Investigating the regulation of APOBEC mutagenesis in cancer

Karim Ahmed

Trinity College

University of Cambridge

This thesis is submitted for the degree of Doctor of Philosophy

August 2023

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Summary

Investigating the regulation of APOBEC mutagenesis in cancer

Karim Ahmed

The sequencing of cancer genomes has revealed that cancers harbour recurrent patterns of mutation, known as mutational signatures. One common mutational signature, known as the APOBEC signature, is found in over 70% of cancer types. The APOBEC signature is thought to be mediated by the activity of the APOBEC enzyme family. However, the underlying cause of APOBEC activity in cancer is not fully understood.

This thesis investigates the regulation of APOBEC mutagenesis in cancer. There is a focus on APOBEC3A and APOBEC3B as likely mediators of APOBEC signature mutations, and a focus on the hypothesis that APOBEC activity might be driven by the activity of LINE-1 retrotransposons.

Firstly, the thesis details experiments to investigate APOBEC activity in cultured cancer cells. The experiments conducted suggest that p53 inactivation leads to the upregulation of LINE-1 and APOBEC3B, and that their expression may be downregulated when p53 activity is promoted. An enzymatic assay for cancer-associated APOBEC activity is established. Reverse transcriptase inhibitors, which inhibit LINE-1 activity, appear to modulate APOBEC3B expression and associated enzymatic activity. This appears to occur when cells are p53-deficient, but not when p53 is intact.

Secondly, the thesis reports exploratory bioinformatic analyses conducted using large genomic datasets. These indicate that APOBEC3A expression is associated with interferon signalling in cancer while APOBEC3B expression is associated with cell cycle signalling in cancer. A deletion of a consensus interferon response factor binding site in the APOBEC3B promoter is identified. Analyses of regulatory data suggest that APOBEC3B might be transcriptionally insulated from syntenic APOBEC3 genes by CTCF. In addition, p53 deficiency in cancer appears to be associated with the upregulation of APOBEC3A and APOBEC3B.

Thirdly, the thesis reports bioinformatic analyses of an RNA sequencing dataset from patients with the rare genetic disease Aicardi-Goutières syndrome. It is thought that the disease process of the genotypes studied might driven by LINE-1 activity, as suggested in part by successful patient trials of reverse transcriptase inhibitors. The analyses conduced appear to identify changes to the transcriptome in Aicardi-Goutières syndrome that might mirror those associated with APOBEC activity in cancer.

In sum, these experiments provide evidence for possible regulators of APOBEC mutagenesis in cancer, including evidence that broadly supports the hypothesis that it may be driven by LINE-1 activity. The experiments also identify a class of drugs that might enable the pharmacological modulation of cancer-associated APOBEC activity. APOBEC mutagenesis is thought to mediate cancer initiation, progression, intratumour heterogeneity and responses to therapy, including immunotherapy. The

findings detailed might therefore contribute to the ability to understand and control the natural history of cancer across multiple cancer types.

Acknowledgements

I am incredibly grateful to Professor Ashok Venkitaraman for his supervision and guidance over a period spanning eleven years, and am very fortunate to have been a student in the Venkitaraman laboratory. I extend my thanks to all our laboratory members for their help and advice, particularly Dr Amy Emery for her support in producing work in the first results chapter. I am also most grateful to Dr Paolo D'Avino for supervising the submission of this thesis and thank the School of Clinical Medicine's MB/PhD Programme for their generous support in funding my PhD and allowing me the opportunity to pursue this work.

For my mother, Abeer.

Table of Contents

List of Abbreviations

3TC	Lamivudine
A	Adenosine
ACTB	Beta-actin
AGS	Aicardi-Goutières syndrome
APC	Adenomatous polyposis coli
APOBEC	Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like
AZT	Azidothymidine
BRCA	Breast cancer susceptibility gene
C	Cytosine
cDNA	Complementary DNA
ChIP	Chromatin immunoprecipitation
CNV	Copy number variant
CTCF	CCCTC-binding factor
d4T	Stavudine
DDR	DNA damage response

DNA	Deoxyribonucleic acid
DSB	Double-strand break
EDTA	Ethylenediaminetetraacetic acid
ENCODE	Encyclopaedia of DNA Elements
FRET	Förster resonance energy transfer
G	Guanine
GATK	Genome Analysis Toolkit
H3K27	Lysine at position 27 in histone H3
HIV	Human immunodeficiency virus
Indel	Insertion/deletion
IRF	Interferon response factor
LINE-1	Long interspersed nuclear element 1
METABRIC	Molecular Taxonomy of Breast Cancer International Consortium
mRNA	Messenger ribonucleic acid
NF-κB	Nuclear factor kappa-light-chain-enhancer of activated B cells
NGS	Next-generation sequencing

NMF	Non-negative matrix factorisation
ORF	Open reading frame
p53	Tumour protein 53
PANTHER	Protein Analysis Through Evolutionary Relationships
PCA	Principal component analysis
PCAWG	Pan-Cancer Analysis of Whole Genomes
PCR	Polymerase chain reaction
RNA	Ribonucleic acid
RNASEH2	Ribonuclease H2
qRT-PCR	Quantitative real-time reverse transcription PCR
RTI	Reverse transcriptase inhibitor
SAMHD1	SAM domain and HD domain-containing protein
SBS	Single base substitution
SNV	Single nucleotide variant
SSB	Single-strand break
ssDNA	Single-stranded DNA
T	Thymine

TAMRA	Tetramethylrhodamine
TBE	Tris-borate-EDTA
TCGA	The Cancer Genome Atlas
TPM	Transcripts per million
TREX1	Three prime repair exonuclease 1
tRNA	Transfer ribonucleic acid
U	Uracil
UDG	Uracil DNA glycosylase
UTR	Untranslated region
VQ	Vector quanisation

1. Introduction

1.1. Cancer: an enduring health problem

The term ‘cancer’ refers to a phenomenon where a bodily cell and its progeny overproliferate, with these cells having the capacity to spread within the body (Weinberg 2007). Cancers arise from cells throughout the body, and are typically named according to their apparent origin.

Cancer is common; approximately 1 in 2 people in the UK (Ahmad, Ormiston-Smith, and Sasieni 2015) and 1 in 6 people worldwide (Bray et al. 2018) are expected to develop cancer. It is the second leading cause of death globally (Wang et al. 2016). Of all known risk factors, cancer is most closely associated with increasing age (Ries et al. 2006). Although advancements in cancer research have led to improvements in patient outcomes in recent decades, around half of cancer patients in England and Wales still die within 5 years of diagnosis (Quaresma, Coleman, and Rachet 2015).

The earliest known description of cancer is thought to have been written in around 3000 BC (Hajdu 2011). Its author - purported to be the physician Imhotep (van Middendorp, Sanchez, and Burridge 2010) - remarks that it is a ‘grave disease’ with ‘no treatment’. Modern epidemiological data reaffirm that cancer evidently remains a grave disease. However, a number of

effective anticancer prevention and treatment strategies have been developed, particularly in the last century (DeVita Jr and Rosenberg 2012).

It is anticipated that further advances in cancer research will lead to further improvements in these strategies for prevention and treatment. A major area of contemporary research is the study of genomic mutation in cancer (Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium 2020), which is the topic of this thesis. The mutation of cellular DNA has been identified as a causative factor - an ‘enabling hallmark of cancer’ (Hanahan and Weinberg 2011), as well as a therapeutic target (Fox and Loeb 2010).

1.2. Mutations as mediators of carcinogenesis

A mutation is defined as a change in a nucleotide sequence (Voet, Voet, and Pratt 2013). Germline mutations are those that can be passed on to offspring, while somatic mutations are those that cannot (Suzuki and Griffiths 1976). Somatic mutations are thought to mediate carcinogenesis - the process of cancer formation (Weinberg 2007).

The evidence supporting the somatic mutation theory of carcinogenesis first emerged in recent centuries. In the nineteenth century, Müller and his student Virchow used microscopy to deduce that cancer arises from cells (Müller 1838; Virchow 1858). Later that century, von Hansemann observed that cancer cells undergo abnormal cell divisions and have unequal numbers of chromosomes (von Hansemann 1890). In the twentieth century, Boveri observed that sea urchins with developmental defects also had abnormal cell divisions and chromosomal abnormalities. He proposed that chromosomal abnormalities could lead to cancer by inducing defects in cellular development (Boveri 1914). The term 'somatic mutation' was first used two years later, referring to 'a permanent modification of somatic tissue' that could be heritably transmitted 'in successive cell generations' of a tumour (Tyzzer 1916).

Experimental and structural studies later identified DNA as the molecule that mediates heredity (Avery, MacLeod, and McCarty 1944; Franklin and Gosling 1953; Watson and Crick 1953). Then, specific alterations in DNA were found to be associated with cancer, as typified by the discovery of a

recurrent translocation (defined in Table 1.2) of chromosomes 9 and 22 in chronic myeloid leukaemia known as the Philadelphia chromosome (Nowell and Hungerford 1960; Rowley 1973). This discovery was enabled by advances in the preparation of samples for microscopy. Studies in this period also made use of viruses that were found to cause cancer (Rous 1910). Such studies led to the finding that a single gene of the Rous sarcoma virus, src, was sufficient for transformation in cultured cells infected with the virus - 'transformation' being the cellular acquisition of cancerous traits (Parker, Varmus, and Bishop 1984). Similarly, it was shown that transferring DNA from cancer cells to non-cancer cells was sufficient for inducing transformation (Shih et al. 1981). Further studies demonstrated that transformation is mediated by the mutation of genes involved in controlling cellular proliferation - 'proto-oncogenes' that mutate to become 'oncogenes' (Perucho et al. 1981; Pulciani et al. 1982). A mutation at a single nucleotide in the HRAS proto-oncogene was subsequently shown to be sufficient for transformation (Reddy et al. 1982).

Boveri had suggested that chromosomal material could promote transformation, as in the case of oncogenes, but might also limit transformation (Boveri 1914). Subsequent studies in the twentieth century supported this hypothesis. In 1971, Knudson used statistical models to propose that familial cases of retinoblastoma that typically occur in younger patients are caused by one inherited germline mutation and a second somatic mutation, while non-familial (sporadic) cases of typically older patients are caused by two somatic mutations (Knudson 1971). This is known as the 'two-hit hypothesis'. The identification of the RB1 gene as the

genomic locus for these two mutations suggested the discovery of the first 'tumour suppressor gene' - a gene that functions to prevent transformation (Friend et al. 1986). It was deduced that, in familial cases of retinoblastoma, one defective RB1 allele is inherited in the germline and the second RB1 allele is inactivated due to somatic mutation. However, in sporadic cases, both RB1 alleles are inactivated somatically. This showed that somatic and germline mutations could contribute to carcinogenesis.

Since the identification of the first oncogene and tumour suppressor gene in the 1980s, over 700 genomic regions that contribute to carcinogenesis when mutated have been identified (Sondka et al. 2018). There have been concurrent efforts to systematically identify the genomic mutations that drive cancer ('driver mutations' - defined in section 1.3.1). The turn of the millennium marked the completion of the first draft of the sequence of the human genome, comprising approximately 3 billion base pairs (Lander et al. 2001). More recently, the large international PCAWG study reported sequencing of the genomes of around 2500 cancers across 38 cancer types - the first time that whole cancer genomes have been sequenced at this scale (PCAWG Consortium 2020). The study has identified several thousand mutations throughout the genome that are thought to drive cancer. This set of mutations was identified by filtering mutations found in these samples according to whether or not they have similar features to known driver mutations. The study's authors comment that this set of mutations is expected to expand as more cancer genomes are sequenced and methods of studying mutations in cancer (described in section 1.3) are

refined. A number of mutation classes have been identified to date (Tables 1.1 and 1.2).

Mutation class	Definition
Single-base substitution, also known as a single nucleotide variant (SNV) or a point mutation	The substitution of the base at an individual nucleotide.
n-base substitution (eg. doublet-base, triplet-base etc.)	The substitution of bases at n contiguous nucleotides.
Small insertion or deletion (Indel)	The gain or loss, respectively, of one or a few nucleotides.
Structural variant, also known as a genomic rearrangement	The gain, loss or reordering of chromosomal segments that typically range in size from kilobases to whole chromosomes.
Aneuploidy	The gain or loss of entire chromosomes.
Whole-genome duplication	The doubling of the number of all chromosomes.

Table 1.1 - Mutation classes identified in cancer genomes (Adapted from Ahmad, Ahmed, and Venkitaraman 2018).

Structural variant class	Definition
Copy number variant (CNV)	The gain or loss of copies of a chromosomal segment.
Amplification	The gain of a chromosomal segment.
Deletion	The loss of a chromosomal segment.
Tandem duplication	The doubling of a chromosomal segment in which the additional copy is inserted immediately adjacent to the original copy and in the same orientation.
Reciprocal inversion	A chromosomal segment that has been reversed in orientation.

Fold-back inversion	An inverted rearrangement between two points typically over 20 kilobases apart on a chromosome, with associated copy number change.
Translocation	The rearrangement of non-homologous chromosomes.
Chromoplexy	The rearrangement of more than two chromosomal segments such that these segments are preserved and shuffled with regards to genomic position and orientation.
Chromothripsis	A cluster of tens or hundreds of structural variants in one or a few chromosomes that are thought to occur as a single event, with associated oscillations in copy number and segment orientation.
Local n-jump	A cluster of n structural variants in a single genomic region.

Templated insertion	The addition, into a single chromosome, of a contiguous string of copies of one or more genomic segments.
Local-distant cluster	A cluster of structural variants that has both local rearrangements and rearrangements to other parts of the genome.

Table 1.2 - Structural variant classes identified in cancer genomes (Adapted from Li et al. 2020).

The PCAWG study also reports that approximately 5-9% of cases studied have ‘no apparent driver mutations’, a finding which could ‘arise from either technical or biological causes’. However, the study also identifies examples where cancers are found to have driver mutations where none were initially apparent. For example, cases of chromophobe renal cell carcinoma demonstrate ‘a remarkably consistent profile of chromosomal aneuploidy’ that appears to be ‘sufficient to initiate a cancer in the absence of more-targeted driver events’. Such a mechanism was first posited by Boveri after observing a range of chromosomal profiles in abnormal sea urchin cells (Boveri 1914).

Together, the observations made from the nineteenth century to the present day indicate that mutations in cellular DNA mediate transformation. It has also been shown that many of the mutations that occur in cancer predispose cells to acquiring additional mutations - a trait described as 'genomic instability'. The increase in genomic instability observed in cancer is thought to enable transformation by enabling subsequent mutations in proto-oncogenes and tumour suppressor genes.

For example, TP53 has been identified the as most frequently mutated gene in cancer (PCAWG 2020). Mutations in TP53 are observed to be 'among the earliest [mutational] events' across cancer types in the PCAWG cohort (Gerstung et al. 2020) and TP53 mutation induces genomic instability that includes structural variation and aneuploidy (Ciriello et al. 2013; Kastenhuber and Lowe 2017). The canonical function of the TP53 gene product, known as p53, is to arrest the proliferative cell cycle or initiate programmed cell death (apoptosis) in response to DNA damage (Weinberg 2007). These functions of p53 were identified upon its initial characterisation in the 1990s, with Lane describing p53 as a 'guardian of the genome' in 1992 (Lane 1992). Many studies on TP53 have followed since. Indeed, TP53 has been identified as 'the most studied human gene of all time' (Hafner et al. 2019), based on the large number of publications that refer to it. In addition to its canonical function, several thousand studies collectively report a wide range of functions for its unmutated (wild-type) and mutant forms (Kastenhuber and Lowe 2017).

TBP3, RB1 and many other genes implicated in carcinogenesis are now understood to be part of the cellular DNA damage response (DDR - Elledge 1996). Failure of the DDR leads to mutagenesis mediated by DNA damage - with DNA damage defined as ‘a change that introduces a deviation from the usual double-helical structure’ (Lewin 2004; Table 1.3). The DDR involves a number of cellular mechanisms (reviewed in Jackson and Bartek 2009). These include the sensing of DNA damage, the repair of DNA damage (Table 1.3), cell signalling cascades, changes to gene expression (transcription), changes to DNA replication, cell cycle arrest, apoptosis and cellular senescence.

DNA repair mechanism	Principal DNA lesions repaired
Direct lesion reversal	O^6 alkylguanine
Mismatch repair	DNA mismatches, insertion/ deletion loops arising from DNA replication
Base excision repair and SSB repair	Abnormal DNA bases, simple base adducts, SSBs generated as base excision repair intermediates
Nucleotide excision repair	Lesions that disrupt the DNA double helix, such as bulky base adducts and UV photoproducts
Translesion synthesis	Base damage blocking replication fork progression
Non-homologous end-joining	DSBs
Homologous recombination	DSBs, stalled replication forks, inter-strand DNA crosslinks
Fanconi anaemia pathway	Inter-strand DNA crosslinks

Table 1.3 - DNA damage and repair processes (adapted from Jackson and Bartek 2009).

Recurrent somatic mutations in DDR genes are found in many cancer types. For example, the ten most frequently mutated cancer genes in the PCAWG cohort include genes implicated in the DDR such as TP53, CDKN2A, ARID1A, PTEN, TERT, CDKN2B, SMAD4 and RB1, with TP53 mutations found in 37% of cases and 96% of cancer types (Negrini, Gorgoulis, and Halazonetis 2010; Katz et al. 2013; Lee, Chen, and Pandolfi 2018; Mathur 2018; PCAWG 2020). In addition, a number of mutations in DDR genes are thought to mediate hereditary cancer syndromes (Table 1.4), including TP53 in Li-Fraumeni syndrome and RB1 in hereditary retinoblastoma.

Disease	Gene(s)	Principal defective response
46BR syndrome	LIG1	Chromosomal stability
Ataxia telangiectasia	ATM	Repair of DNA strand breaks
Basal cell nevus syndrome	PTCH2	Cell signalling
Bloom syndrome	BLM	Resolution of stalled replication/ transcription intermediates

Cowden syndrome and Bannayan-Riley-Ruvalcaba syndrome	PTEN	Cell cycle responses and apoptosis
Cutaneous malignant melanoma	CDKN2A, CDK4	Cell cycle responses and apoptosis
Familial adenomatous polyposis	APC	Cell proliferation and chromosomal stability
Fanconi anaemia	FANCA, FANCB, FANCC, FANCD1, BRCA2, FANCD2, FANCE, FANCF, FANCG, FANCI, FANCJ, FANCL	Chromosomal stability, both spontaneous and in response to cross-linking agents
Hereditary breast and ovarian cancer	BRCA1, BRCA2	Cell cycle response to DNA damage
Hereditary nonpolyposis colon cancer	MLH1, MSH2, MSH6, PMS1, PMS2, MLH3, EXO1	Mismatch repair

Hereditary papillary renal cell carcinoma	MET	Cell signalling
Juvenile polyposis syndrome	SMAD4, BMPR1A	Cell signalling
Li-Fraumeni syndrome	TP53, CHEK2	Cell cycle response to DNA damage
LIG4 syndrome	LIG4	Repair of DNA strand breaks
MYH-associated polyposis	MYH	None noted, despite mutations in a base excision repair gene
Neurofibromatoses type 1 and type 2	NF1, NF2	RAS regulation or cell cytoskeleton maintenance
Nijmegen breakage syndrome	NBS1	Repair of DNA strand breaks
Peutz-Jeghers syndrome	STK11	Cell cycle responses and apoptosis
Retinoblastoma	RB1	Cell cycle response to DNA damage

Rothmund-Thomson syndrome	RECQL4	Resolution of stalled replication/ transcription intermediates
Tuberous sclerosis complex	TSC1, TSC2	Cell cytoskeleton maintainance
von Hippel-Lindau	VHL	Multiple; possibly cell cycle regulation
Werner syndrome	WRN	Resolution of stalled replication/ transcription intermediates
Wilm's tumour	WT1	Transcriptional regulation
Xeroderma pigmentosum	XPA, XPB, XPC, XPD, XPE, XPF, XPG, XPV	Nucleotide excision repair and translesion DNA synthesis

Table 1.4 - Hereditary diseases involving cancer predisposition, many of which are implicated in a defective DNA damage response (adapted from Friedberg et al. 2006).

The DNA damage that the DDR counteracts can occur as a result of processes that are intrinsic to the cell or from exposure to its environment. For example, common oncogene mutations lead to changes in the timing and progression of DNA replication during the cell cycle - such disruptions in replication are known as 'replication stress' and involve DNA lesions that trigger the DDR (Bartkova et al. 2005; Gorgoulis et al. 2005). In addition, external sources of DNA damage such as radiation, chemical agents and biological agents have been shown to cause mutations and increase cancer incidence. Early experimental studies indicated that approximately 90% of identified carcinogens are also mutagenic (McCann and Ames 1976). Epidemiological studies have since identified over 100 carcinogens and over 300 other exposures that are probably or possibly carcinogenic, as determined by the International Agency for Research on Cancer (Cogliano et al. 2011).

In summary, alterations in cellular DNA mediate transformation through the activation of oncogenes and the inactivation of tumour suppressor genes. These changes often lead to genomic instability by compromising the cellular response to DNA damage. In addition, carcinogenic exposures are often mutagenic. Together, these findings support the notion that mutations contribute to carcinogenesis.

1.3. Mutations as modifiers of cellular fitness

1.3.1. Driver mutations and passenger mutations

Efforts have been made to classify the mutations found in cancer according to their effect on cancer development. Mutations may be classified as driver mutations or passenger mutations (Stratton, Campbell, and Futreal 2009). A driver mutation is defined as one that endows a cell with a proliferative advantage. In contrast, a passenger mutation is one that does not confer a proliferative advantage.

The proliferative advantage that occurs as a result of a driver mutation may be viewed as a Darwinian selective advantage that increases the fitness of the cancer cell within its environment (Pepper et al. 2009). A number of selection pressures within the cellular environment have been identified. These include the genetic profile of other cells; lifestyle factors such as external exposures; systemic factors such as hormones, growth factors, immune cells and cytokines; local factors such as oxygen, nutrients, and space; and architectural constraints such as physical compartments, basement membranes and restricted niches (Greaves and Maley 2012).

1.3.2. Cancer as a multi-step evolutionary process

The notion of Darwinian selection as a feature of carcinogenesis was first proposed by Nowell in the 1970s (Nowell 1976). He posited that cancer is clonal (that the cells that form a tumour arise ‘from a single cell of origin’)

and that the sequential selection of more aggressive subclones emerges as a result of 'acquired genetic variability within the original clone'.

The evidence for iterative selection in carcinogenesis was first suggested by an observed delay between exposure to mutagens and cancer formation, and mathematical modelling of the relationship between cancer incidence and age, which together suggested that multiple mutations were required in the process of carcinogenesis (Nordling 1953; Armitage and Doll 1954). Since the incidence of cancer appeared to increase according to the sixth power of age, it was proposed that there were around six rate-limiting mutational events in carcinogenesis.

These observations were supported by histological and molecular studies that indicated that normal cells undergo premalignant changes that enhance their capacity for proliferation before a malignant cancer is formed. For example, histological studies of colon tumours led to their classification as early adenomas (premalignant), intermediate adenomas (premalignant), late adenomas (premalignant) and carcinomas (malignant cancer). Genetic analyses of these lesions indicated that four or five known drivers were found in carcinomas, and that particular drivers were enriched in different stages (Fearon and Vogelstein 1990). For example, loss of chromosome 5q (containing APC) appeared to be important for the formation of early adenomas, KRAS mutations appeared to be important for the formation of intermediate adenomas, loss of chromosome 18q (containing SMAD4) for late adenoma formation and loss of chromosome 17p (containing TP53) for carcinoma formation.

Contemporary whole-genome analyses have supported these early studies, finding that cancers harbour four to five drivers on average (PCAWG 2020). In addition, across cancer types, there are a number of characteristic subclonal or late changes and clonal or early changes, with some driver mutations preceding diagnosis by years or decades (Gerstung et al. 2020). For example, the classical APC-KRAS-SMAD4-TP53 model of colorectal carcinogenesis posed by Fearon and Vogelstein has been expanded, and is presented as a probabilistic process involving several genetic loci, with TP53 inactivation appearing to be an early event. The order of mutations in whole genome sequencing studies such as this may be inferred from the estimated clonality of those mutations in the sample, while timing can be inferred by calibrating these data according to mutational processes that appear to increase linearly with age. This information can also be used to construct evolutionary trees.

Together, the driver events that occur in cancer lead to a range of putatively adaptive changes. Experimental studies interrogating the function of driver mutations have shown that cancer cells subvert a range of processes to overcome the selection pressures in their environment (Hanahan and Weinberg 2011). Cancers may activate proliferative signalling, angiogenesis, metastasis, replicative immortality, tumour-promoting inflammation and genomic instability, while subverting growth suppression, cell death, cellular metabolism and immune destruction. Mutations may also mediate treatment resistance.

Although four to five driver mutations are under positive selection in a typical cancer, thousands of putatively evolutionarily neutral passenger mutations are also present (Martincorena et al. 2017; Alexandrov et al. 2020). In addition, some mutations appear to be under negative selection, such as those that generate a protein that is detected as foreign by the immune system (a neoantigen) or those that occur in genes that are understood to be essential for cellular survival (Martincorena et al. 2017; Rosenthal et al. 2019). These findings highlight that many more mutations occur in cancers than those required to drive them, and that some mutations are deleterious to cancer cells.

1.3.3. The relationship between mutation rate and fitness

The relationship between a cell's rate of mutation and its fitness can be informed by considering cells with mutation rates at their theoretical limits. At one extreme, one might consider a cell whose genome undergoes no mutagenesis whatsoever. Such a cell cannot evolve and therefore lacks the capacity to gain any fitness advantage. At the other extreme, one might consider a cell whose rate of mutation is so great that it entirely degrades the integrity of the instructions encoded in its genome. Such a cell would no longer be able to perform the essential functions required for its survival. These two hypothetical examples, of cells with mutation rates at opposite extremes, illustrate that the rate of mutation is inherently tied to the fitness of a cell. They also indicate that excessively low or excessively high rates of mutation are not advantageous. Indeed, species appear to evolve mutation

rates that are between these two extremes, with multiple determinants appearing to influence the evolved rates of mutation (Lynch et al. 2016).

Experimental and computational studies on species from a range of phylogenetic lineages have indicated that, in general, mutations arise stochastically, a large fraction of mutations are deleterious, and advantageous mutations are rare (Loewe and Hill 2010). These findings support the notion that the rate of mutation is, in itself, a determinant of fitness, and the notion that unduly high rates of mutation are likely to be selected against (Kimura 1967; Loewe and Hill 2010) particularly when the genetic material available to a population is limited due to factors such as a limited population size or a lack of sexual recombination between individuals, as occurs in a population of cancer cells.

Asexually reproducing cells, such as cancer cells, are understood to be susceptible to the phenomenon known as 'Muller's ratchet' - the irreversible accumulation of deleterious mutations in the genome in the absence of sexual recombination (Muller 1964). Muller's ratchet has been observed in asexually reproducing species across taxonomic domains, including fish (Loewe and Lamatsch 2008), worms (Loewe and Cutter 2008), bacteria (Andersson and Hughes 1996) and amoebae (Maciver 2016). It has also has been observed in parts of the human genome that do not undergo recombination in the germline, such as the Y chromosome and the mitochondrial genome (Loewe 2006; Engelstädter 2008). Moreover, Muller's ratchet is particularly prominent in small populations, where it is thought to inevitably lead to extinction - a process that has been termed

‘mutational meltdown’ (Lynch et al. 1993). Mutational meltdown is described to occur as deleterious mutations further reduce the population size and are fixed due to the effect of genetic drift (the change in allele frequency due to random sampling of individuals in the population).

1.3.4. The effect of mutation rate on cancer cell fitness

As described in section 1.3.2, mutations give cancer cells a selective advantage relative to non-cancer cells, increasing their fitness by subverting restrictions on growth, cell death and bodily location as well as mediating treatment resistance (Hanahan and Weinberg 2011). An elevated rate of mutation, as is observed is cancer, increases the probability that such mutations will occur, thereby increasing the probability that cancer cells gain a selective advantage.

On the other hand, since cancer cells act as a limited number of asexually reproducing cells in the body, they are subject to the constraints described in section 1.3.3. As a result, an excessively elevated mutation rate leads to a fitness disadvantage in cancer cells.

A number of findings support the notion that excessively high mutation rates pose a fitness disadvantage to cancer cells. Firstly, there is evidence that whole-genome doubling events in cancer mitigate against the effects of Muller’s ratchet (López et al. 2020). In addition, clonally transmissible cancers, which affect non-human species and grow for many more generations than conventional cancers, avoid mutational meltdown by

reverting to genomic stability and gaining the capacity to replenish their mitochondrial genomes from their host's mitochondria (Ní Leathlobhair and Lenski 2022). In human cancer, genomically unstable BRCA2-deficient cancers acquire secondary mutations that subsequently decrease their chromosomal instability (Lee et al. 1999). Moreover, experimentally-induced exacerbations of genomic instability impair the growth of cancer cells, and abolish their ability to form xenografts (Godek et al. 2016). Lastly, an increased rate of mutation in cancer cells also increases the probability of neoantigen formation, increasing the susceptibility of cancer cells to destruction by the immune system (Niknafs et al. 2023).

Together, the findings detailed in this section indicate that cancer cells are disadvantaged by excessively low or excessively high rates of mutation. It therefore seems plausible that therapeutics that modulate the activity of mutational processes in cancer cells may be of clinical benefit. In keeping with this, very low and very high rates of mutation have been found to be associated with improved patient survival. A large pan-cancer study of 1165 patients with 12 different cancer types indicates that mutation rates at the extremes, in terms of either the rate of point mutations or copy number variation, are associated with an approximately two-fold or seven-fold improvement in survival, respectively (Andor et al. 2015). Improvements in survival have also been found to be associated with very high or very low levels of chromosomal instability (Birkbak et al. 2011; Jamal-Hanjani et al. 2015).

1.4. Next-generation sequencing of cancer

The term 'next-generation sequencing' (NGS) refers to a high-throughput method of nucleic acid sequencing developed in the mid-2000s, characterised by its relatively high speed and low cost (Goodwin, McPherson, and McCombie 2016). It is contrasted with the first generation of nucleic acid sequencing technologies developed by Sanger and others in the 1970s, known as 'Sanger sequencing' (Sanger, Nicklen, and Coulson 1977). By Sanger sequencing cancer genes in DNA extracted from cancer cells, small-scale mutations could be identified in early low-throughput studies (Greenblatt et al. 1994). In addition, large-scale structural variation could be studied through microscopy - a process that was refined by the use of labelled probes that bind to specific DNA regions (Speicher, Ballard, and Ward 1996). Since the sequencing of the first human genome by automated Sanger sequencing in 2001 (Lander et al. 2001), and the sequencing of the first cancer genome by NGS in 2010 (Pleasance et al. 2010), NGS has been used to catalogue small-scale and large-scale DNA alterations in around 2500 whole cancer genomes (PCAWG 2020). This has yielded data on cancer mutagenesis at a scale that has not previously existed - data that is at the resolution of single nucleotides, throughout the length of the genome, and from a large sample of patients.

Through reverse transcription of RNA extracts, NGS has also been applied to measure the transcriptome of cancer cells (RNA-seq) with greater sensitivity and specificity than preceding transcriptomic technologies such as microarrays that, like older DNA sequencing technologies, are also

based on sequence-specific probes (Mardis and Wilson 2009). In addition, specific segments of the genome have been captured and sequenced by NGS, such as in the case of chromatin immunoprecipitation sequencing (ChIP-seq) which identifies DNA regions bound by specific DNA-associated proteins (Park 2009).

The NGS process involves a number of biochemical stages (section 1.4.1) followed by a number of bioinformatic stages (section 1.4.2). Commonly-used methods for studying cancer genomes by NGS are described below. A number of limitations of these process have been identified, which may be used to aid interpretation of the data generated (section 1.4.3).

1.4.1. Biochemical stages

Illumina/Solexa sequencing, a widespread NGS method used by ICGC, the International Cancer Genome Consortium (ICGC 2010), is described in this section and given in Figure 1.1. First, DNA is extracted from cells and fragmented into smaller pieces of approximately 200-500 base pairs mechanically, enzymatically or by sonication. Then, adaptors are ligated to each end of the new fragments (Figure 1.1a). These adaptors are used to adhere the fragment ends to the surface of a flow cell, where the sequencing reaction takes place (Figure 1.1b). First, a polymerase chain reaction (PCR) is used to generate several copies of the fragment. The PCR is primed by a lawn of adaptors adjacent to DNA fragment, and the fragment bends to form a bridge between its binding site and the adjacent primer in a process termed ‘bridge amplification’ (Figure 1.1c). These

copies are adhered to the surface adjacent to the original copy, which increases the size of the surface occupied by each set of copies, to form a spot that is large enough to be observed by an optical microscope if illuminated. To enable detection by such a microscope, a fluorescent DNA synthesis reaction is undertaken. A mixture of free A, C, T and G nucleotides that are fluorescently labelled with a colour that corresponds to each nucleotide is added to provide substrates for the reaction. Once a nucleotide is incorporated into the nascent DNA strand in the synthesis reaction its fluorophore is imaged, then cleaved to permit the next nucleotide to be incorporated. Thus, under the microscope, each spot flashes with a different colour - depending on the nucleotide sequence - at each step of the synthesis reaction. The sequence of colours at each spot can then be decoded to give the sequence of bases of the PCR-amplified fragments in the spot. Imperfections in this process, where the sequence is unclear, are summarised as a quality score that can be accounted for in subsequent analyses. This step concludes the contribution of biochemical processes to the NGS protocol. The information obtained is subsequently used to computationally reconstruct the genome. In order to identify mutations specific to cancer genomes, both cancer and normal cells are processed in this way, to enable downstream comparison.

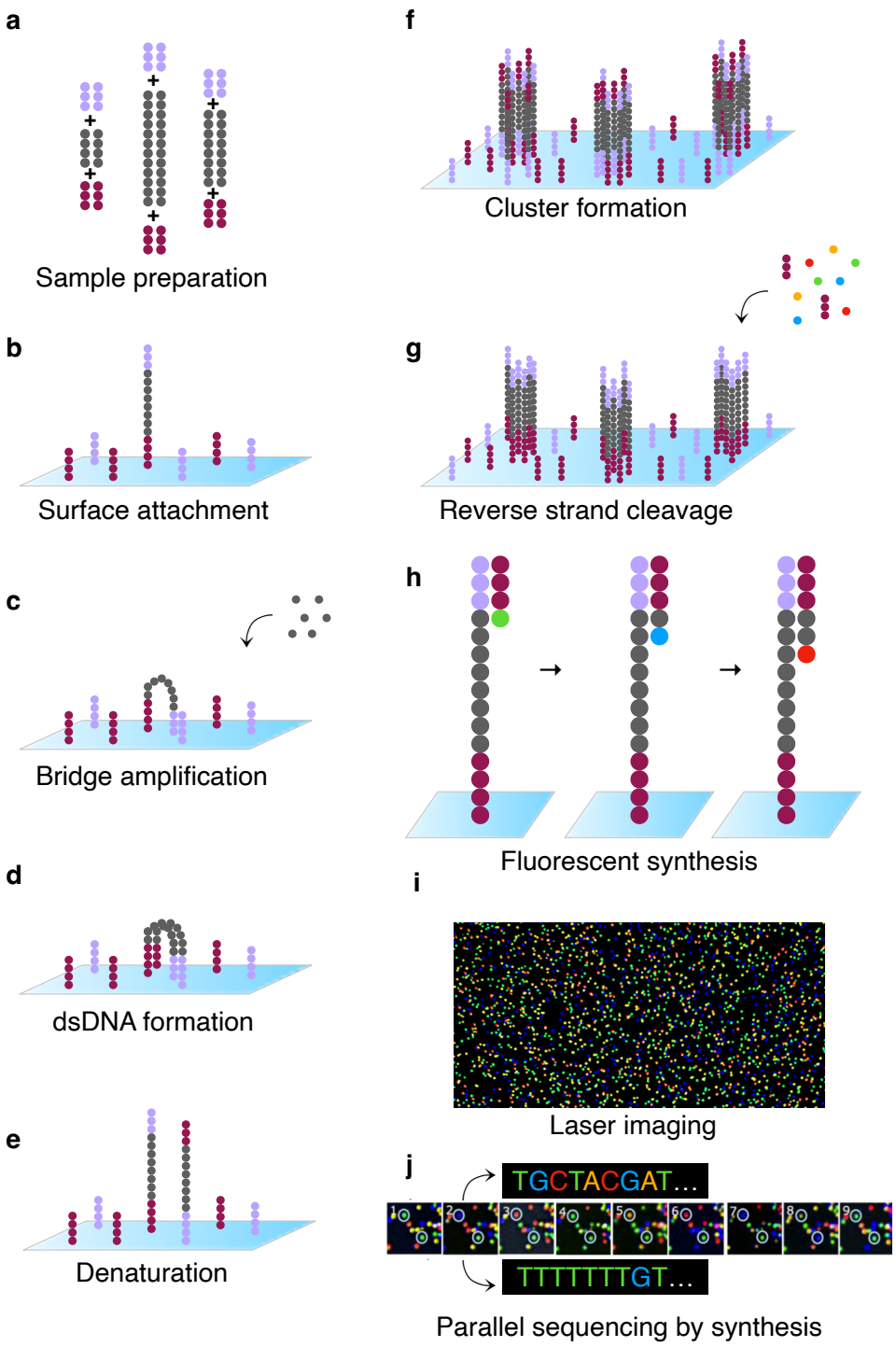


Figure 1.1 - Illumina/Solexa sequencing (described in section 1.4.1). Adapted from materials on the Illumina (**a-h**: illumina.com) Wellcome Genome Campus (**i**: wellcomegenomecampus.org) and European Bioinformatics Institute (**j**: ebi.ac.uk) websites. Fluorescent spots in **i** and **j** are approximately 1.5 μm in diameter.

1.4.2. Bioinformatic stages

The bioinformatic analysis of NGS data also involves multiple steps. The output of next generation sequencing machines - the inferred code of the very many fragments - is known as a set of 'reads'. The first step in the computational analysis is a quality control step, to account for known imperfections in the process. For example, the quality score of bases towards the end of reads is known to tend to be low. Reads are therefore often 'trimmed' to remove these low-quality ends, in order to facilitate downstream interpretation. In most applications, this does not meaningfully affect the coverage of the genome. This is because there are typically several overlapping reads from the many copies of input DNA inserted following DNA extraction.

After quality control, these overlapping reads are stitched together computationally to form a continuous sequence. In order to identify where each read lies along the genome, reads are computationally aligned to the reference human genome. Ideally, reads would align perfectly - to one unique location in the genome, and with no discrepancy between the reference sequence and the read. Such algorithms must, however, account for imperfection - due to the numerous germline and somatic mutations in a cancer genome, and any possible errors or fluctuations in quality due to the sequencing process.

Following read alignment, it is possible to identify putative mutations in the sample relative to the reference genome. In this process, common, normal germline variants are typically taken into account and filtered out, leaving potentially cancer-related variants. Moreover, a subtractive algorithm can compare the constructed sequence from a patient's cancer compared to that of their non-cancer cells (typically their blood) in order to identify cancer-specific mutations. The mutations typically studied are those that occur at a small scale - SNVs or indels. These variants can then be further annotated by referring to manually curated databases. For example, a mutation in a tumour suppressor gene may be known to be inactivating, and can be labelled as such. In contrast, an incidental mutation in a gene that is not associated with cancer may be known to be a benign variant. These mutational catalogues are then ready for subsequent analysis.

1.4.3. Limitations

Although NGS provides, for the first time, a picture of the mutational landscape of many cancers, there are several reasons why that picture may be incomplete or flawed. For example, it has been reported that damage to the input DNA is a major source of sequencing errors (Chen et al. 2017). This has been shown to affect downstream interpretation, and it remains unclear whether input DNA should be repaired artificially prior to sequencing - and if so, to what extent.

Another factor to consider is that the genomic sequence obtained - of, say, a cancer - is actually a composite genome that is an average of many cells'

input DNA. This has been termed a metagenome, and some efforts have been made to deconvolute it (Andor et al. 2015). The mutations detected in the sequence are therefore biased towards those found in all cells. In the instance of cancer, this leads to a greater likelihood of detection of so-called 'founder mutations', present in the founding cancer cell. Mutations occurring in the first few cancer cells are also detected more readily. Indeed, sequencing of cells from different regions of the same tumour paints a different mutational picture, identifying mutations that appear to be specific to those regions (Gerlinger et al. 2012).

Aside from these biases, another factor to consider in considering cancer mutational data is what is considered as normal. The normal cell comparator is central in determining what is cancer-specific, and, as described above, is typically a blood sample. This leaves the possibility that organ-specific mutational processes may be misinterpreted as cancer-specific. For example, mutations detected in a breast cancer may merely be typical of mutations that accumulate in normal breast tissue. However, when compared to a blood sample as the 'normal' reference for that patient, they are deemed cancer-specific. Some efforts have been made to identify mutational processes in normal cells (Behjati et al. 2014). However, this information is, to date, largely incomplete.

The bioinformatic analysis of reads also has a number of potential pitfalls. One example is in how reads that map to highly repetitive regions are aligned. Most reads of this kind are typically discarded, as the algorithm cannot ascertain where they fall. If, therefore, a mutation falls within a highly

repetitive region, it may not be detected. It is also particularly challenging to identify the sequence of highly repetitive genomic regions, such as the centromere.

Moreover, the mathematical problem of mutation calling is still unsolved, and there is a lack of consensus on what methodology should be used. For example, the Cancer Genome Atlas (TCGA) uses four leading mutation calling algorithms in its bioinformatic pipelines, and all are available for analysis. This leads to further heterogeneity in results, and the potential for further computational artefacts.

Accounting for these limitations aids the interpretation of the output of NGS analysis, enabling the measurement of mutation across the genome. Researchers have been able to identify known germline and somatic mutations, and validate their findings of novel mutations using more traditional sequencing methods. With these large-scale pictures of cancer mutagenesis, it has been possible to identify patterns of mutation that point to underlying mutational processes.

1.5. Mutational signatures in cancer

The term 'mutational signature' refers to a pattern of mutations, often attributed to the characteristic activity of a specific mutational process (Alexandrov et al. 2013). The first mutational signature was identified in the 1960s, when it was found that ultraviolet light, a known carcinogen, acted as a mutagen that caused a characteristic pattern of CC>TT mutations (Alexandrov and Stratton 2014). The advent of Sanger sequencing, then NGS, has led to the identification of tens of mutational signatures in cancer, of both known and unknown aetiology (Alexandrov et al. 2020; Steele et al. 2022). Patterns have been identified for changes that occur at a range of mutation sizes, from the level of single-base substitutions to the level of whole genome doublings.

Mutational signatures have been identified in large, contemporary pan-cancer studies by analysing the output data produced by the next-generation sequencing of cancer genomes. As this data accumulated, it became evident that a typical cancer has around 1 somatic mutation per megabase, or around 3,000 mutations in total (Alexandrov et al. 2013). The majority of these mutations are point mutations (Campbell et al. 2017).

In order to identify patterns of mutation representing mutational signatures in these data, point mutations were first classified according to the specific base substitution that was detected. For example, sequencing might identify that a C was substituted for a T, A or G. Of note, this mutation can also be described as equivalent to a G being substituted for an A, T or C,

respectively, on the opposite strand. There are 6 possible types of point mutation that can be identified on this basis - taking all possible combinations into account, and the equivalence of mutations on opposite strands. These are described with reference to the pyrimidine base (C or T), as is convention in the field.

The 6 types of point mutation are:

- C>A
- C>G
- C>T
- T>A
- T>G
- T>C

This 6 type classification was then expanded further, as it is was understood that many mutational processes could be influenced by the sequence context of the mutated base (Friedberg et al. 2006; Alexandrov and Stratton 2014). The six substitution types were further subdivided depending on the base that was found immediately upstream and immediately downstream of the mutated base. For example, a mutated C in a T<u>C</u>A context was delineated as different from to one mutated in a G<u>C</u>G context. Classifying point mutations in this way yields an expanded framework consisting of 96 possible mutation types.

By applying these systems of classifying point mutations, the prevalence of each point mutation type could then be quantified, giving an initial indication

of patterns of mutations between cancers. This is illustrated by data from one of the first NGS studies of multiple cancer genomes (Nik-Zainal et al. 2012). Data from this study indicated that C>T mutations are typically the most common type of point mutation in the 21 breast cancers studied. In particular, C>T mutations that occur in an N<u>C</u>G (also known as CpG) context were the most common point mutations found, according to the 96 type classification. The data also indicate that a wide range of mutation types occur at varying frequencies between patients, a finding consistent with the presence of multiple mutational processes generating multiple mutational signatures in these cancers.

Mathematical methods in machine learning were applied to the 96 type classification data, with the aim of deconvoluting the signals contributed by these putative mutational signatures to the observed data. The algorithm used to identify mutational signatures is known as non-negative matrix factorisation (NMF). NMF is a statistical tool that is part of a family of statistical tools used for unsupervised factor analysis (Hastie et al. 2005). The role of tools such as these is to identify underlying factors in sets of data. 'Supervised' machine learning algorithms learn relationships between input and output data, whereas this is not the case for 'unsupervised' algorithms that instead identify patterns in input data. NMF differs from similar tools used in unsupervised factor analysis because of assumptions that are made in constructing the algorithm. As its name suggests, NMF assumes that the factors (which are mutational processes in the case at hand) are non-negative. In other words, it assumes that factors do not cancel each other out. NMF also assumes that factors will sum linearly to

produce the final output. These assumptions enable NMF to identify distinct features, conceptually equivalent to ‘parts’, that contribute to the observed data. A classical example that is used to illustrate this is that of feature recognition in facial images (Lee and Seung 1999). When given images of faces as input data, NMF identifies distinct facial features, while vector quantisation (VQ - a method equivalent to the commonly-used method of k-means clustering) instead identifies ‘subtypes' of faces and principal component analysis (PCA) forms abstract negative and non-negative features of faces that do not correspond to physical reality.

The pixels of these images are represented as numbers for processing by the NMF algorithm. The pixels of these facial image data can be considered equivalent to the pixels represented in the heatmaps displaying the prevalence of cancer mutations, such as those produced by Nik-Zainal and colleagues. Instead of facial features, applying NMF to cancer mutation heatmaps would be expected to yield ‘parts’ that correspond to mutational signatures. Indeed, when applied to data from the genomes of approximately 23,829 cancers across 32 cancer types in the PCAWG dataset, NMF identifies tens of single-base substitution signatures (Alexandrov et al. 2020). This has yielded signatures with distinct mutational profiles, with subsequent assessments of their proposed aetiologies (shown in Table 1.5) and their prevalence.

