Mann, M.D. Leonetti, and J.S. Weissman. 2020. Pervasive functional translation of noncanonical human open reading frames. *Science*. 367:1140-1146.

Chen, L.L., J.N. DeCerbo, and G.G. Carmichael. 2008. Alu element-mediated gene silencing. *EMBO J.* 27:1694-1705.

Chen, X., and D. Li. 2019. ERVcaller: Identifying polymorphic endogenous retrovirus and other transposable element insertions using whole-genome sequencing data. *Bioinformatics*. 35:3913-3922.

Cherkasova, E., C. Scrivani, S. Doh, Q. Weisman, Y. Takahashi, N. Harashima, H. Yokoyama, R. Srinivasan, W.M. Linehan, M.I. Lerman, and R.W. Childs. 2016. Detection of an immunogenic HERV-E envelope with selective expression in clear cell kidney cancer. *Cancer Res.* 76:2177-2185.

Chong, C., M. Müller, H.S. Pak, D. Harnett, F. Huber, D. Grun, M. Leleu, A. Auger, M. Arnaud, B.J. Stevenson, J. Michaux, I. Bilic, A. Hirsekorn, L. Calviello, L. Simó-Riudalbas, E. Planet, J. Lubiński, M. Bryśkiewicz, M. Wiznerowicz, I. Xenarios, L. Zhang, D. Trono, A. Harari, U. Ohler, G. Coukos, and M. Bassani-Sternberg. 2020. Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. *Nat. Commun.* 11.

Chu, C., R. Borges-Monroy, V.V. Viswanadham, S. Lee, H. Li, E.A. Lee, and P.J. Park. 2021. Comprehensive identification of transposable element insertions using multiple sequencing technologies. *Nat. Commun.* 12.

Chung, N., G.M. Jonaid, S. Quinton, A. Ross, C.E. Sexton, A. Alberto, C. Clymer, D. Churchill, O. Navarro Leija, and M.V. Han. 2019. Transcriptome analyses of tumor-adjacent somatic tissues reveal genes co-expressed with transposable elements. *Mobile DNA*. 10.

Chuong, E.B., N.C. Elde, and C. Feschotte. 2016. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science*. 351:1083-1087.

Cole, W.H., and T.C. Everson. 1956. Spontaneous regression of cancer: preliminary report. *Ann Surg*. 144:366-383.

Core, L.J., J.J. Waterfall, and J.T. Lis. 2008. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*. 322:1845-1848.

Couso, J.P., and P. Patraquim. 2017. Classification and function of small open reading frames. *Nature Reviews Molecular Cell Biology 2017 18:9*. 18:575-589.

D'Anna, F., L. Van Dyck, J. Xiong, H. Zhao, R.V. Berrens, J. Qian, P. Bieniasz-Krzywiec, V. Chandra, L. Schoonjans, J. Matthews, J. De Smedt, L. Minnoye, R. Amorim, S. Khorasanizadeh, Q. Yu, L. Zhao, M. De Borre, S.N. Savvides, M.C. Simon, P. Carmeliet, W. Reik, F. Rastinejad, M. Mazzone, B. Thienpont, and D. Lambrechts. 2020. DNA methylation repels binding of hypoxia-inducible transcription factors to maintain tumor immunotolerance. *Genome Biol.* 21:1-36.

Danecek, P., J.K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M.O. Pollard, A. Whitwham, T. Keane, S.A. McCarthy, R.M. Davies, and H. Li. 2021. Twelve years of SAMtools and BCFtools. *GigaScience*. 10.

De Cecco, M., T. Ito, A.P. Petrashen, A.E. Elias, N.J. Skvir, S.W. Criscione, A. Caligiana, G. Brocculi, E.M. Adney, J.D. Boeke, O. Le, C. Beauséjour, J. Ambati, K. Ambati, M. Simon, A. Seluanov, V. Gorbunova, P.E. Slagboom, S.L. Helfand, N. Neretti, and J.M. Sedivy. 2019. L1 drives IFN in senescent cells and promotes age-associated inflammation. *Nature*. 566:73-78.

de Nonneville, A., P. Finetti, L. Boudin, E. Denicolaï, D. Birnbaum, E. Mamessier, and F. Bertucci. 2023. Prognostic and Predictive Value of LIV1 Expression in Early Breast Cancer and by Molecular Subtype. *Pharmaceutics*. 15:938-938.

Dobin, A., C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T.R. Gingeras. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 29:15-21.

Dolci, M., C. Favero, W. Toumi, E. Favi, L. Tarantini, L. Signorini, G. Basile, V. Bollati, S. D'Alessandro, P. Bagnoli, P. Ferrante, and S. Delbue. 2020. Human Endogenous Retroviruses Long Terminal Repeat Methylation, Transcription, and Protein Expression in Human Colon Cancer. *Front. Oncol.* 10.

Doñate, F., A. Raitano, K. Morrison, Z. An, L. Capo, H. Aviña, S. Karki, K. Morrison, P. Yang, J. Ou, R. Moriya, Y. Shostak, F. Malik, R. Nadell, W. Liu, D. Satpayev, J. Atkinson, I.B.J. Joseph, D.S. Pereira, P.M. Challita-Eid, and D.R. Stover. 2016. AGS16F is a novel antibody drug conjugate directed against ENPP3 for the treatment of renal cell carcinoma. *Clin. Cancer Res.* 22:1989-1999.

DuBridge, R.B., P. Tang, H.C. Hsia, P.M. Leong, J.H. Miller, and M.P. Calos. 1987. Analysis of mutation in human cells by using an Epstein-Barr virus shuttle system. *Mol. Cell. Biol.* 7:379-387.

Dvinge, H., and R.K. Bradley. 2015. Widespread intron retention diversifies most cancer transcriptomes. *Genome Med.* 7:1-13.

Elbarbary, R.A., and L.E. Maquat. 2017. Distinct mechanisms obviate the potentially toxic effects of inverted-repeat Alu elements on cellular RNA metabolism. *Nat. Struct. Mol. Biol.* 24:496-498.

Esnault, C., J. Maestre, and T. Heidmann. 2000. Human LINE retrotransposons generate processed pseudogenes. *Nat. Genet.* 24:363-367.

Everaert, C., H. Helsmoortel, A. Decock, E. Hulstaert, R. Van Paemel, K. Verniers, J. Nuytens, J. Anckaert, N. Nijs, J. Tulkens, B. Dhondt, A. Hendrix, P. Mestdagh, and J. Vandesompele. 2019. Performance assessment of total RNA sequencing of human biofluids and extracellular vesicles. *Sci. Rep.* 9:1-16.

Facciabene, A., X. Peng, I.S. Hagemann, K. Balint, A. Barchetti, L.P. Wang, P.A. Gimotty, C.B. Gilks, P. Lal, L. Zhang, and G. Coukos. 2011. Tumour hypoxia promotes tolerance and angiogenesis via CCL28 and T reg cells. *Nature*. 475:226-230.

Fan, Y., H. Li, X. Ma, Y. Gao, L. Chen, X. Li, X. Bao, Q. Du, Y. Zhang, and X. Zhang. 2015. Prognostic significance of hypoxia-inducible factor expression in renal cell carcinoma: A PRISMA-compliant systematic review and meta-analysis. *Medicine (United States)*. 94:1-8.

Feng, J., Y. Li, L. Zhu, Q. Zhao, D. Li, Y. Li, and T. Wu. 2021. STAT1 mediated long non-coding RNA LINC00504 influences radio-sensitivity of breast cancer via binding to TAF15 and stabilizing CPEB2 expression. *Cancer Biology & Therapy*. 22:630-639.

Foroushani, A.K., B. Chim, M. Wong, A. Rastegar, P.T. Smith, S. Wang, K. Barbian, C. Martens, M. Hafner, and S.A. Muljo. 2020. Posttranscriptional regulation of human endogenous retroviruses by RNA-binding motif protein 4, RBM4. *Proc. Natl. Acad. Sci. U. S. A.* 117:26520-26530.

Foster, K., A. Prowse, A. van den Berg, S. Fleming, M.M. Hulsbeek, P.A. Crossey, F.M. Richards, P. Cairns, N.A. Affara, M.A. Ferguson-Smith, and et al. 1994. Somatic mutations of the von Hippel-Lindau disease tumour suppressor gene in non-familial clear cell renal carcinoma. *Hum. Mol. Genet.* 3:2169-2173.

Frankish, A., M. Diekhans, A.-M. Ferreira, R. Johnson, I. Jungreis, J. Loveland, J.M. Mudge, C. Sisu, J. Wright, J. Armstrong, I. Barnes, A. Berry, A. Bignell, S. Carbonell Sala, J. Chrast, F. Cunningham, T. Di Domenico, S. Donaldson, I.T. Fiddes, C. García Girón, J.M. Gonzalez, T. Grego, M. Hardy, T. Hourlier, T. Hunt, O.G. Izuogu, J. Lagarde, F.J. Martin, L. Martínez, S. Mohanan, P. Muir, F.C P. Navarro, A. Parker, B. Pei, F. Pozo, M. Ruffier, B.M. Schmitt, E. Stapleton, M.-M. Suner, I. Sycheva, B. Uszczynska-Ratajczak, J. Xu, A.

Yates, D. Zerbino, Y. Zhang, B. Aken, J.S. Choudhary, M. Gerstein, R. Guigó, T.J P. Hubbard, M. Kellis, B. Paten, A. Reymond, M.L. Tress, and P. Flicek. 2019. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* 47:D766-D773.

Frankiw, L., D. Baltimore, and G. Li. 2019. Alternative mRNA splicing in cancer immunotherapy. *Nat. Rev. Immunol.* 19:675-687.

Freedman, J.E., M. Gerstein, E. Mick, J. Rozowsky, D. Levy, R. Kitchen, S. Das, R. Shah, K. Danielson, L. Beaulieu, F.C.P. Navarro, Y. Wang, T.R. Galeev, A. Holman, R.Y. Kwong, V. Murthy, S.E. Tanriverdi, M. Koupenova-Zamor, E. Mikhalev, and K. Tanriverdi. 2016. Diverse human extracellular RNAs are widely detected in human plasma. *Nat. Commun.* 7:11106-11106.

Fuentes, D.R., T. Swigut, and J. Wysocka. 2018. Systematic perturbation of retroviral LTRs reveals widespread long-range effects on human gene regulation. *Elife*. 7.