Single-base substitution signature	Proposed aetiology
SBS1	Deamination of 5-methylcytosine
SBS2	APOBEC activity
SBS3	Defective HR DNA repair; BRCA1/2 mutation
SBS4	Tobacco smoking
SBS5	-
SBS6	Defective DNA mismatch repair
SBS7a	Ultraviolet light exposure
SBS7b	Ultraviolet light exposure
SBS7c	Ultraviolet light exposure
SBS7d	Ultraviolet light exposure
SBS8	-
SBS9	In part, polymerase η activity
SBS10a	POLE mutation
SBS10b	POLE mutation
SBS11	Temozolomide treatment

SBS12	-
SBS13	APOBEC activity
SBS14	POLE mutation and mismatch repair deficiency
SBS15	Defective DNA mismatch repair
SBS16	-
SBS17a	-
SBS17b	-
SBS18	Reactive oxygen species
SBS19	-
SBS20	POLD1 mutation and mismatch repair deficiency
SBS21	Defective DNA mismatch repair
SBS22	Aristolochic acid exposure
SBS23	-
SBS24	Aflatoxin exposure
SBS25	Chemotherapy
SBS26	Defective DNA mismatch repair
SBS28	-
SBS29	Tobacco chewing

SBS30	Defective base excision repair; NTHL1 mutation
SBS31	Platinum treatment
SBS32	Azathioprine treatment
SBS33	-
SBS34	-
SBS35	Platinum treatment
SBS36	Defective base excision repair; MUTYH mutation
SBS37	-
SBS38	Indirect effect of ultraviolet light
SBS39	-
SBS40	-
SBS41	-
SBS42	Haloalkane exposure
SBS44	Defective DNA mismatch repair

Table 1.5 - Proposed aetiologies of single-base substitution signatures (as reported in Alexandrov et al. 2020).

The profiles of these signatures and their associated features indicate a consistency with prior knowledge in the field, lending weight to the validity

of the mutational signature discovery process. For example, several mutational signatures have features that are consistent with mutational processes that are known to operate in cancer.

One mutational signature, SBS1, is present in all cancers and consists of C>T mutations in a CpG context. Mutations of this kind are described to occur as result of the spontaneous, hydrolytic deamination of methylated cytosines in the aqueous cellular environment. CpG residues are the classical sites for genomic cytosine methylation. This deamination is understood to be a ubiquitous process that occurs progressively with age and, indeed, the prevalence of mutations attributable to SBS1 is correlated with increasing age.

Other mutational signatures demonstrate consistency with the known underlying aetiology of cancers they are detected in. For example, lung cancer demonstrates signatures with a predominance of C>A mutations that are known to be mediated by carcinogens in tobacco smoke. Similarly, melanomas are found to demonstrate CC>TT mutations that are characteristic of pyrimidine dimer formation following skin cells' exposure to ultraviolet light.

A subset of signatures are associated with identifiable DNA repair defects, such as SBS3 that is found in patients with inactivating mutations in BRCA1 or BRCA2. Other studies have shown that SBS3 and other BRCA-related signatures can be used to identify patients whose cancers have functional defects in homologous recombination, regardless of whether they have

detectable inactivations of BRCA1 or BRCA2. This supports the notion that the inferred mutational signatures are causally produced by specific mechanistic processes, as opposed to simply being statistically correlated to certain inactivating mutations in DNA repair genes, and highlights the potential for these patients to be treated with therapies that take advantage of homologous recombination deficiency such as PARP inhibitors (Davies et al. 2017).

Overall, these data provide, for the first time, what has been termed a 'repertoire' of mutational signatures in cancer derived from 'catalogues' of somatic mutations in cancer genomes. Many mutational signatures were not previously known to occur in cancer. These findings may therefore indicate the discovery of novel mutational processes that operate in cancer cells. Alternatively, they may not truly reflect the mutagenesis that occurs in cancer genomes, as a result of errors inherent to the methods used.

For example, as described in section 1.4.3, a number of errors arise in the sequencing process. Indeed, some of the newly-identified mutational signatures have been attributed to expected sequencing artefacts and subsequently excluded. Other signatures have been attributed to artefacts that are specific to the research centre where the cancer has been sequenced, highlighting the scope for both expected and unexpected artefacts to occur in the NGS process. In addition, the 'catalogues' of somatic mutations may not be entirely complete or accurate, and as a result NMF may be acting on biased input data that skews the patterns detected.

There may also be limitations that arise as a result of the NMF algorithm itself. For example, it may be possible that mutational processes might not be strictly non-negative, as assumed by the algorithm, and may in fact cancel each other out in some contexts. NMF will also generate different results if the precise underlying mathematics is varied, or if user-defined parameters such as the predicted number of signatures is changed (Alexandrov et al. 2020). Other limitations lie in how mutation types are defined for analysis by NMF - it may be the case that features other than a trinucleotide sequence context are relevant to how mutational processes operate in cancer cells. Examples of this are given in section 1.6.

Lastly, the approaches used may not meet the complexity required to fully reflect the biochemical complexity of the problem at hand. The mechanisms of mutagenesis are highly diverse, owing to the wide range of biochemical processes that can lead to DNA damage and the correspondingly wide range of DNA repair mechanisms that may act (or indeed fail) in cancer (Friedberg et al. 2006). As a result, at the biochemical level, a mutational signature produced by a given mutational process may readily be skewed by many of factors. It is known that some factors may, for example, be specific to the tissue of origin or subtype of cancer, which is not something that is accounted for in these methodologies. NMF may consequently more readily identify a mutational signature that is typical of a mutational process in these studies, rather than variants of that mutational signature (Degasperi et al. 2020).

A final point that is of note in interpreting these data is whether the signatures reported are specific to cancer cells. It is possible that some or all of these signatures are also found in normal cells. If certain signatures are found in both normal cells and cancer cells, it may be possible that they are similar or different in prevalence when comparing normal and cancer cells. The sequencing of normal tissues performed thus far appears to demonstrate that, with some exceptions, normal cells generally have a limited number of mutations from a small set of mutational signatures that are widespread in both normal genomes and cancer genomes (Abascal et al. 2021; Moore et al. 2021; Wang et al. 2023). This suggests that the mutational signatures detected in cancer genomes broadly differentiate cancer cells from normal cells and represent sources of genomic instability in cancer.

The first pan-cancer analysis of mutational signatures in 2013, using NMF to identify single-base substitution signatures, informed the work described in this thesis. Since then, the work has expanded to address some of the limitations of initial studies. The increase in cancer genomes available for analysis has been leveraged in the latest studies, which consequently report the discovery of new signatures and the subdivision of previously-identified signatures, including subdivision into 'components that may represent associated - but distinct - DNA damage, repair and/or replication mechanisms'. They also include analyses of doublet-base substitutions, quadruplet-base substitutions, indels and structural abnormalities (Alexandrov et al. 2020; Steele et al. 2022).

1.6. APOBEC activity in cancer

1.6.1. APOBEC signature mutations

SBS2 and SBS13 are thought to be caused by the mutagenic activity of the APOBEC enzyme family (described in section 1.6.2). Mutations attributed to SBS2 and SBS13 have consequently been termed 'APOBEC signature mutations'.

APOBEC signature mutations are characterised by cytosine mutations that occur in a TCW trinucleotide context (where W denotes A or T). SBS2 is characterised by C>T mutations that occur predominantly in this TCW context. In contrast, SBS13 is characterised by C>G mutations, and sometimes C>A mutations, that occur predominantly in this TCW context.

These patterns of mutagenesis are consistent with the known mutational activity of the APOBEC enzyme family. APOBEC enzymes are cytosine deaminases - their enzymatic activity mediates the conversion of cytosine to uracil in nucleic acid sequences. Although first described as mutators of cytosine residues in RNA, many are capable of acting on single-stranded DNA (ssDNA. Substrate preferences in Table 1.6). APOBECs convert cytosine to uracil in genomic ssDNA, leading to APOBEC signature mutations.

APOBEC activity is sensitive to the bases immediately 5' and 3' of the target cytosine, with different APOBECs displaying different trinucleotide

motif preferences (Conticello et al. 2007). The APOBEC signature detected in cancer genomes, of cytosine deamination at TCW motifs, is consistent with the motif preferences of two APOBEC family members, APOBEC3A and APOBEC3B, but not others (see section 1.6.2 and 1.6.3; Taylor et al. 2013).

The mechanism of how APOBEC activity is thought to lead to SBS2 and SBS13 is shown in Figure 1.2 (as described in Morganella et al. 2016).

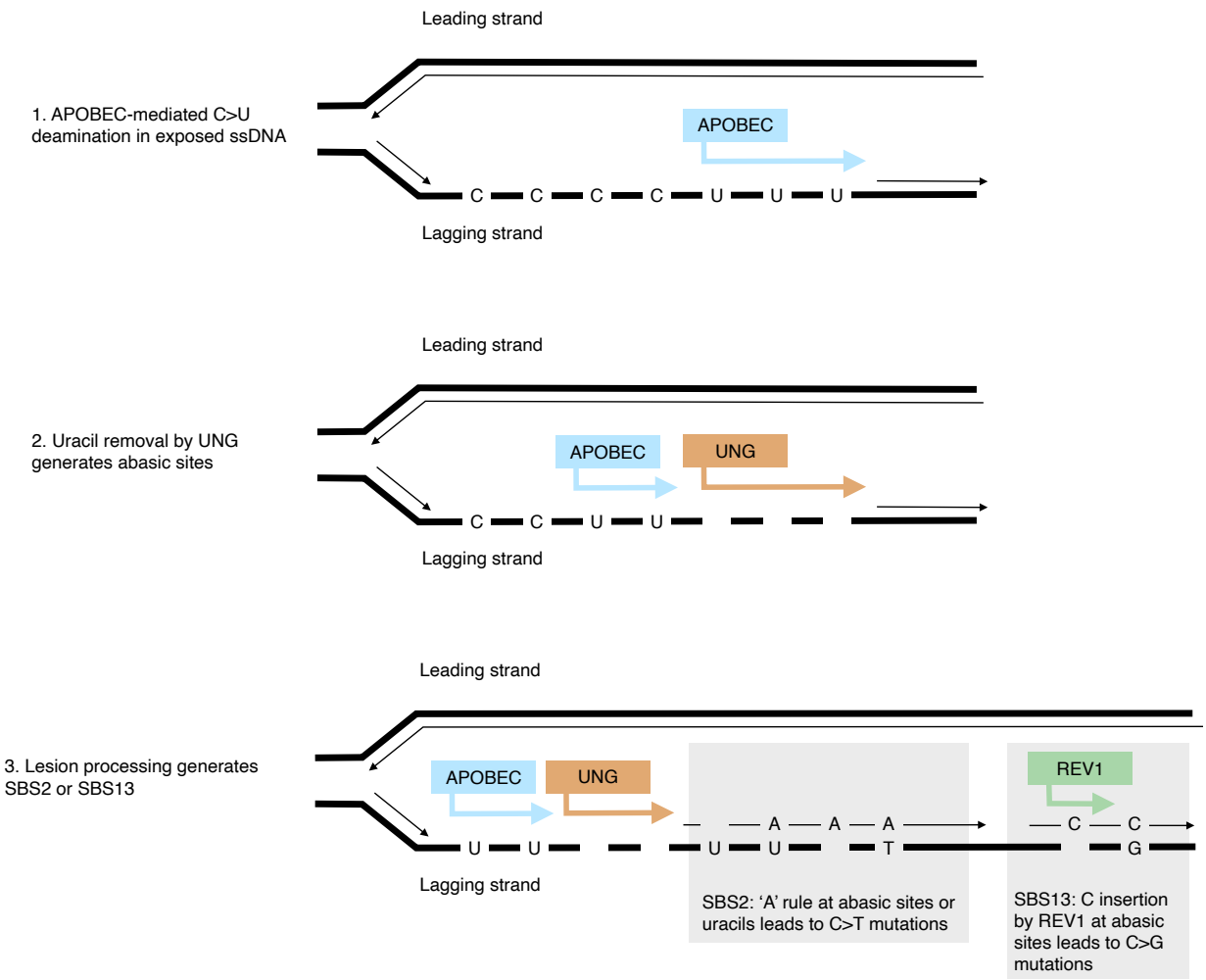


Figure 1.2 - Proposed mechanisms of SBS2 and SBS13 formation resulting from APOBEC cytosine deamination (adapted from Morganella et al. 2016).

The first step involves cytosine deamination in an exposed ssDNA strand, such as the lagging strand that is formed as part of the process of DNA replication. Cytosine deamination entails the hydrolytic cleavage of the amino group in a cytosine molecule, forming ammonia and uracil. Uracil is then removed from DNA by UNG, the uracil DNA glycosylase, which is a component of the base excision repair pathway that leaves only the carbon-

phosphate backbone of the DNA at that position. This lesion is known as an abasic site.

The C>T mutations that characterise SBS2 can be formed either before or after the uracil is excised. If the uracil is not yet excised, it will base pair with adenine. When this uracil is then eventually excised, a thymine will base pair with the adenine that has incorporated into the opposite strand. This therefore completes the process of forming a C>T mutation. Alternatively, a C>T mutation can be formed from an abasic site. Abasic sites are processed by error-prone translesion synthesis DNA polymerases that will frequently pair bases at low fidelity, leading to point mutations (Friedberg et al. 2006). These polymerases tend to preferentially incorporate adenines opposite abasic sites. As a result, thymine would then base pair with the adenine that has incorporated into the opposite strand as before, completing the process of forming a C>T mutation.

Polymerases will not, however, always incorporate adenines opposite abasic sites. For example, the polymerase REV1 is known to insert cytosines opposite abasic sites. As a result, a guanine will ultimately base pair with this cytosine, leading to the C>G mutations that characterise SBS13.

Abasic sites also frequently form strand breaks, a potential source of further mutagenesis and exposed ssDNA. Indeed, APOBEC signature mutations sometimes co-localise with strand breaks and structural abnormalities. A large number of APOBEC signature mutations can occur in a specific region

of the genome, a phenomenon of localised hypermutation known as 'kataegis' (Nik-Zainal et al. 2012). Regions of kataegis co-localise with regions of genomic rearrangements.

The wide range of mutations that might occur as a result of APOBEC mutagenesis support the notion that APOBEC activity might result in mutational processes other than SBS2 and SBS13. In keeping with this, PCAWG data indicate that the number mutations attributed to SBS2 and SBS13 are positively correlated to the number of mutations from all other SBSs (Spearman's $\rho = 0.57$) and that SBS2 and SBS13 are associated with many fold increases in the number of indels and structural variants, putatively through the generation of strand breaks that occur as a result of cytosine deamination (Jakobsdottir et al. 2022).

APOBEC signature mutations are common. They are found in over 70% of cancers (Alexandrov et al. 2020). The only mutational signatures that are found more commonly are those that are ubiquitously detected in all normal cells and cancer cells (SBS1 and the related signatures SBS5 and SBS40). APOBEC signature mutations often contribute to a substantial fraction of the total burden of point mutations in the cancers in which they are detected. For example, around 75% of point mutations in cervix cancer exomes are APOBEC signature mutations, the highest of any cancer type (Alexandrov et al. 2013).

APOBEC signature mutations include common driver mutations that mediate the pathogenesis of many cancer types. For example, APOBEC

signature mutations are found in frequently-mutated TP53 hotspots, and APOBEC3B overexpression in cultured cells has been found to induce these inactivating mutations in TP53 (Burns et al. 2013). APOBEC signature mutations are also found in frequently-mutated PIK3CA helical domain hotspot sites - specifically in cancers that have APOBEC signature mutations throughout their genome, but not in those that do not (Henderson et al. 2014). More broadly, pan-cancer studies associating mutational signatures and driver mutations indicate that APOBEC mutagenesis may be responsible for around a quarter of reported associations (Temko et al. 2018).

Increased APOBEC mutagenesis is also associated with resistance to therapy. For example, bladder cancers with a high number of APOBEC signature mutations were found to be more likely to resist treatment with chemotherapy in patients (Faltas et al. 2016). In addition, high APOBEC3B expression was associated with reduced response to the breast cancer therapy tamoxifen, both in a mouse xenograft model and in patients with the disease (Law et al. 2016). APOBEC activity is also associated with poor response to immunotherapy in lung cancer (McGranahan et al. 2016). In particular, it is though that the genetic heterogeneity mediated by APOBEC in cancers such as lung cancer (de Bruin et al. 2014) might mediate resistance to immunotherapy.

1.6.2. The APOBEC enzyme family

'APOBEC' stands for apolipoprotein B mRNA-editing enzyme catalytic polypeptide-like. This name is derived from the discovery of the function of the first APOBEC, APOBEC1, which edits the mRNA sequence of APOB, the apolipoprotein B gene in cells of the small intestine (Teng, Burant, and Davidson 1993; Hirano et al. 1997). In humans, the APOBEC family contains 11 evolutionarily-related genes (Table 1.6).

Name	Genomic location	Substrate	Cellular localisation
AID	12p13	DNA	Mainly cytoplasmic
APOBEC1	12p13.1	DNA, RNA	Cytoplasmic, nuclear
APOBEC2	6p21	Unkown	Cytoplasmic, nuclear
APOBEC3A	22q13.1	DNA, RNA	Cytoplasmic, nuclear
APOBEC3B	22q13.1	DNA	Mainly nuclear
APOBEC3C	22q13.1	DNA	Cytoplasmic, nuclear
APOBEC3DE	22q13.1	DNA	Cytoplasmic
APOBEC3F	22q13.1	DNA	Cytoplasmic
APOBEC3G	22q13.1	DNA, RNA	Cytoplasmic
APOBEC3H	22q13.1	DNA	Cytoplasmic, nuclear

APOBEC4	1q25.3	Unkown	Unkown

Table 1.6 - Human APOBEC genes, their genomic locations, nucleic acid substrates and cellular localisations (Conticello 2008; Pecori et al. 2022).

APOBECs function as enzymes that deaminate cytosine to form uracil in DNA and RNA substrates, using zinc ions to catalyse the reaction in a core cytosine deaminase domain that is found in all APOBEC family members (Pecori et al. 2022). APOBEC3B, APOBEC3DE, APOBEC3F and APOBEC3G contain two cytosine deaminase domains, while all others have one. Aside from their catalytic domains, all APOBEC proteins also contain cofactor interacting sequences that determine their interactions with substrates. Their sequences also determine their localisation, which in turn also contributes to determining the types of substrates available to each APOBEC. These subcellular localisations, as detailed in Table 1.6, can be influenced by a number of factors. For example, AID and APOBEC1 have both a nuclear localisation and nuclear export sequences that mediate their transit between the nucleus and cytoplasm. APOBEC2, APOBEC3A, APOBEC3C and APOBEC3H passively diffuse between the two compartments. The APOBECs with two cytosine deaminase domains, - APOBEC3B, APOBEC3DE, APOBEC3F and APOBEC3G - are too large to diffuse between compartments. As a result, they are restricted to the cytoplasm, apart from APOBEC3B which is nuclear as a result of its nuclear localisation sequence.

APOBECs are thought to have evolved from the prokaryotic zinc-coordinating tRNA deaminase known as Tad (Figure 1.3, Conticello et al. 2005).

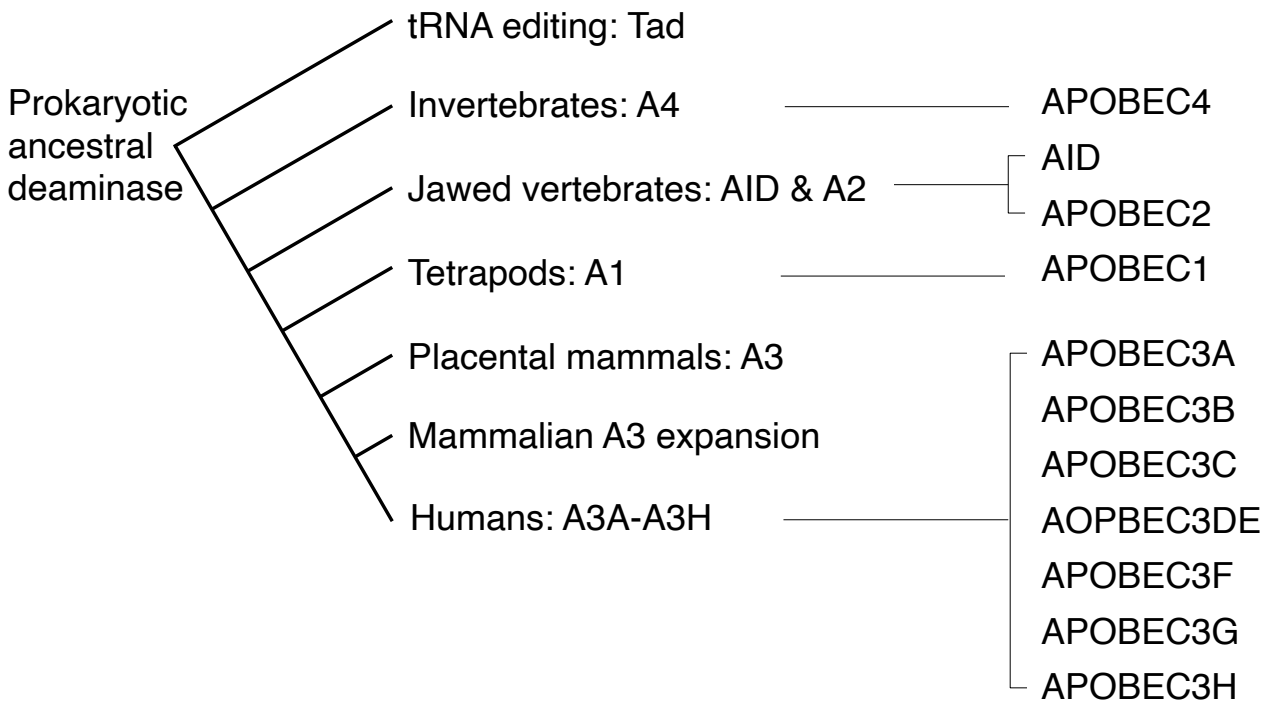


Figure 1.3 - Simplified phylogenetic tree of the APOBEC family (adapted from Pecori et al. 2022)

This prokaryotic RNA deaminase is thought to have been co-opted to perform DNA deamination functions. Interspecies bioinformatic analyses of APOBEC sequences indicate that current APOBEC family members arose through duplications and fusions of ancestral APOBEC genes throughout the process of evolution from prokaryotes to humans. APOBEC4 is the only APOBEC with orthologues found in invertebrates, which suggests that it predates all other APOBECs (Krishnan et al. 2018). AID and APOBEC2 are

found in early vertebrates lineages (Conticello 2008). APOBEC1 and a single APOBEC3 gene emerged through duplications of AID, in tetrapods and placental mammals, respectively (Hayward et al. 2018; Krishnan et al. 2018). APOBEC3 genes then expanded in many mammals. There are seven APOBEC3 paralogues in humans, six in horses, four in cats, two or three in pigs and one in mice (Münk, Willemsen, and Bravo 2012).

The seven human APOBEC3 genes are found immediately next to each other, in order, in a locus in chromosome 22. The APOBEC3 locus has been found to be under positive selection, displaying some of the highest signals of positive selection in the human genome, as determined by the ratio of the rate of non-synonymous mutations (dN, which change the amino acid sequence of a protein) and the rate of synonymous mutations (dS) in these genes (Sawyer, Emerman, and Malik 2004). These dN/dS data, and the recent expansion of APOBEC3s in mammalian lineages, are in keeping with the described function of APOBEC3s as immune genes, since such genes may rapidly evolve to counteract the rapid evolution of pathogens, so are often found to be under positive selection.

APOBEC3s mutate the cytosines found in the DNA of viruses and transposons, leading to both physical and informational degradation of their sequences (Yang et al. 2007). The physiological role of APOBEC3s in multiple species appears to be the restriction of viruses and transposons, with evidence of counteracting mechanisms to evade APOBEC-mediated restriction (Bogerd et al. 2006; Harris and Dudley 2015). For example, APOBEC3G potently mutates the genome of HIV, leading to its inactivation

(Mangeat et al. 2003). To counteract this, HIV expresses the VIF protein which leads to the proteasome-mediated degradation of APOBEC3G, thereby circumventing its antiviral effects (Yu et al. 2003). Although APOBEC3s can restrict both viruses and transposons, it is thought that the expansion of the APOBEC3 locus in humans was driven primarily by increased germline retrotransposition (Ito, Gifford, and Sato 2020; Uriu et al. 2021). It is possible that APOBEC3 activity in cancer might be due to the aberrant activation of the innate immune pathways used to counteract viruses and transposons (see section 1.7.1).

In terms of the physiological function of other APOBECs, AID (activation-induced deaminase) also plays a role in immunity. The canonical role of AID is mutating cytosine residues in immunoglobulin genes, to generate antibody diversity in the process of somatic hypermutation in immunologically-activated B lymphocytes (Muramatsu et al. 2000). The physiological role of APOBEC1 is in RNA editing, as described above. APOBEC2 and APOBEC4 do not appear to have functional cytosine deaminase activity despite retaining their ability to bind DNA. Their physiological functions are unclear. However, APOBEC2 is expressed in cardiac and skeletal muscle, where APOBEC2 knockout models indicate that it plays a role in preventing myopathy (Pecori et al. 2022).

1.6.3. APOBEC3A and APOBEC3B as mediators of APOBEC signature mutations

APOBEC enzymes preferentially deaminate cytosines depending on the sequence context those cytosines occur in. AID prefers WRC, APOBEC1 prefers WC, APOBEC3C prefers TYC, APOBEC3G prefers CCC, while the others prefer TC (W = A or T, R = A or G, Y = C or T; Conticello et al. 2007). APOBEC3A and APOBEC3B have a strong preference (>90%) for a 5' T adjacent to the target C and also show a preference for a 3' W. This pattern matches the pattern found in cancer genomes, and does so much more closely than for other TC-preferring APOBECs (Taylor et al. 2013). APOBEC3A and APOBEC3B were consequently highlighted as likely mediators of APOBEC signature mutations in cancer genomes.

Other lines of evidence support the notion that APOBEC3A and APOBEC3B are likely mediators. For example, they are among the APOBEC family members that can access the cell nucleus, where they are able to act on genomic DNA (Lackey et al. 2013). In addition, pan-cancer analyses have found that the RNA expression levels of both APOBEC3A and APOBEC3B are positively correlated to the number of APOBEC signature mutations, with APOBEC3B showing a stronger positive correlation (Spearman's $\rho = 0.16$ for APOBEC3A and 0.30 for APOBEC3B; Roberts et al. 2013). In these studies, the expression of other APOBECs show no correlation, or weaker correlations, to the number of APOBEC signature mutations. The fact that APOBEC3B showed the strongest positive correlation indicated that it might be a likelier candidate as a

mediator of APOBEC signature mutations in cancer, although it is of note that the coefficient of 0.30 indicates that this correlation is in itself not strong.

RNA-seq data of paired tumour-normal samples also implicated APOBEC3B more than APOBEC3A (Roberts et al. 2013). These data indicated that the expression of both APOBEC3A and APOBEC3B is very low in normal tissues. However, APOBEC3B appears to be overexpressed in a wide range of cancer samples, while the extent of APOBEC3A expression in cancer appears more limited. In addition, APOBEC3B, but not APOBEC3A, is widely expressed in cancer cell lines (Kinomoto et al. 2007; Burns et al. 2013) and APOBEC3A expression is usually restricted to cells of the myeloid lineage (Refsland et al. 2010).

Although these RNA-level data indicate that APOBEC3B is likely to play a leading role, they do not conclusively determine that this is the case. For instance, the relative RNA expression of APOBEC3A and APOBEC3B may not indicate their relative mutational activity, not least since APOBEC3A is a more potent cytosine deaminase enzyme following its translation (Caval et al. 2014; Cortez et al. 2019). In addition, RNA-seq can be understood as providing a 'snapshot' of gene expression when the sample was collected, while DNA sequencing is reflecting the entire mutational history of the cancer and its progenitors. APOBEC3A expression may be transient and the detection of APOBEC3B expression may depend on cell cycle stage, with analyses of single cell sequencing data indicating that their expression may also vary according to tumour cell type (Ng et al. 2019; Oh et al. 2021).

A single gene expression timepoint might therefore be unlikely to explain or represent the history of mutagenesis of a cancer genome.

In keeping with this, DNA sequencing data suggested APOBEC3A might instead play a more important role than APOBEC3B. Overexpression of APOBEC3A and APOBEC3B in yeast cells, followed by genomic NGS, indicated that the motif preferences enzymes might be differentiated by the base that comes immediately before the T<u>C</u>W motif (Chan et al. 2015). APOBEC3A was described to prefer YT<u>C</u>W and APOBEC3B preferred RT<u>C</u>W, although it was not clear whether human cells might yield the same results, owing to possible differences in processing APOBEC-mediated lesions. However, these data were then supported by progress in mutational signature analyses as applied to human cancers. NMF was applied to the PCAWG dataset including the two bases upstream and two bases downstream of a point mutation, expanding the 96 type classification based on three nucleotides to a 1,536 type classification based on 5 nucleotides. This yielded the splitting of SBS2 and SBS13 into signatures with YT<u>C</u>W or RT<u>C</u>W motifs, consistent with the activity of APOBEC3A or APOBEC3B, respectively (Alexandrov et al. 2020). These data indicate that APOBEC3A-like mutations are more commonly found in cancer than APOBEC3B-like mutations. In addition, deletion of APOBEC3A in cultured cancer cells is reported to lead to a greater reduction in the number of new APOBEC signature mutations than deletion of APOBE3B (Petljak et al. 2022).

Other findings indicate that APOBEC3B may not be required for APOBEC signature mutations in cancer. A common germline polymorphism in which a region including the APOBEC3B coding sequence is deleted does not prevent the formation of APOBEC signature mutations in cancer (Nik-Zainal et al. 2014). Indeed, this polymorphism is associated with an increased number of APOBEC signature mutations. The APOBEC3B coding sequence deletion giving rise to a fusion between APOBEC3A and the APOBEC3B 3'UTR. The aberrant relocation of its 3'UTR notwithstanding, this indicates that APOBEC3B is not required for APOBEC mutagenesis in cancer.

Lastly, although these data collectively indicate that APOBEC3A and APOBEC3B mediate APOBEC signature mutations, other APOBECs have also sometimes been implicated. For example, AID, which is normally expressed in B cells, may contribute to the mutation of B cell lymphoma genomes, and APOBEC1 may similarly contribute to mutagenesis in its own physiological context of the small intestine (Pettersen et al. 2015; Wang et al. 2023). In keeping with this, deletion of both APOBEC3A and APOBEC3B from cancer cell lines substantially reduces but does not eliminate the formation of APOBEC signature mutations, suggesting the potential for contributions from other APOBEC family members (Petljak et al. 2022). However, the literature to date nonetheless appears to suggest that most APOBEC signature mutations in cancer are likely to be mediated by APOBEC3A and APOBEC3B.

At the time of preparing the experiments presented in this thesis, it was not yet clear if APOBEC3A or APOBEC3B might act as the main source of APOBEC signature mutations in cancer, or if both might contribute, perhaps to varying degrees in different contexts. APOBEC3B appeared the likelier candidate at the time, owing to its higher expression level, stronger positive correlation with APOBEC signature mutations, the uncertainty around whether APOBEC3A and APOBEC3B might indeed have different motif preferences in humans as opposed to yeast, and the uncertainty around whether the APOBEC3B deletion polymorphism might subject APOBEC3A to unusual regulation through its fusion with the APOBEC3B 3'UTR. As a result, the work in this thesis typically focuses on APOBEC3B in the first instance.

1.7. Mechanisms of APOBEC regulation in cancer

1.7.1. Pathways regulating APOBEC expression in normal cells and cancer cells

As described in section 1.6, APOBEC signature mutations are commonly found in human cancer and likely due to the activity of APOBEC3A or APOBEC3B. APOBEC3s have been described as immune genes involved in the restriction of viruses and transposons. Recurrent driver mutations in APOBEC3 genes or their regulatory regions have not been identified in large pan-cancer analyses, suggesting that APOBEC activity is mediated through the alteration of upstream pathways (Priestley et al. 2019; Elliott and Larsson 2021). These mechanisms leading to their apparent activity in cancer are not yet fully understood.

Mechanistic work in the field of virology has helped elucidate the pathways that regulate APOBEC3 gene activity in normal cells. APOBEC3s are interferon-stimulated genes whose expression increases during states of infection (Stavrou and Ross 2015). A range of molecules, including viral proteins and nucleic acids, trigger this response through a number of cellular sensors including TLR3, TLR4, TLR7, AIM2 and cGAS. Activation of these sensors leads to immune signalling involving IRFs (interferon response factors) that are thought to subsequently trigger APOBEC3 expression through type-I interferons. Not all APOBEC3 genes appear to be recruited in these antiviral responses. For example, infection by HIV is characterised by expression of and mutagenesis by APOBEC3G, with less

significant contributions from APOBEC3D and APOBEC3F (Gillick et al. 2013).

Some cancers have a viral cause that likely leads to APOBEC signature mutations. For example, cervix cancers, which are caused by human papillomavirus (HPV), have high numbers of APOBEC signature mutations, as previously described. Here, HPV is described to lead to APOBEC3B upregulation in a mechanism that requires the inhibition of p53 by the HVP E6 oncoprotein (Vieira et al. 2014). However, in contrast, liver cancers caused by hepatitis viruses do not appear to contain APOBEC signature mutations, despite the fact that APOBEC3s can restrict these viruses (Janahi and McGarvey 2013).

Moreover, most cancers have no known viral aetiology, suggesting that there are other processes that drive APOBEC activity in the majority of cancers. A number of possible causes have been identified. These include NF-κB, which has been described to bind to the to the APOBEC3B promoter to increase APOBEC3B expression following replication stress or exposure to a range of chemotherapy drugs (Periyasamy et al. 2021; Butler and Banday 2023). p53 has also been described to inhibit APOBEC3B expression through recruitment of the repressive DREAM complex to the APOBEC3B promoter (Periyasamy et al. 2017). In addition, nucleic acids derived from the host cell, such as those generated by transposon activity, have been suggested as possible drivers for the induction of APOBEC activity in cancer cells through action on cellular nucleic acid sensors (Petljak and Maciejowski 2020).

At the time of preparing the work in this thesis, the possible mechanistic drivers of APOBEC activity in the majority of cancers were less clear. In terms of regulatory factors, the canonical role of APOBECs was most apparent, as interferon response genes that were thought to have evolved to combat viruses and transposons. A particular focus of the work in this thesis was consequently the possibility that non-viral nucleic acids, such as those derived from transposons, might cause APOBEC activity in cancer.

1.7.2. LINE-1 activity as a possible cause for APOBEC activity in cancer

Around 50% of the human genome is made up of transposon sequences, with roughly 20% of the genome consisting of LINE-1 (long interspersed nuclear element 1) retrotransposon sequences and roughly 30% of the genome consisting of other types of transposons (Kemp and Longworth 2015). LINE-1 elements are part of an ancient family of LINE elements that are found in the genomes of eukaryotes (Ivancevic et al. 2016). Full-length LINE-1 sequences are approximately 6 kb in length (Dombroski et al. 1991). They contain a 5' promoter and two open reading frames - ORF1, which encodes an RNA-binding protein, and ORF2, which encodes a protein with endonuclease and reverse transcriptase functions (Swergold 1990; Mathias et al. 1991; Feng et al. 1996; Hohjoh and Singer 1997). LINE-1 elements also contain an antisense open reading frame, ORF0, whose function is unclear (Denli et al. 2015).

LINE-1 elements are thought to be the only autonomously active transposons in human cells, with around 150 copies - all belonging to the subset of copies specific to *Homo sapiens* termed L1Hs - capable of propagating in the genome (Sassaman et al. 1997; Penzkofer et al. 2016). Other transposons found in the human genome appear to depend on the activity of LINE-1 elements to facilitate their propagation (Dewannieux, Esnault, and Heidmann 2003; Hancks et al. 2011).

As retrotransposons, LINE-1 elements propagate by reverse transcription, which generates new copies through a 'copy and paste' mechanism. This involves transcription of LINE-1 RNA (‘copying’), translation of this RNA into LINE-1 protein, then the formation of a LINE-1 ribonucleoprotein complex that mediates reverse transcription of the RNA into a new genomic location (‘pasting’). LINE-1 elements insert new copies of themselves into the genome in a process known as ‘target-primed reverse transcription’ (Figure 1.4; Cordaux and Batzer 2009). LINE-1 inserts are often truncated at the 5’ end or, in contrast, can involve transduction of downstream sequences due to transcription that goes beyond the 5’ end of the LINE-1 element.

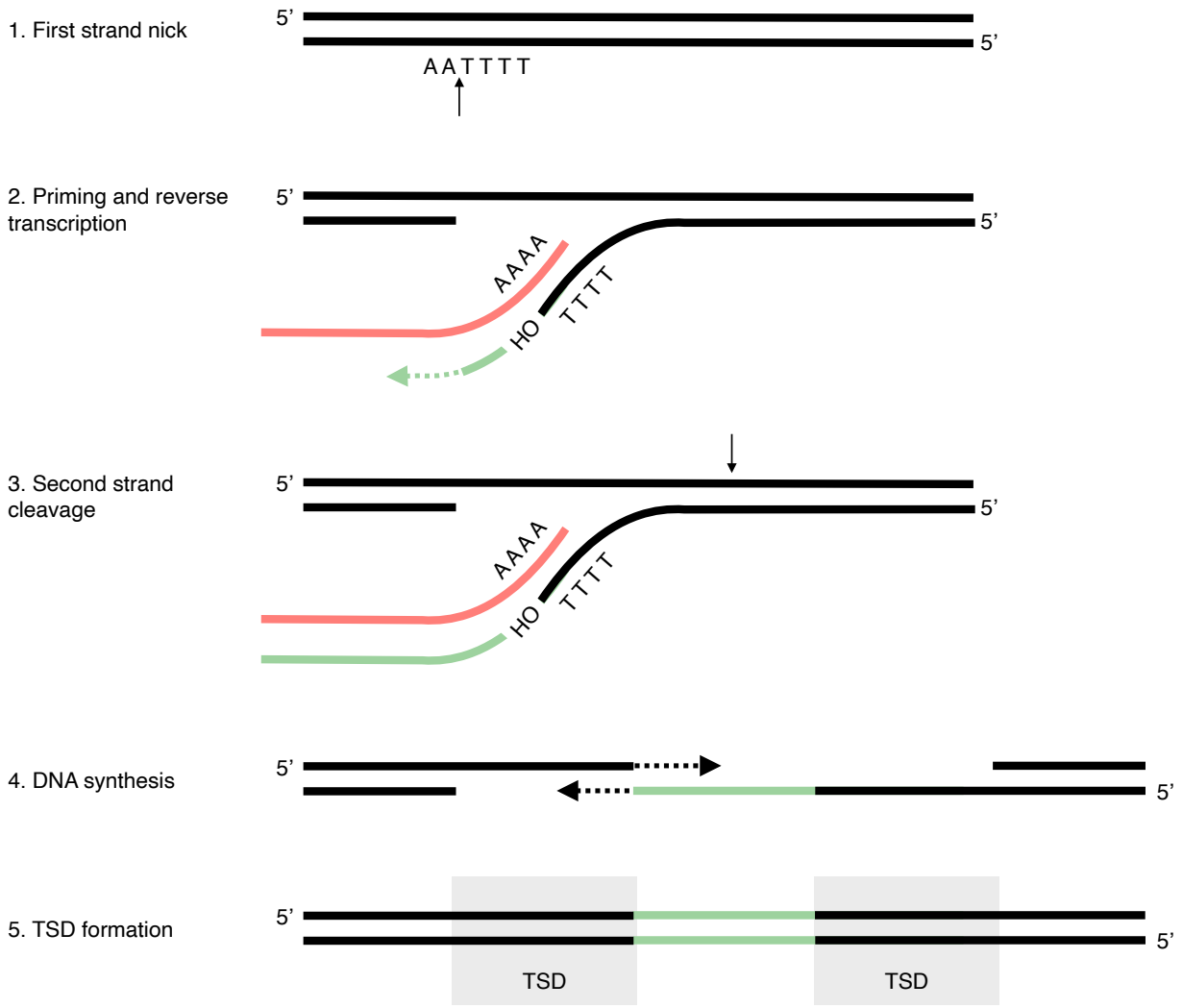


Figure 1.4 - Mechanistic stages of LINE-1 target-primed reverse transcription (adaoted from Cordaux and Batzer 2009). TSD: target site duplication.

First, the ORF2 endonuclease generates a nick in one strand of DNA, with a preference for AA/TTTT target sequences (Jurka 1997). ORF2 then uses the free end of the ssDNA generated to prime the reverse transcription of its RNA, forming LINE-1 cDNA. After this, the second strand of DNA is cleaved by ORF2, leading to a 'sticky-ended' double-strand break with ssDNA

overhangs. This double-strand break is then repaired by cellular DNA repair factors. During the repair process, DNA is synthesised to fill in gaps on opposite strands, leading to LINE-1 dsDNA formation and the formation of target site duplications.

The sticky-ended double-strand breaks that LINE-1 generates are thought to be mutagenic, since they expose DNA to additional damage, as well repair pathways that re-ligate DNA ends in an error-prone manner (Gasior, Roy-Engel, and Deininger 2008; Venkitaraman 2014), with the potential for the generation of point mutations, indels and structural abnormalities. In addition, LINE-1 is capable of causing insertional mutagenesis through the reverse transcription of its own RNA, as well as other RNAs present in the cell (Kemp and Longworth 2015). LINE-1 can therefore threaten genomic integrity by generating DSBs, inserting DNA into functional loci and increasing the copy number of expressed genes.

Many mechanisms exist to inhibit LINE-1 activity in somatic cells at various stages in its cycle of retrotransposition. For example, LINE-1 appears to be subject to a substantial degree of inhibition at the level of its transcription. The LINE-1 promoter contains many CpG sites that are typically highly methylated by host cells, leading to epigenetic silencing (Hata and Sakaki 1997), with further epigenetic silencing occurring through associated histone modifications and heterochromatin formation (Garcia-Perez et al. 2010). Multiple factors are also described to inhibit LINE-1 activity at the post-transcriptional level. These include degradation of LINE-1 RNA through RNA interference involving miR-128 (Hamdorf et al. 2015) and by

the interferon-inducible RNase L (Zhang et al. 2014). Protein complexes including ADAR1, MOV10, SAMHD1 and ZAP proteins are involved in the dissociation of LINE-1 RNA from ORF1, sequestration of LINE-1 RNA in stress granules, and autophagy-mediated degradation of LINE-1 transcripts (Protasova, Andreeva, and Rogaev 2021). LINE-1 RNA degradation products have been described to bind and stimulate RIG-I and MDA5 RNA sensors, leading to type I interferon production (Zhao et al. 2018). Interferon-inducible factors such as APOBEC3 deaminases and the TREX1 exonuclease also interact with LINE-1 ribonucleoproteins, where their inhibitory effects can be mediated through enzymatic or non-enzymatic activity (Li et al. 2017; Orecchini et al. 2018). Other processes of LINE-1 inhibition are active in the germline, such as RNA degradation mediated through the piRNA/PIWI pathway (Pezic et al. 2014).

Studies comparing LINE-1 expression in normal cells and cancer cells indicate that LINE-1 is normally transcriptionally repressed, but that these processes of repression commonly fail in cancer (Menendez, Benigno, and McDonald 2004; Tubio et al. 2014). A pan-cancer immunohistochemical study of around 1000 patients appears to be illustrative, estimating that LINE-1 derepression occurs in around half of all cancers, with little or no LINE-1 expression detected in normal tissues (Rodić et al. 2014). APOBEC3s are among many factors that are used by cells to restrict LINE-1 activity at the post-transcriptional level. It is therefore plausible that uncontrolled LINE-1 expression in cancer might consequently trigger APOBEC activity.

In keeping with the reported increases in LINE-1 expression in many cancers, there is evidence that LINE-1 retrotransposition is a common event in many cancers. PCAWG analyses indicate that around half of all cancers contain somatic insertions of LINE-1 elements, which also contribute to a wide range of small and large scale structural variation, such as LINE-1 mediated deletions that occur as a result of incomplete LINE-1 integration (Rodriguez-Martin et al. 2020). LINE-1 activity can be a significant source of mutagenesis in certain cancers - for example, it is reported that in oesophageal adenocarcinomas, LINE-1 insertions are the most common type of structural variation found.

Although LINE-1 mediated mutagenesis of cancer genomes is found to be common, there is evidence to suggest that findings such as these could represent an underestimate of the full impact of LINE-1 activity. For example, methods of detecting new insertions in genomic NGS data are known to limited by difficulties in resolving repetitive elements such as LINE-1, owing to limitations in the process of aligning sequencing reads with ambiguous genomic origins (Treangen and Salzberg 2012; Ewing 2015). In addition, it is not clear how the impact of LINE-1 mediated double-strand breaks, rather than completed insertions, might be measured using NGS data. This may be of relevance given the observation that LINE-1 generates many more double-strand breaks than new insertions when overexpressed in cultured cells (Gasior et al. 2006).

1.7.3. LINE-1 inhibition as a possible physiological role for APOBEC3B

In keeping with their role as inhibitors of transposons, several APOBECs have been shown to inhibit LINE-1 activity, typically in assays where they are overexpressed alongside LINE-1 (Muckenfuss et al. 2006; Kinomoto et al. 2007; Conticello 2008). In cell culture, overexpression of AID, APOBEC1, APOBEC3A, AOBEC3B, APOBEC3C, APOBEC3DE and APOBEC3F leads to LINE-1 inhibition in a manner that does not require their cytosine deaminase activity, despite evidence of APOBEC signature mutations in genomic LINE-1 sequences (Orecchini et al. 2018).

APOBEC enzymes are known to be recruited in a specific physiological contexts. For example, as previously described, AID is recruited in the context of B cell somatic hypermutation, and APOBEC3G is typically recruited in the context of HIV infection. It is not yet known if one or more APOBECs might be specifically recruited for the process of LINE-1 inhibition. There is, however, evidence to suggest that endogenous APOBEC3B might be recruited as a physiological defence against LINE-1, while other APOBECs are not.

For example, a study by Marchetto and colleagues looked to find differences in the RNA-seq profiles of human and non-human primate induced pluripotent stem cells (iPSCs). APOBEC3B was among the top differentially expressed genes. The expression of APOBEC3B was around 30-fold higher in human iPSCs, while the expression of other APOBECs

was broadly similar. This increase in APOBEC3B expression was associated with a substantial reduction in LINE-1 activity in human iPSCs when LINE-1 was overexpressed, relative to non-human primate iPSCs. LINE-1 activity could be inhibited by APOBEC3B overexpression, but not by APOBEC3G. These findings were also associated with reduced LINE-1 activity in the human germline since the evolutionary divergence from non-human primates. Together, these data suggest that APOBEC3B, but not other APOBECs, has played the physiological role of inhibiting LINE-1 activity in human evolution.

Knockdown studies of different APOBECs in cell culture also support the notion that APOBEC3B may act to inhibit LINE-1 in physiological contexts. These data suggest that knockdown of APOBEC3B increases LINE-1 activity, while knockdown of other APOBECs (A3C, A3DE, A3F or A3G) does not (Wissing et al. 2011). In these studies, APOBEC3A expression was undetectable, so could not be knocked down.