Gardner, L.B. 2008. Hypoxic Inhibition of Nonsense-Mediated RNA Decay Regulates Gene Expression and the Integrated Stress Response. *Mol. Cell. Biol.* 28:3729-3741.

Gavin, A.L., D. Huang, T.R. Blane, T.C. Thinnes, Y. Murakami, R. Fukui, K. Miyake, and D. Nemazee. 2021. Cleavage of DNA and RNA by PLD3 and PLD4 limits autoinflammatory triggering by multiple sensors. *Nat. Commun.* 12:5874.

Gavin, A.L., D. Huang, C. Huber, A. Martensson, V. Tardif, P.D. Skog, T.R. Blane, T.C. Thinnes, K. Osborn, H.S. Chong, F. Kargaran, P. Kimm, A. Zeitjian, R.L. Sielski, M. Briggs, S.R. Schulz, A. Zarpellon, B. Cravatt, E.S. Pang, J. Teijaro, J.C. de la Torre, M. O'Keeffe, H. Hochrein, M. Damme, L. Teyton, B.R. Lawson, and D. Nemazee. 2018. PLD3 and PLD4 are single-stranded acid exonucleases that regulate endosomal nucleic-acid sensing. *Nat. Immunol.* 19:942-953.

Gezer, U., A.J. Bronkhorst, and S. Holdenrieder. 2022. The Utility of Repetitive Cell-Free DNA in Cancer Liquid Biopsies. *In* Diagnostics. Vol. 12. 1363-1363.

Gijsbers, R., H. Ceulemans, and M. Bollen. 2003. Functional characterization of the non-catalytic ectodomains of the nucleotide pyrophosphatase/phosphodiesterase NPP1. *Biochem. J.* 371:321-330.

Giraldez, M.D., R.M. Spengler, A. Etheridge, P.M. Godoy, A.J. Barczak, S. Srinivasan, P.L. De Hoff, K. Tanriverdi, A. Courtright, S. Lu, J. Khoory, R. Rubio, D. Baxter, T.A.P. Driedonks, H.P.J. Buermans, E.N.M. Nolte-'T Hoen, H. Jiang, K. Wang, I. Ghiran, Y.E. Wang, K. Van Keuren-Jensen, J.E. Freedman, P.G. Woodruff, L.C. Laurent, D.J. Erle, D.J. Galas, and M. Tewari. 2018. Comprehensive multi-center assessment of small RNA-seq methods for quantitative miRNA profiling. *Nat. Biotechnol.* 36:746-757.

Giraldez, M.D., R.M. Spengler, A. Etheridge, A.J. Goicochea, M. Tuck, S.W. Choi, D.J. Galas, and M. Tewari. 2019. Phospho‐RNA‐seq: a modified small RNA‐seq method that reveals circulating mRNA and lncRNA fragments as potential biomarkers in human plasma. *EMBO J.* 38.

Gnarra, J.R., K. Tory, Y. Weng, L. Schmidt, M.H. Wei, H. Li, F. Latif, S. Liu, F. Chen, F.M. Duh, and et al. 1994. Mutations of the VHL tumour suppressor gene in renal carcinoma. *Nat. Genet.* 7:85-90.

Goke, J., X. Lu, Y.S. Chan, H.H. Ng, L.H. Ly, F. Sachs, and I. Szczerbinska. 2015. Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. *Cell Stem Cell*. 16:135-141.

Gonzalez, A.C., M. Schweizer, S. Jagdmann, C. Bernreuther, T. Reinheckel, P. Saftig, and M. Damme. 2018. Unconventional Trafficking of Mammalian Phospholipase D3 to Lysosomes. *Cell Rep.* 22:1040-1053.

Goubert, C., N.A. Zevallos, and C. Feschotte. 2020. Contribution of unfixed transposable element insertions to human regulatory variation. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 375.

Goyal, A., J. Bauer, J. Hey, D.N. Papageorgiou, E. Stepanova, M. Daskalakis, J. Scheid, M. Dubbelaar, B. Klimovich, D. Schwarz, M. Märklin, M. Roerden, Y.-Y. Lin, T. Ma, O. Mücke, H.-G. Rammensee, M. Lübbert, F. Loayza-Puch, J. Krijgsveld, J.S. Walz, and C. Plass. 2023. DNMT and HDAC inhibition induces immunogenic neoantigens from human endogenous retroviral element-derived transcripts. *Nat. Commun.* 14:6731-6731.

Grabherr, M.G., B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. Di

Palma, B.W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29:644-652.

Grabski, D.F., A. Ratan, L.R. Gray, S. Bekiranov, D. Rekosh, M.L. Hammarskjold, and S.K. Rasmussen. 2021. Upregulation of human endogenous retrovirus-K (HML-2) mRNAs in hepatoblastoma: Identification of potential new immunotherapeutic targets and biomarkers. *Journal of Pediatric Surgery*. 56:286-292.

Graham, F.L., J. Smiley, W.C. Russell, and R. Nairn. 1977. Characteristics of a human cell line transformed by DNA from human adenovirus type 5. *J. Gen. Virol.* 36:59-74.

Groot, M., and H. Lee. 2020. Sorting Mechanisms for MicroRNAs into Extracellular Vesicles and Their Associated Diseases. *Cells*. 9:1-16.

Groppe, D.M. 2010. Bonferroni-Holm (1979) correction for multiple comparisons, University of California, San Diego.

Gumireddy, K., A. Li, A.V. Kossenkov, M. Sakurai, J. Yan, Y. Li, H. Xu, J. Wang, P.J. Zhang, L. Zhang, L.C. Showe, K. Nishikura, and Q. Huang. 2016. The mRNA-edited form of GABRA3 suppresses GABRA3-mediated Akt activation and breast cancer metastasis. *Nat. Commun.* 7:10715.

Guo, Z.H., L.T. Yao, and A.Y. Guo. 2020. Clinical and biological impact of LINC02544 expression in breast cancer after neoadjuvant chemotherapy. *European review for medical and pharmacological sciences*. 24:10573-10585.

Happel, C., A. Ganguly, and D.A. Tagle. 2020. Extracellular RNAs as potential biomarkers for cancer. *Journal of Cancer Metastasis and Treatment*. 2020.

Harris, C.R., K.J. Millman, S.J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N.J. Smith, R. Kern, M. Picus, S. Hoyer, M.H. van Kerkwijk, M. Brett, A. Haldane, J.F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T.E. Oliphant. 2020. Array programming with NumPy. *Nature*. 585:357-362.

Hartl, D.L. 2000. Molecular melodies in high and low C. *Nat. Rev. Genet.* 1:145-149.

Herman, J.G., F. Latif, Y. Weng, M.I. Lerman, B. Zbar, S. Liu, D. Samid, D.S. Duan, J.R. Gnarra, W.M. Linehan, and et al. 1994. Silencing of the VHL tumor-suppressor gene by DNA methylation in renal carcinoma. *Proc Natl Acad Sci U S A*. 91:9700-9704.

Hinger, S.A., D.J. Cha, J.L. Franklin, J.N. Higginbotham, Y. Dou, J. Ping, L. Shu, N. Prasad, S. Levy, B. Zhang, Q. Liu, A.M. Weaver, R.J. Coffey, and J.G. Patton. 2018. Diverse Long RNAs Are Differentially Sorted into Extracellular Vesicles Secreted by Colorectal Cancer Cells. *Cell Rep.* 25:715-725.e714.

Hirschfeld, M., A.Z. Hausen, H. Bettendorf, M. Jäger, and E. Stickeier. 2009. Alternative splicing of Cyr61 Is regulated by hypoxia and significantly changed in breast cancer. *Cancer Res.* 69:2082-2090.

Honorat, M., A. Mesnier, J. Vendrell, J. Guitton, I. Bieche, R. Lidereau, G.D. Kruh, C. Dumontet, P. Cohen, and L. Payen. 2008. ABCC11 expression is regulated by estrogen in MCF7 cells, correlated with estrogen receptor expression in postmenopausal breast tumors and overexpressed in tamoxifen-resistant breast cancer cells. *Endocrine Related Cancer*. 15:125-138.

Hoyt, S.J., J.M. Storer, G.A. Hartley, P.G.S. Grady, A. Gershman, L.G. de Lima, C. Limouse, R. Halabian, L. Wojenski, M. Rodriguez, N. Altemose, A. Rhie, L.J. Core, J.L. Gerton, W. Makalowski, D. Olson, J. Rosen, A.F.A. Smit, A.F. Straight, M.R. Vollger, T.J. Wheeler, M.C. Schatz, E.E. Eichler, A.M. Phillippy, W. Timp, K.H. Miga, and R.J. O'Neill. 2022. From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science*. 376.

Hu, Z., J. Yuan, M. Long, J. Jiang, Y. Zhang, T. Zhang, M. Xu, Y. Fan, J.L. Tanyi, K.T. Montone, O. Tavana, H.M. Chan, X. Hu, R.H. Vonderheide, and L. Zhang. 2021. The Cancer Surfaceome Atlas integrates genomic, functional and drug response data to identify actionable targets. *Nature Cancer*.

Huang, G., L. Tao, S. Shen, and L. Chen. 2016. Hypoxia induced CCL28 promotes angiogenesis in lung adenocarcinoma by targeting CCR3 on endothelial cells. *Sci. Rep.* 6:1-11.

Hunter, J.D. 2007. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*. 9:90-95.

Ivanov, A.V., H. Peng, V. Yurchenko, K.L. Yap, D.G. Negorev, D.C. Schultz, E. Psulkowski, W.J. Fredericks, D.E. White, G.G. Maul, M.J. Sadofsky, M.M. Zhou, and F.J. Rauscher, 3rd. 2007. PHD domain-mediated E3 ligase activity directs intramolecular sumoylation of an adjacent bromodomain required for gene silencing. *Mol. Cell*. 28:823-837.

Jang, H.S., N.M. Shah, A.Y. Du, Z.Z. Dailey, E.C. Pehrsson, P.M. Godoy, D. Zhang, D. Li, X. Xing, S. Kim, D. O'Donnell, J.I. Gordon, and T. Wang. 2019. Transposable elements drive widespread expression of oncogenes in human cancers. *In* Nat. Genet. Vol. 51. Nature Publishing Group. 611-617.

Janiszewska, A.D., S. Poletajew, and A. Wasiutynski. 2013. Spontaneous regression of renal cell carcinoma. *Contemp Oncol (Pozn)*. 17:123-127.

Jia, M., Z. Li, M. Pan, M. Tao, J. Wang, and X. Lu. 2020. LINC-PINT Suppresses the Aggressiveness of Thyroid Cancer by Downregulating miR-767-5p to Induce TET2 Expression. *Mol Ther Nucleic Acids*. 22:319-328.

Jia, Y., J. Li, H. Wu, W. Wang, S. Sun, C. Feng, X. Liu, C. Li, Y. Zhang, Y. Cai, X. Wei, P. Yao, X. Liu, S. Zhang, and F. Wu. 2023. Comprehensive analysis of NT5DC family prognostic and immune significance in breast cancer. *Medicine*. 102:e32927-e32927.

Jin, X., X.E. Xu, Y.Z. Jiang, Y.R. Liu, W. Sun, Y.J. Guo, Y.X. Ren, W.J. Zuo, X. Hu, S.L. Huang, H.J. Shen, F. Lan, Y.F. He, G.H. Hu, G.H. Di, X.H. He, D.Q. Li, S. Liu, K.D. Yu, and Z.M. Shao. 2019. The endogenous retrovirus-derived long noncoding RNA TROJAN promotes triple-negative breast cancer progression via ZMYND8 degradation. *Science Advances*. 5.

Kazachenka, A., J.H. Loong, J. Attig, G.R. Young, P. Ganguli, G. Devonshire, N. Grehan, R.C. Fitzgerald, P.A.W. Edwards, N. Grehan, B. Nutzinger, E. Fidziukiewicz, A.M. Redmond, S. Abbas, A. Freeman, E.C. Smyth, M. O'Donovan, A. Miremadi, S. Malhotra, M. Tripathi, C. Cheah, H. Coles, C. Flint, M. Eldridge, M. Secrier, G. Devonshire, S. Jammula, J. Davies, C. Crichton, N. Carroll, R.H. Hardwick, P. Safranek, A. Hindmarsh, V. Sujendran, S.J. Hayes, Y. Ang, A. Sharrocks, S.R. Preston, I. Bagwan, V. Save, R.J.E. Skipworth, T.R. Hupp, J.R. O'Neill, O. Tucker, A. Beggs, P. Taniere, S. Puig, G. Contino, T.J. Underwood, R.C. Walker, B.L. Grace, J. Lagergren, J. Gossage, A. Davies, F. Chang, U. Mahadeva, V. Goh, F.D. Ciccarelli, G. Sanders, R. Berrisford, D. Chan, E. Cheong, B. Kumar, L. Sreedharan, S.L. Parsons, I. Soomro, P. Kaye, J. Saunders, L. Lovat, R. Haidry, M. Scott, S. Sothi, S. Lishman, G.B. Hanna, C.J. Peters, K. Moorthy, A. Grabowska, R. Turkington, D. McManus, H. Coleman, R.D. Petty, F. Bartlett, F.D. Ciccarelli, R.C. Fitzgerald, and G. Kassiotis. 2023. The transcriptional landscape of endogenous retroelements delineates esophageal adenocarcinoma subtypes. *NAR cancer*. 5.

Kazachenka, A., G.R. Young, J. Attig, C. Kordella, E. Lamprianidou, E. Zoulia, G. Vrachiolias, M. Papoutselis, E. Bernard, E. Papaemmanuil, I. Kotsianidis, and G. Kassiotis. 2019. Epigenetic therapy of myelodysplastic syndromes connects to cellular differentiation independently of endogenous retroelement derepression. *Genome Med.* 11:86.

Kent, W.J. 2002. BLAT - The BLAST-Like Alignment Tool. *Genome Res.* 12:656-664.

Keup, C., V. Suryaprakash, M. Storbeck, O. Hoffmann, R. Kimmig, and S. Kasimir‐bauer. 2021. Longitudinal multi‐parametric liquid biopsy approach identifies unique features of circulating tumor cell, extracellular vesicle, and cell‐free DNA characterization for disease monitoring in metastatic breast cancer patients. *Cells*. 10:1-22.

Kim, H.S., M.G. Kim, K.-W. Min, U.S. Jung, and D.-H. Kim. 2021. High MMP-11 expression associated with low CD8+ T cells decreases the survival rate in patients with breast cancer. *PLoS ONE*. 16:e0252052-e0252052.

Kong, Y., C.M. Rose, A.A. Cass, A.G. Williams, M. Darwish, S. Lianoglou, P.M. Haverty, A.J. Tong, C. Blanchette, M.L. Albert, I. Mellman, R. Bourgon, J. Greally, S. Jhunjhunwala, and H. Chen-Harris. 2019. Transposable element expression in tumors is associated with immune infiltration and increased antigenicity. *Nat. Commun.* 10.

Korekane, H., J.Y. Park, A. Matsumoto, K. Nakajima, S. Takamatsu, K. Ohtsubo, Y. Miyamoto, S. Hanashima, K. Kanekiyo, S. Kitazume, Y. Yamaguchi, I. Matsuo, and N. Taniguchi. 2013. Identification of ectonucleotide pyrophosphatase/phosphodiesterase 3 (ENPP3) as a regulator of N-acetylglucosaminyltransferase GnT-IX (GnT-Vb). *J. Biol. Chem.* 288:27912-27926.

Kretzmer, H., A. Biran, N. Purroy, C.K. Lemvigh, K. Clement, M. Gruber, H. Gu, L. Rassenti, A.W. Mohammad, C. Lesnick, S.L. Slager, E. Braggio, T.D. Shanafelt, N.E. Kay, S.M.

Fernandes, J.R. Brown, L. Wang, S. Li, K.J. Livak, D.S. Neuberg, S. Klages, B. Timmermann, T.J. Kipps, E. Campo, A. Gnirke, C.J. Wu, and A. Meissner. 2021. Preneoplastic Alterations Define CLL DNA Methylome and Persist through Disease Progression and Therapy. *Blood Cancer Discovery*. 2:54-69.

Krogh, A., B. Larsson, G. Von Heijne, and E.L.L. Sonnhammer. 2001. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J. Mol. Biol.* 305:567-580.

Ku, J., K. Lee, D. Ku, S. Kim, J. Lee, H. Bang, N. Kim, H. Do, H. Lee, C. Lim, J. Han, Y.S. Lee, and Y. Kim. 2024. Alternative polyadenylation determines the functional landscape of inverted Alu repeats. *Mol. Cell*.

Lamprecht, B., K. Walter, S. Kreher, R. Kumar, M. Hummel, D. Lenze, K. Köchert, M.A. Bouhlel, J. Richter, E. Soler, R. Stadhouders, K. Jöhrens, K.D. Wurster, D.F. Callen, M.F. Harte, M. Giefing, R. Barlow, H. Stein, I. Anagnostopoulos, M. Janz, P.N. Cockerill, R. Siebert, B. Dörken, C. Bonifer, and S. Mathas. 2010. Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. *Nat. Med.* 16:571-579.

Lanciano, S., and G. Cristofari. 2020. Measuring and interpreting transposable element expression. *Nat. Rev. Genet.* 21:721-736.

Lanciano, S., C. Philippe, A. Sarkar, D. Pratella, C. Domrane, A.J. Doucet, D. van Essen, S. Saccani, L. Ferry, P.A. Defossez, and G. Cristofari. 2024. Locus-level L1 DNA methylation profiling reveals the epigenetic and transcriptional interplay between L1s and their integration sites. *Cell Genom*. 4:100498.

Lander, E.S., L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J.P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, Y. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J.C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R.H. Waterston, R.K. Wilson, L.W. Hillier, J.D. McPherson, M.A. Marra, E.R. Mardis, L.A. Fulton, A.T. Chinwalla, K.H. Pepin, W.R. Gish, S.L. Chissoe, M.C. Wendl, K.D. Delehaunty, T.L. Miner, A. Delehaunty, J.B. Kramer, L.L. Cook, R.S. Fulton, D.L. Johnson, P.J. Minx, S.W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J.F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, et al. 2001. Initial sequencing and analysis of the human genome. *Nature*. 409:860-921.

Larouche, J.D., A. Trofimov, L. Hesnard, G. Ehx, Q. Zhao, K. Vincent, C. Durette, P. Gendron, J.P. Laverdure, É. Bonneil, C. Côté, S. Lemieux, P. Thibault, and C. Perreault. 2020. Widespread and tissue-specific expression of endogenous retroelements in human somatic tissues. *Genome Med.* 12:1-16.

Larson, M.H., W. Pan, H.J. Kim, R.E. Mauntz, S.M. Stuart, M. Pimentel, Y. Zhou, P. Knudsgaard, V. Demas, A.M. Aravanis, and A. Jamshidi. 2021. A comprehensive characterization of the cell-free transcriptome reveals tissue- and subtype-specific biomarkers for cancer detection. *Nat. Commun.* 12.

Laumont, C.M., K. Vincent, L. Hesnard, É. Audemard, É. Bonneil, J.P. Laverdure, P. Gendron, M. Courcelles, M.P. Hardy, C. Côté, C. Durette, C. St-Pierre, M. Benhammadi, J. Lanoix, S. Vobecky, E. Haddad, S. Lemieux, P. Thibault, and C. Perreault. 2018. Noncoding regions are the main source of targetable tumor-specific antigens. *Sci. Transl. Med.* 10.

Law, A.Y., and C.K. Wong. 2010. Stanniocalcin-2 is a HIF-1 target gene that promotes cell proliferation in hypoxia. *Exp. Cell Res.* 316:466-476.

Li, J., W. Zhang, X. Ma, Y. Wei, F. Zhou, J. Li, C. Zhang, and Z. Yang. 2023. Cuproptosis/ferroptosis-related gene signature is correlated with immune infiltration and predict the prognosis for patients with breast cancer. *Front. Pharmacol.* 14.

Li, W., J. Freudenberg, and P. Miramontes. 2014. Diminishing return for increased Mappability with longer sequencing reads: Implications of the k-mer distributions in the human genome. *BMC Bioinformatics*. 15.

Li, W., L. Lin, R. Malhotra, L. Yang, R. Acharya, and M. Poss. 2019. A computational framework to assess genome-wide distribution of polymorphic human endogenous retrovirus-K In human populations. *PLoS Comput. Biol.* 15:e1006564.