It is possible that the elevated APOBEC3B expression observed in cancer might therefore reflect cells attempting to recruit APOBEC3B as a response to LINE-1 expression. There is some evidence from studies of cancer to suggest that this hypothesis is plausible. For example, many cancers with evidence of LINE-1 activity also have evidence of APOBEC activity (Alexandrov et al. 2020; Rodriguez-Martin et al. 2020). However, this may simply reflect the fact that both phenomena are common, and there are examples of cancers where both do not co-occur. The loss of p53 function, another common feature of cancer, has been associated with LINE-1

derepression in fruit flies, zebrafish, mice and humans, suggesting an evolutionarily conserved mechanism (Wylie et al. 2016). p53 inactivation is also associated with APOBEC3B upregulation (Burns et al. 2013; Periyasamy et al. 2017). Few studies have specifically examined the relationship between APOBEC activity and LINE-1 activity in cancer. However, one study has found that APOBEC mutagenesis in cultured cancer cells is episodic (Petljak et al. 2019). It was noted that this is consistent with the activity of transposons, that also become active episodically. The authors find a positive correlation between LINE-1 transposition events and APOBEC activity in cell lines, but no significant correlation when this is extended to an analysis of patient cancers.

1.7.4. Aicardi-Goutières syndrome as a possible model for studying APOBEC activity

Aicardi-Goutières syndrome (AGS) is a rare, severe paediatric genetic condition, initially characterised in the 1980s as a disease of neuroinflammation involving white matter disease, intracranial calcification and cerebrospinal fluid lymphocytosis (Aicardi and Goutières 1984). It is now understood that AGS causes systemic autoinflammation, with patients also developing skin lesions, lupus-like disease, glaucoma and hypothyroidism (Crow, Shetty, and Livingston 2020). The disease presents *in utero*, where it mimics a congenital viral infection in the absence of a causative virus (so-called 'sterile inflammation') and has high rates of death or disability.

A number of germline genetic alterations have been associated with Aicardi-Goutières syndrome (Table 1.7). These are thought to cause chronic activation of type I interferon that may be central to the pathophysiology of the disease. Indeed, AGS is among a number of conditions that were termed 'type I interferonopathies' in the early 2010s (Crow 2011).

Gene	Protein function	Proposed link to type I interferon signalling	Mutation effect
TREX1	DNase	Cytosolic DNA	LOF (autosomal recessive or dominant negative)
SAMHD1	Control of dNTP pool	Cytosolic DNA	LOF (autosomal recessive)
RNASEH2A	RNase	Cytosolic RNA–DNA hybrids	LOF (autosomal recessive)
RNASEH2B	RNase	Cytosolic RNA–DNA hybrids	LOF (autosomal recessive)
RNASEH2C	RNase	Cytosolic RNA–DNA hybrids	LOF (autosomal recessive)

ADAR1	RNA editing	Cytosolic dsRNA	LOF (autosomal recessive or dominant negative)
IFIH1	dsRNA sensor	Cytosolic dsRNA	GOF (autosomal dominant)
LSM 11	RDH pre-mRNA processing	Histone stoichiometry/ genomic DNA	LOF (autosomal recessive)
RNU7-1	RDH pre-mRNA processing	Histone stoichiometry/ genomic DNA	LOF (autosomal recessive)

Table 1.7 - Genotypes implicated in Aicardi-Goutières syndrome (adapted from Crow and Stetson 2022).
dsRNA: double-stranded RNA, RDH: replication-dependent histone, LOF: loss-of-function, GOF: gain-of-function.

The genes implicated in Aicardi-Goutières syndrome are those whose products are involved in nucleic acid metabolism or nucleic acid sensing. This is a theme among type I interferonopathies, where pathways involved in the immune response to 'non-self' nucleic acids, such as those

originating from viruses, are active in the absence of such stimuli (Crow and Stetson 2022).

Nine genes are implicated in AGS. LSM11 and RNU-7 are the most recently identified genes. They are described to promote the physical separation of genomic DNA and the DNA sensor cGAS by maintaining histone stoichiometry (Uggenti et al. 2020).

ADAR1 encodes an adenosine deaminase involved in RNA editing. IFIH1 encodes the cellular sensor of dsRNA known as MDA5. It has been suggested that these two genes work in tandem in the cellular response to sequences from the inactive Alu retrotransposon, which make up around 10% of the human genome. ADAR1 is described to deaminate adenosine residues in transcribed Alu sequences to mark them as 'self' (Chung et al. 2018). Failure of this process leads to MDA5 activation, and gain-of-function IFIH1 mutations found in AGS are described to make MDA5 molecules hypersensitive to cellular Alu RNAs (Ahmad et al. 2018).

TREX1 encodes an exonuclease that degrades ssDNA and dsDNA. SAMHD1 encodes a dNTPase that is thought to inhibit the synthesis of cytosolic DNA. RNASEH2A, RNASEH2B and RNASEH2C encode subunits of the RNase H2 complex, which degrades RNA in RNA-DNA hybrids. For these genotypes, nucleic acids arising from DNA damage or from retrotransposon activity have been posited as possible stimuli that trigger sterile inflammation (Crow and Manel 2015).

In terms of evidence supporting the role of DNA damage as a source of aberrant nucleic acids in these genotypes, TREX1 knockout cells have been described to display chronic activation of ATM-mediated DNA damage checkpoint signalling, with cytosolic ssDNAs generated during S phase that could represent replication intermediates (Yang, Lindahl, and Barnes 2007). SAMHD1 deficiency is also associated with chronic activation of DNA damage signalling, with evidence of replication stress (Kretschmer et al. 2015). In addition, RNase2 deficiency is described to lead to replication stress, genomic instability and the formation of micronuclei, as it is involved in the processing of RNA-DNA hybrids formed during DNA synthesis (Pizzi et al. 2015; Mackenzie et al. 2017). However, LINE-1 activity alone has the potential to cause DNA damage signalling, cytosolic ssDNA accumulation, replication stress, genomic instability and micronuclei (Section 1.7.2, McKerrow et al. 2022).

TREX1, SAMHD1 and RNase H2 are described to be inhibitors of LINE-1 activity, as well as ADAR1 which is also an AGS gene (section 1.7.2). The defects in their activity found in AGS have been reported to lead to the accumulation of LINE-1 RNA and cDNA (Stetson et al. 2008; Pokatayev et al. 2016; Herrmann et al. 2018). SAMHD1 and TREX1 are thought to limit the number LINE-1 cDNAs, while RNaseH2 is thought to limit the number of LINE-1 RNAs. In their absence, accumulated LINE1 ssDNA, dsDNA and RNA-DNA hybrids can activate the cGAS/STING/IRF3 pathway, leading to type I interferon activation. Although RNase H2 can aid LINE-1 retrotransposition by removing RNA during target-primed reverse transcription (Benitez-Guijarro et al. 2018) it is likely that the accumulation

of LINE-1 RNA-DNA hybrids in states of RNase H2 deficiency are immunostimulatory (Uggenti et al. 2020).

Since LINE-1 encodes a reverse transcriptase, it is susceptible to inhibition by nucleoside reverse transcriptase inhibitors (RTIs), such as those used in treating HIV (Dai, Huang, and Boeke 2011). RTIs have consequently been used to test the hypothesis that retrotransposition contributes to the pathophysiology of AGS, and to test the hypothesis that AGS could be treated with RTIs.

Results from TREX1-deficient mouse models have given conflicting results in these regards. Treatment with azidothymidine (AZT), a classical nucleoside RTI and the first RTI developed for HIV treatment, was not able to alleviate a TREX1-deficient mouse model of AGS (Stetson et al. 2008). In addition, while one study found that treatment with tenofovir, emtricitabine and nevirapine RTIs successfully alleviated the lethal cardiomyopathy that occurs in these mice, another found that these drugs had no such effect (Beck-Engeser, Eilat, and Wabl 2011; Achleitner et al. 2017).

Unlike the conflicting results reported for this mouse model, studies in human cells and clinical trials indicated that LINE-1 activity could be inhibited by RTIs to alleviate AGS pathophysiology. These differences may be in part due to the fact that the mouse genome has multiple active retrotransposons aside from LINE-1 that may also be expressed in different tissues (Crow and Manel 2015). TREX1-deficient human cells are reported to accumulate LINE-1 ssDNA in the cytosol, leading to interferon-associated

toxicity that could be blocked with the nucleoside RTIs 3TC and d4T (Thomas et al. 2017). In addition, treatment with the nucleoside RTIs abacavir, 3TC and AZT in a clinical trial involving AGS patients with germline TREX1, SAMHD1 and RNASEH2A-C mutations was reported to result in reduced interferon signalling and improved neurovascular function, indicating alleviation of the disease process (Rice et al. 2018). These data support the notion that LINE-1 activity drives AGS pathophysiology, and that RTIs can be used to treat AGS. It may be the case that these human studies could better represent what is true for human AGS than the studies based on mouse models that are more equivocal in their results. It is possible that RTIs could also inhibit inflammation through a known off-target effect of inhibiting inflammasome activity (Fowler et al. 2014). However, a role for inflammasome activity in AGS has not been established (Crow and Manel 2015).

AGS therefore mimics viral infection in the absence of a causative virus, and may be triggered by endogenous nucleic acids such as those deriving from LINE-1 elements. This is reminiscent of what is observed in cancer, where there is an apparent activation of APOBEC3s in the absence of viral infection, with LINE-1 activity as a putative cause. It was consequently reasoned that LINE-1 associated Aicardi-Goutières syndrome genotypes could be used as a possible model to study mechanisms of APOBEC regulation in this thesis. Of note, patients with Aicardi-Goutières syndrome are not thought to have an increased propensity for developing cancer. However, AGS patients are reported to die at a young age, typically before

early adulthood (Aicardi and Goutières 2000) and therefore may not have the opportunity to accumulate the mutations required for carcinogenesis.

1.8. The multiple functions of p53

TP53 is the most frequently mutated gene in cancer (PCAWG 2020). As described in section 1.2 and section 1.7, mutations in TP53 canonically lead to impairments in cell cycle arrest and apoptosis in response to DNA damage (Weinberg 2007) and are also associated with LINE-1 upregulation (Wylie et al. 2016) and APOBEC3B upregulation (Periyasamy et al. 2017).

p53 was first identified as a 53 kDa protein that binds to the large T antigen of simian virus 40 (Lane and Crawford 1979). The p53 protein acts as a tetramer of four identical subunits (Jeffrey, Gorina, and Pavletich 1995) which each contain a DNA-binding domain that is used to enact its function as a transcriptional regulator (Laptenko and Prives 2006). The vast majority of TP53 mutations that occur in cancer are found in the DNA-binding domain.

Early work on p53 showed that its overexpression led to the suppression of cellular growth and transformation (Finlay, Hinds, and Levine 1989). It was found that p53 protein levels are maintained at low levels by regulators that include MDM2, which promotes p53 degradation (Haupt et al. 1997) and that p53 levels are stabilised in response to a stresses including DNA damage and oncogene activity in part through MDM2 inhibition (Haupt et al. 1997; Pomerantz et al. 1998). p53 enforces reversible cell cycle arrest in the G1 phase by transcriptionally activating the gene encoding p21 (Harper et al. 1993) stably arrests the cell cycle by activating senescence in tandem

with Rb (Shay, Pereira-Smith, and Wright 1991) and activates apoptotic signalling by inducing the activity of BCL-2 genes (Miyashita et al. 1994).

These functions reflect p53's canonical role as the 'guardian of the genome' which acts to hinder growth and transformation in cells with oncogenic signalling or potentially mutagenic DNA damage (Lane 1992). In keeping with this, other work has identified its role in maintaining chromosomal stability by promoting the fidelity of the G2/M transition (Vitre and Cleveland 2012) and in promoting DNA repair signalling (Williams and Schumacher 2016). Several non-canonical functions of p53 have also been described, which include the regulation of metabolism, autophagy, cellular development, tissue remodelling and responses to reactive oxygen species (Kastenhuber and Lowe 2017). Non-canonical p53 functions may occur in a context-dependent manner.

2. Thesis aims

The overall aim of the work presented in this thesis is to investigate the regulation of APOBEC mutagenesis in cancer. As previously discussed, aberrant signalling appears likely to drive aberrant APOBEC3A and/or APOBEC3B expression in cancer, leading to APOBEC signature mutations. The thesis explores possible drivers of cancer-associated APOBEC activity, with a particular focus on LINE-1 activity as a possible driver and APOBEC3B as a possible mediator.

The work in first results chapter describes cell culture experiments that look to identify models and measures for LINE-1 and APOBEC activity, as well as testing whether RTI treatment could modulate APOBEC activity in cancer cells. The second results chapter looks to identify possible regulators of cancer-associated APOBEC activity, primarily by examining NGS data from patient cancers. The work in the third and final results chapter looks to obtain similar insights, drawing primarily on NGS data from patients with Aicardi-Goutières syndrome.

These experiments were performed according to the overarching rationale summarised in Figure 2.1. It was reasoned that TP53 inactivation might lead to LINE-1 activity, which in turn could lead to cancer-associated APOBEC activity, perhaps through AGS-like signalling. RTIs inhibit LINE-1, which under this frame of reference would lead to reduced genomic instability both by directly reducing metagenesis by LINE-1 and indirectly

reducing mutagenesis from APOBEC activity, with consequences for cancer evolution.

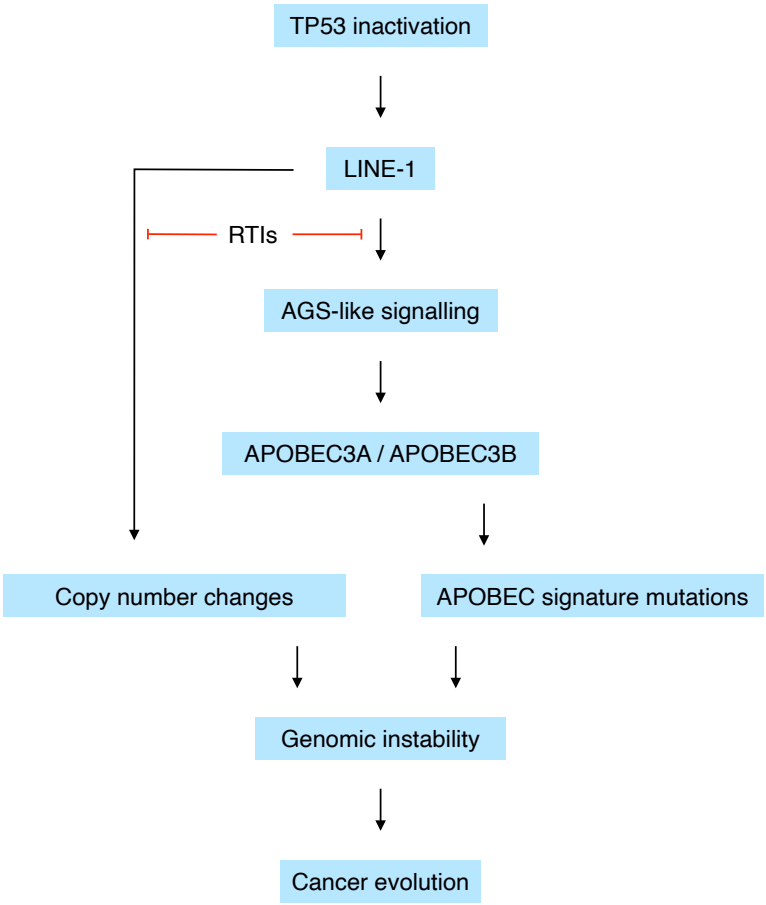


Figure 2.1 - Schematic diagram illustrating the overarching rationale for experiments described in the thesis.

3. Results

3.1. Results Chapter 1: Experiments investigating APOBEC activity in cultured cancer cells

3.1.1. Introduction

The specific focus of this first results chapter is to test the hypothesis that LINE-1 activity promotes APOBEC activity in cancer. The use of cell biology experiments in this chapter was intended to complement results gained from bioinformatic analyses. The bioinformatic analyses described in the second and third results chapters provide data that is primarily descriptive and correlative in nature. In contrast, the experimental methods described in this chapter allow, in principle, for the identification of causal relationships between cellular factors - such as the putatively causal relationship between LINE-1 activity and APOBEC activity - to be ascertained from the perturbation of specific cellular mechanisms. It was reasoned that descriptive genomic data from patient samples might provide a more accurate overall summary of APOBEC biology *in vivo*, while experiments on cultured cancer cells would allow for the direct testing of suggested associations, albeit in a model system that might deviate from the true disease state in several ways.

A number of methodological considerations were reviewed in the process of designing the experiments described in this chapter. Efforts were made to attempt to reflect the reality of the disease state as accurately as possible and test the hypothesis at hand as reliably as possible, taking any practical or methodological constraints into account. The factors considered will be discussed in turn in order to provide the rationale that informed and constrained the experiments conducted. These factors were:

- what cellular material to use for these experiments
- how best to measure LINE-1 activity
- how best to measure APOBEC activity, and
- how best to perturb the cellular system in order to test the hypothesis that LINE-1 activity promotes APOBEC activity.

3.1.1.1. Choice of cellular material

The first factor considered was what cellular material should be used. Cancer cell lines appear to be used widely in the field. Indeed, experiments using cancer cell lines have contributed to many seminal papers in cancer and in other fields. However, they do not seem to be completely reliable model systems. They have been described to show continuous genomic instability in culture (Petljak et al. 2019), which may distort their phenotypes beyond the degree of distortion caused by the level of mutagenesis observed *in vivo.* They are also subjected to non-physiological selection pressures; such cells are typically grown as a monolayer in plastic dishes in the absence of supporting stromal tissue. Factors such as these are

thought to contribute to the limited reproducibility observed for studies on cancer cell lines (Liu et al. 2019) and the fact that findings on possible therapeutic strategies *in vitro* rarely translate to efficacious treatments in clinical settings (Hingorani et al. 2019).

Given the potential disadvantages of using standard cancer cell lines, other types of cellular material were considered for the experiments described in this chapter. These were the use of three-dimensional organoid culture, the use of co-culture with stromal cells, the use of animal models, and the use of human trials. However, these were ultimately not pursued. The use of organoids and stromal cells was not pursued as these methods had not been previously established in our laboratory, and there was no clear suggestion from the literature that these factors might meaningfully contribute to LINE-1 and APOBEC regulation in cancer. Animal models and human trials were not pursued owing the the ethical and practical limitations of setting up such studies, and that such studies are typically preceded by in *vitro* data that justify such *in vivo* investigation. In addition, animal models of LINE-1 and APOBEC regulation are limited by the fact that their genetic loci remain under selection and vary significantly between (and indeed within) species. For example, the commonly used animal model, *Mus muluscus*, has only one APOBEC3 gene in place of the seven human APOBEC3s, and has multiple active LINE-1 element families of differing genetic sequences (Sookdeo et al. 2013) in contrast to the single active LINE-1 family specific to *Homo sapiens*. It is therefore unclear to what extent conclusions derived from a mouse model, including one that could be partly humanised by genetic modification, may yield results that are

applicable to the hypothesis at hand pertaining to human carcinogenesis - particularly when other mechanisms of LINE-1 repression might also differ significantly. As a result, cultured cancer cells were ultimately chosen for these experiments despite the limitations that have been described for their use.

Another consideration was how to use this model system to best approximate the LINE-1 and APOBEC activation that is observed in cancer *in vivo*. One approach to this would be to culture cancer cells and matched normal cells from the same donor tissue - for example, breast cancer cells and adjacent normal breast cells from the same patient. However, practical considerations meant that this approach was not pursued - these materials were not readily available, and non-cancer cells are typically resistant to continual growth *in vitro*. Instead, activation of LINE-1 and APOBEC was pursued by studying the effect of TP53 deficiency in otherwise isogenic cancer cells. This approach was pursued as a review of the literature indicated that p53 deficiency might be a cause of LINE-1 and APOBEC activation in cancer, in keeping with p53's function of counteracting many mediators of genomic instability (described in the Thesis introduction). The relationship between p53 inactivation and LINE-1 activity was suggested by data including the finding that LINE-1 elements contain multiple p53 binding sites (Harris et al. 2009) and that LINE-1 activity induces DNA damage that leads to a p53 response (Haoudi et al. 2004). Similarly, it was shown by Burns et al. that TP53 mutations in TCGA breast cancers were associated with elevated APOBEC3B expression, a correlation that suggested that a causative relationship between p53 inactivation and APOBEC activity might

be found if tested experimentally. However, at the time of conducting these experiments, no studies to our knowledge had demonstrated that p53 inactivation would lead to the activation of LINE-1 or APOBEC3B. This has since been reported in the literature (see sections 1.7.1, 1.7.3 and 3.1.6).

3.1.1.2. Measure of LINE-1 activity

Another factor considered in the design of these experiments was how best to measure LINE-1 activity. LINE-1 replication is a multi-step process, and it is possible that one or multiple LINE-1 replication steps could act as a signal that triggers a putative host cell response involving APOBEC activity. Given the abundance of LINE-1 elements in the genome, it was reasoned that a signal that might trigger such a response would likely be specific to LINE-1 elements that were no longer quiescent. The features of non-quiescent LINE-1 elements include DNA demethylation, RNA expression, translation, LINE-1 nucleoprotein formation, cDNA synthesis and integration. A measure of LINE-1 activity that could trigger APOBEC activity might therefore include one or more of these features.

Both DNA demethylation and integration were deemed unsuitable as measures of LINE-1 activity. DNA demethylation was not chosen as it was reasoned that demethylated LINE-1 elements may not necessarily be expressed or otherwise active. In addition, LINE-1 elements contain a significant fraction of CpG sites in the genome, such that LINE-1 methylation is often used as a surrogate for global methylation in epigenetic

studies (Vryer and Saffery 2017). Measuring LINE-1 methylation might therefore reflect global epigenetic change, and global, non-specific changes to gene expression could confound interpretation of results in this type of analysis.

LINE-1 integration was also not chosen. LINE-1 integration occurs when the host cell fails to repress LINE-1 activity at every stage of repression. On considering this further, and on review of the literature, it was not clear what results from cells with high numbers of *de novo* LINE-1 integration events indicated about the dynamics of LINE-1 activity in these cells, particularly in relation to defences against LINE-1. It was not clear if these cells might have weaker defences against LINE-1 (for example, having weaker APOBEC expression) or if such cells are overwhelmed by LINE-1 activity despite a maximal repressive response - or if indeed both are true in different contexts. As a result, LINE-1 integration was deemed a potentially unreliable measure for testing the hypothesis that LINE-1 activity promotes APOBEC activity.

The intermediate steps of LINE-1 mobilisation were therefore subsequently considered. Antibodies for LINE-1 remain in active development (Sharma et al. 2016; Ardeljan et al. 2020) with no commercially available antibody in widespread use at the time of conducting these experiments. Detection of protein levels of LINE-1 were consequently also deemed unsuitable, as a method for identifying LINE-1 protein levels appeared not to be firmly established in the field. cDNA levels were also considered - however, measuring LINE-1 cDNA is complicated by the high number of genomic

copies of LINE-1 that could contaminate the cDNA sample and falsely elevate cDNA quantification. Ultimately, quantification of LINE-1 mRNA by qRT-PCR was chosen as the measure of LINE-1 activity in these experiments given the disadvantages of the other options described above, and primers specific to L1Hs were chosen in an attempt to selectively measure the expression of potentially active elements (as described in Marchetto et al. 2013).

3.1.1.3. Measure of APOBEC activity

On considering how best to measure APOBEC activity, it was reasoned that directly measuring the APOBEC signature observed in cancer genomes would likely be the most representative measure that could be obtained in cultured cancer cells *in vitro.* However, this was not pursued in these experiments for several reasons. These were that it was not clear whether cells would need to be grown and treated for a prolonged period of time in order for mechanistic changes to yield measurable mutations, multiple individual clones would need to be grown out to identify new mutations that had occurred in the cell culture, and the sequencing required for such experiments would be relatively costly.

Given that APOBEC expression is associated with APOBEC activity, measures of APOBEC expression levels were also considered as surrogates of APOBEC activity. The high sequence homology of the APOBEC genes leads to challenges in measuring them at the protein and

RNA level. Primers that are highly specific to each APOBEC3 gene have been developed (Refsland et al. 2010), while specific antibodies are still under development. For example, a monoclonal antibody with improved specificity showed cross-reactivity between A3A, A3B and A3G, and was described in late 2019 (Brown et al. 2019). It was consequently decided that measurement by qRT-PCR would again be preferable to measurement by Western blot. On considering which genes to measure, it was reasoned that it might be more practical to focus on a single APOBEC gene, and thus APOBEC3B was chosen for study rather than APOBEC3A. This is because of its place as a likely mediator of the APOBEC signature, with RNA expression that is more clearly elevated and more highly correlated with the number of APOBEC signature mutations than APOBEC3A (Thesis Introduction; Burns et al. 2013; Roberts et al. 2013). APOBEC3B RNA expression appears therefore to be a marker of APOBEC activity in cancer, notwithstanding the likely role of APOBEC3B as an enzymatic mediator of genomic cytosine deamination.

In addition to measuring APOBEC3B RNA expression, efforts were made to quantify the biochemical activity of APOBEC enzymes in cell extracts. This is detailed as part of the results given in this chapter, where a cytosine deaminase assay is developed for this purpose. It was reasoned that measuring the APOBEC-mediated cytosine deaminase activity of cell extracts might complement and expand upon RNA expression data by approximating both measures of protein expression as measured by Western blot and the APOBEC mutational signature as measured by NGS. It was also reasoned that the deaminase assay would reflect APOBEC

protein abundance regardless of whether or not inhibition of LINE-1 was mediated via deaminase-dependent or deaminase-independent mechanisms.

3.1.1.4. Method of perturbation

The final factor considered was what experimental interventions could be used to test the hypothesis at hand. If, as hypothesised, LINE-1 activity does promote APOBEC activity, then methods of modulating LINE-1 activity can be used to elicit a concordant modulation of APOBEC activity. Methods for either increasing or decreasing LINE-1 activity were therefore considered. It was reasoned that these artificial interventions should ideally closely resemble the LINE-1 deregulation that occurs naturally in cancer, in order to avoid conditions that would be more likely to yield unrepresentative results. In addition, attempts were made to avoid experimental interventions that appeared to have a high risk of affecting APOBEC activity directly, rather than indirectly though their effect on LINE-1 activity.

A review of the literature indicated that interventions that could be considered for increasing LINE-1 activity were the use of DNA methyltransferase inhibitors, LINE-1 overexpression plasmids and p53 inactivation. Although LINE-1 elements are silenced by methylation, DNA methyltransferase inhibitors were not used owing to concerns that inhibiting all cellular methylation could affect APOBEC activity by the derepression of genes other than LINE-1, including the APOBEC genes themselves.

Although LINE-1 overexpression could have been suitable, p53 inactivation was ultimately selected the preferred method for increasing LINE-1 activity. This was because LINE-1 overexpression using a transfected plasmid would likely lead to LINE-1 expression levels that were higher than those observed in cancer, and could elicit significant LINE-1 mediated DNA damage that might plausibly limit the cell's recruitment of additional mutagens such as the APOBEC family. Given the purported link between p53 inactivation and LINE-1 activity, and the commonality of p53 inactivation in cancer, it was reasoned that p53 inactivation might represent the best method of increasing LINE-1 activity from the options considered, in terms of recapitulating disease pathophysiology, despite the additional functions of p53 (reviewed in Kastenhuber and Lowe 2017) that could also be disrupted.

The interventions considered for decreasing LINE-1 activity were p53 activation, RNA interference and the use of reverse transcriptase inhibitors. It was reasoned that p53 activation might yield representative results for the reasons described above, and could be readily achieved by the use of the MDM2 inhibitor Nutlin, which stabilises cellular p53 (Vassilev et al. 2004). This strategy was consequently pursued in the experiments described in this chapter. In contrast, RNA interference was not pursued as siRNAs against LINE-1 would, in principle, lead to genome-wide hypermethylation that might affect APOBEC expression indirectly. In support for this concern, it has been reported, for example, that siRNAs targeting LINE-1 typically lead to a 2-3 fold increase in global DNA methylation across a range of cell types (Ohms and Rangasamy 2014). Reverse transcriptase inhibitors were

also chosen for use in these experiments. RTIs have been used as a life-long therapeutics in patients with HIV, with a relatively favourable side effect profile. Their side effect profile may be due to the relatively low concentrations of RTIs required to inhibit HIV and the apparent paucity of physiological roles for reverse transcriptases in human cells. Instead, a significant burden of side effects are thought to be mediated through off-target inhibition of mitochondrial polymerases (Lewis, Day, and Copeland 2003). Nucleoside reverse transcriptase inhibitors inhibit LINE-1 activity at concentrations comparable to those used to treat HIV, with an IC_{50} for the canonical RTI azidothymidine (AZT) on the order of 1 nM (Dai, Huang, and Boeke 2011). In contrast, telomerase reverse transcriptase is inhibited at much higher concentrations, with an IC_{50} for AZT on the order of 100 μM (Hukezalie et al. 2012). RTIs were therefore selected as relatively selective inhibitors of LINE-1. It was reasoned that if RTIs were found to modulate LINE-1 and APOBEC mutagenesis *in vitro*, then they might be more readily indicated as possible therapeutics for modulating mutagenesis in cancer.

3.1.1.5. Summary

To summarise, the aim of the experiments described in this chapter are to test the hypothesis that LINE-1 activity promotes APOBEC activity in cancer. Techniques in cell biology and biochemistry, involving the use of cultured cancer cells, were chosen to conduct these experiments. LINE-1 activity was measured by qRT-PCR and APOBEC activity was measured by qRT-PCR of APOBEC3B and the use of a cytosine deaminase assay. p53

inactivation was chosen both to model LINE-1 and APOBEC derepression and as a method of elevating LINE-1 activity. Attempts to inhibit LINE-1 activity by p53 activation via Nultin and RTIs were also pursued. The results of these experiments, as well as the rationale underpinning their iterative progression and modification, are given in the sections that follow.

3.1.2. Investigating the effect of p53 activation and p53 inactivation on LINE-1 expression and APOBEC3B expression in the HCT116 cancer cell line

The introduction to this results chapter describes that cultured cancer cells were chosen as the system in which to test the hypothesis at hand, and that p53 inactivation was chosen as a putative method of recapitulating the LINE-1 and APOBEC3B derepression observed in cancer. A cell line that might be well-suited for p53 inactivation was therefore sought for these experiments. The HCT116 colorectal cancer cell line was chosen because a well-characterised p53 homozygous null version of the line was available. The pair of cell lines include HCT116 $p53^{wt/wt}$ cells and their otherwise isogenic $p53^{-/-}$ counterparts. This isogenic pair of cell lines was first generated by Vogelstein and colleagues by replacing the first codon of *TP53* with its second intron, for use in some of the earliest studies of p53 function, yielding seminal findings (Bunz et al. 1998). These cells were then made available for widespread use by other researchers in the field. Our laboratory is among those that have made use of these cells. We have validated their status and used them for a number of years to investigate mechanisms of DNA damage and repair (Hattori et al. 2014). A figure produced in collaboration with a postdoctoral colleague in our laboratory, Dr Amy Emery, validating the p53 status of the batch of cells used for these experiments by Western blot, is given below (Figure 3.1.1). The experiment was performed by Dr Emery during a period of restricted working in the Covid-19 pandemic.

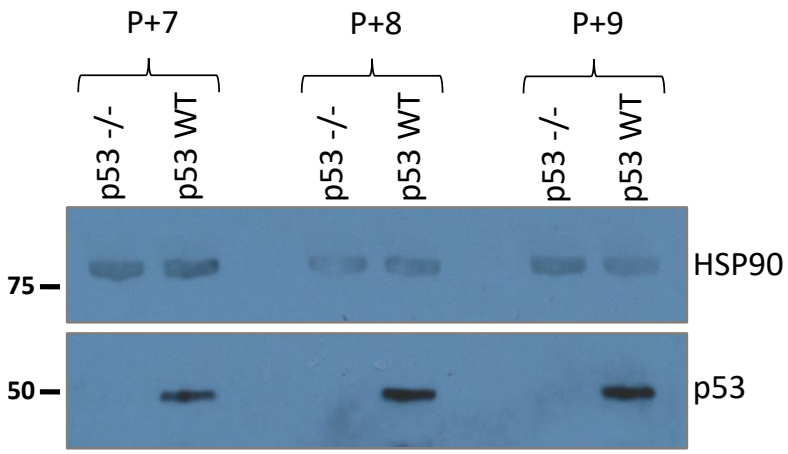


Figure 3.1.1 - p53 protein expression is present in HCT116 p53$^{wt/wt}$ cells and absent in HCT116 p53$^{-/-}$ cells, as determined by Western blot of three serial passages of these cell lines. HSP90 protein expression is measured as a loading control.

The first experiment conducted looked to ascertain whether mRNA expression of LINE-1 and APOBEC3B, as determined by qRT-PCR, was elevated in HCT116 p53$^{-/-}$ cells relative to HCT116 p53$^{wt/wt}$ cells. LINE-1 qRT-PCR primers specific to L1Hs targeting ORF1 and ORF2, as well as qRT-PCR primers specific to APOBEC3B were identified from the literature and used in these experiments (Refsland et al. 2010; Marchetto et al. 2013). The results are given below (Figure 3.1.2).

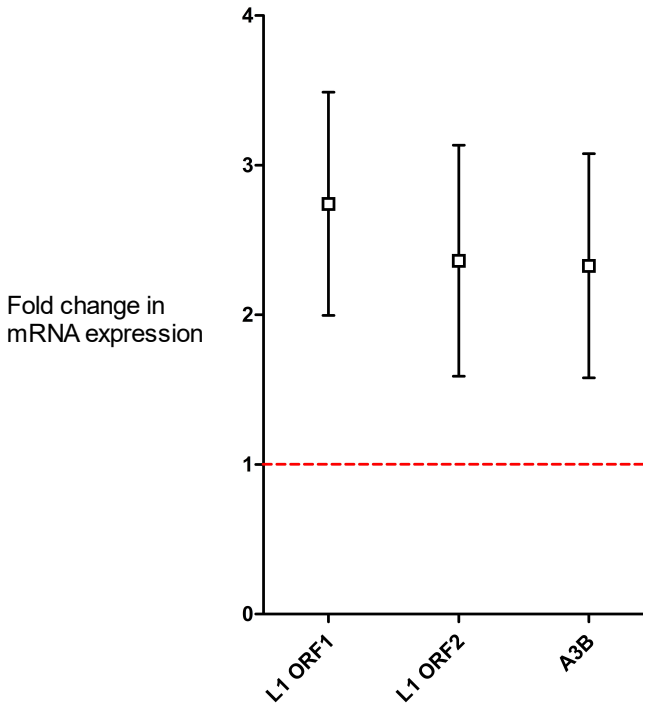


Figure 3.1.2 - p53 loss is associated with a two-fold to three-fold increase in LINE-1 and APOBEC3B mRNA levels in HCT116 cells. Expression levels are normalised to ACTB mRNA expression and given as a fold change in HCT116 $p53^{-/-}$ cells relative to HCT116 $p53^{wt/wt}$ cells. Mean ± standard error of the mean. Three experimental repeats. Dashed line: fold change value of 1. t-test p-values for ORF1 = 0.0196, ORF2 = 0.0414, APOBEC3B = 0.0460.

The results in Figure 3.1.2 indicate that p53 inactivation in HCT116 cells does indeed lead to an increase in LINE-1 and APOBEC3B mRNA expression. The two-fold to three-fold increase observed in these cells *in vitro* is in keeping with the fold change that is observed in bioinformatic correlations of elevated LINE-1 mRNA expression in p53-deficient cancers (Wylie et al. 2016) and elevated APOBEC3B mRNA expression in p53-

deficient cancers from *in vivo* TCGA samples (Results Chapter 2; Figure 3.2.3).

These results suggest two findings. Firstly, that p53 inactivation in HCT116 cells may indeed be a model that approximates LINE-1 and APOBEC3B derepression in cancer. And secondly, that p53 inactivation appears to lead to LINE-1 and APOBE3B activation.

These cell lines seemed likely to be an appropriate model system for LINE-1 and APOBEC3B derepression based on the results given in Figure 3.1.2 and the rationale given in the introduction to this chapter. These cells were consequently used for further experimentation. The next experiment conducted aimed to examine whether increasing p53 activity might reduce LINE-1 and APOBEC3B expression - the inverse relationship to the first experiment given in Figure 3.1.2. To test this, p53 activity was increased using the MDM2 inhibitor Nutlin-3a, which stabilises cellular p53 protein levels. These experiments were conducted in the HCT116 $p53^{wt/wt}$ cells that express p53. First, the dose of Nutlin-3a that would lead to cell death was established, in order to identify a Nutlin-3a dose that would promote p53 activity in these cells without causing cell death. This was achieved by treating cells with a dilution series of Nutlin-3a concentrations, and cellular survival was measured three days after treatment by the Sulforhodamine B Biomass Assay. The results are given below in Figure 3.1.3.

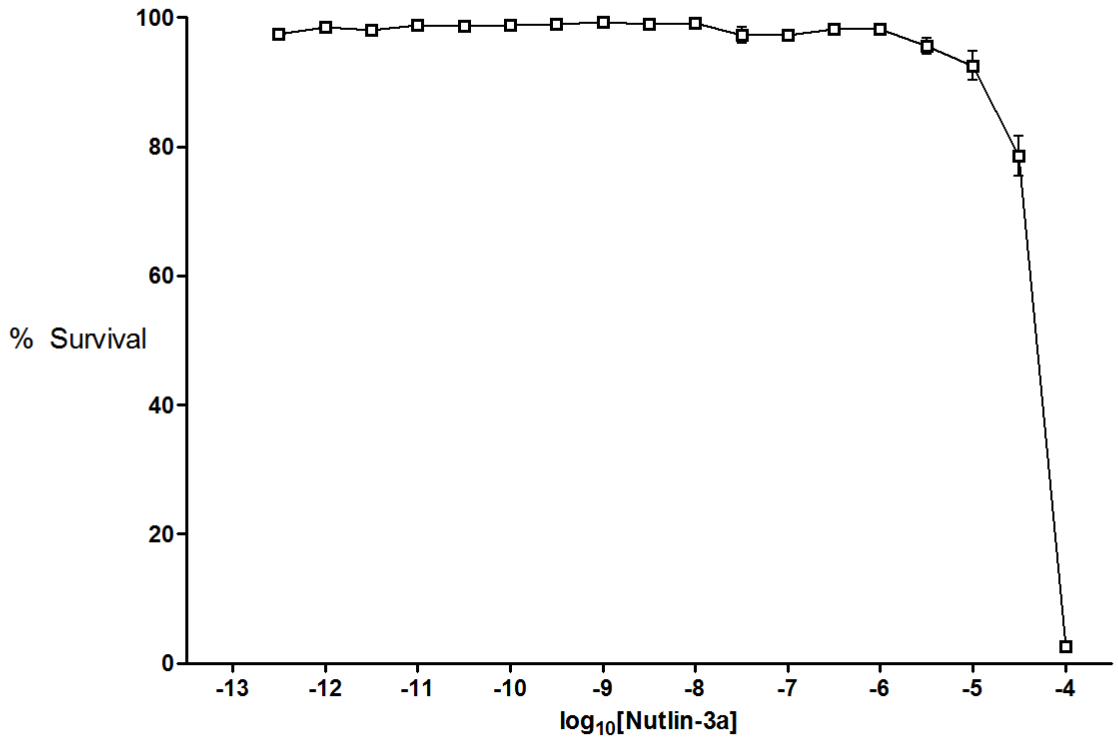


Figure 3.1.3 - Dose-survival curve for HCT116 p53$^{wt/wt}$ cells three days after treatment with Nutlin-3a. Mean ± standard error of the mean. Three experimental repeats.

The results in Figure 3.1.3 indicate that Nutlin-3a concentrations greater than 10^{-6} M appear to lead to increasing cell death, with almost complete cell death at 10^{-4} M. This cell death is presumably due to p53-mediated apoptosis, with an additional contribution from the effect Nutlin-3a toxicity. Based on the data given in Figure 3.1.3, a Nutlin-3a dose of 10^{-6} M (1 μM) was chosen as the dose that would promote p53 activity without causing cell death.

In the next experiment, HCT116 p53$^{wt/wt}$ cells were treated with 1 μM Nutlin-3a, and expression levels of LINE-1 and APOBEC3B were measured by qRT-PCR after three days. The results are given below in Figure 3.1.4.

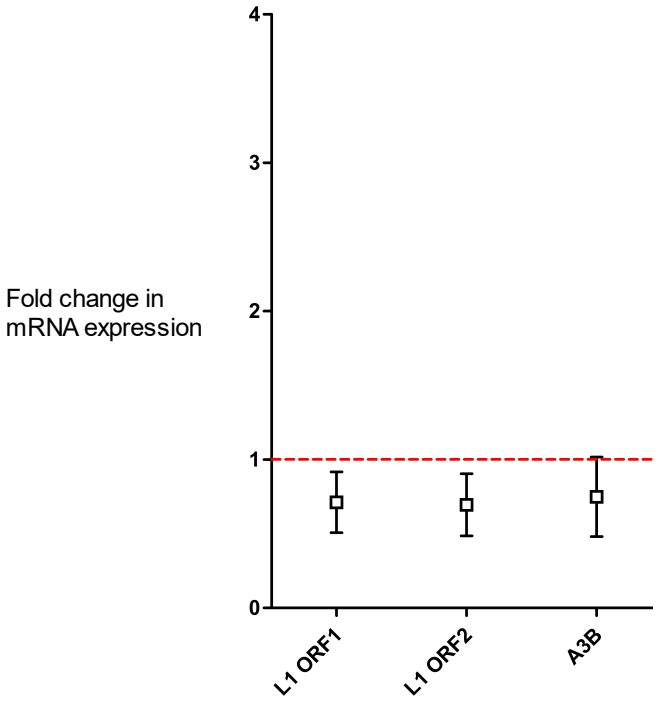


Figure 3.1.4 - Treatment with 1 μM Nutlin-3a leads to a moderate reduction in LINE-1 and APOBEC3B mRNA levels in HCT116 p53$^{wt/wt}$ cells. Expression levels are normalised to ACTB expression. Mean ± standard error of the mean. Three experimental repeats. Dashed line: fold change value of 1. t-test p-values for ORF1 = 0.0865, ORF2 = 0.0939, APOBEC3B = 0.184.

The data given in Figure 3.1.4 suggest that increasing p53 activity might lead to a decrease in LINE-1 and APOBEC3B mRNA expression in these

cells. However, these changes do not individually reach statistical significance. Taken together with the results given in Figure 3.1.2 (which show an inverse pattern when p53 is inactivated) the overall trends observed in both figures are in keeping with the notion that p53 is an inhibitor of LINE-1 and APOBEC3B mRNA expression in HCT116 cells. This would, in turn, be in keeping with the notion that p53 inactivation might lead to the activation of LINE-1 and APOBEC3B, and lend weight to the use of p53 inactivation in these cells as a means to recapitulate LINE-1 and APOBEC3B derepression as is observed *in vivo.* Based on this rationale, these cells were taken forward as a model system for further experimental interventions and mechanistic investigation.

3.1.3. Investigating the effect of the reverse transcriptase inhibitor azidothymidine on LINE-1 expression and APOBEC3B expression in the HCT116 cell line

The data presented in the previous section, section 3.1.2, do not provide insight into the extent to which APOBEC3B modulation is mediated by p53 alone, or whether LINE-1 may mediate some or all of the effects observed on APOBEC3B. In an attempt to disentangle these two factors, a LINE-1 inhibitor was used in both HCT116 $p53^{wt/wt}$ and HCT116 $p53^{-/-}$ cells, and mRNA expression was again measured by qRT-PCR. As described in the introduction to this chapter, the method of LINE-1 inhibition selected was pharmacological inhibition by nucleoside reverse transcriptase inhibitors (RTIs). In the experiments in this section, the RTI azidothymidine (AZT) was used to inhibit LINE-1. AZT was chosen as it plays a canonical role as the first RTI developed for use in patients with HIV, and was described to be an effective inhibitor of LINE-1 activity (Jones et al. 2008; Dai, Huang, and Boeke 2011).

First, the toxicity of AZT in these cell lines was established. This was determined by treating cells with a dilution series of AZT, and measuring cell death by Sulforhodamine B Biomass Assay. The results are given below in Figure 3.1.5.

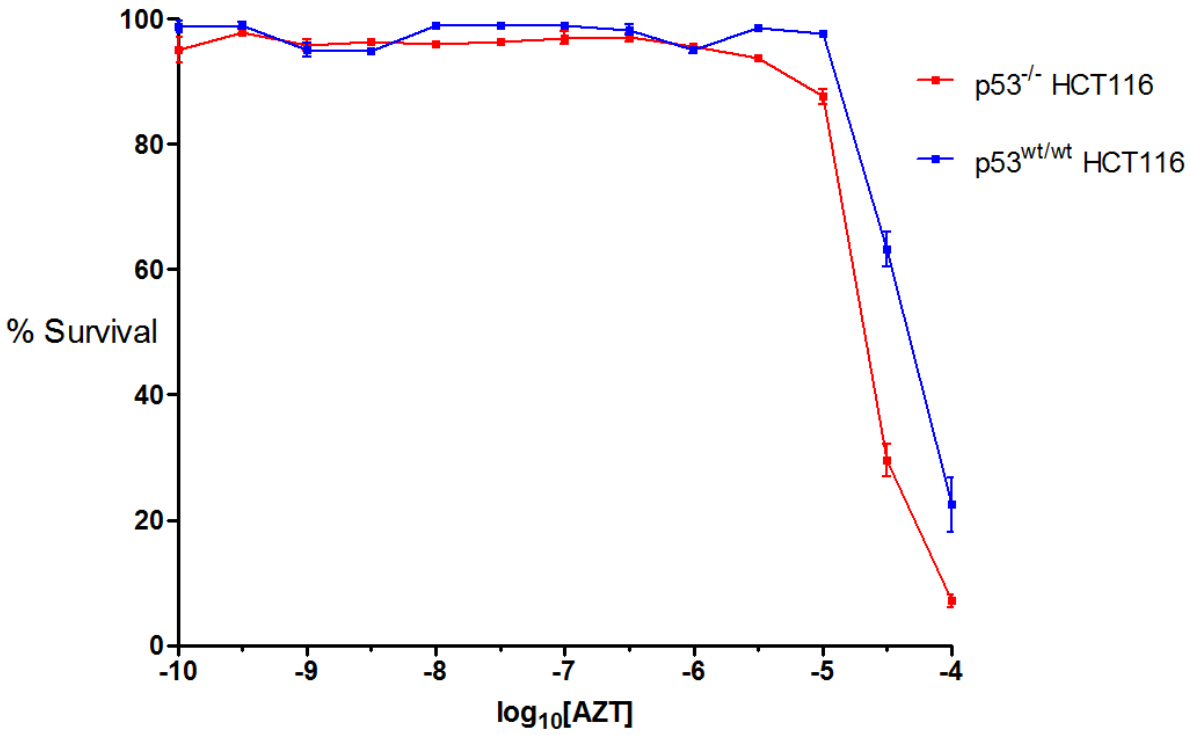


Figure 3.1.5 - Dose-survival curves for HCT116 p53$^{wt/wt}$ cells and HCT116 p53$^{-/-}$ cells two days after treatment with AZT. Mean ± standard error of the mean. Three experimental repeats.

The results in Figure 3.1.5 indicate that AZT concentrations above 10^{-5} M (10 μM) appear to lead to cell death. Based on these data, concentrations of AZT above 10 μM were not used in subsequent experiments.

Next, HCT116 p53$^{wt/wt}$ cells and HCT116 p53$^{-/-}$ cells were again treated with a dilution series of AZT at concentrations of 10μM or below. APOBEC3B mRNA levels were measured by qRT-PCR. The results are given below in Figure 3.1.6.

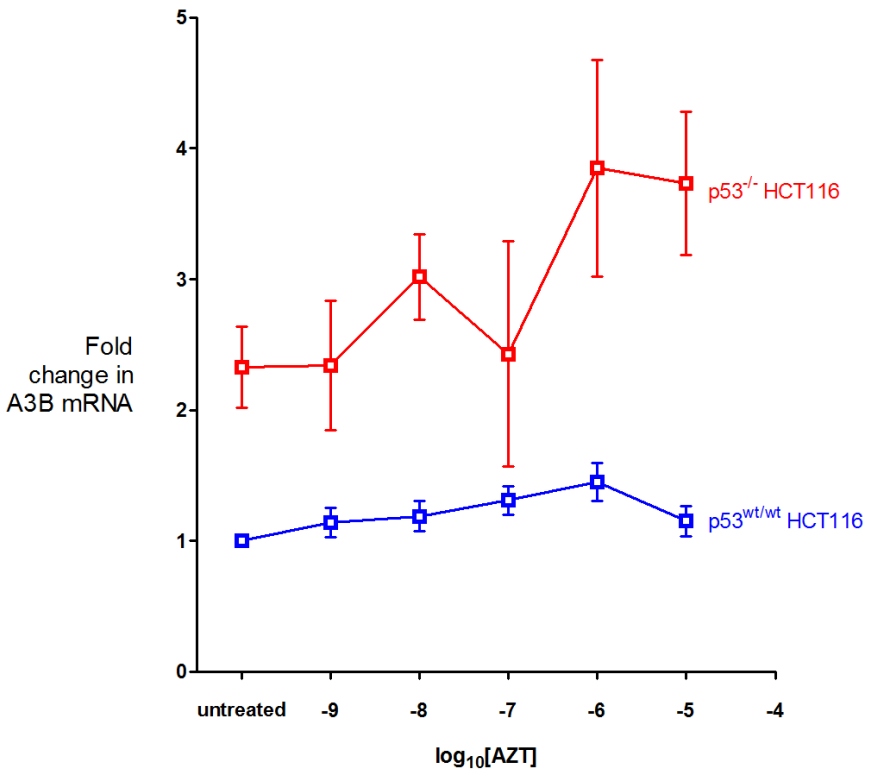


Figure 3.1.6 - Dose-response curves for APOBEC3B mRNA expression in HCT116 $p53^{wt/wt}$ cells and HCT116 $p53^{-/-}$ cells, two days after treatment with AZT. Expression levels are normalised to ACTB expression and the untreated $p53^{wt/wt}$ group. Mean ± standard error of the mean. Three experimental repeats. t-test p-value comparing 10 μM conditions = 0.0056.

The results in Figure 3.1.6 indicate that treatment with AZT leads to an increase in APOBEC3B mRNA expression in HCT116 $p53^{-/-}$ cells. In general, this effect appears to increase as the dose of AZT increases, although the data point for HCT116 $p53^{-/-}$ cells treated with 10^{-7} M AZT does not appear to be in keeping with this general trend. The maximal increase in APOBE3B mRNA expression in the HCT116 $p53^{-/-}$ cells is approximately two-fold, when treated with the highest doses of AZT, However, there is little to no increase in APOBEC3B mRNA expression in HCT116 $p53^{wt/wt}$ cells as

AZT concentrations increase. These data appear to contradict the hypothesis at hand, and instead support the notion that LINE-1 inhibition promotes (rather than inhibits) APOBEC activity in states of LINE-1 derepression.

After conducting the experiment shown in Figure 3.1.6, it was reasoned that the effect of AZT on APOBEC3B mRNA expression might be more pronounced if samples were collected three days following treatment with AZT rather than after two days. This was in keeping with the experience of laboratory colleagues, who reported that gene expression changes in their work could peak after three days rather than two. A prolonged treatment period is also in keeping with reports in the literature. For example, Kanu and colleagues treat cells for 4-10 days before detecting changes in APOBEC3B expression in cancer cells (Kanu et al. 2016).

The experiment was repeated using the longer time interval of three days, using the highest AZT concentration of 10 μM. In addition, LINE-1 mRNA levels were measured to investigate whether the unexpected result was due to any unexpected changes in LINE-1 mRNA expression mediated by AZT treatment. The results are given below in Figure 3.1.7.

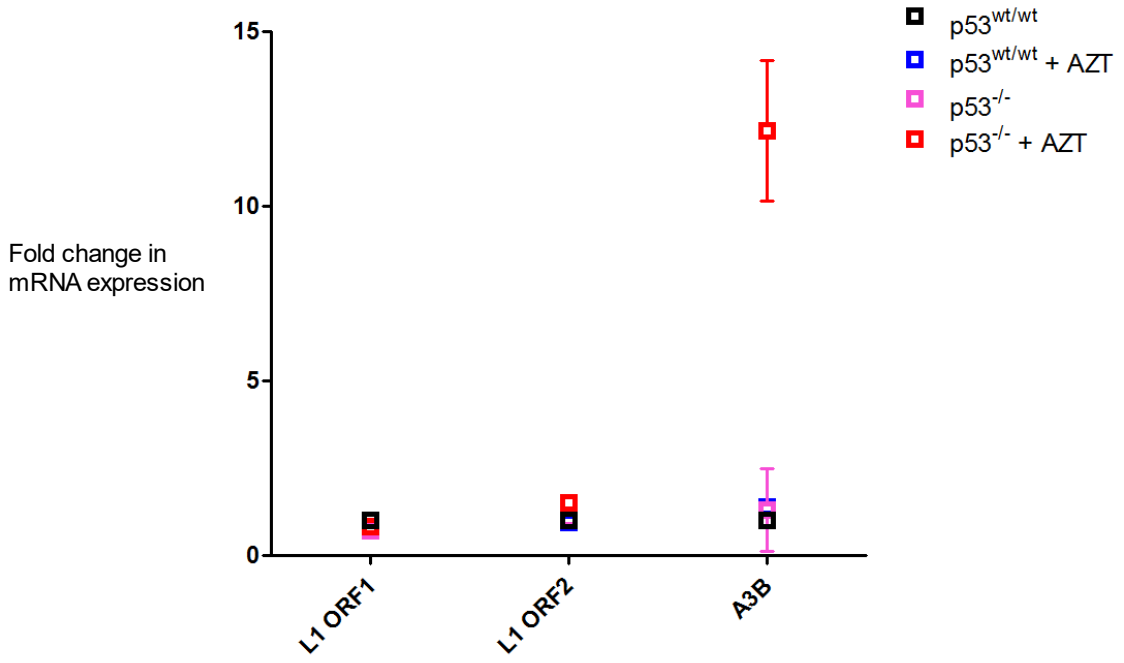


Figure 3.1.7 - Treatment with 10 μM AZT leads to an increase in APOBEC3B mRNA levels in HCT116 p53$^{-/-}$ cells but not their HCT116 p53$^{wt/wt}$ counterparts, as measured by qRT-PCR three days after treatment. LINE-1 expression shows no substantial change. Expression levels are normalised to ACTB expression and the untreated group in all four conditions. Mean ± standard error of the mean. Three experimental repeats. t-test p-value for p53$^{-/-}$ + 10 μM AZT = 0.000537.

The results shown in Figure 3.1.7 indicate that LINE-1 mRNA expression is not substantially altered by treatment with 10 μM AZT after three days. In contrast, APOBEC3B mRNA expression is elevated by a factor of around twelve after three days. This is again observed to be present in HCT116 p53$^{-/-}$ cells but not their HCT116 p53$^{wt/wt}$ counterparts. Measurement after an additional day therefore does indeed lead to a larger increase than the two-

fold increase in APOBEC3B mRNA expression observed in the previous experiment given in Figure 3.1.6. This increase does not appear to be attributable to unexpected increases in LINE-1 mRNA expression following AZT treatment. The data given here in Figure 3.1.17 hence further contradict the hypothesis at hand, and instead support the notion that LINE-1 inhibition promotes APOBEC activity in states of LINE-1 derepression.

To summarise, the results given thus far in this chapter together indicate that, in HCT116 cells, APOBEC3B expression can be modulated by modulating p53 activity, and can also be modulated by modulating the reverse transcriptase activity of LINE-1 when p53 is inactive. These results achieved aims described in the introduction to this chapter and represented novel observations, some of which have since also been shown by other laboratories (see section 3.1.6).

While the results from modulating p53 activity were expected, the results from treating cells with AZT were not. This prompted an attempt to investigate these observations in greater detail, using alternative techniques to test the same hypothesis. Specifically, other methods of measuring APOBEC activity and inhibiting LINE-1 activity were pursued. A biochemical APOBEC assay and the use of other RTIs other than AZT were used to address each of these aims, respectively. The results obtained from these two pursuits are given in the remaining two sections of this chapter.

3.1.4. High-throughput cytosine deaminase assay development

In the experiments given thus far in this chapter, cancer-associated APOBEC activity has been approximated by using APOBEC3B mRNA expression as a surrogate measure. Section 3.1.1.3 in the introduction to this chapter gives the rationale for this. The reasons for this choice include the fact that APOBEC3B mRNA expression is most clearly elevated in cancer relative to the expression of other APOBEC3 genes, APOBEC3B mRNA expression is most closely correlated to the number of APOBEC signature mutations relative to the expression of other APOBEC3 genes, and the C>T mutations at T<u>C</u>W residues (W = A or T) that define the APOBEC signature are in keeping with the enzymatic sequence preference of APOBEC3B or APOBEC3A cytosine deamination (Burns et al. 2013; Roberts et al. 2013; Taylor et al. 2013). This evidence indicates that APOBEC3B mRNA expression is, at least, a marker of APOBEC activity in cancer, notwithstanding the likely role of deamination by expressed APOBEC3B protein in causing APOBEC signature mutations.

Despite the evidence in support of this choice, there are also flaws evident in using APOBEC3B mRNA expression as surrogate a measure of APOBEC activity. These include the fact that APOBEC3B mRNA expression may not accurately represent the extent of APOBEC3B protein expression or the extent of any resultant APOBEC signature mutations in genomic DNA. In addition, mutagenic APOBEC activity might also be mediated by APOBEC3A rather than APOBEC3B. A method of measuring

APOBEC activity that could account for some or all of these flaws was consequently pursued.

Measuring APOBEC protein levels by Western blot or the number of APOBEC signature mutations by DNA sequencing were not pursued owing to practical constraints, as is also described in section 3.1.1.3. Instead, a biochemical method, measuring the cytosine deaminase activity of expressed APOBEC3A and APOBE3B enzymes, was pursued.

This approach in centred upon measuring the extent to which cell extracts can deaminate TCW residues in synthetic DNA probes. This is thought to be a specific marker of the enzymatic activity of APOBEC3A and APOBEC3B, and is therefore an indirect measure of APOBEC3A and APOBEC3B protein levels in these cells. Deamination of exogenous TCW residues in synthetic probes can also be used to approximate the formation of APOBEC signature mutations at endogenous TCW residues in the genome, which is the measure of APOBEC activity that appears to be most directly implicated in the pathogenesis of cancer. The features of such an assay therefore address some of the flaws highlighted above in using only APOBEC3B mRNA expression when investigating APOBEC activity.

This section describes the process of developing this type of assay in our laboratory. The methods draw in large part from, and attempt to expand upon, assays for APOBEC cytosine deaminase activity described in the literature (Burns et al. 2013; Holtz, Sadler, and Mansky 2013; Vieira et al. 2014). The biochemical reactions required for the cytosine deaminase

assay described in this section are presented diagrammatically in Figure 3.1.8 below, with the protocol detailed in the Materials and Methods section.

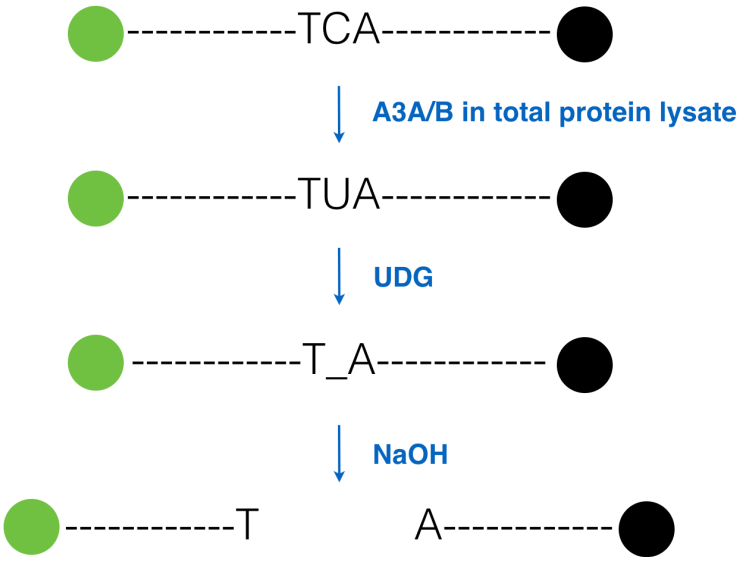


Figure 3.1.8 - A diagrammatic representation of reaction steps in the cytosine deaminase assay.

The assay is conducted by mixing a cell lysate containing APOBEC enzymes with a labelled ssDNA substrate. The ssDNA is labelled with a green 5' fluorescein fluorophore and a 3' TAMRA fluorescence quencher. The two labels act as a Förster resonance energy transfer (FRET) pair. The green fluorescence of the intact ssDNA probe is quenched, since the fluorescein emission spectrum overlaps with the TAMRA excitation spectrum when the two are in close proximity (Edelman, Cheong, and Kahn 2003). However, if this ssDNA is cleaved, FRET is effectively eliminated by

the loss of proximity of the two labels (Edelman, Cheong, and Kahn 2003; Burns et al. 2013). This leads to a loss of quenching, and an increase in green fluorescence that is proportional to the amount of cleaved ssDNA.

The labelled ssDNA probe used in these experiments contains a single cytosine in an TCA motif, as is described to be favoured by APOBEC3A and APOBEC3B (Taylor et al. 2013). Deamination of this cytosine by APOBEC3A or APOBEC3B in cell extracts would consequently lead to the formation of a single uracil. The addition of exogenous uracil DNA glycosylase (UDG) then leads to uracil excision and the formation an abasic site. Lastly, the addition of exogenous NaOH leads to ssDNA cleavage at this abasic site, liberating fluorescein from its TAMRA quencher. As result, APOBEC activity leads to green fluorescence once all reaction steps are complete.

The assay format that was first trialled was based on protocols given in Burns et al. (2013) and Vieira at al. (2014). These were selected as they were designed for use in high-throughput workflows. The protocols use a small reaction volume, with 20 μl reactions enabling processing on 384-well plates. A high-throughput assay protocol was pursued in order to facilitate the investigation of APOBEC activity in our laboratory in a wide range of treatment conditions in future work.

This section describes the successful development of a high-throughput assay, through optimisation and the inclusion of control conditions that were not previously described in the literature reviewed and which enable assay

interpretation. However, a low-throughput deaminase assay is then ultimately used in the subsequent section, in order to better resolve the small difference between APOBEC-specific deaminase activity and non-specific nuclease activity observed in the HCT116 cells used. The experiments conducted in developing the high-throughput assay will now be described in turn.

The first experiment conducted with this assay aimed to vary a number of protocol conditions simultaneously in order to identify optimal reaction conditions. A number of control conditions that were not present in the protocols found in the literature were included in an attempt to aid assay interpretation. Firstly, the protein concentration of the cell lysate was varied across a dilution series. Incubation times were also varied, so that results were measured after either one, two or three hours of incubation. In addition, a set of reactions lacking the fluorescently labelled ssDNA were used as a negative control to determine if there was any endogenous fluorescence unexpectedly arising from the remaining components of the reaction. Lastly, another set of reactions lacking the addition of exogenous UDG were used as another negative control, aiming to test the premise that exogenous UDG would be required for the reaction to proceed fully to ssDNA cleavage.

The cells used for the experiments described in this section are the HCT116 $p53^{-/-}$ cells. These were chosen because they were found to have elevated APOBEC3B mRNA, as described in experiments given earlier in this chapter, and were consequently deemed more likely to yield a detectable

signal of APOBEC cytosine deaminase activity. The results of the first deaminase assay experiment are given below in Figure 3.1.9.

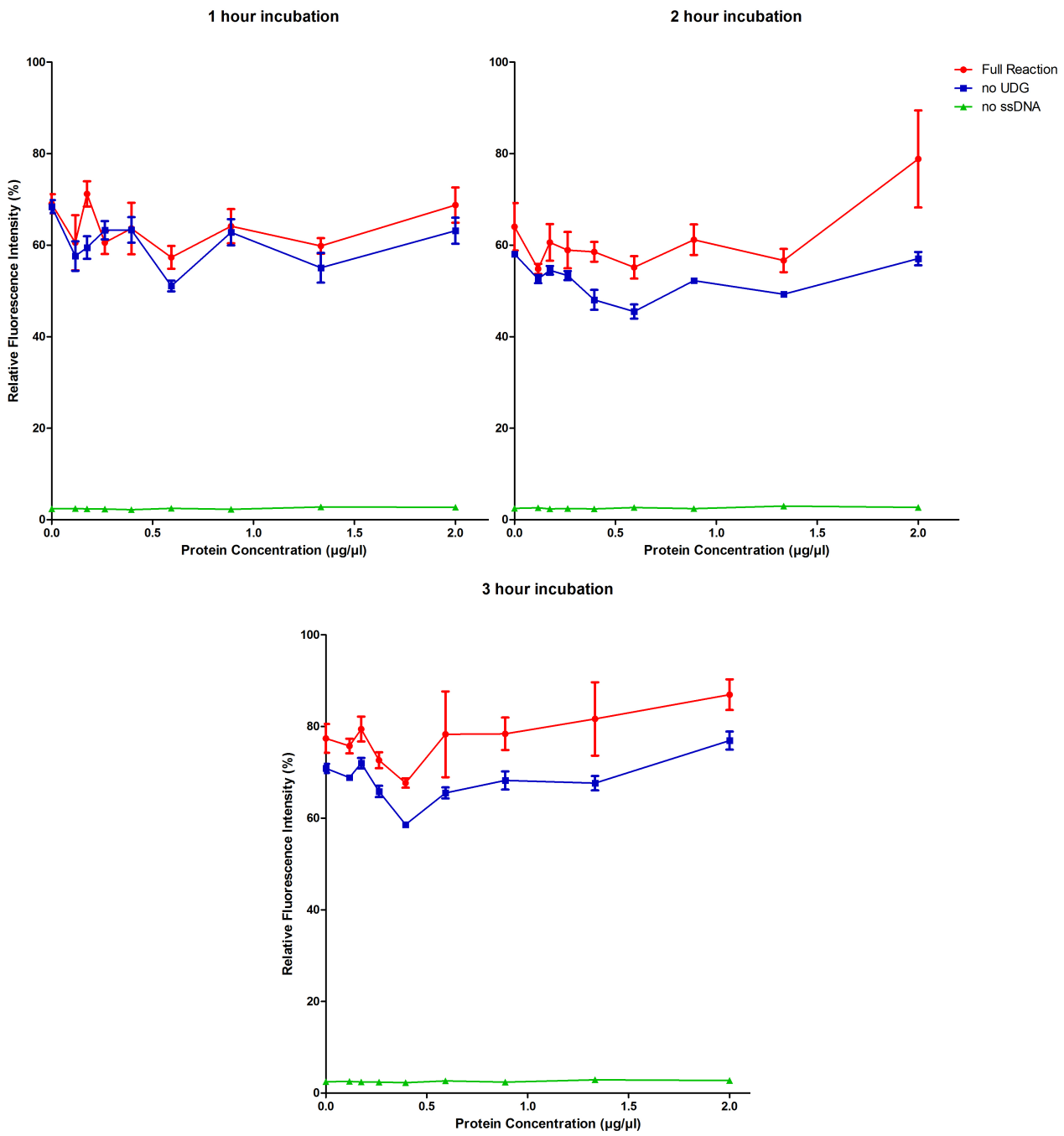


Figure 3.1.9 - Results of the initial cytosine deaminase assay experiment. The data are normalised to the maximum signal measured across all conditions. Mean ± standard error of the mean. Three experimental repeats.

The results shown in Figure 3.1.9 indicate a number of findings that were used to guide assay optimisation. These will be described in turn. Firstly, it

appears that reactions lacking the fluorescent ssDNA probe have negligible levels of fluorescence in the excitation and emission wavelengths used for the assay. This indicates that fluorescence arising from these remaining components of the reaction are unlikely to affect assay interpretation. This control condition was therefore consequently excluded from subsequent experiments.

Incubation times were also considered. These data suggest that incubation times of two hours or more lead to a greater degree of fluorescence, with the largest values for fluorescence arising in the two hour and three hour incubation conditions. As a result, a reaction time of two hours was chosen as the incubation time, in keeping with the incubation period used by Burns et al. and Vieira et al.

In terms of protein concentration of the lysate, the data in Figure 3.1.9 suggest that the reaction proceeds more readily at higher concentrations. However, this appeared to be a weak trend, and the data points did not form an asymptotic curve that would be indicative of a reaction series reaching completion. As a result, higher protein concentrations were used in subsequent experiments in an attempt to promote reaction completion.

Lastly, the data showing the control reactions lacking UDG were considered. These data indicate that, under these conditions, the vast majority of ssDNA probe cleavage does not depend on the addition of exogenous UDG. This is not in keeping with the assumptions made for the assay as used in the literature and outlined in the text pertaining to Figure

3.1.8. It was speculated that this may be to the action of endogenous UDG in the cell lysate, or may also be due to ssDNA probe degradation that was not dependent upon APOBEC activity, but may instead be mediated by endogenous DNAses in the cell lysate. As a result, additional control conditions attempting to measure probe degradation that is not specific to the activity of APOBEC3A and APOBEC3B were introduced.

To summarise the conclusions drawn from the data in Figure 3.1.9, autofluorescence was shown to be negligible and a two hour incubation time appeared to be optimal. Higher protein concentrations and the use of control conditions for measuring non-specific probe degradation should be introduced. These conclusions were incorporated in designing the subsequent experiment, shown in Figure 3.1.10 and detailed below.

In the experiment shown in Figure 3.1.10, attempts were made to characterise and quantify non-specific degradation arising from the samples in the experiment. This was pursued by sourcing two additional negative control ssDNAs. The first new ssDNA contains no cytosines (TTA in place of TCA). Since this ssDNA contains no cytosines, it cannot be subject to APOBEC-mediated cytosine deamination, and therefore all degradation of this control probe must be due non-specific causes. Degradation of this probe does not depend on uracil formation, and any increase in fluorescence is therefore independent of UDG activity.

The second new ssDNA control used contains a single cytosine in an ACA rather than a TCA sequence context. This cytosine is in a context that is not

preferred by APOBEC3A, APOBEC3B, or indeed any of the APOBEC3s, which all prefer a 5' thymidine (Conticello et al. 2007; Olson, Harris, and Harki 2018). Studies where APOBEC3A or APOBEC3B are overexpressed in yeast with subsequent genomic sequencing indicate that the vast majority of mutations occur at TCA residues, with ACA residues accounting for the site of approximately 5% of mutations mediated by APOBEC3A and 1% of mutations mediated by APOBEC3B (compared in Chan et al. 2015). This ACA control probe is therefore included in an attempt to characterise any degradation that might arise from the presence of cytosine alone. This may be useful since cytosine is known to be susceptible to additional damage in commonly-used biochemical protocols. For example, it is known that cytosine deamination that occurs in the process of preparing samples for NGS is a major cause of sequencing artefacts (Chen et al. 2014).

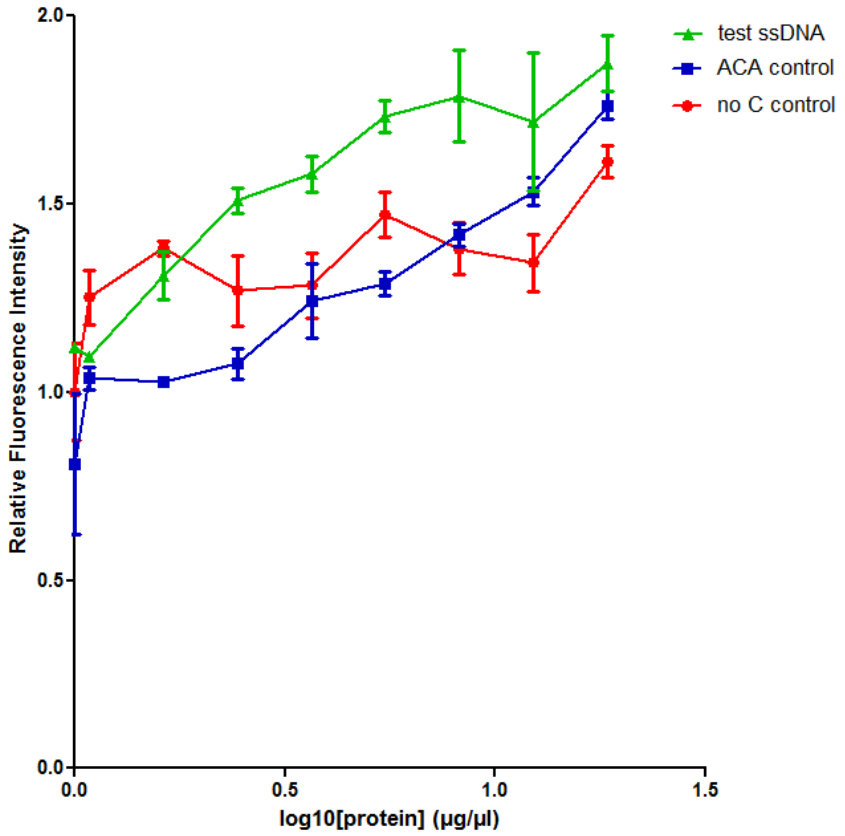


Figure 3.1.10 - Cytosine deaminase assay results following one round of optimisation and the addition of negative controls. The data are normalised to the no C control reaction containing lysis buffer alone (i.e. lysate protein concentration of zero). Test ssDNA = T<u>C</u>A, ACA control = A<u>C</u>A, no C control = T<u>T</u>A. Mean ± standard error of the mean. Three experimental repeats.

The results in Figure 3.1.10 indicate that increasing the protein concentration produces a more prominent degradation curve, as predicted from the previous experiment. They also indicate that it is possible to identify non-specific degradation, and reaffirm that non-specific degradation is a significant contributor to total degradation. For example, there is a notable degree of 'no C' ssDNA degradation. This degradation is UDG-

independent, and is therefore due to the activity of nucleases or other DNA-degrading factors in the lysate. These data appear to reaffirm the need to include a measure of non-specific degradation in this assay in future.

Figure 3.1.10 also shows that the test ssDNA generally produces more fluorescence than the two control ssDNA under these reaction conditions. This suggests that this additional fluorescence represents detectable APOBEC3A/APOBEC3B activity. However, this trend is not entirely clear given the errors in the readings made and the fact that non-specific degradation appeared to sometimes exceed degradation that was thought to be specific. It was speculated that this was due to a low signal-to-noise ratio. As a result, efforts were made to increase the dynamic range of the assay in order to clarify the trend suggested by Figure 3.1.10.

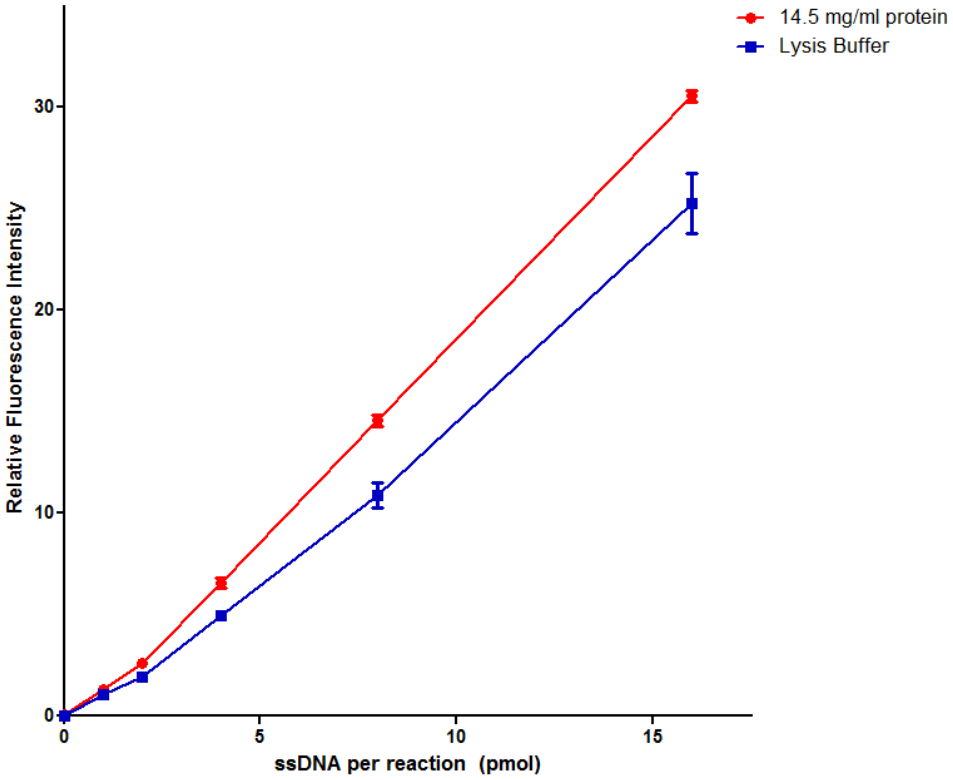
To summarise, the data in Figure 3.1.10 indicate that increasing protein concentration increases ssDNA degradation, that a significant amount of ssDNA degradation is non-specific, and that negative control ssDNA degradation sometimes overlapped with or exceeded degradation specific to cytosine deamination, which was thought could be in part due to a low dynamic range. High protein concentrations and negative control ssDNAs were therefore used in subsequent experiments. The remaining experiments in this section attempt to identify and improve upon the factors affecting the dynamic range of the assay. Three experiments were conducted, given in Figure 3.1.11, Figure 3.1.12 and Figure 3.1.13. These were to test the role of ssDNA concentration, ssDNA length and uracil formation, respectively, in determining the dynamic range.

The dynamic range of the assay is the ratio of its maximum signal to its minimum signal. The dynamic range may therefore be increased either by increasing the assay's maximum signal or by decreasing the assay's minimum signal.

The first factor considered was whether ssDNA concentration could be varied in order to improve the assay's dynamic range. Increasing ssDNA concentration could increase the dynamic range of the assay if ssDNA availability is the limiting factor in the reaction. However, this comes at the expense of increased basal fluorescence when no ssDNA cleavage has occurred, such as in the control condition containing lysis buffer alone which lacks any cell lysate. These factors contribute to a trade-off that increasing the ssDNA concentration brings to potentially improving the dynamic range of the assay. Decreasing ssDNA concentration, rather than increasing it, in order to improve the dynamic range of the assay is also subject to the same trade-off. If ssDNA availability is not the limiting factor in the reaction, then reducing ssDNA concentration may improve the dynamic range of the assay.

The data for the experiments conducted in Figure 3.1.10 were considered when evaluating whether ssDNA concentration should be decreased or increased in an attempt to improve the dynamic range. It was thought that the curves produced should become asymptotic if the reaction reached completion, indicating that ssDNA availability might not be a limiting factor. The shape of the 'test ssDNA' curve was in keeping with possible reaction

completion, but the other conditions did not show this. As a result, it was not clear whether these experiments supported increasing or decreasing the concentration of ssDNA. The data shown in Figure 3.1.9 were also reviewed. These data indicated that ssDNA fluorescence was several fold higher than that of the remaining reaction materials at baseline, supporting the notion that reducing ssDNA concentration could be beneficial. Since it was not clear from the experiments conducted how ssDNA should be varied, both increasing and decreasing ssDNA concentration were trialled in an attempt to improve the dynamic range. This was done using reactions at either the maximum protein concentration that could be achieved by the protocol, or using lysis buffer alone. The results are given in Figure 3.1.11.



ssDNA per reaction (pmol)	1	2	4	8	16
Dynamic range	1.23	1.35	1.32	1.34	1.21

Figure 3.1.11 - ssDNA concentration does not significantly alter the assay's dynamic range. The data are normalised to the 1 pmol, lysis buffer only reaction. Mean ± standard error of the mean. Three experimental repeats.

Figure 3.1.11 shows that the dynamic range of the assay does not markedly vary as the ssDNA concentration is varied. The protocols given in the literature use 4 pmol of ssDNA per 20 μl reaction. The data above indicate that the dynamic range reduces from 1.32 to approximately 1.22 if the amount of ssDNA is either increased or decreased four-fold, to 1 pmol or 16 pmol respectively. The dynamic range at the 2 pmol or 8 pmol conditions

appears to increase by 0.03 and 0.02 respectively compared to the 4 pmol condition. It was reasoned that any possible improvements suggested by these differences are negligible, given that the standard error of the mean for the data points given in Figure 3.1.10 were comparatively large. As a result, the original 4 pmol condition was chosen as the preferred ssDNA amount in subsequent experiments, in order to maintain consistency and comparability with published protocols where possible.

After assessing ssDNA concentration, the next factor considered was whether ssDNA length could affect the dynamic range of the assay. As described above and shown in Figure 3.1.9, background ssDNA fluorescence is much higher than the fluorescence of the remaining reaction materials. One way to reduce this basal fluorescence, aside from reducing the amount of ssDNA in each reaction, is to improve the efficiency of FRET quenching. FRET efficiency is dependent upon the sixth power of the distance between two labels (Edelman, Cheong, and Kahn 2003), and as a result, FRET efficiency can be markedly altered as this distance is changed. A shorter 17 nucleotide ssDNA was identified in the literature (Burns et al. 2013), for comparison with the 40 nucleotide ssDNA used up to this point (Vieira et al. 2014). In order to assess FRET quenching efficiencies, three ssDNAs were used. These were the new 17-mer, the old 40-mer, and the 40-mer labelled with a 5' fluorescein alone (missing a 3' TAMRA quencher). Their relative fluorescence is given in Figure 3.1.12.

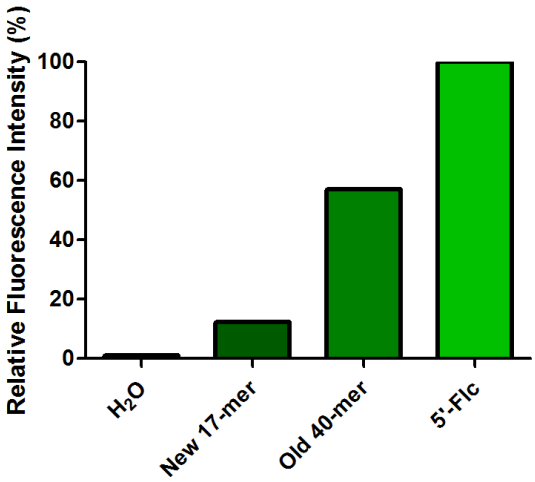


Figure 3.1.12 - Relative fluorescence intensity of 4 pmol labelled ssDNA, untreated and dissolved in 20 μl H_2O. H_2O = water only negative control, 5'-Flc = 40-mer labelled with fluorescein only - positive control. Values normalised to 5'-Flc.

The results in Figure 3.1.12 indicate that the FRET quenching efficiency was improved from approximately 40% to approximately 90% by reducing the distance between the 5' and 3' labels. This was thought to be likely to lead to an improvement in the assay's dynamic range. The possibility of further shortening the ssDNA was considered. However, this was not pursued, owing to concerns that an unduly short ssDNA might no longer act as a substrate for APOBEC-mediated deamination (as described in Thielen et al. 2007 for APOBEC3G), unlike the 17-mer which has been purported to be subject to such deamination in the literature.

After varying ssDNA length and concentration, the final factor considered in assessing the dynamic range of the assay was the role of uracil formation. This was achieved by the inclusion of ssDNA that, in place of having a single cytosine, instead included a single uracil. This was used in order to assess the extent to which action by UDG and NaOH-mediated hydrolysis were limiting the reaction, rather than cytosine deamination. It was reasoned that this uracil-containing ssDNA would in effect be a positive control, indicating the maximum degree of ssDNA cleavage achievable in these reaction conditions.

The uracil-containing positive control ssDNA was used in the experiments given in Figure 3.1.13. This figure repeats the experiments done in Figure 3.1.10, but uses a new set of shorter 17 nucleotide ssDNAs, as discussed in relation to Figure 3.1.12. The results are given below, in the final figure of this section.

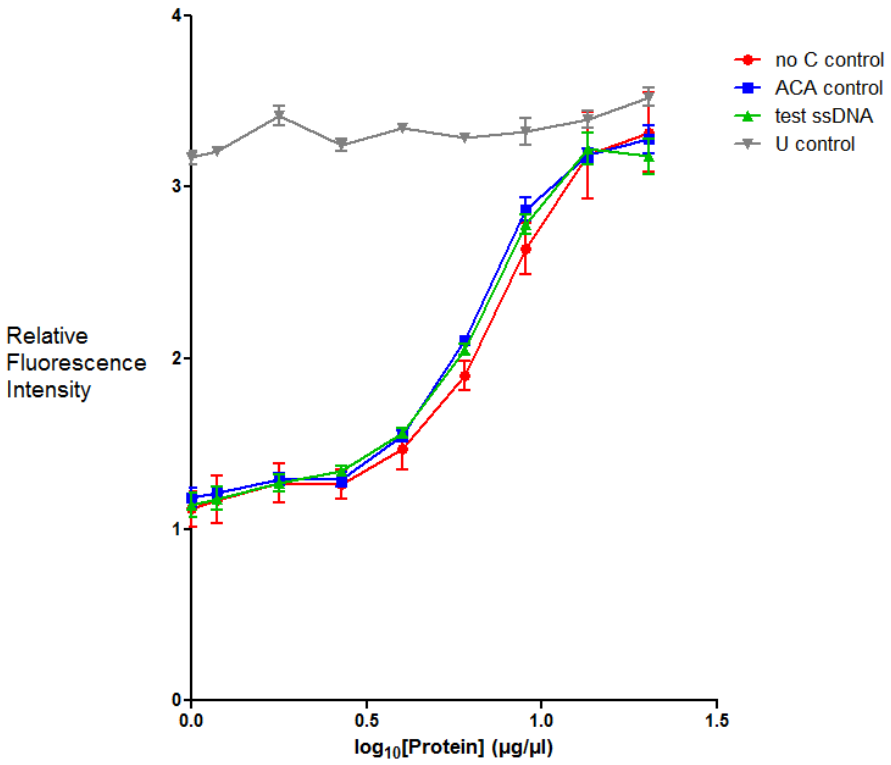


Figure 3.1.13 - HCT 116 p53$^{-/-}$ cells are subjected to the final iteration of the high throughput deaminase assay. Test ssDNA = T<u>C</u>A, ACA control = A<u>C</u>A, no C control = T<u>T</u>A, U control = T<u>U</u>A. Mean ± standard error of the mean. Three experimental repeats.

Figure 3.1.13 shows that the new, shorter ssDNAs contribute to an improved dynamic range. Here, the dynamic range is approximately 3.5-fold. This can be compared to the dynamic range using the longer ssDNAs in Figure 3.1.10, which is approximately 2-fold. These data also show that the new, uracil-containing positive control ssDNA facilitates the interpretation of the assay's results by giving a measure of the maximum limit of the reaction. The data in Figure 3.1.13 indicate that the reaction nears completion at high protein concentrations, as might be speculated from the shape of the test ssDNA curve in Figure 3.1.10. The data in Figure

3.1.13 therefore confirm that the changes made to the assay lead to an improved dynamic range, with control ssDNAs that enable assay interpretation by defining the upper and lower bounds of the reaction.

In terms of assessing APOBEC activity, the data in Figure 3.1.13 indicate that the vast majority of ssDNA degradation that occurs under these conditions is non-specific; there is little difference between the curves for the test ssDNA, ACA ssDNA and no C ssDNA. This suggests that cytosine deaminase activity that is specific to APOBEC3A and APOBEC3B is not detectable under these conditions. This is contrast to the results in Figure 3.1.10, where the test ssDNA curve was distinguishable from the ACA and no C ssDNA curves. To speculate, this may indicate that the 17-mer used to improve the dynamic range may have lost some or all of its substrate specificity to APOBEC3A and APOBEC3B. Alternatively, the difference apparent in Figure 3.1.10 may be a function of statistical noise owing to a low dynamic range - this would suggest that the result in Figure 3.1.13 is a more accurate one, and that perhaps the HCT116 cells used have too low a rate of APOBEC activity or too high a rate of non-specific nuclease activity to elicit a prominent APOBEC-specific signal.

In summary, this section details the development of a high-throughput cytosine deaminase assay in our laboratory. The experiments conducted identify a number of factors that influence the assay's ability to detect APOBEC activity. The variables investigated include incubation time, protein concentration, autofluorescence, ssDNA concentration and ssDNA length. The inclusion of novel ssDNA controls enable interpretation of the

assay, and confirm that the variables optimised lead to an assay with an improved dynamic range. These results optimise and expand upon what has previously been published on high throughput cytosine deaminase assays. In the next and final section of this chapter, these results are used to inform the use of a low-throughput cytosine deaminase assay, which is used in an attempt investigate APOBEC activity as it pertains to LINE-1 activity.

3.1.5. Investigating the effect of the reverse transcriptase inhibitors 3TC and d4T on APOBEC cytosine deaminase activity in the HCT116 cell line

In this final section, the results accumulated thus far in the chapter are taken together to design an experiment that attempts to test the hypothesis that LINE-1 activity promotes APOBEC activity. APOBEC activity is measured using a low-throughput deaminase assay, while LINE-1 activity is modulated using a combination of the reverse transcriptase inhibitors 3TC and d4T. The rationale for these two choices is given below.

In terms of the deaminase assay, the results in the previous section indicated a low capacity for the FRET-based, high-throughput assay to discriminate APOBEC activity from non-specific activity in the conditions used. It was speculated that the length of the ssDNA used could be a key factor. On one hand, shortening the ssDNA promotes FRET quenching efficiency. However, on the other hand, shortening the ssDNA could reduce the ability for APOBEC enzymes in the lysate to recognise the ssDNA as a substrate (Thielen et al. 2007). One way to address this trade-off would be to design a series of ssDNAs between 17 nucleotides and 40 nucleotides in length, to identify if there is an intermediate ssDNA length that retains both a high dynamic range and, putatively, retains the ability to be metabolised by APOBEC enzymes in the lysate. This was not pursued, as an alternative explanation for the observations made is that, in the HCT116 p53$^{-/-}$ cells used, APOBEC activity could simply be too low relative to a high activity of non-specific nuclease activity; the results of a length series of ssDNAs

might not be able to indicate whether this is indeed the case if no ssDNA that optimises the putative length trade-off is identified.

An alternative deaminase assay was consequently pursued. This version of the assay is not FRET-based. It is based on the use of the 40-mer ssDNA which is only labelled with a 5-fluorescein. The fluorescently-labelled ssDNA reaction products are size-separated on a TBE-Urea gel. This gel-based method is not a high-throughput one. However, it does enable the detection of a 12 nucleotide-long band that is specific to ssDNA cleavage at the site of the single cytosine within it. Given that non-specific ssDNA degradation could generate a wide range of ssDNAs from 1 to 39 nucleotides in length, size-separating the reaction products should, in principle, enable the detection of APOBEC-specific activity at higher sensitivity than the high-throughput FRET-based assay. This assay should enable direct visualisation of the sites of specific and non-specific ssDNA degradation, with the dynamic range in effect enhanced by limiting the range of ssDNAs examined to only those of 12 nucleotides in length. The control ssDNAs chosen for this assay were the no C and U controls, to enable identification of the uncleaved and specifically-cleaved bands, respectively. The ACA control was not included, as it was observed that there was no substantial difference between this ssDNA and the no C ssDNA in the experiments given in Figure 3.1.10 and Figure 3.1.13.

The role of the deaminase assay is to measure APOBEC activity as it pertains to LINE-1 activity. In this final section, LINE-1 activity is modulated using 3TC and d4T. These were chosen as an alternative to AZT, which in

the experiments conducted led to an unexpected increase in APOBEC3B mRNA expression in HCT116 p53$^{-/-}$ cells, as shown in Figure 3.1.6 and Figure 3.1.7. The introduction and third results chapter of this thesis discuss evidence that the autoinflammatory pathways in Aicardi-Gouitères syndrome might be analagous to the pathways associated with in APOBEC activation in cancer. RTIs have been used in an attempt to rescue AGS phenotypes, both in model systems and in clinical trials, in keeping with the hypothesis that AGS is caused by a defect in retrotransposon repression. In a mouse model of AGS, AZT monotherapy fails to rescue the disease phenotype (Stetson et al. 2008), and it has been proposed that the chain-terminating effect of AZT on reverse transcriptases leads to the release of short ssDNAs that ultimately promote, rather than inhibit, innate immune pathways (Beck-Engeser, Eilat, and Wabl 2011). Instead, subsequent studies report that the use of a combination of RTIs is effective in rescuing AGS phenotypes. For example, Thomas et al. describe the use of 10 μM 3TC and 1 μM d4T in rescuing a LINE-1 mediated autoinflammatory phenotype in an human AGS neuronal cell culture model. The three RTIs share a common mechanism of action of inhibiting the active site of the LINE-1 reverse transcriptase - however, *in vitro* inhibition studies of purified ORF2 indicate that d4T is the most potent, followed by 3TC and then AZT, which may explain their differential efficacy in these contexts (Baldwin et al. 2023).

Given the use of a human model system, and the demonstration that the effect of these RTIs was mediated through LINE-1 in the Thomas et al. study, 3TC and d4T were chosen for use in the experiments in this section,

at the concentrations reported. Both p53-proficient and p53-deficient HCT116 cells were compared using the gel-based deaminase assay described above, either treated or untreated with these new RTIs. The results are given below, in Figure 3.1.14. Since ssDNAs used are conjugated to a fluorescein marker that may alter ssDNA migration, the no C and U controls were used as markers for interpreting the output of the instead of using a ssDNA ladder without conjugated fluorescein.