Li, W., C. Yang, J. Li, X. Li, and P. Zhou. 2022. MicroRNA-217 aggravates breast cancer through activation of NF1-mediated HSF1/ATG7 axis and c-Jun/ATF3/MMP13 axis. *Hum. Cell*. 36:377-392.

Li, X., X. Song, J. Ma, Y. Zhao, Q. Jiang, Z. Zhao, and M. Li. 2020a. FSIP1 is correlated with estrogen receptor status and poor prognosis. *Mol. Carcinog.* 59:126-135.

Li, Y., E. Salo-Mullen, A. Varghese, M. Trottier, Z.K. Stadler, and L. Zhang. 2020b. Insertion of an Alu-like element in MLH1 intron 7 as a novel cause of Lynch syndrome. *Mol Genet Genomic Med*. 8:e1523.

Liao, Y., G.K. Smyth, and W. Shi. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. 30:923-930.

Lidgren, A., Y. Hedberg, K. Grankvist, T. Rasmuson, J. Vasko, and B. Ljungberg. 2005. The expression of hypoxia-inducible factor 1α is a favorable independent prognostic factor in renal cell carcinoma. *Clin. Cancer Res.* 11:1129-1135.

Limsakul, P., P. Choochuen, G. Charupanit, and K. Charupanit. 2023. Transcriptomic Analysis of Subtype-Specific Tyrosine Kinases as Triple Negative Breast Cancer Biomarkers. *Cancers*. 15:403-403.

Linsley, P.S., C. Speake, E. Whalen, and D. Chaussabel. 2014. Copy number loss of the interferon gene cluster in melanomas is linked to reduced T cell infiltrate and poor patient prognosis. *PLoS ONE*. 9:e109760.

Liu, B., and C. Wei. 2021. Hypoxia Induces Overexpression of CCL28 to Recruit Treg Cells to Enhance Angiogenesis in Lung Adenocarcinoma.

Liu, J., J. Li, P. Li, Y. Jiang, H. Chen, R. Wang, F. Cao, and P. Liu. 2019. DLG5 suppresses breast cancer stem cell‐like characteristics to restore tamoxifen sensitivity by inhibiting TAZ expression. *J. Cell. Mol. Med.* 23:512-521.

Liu, L., C. Yang, J. Shen, L. Huang, W. Lin, H. Tang, W. Liang, W. Shao, H. Zhang, and J. He. 2016. GABRA3 promotes lymphatic metastasis in lung adenocarcinoma by mediating upregulation of matrix metalloproteinases. *Oncotarget*. 7:32341-32350.

Liu, N., C.H. Lee, T. Swigut, E. Grow, B. Gu, M.C. Bassik, and J. Wysocka. 2018. Selective silencing of euchromatic L1s revealed by genome-wide screens for L1 regulators. *Nature*. 553:228-232.

Liu, Y., F. Guo, M. Dai, D. Wang, Y. Tong, J. Huang, J. Hu, and G. Li. 2009. Gammaaminobutyric acid A receptor alpha 3 subunit is overexpressed in lung cancer. *Pathol Oncol Res*. 15:351-358.

Liu, Y., Y.H. Li, F.J. Guo, J.J. Wang, R.L. Sun, J.Y. Hu, and G.C. Li. 2008. Gamma-aminobutyric acid promotes human hepatocellular carcinoma growth through overexpressed gamma-aminobutyric acid A receptor alpha 3 subunit. *World J Gastroenterol*. 14:7175-7182.

Lombardi, O., R. Li, S. Halim, H. Choudhry, P.J. Ratcliffe, and D.R. Mole. 2022. Pan-cancer analysis of tissue and single-cell HIF-pathway activation using a conserved gene signature. *Cell Rep*. 41.

Long, M., M. Zhan, S. Xu, R. Yang, W. Chen, S. Zhang, Y. Shi, Q. He, M. Mohan, Q. Liu, and J. Wang. 2017. miR-92b-3p acts as a tumor suppressor by targeting Gabra3 in pancreatic cancer. *Mol. Cancer*. 16:167.

Lopes, I., G. Altab, P. Raina, and J.P. de Magalhaes. 2021. Gene Size Matters: An Analysis of Gene Length in the Human Genome. *Front Genet*. 12:559998.

Loriot, A., A. Van Tongelen, J. Blanco, S. Klaessens, J. Cannuyer, N. van Baren, A. Decottignies, and C. De Smet. 2014. A novel cancer-germline transcript carrying pro-metastatic miR-105 and TET-targeting miR-767 induced by DNA hypomethylation in tumors. *Epigenetics*. 9:1163-1171.

Lower, R., J. Lower, C. Tondera-Koch, and R. Kurth. 1993. A general method for the identification of transcribed retrovirus sequences (R-U5 PCR) reveals the expression of the human endogenous retrovirus loci HERV-H and HERV-K in teratocarcinoma cells. *Virology*. 192:501-511.

Lower, R., R.R. Tonjes, C. Korbmacher, R. Kurth, and J. Lower. 1995. Identification of a Rev-related protein by analysis of spliced transcripts of the human endogenous retroviruses HTDV/HERV-K. *J. Virol.* 69:141-149.

Lu, S., J. Zhang, X. Lian, L. Sun, K. Meng, Y. Chen, Z. Sun, X. Yin, Y. Li, J. Zhao, T. Wang, G. Zhang, and Q.-Y. He. 2019. A hidden human proteome encoded by 'non-coding' genes. *Nucleic Acids Res.* 47:8111-8125.

Mackenzie, K.J., P. Carroll, C.A. Martin, O. Murina, A. Fluteau, D.J. Simpson, N. Olova, H. Sutcliffe, J.K. Rainger, A. Leitch, R.T. Osborn, A.P. Wheeler, M. Nowotny, N. Gilbert, T. Chandra, M.A.M. Reijns, and A.P. Jackson. 2017. cGAS surveillance of micronuclei links genome instability to innate immunity. *Nature*. 548:461-465.

Magin, C., R. Lower, and J. Lower. 1999. cORF and RcRE, the Rev/Rex and RRE/RxRE homologues of the human endogenous retrovirus family HTDV/HERV-K. *J. Virol.* 73:9496-9507.

Mao, J., Q. Zhang, and Y.S. Cong. 2021. Human endogenous retroviruses in development and disease. *Comput Struct Biotechnol J*. 19:5978-5986.

Maranchie, J.K., J.R. Vasselli, J. Riss, J.S. Bonifacino, W.M. Linehan, and R.D. Klausner. 2002. The contribution of VHL substrate binding and HIF1-alpha to the phenotype of VHL loss in renal cell carcinoma. *Cancer Cell*. 1:247-255.

Marasca, F., S. Sinha, R. Vadalà, B. Polimeni, V. Ranzani, E.M. Paraboschi, F.V. Burattin, M. Ghilotti, M. Crosti, M.L. Negri, S. Campagnoli, S. Notarbartolo, A. Sartore-Bianchi, S. Siena, D. Prati, G. Montini, G. Viale, O. Torre, S. Harari, R. Grifantini, G. Soldà, S. Biffo, S. Abrignani, and B. Bodega. 2022. LINE1 are spliced in non-canonical transcript variants to regulate T cell quiescence and exhaustion. *Nat. Genet.* 54:180-193.

Mardjuki, R., S. Wang, J.A. Carozza, G.C. Abhiraman, X. Lyu, and L. Li. 2024. Identification of extracellular membrane protein ENPP3 as a major cGAMP hydrolase, cementing cGAMP's role as an immunotransmitter. *bioRxiv*.

Margolin, J.F., J.R. Friedman, W.K. Meyer, H. Vissing, H.J. Thiesen, and F.J. Rauscher, 3rd. 1994. Kruppel-associated boxes are potent transcriptional repression domains. *Proc Natl Acad Sci U S A*. 91:4509-4513.

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

Martin, S.L. 2006. The ORF1 protein encoded by LINE-1: structure and function during L1 retrotransposition. *J Biomed Biotechnol*. 2006:45621.

Mathe, A., M. Wong-Brown, B. Morten, J.F. Forbes, S.G. Braye, K.A. Avery-Kiejda, and R.J. Scott. 2015. Novel genes associated with lymph node metastasis in triple negative breast cancer. *Sci. Rep.* 5:15832-15832.

Max, K.E.A., K. Bertram, K.M. Akat, K.A. Bogardus, J. Li, P. Morozov, I.Z. Ben-Dov, X. Li, Z.R. Weiss, A. Azizian, A. Sopeyin, T.G. Diacovo, C. Adamidi, Z. Williams, and T. Tuschl. 2018. Human plasma and serum extracellular small RNA reference profiles and their clinical utility. *Proc. Natl. Acad. Sci. U. S. A.* 115:E5334-E5343.

Maxwell, P.H., M.S. Wiesener, G.W. Chang, S.C. Clifford, E.C. Vaux, M.E. Cockman, C.C. Wykoff, C.W. Pugh, E.R. Maher, and P.J. Ratcliffe. 1999. The tumour suppressor protein VHL targets hypoxia-inducible factors for oxygen-dependent proteolysis. *Nature*. 399:271-275.

Mayer, J., J. Blomberg, and R.L. Seal. 2011. A revised nomenclature for transcribed human endogenous retroviral loci. *Mobile DNA*. 2:1-8.

McDermott, D.F., M.A. Huseni, M.B. Atkins, R.J. Motzer, B.I. Rini, B. Escudier, L. Fong, R.W. Joseph, S.K. Pal, J.A. Reeves, M. Sznol, J. Hainsworth, W.K. Rathmell, W.M. Stadler, T. Hutson, M.E. Gore, A. Ravaud, S. Bracarda, C. Suarez, R. Danielli, V. Gruenwald, T.K. Choueiri, D. Nickles, S. Jhunjhunwala, E. Piault-Louis, A. Thobhani, J. Qiu, D.S. Chen, P.S. Hegde, C. Schiff, G.D. Fine, and T. Powles. 2018. Clinical activity and molecular correlates of response to atezolizumab alone or in combination with bevacizumab versus sunitinib in renal cell carcinoma. *Nat. Med.* 24:749-757.

Mehdipour, P., S.A. Marhon, I. Ettayebi, A. Chakravarthy, A. Hosseini, Y. Wang, F.A. de Castro, H. Loo Yau, C. Ishak, S. Abelson, C.A. O'Brien, and D.D. De Carvalho. 2020. Epigenetic therapy induces transcription of inverted SINEs and ADAR1 dependency. *Nature*:1-5.

Memon, D., K. Dawson, C.S.F. Smowton, W. Xing, C. Dive, and C.J. Miller. 2016. Hypoxia-driven splicing into noncoding isoforms regulates the DNA damage response. *npj Genomic Medicine*. 1.

Merlotti, A., B. Sadacca, Y.A. Arribas, M. Ngoma, M. Burbage, C. Goudot, A. Houy, A. Rocañín-Arjó, A. Lalanne, A. Seguin-Givelet, M. Lefevre, S. Heurtebise-Chrétien, B. Baudon, G. Oliveira, D. Loew, M. Carrascal, C.J. Wu, O. Lantz, M.-H. Stern, N. Girard, J.J. Waterfall,

and S. Amigorena. 2023. Noncanonical splicing junctions between exons and transposable elements represent a source of immunogenic recurrent neo-antigens in patients with lung cancer.

Mi, S., X. Lee, X.-p. Li, G.M. Veldman, H. Finnerty, L. Racie, E. LaVallie, X.Y. Tang, P. Edouard, S. Howes, J.C. Keith, and J.M. McCoy. 2000. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature*. 403:785-789.

Mithraprabhu, S., R. Morley, T. Khong, A. Kalff, K. Bergin, J. Hocking, I. Savvidou, K.M. Bowen, M. Ramachandran, K. Choi, B.K.L. Wong, J. Reynolds, and A. Spencer. 2019. Monitoring tumour burden and therapeutic response through analysis of circulating tumour DNA and extracellular RNA in multiple myeloma patients. *Leukemia*. 33:2022-2033.

Mommert, M., O. Tabone, G. Oriol, E. Cerrato, A. Guichard, M. Naville, P. Fournier, J.N. Volff, A. Pachot, G. Monneret, F. Venet, K. Brengel-Pesce, J. Textoris, and F. Mallet. 2018. LTR-retrotransposon transcriptome modulation in response to endotoxin-induced stress in PBMCs. *BMC Genomics*. 19:522.

Montagner, M., E. Enzo, M. Forcato, F. Zanconato, A. Parenti, E. Rampazzo, G. Basso, G. Leo, A. Rosato, S. Bicciato, M. Cordenonsi, and S. Piccolo. 2012. SHARP1 suppresses breast cancer metastasis by promoting degradation of hypoxia-inducible factors. *Nature*. 487:380-384.

Motzer, R.J., B.I. Rini, D.F. McDermott, B.G. Redman, T.M. Kuzel, M.R. Harrison, U.N. Vaishampayan, H.A. Drabkin, S. George, T.F. Logan, K.A. Margolin, E.R. Plimack, A.M. Lambert, I.M. Waxman, and H.J. Hammers. 2015. Nivolumab for Metastatic Renal Cell Carcinoma: Results of a Randomized Phase II Trial. *J. Clin. Oncol.* 33:1430-1437.

Motzer, R.J., P.B. Robbins, T. Powles, L. Albiges, J.B. Haanen, J. Larkin, X.J. Mu, K.A. Ching, M. Uemura, S.K. Pal, B. Alekseev, G. Gravis, M.T. Campbell, K. Penkov, J.L. Lee, S. Hariharan, X. Wang, W. Zhang, J. Wang, A. Chudnovsky, A. di Pietro, A.C. Donahue, and T.K. Choueiri. 2020. Avelumab plus axitinib versus sunitinib in advanced renal cell carcinoma: biomarker analysis of the phase 3 JAVELIN Renal 101 trial. *Nat. Med.* 26:1733-1741.

Mugoni, V., Y. Ciani, C. Nardella, and F. Demichelis. 2022. Circulating RNAs in prostate cancer patients. *Cancer Lett.* 524:57-69.

Murillo, O.D., W. Thistlethwaite, J. Rozowsky, S.L. Subramanian, R. Lucero, N. Shah, A.R. Jackson, S. Srinivasan, A. Chung, C.D. Laurent, R.R. Kitchen, T. Galeev, J. Warrell, J.A. Diao, J.A. Welsh, K. Hanspers, A. Riutta, S. Burgstaller-Muehlbacher, R.V. Shah, A. Yeri, L.M. Jenkins, M.E. Ahsen, C. Cordon-Cardo, N. Dogra, S.M. Gifford, J.T. Smith, G. Stolovitzky, A.K. Tewari, B.H. Wunsch, K.K. Yadav, K.M. Danielson, J. Filant, C. Moeller, P. Nejad, A. Paul, B. Simonson, D.K.T.W. Wong, X. Zhang, L. Balaj, R. Gandhi, A.K. Sood, R.P. Alexander, L. Wang, C. Wu, D.K.T.W. Wong, D.J. Galas, K. Van Keuren-Jensen, T. Patel, J.C. Jones, S. Das, K.H. Cheung, A.R. Pico, A.I. Su, R.L. Raffai, L.C. Laurent, M.E. Roth, M.B. Gerstein, and A. Milosavljevic. 2019. exRNA Atlas Analysis Reveals Distinct Extracellular RNA Cargo Types and Their Carriers Present across Human Biofluids. *Cell*. 177:463-477.e415.

Neph, S., M.S. Kuehn, A.P. Reynolds, E. Haugen, R.E. Thurman, A.K. Johnson, E. Rynes, M.T. Maurano, J. Vierstra, S. Thomas, R. Sandstrom, R. Humbert, and J.A. Stamatoyannopoulos. 2012. BEDOPS: high-performance genomic feature operations. *Bioinformatics*. 28:1919-1920.

Ng, K.W., J. Attig, G.R. Young, E. Ottina, S.I. Papamichos, I. Kotsianidis, and G. Kassiotis. 2019. Soluble PD-L1 generated by endogenous retroelement exaptation is a receptor antagonist. *Elife*. 8.

Nichols, J., B. Zevnik, K. Anastassiadis, H. Niwa, D. Klewe-Nebenius, I. Chambers, H. Scholer, and A. Smith. 1998. Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. *Cell*. 95:379-391.

Nikitina, A.S., V.V. Babenko, K.A. Babalyan, A.O. Vasiliev, A.V. Govorov, E.A. Prilepskaya, S.A. Danilenko, O.V. Selezneva, and E.I. Sharova. 2016. Primary screening of candidate RNA biomarkers for diagnostics of prostate cancer. *Biochemistry (Moscow) Supplement Series B: Biomedical Chemistry*. 10:180-183.

O'Neill, C.P., K.E. Gilligan, and R.M. Dwyer. 2019. Role of extracellular vesicles (EVs) in cell stress response and resistance to cancer therapy. *Cancers*. 11.

Ohtani, H., M. Liu, W. Zhou, G. Liang, and P.A. Jones. 2018. Switching roles for DNA and histone methylation depend on evolutionary ages of human endogenous retroviruses. *Genome Res.* 28:1147-1157.

Orioli, A., C. Pascali, A. Pagano, M. Teichmann, and G. Dieci. 2012. RNA polymerase III transcription control elements: themes and variations. *Gene*. 493:185-194.

Ostertag, E.M., J.L. Goodier, Y. Zhang, and H.H. Kazazian, Jr. 2003. SVA elements are nonautonomous retrotransposons that cause disease in humans. *Am. J. Hum. Genet.* 73:1444-1451.

Ouspenskaia, T., T. Law, K.R. Clauser, S. Klaeger, S. Sarkizova, F. Aguet, B. Li, E. Christian, B.A. Knisbacher, P.M. Le, C.R. Hartigan, H. Keshishian, A. Apffel, G. Oliveira, W. Zhang, Y.T. Chow, Z. Ji, S.A. Shukla, P. Bachireddy, G. Getz, N. Hacohen, D.B. Keskin, S.A. Carr, C.J. Wu, and A. Regev. 2020. Thousands of novel unannotated proteins expand the MHC I immunopeptidome in cancer. *bioRxiv*:2020.2002.2012.945840-942020.945802.945812.945840.

Panda, A., A.A. de Cubas, M. Stein, G. Riedlinger, J. Kra, T. Mayer, C.C. Smith, B.G. Vincent, J.S. Serody, K.E. Beckermann, S. Ganesan, G. Bhanot, and W.K. Rathmell. 2018. Endogenous retrovirus expression is associated with response to immune checkpoint blockade in clear cell renal cell carcinoma. *JCI insight*. 3.

Papamichos, S.I. 2021. Endogenous Retroelement-Driven Expression of OCT4B mRNA Variants. *Stem Cell Reviews and Reports*.

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

Payer, L.M., J.P. Steranka, M.S. Kryatova, G. Grillo, M. Lupien, P.P. Rocha, and K.H. Burns. 2021. Alu insertion variants alter gene transcript levels. *Genome Res.*

Price, E., O. Gianfrancesco, P.T. Harrison, B. Frank, V.J. Bubb, and J.P. Quinn. 2021. CRISPR Deletion of a SVA Retrotransposon Demonstrates Function as a cis-Regulatory Element at the TRPV1/TRPV3 Intergenic Region. *Int. J. Mol. Sci.* 22:1911-1911.

Qin, Y., J. Yao, D.C. Wu, R.M. Nottingham, S. Mohr, S. Hunicke-Smith, and A.M. Lambowitz. 2016. High-throughput sequencing of human plasma RNA by using thermostable group II intron reverse transcriptases. *RNA*. 22:111-128.

Quinlan, A.R., and I.M. Hall. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 26:841-842.

Quinn, J.P., and V.J. Bubb. 2014. SVA retrotransposons as modulators of gene expression. *Mob Genet Elements*. 4:e32102.

Raviram, R., P.P. Rocha, V.M. Luo, E. Swanzey, E.R. Miraldi, E.B. Chuong, C. Feschotte, R. Bonneau, and J.A. Skok. 2018. Analysis of 3D genomic interactions identifies candidate host genes that transposable elements potentially regulate. *Genome Biol.* 19.

Reggiardo, R.E., S.V. Maroli, V. Peddu, A.E. Davidson, A. Hill, E. LaMontagne, Y.A. Aaraj, M. Jain, S.Y. Chan, and D.H. Kim. 2023. Profiling of repetitive RNA sequences in the blood plasma of patients with cancer. *Nature biomedical engineering*.

Ren, L., Y. Yu, L. Wang, Z. Zhu, R. Lu, and Z. Yao. 2016. Hypoxia-induced CCL28 promotes recruitment of regulatory T cells and tumor growth in liver cancer. *Oncotarget*. 7.