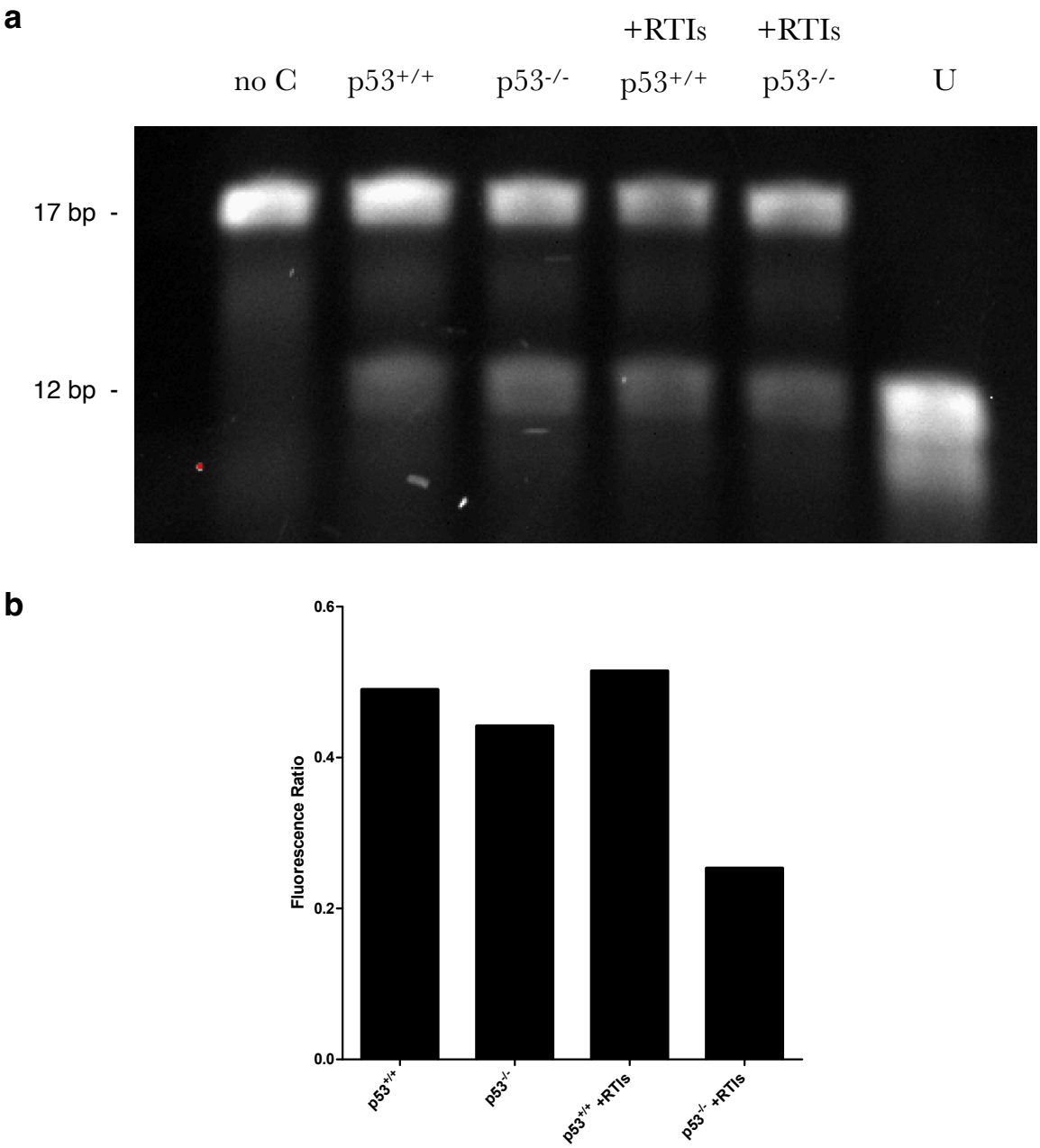


Figure 3.1.14 - A representative TBE-Urea gel (a) and its quantification (b) indicating that 3TC and d4T inhibits cancer-associated APOBEC activity in HCT116 p53$^{-/-}$ cells. The fluorescence ratio is calculated by dividing the intensity of the lower (cleaved) band by the intensity of the upper (uncleaved) band. Three experimental repeats.

The results in Figure 3.1.14a indicate that the TBE-Urea gel separates the ssDNA in the reaction mixture as expected, with the uncleaved input ssDNA and cleaved APOBEC-specific product identifiable as distinct bands. The no C lane shows a strong band of uncleaved ssDNA with a relatively low intensity tail of bands, suggesting that the non-specific nuclease activity is low. The U lane shows a strong APOBEC-specific band, with almost no uncleaved ssDNA, indicating that the reaction proceeds to completion if uracil is generated. Comparing the APOBEC-specific band in the U lane to the lanes containing the test ssDNA shows that, in general, the band in the U lane is of notably higher intensity. These data suggest that the rate of APOBEC activity in the HCT116 cells is low, rather than there being a particularly high rate of non-specific nuclease activity.

Figure 3.1.14b gives the digitally-quantified ratio of the intensity lower cleaved band to the upper uncleaved band in the test ssDNA conditions. These data indicate that the RTI combination of 3TC and d4T inhibits cancer-associated APOBEC activity in HCT116 $p53^{-/-}$ cells. In this condition, around a quarter of the input ssDNA is cleaved to form the 12 nucleotide product. In contrast, the other three conditions show a fraction of around a half.

These data can be compared to the previous results in this chapter, where APOBEC3B mRNA expression was measured in relation to p53 inactivation and AZT treatment. Previously, APOBEC3B mRNA expression appeared two-fold to three-fold higher in HCT116 $p53^{-/-}$ cells compared to HCT116 $p53^{+/+}$ cells. No such difference in APOBEC activity is apparent in Figure

3.1.14. This could be due to a discordance between APOBEC3B mRNA expression and its expression at protein level or deaminase activity of the expressed protein. This could also be due to the contribution of APOBEC3A; APOBEC3A mRNA expression levels were not measured in previous experiments. However, as before, RTI treatment leads to an APOBEC response only when p53 is deficient and LINE-1 is derepressed. Here, enzymatic activity is inhibited by around a half, whereas AZT treatment previously increased APOBEC3B mRNA expression by around twelve-fold.

To summarise, the results of the experiment shown in Figure 3.1.14 suggest that a combination of 3TC and d4T can be used to inhibit APOBEC activity in p53-deficient HCT116 cells. Since these cells exhibit impaired LINE-1 repression, these results in Figure 3.1.14 also suggest that LINE-1 activity may promote cancer-associated APOBEC activity.

3.1.6. Discussion

The aim of the experiments conducted in this chapter are to test the hypothesis that LINE-1 activity promotes APOBEC activity in cancer. This is in line with the overall aim of the thesis to investigate the regulation of APOBEC activity in cancer. The aim of the chapter was to be pursued by using experimental methods to attempt to identify causal relationships between cellular factors that could mediate cancer-associated APOBEC activity.

To this end, experiments were performed using the HCT116 cell line, and these experiments yielded the following results:

- p53 deficiency in HCT116 cells leads to LINE-1 mRNA and APOBEC3B mRNA upregulation.
- p53 stabilisation with Nutlin-3a in HCT116 cells might lead to LINE-1 mRNA and APOBEC3B mRNA downregulation.
- AZT treatment leads to APOBEC3B upregulation in p53-deficient HCT116 cells.
- The development of a high-throughput cytosine deaminase assay, through its optimisation and use of novel controls.
- 3TC and d4T treatment inhibits cancer-associated APOBEC cytosine deaminase activity in p53-deficient HCT116 cells.

Together, these experiments provide new evidence that can be used to evaluate the hypothesis that LINE-1 activity promotes APOBEC activity in cancer. On beginning to evaluate this evidence, it is noted that the evidence

is intrinsically limited by the fact that it has been accrued using only one cell type, the HCT116 cell line. Aside from the limitations of using a cancer cell line described in the introduction to this chapter, there are other factors that might potentially limit the generalisability of the findings produced from these cells. For example, HCT116 cells are colorectal cancer cells. Colorectal cancers do not typically display APOBEC signature mutations (Alexandrov et al. 2020). Moreover, on comparing different cancer types, colorectal cancer has been observed to have among the highest rates of LINE-1 retrotransposition events (Rodriguez-Martin et al. 2020). Therefore, as colorectal cancer cells, HCT116 cells may not have APOBEC activity and LINE-1 activity that is typical or representative of most cancers. Indeed, they appear to display observable but low levels of APOBEC activity in the experiments in this chapter. Ovarian or pancreatic cancers appear to more stably display both APOBEC activity and LINE-1 activity, and might therefore be more suitable models in this regard. In general, the evidence gathered in this chapter would likely be strengthened if repeated in a range of cancer cell types, as it would not be limited to the potentially atypical context of colorectal cancer, or indeed any atypical phenotypes that the HCT116 cell line might uniquely display when compared to other cell lines. In addition, assaying the developmental lineage of particular cell type - such as its stem cells and early progenitor cells - using single cell techniques and organoid cultures may yield data that may be of relevance to putative mechanisms of carcinogenesis *in vivo* involving mechanisms of cellular development. These limitations notwithstanding, these data can nonetheless be used to evaluate the hypothesis at hand, not least since this

specific cell line has been used to demonstrate seminal and generalisable findings in cancer biology, as previously described.

The experiments in this chapter are based on the premise that p53 inhibits LINE-1. The results of the experiments indicate that this does appear to be the case. p53 inactivation leads to LINE-1 upregulation at the mRNA level, while the opposite might be true when p53 is stabilised by Nutlin-3a. When these HCT116 cells are treated with reverse transcriptase inhibitors, effects on APOBEC activity, a likely downstream target of LINE-1 activity, occur only when cells are p53-deficient. The only active reverse transcriptase that can be targeted by the drug concentrations used is that of LINE-1, notwithstanding any previously undescribed off-target effects. Therefore, the observations from the p53 inactivation, p53 stabilisation, AZT treatment and 3TC/d4T treatment experiments all provide evidence supporting the notion that p53 inhibits of LINE-1 activity. As described in the introduction to this chapter, LINE-1 elements contain multiple p53 binding sites (Harris et al. 2009) and LINE-1 activity induces DNA damage that leads to a p53 response (Haoudi et al. 2004), suggesting that p53 could directly inhibit LINE-1.

Given this evidence, the findings above suggest that HCT116 $p53^{-/-}$ cells can reasonably be considered as a model that approximates the LINE-1 derepression that is observed in cancer *in vivo*. When considering the hypothesis that is the focus of this chapter, it appears to be the case that LINE-1 activity and APOBEC activity are not unrelated in this p53-deficient state. However, whether this relationship has the expected directionality,

with LINE-1 promoting APOBEC activity, is a question that is not answered simply by the data. On one hand, the results of the qRT-PCR experiments are in keeping with this hypothesis. They indicate that LINE-1 mRNA upregulation is accompanied by APOBEC3B mRNA upregulation when p53 is inactivated, and that the inverse might also true, where cells are treated with Nutlin-3a. Since generating these data, the results on APOBEC3B mRNA expression using p53 inactivation and stabilisation have also been reproduced by an independent group, both in HCT116 cells and in breast cancer cell lines (Periyasamy et al. 2017). In addition, treatment with the LINE-1 inhibitors 3TC and d4T leads to a reduction in cytosine deaminase activity, only in p53-deficient cells where LINE-1 derepression has occurred.

However, one conflicting experiment is where cells are treated with AZT. Only p53-deficient cells with LINE-1 derepression show a change in APOBEC3B mRNA expression when treated with AZT, supporting the notion that LINE-1 and APOBEC3B are mechanistically linked. However, the effect is strongly in the opposite direction to that predicted. As described in the text giving the rationale for trialling 3TC and d4T, this unexpected result seems likely to be attributable to the formation of short ssDNAs that are generated when AZT is used as a monotherapy in states of LINE-1 derepression. This paradoxically promotes, rather than inhibits, innate immune mechanisms implicated in autoinflammation (Beck-Engeser, Eilat, and Wabl 2011) such as those that could mediate APOBEC3B upregulation. However, while this reasoning might be plausible, it is as yet unconfirmed by experimentation in the experimental conditions described in this chapter.

Another conflicting piece of evidence is that HCT116 $p53^{+/+}$ and HCT116 $p53^{-/-}$ cells show similar levels of cytosine deaminase activity, despite the p53-deficient cells showing elevated levels of APOBEC3B mRNA. From the evidence gathered so far, it is not clear why this is the case. Although HCT116 cells have been reported to express APOBEC3B rather than APOBEC3A (Papatheodorou et al. 2017), one explanation is that APOBEC3A expression, which was not measured at the mRNA level, could be contributing to the enzymatic activity measured in the deaminase assay.

Looking retrospectively at the data accumulated, further experimentation could help to clarify the points raised by these conflicting results. For example, no data has been gathered to assess how APOBEC3A varies at the mRNA level as p53 and LINE-1 are varied, or as cells are treated with AZT. This is also true for the question of whether APOBEC3A or APOBEC3B mRNA expression is affected by treatment with 3TC and d4T, or whether AZT or Nutlin-3a treatment leads to a change in cytosine deaminase activity concordant with the observed changes in APOBEC3B mRNA they appear to elicit. Such experiments would provide greater context for the dynamics of how APOBEC3A and APOBEC3B expression combine to produced cytosine deaminase activity in these cells, as p53 status and drug treatments are varied. However, such experiments might not provide definitive clarity as to whether LINE-1 activity promotes APOBEC activity, with progress in methods and knowledge in the field providing alternative routes that could be pursued.

On balance, considering the evidence accumulated thus far, it seems that the results provide some support for the notion that LINE-1 activity promotes APOBEC activity, but that that support can be considered equivocal. However, with the exception of the observation that p53 deficiency appears not to influence cytosine deaminase activity, the data gathered support the notion that LINE-1 activity is at least mechanistically linked to APOBEC activity in cancer.

Aside from the hypothesis investigated as the aim of this chapter, it is of note in and of itself that reverse transcriptase inhibitors appear to modulate APOBEC activity in cancer cells, in a manner that appears to be dependent on LINE-1 activity. This is observed using two different measures of APOBEC activity, and it appears that APOBEC activity can be promoted or inhibited depending on the RTI used. This observation could have clinical utility, as it provisionally identifies RTI treatment as a novel method of modulating mutagenesis in a wide range of cancer types. An experiment that could be used to evaluate the robustness of the observation that RTIs modulate APOBEC activity is to use the high-throughput cytosine deaminase assay developed to test a range of RTIs, either alone or in combination, on their ability to modulate APOBEC activity in p53-deficient cancer cell lines from a range of tissues of origin.

It appears likely that the high-throughput cytosine deaminase assay, in its final form presented towards the end of the chapter, could be used to produce such results. The use of novel ssDNA controls expand upon what is described in the literature to enable interpretation of the assay, and

confirm that the variables optimised lead to an assay with an improved dynamic range. The results of the gel-based deaminase assay suggest that the HCT116 cells used might exhibit a low level of APOBEC activity. This could be evaluated by comparing the activity from these cells to that elicited using a recombinant APOBEC protein positive control. Using other cell lines with high endogenous APOBEC overexpression would confirm the ability of the high-throughput assay to detect APOBEC activity in cells, and therefore any modulation that might occur with RTI treatment. Looking further, if the ability of RTIs to modulate APOBEC activity is confirmed, then subsequent experiments could include testing whether continuous RTI treatment leads to a detectable change in the APOBEC signature as measured by NGS, or testing whether any modulation in genomic mutagenesis results in effects on cellular fitness.

A number of experiments could also be performed in future to strengthen the results already detailed in this chapter. Firstly, given risks of genetic drift and selection in continuous cell culture, genotyping both the wild-type and knockout HCT116 cells would have benefit in confirming the copy number of genes of interest, particularly in p53-deficient cells with heightened genomic instability. For example, it may be the case that LINE-1 and APOBEC3B expression is elevated in p53-deficient cells simply because there are more copies of each, with implications for the interpretation of the apparent differences in fold change seen in qRT-PCR experiments. To mitigate against this, experiments involving transient p53 inactivation, for example through the use of siRNA targeting the TP53 gene or overexpression of MDM2, would allow for p53 signalling to be disrupted in

the short term without allowing time for substantial genomic changes to occur. Alternatively, wild-type p53 could be transiently expressed using an exogenous construct in knockout cells, which would be expected to rescue the phenotypes observed. In general, experiments involving p53, including the experiments in this chapter where cells are treated with Nutlin-3a, would benefit from assaying whether cells display MDM2/4 amplification, Western blots assaying the abundance of p53 protein in different experimental conditions, and an assessment of whether changes in proliferation or cell cycle profiles might confound the results obtained. In addition, the capacity for the RTIs used to inhibit LINE-1 activity described in the literature could be confirmed in the HCT116 cells used. Lastly, repeating the experiments conducted would yield a greater sample size, thereby improving the statistical confidence of the results obtained.

3.2. Results Chapter 2: Bioinformatic analyses investigating APOBEC regulation in large genomic datasets

3.2.1. Introduction

The work presented in this chapter was performed in an attempt to identify possible regulators of APOBEC activity in cancer. The chapter describes exploratory analyses of large genomic datasets. These datasets were produced and curated by a number of consortia as resources that could be used for further study by researchers in the field. The work in this chapter makes use of bioinformatic tools designed for the analysis of these datasets. The datasets and tools used in this chapter are summarised in this section. Each analysis is then described in greater detail in subsequent sections. The analyses conducted in this chapter aimed to determine:

- What gene expression is associated with APOBEC3A expression, APOBEC3B espression and the number of APOBEC signature mutations in cancer.
- Whether APOBEC3 expression varies with p53 deficiency in cancer.
- Whether human APOBEC3 promoters might differ from those of chimpanzees and bonobos.
- Whether there might be transcriptional regulators of APOBEC3B that recurrently bind to the locus in multiple human cell types.

Sections 3.2.2 and 3.2.3 describe the analysis of data from the Cancer Genome Atlas pan-cancer analysis project (TCGA; Weinstein et al. 2013). The TCGA dataset was generated by an international consortium that has profiled thousands of cancers, across a range of cancer types, at genomic, epigenomic, transcriptomic, and proteomic levels. This dataset was chosen for study as it represented the most comprehensive molecular characterisation of patient cancers available at the time of conducting the analyses in this chapter. It was reasoned that associations derived from these descriptive data could plausibly be used to infer possible mechanisms of APOBEC regulation in cancer. Such inferences were contrasted with data that could be gained from experiments on cultured cancer cells. It was reasoned that causative regulatory relationships could in principle more readily be drawn from experiments in cell culture, but such data might not reflect biological process that occur *in vivo* as accurately as data drawn directly from patient samples in the TCGA cohort (this elaborated on in section 3.1.1, the introduction to the first results chapter). Somatic mutation and RNA sequencing data for multiple cancers in the TCGA dataset were used in the analyses described in this chapter, in an attempt to identify possible regulators of APOBEC activity. Analysis of gene expression using RNA-seq data in section 3.2.2 makes use of the PANTHER (Protein Analysis Through Evolutionary Relationships) classification tool, which allows for the identification of statistically overrepresented gene ontologies in lists of candidate genes (Mi et al. 2013).

Sections 3.2.4 and 3.2.5 make use of another resource, the University of California Santa Cruz (UCSC) genome browser, to study genetic and epigenetic data. These data include the reference sequence of the human genome and location of human genes, as well as the reference sequences of the chimpanzee and bonobo genomes. Epigenetic data were derived from either the Encyclopaedia of DNA Elements (ENCODE) dataset (ENCODE Project Consortium 2012), the Chromatin Immunoprecipitation Sequencing Atlas (ChIP-Atlas) dataset (Oki et al. 2018) and the ReMap dataset (Hammal et al. 2022). These datasets collate publicly available epigenetic data, such as ChIP-seq data, in attempt to build curated, high-quality catalogues of regulatory genetic regions in the human genome. These data were used in the analyses in this chapter in an attempt to identify putative APOBEC3 regulatory regions and associated transcriptional regulators.

Lastly, the design of the analyses in this chapter is premised on the rationale described in the thesis introduction. That is, that APOBEC3s, particularly APOBEC3A and APOBEC3B, have been implicated as likely mediators of APOBEC signature mutations in cancer that might be overexpressed as a result of LINE-1 activity. As APOBEC3B was thought likely to play a leading role in causing APOBEC signature mutations in cancer, there is a tendency towards investigating APOBEC3B in the design of analyses in this chapter. In addition, in the discussion section of this chapter, the data are discussed in relation to the possible role of LINE-1 as a driver of APOBEC activity.

3.2.2. Investigating gene expression associated with APOBEC3A expression, APOBEC3B expression or the number of APOBEC signature mutations in the TCGA dataset

This section describes bioinformatic analyses conducted using somatic mutation and RNA-seq data from the TCGA dataset. Here, an exploratory analysis was performed to investigate if certain gene expression profiles were associated with either APOBEC3A expression, APOBEC3B expression or the number of APOBEC signature mutations. At the time of conducting this analysis, results such as these had not been previously reported. It was hypothesised that generating these results might, for example, reveal gene expression profiles that differentiated cancers that express APOBEC3A from those that express APOBEC3B. It was also reasoned that the associations produced by such an analysis might point towards possible mechanisms of APOBEC3A and APOBEC3B regulation in cancer - confounders notwithstanding.

RNA-seq and somatic mutation data were sourced from TCGA for 2,840 patients (summarised in Table 3.2.1). The patients selected were diagnosed with one of of five cancer types that are described to have a relatively high rate of APOBEC signature mutations. Cancers associated with viral infection, such as head and neck cancers and cervix cancers, were excluded in attempt to find APOBEC regulators of non-viral aetiology.

Cancer type	Number of patients
Bladder	427
Breast	979
Lung adenocarcinoma	576
Lung squamous cell carcinoma	551
Ovary	307

Table 3.2.1 - Cancer types and patient numbers used for the analysis of TCGA data.

The expression of APOBEC3s in these samples, as determined by RNA-seq, is presented in Figure 3.2.1. These data underwent normalisation by TCGA prior to download, using the transcripts per million (TPM) method which accounts for technical variability due to gene length and sequencing depth (Zhao et al. 2021). For this analysis, APOBEC3 expression was then also normalised to the expression of TBP (the TATA-Box Binding Protein housekeeping gene) in each sample. Although housekeeping genes such as TBP are thought to represent genes whose expression is constant across conditions, there is some variability when RNA-seq normalisation using different housekeeping genes is compared (Eisenberg and Levanon 2013). TBP was chosen to allow for comparison with reports in the literature, which typically use TBP for normalisation of APOBEC3 expression in RNA-seq or qRT-PCR data.

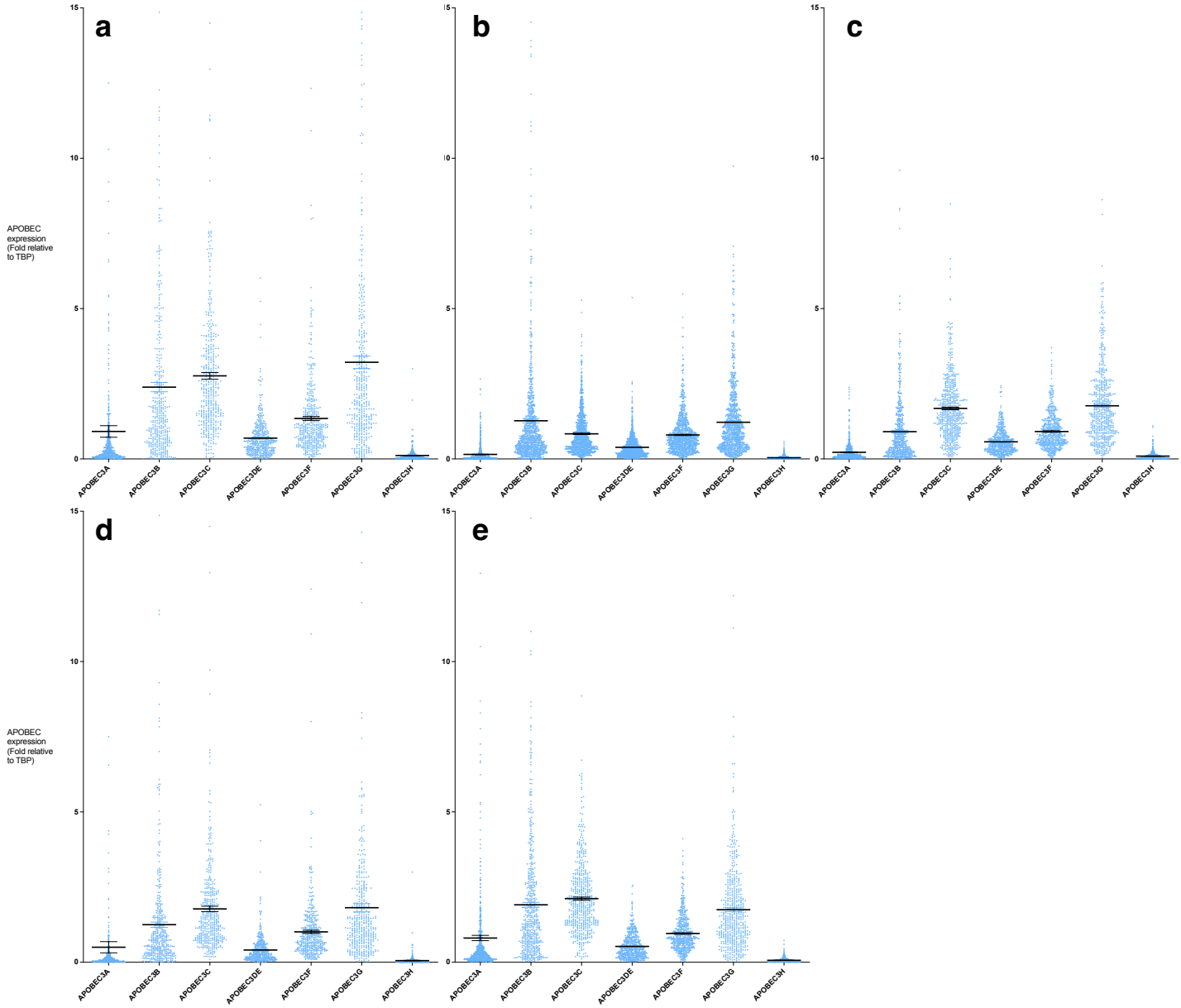


Figure 3.2.1 - TPM-normalised expression levels of APOBEC3 genes normalised to the expression of TBP across the five cancer types studied in the TCGA dataset. a = breast, b = bladder, c = lung adenocarcinoma, d = lung squamous cell carcinoma, e = ovarian.

Somatic mutation data were derived from exome sequencing performed by the TCGA study. A total of 542,170 single base substitutions were identified across all 2,840 samples. Mutation calling was performed by TCGA prior to download using Mutect2, a widely-used variant calling method that is part of

the GATK best practices pipeline that is also used in this thesis (DePristo et al. 2011). Other methods are also used in the field as there is presently some variability in the performance of variant calling algorithms (Supernat et al. 2018).

The number of mutations at TCA residues in each cancer type is given in Figure 3.2.2. Mutations at TCA residues were chosen as a measure of the number of APOBEC signature mutations, as opposed to mutations at TCW residues, due to limited computing power available at the time of completing this analysis (the time taken for processing results for a single cancer type was on the order of weeks). It was noted that APOBEC activity could potentially cause C>T, C>G or C>A mutations (Morganella et al. 2016) and APOBEC expression was associated with mutations at C residues in general as well as at TCW residues (Burns et al. 2013). As a result, any point mutation occurring at a TCA residue was chosen as a measure of APOBEC signature mutations, rather than the more restricted definition of C>T or C>G mutations at these residues that characterise SBS2 and SBS13.

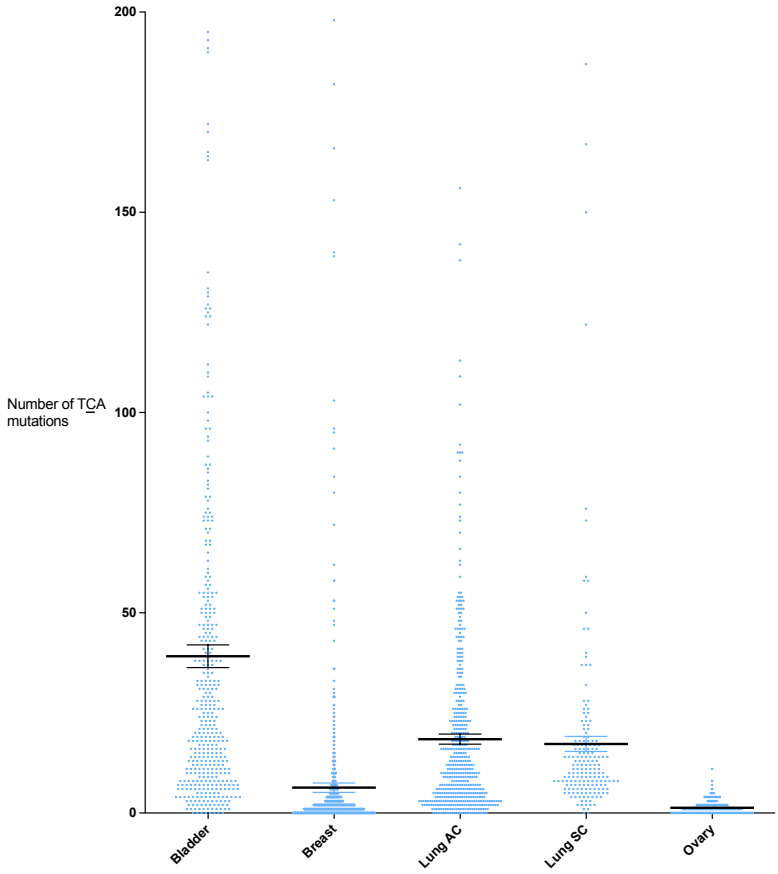


Figure 3.2.2 - Number of point mutations at TCA residues across the five cancer types studied in the TCGA dataset. AC: adenocarcinoma, SC: squamous cell carcinoma.

APOBEC3A expression, APOBEC3B expression and the number of APOBEC signature TCA mutations were each separately correlated to the expression of all other genes using a Spearman's rank correlation. A Spearman correlation was chosen over a Pearson correlation in an attempt to account for non-linear differences in the distribution of the expression levels of all genes. Genes with statistically significant correlations ($\sigma_{rs} <$ 0.04) and which were amongst the top 20 most highly correlated genes were then analysed using the PANTHER tool to identify statistically

significant enrichments ($p < 0.05$) in gene ontologies. Statistically significant enrichments are determined using a binomial test comparing the expected and observed frequencies of gene ontologies in lists of candidate genes (Mi et al. 2013). The results are presented below.

Bladder		Breast		Lung Adenocarcinoma		Lung Squamous Cell Carcinoma		Ovarian	
Gene	ρ	Gene	ρ	Gene	ρ	Gene	ρ	Gene	ρ
ISG20	0.472	OASL	0.699	IFIT3	0.545	IL1RN	0.621	OASL	0.672
PLAUR	0.467	CCL8	0.680	RSAD2	0.532	S100A12	0.569	CCL8	0.659
HBEGF	0.462	CXCL10	0.664	OASL	0.515	FFAR2	0.530	IFI35	0.650
OASL	0.459	CXCL11	0.662	LILRB2	0.510	TGM1	0.527	IFIT3	0.648
MXD1	0.451	LAG3	0.659	GBP1	0.509	S100A8	0.522	RSAD2	0.647
S100A12	0.451	CXCR2P1	0.657	LILRA5	0.508	SAMD9	0.522	LILRA5	0.636
RSAD2	0.448	EPSTI1	0.654	IFI30	0.508	SPRR2E	0.506	CMPK2	0.635
BCL2A1	0.445	TAP1	0.653	IFIT2	0.504	EMP1	0.503	PLSCR1	0.630
LAMP3	0.445	IFI44	0.652	STX11	0.501	IL1F6	0.500	GPBAR1	0.628
PRDM1	0.444	MX2	0.651	PILRA	0.499	MXD1	0.494	CD80	0.626

Table 3.2.2 - 10 genes showing the strongest positive correlation with APOBEC3A expression and their Spearman coefficients in each cancer.

Bladder	Breast	Lung Adenocarcinoma	Lung Squamous Cell Carcinoma	Ovarian
negative regulation of viral genome replication	lymphocyte chemotaxis	type I interferon signaling pathway	neutrophil aggregation	type I interferon signaling pathway
type I interferon signaling pathway	type I interferon signaling pathway	cellular response to type I interferon	sequestering of zinc ion	cellular response to type I interferon
cellular response to type I interferon	cellular response to type I interferon	response to type I interferon	chemokine production	response to type I interferon
response to type I interferon	response to type I interferon	cellular response to interferon-gamma	leukocyte migration involved in inflammatory response	negative regulation of viral genome replication
regulation of viral genome replication	defense response to virus	defense response to virus	chronic inflammatory response	lymphocyte chemotaxis
negative regulation of viral life cycle	response to virus	response to virus	defense response to fungus	regulation of viral genome replication
negative regulation of viral process	defense response to other organism	defense response to other organism	positive regulation of inflammatory response	negative regulation of viral process
negative regulation of multi-organism process	cytokine-mediated signaling pathway	immune effector process	keratinocyte differentiation	negative regulation of viral life cycle
defense response to virus	response to other organism	cytokine-mediated signaling pathway	leukocyte chemotaxis	defense response to virus
response to virus	response to external biotic stimulus	response to other organism	positive regulation of response to wounding	interferon-gamma-mediated signaling pathway

Table 3.2.3 - 10 most overrepresented gene ontology terms for APOBEC3A-linked genes in Table 3.2.2.

Bladder		Breast		Lung Adenocarcinoma		Lung Squamous Cell Carcinoma		Ovarian	
Gene	ρ	Gene	ρ	Gene	ρ	Gene	ρ	Gene	ρ
C1orf135	0.551	GTSE1	0.638	ASF1B	0.552	PLEKHG6	0.614	RACGAP1	0.523
MED8	0.524	RAD51	0.621	MLF1IP	0.541	CENPM	0.567	APOBEC3A	0.511
SLC35A2	0.524	CDC45	0.614	CDCA8	0.536	CDC45	0.559	MCM5	0.508
SEC13	0.507	CDCA8	0.610	RPL39L	0.529	MCM5	0.549	KIF4A	0.508
RAD51	0.503	CENPA	0.609	CENPM	0.529	EPT1	0.540	KIFC1	0.507
CDK2	0.503	HJURP	0.607	MCM2	0.525	TUBA1C	0.538	NCAPG	0.500
CDC6	0.500	TPX2	0.605	STIL	0.525	ACTL6A	0.538	TCF19	0.499
PVRL2	0.498	CCNB2	0.605	C16orf75	0.524	SFXN1	0.535	NUSAP1	0.495
MASTL	0.497	UBE2C	0.604	MCM6	0.520	RAD51	0.534	TPX2	0.493
RAD18	0.494	KIF2C	0.604	GINS3	0.519	THOC3	0.530	PRC1	0.492

Table 3.2.4 - 10 genes showing the strongest positive correlation with APOBEC3B expression and their Spearman coefficients in each cancer.

Bladder	Breast	Lung Adenocarcinoma	Lung Squamous Cell Carcinoma	Ovarian
meiotic nuclear division	activation of anaphase-promoting complex activity	DNA replication initiation	DNA strand elongation involved in DNA replication	mitotic spindle elongation
meiotic cell cycle process	mitotic metaphase plate congression	DNA strand elongation involved in DNA replication	DNA strand elongation	spindle elongation
meiotic cell cycle	mitotic spindle assembly checkpoint	DNA strand elongation	DNA-dependent DNA replication	mitotic spindle midzone assembly
mitotic cell cycle phase	spindle assembly checkpoint	mitotic prometaphase	mitotic S phase	spindle midzone assembly
cell cycle phase	mitotic spindle checkpoint	DNA-dependent DNA replication	S phase	mitotic chromosome condensation
biological phase	negative regulation of mitotic metaphase/ anaphase transition	mitotic cell cycle phase	mitotic interphase	mitotic spindle assembly
nuclear division	negative regulation of metaphase/anaphase transition of cell cycle	cell cycle phase	interphase	microtubule cytoskeleton organization involved in mitosis
mitotic nuclear division	negative regulation of mitotic sister chromatid separation	biological phase	mitotic cell cycle phase	mitotic cytokinesis
regulation of DNA metabolic process	negative regulation of mitotic sister chromatid segregation	DNA conformation change	cell cycle phase	chromosome condensation
organelle fission	negative regulation of sister chromatid segregation	mitotic M phase	biological phase	cytoskeleton-dependent cytokinesis

Table 3.2.5 - 10 most overrepresented gene ontology terms for APOBEC3B-linked genes in Table 3.2.4.

Bladder		Breast		Lung Adenocarcinoma		Lung Squamous Cell Carcinoma		Ovarian	
Gene	ρ	Gene	ρ	Gene	ρ	Gene	ρ	Gene	ρ
KIAA1841	0.318	BEND3	0.294	MCM10	0.312	PPIAL4G	0.545	TMEM93	0.296
UAP1	0.298	CBX2	0.294	PRR11	0.308	TMEM183A	0.437	PPA1	0.286
KLRD1	0.291	CENPW	0.289	TTK	0.306	FAM58B	0.436	SSSCA1	0.282
SMC1B	0.289	FAM54A	0.287	DIAPH3	0.303	RPS3A	0.423	TMEM223	0.281
SGOL1	0.278	A2ML1	0.284	CKAP2L	0.298	C12orf41	0.412	RHEB	0.278
FAM54A	0.277	NXPH4	0.275	FAM54A	0.293	ZNF410	0.411	ZCRB1	0.278
TCAM1P	0.277	CDC20	0.275	KIAA1524	0.291	LOC441089	0.411	C11orf83	0.271
MSH2	0.274	CENPA	0.270	C11orf82	0.291	RPS5	0.411	LOC150381	0.270
RAD18	0.272	CCNE1	0.268	KPNA2	0.289	SNX11	0.409	NDUFS8	0.270
C1orf135	0.271	UBE2C	0.268	SGOL1	0.288	PRKAG1	0.399	NDUFB11	0.267

Table 3.2.6 - 10 genes showing the strongest positive correlation with the number of TCA mutations and their Spearman coefficients in each cancer.

Bladder	Breast	Lung Adenocarcinoma	Lung Squamous Cell Carcinoma	Ovarian
sister chromatid cohesion	CENP-A containing nucleosome assembly	sister chromatid cohesion	(none)	organonitrogen compound metabolic process
	organelle organization	cell cycle process		
	CENP-A containing chromatin organization	cell cycle		
	centromere complex assembly	organelle organization		
	chromatin remodeling at centromere	sister chromatid segregation		
	chromosome segregation	nuclear chromosome segregation		
	mitotic nuclear division	chromosome segregation		
	mitotic cell cycle process	single-organism organelle organization		
	mitotic cell cycle	mitotic prometaphase		
	cell cycle	mitotic nuclear division		

Table 3.2.7 - Overrepresented gene ontology terms for TCA mutation-linked genes in Table 3.2.6.

These results indicate that APOBEC3B expression is associated with cell cycle-related gene expression ($\rho \approx 0.54$, Tables 3.2.4 and 3.2.5), while APOBEC3A expression is associated with interferon-related gene expression ($\rho \approx 0.55$, Table 3.2.2 and Table 3.2.3). APOBEC signature mutations show a weaker association ($\rho \approx 0.28$) with cell cycle-related gene expression in three of the five cancers examined (Table 3.2.7 and Table 3.2.8).

Since both APOBEC3B expression and APOBEC signature mutations are associated with cell-cycle related gene expression, but APOBEC3A expression is not, these results support the notion that APOBEC3B may be more likely to mutate cancer genomes than APOBEC3A - at least in bladder cancer, breast cancer and lung adenocarcinoma. However, there are many possible confounding factors in interpreting correlations such as these that limit any such mechanistic inferences.

3.2.3. Investigating the association between p53 deficiency and APOBEC3 expression in the TCGA dataset

This section describes the use of data from the TCGA dataset to examine whether p53 inactivation is associated with changes in the expression of any APOBEC3s. This analysis was prompted by the observation that p53 deficiency was associated with APOBE3B upregulation in breast cancer (Burns et al. 2013). However, it was not clear whether this was true of other APOBEC3s, or for other cancer types with a relatively high prevalence of APOBEC signature mutations. p53 deficiency was also described to lead to LINE-1 activity in a manner that is conserved between species (Section 1.3.7; Wylie et al. 2016). As a result, it was reasoned that p53 deficiency in this dataset might act as a possible surrogate for LINE-1 activity, enabling investigation of the relationship between LINE-1 activity and APOBEC3 expression in patient cancers.

The samples from patients analysed in the previous section (detailed in Table 3.2.1 with APOBEC3 expressions distributions given in Figure 3.2.1) were also used in this analysis. However, this analysis excludes the use of ovarian cancer samples. p53 inactivation is near-ubiquitous in ovarian cancers in this dataset, in keeping with prior findings (Cancer Genome Atlas Research Network 2011). As a result, it was reasoned that comparing wild-type and mutant p53 samples may not yield representative results. Samples with wild-type p53 in this dataset may represent ovarian cancers with unusual sets of genetic alterations, which might make comparisons to other more typical ovarian cancers unrepresentative.

Samples with TP53 mutations were identified in the TCGA somatic mutation calls. This was used to calculate the fold change in APOBEC3 expression in samples with inactive p53 relative to those with intact p53. APOBEC3 expression was again normalised to TBP expression. The results are shown in Figure 3.2.3.

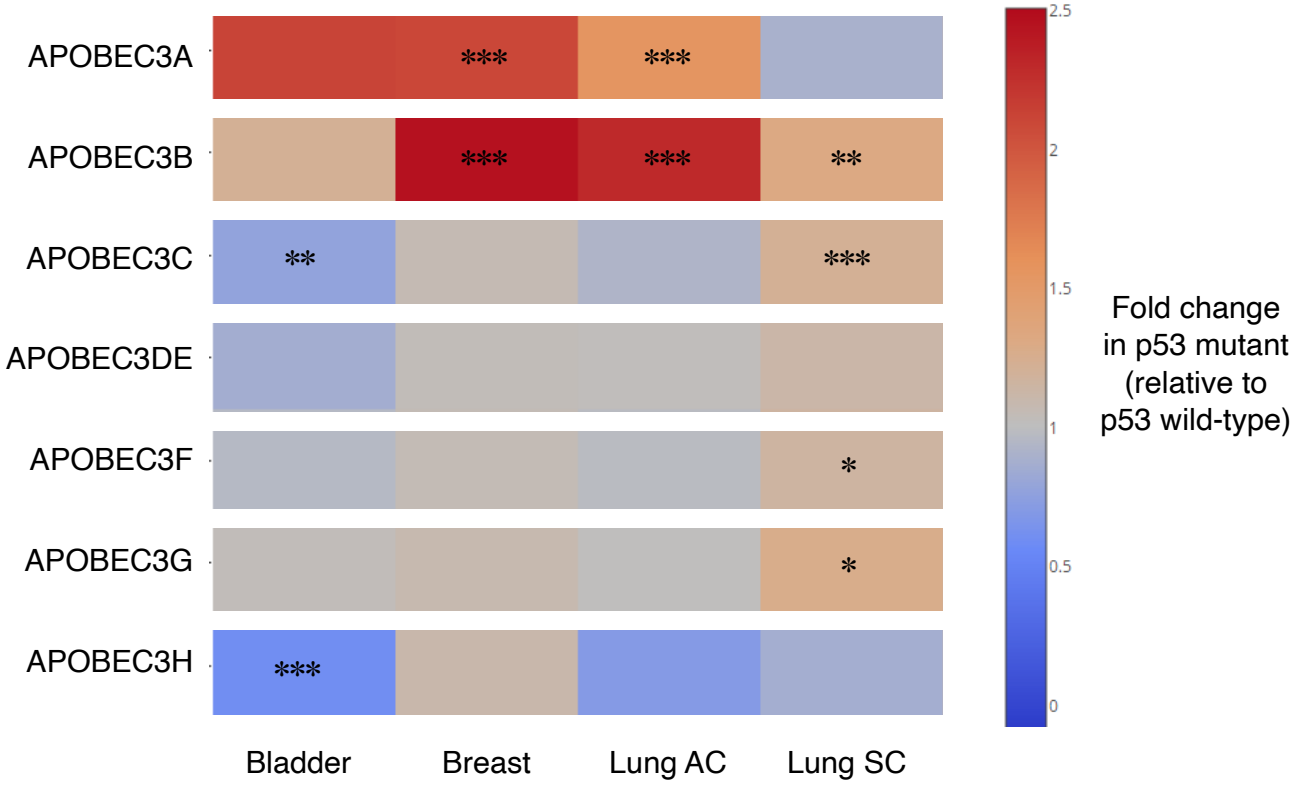


Figure 3.2.3 - Fold change in APOBEC3 expression, normalised to TBP expression, in p53-deficient cancers relative to those with intact p53. Bonferroni-corrected t-test p-values: * <0.05, ** <0.01, *** <0.001.

Figure 3.2.3 indicates that there is a statistically significant upregulation of APOBEC3B in three of four cancer types when p53 is inactivated. This relationship is found in breast cancer, as previously identified. APOBEC3A is upregulated in two of four cancers. The other APOBEC3s do not appear to be consistently upregulated or downregulated across cancers at statistical significance. These data suggest that p53 inactivation is associated with increased APOBEC3A expression and APOBEC3B expression in multiple cancers, although this association may or may not be causal.

3.2.4. Identification of APOBEC3 promoter regions in the human genome and comparison to chimpanzee and bonobo genome sequences

As described in the thesis introduction, a review of the literature accumulated indicated that APOBEC3B was a likely mediator of APOBEC signature mutations in cancer (section 1.6.3). APOBEC3B regulation was also reported to differ between humans and our nearest evolutionary relatives, chimpanzees and bonobos (section 1.7.3; Marchetto et al. 2013). This prompted the analysis in this section, which aimed to identify the APOBEC3B promoter region and compare its sequence between species. This was achieved using the UCSC genome browser. In this section, ChIP-seq data is used to verify that sequences in the immediate proximity of APOBEC3 start codons are associated with regulatory activity.

As shown in Figure 3.2.4, the APOBEC3B gene was identified within the APOBEC3 locus on chromosome 22 q13.1 using the latest reference version of the human genome (build 38). The putative promoter region is (highlighted with red arrows in Figure 3.2.4) is found immediately upstream of the region encoding APOBEC3B transcript and demonstrates heightened H3K27 acetylation (a marker associated with transcriptional promotion), consistent with a role as a regulatory region.