Renaudineau, Y., S. Hillion, A. Saraux, R.A. Mageed, and P. Youinou. 2005. An alternative exon 1 of the CD5 gene regulates CD5 expression in human B lymphocytes. *Blood*. 106:2781-2789.

Rezaei, S.D., J.A. Hayward, S. Norden, J. Pedersen, J. Mills, A.C. Hearps, G. Tachedjian, R. Sd, H. Ja, N. S, P. J, M. J, H. Ac, and T. G. 2021. HERV-K Gag RNA and Protein Levels Are Elevated in Malignant Regions of the Prostate in Males with Prostate Cancer. *Viruses*. 13:1-12.

Ricketts, C.J., A.A. De Cubas, H. Fan, C.C. Smith, M. Lang, E. Reznik, R. Bowlby, E.A. Gibb, R. Akbani, R. Beroukhim, D.P. Bottaro, T.K. Choueiri, R.A. Gibbs, A.K. Godwin, S. Haake, A.A. Hakimi, E.P. Henske, J.J. Hsieh, T.H. Ho, R.S. Kanchi, B. Krishnan, D.J. Kwiatkowski, W. Lui, M.J. Merino, G.B. Mills, J. Myers, M.L. Nickerson, V.E. Reuter, L.S. Schmidt, C.S. Shelley, H. Shen, B. Shuch, S. Signoretti, R. Srinivasan, P. Tamboli, G. Thomas, B.G. Vincent, C.D. Vocke, D.A. Wheeler, L. Yang, W.Y. Kim, A.G. Robertson, N. Cancer Genome Atlas Research, P.T. Spellman, W.K. Rathmell, and W.M. Linehan. 2018. The Cancer Genome Atlas Comprehensive Molecular Characterization of Renal Cell Carcinoma. *Cell Rep.* 23:3698.

Robinson, J.T., H. Thorvaldsdóttir, W. Winckler, M. Guttman, E.S. Lander, G. Getz, and J.P. Mesirov. 2011. Integrative genomics viewer. *Nat. Biotechnol.* 29:24-26.

Rodriguez-Martin, B., E.G. Alvarez, A. Baez-Ortega, J. Zamora, F. Supek, J. Demeulemeester, M. Santamarina, Y.S. Ju, J. Temes, D. Garcia-Souto, H. Detering, Y. Li, J. Rodriguez-Castro, A. Dueso-Barroso, A.L. Bruzos, S.C. Dentro, M.G. Blanco, G. Contino, D. Ardeljan, M. Tojo, N.D. Roberts, S. Zumalave, P.A.W. Edwards, J. Weischenfeldt, M. Puiggròs, Z. Chong, K. Chen, E.A. Lee, J.A. Wala, K. Raine, A. Butler, S.M. Waszak, F.C.P. Navarro, S.E. Schumacher, J. Monlong, F. Maura, N. Bolli, G. Bourque, M. Gerstein, P.J. Park, D.C. Wedge, R. Beroukhim, D. Torrents, J.O. Korbel, I. Martincorena, R.C. Fitzgerald, P. Van Loo, H.H. Kazazian, K.H. Burns, K.C. Akdemir, E.G. Alvarez, A. Baez-Ortega, R. Beroukhim, P.C. Boutros, D.D.L. Bowtell, B. Brors, K.H. Burns, P.J. Campbell, K. Chan, K. Chen, I. Cortés-Ciriano, A. Dueso-Barroso, A.J. Dunford, P.A. Edwards, X. Estivill, D. Etemadmoghadam, L. Feuerbach, J.L. Fink, M. Frenkel-Morgenstern, D.W. Garsed, M. Gerstein, D.A. Gordenin, D. Haan, J.E. Haber, J.M. Hess, B. Hutter, M. Imielinski, D.T.W. Jones, Y.S. Ju, M.D. Kazanov, L.J. Klimczak, Y. Koh, J.O. Korbel, K. Kumar, E.A. Lee, J.J.K. Lee, Y. Li, A.G. Lynch, G. Macintyre, F. Markowetz, I. Martincorena, A. Martinez-Fundichely, M. Meyerson, S. Miyano, H. Nakagawa, F.C.P. Navarro, S. Ossowski, P.J. Park, J.V. Pearson, M. Puiggròs, et al. 2020. Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nat. Genet.* 52:306-319.

Rooney, M.S., S.A. Shukla, C.J. Wu, G. Getz, and N. Hacohen. 2015. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell*. 160:48-61.

Sakurai, M., Y. Shiromoto, H. Ota, C. Song, A.V. Kossenkov, J. Wickramasinghe, L.C. Showe, E. Skordalakes, H.Y. Tang, D.W. Speicher, and K. Nishikura. 2017. ADAR1 controls apoptosis of stressed cells by inhibiting Staufen1-mediated mRNA decay. *Nat. Struct. Mol. Biol.* 24:534-543.

Salama, R., N. Masson, P. Simpson, L.K. Sciesielski, M. Sun, Y.M. Tian, P.J. Ratcliffe, and D.R. Mole. 2015. Heterogeneous effects of direct hypoxia pathway activation in kidney cancer. *PLoS ONE*. 10:1-19.

Samanta, D., N.R. Prabhakar, and G.L. Semenza. 2017. Systems biology of oxygen homeostasis. *Wiley Interdiscip Rev Syst Biol Med*. 9.

Savelyeva, A.V., E.V. Kuligina, D.N. Bariakin, V.V. Kozlov, E.I. Ryabchikova, V.A. Richter, and D.V. Semenov. 2017. Variety of RNAs in Peripheral Blood Cells, Plasma, and Plasma Fractions. *BioMed Research International*. 2017.

Schenk, U., M. Frascoli, M. Proietti, R. Geffers, E. Traggiai, J. Buer, C. Ricordi, A.M. Westendorf, and F. Grassi. 2011. ATP inhibits the generation and function of regulatory T cells through the activation of purinergic P2X receptors. *Sci. Signal.* 4:ra12.

Schulz, L., M. Torres-Diz, M. Cortés-López, K.E. Hayer, M. Asnani, S.K. Tasian, Y. Barash, E. Sotillo, K. Zarnack, J. König, and A. Thomas-Tikhonenko. 2021. Direct long-read RNA sequencing identifies a subset of questionable exitrons likely arising from reverse transcription artifacts. *Genome Biol.* 22:1-12.

Scott, E.C., E.J. Gardner, A. Masood, N.T. Chuang, P.M. Vertino, and S.E. Devine. 2016. A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. *Genome Res.* 26:745-755.

Sela, N., B. Mersch, A. Hotz-Wagenblatt, and G. Ast. 2010. Characteristics of transposable element exonization within human and mouse. *PLoS ONE*. 5:10907-10907.

Sena, J.A., L. Wang, L.E. Heasley, and C.J. Hu. 2014a. Hypoxia regulates alternative splicing of HIF and non-HIF target genes. *Molecular Cancer Research*. 12:1233-1243.

Sena, J.A., L. Wang, M.R. Pawlus, and C.J. Hu. 2014b. HIFs enhance the transcriptional activation and splicing of adrenomedullin. *Molecular Cancer Research*. 12:728-741.

Shah, N.M., H.J. Jang, Y. Liang, J.H. Maeng, S.-C. Tzeng, A. Wu, N.L. Basri, X. Qu, C. Fan, A. Li, B. Katz, D. Li, X. Xing, B.S. Evans, and T. Wang. 2023. Pan-cancer analysis identifies tumor-specific antigens derived from transposable elements. *Nat. Genet.* 55:631-639.

Shenoy, N. 2020. HIF1α is not a target of 14q deletion in clear cell renal cancer. *Sci. Rep.* 10:1-12.

Shi, A., G.G. Kasumova, W.A. Michaud, J. Cintolo-Gonzalez, M. Díaz-Martínez, J. Ohmura, A. Mehta, I. Chien, D.T. Frederick, S. Cohen, D. Plana, D. Johnson, K.T. Flaherty, R.J. Sullivan, M. Kellis, and G.M. Boland. 2020. Plasma-derived extracellular vesicle analysis

and deconvolution enable prediction and tracking of melanoma checkpoint blockade outcome. *Science Advances*. 6:1-13.

Shi, W., Z. Chen, H. Liu, C. Miao, R. Feng, G. Wang, G. Chen, Z. Chen, P. Fan, W. Pang, and C. Li. 2022. COL11A1 as an novel biomarker for breast cancer with machine learning and immunohistochemistry validation. *Front. Immunol.* 13.

Siebenthall, K.T., C.P. Miller, J.D. Vierstra, J. Mathieu, M. Tretiakova, A. Reynolds, R. Sandstrom, E. Rynes, E. Haugen, A. Johnson, J. Nelson, D. Bates, M. Diegel, D. Dunn, M. Frerker, M. Buckley, R. Kaul, Y. Zheng, J. Himmelfarb, H. Ruohola-Baker, and S. Akilesh. 2019. Integrated epigenomic profiling reveals endogenous retrovirus reactivation in renal cell carcinoma. *EBioMedicine*. 41:427-442.

Simonti, C.N., M. Pavličev, and J.A. Capra. 2017. Transposable element exaptation into regulatory regions is rare, influenced by evolutionary age, and subject to pleiotropic constraints. *Mol. Biol. Evol.* 34:2856-2869.

Singh, M., H. Cai, M. Bunse, C. Feschotte, and Z. Izsvák. 2020. Human Endogenous Retrovirus K Rec Forms a Regulatory Loop with MITF that Opposes the Progression of Melanoma to an Invasive Stage. *Viruses*. 12:1303-1303.

Siravegna, G., S. Marsoni, S. Siena, and A. Bardelli. 2017. Integrating liquid biopsies into the management of cancer. *In* Nat Rev Clin Oncol. Vol. 14. Nature Publishing Group. 531-548.

Skog, J., T. Würdinger, S. van Rijn, D.H. Meijer, L. Gainche, W.T. Curry, B.S. Carter, A.M. Krichevsky, and X.O. Breakefield. 2008. Glioblastoma microvesicles transport RNA and proteins that promote tumour growth and provide diagnostic biomarkers. *Nat. Cell Biol.* 10:1470-1476.

Smith, C.C., K.E. Beckermann, D.S. Bortone, A.A. Cubas, L.M. Bixby, S.J. Lee, A. Panda, S. Ganesan, G. Bhanot, E.M. Wallen, M.I. Milowsky, W.Y. Kim, K. Rathmell, R. Swanstrom, J.S. Parker, J.S. Serody, S.R. Selitsky, and B.G. Vincent. 2018. Endogenous retroviral signatures predict immunotherapy response in clear cell renal cell carcinoma. *J. Clin. Invest.* 128:4804-4820.

Smith, C.C., S.R. Selitsky, S. Chai, P.M. Armistead, B.G. Vincent, and J.S. Serody. 2019. Alternative tumour-specific antigens. *Nat. Rev. Cancer*. 19:465-478.

Smythies, J.A., M. Sun, N. Masson, R. Salama, P.D. Simpson, E. Murray, V. Neumann, M.E. Cockman, H. Choudhry, P.J. Ratcliffe, and D.R. Mole. 2019. Inherent DNA‐binding specificities of the HIF‐1α and HIF‐2α transcription factors in chromatin. *EMBO Rep.* 20.

Snow, R.M., and P.F. Schellhammer. 1982. Spontaneous regression of metastatic renal cell carcinoma. *Urology*. 20:177-181.

Solassol, J., M. Larrieux, J. Leclerc, V. Ducros, C. Corsini, J. Chiesa, P. Pujol, and J.M. Rey. 2019. Alu element insertion in the MLH1 exon 6 coding sequence as a mutation predisposing to Lynch syndrome. *Hum. Mutat.* 40:716-720.

Sproviero, D., S. Gagliardi, S. Zucca, M. Arigoni, M. Giannini, M. Garofalo, M. Olivero, M. Dell'orco, O. Pansarasa, S. Bernuzzi, M. Avenali, M.C. Ramusino, L. Diamanti, B. Minafra, G. Perini, R. Zangaglia, A. Costa, M. Ceroni, N.I. Perrone‐bizzozero, R.A. Calogero, and C. Cereda. 2021. Different mirna profiles in plasma derived small and large extracellular vesicles from patients with neurodegenerative diseases. *Int. J. Mol. Sci.* 22:1-17.

Stacey, S.N., B. Kehr, J. Gudmundsson, F. Zink, A. Jonasdottir, S.A. Gudjonsson, A. Sigurdsson, B.V. Halldorsson, B.A. Agnarsson, K.R. Benediktsdottir, K.K. Aben, S.H. Vermeulen, R.G. Cremers, A. Panadero, B.T. Helfand, P.R. Cooper, J.L. Donovan, F.C. Hamdy, V. Jinga, I. Okamoto, J.G. Jonasson, L. Tryggvadottir, H. Johannsdottir, A.M. Kristinsdottir, G. Masson, O.T. Magnusson, P.D. Iordache, A. Helgason, H. Helgason, P. Sulem, D.F. Gudbjartsson, A. Kong, E. Jonsson, R.B. Barkardottir, G.V. Einarsson, T. Rafnar, U. Thorsteinsdottir, I.N. Mates, D.E. Neal, W.J. Catalona, J.I. Mayordomo, L.A. Kiemeney, G. Thorleifsson, and K. Stefansson. 2016. Insertion of an SVA-E retrotransposon into the CASP8 gene is associated with protection against prostate cancer. *Hum. Mol. Genet.* 25:1008-1018.

Storvall, H., D. Ramsköld, and R. Sandberg. 2013. Efficient and Comprehensive Representation of Uniqueness for Next-Generation Sequencing by Minimum Unique Length Analyses. *PLoS ONE*. 8.

Sun, L., J. Wu, F. Du, X. Chen, and Z.J. Chen. 2013. Cyclic GMP-AMP synthase is a cytosolic DNA sensor that activates the type I interferon pathway. *Science*. 339:786-791.

Takahashi, Y., N. Harashima, S. Kajigaya, H. Yokoyama, E. Cherkasova, J.P. McCoy, K. Hanada, O. Mena, R. Kurlander, A. Tawab, R. Srinivasan, A. Lundqvist, E. Malinzak, N. Geller, M.I. Lerman, and R.W. Childs. 2008. Regression of human kidney cancer following allogeneic stem cell transplantation is associated with recognition of an HERV-E antigen by T cells. *J. Clin. Invest.* 118:1099-1109.

Tange, O. 2023. GNU Parallel 20230722 ('Пригóжин'). Zenodo.

Tavakolian, S., H. Goudarzi, and E. Faghihloo. 2019. Evaluating the expression level of HERV-K env, np9, rec and gag in breast tissue. *Infectious Agents and Cancer*. 14.

Thomas, C.A., L. Tejwani, C.A. Trujillo, P.D. Negraes, R.H. Herai, P. Mesci, A. Macia, Y.J. Crow, and A.R. Muotri. 2017. Modeling of TREX1-Dependent Autoimmune Disease using Human Stem Cells Highlights L1 Accumulation as a Source of Neuroinflammation. *Cell Stem Cell*. 21:319-331 e318.

Thompson, J.A., R.J. Motzer, A.M. Molina, T.K. Choueiri, E.I. Heath, B.G. Redman, R.S. Sangha, D.S. Ernst, R. Pili, S.K. Kim, L. Reyno, A. Wiseman, F. Trave, B. Anand, K. Morrison, F. Doñate, and C.K. Kollmannsberger. 2018. Phase I trials of anti-ENPP3 antibody–drug conjugates in advanced refractory renal cell carcinomas. *Clin. Cancer Res.* 24:4399-4406.

Thorsson, V., D.L. Gibbs, S.D. Brown, D. Wolf, D.S. Bortone, T.H. Ou Yang, E. Porta-Pardo, G.F. Gao, C.L. Plaisier, J.A. Eddy, E. Ziv, A.C. Culhane, E.O. Paull, I.K.A. Sivakumar, A.J. Gentles, R. Malhotra, F. Farshidfar, A. Colaprico, J.S. Parker, L.E. Mose, N.S. Vo, J. Liu, Y. Liu, J. Rader, V. Dhankani, S.M. Reynolds, R. Bowlby, A. Califano, A.D. Cherniack, D. Anastassiou, D. Bedognetti, A. Rao, K. Chen, A. Krasnitz, H. Hu, T.M. Malta, H. Noushmehr, C.S. Pedamallu, S. Bullman, A.I. Ojesina, A. Lamb, W. Zhou, H. Shen, T.K. Choueiri, J.N. Weinstein, J. Guinney, J. Saltz, R. Holt, C.E. Rabkin, S.J. Caesar-Johnson, J.A. Demchok, I. Felau, M. Kasapi, M.L. Ferguson, C.M. Hutter, H.J. Sofia, R. Tarnuzzer, Z. Wang, L. Yang, J.C. Zenklusen, J. Zhang, S. Chudamani, J. Liu, L. Lolla, R. Naresh, T. Pihl, Q. Sun, Y. Wan, Y. Wu, J. Cho, T. DeFreitas, S. Frazer, N. Gehlenborg, G. Getz, D.I. Heiman, J. Kim, M.S. Lawrence, P. Lin, S. Meier, M.S. Noble, G. Saksena, D. Voet, H. Zhang, B. Bernard, N. Chambwe, V. Dhankani, T. Knijnenburg, R. Kramer, K. Leinonen, Y. Liu, M. Miller, S. Reynolds, I. Shmulevich, V. Thorsson, W. Zhang, R. Akbani, B.M. Broom, A.M. Hegde, Z. Ju, R.S. Kanchi, et al. 2018. The Immune Landscape of Cancer. *Immunity*. 48:812-830.e814.

Tian, L., L. Yang, W. Zheng, Y. Hu, P. Ding, Z. Wang, D. Zheng, L. Fu, B. Chen, T. Xiao, Y. Wang, F. Chen, J. Liu, K. Gao, S. Shen, and R. Zhai. 2020. RNA sequencing of exosomes revealed differentially expressed long noncoding RNAs in early-stage esophageal squamous cell carcinoma and benign esophagitis. *Epigenomics*. 12:525-541.

Tie, C.H.C., L. Fernandes, L. Conde, L. Robbez-Masson, R.P. Summer, T. Peacock, M.T. Rodriguez-Plata, G. Mickute, R. Gifford, G.J. Towers, J. Herroro, and H.M. Rowe. 2018. KAP1 regulates endogenous retroviruses in adult human cells and contributes to innate immune control. *Frontiers in Chemistry*. 30.

Tiessen, A., P. Pérez-Rodríguez, and L. Delaye-Arredondo. 2012. Mathematical modeling and comparison of protein size distribution in different plant, animal, fungal and microbial species reveals a negative correlation between protein size and protein number, thus providing insight into the evolution of proteomes. *BMC Res. Notes*. 5:85-85.

Topham, J.T., E. Titmuss, E.D. Pleasance, L.M. Williamson, J.M. Karasinska, L. Culibrk, M.K.C. Lee, S. Mendis, R.E. Denroche, G.H. Jang, S.E. Kalloger, H.L. Wong, R.A. Moore, A.J. Mungall, G.M. O'Kane, J.J. Knox, S. Gallinger, J.M. Loree, D.L. Mager, J. Laskin, M.A. Marra, S.J.M. Jones, D.F. Schaeffer, and D.J. Renouf. 2020. Endogenous retrovirus transcript levels are associated with immunogenic signatures in multiple metastatic cancer types. *Mol. Cancer Ther.* 19:1889-1897.

Trapnell, C., B.A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M.J. van Baren, S.L. Salzberg, B.J. Wold, and L. Pachter. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28:511-515.

Tretyakov, K. 2024. matplotlib_venn.

Tsai, S.H., M. Kinoshita, T. Kusu, H. Kayama, R. Okumura, K. Ikeda, Y. Shimada, A. Takeda, S. Yoshikawa, K. Obata-Ninomiya, Y. Kurashima, S. Sato, E. Umemoto, H. Kiyono, H.

Karasuyama, and K. Takeda. 2015. The ectoenzyme E-NPP3 negatively regulates ATP-dependent chronic allergic responses by basophils and mast cells. *Immunity*. 42:279-293.