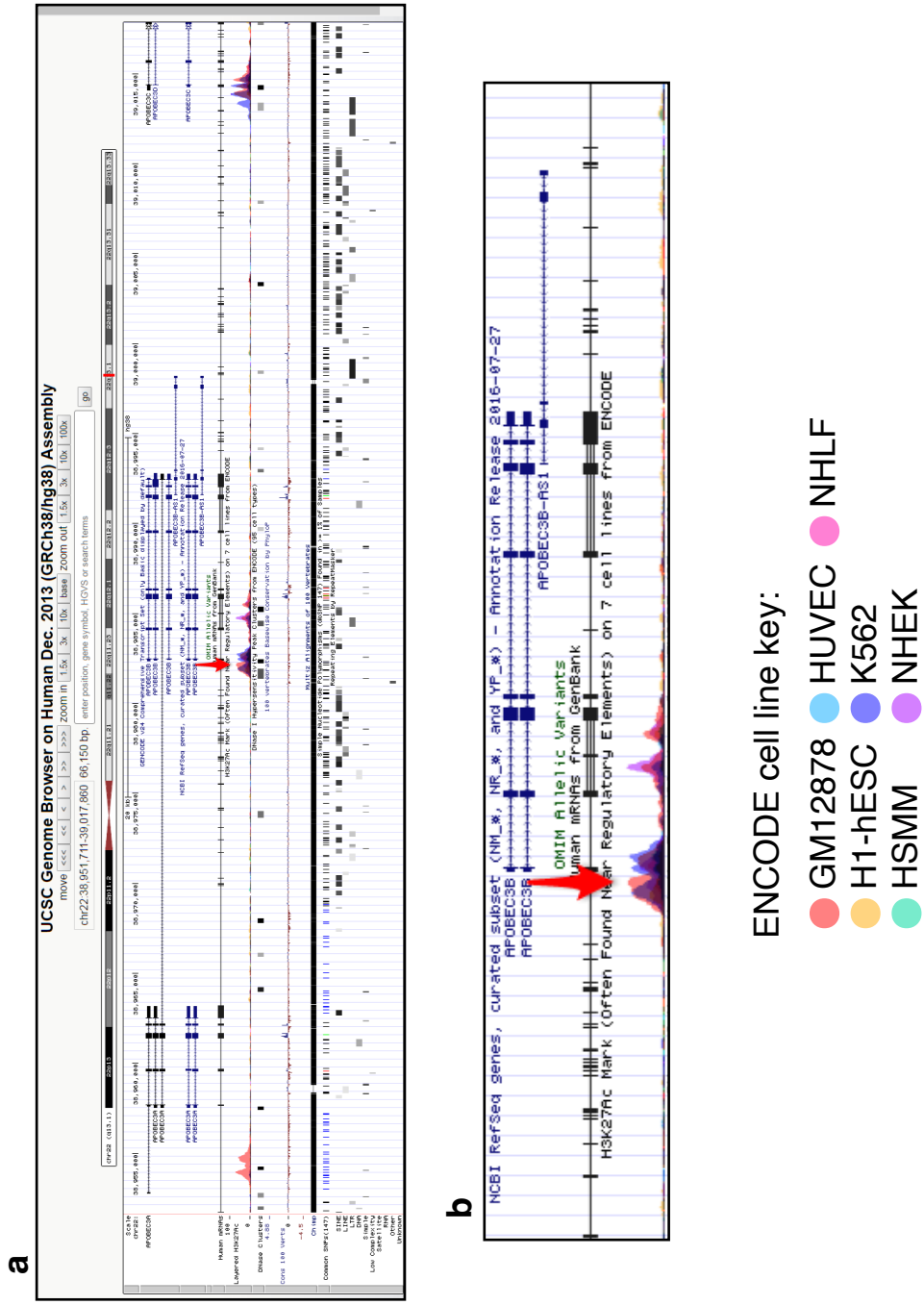
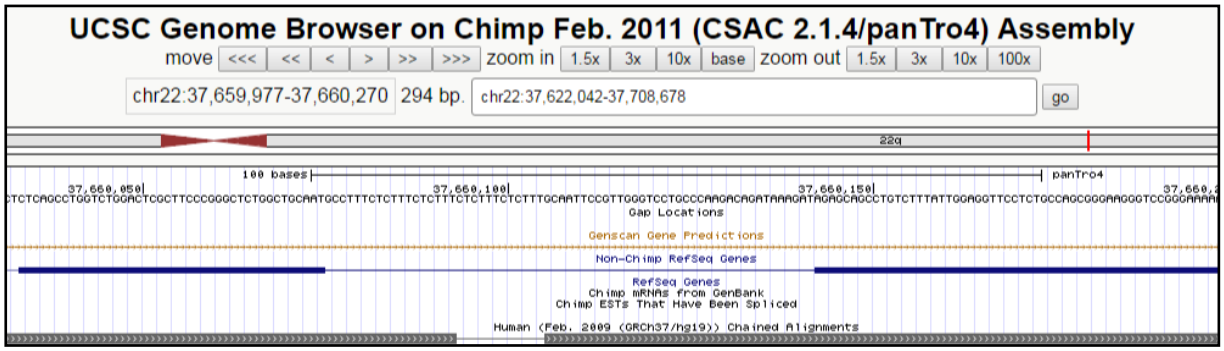


Figure 3.2.4 - UCSC Genome Browser view of the APOBEC3B region in the human genome reference build 38. (a) A broad view of the region including the upstream APOBEC3A locus and part of the downstream APOBEC3C locus. (b) A focused view of the key data from (a) for identifying the putative APOBEC3B promoter: the genomic location of the transcript as well as ENCODE data for H3K27 acetylation from seven cell lines. Red arrows in both images highlight the identified promoter region.

This region was subsequently compared to equivalent sequences in chimpanzees and bonobos, yielding identification of a deletion in the human promoter (Figure 3.2.5).

a



b

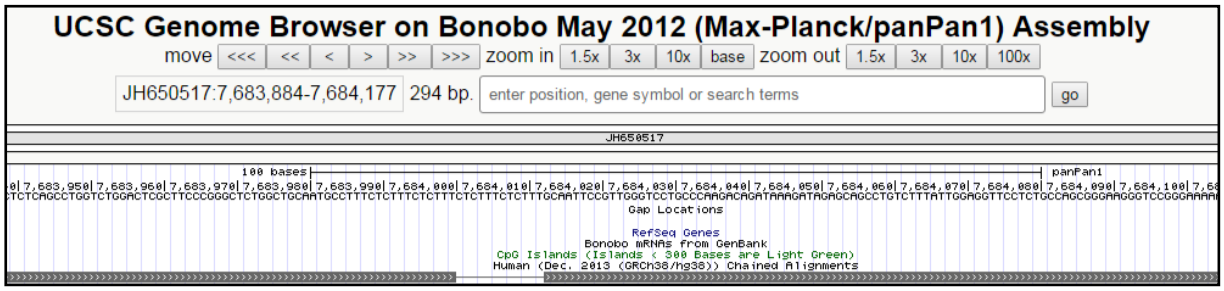


Figure 3.2.5 - UCSC Genome Browser view of deletion in the human APOBEC3B promoter region relative to the chimp (a) and bonobo (b) genomes.

The deleted sequence comprises the loss of 12 nucleotides around 2 kb upstream of human APOBEC3B start codon. The deletion eliminates approximately half of a CT-rich region. This sequence is otherwise intact in both chimpanzees and bonobos.

The sequence lost is consistent with the consensus sequence for interferon response factors (IRFs). The IRF consensus sequence is 5'-TTT-CN-NTT-3'

(Yanai, Negishi, and Taniguchi 2012). The deletion leads to the elimination of two out of four overlapping IRF consensus sequences in the CT-rich region. This is illustrated below, with the deleted region shown in red:

5'-~~CTTT-CT-CTTT-CT~~-CTTT-CT-CTTT-CT-CTTT-3'

Repeating this process systematically for all APOBEC3s, and including up-to-date ChIP-Seq data from the ReMap project, showed that there were no such deletions in the promoters of the other APOBEC3 genes. This is shown in a series of figures below, Figure 3.2.6.1 to Figure 3.2.6.7. These figures show UCSC genome browser views of human APOBEC3 promoter regions and comparisons to chimpanzee and bonobo genome sequences. ReMap density plots show the number of transcriptional regulators found to bind to that portion of DNA in the ReMap dataset. The ReMap data are filtered for quality according to cross-correlation and the FRiP (fraction of reads in peaks) metrics developed by the ENCODE Consortium (Hammal et al. 2022).

Together, these data suggest that, in the recent evolution of the human genome, APOBEC3B may have evolved to become less responsive to interferon response factors. However, this is only inferred bioinformatically, and not verified experimentally.

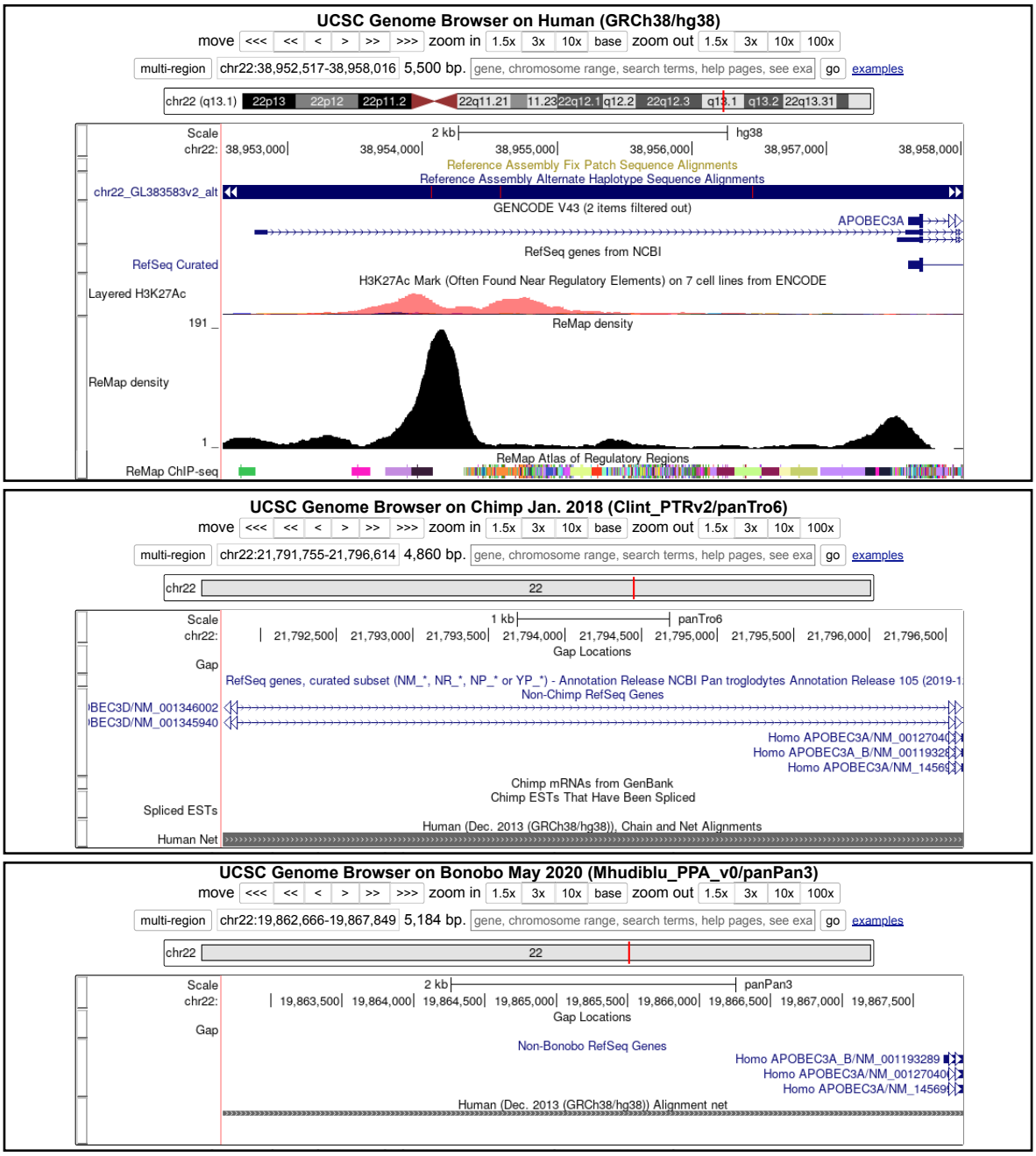


Figure 3.2.6.1 - UCSC genome browser view of human APOBEC3A promoter region and comparison to chimpanzee and bonobo sequences.

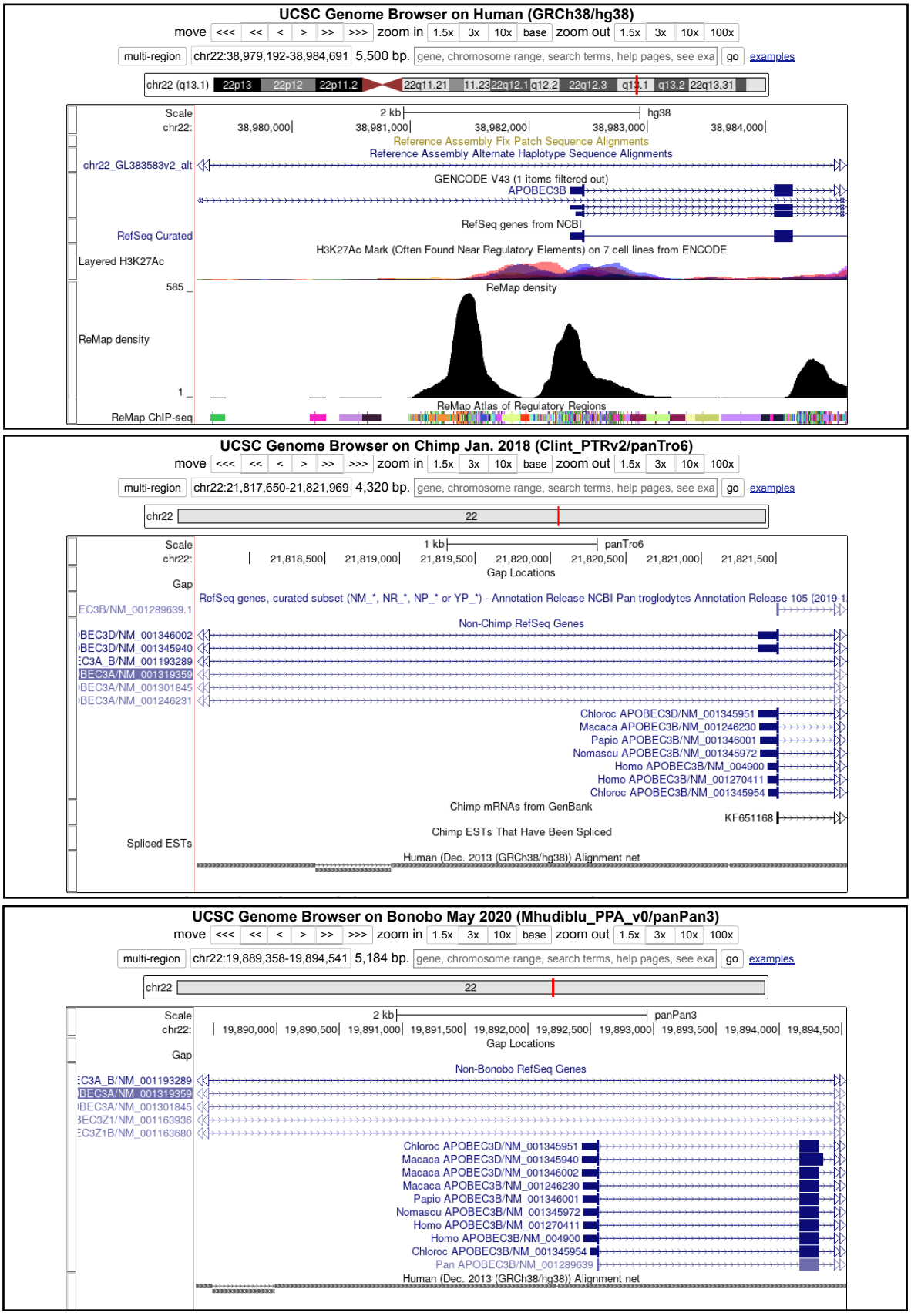


Figure 3.2.6.2 - UCSC genome browser view of human APOBEC3B promoter region and comparison to chimpanzee and bonobo sequences.

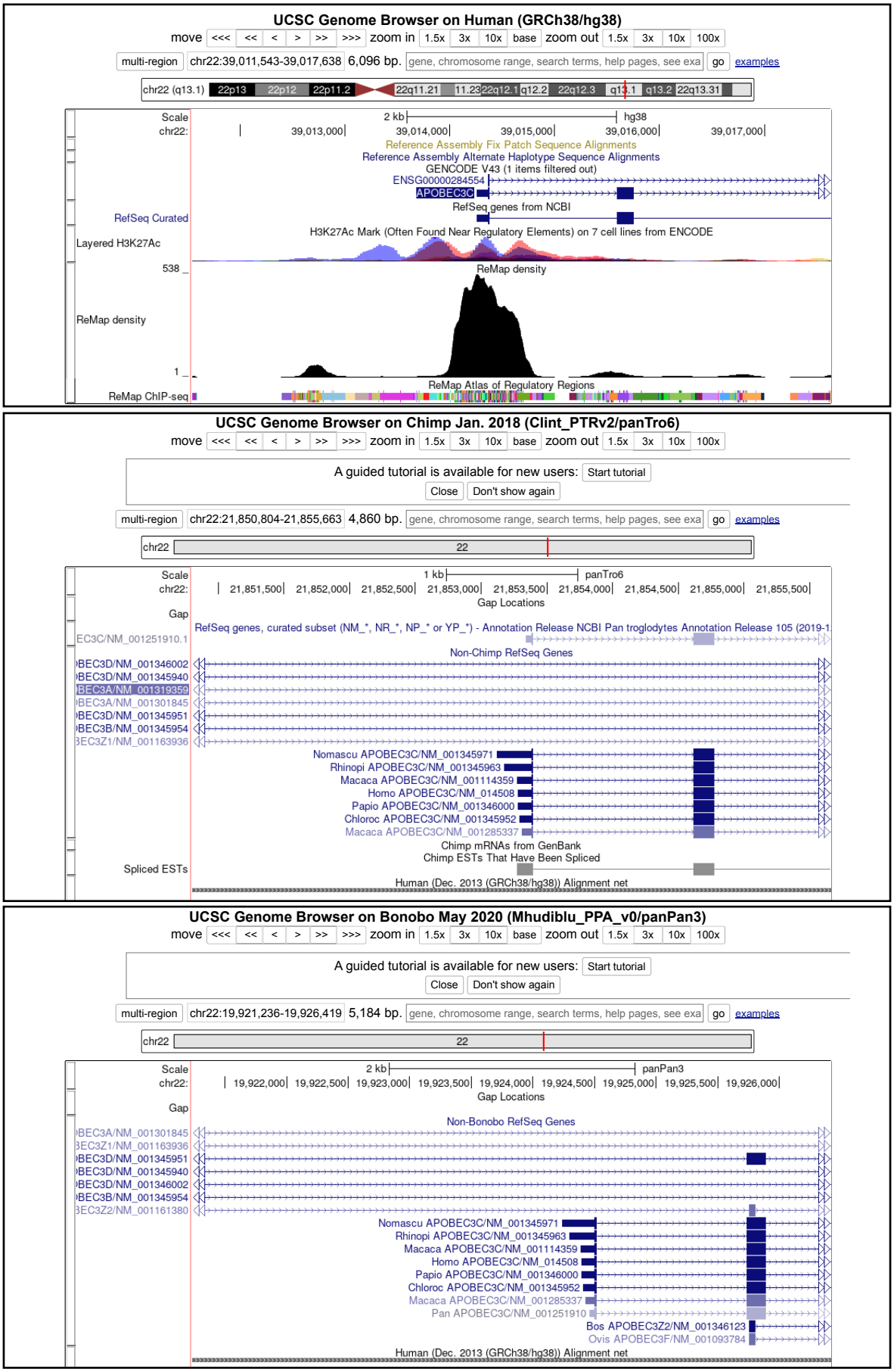


Figure 3.2.6.3 - UCSC genome browser view of human APOBEC3C promoter region and comparison to chimpanzee and bonobo sequences.

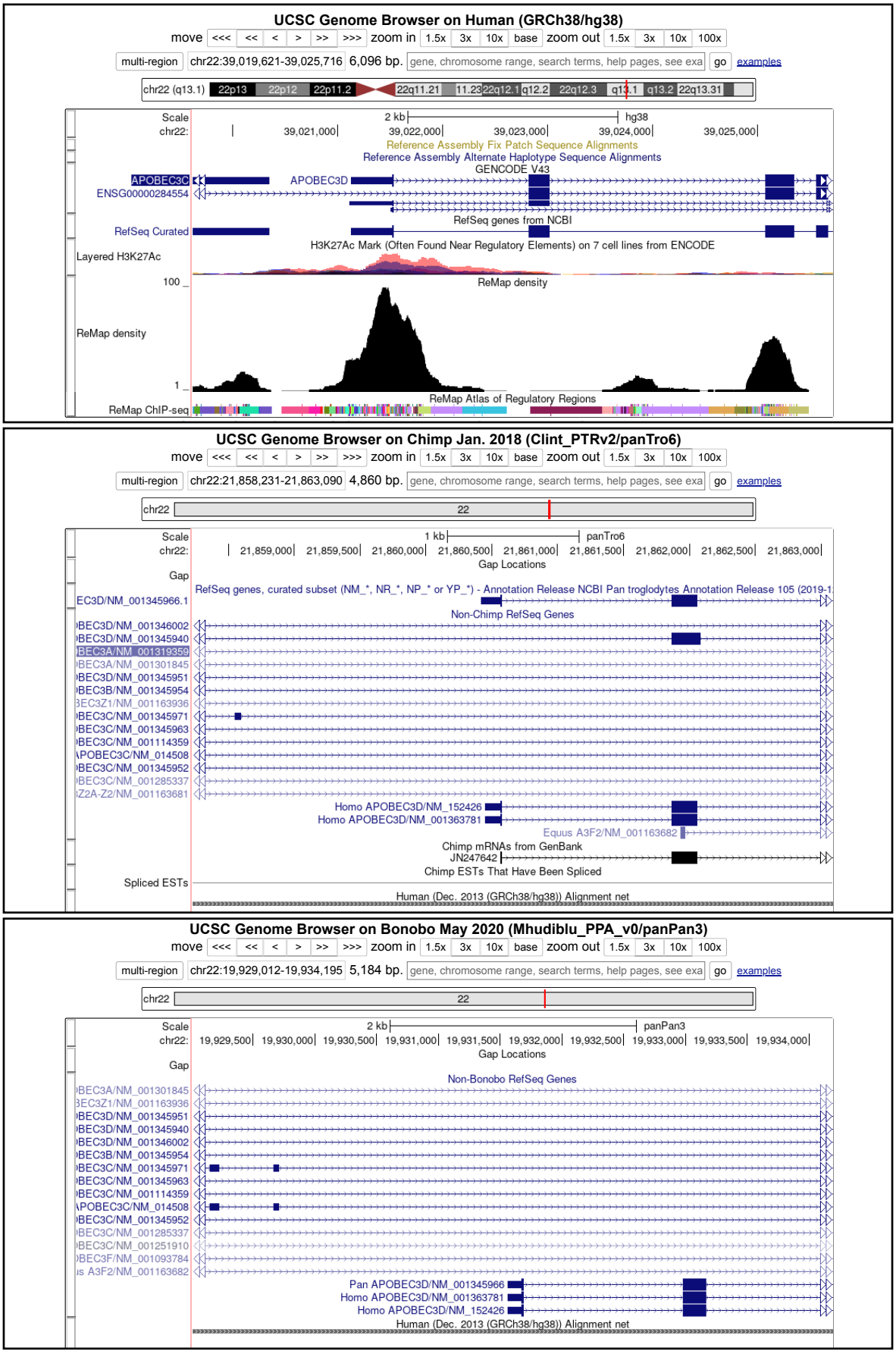


Figure 3.2.6.4 - UCSC genome browser view of human APOBEC3DE promoter region and comparison to chimpanzee and bonobo sequences.

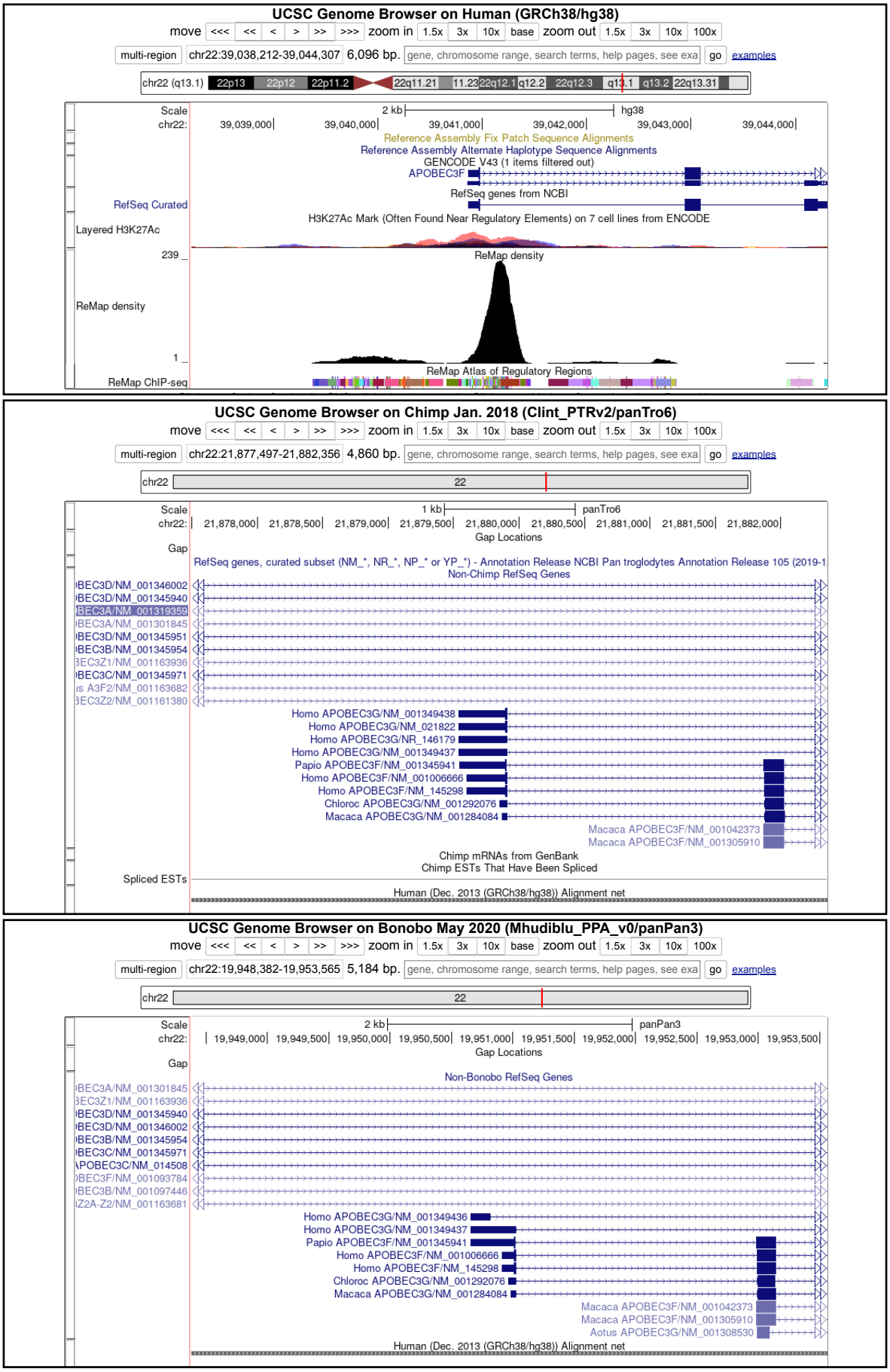


Figure 3.2.6.5 - UCSC genome browser view of human APOBEC3F promoter region and comparison to chimpanzee and bonobo sequences.

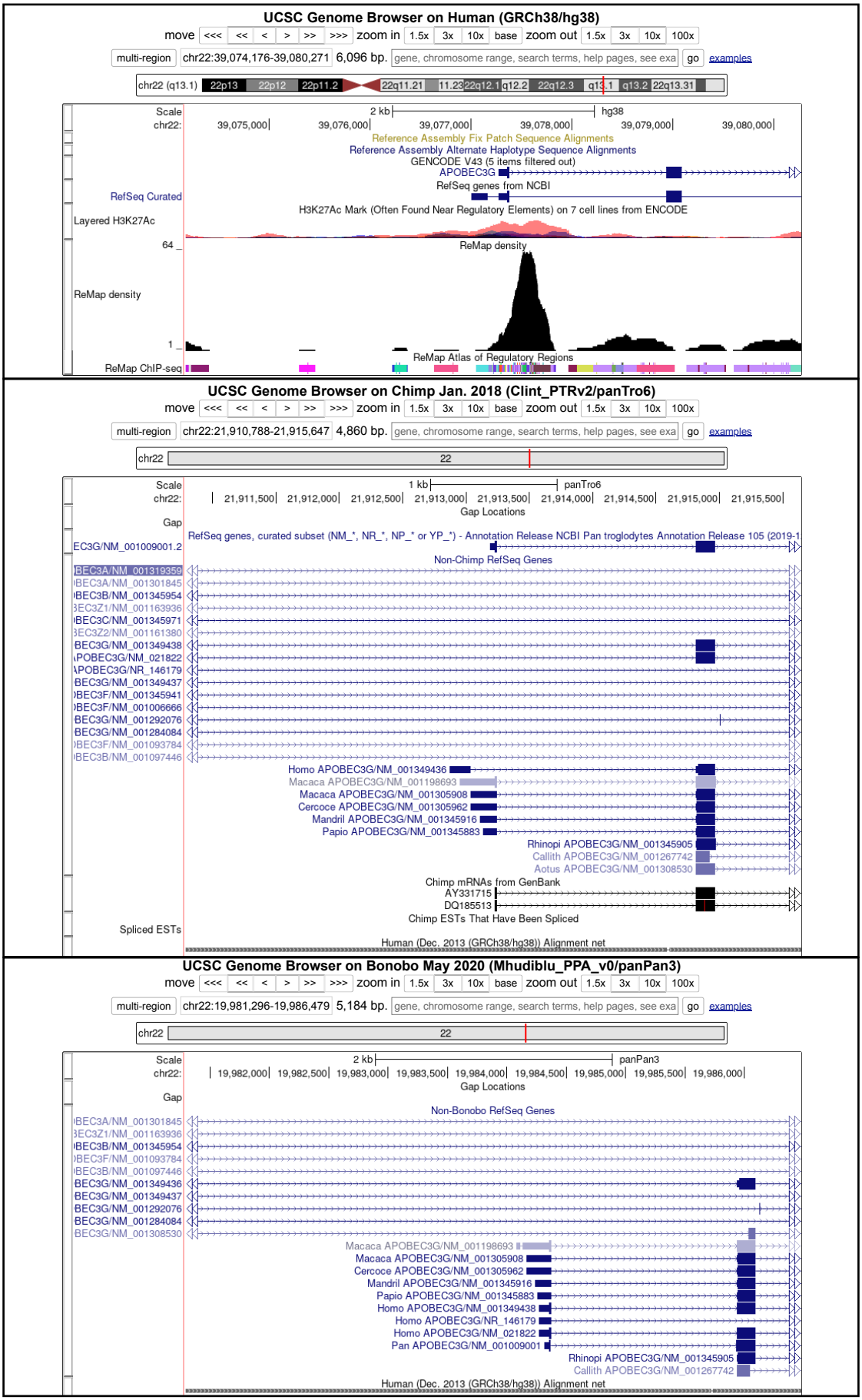


Figure 3.2.6.6 - UCSC genome browser view of human APOBEC3G promoter region and comparison to chimpanzee and bonobo sequences.

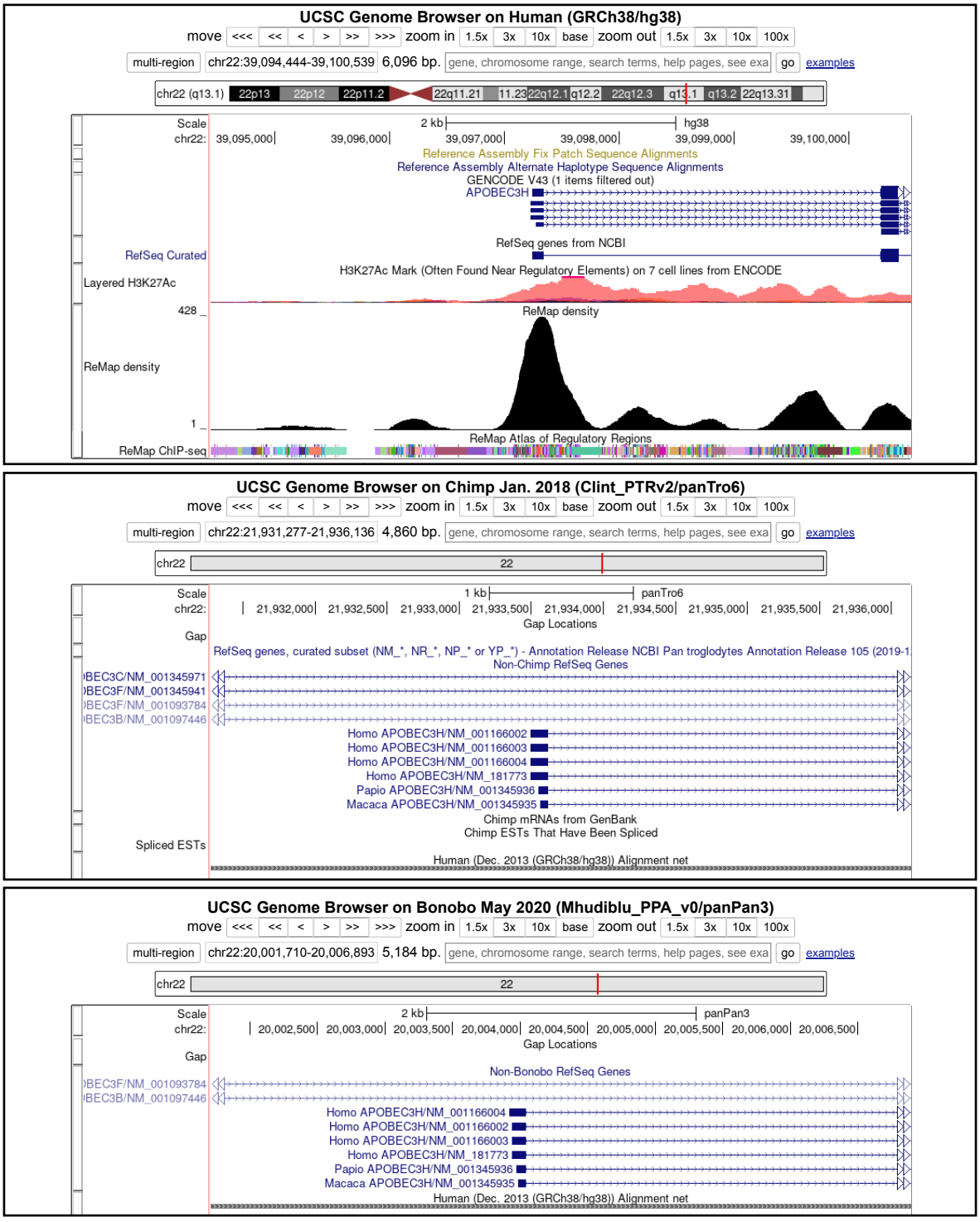


Figure 3.2.6.7 - UCSC genome browser view of human APOBEC3H promoter region and comparison to chimpanzee and bonobo sequences.

3.2.5. Identifying candidate transcriptional regulators of APOBEC3B using ChIP-Seq data

This section describes further exploratory analyses of regulatory data to identify possible regulators of APOBEC3B. Here, the UCSC genome browser is used to summarise ChIP-seq data from ENCODE project. Instead of transcription regulator binding being summarised as a density plot, as in Figure 3.2.6 in the previous section, the specific transcriptional regulators measured in a range of cancer and non-cancer cell lines are listed. This is shown for the APOBEC3B locus in Figure 3.2.7. It was reasoned that proteins that were found to bind to this region recurrently across a range of cell lines were more likely to be *bona fide* regulators of APOBEC3B expression.

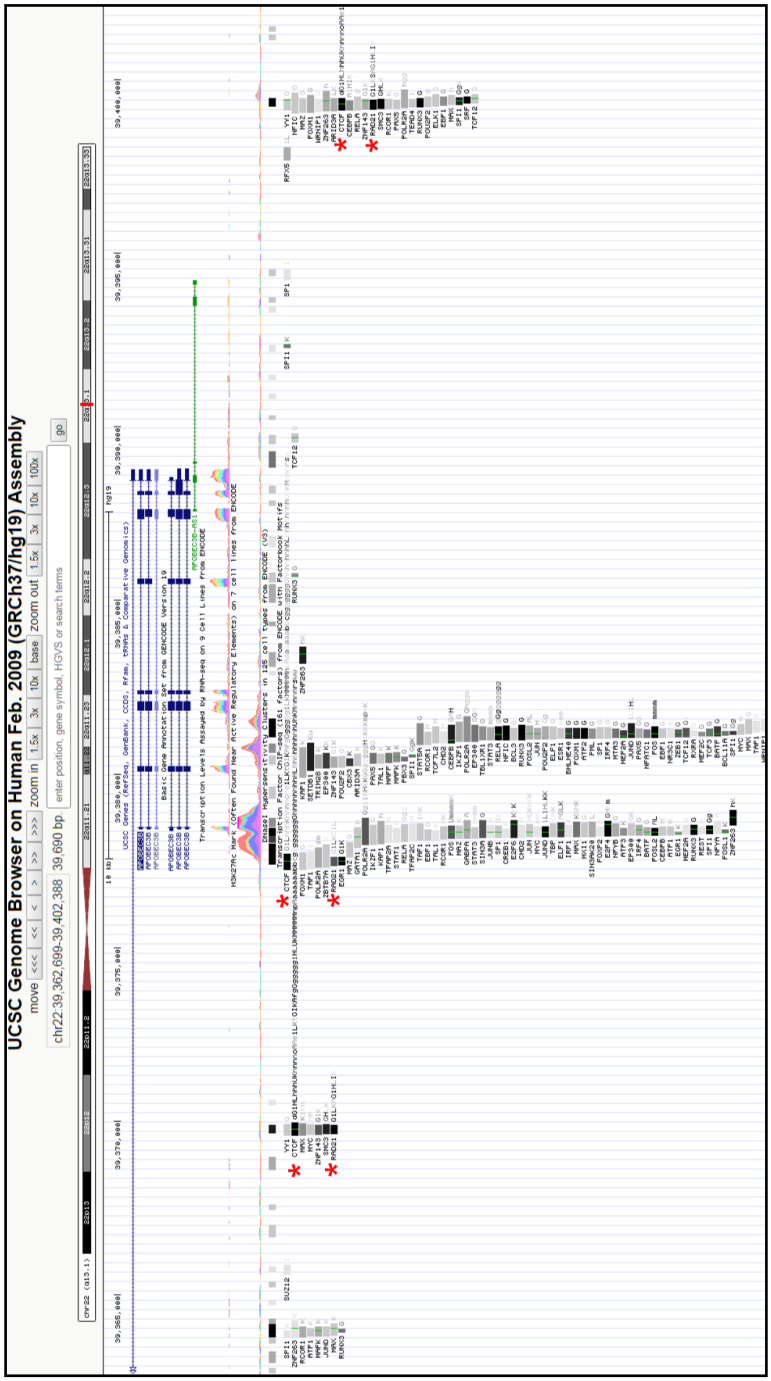


Figure 3.2.7 - UCSC genome browser view of the APOBEC3B locus, with ENCODE ChIP-seq data loaded. For each transcription factor detected to bind to the region, a grey box is given to indicate a peak cluster of transcription factor occupancy. For each transcription factor, a set of letters is given. Each letter represents a different cell line tested. The darkness of each box is proportional to the maximum signal strength observed in any cell type contributing to the cluster. Red stars indicate CTCF and its binding partner Rad21, which are discussed below.

The data in Figure 3.2.7 show that many different transcriptional regulators that appear to bind to the region. However, most do not appear to be detected consistently across different cell lines. As a result, it may be difficult to ascertain whether this might reflect cell line-specific regulation, or non-specific binding by these transcription factors.

However, in contrast, CTCF appears to bind to the region reproducibly in a wide range of cell lines, showing by far the most consistent results of the transcriptional regulators observed. Moreover, its binding partner Rad21 also appears to bind to the region in multiple cell lines. These findings support the notion that CTCF might bind and transcriptionally regulate the APOBEC3B locus.

CTCF has been characterised as a transcriptional insulator that is used to demarcate functional regions of gene expression (Ong and Corces 2014). This suggested that APOBEC3B might be transcriptionally insulated from neighbouring genes. To examine this further, these CTCF Chip-Seq data were downloaded from the ChIP-Atlas. CTCF results were filtered for high significance threshold (>500, as determined by the peak-calling algorithm MACS2) and selected for data from all cell types available. These data were visualised in the region encoding all APOBEC3 genes using the Integrative Genomics Viewer (Figure 3.2.8).

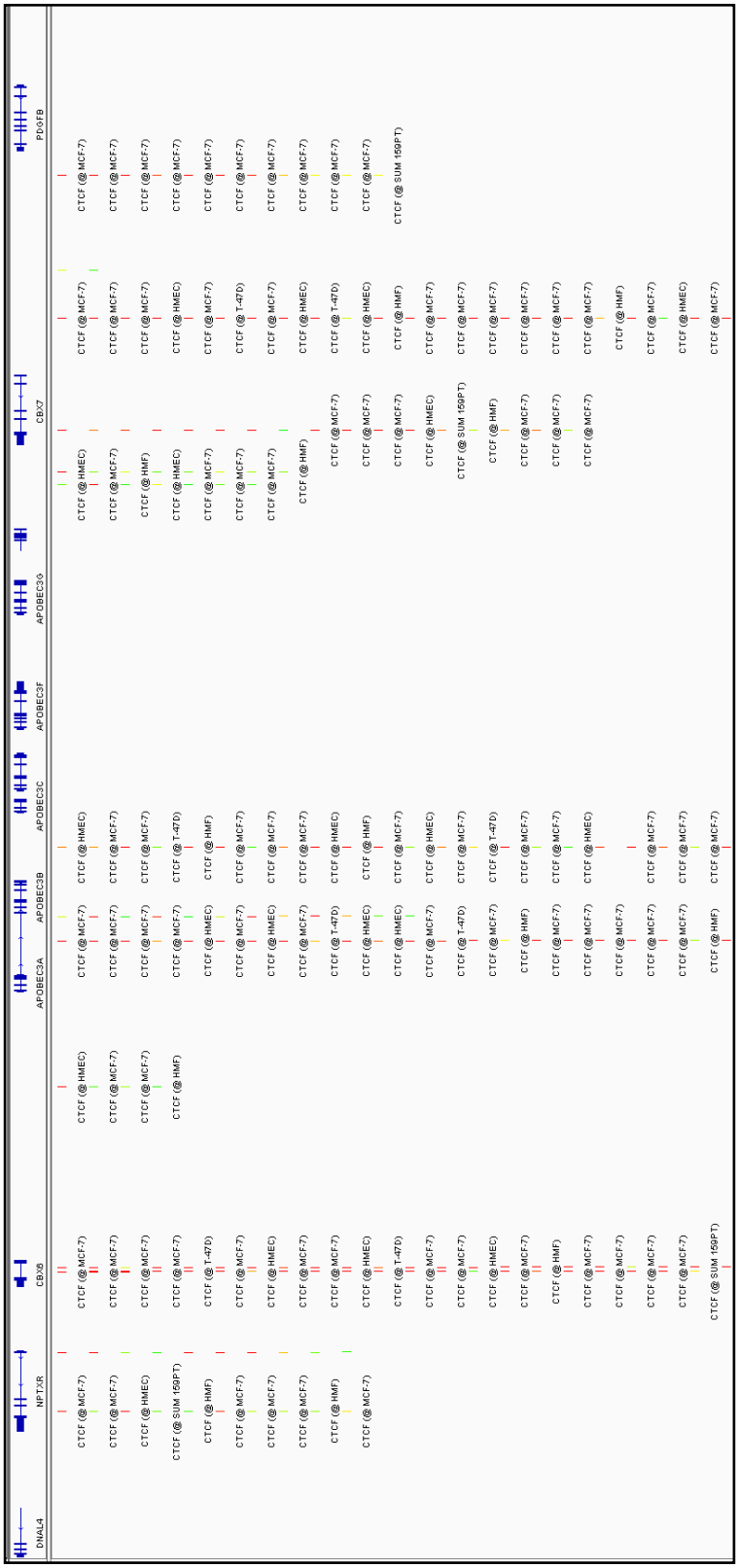


Figure 3.2.8 - A representative Integrative Genomics Viewer (IGV) view of ChIP-Atlas CTCF ChIP-Seq peaks in the region encoding APOBEC3A-APOBEC3H.

The data shown in Figure 3.2.8 indicate that CTCF is found to bind to the APOBEC3B locus reproducibly in multiple cell lines, binding immediately upstream and downstream of the APOBEC3B gene. This does not appear to be the case for any of the other APOBEC3 genes. Instead, CTCF appears to demarcate the start and end of the entire APOBEC3 locus. Given the insulatory function of CTCF, these data suggest that the entire APOBEC3 region might be subject to common transcriptional regulation. However, APOBEC3B might be insulated from this regulation of syntenic APOBEC3s, and subject to different mechanisms of transcriptional control.

3.2.6. Discussion

These exploratory analyses were conducted in order to identify possible regulators of cancer-associated APOBEC activity. To summarise, the analyses conducted in this chapter suggest the following findings:

- APOBEC3A expression is associated with the expression of interferon-related genes.
- APOBEC3B expression is associated with the expression of cell cycle-related genes.
- The number of APOBEC signature mutations is weakly associated with the expression of cell cycle-related genes.
- APOBEC3A expression and APOBEC3B expression are associated with p53 deficiency.
- The APOBEC3B promoter has lost a consensus IRF binding region, while other APOBEC3s have not.
- APOBEC3B may be transcriptionally insulated from syntenic APOBEC3s by CTCF.

Across the 2,840 TCGA cancers studied, APOBEC3A expression was found to be associated with the expression of interferon-related genes in all five cancer types tested while APOBEC3B expression was associated with the expression of cell cycle-related genes in all five cancer types tested. One strength of the analysis of the TCGA dataset is the large number of samples studied across a range of cancer types. The reproducibility of these associations across cancer types suggests that the underlying mechanisms - causal or confounding - that lead to these associations may be consistent

across tissue types studies. APOBEC3 expression, particularly the expression of APOBEC3A, is known to be stimulated by interferon in experimental studies (Stavrou and Ross 2015). In addition, experimental data from cultured cells indicate that APOBEC3B expression depends on cell cycle phase (Hirabayashi et al. 2021). These studies lend weight to the possibility that the associations found in the TCGA dataset might represent causal regulatory mechanisms.

These associations, described in section 3.2.2, have been reproduced by other groups - first in breast cancer, then in a pan-cancer study (Cescon, Haibe-Kains, and Mak 2015; Ng et al. 2019). These studies also expand upon the findings described in this section by assessing gene expression associated with all APOBEC3s, not just APOBEC3A and APOBEC3B. While APOBEC3B is associated with cell cycle-related gene expression, they find that the expression of each of the other APOBEC3s is associated with interferon-related gene expression.

This is keeping with other results presented in this chapter. Firstly, that the APOBEC3B promoter region appears to have lost a consensus IRF binding region in the divergence from non-human primates, while other human APOBEC3 promoters do not appear to have any deletions. And secondly, that APOBEC3B may be insulated from the transcriptional regulation of syntenic APOBEC3s by CTCF.

Together, these data indicate that APOBEC3B expression is subject to regulation that may differ from other APOBEC3s. These data indicate that,

in general, APOBEC3 expression appears to be associated with interferon signalling in cancer. However, APOBEC3B expression appears to be associated with the expression of cell cycle genes rather than interferon-related genes. This difference could be influenced by the identified IRF binding site deletion or the insulatory action of CTCF.

Direct experimentation to test whether these relationships might be causal would be needed to verify this. For example, gene editing technology such as CRISPR/Cas9 could be used to restore the deleted IRF binding sequence in cultured cells in order to test whether this sequence might influence the interferon-responsiveness of APOBEC3B expression. CTCF function could also be disrupted in such a system to examine whether it has an influence on the interferon-responsiveness of APOBEC3B expression.

Unless there are errors in the reference sequences of the human, chimpanzee or bonobo genomes, the apparent deletion in the APOBEC3B promoter is likely to be a true observation. The CTCF binding observed is reproducible across many experiments in multiple cell types, and is associated with Rad21 binding, which suggests that that CTCF is truly likely to bind to the sequences of interest as observed. Other transcriptional regulators that may have been found to bind to the APOBEC3B promoter less reproducibly were not considered in this analysis in attempt to assess the regulatory data with a high degree of stringency. However, they may represent true regulators of APOBEC3B in particular contexts and may be of relevance to APOBEC3B regulation in cancer *in vivo*.

The results in section 3.2.2 also include data that suggest that the number of APOBEC signature mutations for cancers in the TCGA dataset are associated with cell cycle-related gene expression rather than interferon-related gene expression. This might be indicative of an association between APOBEC3B expression, which is also associated with cell cycle-related gene expression, and the number of APOBEC signature mutations, but there are multiple plausible confounding factors that could cast doubt on this possibility. There are a number of limitations of this analysis that make these suggestions less likely.

For example, it is noted that the number of APOBEC signature mutations is associated with cell cycle-related gene expression in bladder cancer, breast cancer and lung adenocarcinomas but not lung squamous cell carcinomas or ovarian cancer. Where these associations are found, the Spearman correlation coefficients are around 0.28, which is not suggestive of a strong positive correlation. The measure of APOBEC signature mutations used was mutations at T<u>C</u>A residues rather than mutations at <u>TC</u>W residues, owing to limited computing power. Using T<u>C</u>W mutations as a measure would likely be more representative of APOBEC activity and might yield greater statistical power to detect associations. To aid interpretation of this analysis, greater computing power could be used to assess what gene expression profiles are associated with different types of point mutations. This might help account for one possible confounder, which is that cells that express high levels of cell cycle genes could plausibly have more point mutations of all types.

One purpose of this analysis was to identify if an APOBEC3A-associated gene expression profile or an APOBEC3B-associated gene expression profile was also associated with the number of APOBEC signature mutations, in an attempt to investigate the possibility that either APOBEC3A or APOBEC3B might be more likely to mediate the APOBEC mutagenesis observed in cancer. However, the field has advanced since this analysis was conducted. There is now an understanding that APOBEC3A (YTCW) and APOBEC3B (RTCW) activity can be differentiated by the base two positions 5' of the target cytosine, indicating that APOBEC3A contributes more APOBEC signature mutations to cancer genomes (Alexandrov et al. 2020). As a result, an expanded analysis of the type conducted in this chapter might not provide information that would add meaningfully to what is presently known. In addition, the results obtained via PANTHER analysis, which depends on the detection of gene ontology enrichment in a pre-determined list of genes of interest, might be complemented and validated by the use of a technique such as Gene Set Enrichment Analysis (GSEA), which instead looks to see if there are statistically significant differences in gene expression for all genes represented by specific gene ontologies (Subramanian et al. 2005).

Section 3.2.3 examines the association between TP53 mutations and APOBEC3 expression. APOBEC3A expression is upregulated around 1.5-fold to 2-fold in breast cancer and lung adenocarcinoma when TP53 is mutated, while APOBEC3B expression is upregulated around 1.5-fold to 2.5-fold in breast cancer, lung adenocarcinoma and lung squamous cell carcinoma. Other APOBEC3s show no consistent pattern of upregulation or

downregulation. These results appear to confirm and expand upon the association between TP53 mutation and APOBEC3B upregulation in breast cancer reported by Burns et al. They indicate that TP53 mutations are associated with APOBEC3A and APOBEC3B in multiple cancers, with a more marked association for APOBEC3B which is found in three of four cancer types examines (as opposed to two of four for APOBEC3A) and has the highest fold change values found in all conditions.

Since TP53 mutations commonly occur in the process of carcinogenesis, these data further implicate APOBEC3A and APOBEC3B, but not other APOBEC3s, in the process of carcinogenesis. These associations suggest that APOBEC3A or APOBEC3B activity might lead to TP53 mutations, as demonstrated experimentally for APOBEC3B by Burns et al. Alternatively, these associations could also suggest that TP53 mutations might lead to APOBEC3A and APOBEC3B upregulation, as indicated by the results for APOBEC3B in the first results chapter and by Periyasamy et al. It is noted that the inactivation of p53 in HCT116 leads to an approximately 2.5-fold upregulation in APOBEC3B expression in the experiments described in the first results chapter, which a similar degree of upregulation found in the analysis of the TCGA dataset.

The analyses in this chapter were also, in part, conducted to assess the possibility that LINE-1 activity might promote cancer-associated APOBEC activity. The results in this chapter are consistent with this possibility. For example, LINE-1 activity is described to lead to an interferon response, and APOBEC3A expression is found to be associated with interferon-related

gene expression. In addition, p53 inactivation is described to lead to LINE-1 derepression, and is also associated with APOBEC3A and APOBEC3B upregulation in the TCGA data analysed.

It is thought that expression of potentially active LINE-1 elements cannot be accurately measured using standard RNA-seq methods, owing to difficulties in resolving repetitive sequences using bioinformatic tools, leading to a need for customised sequencing protocols (Deininger et al. 2016). As discussed in the introduction to the first results chapter, LINE-1 RNA expression might be a stimulus that triggers APOBEC recruitment. It might therefore be a good measure of LINE-1 for identifying a putative correlation between LINE-1 activity and APOBEC activity in cancer. This is in contrast to other measures of LINE-1 activity, such as LINE-1 methylation or the number of new LINE-1 insertions, which may be less sensitive in detecting this putative correlation. As a result, future work using bespoke RNA-seq protocols for LINE-1 RNA measurement in patient cancer samples may prove useful in assessing this relationship in greater detail.

3.3. Results Chapter 3: Bioinformatic analyses investigating APOBEC activity in Aicardi-Goutières syndrome

3.3.1. Introduction

This chapter presents the results of bioinformatic analyses that were performed to investigate whether cells from patients with Aicardi-Goutières syndrome might show evidence of cancer-associated APOBEC activity. The AGS phenotype mimics that of a viral infection, but in the absence of a causative virus. It is instead thought to be triggered by endogenous nucleic acids such as those deriving from LINE-1 elements. This is reminiscent of what is observed in cancer, where there is an apparent activation of the antiviral APOBEC3 enzymes in the absence of viral infection, and LINE-1 activity might instead be an underlying cause.

It was reasoned that LINE-1 associated Aicardi-Goutières syndrome genotypes might therefore be used as a possible model to study mechanisms of APOBEC regulation. AGS has been associated with LINE-1 activity when the underlying defects are in the function TREX1, RNase H2 or SAMHD1 (see section 1.7.4). At the time of preparing the analyses in this chapter, Lim and colleagues had published what was, to our knowledge, the first NGS analysis of AGS patient samples (Lim et al. 2015). As part of their study, they performed RNA sequencing on 14 fibroblast samples from four

patients with LINE-1 associated AGS genotypes and one age-matched health control. The four patients had pathogenic mutations in TREX1, RNASEH2A, RNASEH2B and SAMHD1, respectively. Four samples were derived from the patient with RNASEH2A deficiency and four were from the patient with RNASEH2B deficiency. The other individuals provided two samples each.

It was reasoned that these samples could be used to examine whether the pattern of APOBEC expression in AGS was similar or different to the pattern observed in cancer, which is characterised by APOBEC3B overexpression and a less marked APOBEC3A expression. Global gene expression profiles could also be assessed to determine if AGS samples show any gene expression profiles linked to cancer-associated APOBEC activity. In addition, in the absence of standard genome sequencing data, reads from these RNA-seq data could be examined for the presence of APOBEC signature mutations of the type that are found in cancer. If such mutations were indeed found, then it would be possible to test if their prevalence is positively correlated with the expression levels of APOBEC3A and APOBEC3B, as is observed in cancer.

Data from Lim et al. and the TCGA studies were downloaded in order to perform these analyses. The PANTHER classification tool, which allows for the identification of statistically overrepresented gene ontologies in lists of candidate genes, was again used, as in the previous results chapter. In addition, the Genome Analysis Toolkit (GATK) Best Practices pipeline for somatic mutation calling from RNA-seq data was used to identify whether

APOBEC signature mutations were present in these samples (DePristo et al. 2011).

After conducting the analyses in this chapter, an additional AGS RNA-seq study was published (Rice et al. 2018). This was not chosen for further study as the samples were derived from whole blood. Blood cells are known to have substantially different APOBEC3 regulation to cells from solid tissues (Refsland et al. 2010). For example, APOBEC3A is described to show little to no expression in solid tissues but is markedly expressed in blood cells, where it appears to be the most highly expressed APOBEC3 gene. Despite this, APOBEC signature mutations are usually prevalent in solid tumours rather than liquid tumours (Alexandrov et al. 2020).

3.3.2. APOBEC expression in Aicardi-Goutières syndrome

The first analysis in this chapter was conduced to assess whether APOBEC expression in AGS might be similar to or different from that observed in cancer. Expression data were obtained for all 14 samples. Expression levels were normalised to the expression of the TBP housekeeping gene, as in the previous results chapter, to allow for comparison with reports in the literature that also typically use TBP for normalisation. The expression level of each APOBEC gene in the AGS samples was then also normalised to its mean expression level in the two control samples. Given the small sample size, it was reasoned that samples from different AGS genotypes should be grouped together to achieve greater statistical power to detect expression changes, as it was reasoned that these samples would be expected to display common phenotypes owing to shared underlying pathophysiology. The data from this analysis are shown below, in Figure 3.3.1.

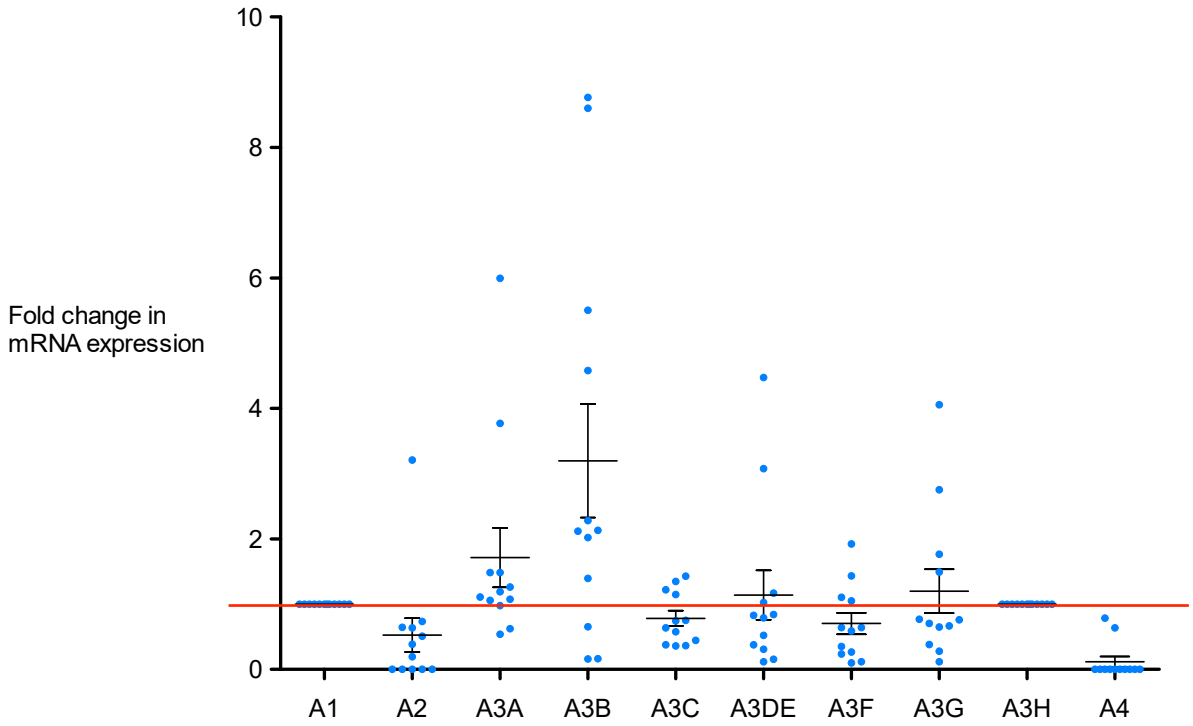


Figure 3.3.1 - Fold change in APOBEC expression in AGS samples relative to control samples, as measured by RNA-seq. Expression levels for each APOBEC gene are normalised to the mean of the control samples and to TBP expression. Error bars: Standard error of the mean. Red line: fold change value of 1.

The pattern of APOBEC expression in Figure 3.3.1 appears to be similar to that found in cancer. APOBEC3B expression appears elevated around 3-fold. APOBEC3A appears to upregulated to a lesser extent, showing an approximately 1.5-fold increase in expression. The other APOBEC genes are not comparably overexpressed. While the pattern observed is reminiscent of that observed in cancer, none of the APOBEC genes are individually upregulated at a level of statistical significance in these AGS

samples. For example, the t-test p-value for APOBEC3B, which is shows the highest fold change in expression, is 0.170.

3.3.3. Mutational profiles of Aicardi-Goutières syndrome RNA and association of APOBEC signature mutations with APOBEC expression

RNA-Seq data were downloaded and processed according to GATK best practices for mutation calling in RNA. This was performed in order to assess whether AGS samples might have evidence of APOBEC signature mutations that exceeds that of control samples. Raw data files were downloaded from the Sequence Read Archive. Quality control was performed on these raw data using FastQC, to evaluate read quality scores, and TrimGalore, to perform read trimming using a quality score threshold. Reads were mapped to the reference human genome (build 38). Additional data curation steps were performed, including base recalibration to correct for systematic errors in the base quality scores, and the filtering of known germline variants from the set of mutations detected. The output data are presented in Table 3.3.1 as the number of point mutations of each type across the 14 fibroblast samples, using the 96 type classification system. The data are also summarised in Figure 3.3.2, which plots the mean proportion of each point mutation type in control samples and AGS samples, respectively.

	C1	C2	T1	T2	RB1	RB2	RB3	RB4	RA1	RA2	RA3	RA4	S1	S2
A[C>A]A	4	4	3	73	2	104	4	4	4	52	3	2	4	127
A[C>A]C	0	1	2	31	1	46	4	2	3	20	5	1	4	53
A[C>A]G	1	1	2	7	1	9	1	0	0	5	1	0	0	8
A[C>A]T	0	0	3	16	3	25	3	1	3	18	2	0	3	23
A[C>G]A	5	1	6	14	4	23	6	4	1	8	8	5	4	26
A[C>G]C	3	2	1	12	0	11	4	1	1	10	1	0	4	15
A[C>G]G	1	0	2	1	1	10	2	0	1	3	1	1	1	3
A[C>G]T	2	1	2	11	3	17	2	1	2	7	2	0	5	20
A[C>T]A	1	0	18	55	9	76	9	4	20	29	7	4	9	58
A[C>T]C	1	0	15	35	15	62	6	5	10	31	10	4	13	62
A[C>T]G	1	0	0	0	0	5	3	0	3	1	3	0	0	1
A[C>T]T	3	2	8	49	7	75	7	5	9	24	10	4	8	50
A[T>A]A	5	0	3	63	6	105	2	1	7	43	3	1	2	76
A[T>A]C	3	3	2	31	0	36	2	2	1	21	2	0	1	40
A[T>A]G	1	1	4	26	4	23	5	0	2	21	5	0	6	34
A[T>A]T	1	1	0	19	0	26	1	0	1	15	3	1	2	25
A[T>C]A	6	2	163	95	18	131	104	80	61	73	169	63	136	153
A[T>C]C	3	1	22	54	16	86	20	14	23	42	35	12	30	90
A[T>C]G	3	1	137	92	36	123	87	71	92	59	137	66	100	148
A[T>C]T	1	0	306	193	55	223	203	170	248	122	336	115	221	215
A[T>G]A	4	1	3	12	1	12	1	0	1	9	5	0	2	10
A[T>G]C	0	1	7	11	1	11	3	1	4	5	5	2	0	13
A[T>G]G	4	4	10	20	2	27	3	3	4	7	4	3	5	25
A[T>G]T	4	1	1	16	2	16	2	1	0	10	1	1	4	15
C[C>A]A	1	1	3	62	1	95	3	2	4	53	7	3	4	92
C[C>A]C	3	1	3	44	3	56	4	1	3	29	3	2	2	59
C[C>A]G	1	1	1	11	1	17	0	2	1	5	0	0	2	14
C[C>A]T	3	2	1	25	1	32	2	2	4	17	1	1	3	28
C[C>G]A	4	1	2	11	2	15	4	1	4	9	4	2	5	15
C[C>G]C	6	2	0	10	1	11	0	0	2	7	1	2	4	17
C[C>G]G	3	1	0	8	1	15	0	2	1	3	3	1	1	11
C[C>G]T	6	1	1	7	2	16	4	3	2	6	4	3	6	17
C[C>T]A	7	6	13	78	15	89	7	6	16	49	11	4	9	63
C[C>T]C	20	13	12	67	12	113	7	3	14	50	7	2	9	102
C[C>T]G	0	1	3	7	3	11	3	1	2	6	4	2	2	8
C[C>T]T	9	6	16	37	13	95	9	3	20	38	13	3	9	69
C[T>A]A	19	22	1	37	2	51	2	0	2	33	2	1	1	58
C[T>A]C	24	19	4	50	5	50	2	2	5	21	3	0	1	74
C[T>A]G	1	1	5	31	6	41	2	1	9	20	4	3	4	51
C[T>A]T	35	24	3	14	2	28	2	3	2	10	4	0	2	38
C[T>C]A	13	4	1125	523	147	507	725	641	589	343	1091	431	752	569
C[T>C]C	14	8	174	148	51	230	126	71	140	128	174	57	114	251
C[T>C]G	6	3	894	418	138	522	593	517	494	273	931	350	578	499
C[T>C]T	10	3	552	322	129	359	434	347	342	181	627	230	445	413
C[T>G]A	24	15	4	13	0	5	2	0	1	2	4	1	2	11
C[T>G]C	65	49	4	13	5	23	1	1	2	9	2	0	4	29
C[T>G]G	5	3	3	21	4	25	3	1	5	13	4	3	8	25
C[T>G]T	34	31	5	13	0	16	2	0	0	6	2	2	2	16
G[C>A]A	7	2	4	19	4	28	1	0	2	14	3	0	3	26
G[C>A]C	0	0	2	22	3	23	2	3	2	14	2	2	4	35
G[C>A]G	4	2	2	10	0	24	2	2	1	11	1	1	1	19
G[C>A]T	0	0	2	22	4	28	1	3	1	10	0	0	2	30
G[C>G]A	1	0	2	18	2	21	5	0	0	8	1	1	6	19
G[C>G]C	4	1	3	21	0	42	3	1	6	16	4	1	10	33
G[C>G]G	6	4	1	5	1	16	1	1	3	2	3	1	4	14
G[C>G]T	1	1	4	16	1	28	7	3	3	11	2	1	4	22
G[C>T]A	0	0	9	52	11	87	2	3	14	37	5	3	7	69
G[C>T]C	2	2	10	55	15	73	8	5	12	30	9	6	8	75
G[C>T]G	7	0	5	17	2	17	5	3	4	7	3	2	2	16
G[C>T]T	2	3	5	58	12	70	6	1	16	43	8	1	8	89
G[T>A]A	2	2	3	15	0	19	1	1	2	9	1	0	3	15

G[T>A]C	3	1	1	11	3	25	0	1	4	8	1	0	1	25
G[T>A]G	1	0	3	51	5	91	3	0	1	45	4	0	2	93
G[T>A]T	4	2	0	25	2	21	0	0	1	6	2	0	3	21
G[T>C]A	116	85	428	213	76	274	339	279	268	138	452	186	358	279
G[T>C]C	32	15	35	49	13	53	18	17	20	29	33	13	19	76
G[T>C]G	120	54	411	221	73	276	298	238	250	131	460	156	267	287
G[T>C]T	262	168	210	140	44	184	133	127	185	92	224	82	153	175
G[T>G]A	980	548	0	7	1	17	3	1	1	5	2	0	6	12
G[T>G]C	133	80	2	12	1	25	2	2	1	9	3	2	2	11
G[T>G]G	755	493	3	19	3	31	2	3	2	7	3	2	5	37
G[T>G]T	544	357	2	17	0	16	2	1	2	6	2	0	1	10
T[C>A]A	403	263	0	22	2	29	1	0	3	13	1	0	1	30
T[C>A]C	30	11	0	20	0	14	2	2	0	16	1	0	2	36
T[C>A]G	333	230	1	2	0	4	1	1	1	1	0	1	1	6
T[C>A]T	174	103	0	17	3	19	0	1	2	12	0	1	1	21
T[C>G]A	297	185	4	23	3	25	4	0	5	11	2	1	3	35
T[C>G]C	30	20	5	16	5	14	5	2	6	12	6	2	5	28
T[C>G]G	229	164	1	3	0	6	0	0	2	4	0	0	1	12
T[C>G]T	207	131	5	26	2	37	3	3	1	16	8	0	5	36
T[C>T]A	1	0	12	58	10	86	7	5	16	38	15	4	10	89
T[C>T]C	1	1	19	50	17	86	16	7	15	40	18	5	20	64
T[C>T]G	3	1	2	14	4	12	1	3	4	11	5	1	2	13
T[C>T]T	1	2	19	98	21	128	9	6	31	58	13	9	20	114
T[T>A]A	3	1	4	18	2	25	6	1	4	16	3	0	1	44
T[T>A]C	1	1	2	17	1	32	1	1	0	16	1	0	1	35
T[T>A]G	5	4	0	17	1	42	0	1	6	11	3	2	3	23
T[T>A]T	1	2	3	40	8	39	7	1	3	18	1	0	11	43
T[T>C]A	1	1	326	203	49	253	222	205	266	146	344	115	249	268
T[T>C]C	2	1	42	96	21	114	36	27	50	53	57	16	34	112
T[T>C]G	6	1	268	158	63	198	194	161	179	93	281	102	175	233
T[T>C]T	1	1	192	172	74	233	157	116	269	101	207	79	159	230
T[T>G]A	3	1	3	29	2	34	5	2	3	15	4	1	3	32
T[T>G]C	6	3	4	20	5	20	3	1	0	20	2	1	3	20
T[T>G]G	6	3	5	15	2	26	1	3	3	15	3	4	4	31
T[T>G]T	5	3	2	61	1	65	1	1	2	42	1	3	7	85

Table 3.3.1 - Number of point mutations of each type (using the 96 type classification system) in the AGS RNA-seq dataset.

C: control, RA: RNSASEH2A, RB: RNASEH2B, S: SAMHD1, T: TREX1.

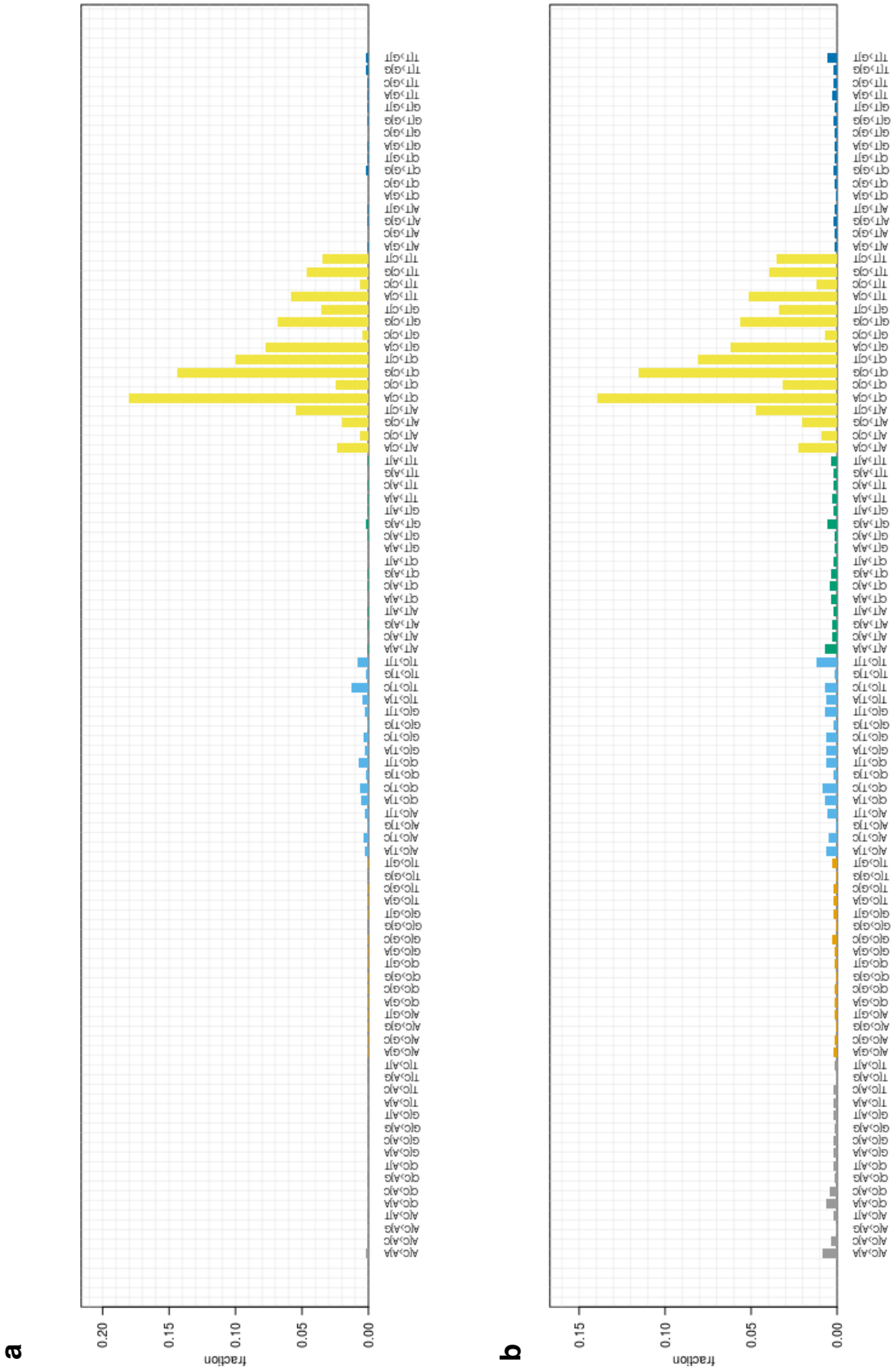


Figure 3.3.2 - The mean proportion of each of the 96 mutation types in the RNA-seq samples analysed. (a) control sample means. (b) AGS sample means.

These data indicate that the mutational profiles of both control and AGS samples are marked by a predominance of T>C mutations. This is in keeping with the physiological activity of ADAR deaminases that perform RNA editing of the human transcriptome (Lerner, Papavasiliou, and Pecori 2019). In addition to this, AGS samples appear to have slightly more point mutations of all other types. Both control and AGS samples were found to have mutations at TCW residues that could be consistent with the activity of APOBEC3A or APOBEC3B. The number of these TCW mutations in control and AGS samples is shown in Figure 3.3.3. Here, TCW mutations were chosen as a measure of APOBEC signature mutations as it was noted that cancer-associated APOBEC activity could potentially cause C>T, C>G or C>A mutations (Morganella et al. 2016) and APOBEC expression was associated with mutations at C residues in general as well as at TCW residues (Burns et al. 2013), suggesting that cancer-associated APOBEC activity might be more promiscuous than suggested by the NMF-generated SBS2 and SBS13 (see sections 1.5 and 1.6).

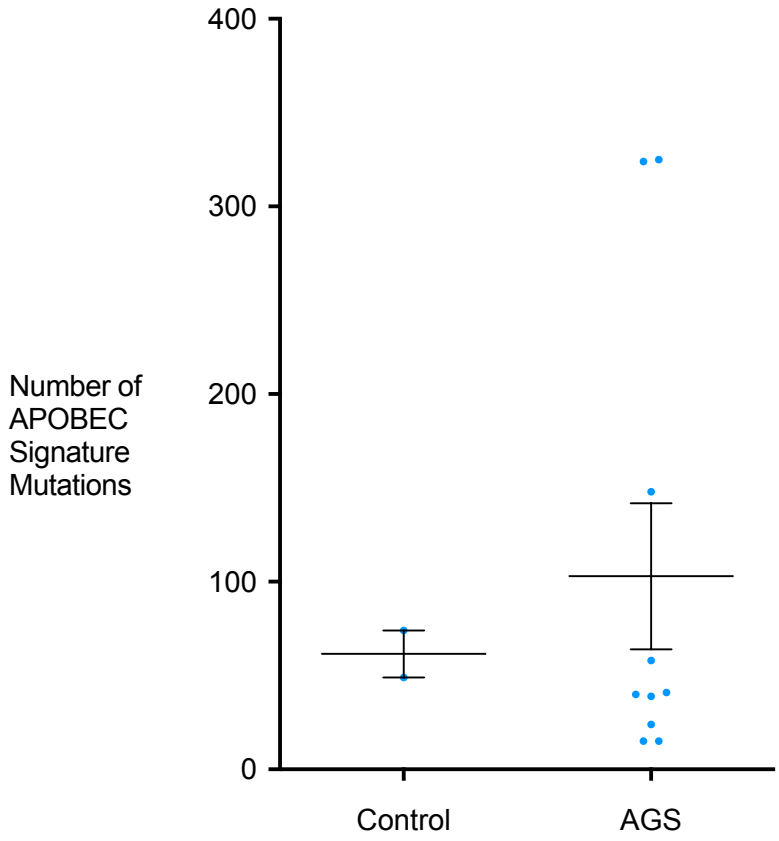


Figure 3.3.3 - The number of APOBEC signature TCW mutations detected in RNA-seq of AGS samples and control samples. Error bars: Standard error of the mean.

The data in Figure 3.3.3 showed an increase in the number of mutations at TCW residues that was not statistically significant (control mean: 61.5, AGS mean: 102.9 , t-test p-value: 0.328).

Given that mutations of all types were slightly elevated in AGS samples relative to controls, the percentage of mutations occurring at TCW residues

in control and AGS samples, rather than simply their number, was also compared (Figure 3.3.4).

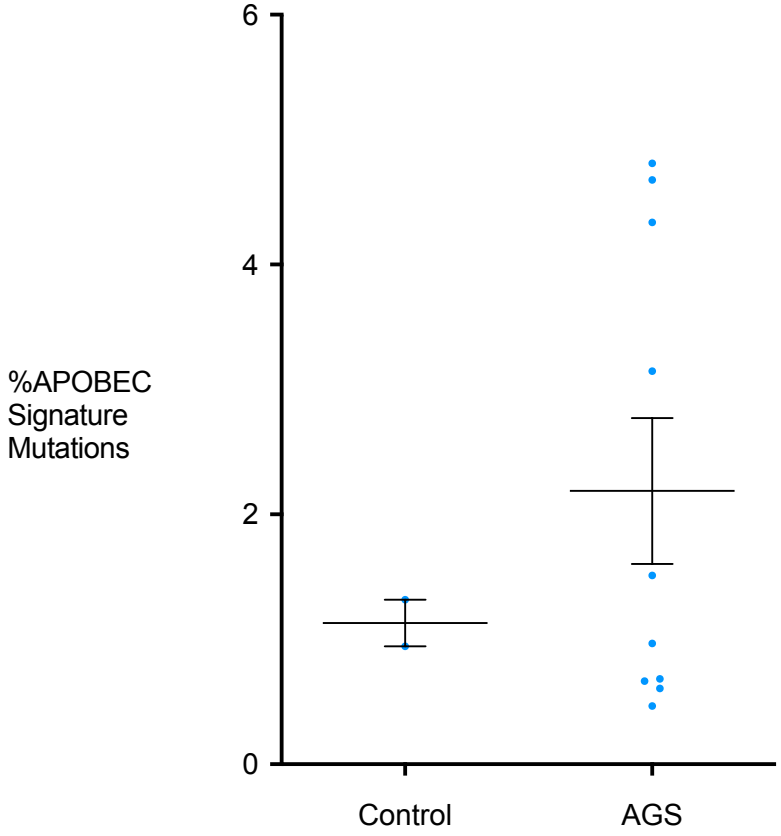


Figure 3.3.4 - The percentage of APOBEC signature TCW mutations detected in RNA-seq of AGS samples and control samples. Error bars: Standard error of the mean.

The data in Figure 3.3.4 showed an increase in the percentage of mutations at TCW residues that was again not statistically significant (control mean: 1.1%, AGS mean: 2.1% , t-test p-value: 0.227).

It was noted that there was some variability in data points for APOBEC3A and APOBEC3B expression, the number of TCW mutations and percentage of TCW mutations. Their distribution did not appear to depend on the underlying genotypes of the samples (Table 3.3.2).

	AGS Sample											
	TREX1		RNASEH2B				RNASEH2A				SAMHD1	
	1	2	1	2	3	4	1	2	3	4	1	2
APOBEC3A expression	1.1	3.8	1.5	1.3	1.2	1.1	6.0	0.6	1.1	1.5	0.5	1.0
APOBEC3B expression	0.2	8.8	2.1	5.5	0.7	2.1	8.6	4.6	0.2	2.3	2.0	1.4
TCW number	40	244	41	324	24	15	58	148	39	15	40	325
TCW percentage	0.7	4.7	3.1	4.8	0.6	0.5	1.5	4.3	0.7	0.7	1.0	4.7

Table 3.3.2 - The expression levels of APOBEC3A and APOBEC3B and the number and percentage of TCW mutations in each AGS sample. Expression data are normalised to TBP and the mean of the control samples.

Next, Spearman rank correlations were performed using data for APOBEC gene expression and either the number or percentage of TCW mutations in the AGS samples (Table 3.3.3). A Spearman correlation was chosen over a Pearson correlation in an attempt to account for non-linear differences in the distributions of these sets of data.

	Number of TCW mutations		Percentage of TCW mutations	
APOBEC gene	Spearman's ρ	p-value	Spearman's ρ	p-value
APOBEC1	NA	NA	NA	NA
APOBEC2	-0.13	0.677	-0.04	0.911
APOBEC3A	0.01	0.965	0.13	0.681
APOBEC3B	0.38	0.220	0.54	0.071
APOBEC3C	-0.62	0.030	-0.65	0.022
APOBEC3DE	-0.49	0.102	-0.58	0.479
APOBEC3F	-0.61	0.036	-0.64	0.240
APOBEC3G	-0.42	0.173	-0.47	1.245
APOBEC3H	NA	NA	NA	NA
APOBEC4	-0.30	0.335	-0.40	0.194

Table 3.3.3 - Spearman correlations using expression levels of APOBEC genes and either the number or percentage of APOBEC signature TCW mutations in AGS samples.

The results in Table 3.3.3 indicate that expression of APOBEC3A and APOBEC3B are positively correlated with the number of TCW mutations in AGS samples (ρ = 0.01 and 0.38, respectively) and the percentage of TCW mutations in AGS samples (ρ = 0.13 and 0.54, respectively) with APOBEC3B showing a stronger correlation than APOBEC3A and no other APOBEC genes showing positive correlations. This mimics the pattern observed in cancer. However, as indicated by Table 3.3.3, none of the positive correlations calculated are statistically significant. The reasons underlying negative correlations of approximately ρ = -0.6 for APOBEC3C

and APOBEC3F are not clear. Again, these correlations do not meet statistical significance if a Bonferroni correction is applied.

3.3.4. Gene ontology analysis of RNA-seq data in Aicardi-Goutières syndrome and APOBEC3B-deficient breast cancer

Gene ontology analysis was performed using the PANTHER tool on the list of ~100 genes reported by Lim et al to be upregulated at statistical significance in AGS patients relative to controls. Gene ontology terms which are overrepresented at statistical significance ($p < 0.05$) are shown in Table 3.3.4.

AGS upregulated
positive regulation of neutrophil chemotaxis
chemokine-mediated signaling pathway
positive regulation of leukocyte chemotaxis
mitotic prometaphase
regulation of leukocyte chemotaxis
anaphase
mitotic anaphase
regulation of leukocyte migration
M phase
mitotic M phase
mitotic cell cycle phase
cell cycle phase
biological phase
positive regulation of response to external stimulus
cytokine-mediated signaling pathway
mitotic nuclear division

Table 3.3.4 - Overrepresented gene ontology terms for genes reported to be most significantly upregulated in AGS, as reported by Lim et al.

The results shown in Table 3.3.4 indicate that, although AGS is described to be an interferonopathy, interferon-related genes are not overrepresented in the list of overexpressed AGS genes. However, a number of immune-related and cell cycle-related gene ontologies are overrepresented.

It was not clear why cell cycle genes might be upregulated as part of the pathophysiology of AGS, or why more general immune gene ontologies might be overrepresented rather than interferon-related gene ontologies. However, it was noted that APOBEC3B expression is associated with the expression of cell cycle-related genes in cancer (see section 3.2.2). It was also noted that the APOBEC3B deletion polymorphism was previously associated with the upregulation of immune gene expression in two breast cancer datasets, from TCGA and METABRIC (Molecular Taxonomy of Breast Cancer International Consortium; Curtis et al. 2012; Cescon, Haibe-Kains, and Mak 2015).

To further characterise the possible similarity between gene expression in AGS and APOBEC3B-deficient cancers, PANTHER gene ontology analysis was conducted on upregulated and downregulated genes using the Lim et al and TCGA datasets. TCGA patient IDs from Nik-Zainal et al. 2014 were used to identify the APOBEC3B-deficient cancers in the TCGA dataset. Individuals that were either homozygous or heterozygous for the APOBEC3B deletion allele were grouped in an attempt to achieve greater statistical power to detect expression changes, given observed similarities in their resultant phenotypes. The top and bottom 100 differentially-expressed genes in the AGS and breast cancer datasets were analysed using the PANTHER gene ontology tool. Gene ontology terms which are overrepresented at statistical significance ($p < 0.05$) are shown in Table 3.3.5.

a

AGS upregulated	APOBECB-deficient cancer upregulated
positive regulation of neutrophil chemotaxis	immune response
chemokine-mediated signaling pathway	biological process
positive regulation of leukocyte chemotaxis	tissue homeostasis
mitotic prometaphase	regulated exocytosis
regulation of leukocyte chemotaxis	retina homeostasis
anaphase	defense response
mitotic anaphase	multi-organism process
regulation of leukocyte migration	response to external stimulus
M phase	secretion
mitotic M phase	single-multicellular organism process
mitotic cell cycle phase	humoral immune response
cell cycle phase	defense response to other organism
biological phase	response to biotic stimulus
positive regulation of response to external stimulus	response to external biotic stimulus
cytokine-mediated signaling pathway	response to bacterium
mitotic nuclear division	response to other organism
	single-organism process
	defense response to bacterium
	antimicrobial humoral response

b

AGS downregulated	APOBECB-deficient cancer downregulated
extracellular matrix organization	multicellular organismal process
extracellular structure organization	system development
animal organ morphogenesis	cell-cell adhesion
animal organ development	extracellular structure organization
anatomical structure development	extracellular matrix organization
developmental process	single-multicellular organism process
system development	cell adhesion
multicellular organism development	biological adhesion
multicellular organismal process	
anatomical structure morphogenesis	
cell adhesion	
biological adhesion	

Table 3.3.5 - Overrepresented gene ontology terms for genes in AGS and APOBEC3B-deficient breast cancers that are either upregulated (a) or downregulated (b).

The results in Table 3.3.5 indicate that AGS and APOBEC3B-deficient breast cancers appear to share similar gene expression changes. In terms of upregulated genes, both APOBEC3B-deficient breast cancers AGS show an overrepresentation of immune response genes. Interferon-related gene ontologies are not overrepresented. The immune gene ontology terms that are overrepresented in the two datasets are qualitatively similar but not identical. In terms of downregulated genes, both APOBEC3B-deficient breast cancers and AGS samples show an overrepresentation of developmental, cell adhesion and extracellular structure genes. Here, there are many identical gene ontology terms between the two groups.

3.3.5. Discussion

These analyses were conducted in order to investigate whether cells from patients with Aicardi-Goutières syndrome might show evidence of cancer-associated APOBEC activity. On one hand, the overall trends shown in each analysis performed are consistent with the notion that cancer-associated APOBEC activity might indeed be present in Aicardi-Goutières syndrome. But on the other hand, multiple individual results within these trends were found not to be statistically significant. The findings suggested by these data in particular are therefore equivocal. To summarise:

- The pattern of APOBEC expression in AGS appears similar to that observed in cancer, where APOBEC3B and to a lesser extent APOBEC3A are overexpressed. However, the apparent overexpression of each of these genes is not statistically significant.
- APOBEC signature mutations appear to be detectable in RNA from AGS. The mean number and mean percentage of APOBEC signature mutations appears to be around 2-fold higher in AGS samples relative to controls, but this apparent increase is not statistically significant.
- When the number of APOBEC signature mutations is correlated to the expression levels of APOBEC genes in AGS, the pattern appears similar to that observed in cancer, where APOBEC3B and to a lesser extent APOBEC3A show the highest positive correlations. However, these correlations did not individually reach statistical significance.
- Although AGS is described to be an interferonopathy, interferon-related gene ontologies are not overrepresented in the list of significantly

upregulated AGS genes. Instead, more general immune gene ontologies are overrepresented, alongside cell cycle-related gene ontologies.

- Germline APOBEC3B deletions in breast cancer appear to show similar gene expression patterns to those observed in AGS. In both cases, immune genes are upregulated, while developmental, cell adhesion and extracellular structure genes are downregulated.

One explanation for the fact that some results did not reach statistical significance is that the analyses in section 3.3.2 and 3.3.3 were limited by the small sample size of the Lim et al. study. This possibility appears to be reaffirmed by considering the probability of detecting APOBEC3B upregulation in this dataset, for example. In this thesis and in work by others, APOBEC3B expression appears to increase by a magnitude of 2-3 fold in cancer. A power calculation suggests that the analysis conducted may have been underpowered in attempting to detect a fold change of this magnitude. Assuming equal numbers of cases and controls, a fold change of 2.5 with a standard deviation of 1, a false positive rate of 0.05 and a power threshold of 90%, a study of 9 cases and 9 controls would be required to detect APOBEC3B upregulation at statistical significance. This suggests that future work would likely benefit from a greater sample size. To our knowledge, an RNA-seq study of AGS patients and matched controls of a larger sample size had not been conducted at the time of preparing these analyses.

Another limitation of the analyses conducted is the use of RNA-seq to identify APOBEC signature mutations rather then sequencing of the

genome. Firstly, cancer-associated APOBEC signature mutations are found in DNA rather than RNA. Mutations in DNA were understood to be detectable in expressed RNA, and APOBEC3A and APOBEC3B had been characterised predominantly as ssDNA mutators. However, it is not possible to definitively determine whether the APOBEC signature mutations detected in this analysis are due to DNA editing or RNA editing using these data alone. Another limitation of using RNA-seq data for this analysis lies in the limited amount of the genome covered when using this technique. The coverage of the genome is broadly similar to that of exome sequencing, at around 1%. In keeping with this, there are on the order of tens of APOBEC signature mutations found in the Lim et al. RNA-seq dataset (Figure 3.3.3) and the TCGA exome sequencing dataset (Figure 3.2.2). In contrast, whole genome sequencing might be expected to identify hundreds or thousands of APOBEC signature mutations, which would provide additional statistical power for some of the analyses conducted. In addition, an examination of whether or not the mutations called in the RNA sequencing data might reflect pathogenic changes in potentially relevant genes, such as TP53, was not conducted. This is because it is was not clear whether RNA-seq could accurately identify specific DNA-level variants in specific genes, despite its use in approximating this mutagenesis at an exome-wide scale. Examining the functional consequences of specific variants might be particularly beneficial in future DNA sequencing work

Another factor that can be considered in this analysis is the heterogeneity of AGS mutations in the samples studied. Given the rarity of AGS as a disease, studies using samples from multiple patients typically consist of

multiple genotypes, given that they are thought to share phenotypic pathophysiology (such as in Rice et al. 2018). Although all such patients display AGS phenotypes, the precise mechanistic changes that occur with each type of AGS mutation are not identical. It may therefore be possible possible that different AGS genes might influence APOBEC regulation in different ways in these analyses. However, Table 3.3.2 suggested that this might not be the case for the key data studied in this chapter. To speculate, there may be variability due to an inherent heterogeneity in APOBEC activity in these cells – in a manner that is analogous to previous descriptions of the episodic and heterogeneous nature of LINE-1 and APOBEC activity in cancer cells (Petljak et al. 2019; Rodriguez-Martin et al. 2020).

Despite the limitations of the analysis described above, the Lim et al. dataset nonetheless appears to represent a rare and valuable resource. This dataset allows, in principle, for the study of the dysregulation of the metabolism of endogenous nucleic acids that are likely derived from LINE-1 elements *in vivo*. The value of data such as these is emphasised by the fact that both LINE-1 elements and their regulation vary substantially between species, potentially limiting the utility of studying similar genetic defects in animal models. Although many of the results in this chapter equivocal, the trends observed throughout the analyses are repeatedly consistent with the presence of cancer-associated APOBEC activity, a feature of the data which in itself might be unlikely to occur due to chance. It may therefore be possible that further study indicates more rigorously that LINE-1 activity in AGS leads to ABOBEC activity that is analogous to that observed in cancer.

It is also possible that further study indicates that the endogenous nucleic acids other than LINE-1 might contribute to any such observations (discussed in section 1.7.4).

The gene ontology results in this chapter were statistically significant. They appear to provide evidence that gene expression in AGS shows parallels with gene expression that is associated with APOBEC3B in cancer. For example, cell cycle-related genes appear to be upregulated in AGS. Cell cycle-related genes also appear to be associated with APOBEC3B expression in cancer (see section 3.2.2). In addition, the AGS data indicate that genes with immune gene ontologies are upregulated and genes with developmental, cell adhesion and extracellular structure gene ontologies are downregulated. This also appears to be the case in breast cancers with APOBEC3B deficiency. Of note, although AGS is reported to be an interferonopathy, interferon-related genes are not overrepresented in the set significantly upregulated genes. As with the PANTHER gene ontology results in the previous results chapter, the ontology results in this chapter might be complemented and validated by the use of techniques such as Gene Set Enrichment Analysis, which identifies statistically significant differences in gene expression for all genes represented by specific gene ontologies (Subramanian et al. 2005).