Tubio, J.M.C., Y. Li, Y.S. Ju, I. Martincorena, S.L. Cooke, M. Tojo, G. Gundem, C.P. Pipinikas, J. Zamora, K. Raine, A. Menzies, P. Roman-Garcia, A. Fullam, M. Gerstung, A. Shlien, P.S. Tarpey, E. Papaemmanuil, S. Knappskog, P. Van Loo, M. Ramakrishna, H.R. Davies, J. Marshall, D.C. Wedge, J.W. Teague, A.P. Butler, S. Nik-Zainal, L. Alexandrov, S. Behjati, L.R. Yates, N. Bolli, L. Mudie, C. Hardy, S. Martin, S. McLaren, S. O'Meara, E. Anderson, M. Maddison, S. Gamble, C. Foster, A.Y. Warren, H. Whitaker, D. Brewer, R. Eeles, C. Cooper, D. Neal, A.G. Lynch, T. Visakorpi, W.B. Isaacs, L.V. Veer, C. Caldas, C. Desmedt, C. Sotiriou, S. Aparicio, J.A. Foekens, J.E. Eyfjord, S.R. Lakhani, G. Thomas, O. Myklebost, P.N. Span, A.L. Borresen-Dale, A.L. Richardson, M. Van de Vijver, A. Vincent-Salomon, G.G. Van den Eynden, A.M. Flanagan, P.A. Futreal, S.M. Janes, G.S. Bova, M.R. Stratton, U. McDermott, P.J. Campbell, I.B.C. Group, I.B.C. Group, and I.P.C. Group. 2014. Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science*. 345:1251343.

Tunbak, H., R. Enriquez-Gasca, C.H.C.C. Tie, P.A. Gould, P. Mlcochova, R.K. Gupta, L. Fernandes, J. Holt, A.G. Van Der Veen, E. Giampazolias, K.H. Burns, P.V. Maillard, and H.M. Rowe. 2020. The HUSH complex is a gatekeeper of type I interferon through epigenetic regulation of LINE-1s. *Nat. Commun.* 11.

Turajlic, S., K. Litchfield, H. Xu, R. Rosenthal, N. McGranahan, J.L. Reading, Y.N.S. Wong, A. Rowan, N. Kanu, M. Al Bakir, T. Chambers, R. Salgado, P. Savas, S. Loi, N.J. Birkbak, L. Sansregret, M. Gore, J. Larkin, S.A. Quezada, and C. Swanton. 2017. Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. *Lancet Oncol.* 18:1009-1021.

Van, R., Guido, Drake, and L. Fred. 2009. Python 3 Reference Manual. CreateSpace, Scotts Valley, CA.

Vargiu, L., P. Rodriguez-Tomé, G.O. Sperber, M. Cadeddu, N. Grandi, V. Blikstad, E. Tramontano, and J. Blomberg. 2016. Classification and characterization of human endogenous retroviruses mosaic forms are common. *Retrovirology*. 13.

Wang-Johanning, F., M. Li, F.J. Esteva, K.R. Hess, B. Yin, K. Rycaj, J.B. Plummer, J.G. Garza, S. Ambs, and G.L. Johanning. 2013. Human endogenous retrovirus type K antibodies and mRNA as serum biomarkers of early-stage breast cancer. *Int. J. Cancer*. 134:587-595.

Wang-Johanning, F., K. Rycaj, J.B. Plummer, M. Li, B. Yin, K. Frerich, J.G. Garza, J. Shen, K. Lin, P. Yan, S.A. Glynn, T.H. Dorsey, K.K. Hunt, S. Ambs, and G.L. Johanning. 2012. Immunotherapeutic potential of anti-human endogenous retrovirus-k envelope protein antibodies in targeting breast tumors. *Journal of the National Cancer Institute*. 104:189-210.

Wang, H., J. Xing, D. Grover, D.J. Hedges, K. Han, J.A. Walker, and M.A. Batzer. 2005. SVA elements: a hominid-specific retroposon family. *J. Mol. Biol.* 354:994-1007.

Wang, L., J. Wang, E. Jia, Z. Liu, Q. Ge, and X. Zhao. 2020. Plasma RNA sequencing of extracellular RNAs reveals potential biomarkers for non-small cell lung cancer. *Clinical Biochemistry*. 83:65-73.

Wang, S., V. Bohnert, A.J. Joseph, V. Sudaryo, G. Skariah, J.T. Swinderman, F.B. Yu, V. Subramanyam, D.M. Wolf, X. Lyu, L.A. Gilbert, L.J. Van't Veer, H. Goodarzi, and L. Li. 2023. ENPP1 is an innate immune checkpoint of the anticancer cGAMP-STING pathway in breast cancer. *Proc Natl Acad Sci U S A*. 120:e2313693120.

Wang, T.-Y., Q. Liu, Y. Ren, S.K. Alam, L. Wang, Z. Zhu, L.H. Hoeppner, S.M. Dehm, Q. Cao, and R. Yang. 2021. A pan-cancer transcriptome analysis of exitron splicing identifies novel cancer driver genes and neoepitopes. *Mol. Cell*. 81:2246-2260.e2212.

Wei, W., N. Gilbert, S.L. Ooi, J.F. Lawler, E.M. Ostertag, H.H. Kazazian, J.D. Boeke, and J.V. Moran. 2001. Human L1 retrotransposition: cis preference versus trans complementation. *Mol. Cell. Biol.* 21:1429-1439.

Wentzensen, N., J.F. Coy, H.-P. Knaebel, M. Linnebacher, B. Wilz, J. Gebert, and M. von Knebel Doeberitz. 2007. Expression of an endogenous retroviral sequence from the HERV-H group in gastrointestinal cancers. *Int. J. Cancer*. 121:1417-1423.

Witzgall, R., E. O'Leary, A. Leaf, D. Onaldi, and J.V. Bonventre. 1994. The Kruppel-associated box-A (KRAB-A) domain of zinc finger proteins mediates transcriptional repression. *Proc Natl Acad Sci U S A*. 91:4514-4518.

Wu, H., J. Feng, W. Zhong, X. Zouxu, Z. Xiong, W. Huang, C. Zhang, X. Wang, and J. Yi. 2023. Model for predicting immunotherapy based on M2 macrophage infiltration in TNBC. *Front. Immunol.* 14.

Xu, W., M.B. Atkins, and D.F. McDermott. 2020. Checkpoint inhibitor immunotherapy in kidney cancer. *Nat. Rev. Urol.* 17:137-150.

Yamamoto, G., I. Miyabe, K. Tanaka, M. Kakuta, M. Watanabe, S. Kawakami, H. Ishida, and K. Akagi. 2021. SVA retrotransposon insertion in exon of MMR genes results in aberrant RNA splicing and causes Lynch syndrome. *Eur. J. Hum. Genet.* 29:680-686.

Yang, C., Y. Li, M. Trottier, M.P. Farrell, V.K. Rai, E.S.-M. E, D.J. Gallagher, Z.K. Stadler, H.M. van der Klift, and L. Zhang. 2021a. Insertion of an SVA element in MSH2 as a novel cause of Lynch syndrome. *Genes Chromosomes Cancer*. 60:571-576.

Yang, F., B. Tanasa, R. Micheletti, K.A. Ohgi, A.K. Aggarwal, and M.G. Rosenfeld. 2021b. Shape of promoter antisense RNAs regulates ligand-induced transcription activation. *Nature*:1-6.

Young, G., and J. Attig. 2019. orf_scanner. Github. outputs CDS predictions from input sequences.

Yuan, T., X. Huang, M. Woodcock, M. Du, R. Dittmar, Y. Wang, S. Tsai, M. Kohli, L. Boardman, T. Patel, and L. Wang. 2016. Plasma extracellular RNA profiles in healthy and cancer patients. *Sci. Rep.* 6:19413-19413.

Zare, M., S. Mostafaei, A. Ahmadi, S. Azimzadeh Jamalkandi, A. Abedini, Z. Esfahani-Monfared, R. Dorostkar, and M. Saadati. 2018. Human endogenous retrovirus env genes: Potential blood biomarkers in lung cancer. *Microb. Pathog.* 115:189-193.

Zhang, X.O., H. Pratt, and Z. Weng. 2021. Investigating the Potential Roles of SINEs in the Human Genome. *Annu. Rev. Genomics Hum. Genet.* 22:199-218.

Zhou, F., J. Krishnamurthy, Y. Wei, M. Li, K. Hunt, G.L. Johanning, L.J.N. Cooper, and F. Wang-Johanning. 2015. Chimeric antigen receptor T cells targeting HERV-K inhibit breast cancer and its metastasis through downregulation of Ras. *Oncoimmunology*. 4.

Zhou, F., M. Li, Y. Wei, K. Lin, Y. Lu, J. Shen, G.L. Johanning, and F. Wang-Johanning. 2016. Activation of HERV-K Env protein is essential for tumorigenesis and metastasis of breast cancer cells. *Oncotarget*. 7:84093-84117.

Zhou, R., Q. Zhang, and P. Xu. 2020. TBK1, a central kinase in innate immune sensing of nucleic acids and beyond. *Acta Biochim Biophys Sin (Shanghai)*. 52:757-767.

Zhou, W., M.Y. Fong, Y. Min, G. Somlo, L. Liu, M.R. Palomares, Y. Yu, A. Chow, S.T. O'Connor, A.R. Chin, Y. Yen, Y. Wang, E.G. Marcusson, P. Chu, J. Wu, X. Wu, A.X. Li, Z. Li, H. Gao, X. Ren, M.P. Boldin, P.C. Lin, and S.E. Wang. 2014. Cancer-secreted miR-105 destroys vascular endothelial barriers to promote metastasis. *Cancer Cell*. 25:501-515.

Zhou, W., Y. Li, D. Gu, J. Xu, R. Wang, H. Wang, and C. Liu. 2022. High expression COL10A1 promotes breast cancer progression and predicts poor prognosis. *Heliyon*. 8:e11083-e11083.

Zhou, Z., Q. Wu, Z. Yan, H. Zheng, C.J. Chen, Y. Liu, Z. Qi, R. Calandrelli, Z. Chen, S. Chien, H. Irene Su, and S. Zhong. 2019. Extracellular RNA in a single droplet of human serum reflects physiologic and disease states. *Proc. Natl. Acad. Sci. U. S. A.* 116:19200-19208.

Zhu, M., X. Chen, H. Zhang, N. Xiao, C. Zhu, Q. He, W. Guo, Z. Cai, H. Shen, and Y. Wang. 2011. AluYb8 insertion in the MUTYH gene and risk of early-onset breast and gastric cancers in the Chinese population. *Asian Pac J Cancer Prev*. 12:1451-1455.