These parallels implicate the mechanistic changes that occur in AGS with those that are associated with APOBEC3B activity cancer. The mechanistic basis of the expression changes detected are not immediately clear. For example, on review of the literature, it is not clear why cell cycle-related

genes might be upregulated in AGS. Further experimental work is required to determine whether the parallels observed are due to shared pathophysiology, or instead due to confounding factors or chance.

4. Discussion

Introduction

The stated aim of this thesis is to investigate the regulation of APOBEC mutagenesis in cancer. The approach taken for the work conducted was based on the understanding that aberrant signalling appears likely to drive the aberrant expression of APOBEC3A and APOBEC3B in cancer, leading to APOBEC signature mutations. A focus of the work conducted was the hypothesis that APOBEC activity might be driven by LINE-1 upregulation. Inactivation of either p53 or AGS genes were identified as possible ways to model the upregulation of LINE-1 activity that occurs in cancer. Here, the key findings in each chapter of the thesis are listed in turn and critically analysed in light of the hypothesis at hand to inform suggestions for future work.

1. Experiments investigating APOBEC activity in cultured cancer cells

In terms of the first results chapter, the experiments conducted suggest that p53 inactivation leads to the upregulation of LINE-1 and APOBEC3B, and that the opposite might occur when p53 activity is promoted by Nutlin-3a. Reverse transcriptase inhibitors of LINE-1 activity appear to modulate APOBEC3B expression and associated enzymatic activity. This appears to occur when cells are p53-deficient, but not when p53 is intact. In addition, a high-throughput deaminase assay is established.

These experiments do not determine whether or not LINE-1 activity drives APOBEC activity, although the findings are broadly consistent with this hypothesis. An approach to testing this hypothesis in future could entail the use of gene editing to inactivate the 150 or so reportedly active LINE-1 elements in the genome (Penzkofer et al. 2016). This would provide a genetic approach to LINE-1 inhibition to complement the pharmacological approach pursued in these experiments. It would be of use to examine whether APOBEC activity is reduced or abolished if LINE-1 is genetically inactivated, and to examine what global expression changes occur by RNA-seq, including expression changes that mirror those found in AGS.

In this section, p53 inactivation is used as a model for LINE-1 activation. Since these experiments were designed and conducted, it has been reported that p53 directly regulates the APOBEC3B promoter (Periyasamy et al. 2017). It is therefore unclear whether this model is as valid to the extent that was assumed. The TP53 gene might regulate APOBEC activity through direct epigenetic control or indirectly through LINE-1 activation as posited - possibly through pathways that are of relevance to AGS pathophysiology. LINE-1 inactivation experiments, as described above, performed in the context of p53 inactivation would enable the detection of the extent to which p53's impact on APOBEC activity is LINE-1 mediated and an assessment of global gene expression changes.

Although the conflicting results concerning AZT versus d4T and 3TC were rationalised in the context of a possible parallel with AGS, the experiments

conducted would be strengthened by assessing possible changes in APOBEC activity using a wider range of RTIs. Recently reported structures of LINE-1 ORF2 may shed light on why there may be differential effects of different RTIs, or inform the development of more specific inhibitors that could be used in their place to test the hypothesis at hand (Baldwin et al. 2023; Thawani et al. 2023).

2. Bioinformatic analyses investigating APOBEC regulation in large genomic datasets

In terms of the second results chapter, the analyses conducted indicate that APOBEC3A expression is associated with interferon signalling in cancer while APOBEC3B expression is associated with cell cycle signalling in cancer. A deletion of a consensus interferon response factor binding site in the human APOBEC3B promoter is identified. Analyses of regulatory data suggest that APOBEC3B might be transcriptionally insulated from syntenic APOBEC3 genes by CTCF. In addition, p53 deficiency in cancer appears to be associated with the upregulation of APOBEC3A and APOBEC3B.

The results of these exploratory analyses are not inconsistent with the hypothesis that LINE-1 activity promotes APOBEC activity in cancer. However, they again do provide definitive answers. Further experimental work is required to develop an understanding of the observations gained from these data, and then test their relevance in the context of LINE-1. For example, the significance of the deletion in the APOBEC3B promoter could be evaluated in future work by using gene editing techniques to delete this

region in chimpanzee or bonobo cells and examine how APOBEC3B expression varies when these cells are exposed to interferon, or indeed exposed to LINE-1 upregulation. Similar functional experiments could be performed to examine the significance of CTCF from putatively insulating APOBEC3B from syntenic interferon responsiveness by knocking out these CTCF binding sites in human cells. While APOBEC3s are known to be interferon-responsive, work published since the completion of this work indicates that the association of cell cycle signalling to APOBEC3B may reflect a causal relationship, with APOBEC3B expression varying as a result of cell cycle phase (Hirabayashi et al. 2021)

In addition, while p53 deficiency is associated with APOBEC3A and APOBEC3B upregulation, it is not clear whether this is a potentially causal relationship. One method of addressing this in future would be to attempt to assess the temporality of the relationship between p53 inactivation and APOBEC activity. Since the clonality of mutations in a genetically heterogenous tumour can be used to infer timing, it may be possible to examine computationally whether APOBEC signature mutations (and indeed LINE-1 insertions) occur before or after TP53 mutation in a range of cancers.

3. Bioinformatic analyses investigating APOBEC activity in Aicardi-Goutières syndrome

In terms of the third results chapter, the analyses conduced suggest that changes to the transcriptome in Aicardi-Goutières syndrome might mirror those associated with APOBEC activity in cancer. AGS cells show an upregulation of genes with cell cycle gene ontologies, mimicking the association of APOBEC3B expression in cancer. AGS cells also show the upregulation of immune genes and the downregulation of developmental, cell adhesion and extracellular structure genes, which mimics the changes that occur when APOBEC3B deletions are found in breast cancer.

These results are also broadly consistent with the hypothesised role played by LINE-1 and could also be strengthened by future work. Obtaining a larger number of AGS samples would likely be required for sufficient statistical power to detect the changes of interest. Performing genome sequencing in future work would allow for the detection of APOBEC signature mutations with greater accuracy than can be achieved with mutation calling in RNA-seq data, including the detection of potentially relevant somatic mutations of specific genes such as TP53. In addition, using techniques specific for the measurement of LINE-1 expression in RNA-seq data would help confirm the presence of increased LINE-1 activity. LINE-1 knockout through gene editing or treatment of RTIs prior to sequencing would have utility in determining the contribution of LINE-1 activity to any APOBEC-related phenotypes observed.

It would also be of use to perform experimental work to determine the significance of the APOBEC3B deletion polymorphism in breast cancer to LINE-1 and AGS biology. The deletion polymorphism could be generated in

cancer cells using gene editing. Based on the results of the analysis in this chapter, one might expect that this would lead to elevated LINE-1 activity that in turn leads to AGS-related signalling. It would also be expected that this phenotype could be rescued by inhibiting LINE-1 genetically or pharmacologically.

It is of note that the RTI treatments that alleviate AGS models also appear to reduce measures APOBEC activity in HCT116 cells, while the opposite is true for those that do not alleviate AGS models, and that both such effects occur when p53 is deficient and LINE-1 expression is elevated. If reverse transcriptase inhibitors might influence cancer cells *in vivo* in similar ways, then it might be expected that cancer patients on RTI therapy as part of a long term antiretroviral treatment might have cancer genomes with atypical numbers of APOBEC signature mutations. Such genome sequences are currently being generated and analysed as part of the HIV+ Tumor Molecular Characterization Project performed by the United States National Cancer Institute (NCI 2024).

Conclusion

The hypothesis that LINE-1 activity promotes APOBEC activity in cancer appears to remain of importance to the field. Both LINE-1 activity and APOBEC activity are highly prevalent and likely underestimated sources of genomic instability in cancer that are likely to mediate cancer evolution. Indeed, recent reports indicate that APOBEC activity plays a causal role in the evolutionary trajectory of cancer cells, with experimental work indicating

that APOBEC activity mediates the acquisition of resistance in cells subjected to targeted therapies (Isozaki et al. 2023). Although multiple mechanisms have now been described in the literature as leading to cancer-associated APOBEC activity *in vitro*, it not yet clear which of these operates *in vivo. I*n the event there are multiple such mechanisms, it is not clear which is most common and which might be therapeutically relevant. LINE-1 activity appears likely to play a role in driving APOBEC activity at least in some contexts. It is commonly found in range of cancer types, is therapeutically actionable, and its inhibition would be expected to reduce genomic instability mediated by both LINE-1 activity and APOBEC activity.

The findings described in this thesis provide evidence for possible regulators of APOBEC mutagenesis in cancer, including evidence that broadly supports the hypothesis that it may be driven by LINE-1 activity. The experiments conducted identify a class of drugs that might enable the pharmacological modulation of cancer-associated APOBEC activity through the modulation of LINE-1 activity. APOBEC mutagenesis is thought to mediate cancer initiation, progression, intratumour heterogeneity and responses to therapy, including immunotherapy. It is therefore hoped that the work conducted might contribute to the ability to understand and control the natural history of cancer across multiple cancer types.

5. Materials and Methods

5.1. Cell culture

HCT116 $p53^{wt/wt}$ and $p53^{-/-}$ cells (Dr. Bert Vogelstein, Howard Hughes Medical Center, USA) were grown in McCoy's 5A (1x) + GlutaMAX™-I Modified Medium (Gibco) supplemented with 10% Foetal Bovine Serum (Gibco) and passaged by trypsinisation using trypsin-EDTA solution (Gibco). $p53^{-/-}$ knockout cells were generated by Vogelstein and colleagues by replacing the first codon of TP53 with its second intron (Bunz et al. 1998).

Cells were incubated in humid conditions at 37 °C and 5% CO_2 in Nunclon™ Delta treated flasks (Thermo Fisher). Cultured cells were treated with up to 100 μM Nutlin-3a (Sigma Aldrich), 100 μM AZT (Sigma Aldrich), a combination of 1 μM d4T (Sigma Aldrich) and 10 μM 3TC (Sigma Aldrich) or given mock treatment. Inhibitors stocks were stored at -20˚C in 10 μl aliquots. Cells were assayed 2-3 days after a single dose of inhibitor treatment.

5.2. Western blot

HCT116 $p53^{wt/wt}$ and $p53^{-/-}$ cells were harvested over three consecutive passages. Cells were lysed in lysis buffer (20mM Tris (pH8), 150mM NaCl,

0.5% NP-40, 1mM EDTA) plus protease inhibitor cocktail. Protein content was measured by Bradford and 15ug of total cell lysate was loaded for each sample on a 4-12% Bis-Tris gel and run in MOPS running buffer. The protein was transferred to PVDF, 100v for 1hr, in transfer buffer containing 7.5% methanol.

The membrane was cut in half between 50kDa and 75kDa and blocked in TBST + 5%milk for 1hr. The upper section was probed with anti-HSP90 (Cell Signalling Technology) diluted 1:2000 in TBST + 5% milk and the lower section was probed with anti-p53 (Santa Cruz) diluted 1:500 in TBST + 5% milk. The blot was washed and incubated with anti-mouse or anti-rabbit HRP-conjugated secondary antibodies then developed with ECL reagent.

5.3. Sulforhodamine B biomass assay

Cells were seeded in 96-well plates (Thermo Fisher) at a seeding density of 2000 cells per well and incubated with or without varying concentrations of AZT (Sigma Aldrich) or Nutlin-3a (Sigma Aldrich) for 72 hours. Cells were fixed with 1% trichloroacetic acid for 30 minutes. Then, plates were washed with deionised water before staining with 0.057% Sulforhodamine B (Sigma Aldrich) solution for 30 minutes. After washing with 1% acetic acid, plates were allowed to air dry overnight.

The Sulforhodamine B stain was then solubilised by 10 mM Tris for 10 mins while agitating. Fluorescence intensity was measured at excitation 540 nm and emission 590 nm on a PHERAstar Plus plate reader (BMG Labtech).

5.4. qRT-PCR

Experiments were performed in triplicates. RNA extraction was performed using the RNeasy Mini Kit (Qiagen), including Qiazol and DNase (Qiagen) treatment, according to the manufacturer's instructions. RNA was stored at -70 °C once extracted. cDNA synthesis was performed using the SuperScript III reverse transcriptase (Invitrogen) according to the manufacturer's instructions. Then, two PCR reactions for each triplicate were performed using the SYBR® Green PCR Kit (Qiagen). PCR reactions were performed in 96-well optical plates (Invitrogen) using a StepOne Plus RT-PCR machine (Applied Biosystems). Expression levels were quantified using the $2^{-\Delta\Delta CT}$ method, normalised to ACTB expression. Data were analysed using GraphPad Prism software. The following forward (F) and reverse (R) primers were used:

Target	Source		Sequence	Concentration
LINE-1 ORF1	Marchetto et al.	**F**	ATGAGCAAAGCCTCCAAGAA	500 nM
		R	TTCTCCCCATCACTTTCAGG	
LINE-1 ORF2		**F**	TGGAGGCATCACACTACCTG	500 nM
		R	ATGCGGCATTATTTCTGAGG	
APOBEC3B	Burns et al.	**F**	GACCCTTTGGTCCTTCGAC	500 nM
		R	GCACAGCCCCAGGAGAAG	
ACTB	Qiagen	**F**	N/A: proprietary information. Primers are designed to amplify exons 3 and 4.	N/A: proprietary information.
		R		

5.5. Cytosine deaminase assay

The protocol used was adapted from the Burns et al. (2013) and Vieira et al. (2014) studies. Cultured cells were harvested by trypsinisation and

washed in PBS. Cells lysis was then performed using 400 μl lysis buffer (25 mM HEPES pH 7.4, 250 mM NaCl, 0.5% Triton X-100 (Thermo Fisher), 1 mM EDTA, 1 mM MgCl2, 1 mM $ZnCl_2$, 10% glycerol, 2x Pierce™ EDTA-free protease inhibitor cocktail (Thermo Fisher)) using a 25 gauge needle (BD Microlance) and syringe. Lysates were sonicated at 30 Hz for 10 seconds and left to rotate for 30 minutes at 4 °C. Lysates were then centrifuged at 12000x g for 10 minutes at 4 °C. The supernatant was extracted and protein concentration determined by absorbance at 280 nm using a NanoDrop 2000 spectrophotometer (Thermo Fisher).

20 μl reactions were performed in black low volume 384-well plates (Corning) using 16 μl protein extract and 4 μl reaction mixture (4 pmol ssDNA probe, 0.025 U Uracil DNA Glycosylase (NEB), 5x UDG buffer (NEB), 1.75 U RNase H (Sigma Aldrich)). Where dilutions of lysates were used, these were diluted in lysis buffer. Plates were sealed and incubated in dark conditions at 37 °C for 2 hours. Reactions were supplemented with 20 μl 0.2 M NaOH and incubated for a further 30 minutes at 37 °C. Plates were read at excitation 485 nm and emission 520 nm on a PHERAstar Plus plate reader (BMG Labtech). ssDNA probes (Sigma Aldrich) used had the following sequences:

5’-6FAM-ATTATTATTAT<u>NNN</u>AATGGATTTATTTATTTATTTATTTATTT-TAMRA-3’

5’-6FAM-ATTATTATTAT<u>NNN</u>AATGGATTTATTTATTTATTTATTTATTT-3’

5’-6FAM-ATAANNNAATAGATAAT-TAMRA-3’

(NNN: TTA, ACA, TCA or TUA).

Where a gel was used to separate fragments, the gel used was a 15-well 15% Tris-Borate-EDTA (TBE)-Urea gel (Thermo Fisher). Gels were imaged using a Gel Doc XR+ machine (Bio-Rad). Imaged bands were quantified using ImageJ software (NIH - https://imagej.nih.gov/ij/).

5.6. Bioinformatic analyses of large datasets

Mutect2-called somatic mutation data and TPM-normalised RNA-seq data from 2840 patients were sourced from The Cancer Genome Atlas (TCGA) data portal (tcga-data.nci.nih.gov). Microsoft Excel and R (r-project.org) functions were used to link datasets, while R was used to compute correlations and complete arithmetic.

The bases immediately 5' and 3' of point mutations called in the TCGA dataset were extracted using the UCSC human genome builds 36 and 37 (genome.ucsc.edu) in order to identify APOBEC signature TCA point mutations. Then, the number of APOBEC signature mutations for each patient was linked to their RNA-seq profile. The gene expression data for all genes was normalised to that of the TATA binding protein (TBP) housekeeping gene. Then, APOBEC3A expression, APOBEC3B expression and the number of APOBEC signature TCA mutations were correlated to

the expression to all other genes using a Spearman's rank correlation. The standard error of the coefficient (σ_{rs}) is given by the formula $0.6325/\sqrt{(n-1)}$, where n is the number of patients. The top 20 most highly correlated genes were analysed for gene ontology enrichment using the Protein Analysis Through Evolutionary Relationships (PANTHER) tool (http://geneontology.org/ - Bonferroni-corrected $p < 0.05$ in all cases).

Concerning the p53 fold change analysis, patients with TP53 mutations were first identified using the somatic mutation data sourced from the TCGA data portal. The corresponding APOBEC expression data for the two groups were then identified in the linked RNA-seq data, and fold change calculated directly between these two groups. Calculations were performed in Microsoft Excel, GraphPad Prism and R.

To compare genomes of human and non-human primates, UCSC genome builds hg38 (human), panTro4 (chimpanzee) and panPan1 (bonobo) were viewed on the UCSC genome browser to identity sequence differences in the APOBEC3B promoter. ENCODE H3K27 acetylation data from all 7 cell lines available (GM12878, H1-hESC, HSMM, HUVEC, K562, NHEK, NHLF) were used to verify the regulatory role of putative promoter regions in human APOBEC3 genes, alongside ReMap peaks. Peak data were filtered for quality according to cross-correlation and the FRiP (fraction of reads in peaks) metrics developed by the ENCODE Consortium (Hammal et al. 2022).

ChIP Atlas Peak Browser (https://chip-atlas.org/) CTCF results (n=490) were filtered for high significance threshold (>500) and selected for data from all cell types available (n=43086). These data were downloaded as BED files. Indexing and viewing of these BED files were performed on the Broad Institute Integrative Genomics Viewer (IGV) (https://software.broadinstitute.org/software/igv/).

5.7. Bioinformatic analyses of Aicardi-Goutières syndrome data

These analyses made use of the RNA-Seq data from Lim et al. study. These authors performed RNA sequencing on 14 fibroblast samples from four patients with LINE-1 associated AGS genotypes and one age-matched health control. The four patients had pathogenic mutations in TREX1, RNASEH2A, RNASEH2B and SAMHD1, respectively. Four samples were derived from the patient with RNASEH2A deficiency and four were from the patient with RNASEH2B deficiency. The other individuals provided two samples each.

For mutational analyses, raw RNA-seq files from the study were sourced from the NCBI SRA repository (https://trace.ncbi.nlm.nih.gov/Traces/study/?acc=SRP041718). The samples were subsequently processed according to GATK best practices for variant calling from RNA-Seq data (Figure 5.1; https://software.broadinstitute.org/gatk/best-practices/workflow?id=11164).

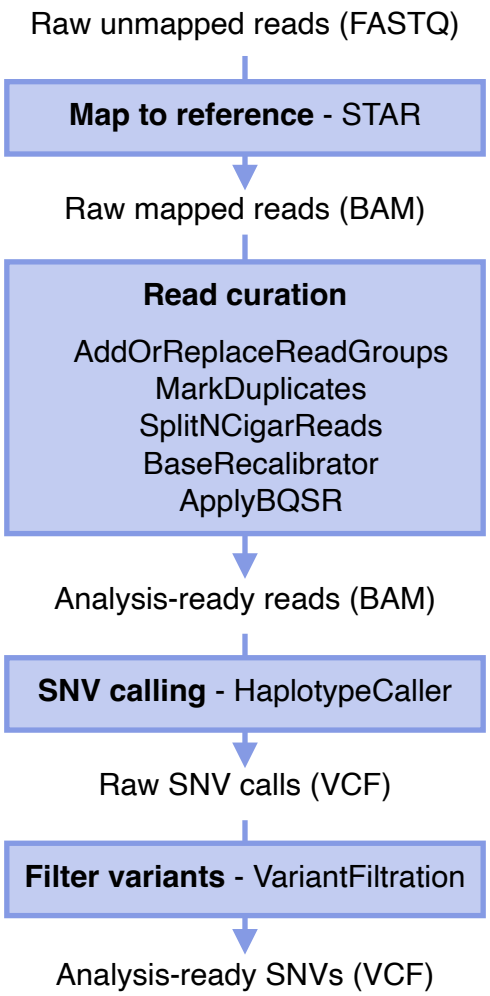


Figure 5.1 - A diagram summarising the processing of RNA-Seq data using the GATK best practices pipeline.

Quality control was performed on the raw FASTQ format RNA-seq data using FastQC, to evaluate read quality scores, and TrimGalore, to perform read trimming using a quality threshold of 20. Reads were mapped to the reference human genome build 38 using the STAR aligner to generate BAM files. Then, a number of data curation steps are performed using GATK tools, including the identification of artefacts arising from duplicate reads and base recalibration to correct for systematic errors in the base quality

scores. Then, variants were called into VCF files using the GATK HaplotypeCaller, and filtered with high threshold GATK VariantFiltration (using parameters -window 35 -cluster 3 -filterName FS -filter "FS > 30.0" -filterName QD -filter "QD < 2.0”).

VCF files were analysed using the deconstructSigs R package (Rosenthal et al. 2016) to identify the proportion of TCW mutations in each sample and generate mutation proportion graphs across all 96 mutation types.

For expression analyses, genome-wide expression data were sourced from NCBI GEO (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE57353). Calculations were performed in Microsoft Excel, GraphPad Prism and R.

In comparing AGS to APOBEC3B-deficient breast cancers, a list of APOBEC3B-deficient breast cancers that were either homozygous or heterozygous for the deletion polymorphism were sourced from Nik-Zainal et al. Fold change in expression was calculated using (mean[APOBEC3B-deficient cancers] + 0.5) / (mean[other cancers] + 0.5). The top 100 most highly upregulated and downregulated genes were subjected to PANTHER gene ontology analysis (http://geneontology.org/ Bonferroni-corrected $p < 0.05$ in all cases).

The formula for the power calculation described in the discussion section of the third results chapter is given in *Fundamentals of Biostatistics* (Rosner

2015) and was calculated online (https://clincalc.com/stats/samplesize.aspx).

List of Tables

List of Figures

Bibliography

Abascal, Federico, Luke MR Harvey, Emily Mitchell, Andrew RJ Lawson, Stefanie V Lensing, Peter Ellis, Andrew JC Russell, Raul E Alcantara, Adrian Baez-Ortega, and Yichen Wang. 2021. 'Somatic mutation landscapes at single-molecule resolution', *Nature*, 593: 405-10.

Achleitner, Martin, Martin Kleefisch, Alexander Hennig, Katrin Peschke, Anastasia Polikarpova, Reinhard Oertel, Benjamin Gabriel, Livia Schulze, Dirk Lindeman, and Alexander Gerbaulet. 2017. 'Lack of Trex1 causes systemic autoimmunity despite the presence of antiretroviral drugs', *The Journal of Immunology*, 199: 2261-69.

Ahmad, AS, N Ormiston-Smith, and PD Sasieni. 2015. 'Trends in the lifetime risk of developing cancer in Great Britain: comparison of risk for those born from 1930 to 1960', *British journal of cancer*, 112: 943-47.

Ahmad, Sadeem, Xin Mu, Fei Yang, Emily Greenwald, Ji Woo Park, Etai Jacob, Cheng-Zhong Zhang, and Sun Hur. 2018. 'Breaching self-tolerance to Alu duplex RNA underlies MDA5-mediated inflammation', *Cell*, 172: 797-810. e13.

Ahmad, SS, K Ahmed, and AR Venkitaraman. 2018. 'Science in Focus: Genomic instability and its implications for clinical cancer care', *Clinical Oncology*, 30: 751-55.

Aicardi, J, and F Goutières. 1984. 'A progressive familial encephalopathy in infancy with calcifications of the basal ganglia and chronic cerebrospinal fluid lymphocytosis', *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 15: 49-54.

Aicardi, J, and Françoise Goutières. 2000. 'Systemic lupus erythematosus or Aicardi-Goutieres syndrome?', *Neuropediatrics*, 31: 113-13.

Alexandrov, Ludmil B, Jaegil Kim, Nicholas J Haradhvala, Mi Ni Huang, Alvin Wei Tian Ng, Yang Wu, Arnoud Boot, Kyle R Covington, Dmitry A Gordenin, and Erik N Bergstrom. 2020. 'The repertoire of mutational signatures in human cancer', *Nature*, 578: 94-101.

Alexandrov, Ludmil B, and Michael R Stratton. 2014. 'Mutational signatures: the patterns of somatic mutations hidden in cancer genomes', *Current opinion in genetics & development*, 24: 52-60.

Alexandrov, Ludmil B., Serena Nik-Zainal, David C. Wedge, Samuel A. J. R. Aparicio, Sam Behjati, Andrew V. Biankin, Graham R. Bignell, Niccol Bolli, Ake Borg, Anne-Lise Brresen-Dale, Sandrine Boyault, Birgit Burkhardt, Adam P. Butler, Carlos Caldas, Helen R. Davies, Christine Desmedt, Roland Eils, Jrunn Erla Eyfjrd, John A. Foekens, Mel Greaves, Fumie Hosoda, Barbara Hutter, Tomislav Ilicic, Sandrine Imbeaud, Marcin Imielinski, Marcin Imielinsk, Natalie Jger, David T. W. Jones, David Jones, Stian Knappskog, Marcel Kool, Sunil R. Lakhani, Carlos Lpez-Otn, Sancha Martin, Nikhil C. Munshi, Hiromi Nakamura, Paul A. Northcott, Marina Pajic, Elli Papaemmanuil, Angelo Paradiso, John V. Pearson, Xose S. Puente, Keiran Raine, Manasa Ramakrishna, Andrea L. Richardson, Julia Richter, Philip Rosenstiel, Matthias Schlesner, Ton N. Schumacher, Paul N. Span, Jon W. Teague, Yasushi Totoki, Andrew N. J. Tutt, Rafael Valds-Mas, and Marit M. and van Buuren. 2013. 'Signatures of mutational processes in human cancer.', *Nature*, 500: 415--21.

Andersson, Dan I, and Diarmaid Hughes. 1996. 'Muller's ratchet decreases fitness of a DNA-based microbe', *Proceedings of the National Academy of Sciences*, 93: 906-07.

Andor, Noemi, Trevor A. Graham, Marnix Jansen, Li C. Xia, C. Athena Aktipis, Claudia Petritsch, Hanlee P. Ji, and Carlo C. Maley. 2015. 'Pan-cancer analysis of the extent and consequences of intratumor heterogeneity', *Nature Medicine*, 22: 105--13.

Ardeljan, Daniel, Xuya Wang, Mehrnoosh Oghbaie, Martin S Taylor, David Husband, Vikram Deshpande, Jared P Steranka, Mikhail Gorbounov, Wan Rou Yang, and Brandon Sie. 2020. 'LINE-1 ORF2p expression is nearly imperceptible in human cancers', *Mobile DNA*, 11: 1-19.

Armitage, Peter, and Richard Doll. 1954. 'The age distribution of cancer and a multi-stage theory of carcinogenesis', *British journal of cancer*, 8: 1.

Avery, Oswald T, Colin M MacLeod, and Maclyn McCarty. 1944. 'Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III', *The Journal of Experimental Medicine*, 79: 137-58.

Baldwin, Eric T., Trevor van Eeuwen, David Hoyos, Arthur Zalevsky, Egor P. Tchesnokov, Roberto Sánchez, Bryant D. Miller, Luciano H. Di Stefano, Francesc Xavier Ruiz, Matthew Hancock, Esin Işik, Carlos Mendez-Dorantes, Thomas Walpole, Charles Nichols, Paul Wan, Kirsi Riento, Rowan Halls-Kass, Martin Augustin, Alfred Lammens, Anja Jestel, Paula Upla, Kera Xibinaku, Samantha Congreve, Maximiliaan Hennink, Kacper B. Rogala, Anna M. Schneider, Jennifer E. Fairman, Shawn M. Christensen, Brian Desrosiers, Gregory S. Bisacchi, Oliver L. Saunders, Nafeeza

Hafeez, Wenyan Miao, Rosana Kapeller, Dennis M. Zaller, Andrej Sali, Oliver Weichenrieder, Kathleen H. Burns, Matthias Götte, Michael P. Rout, Eddy Arnold, Benjamin D. Greenbaum, Donna L. Romero, John LaCava, and Martin S. Taylor. 2023. 'Structures, functions, and adaptations of the human LINE-1 ORF2 protein', *Nature*.

Bartkova, Jirina, Zuzana Hořejší, Karen Koed, Alwin Krämer, Frederic Tort, Karsten Zieger, Per Guldberg, Maxwell Sehested, Jahn M Nesland, and Claudia Lukas. 2005. 'DNA damage response as a candidate anti-cancer barrier in early human tumorigenesis', *Nature*, 434: 864-70.

Beck-Engeser, Gabriele B., Dan Eilat, and Matthias Wabl. 2011. 'An autoimmune disease prevented by anti-retroviral drugs.', *Retrovirology*, 8: 91.

Behjati, Sam, Meritxell Huch, Ruben van Boxtel, Wouter Karthaus, David C Wedge, Asif U Tamuri, Iñigo Martincorena, Mia Petljak, Ludmil B Alexandrov, and Gunes Gundem. 2014. 'Genome sequencing of normal cells reveals developmental lineages and mutational processes', *Nature*, 513: 422.

Benitez-Guijarro, Maria, Cesar Lopez-Ruiz, Žygimantė Tarnauskaitė, Olga Murina, Mahwish Mian Mohammad, Thomas C Williams, Adeline Fluteau, Laura Sanchez, Raquel Vilar-Astasio, and Marta Garcia-Canadas. 2018. 'RNase H2, mutated in Aicardi-Goutières syndrome, promotes LINE-1 retrotransposition', *The EMBO Journal*, 37: e98506.

Birkbak, Nicolai J., Aron C. Eklund, Qiyuan Li, Sarah E. McClelland, David Endesfelder, Patrick Tan, Iain B. Tan, Andrea L. Richardson, Zoltan Szallasi, and Charles Swanton. 2011. 'Paradoxical relationship between

chromosomal instability and survival outcome in cancer.', *Cancer Research*, 71: 3447--52.

Bogerd, Hal P, Heather L Wiegand, Brian P Doehle, Kira K Lueders, and Bryan R Cullen. 2006. 'APOBEC3A and APOBEC3B are potent inhibitors of LTR-retrotransposon function in human cells', *Nucleic Acids Research*, 34: 89-95.

Boveri, Theodor. 1914. 'Zur frage der entstehung maligner tumoren'.

Bray, Freddie, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. 2018. 'Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries', *CA: a cancer journal for clinicians*, 68: 394-424.

Brown, William L, Emily K Law, Prokopios P Argyris, Michael A Carpenter, Rena Levin-Klein, Alison N Ranum, Amy M Molan, Colleen L Forster, Brett D Anderson, and Lela Lackey. 2019. 'A rabbit monoclonal antibody against the antiviral and cancer genomic DNA mutating enzyme APOBEC3B', *Antibodies*, 8: 47.

Bunz, Fred, A Dutriaux, C Lengauer, T Waldman, Shibin Zhou, JP Brown, JM Sedivy, Kenneth W Kinzler, and Bert Vogelstein. 1998. 'Requirement for p53 and p21 to sustain G2 arrest after DNA damage', *Science*, 282: 1497-501.

Burns, Michael B., Lela Lackey, Michael A. Carpenter, Anurag Rathore, Allison M. Land, Brandon Leonard, Eric W. Refsland, Delshanee Kotandeniya, Natalia Tretyakova, Jason B. Nikas, Douglas Yee, Nuri A. Temiz, Duncan E. Donohue, Rebecca M. McDougle, William L. Brown, Emily K. Law, and Reuben S. Harris. 2013. 'APOBEC3B is an enzymatic source of mutation in breast cancer.', *Nature*, 494: 366--70.

Butler, Kelly, and A Rouf Banday. 2023. 'APOBEC3-mediated mutagenesis in cancer: causes, clinical significance and therapeutic potential', *Journal of Hematology & Oncology*, 16: 1-25.

Campbell, Peter J, Gad Getz, Joshua M Stuart, Jan O Korbel, and Lincoln D Stein. 2017. 'Pan-cancer analysis of whole genomes', *BioRxiv*: 162784.

Caval, V., R. Suspne, M. Shapira, J. P. Vartanian, and S. Wain-Hobson. 2014. 'A prevalent cancer susceptibility APOBEC3A hybrid allele bearing APOBEC3B 3′UTR enhances chromosomal DNA damage', *Nature Communications*, 5.

Cescon, David W., Benjamin Haibe-Kains, and Tak W. Mak. 2015. 'APOBEC3B expression in breast cancer reflects cellular proliferation, while a deletion polymorphism is associated with immune activation.', *Proceedings of the National Academy of Sciences of the United States of America*, 112: 2841--6.

Chan, Kin, Steven A. Roberts, Leszek J. Klimczak, Joan F. Sterling, Natalie Saini, Ewa P. Malc, Jaegil Kim, David J. Kwiatkowski, David C. Fargo, Piotr A. Mieczkowski, Gad Getz, and Dmitry A. Gordenin. 2015. 'An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers.', *Nature Genetics*, 47: 1067--72.

Chen, Guoli, Stacy Mosier, Christopher D Gocke, Ming-Tseh Lin, and James R Eshleman. 2014. 'Cytosine deamination is a major cause of baseline noise in next-generation sequencing', *Molecular diagnosis & therapy*, 18: 587-93.

Chen, Lixin, Pingfang Liu, Thomas C Evans, and Laurence M Ettwiller. 2017. 'DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification', *Science*, 355: 752-56.

Chung, Hachung, Jorg JA Calis, Xianfang Wu, Tony Sun, Yingpu Yu, Stephanie L Sarbanes, Viet Loan Dao Thi, Abigail R Shilvock, H-Heinrich Hoffmann, and Brad R Rosenberg. 2018. 'Human ADAR1 prevents endogenous RNA from triggering translational shutdown', *Cell*, 172: 811-24. e14.

Ciriello, Giovanni, Martin L Miller, Bülent Arman Aksoy, Yasin Senbabaoglu, Nikolaus Schultz, and Chris Sander. 2013. 'Emerging landscape of oncogenic signatures across human cancers', *Nature Genetics*, 45: 1127-33.

Cogliano, Vincent James, Robert Baan, Kurt Straif, Yann Grosse, Béatrice Lauby-Secretan, Fatiha El Ghissassi, Véronique Bouvard, Lamia Benbrahim-Tallaa, Neela Guha, and Crystal Freeman. 2011. 'Preventable exposures associated with human cancers', *Journal of the National Cancer Institute*, 103: 1827-39.

Consortium, ENCODE Project. 2012. 'An integrated encyclopedia of DNA elements in the human genome', *Nature*, 489: 57.

Consortium, International Cancer Genome. 2010. 'International network of cancer genome projects', *Nature*, 464: 993.

Conticello, Silvestro G, Cornelia JF Thomas, Svend K Petersen-Mahrt, and Michael S Neuberger. 2005. 'Evolution of the AID/APOBEC family of polynucleotide (deoxy) cytidine deaminases', *Molecular biology and evolution*, 22: 367-77.

Conticello, Silvestro G. 2008. 'The AID/APOBEC family of nucleic acid mutators.', *Genome Biology*, 9: 229.

Conticello, Silvestro G., Marc-Andre Langlois, Zizhen Yang, and Michael S. Neuberger. 2007. 'DNA Deamination in Immunity: AID in the Context of Its APOBEC Relatives', *Advances in Immunology*, 94: 37--73.

Cordaux, Richard, and Mark A Batzer. 2009. 'The impact of retrotransposons on human genome evolution', *Nature Reviews Genetics*, 10: 691.

Cortez, Luis M, Amber L Brown, Madeline A Dennis, Christopher D Collins, Alexander J Brown, Debra Mitchell, Tony M Mertz, and Steven A Roberts. 2019. 'APOBEC3A is a prominent cytidine deaminase in breast cancer', *PLOS Genetics*, 15: e1008545.

Crow, Yanick J. 2011. 'Type I interferonopathies: a novel set of inborn errors of immunity', *Annals of the New York Academy of Sciences*, 1238: 91-98.

Crow, Yanick J, and Nicolas Manel. 2015. 'Aicardi–Goutières syndrome and the type I interferonopathies', *Nature Reviews Immunology*, 15: 429.

Crow, Yanick J, Jayakara Shetty, and John H Livingston. 2020. 'Treatments in aicardi–goutières syndrome', *Developmental Medicine & Child Neurology*, 62: 42-47.

Crow, Yanick J, and Daniel B Stetson. 2022. 'The type I interferonopathies: 10 years on', *Nature Reviews Immunology*, 22: 471-83.

Curtis, Christina, Sohrab P. Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M. Rueda, Mark J. Dunning, Doug Speed, Andy G. Lynch, Shamith Samarajiwa, Yinyin Yuan, Stefan Grf, Gavin Ha, Gholamreza Haffari, Ali Bashashati, Roslin Russell, Steven McKinney, Anita Langerd, Andrew Green, Elena Provenzano, Gordon Wishart, Sarah Pinder, Peter Watson,

Florian Markowetz, Leigh Murphy, Ian Ellis, Arnie Purushotham, Anne-Lise Brresen-Dale, James D. Brenton, and Tavar. 2012. 'The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.', *Nature*, 486: 346--52.

Dai, Lixin, Qing Huang, and Jef D. Boeke. 2011. 'Effect of reverse transcriptase inhibitors on LINE-1 and Ty1 reverse transcriptase activities and on LINE-1 retrotransposition.', *BMC biochemistry*, 12: 18.

Davies, Helen, Dominik Glodzik, Sandro Morganella, Lucy R. Yates, Johan Staaf, Xueqing Zou, Manasa Ramakrishna, Sancha Martin, Sandrine Boyault, Anieta M. Sieuwerts, Peter T. Simpson, Tari A. King, Keiran Raine, Jorunn E. Eyfjord, Gu Kong, ke Borg, Ewan Birney, Hendrik G. Stunnenberg, Marc J. van de Vijver, Anne-Lise Brresen-Dale, John W. M. Martens, Paul N. Span, Sunil R. Lakhani, Anne Vincent-Salomon, Christos Sotiriou, Andrew Tutt, and Alastair M. and Thompson. 2017. 'HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures', *Nature Medicine*, 23: 517--25.

de Bruin, Elza C, Nicholas McGranahan, Richard Mitter, Max Salm, David C Wedge, Lucy Yates, Mariam Jamal-Hanjani, Seema Shafi, Nirupa Murugaesu, and Andrew J Rowan. 2014. 'Spatial and temporal diversity in genomic instability processes defines lung cancer evolution', *Science*, 346: 251-56.

Degasperi, Andrea, Tauanne Dias Amarante, Jan Czarnecki, Scott Shooter, Xueqing Zou, Dominik Glodzik, Sandro Morganella, Arjun S Nanda, Cherif Badja, and Gene Koh. 2020. 'A practical framework and online tool for mutational signature analyses show intertissue variation and driver dependencies', *Nature cancer*, 1: 249-63.

Deininger, Prescott, Maria E Morales, Travis B White, Melody Baddoo, Dale J Hedges, Geraldine Servant, Sudesh Srivastav, Madison E Smither, Monica Concha, and Dawn L DeHaro. 2016. 'A comprehensive approach to expression of L1 loci', *Nucleic Acids Research*, 45: e31-e31.

Denli, Ahmet M, Iñigo Narvaiza, Bilal E Kerman, Monique Pena, Christopher Benner, Maria CN Marchetto, Jolene K Diedrich, Aaron Aslanian, Jiao Ma, and James J Moresco. 2015. 'Primate-specific ORF0 contributes to retrotransposon-mediated diversity', *Cell*, 163: 583-93.

DePristo, Mark A, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo Del Angel, Manuel A Rivas, and Matt Hanna. 2011. 'A framework for variation discovery and genotyping using next-generation DNA sequencing data', *Nature Genetics*, 43: 491.

DeVita Jr, Vincent T, and Steven A Rosenberg. 2012. 'Two hundred years of cancer research', *New England journal of medicine*, 366: 2207-14.

Dewannieux, Marie, Cécile Esnault, and Thierry Heidmann. 2003. 'LINE-mediated retrotransposition of marked Alu sequences', *Nature Genetics*, 35: 41-48.

Dombroski, Beth A, Stephen L Mathias, Elizabeth Nanthakumar, Alan F Scott, and Haig H Kazazian Jr. 1991. 'Isolation of an active human transposable element', *Science*, 254: 1805-08.

Edelman, Laurence M., Raymond Cheong, and Jason D. Kahn. 2003. 'Fluorescence resonance energy transfer over approximately 130 basepairs in hyperstable lac repressor-DNA loops.', *Biophysical journal*, 84: 1131--45.

Eisenberg, Eli, and Erez Y Levanon. 2013. 'Human housekeeping genes, revisited', *TRENDS in Genetics*, 29: 569-74.

Elledge, Stephen J. 1996. 'Cell cycle checkpoints: preventing an identity crisis', *Science*, 274: 1664-72.

Elliott, Kerryn, and Erik Larsson. 2021. 'Non-coding driver mutations in human cancer', *Nature Reviews Cancer*, 21: 500-09.

Engelstädter, Jan. 2008. 'Muller's ratchet and the degeneration of Y chromosomes: a simulation study', *Genetics*, 180: 957-67.

Ewing, Adam D. 2015. 'Transposable element detection from whole genome sequence data.', *Mobile DNA*, 6: 24.

Faltas, Bishoy M, Davide Prandi, Scott T Tagawa, Ana M Molina, David M Nanus, Cora Sternberg, Jonathan Rosenberg, Juan Miguel Mosquera, Brian Robinson, and Olivier Elemento. 2016. 'Clonal evolution of chemotherapy-resistant urothelial carcinoma', *Nature Genetics*, 48: 1490.

Fearon, Eric R, and Bert Vogelstein. 1990. 'A genetic model for colorectal tumorigenesis', *Cell*, 61: 759-67.

Feng, Qinghua, John V Moran, Haig H Kazazian, and Jef D Boeke. 1996. 'Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition', *Cell*, 87: 905-16.

Finlay, CA, PW Hinds, and AJ Levine. 1989. 'The p53 proto-oncogene can act as a suppressor of transformation', *Cell*, 57: 1083-93.

Fowler, Benjamin J, Bradley D Gelfand, Younghee Kim, Nagaraj Kerur, Valeria Tarallo, Yoshio Hirano, Shoba Amarnath, Daniel H Fowler, Marta Radwan, and Mark T Young. 2014. 'Nucleoside reverse transcriptase inhibitors possess intrinsic anti-inflammatory activity', *Science*, 346: 1000-03.

Fox, Edward J, and Lawrence A Loeb. 2010. "Lethal mutagenesis: targeting the mutator phenotype in cancer." In *Seminars in cancer biology*, 353-59. Elsevier.

Franklin, Rosalind E, and Raymond G Gosling. 1953. 'Molecular configuration in sodium thymonucleate', *Nature*, 171: 740-41.

Friedberg, Errol C., Richard D. Wood, Graham C. Walker, Roger A. Schultz, Wolfram Siede, Tom Ellenberger, and Errol C. Friedberg. 2006. *DNA repair and mutagenesis* (ASM Press).

Friend, Stephen H, Rene Bernards, Snezna Rogelj, Robert A Weinberg, Joyce M Rapaport, Daniel M Albert, and Thaddeus P Dryja. 1986. 'A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma', *Nature*, 323: 643-46.

Garcia-Perez, Jose L, Maria Morell, Joshua O Scheys, Deanna A Kulpa, Santiago Morell, Christoph C Carter, Gary D Hammer, Kathleen L Collins, K Sue O'Shea, and Pablo Menendez. 2010. 'Epigenetic silencing of engineered L1 retrotransposition events in human embryonic carcinoma cells', *Nature*, 466: 769-73.

Gasior, Stephen L., Astrid M. Roy-Engel, and Prescott L. Deininger. 2008. 'ERCC1/XPF limits L1 retrotransposition', *DNA Repair*, 7: 983--89.

Gasior, Stephen L., Timothy P. Wakeman, Bo Xu, and Prescott L. Deininger. 2006. 'The human LINE-1 retrotransposon creates DNA double-strand breaks.', *Journal of Molecular Biology*, 357: 1383--93.

Gerlinger, Marco, Andrew J Rowan, Stuart Horswell, James Larkin, David Endesfelder, Eva Gronroos, Pierre Martinez, Nicholas Matthews, Aengus Stewart, and Patrick Tarpey. 2012. 'Intratumor heterogeneity and branched

evolution revealed by multiregion sequencing', *New England journal of medicine*, 366: 883-92.

Gerstung, Moritz, Clemency Jolly, Ignaty Leshchiner, Stefan C Dentro, Santiago Gonzalez, Daniel Rosebrock, Thomas J Mitchell, Yulia Rubanova, Pavana Anur, and Kaixian Yu. 2020. 'The evolutionary history of 2,658 cancers', *Nature*, 578: 122-28.

Gillick, Kieran, Darja Pollpeter, Prabhjeet Phalora, Eun-Young Kim, Steven M Wolinsky, and Michael H Malim. 2013. 'Suppression of HIV-1 infection by APOBEC3 proteins in primary human CD4+ T cells is associated with inhibition of processive reverse transcription as well as excessive cytidine deamination', *Journal of Virology*, 87: 1508-17.

Godek, Kristina M., Monica Venere, Quilian Wu, Kevin D. Mills, William F. Hickey, Jeremy N. Rich, and Duane A. Compton. 2016. 'Chromosomal Instability Affects the Tumorigenicity of Glioblastoma Tumor-Initiating Cells.', *Cancer discovery*, 6: 532--45.

Goodwin, Sara, John D McPherson, and W Richard McCombie. 2016. 'Coming of age: ten years of next-generation sequencing technologies', *Nature Reviews Genetics*, 17: 333.

Gorgoulis, Vassilis G, Leandros-Vassilios F Vassiliou, Panagiotis Karakaidos, Panayotis Zacharatos, Athanassios Kotsinas, Triantafillos Liloglou, Monica Venere, Richard A DiTullio, Nikolaos G Kastrinakis, and Brynn Levy. 2005. 'Activation of the DNA damage checkpoint and genomic instability in human precancerous lesions', *Nature*, 434: 907-13.

Greaves, Mel, and Carlo C Maley. 2012. 'Clonal evolution in cancer', *Nature*, 481: 306.

Greenblatt, MS, William P Bennett, M Hollstein, and CC Harris. 1994. 'Mutations in the p53 tumor suppressor gene: clues to cancer etiology and molecular pathogenesis', *Cancer Research*, 54: 4855-78.

Hafner, Antonina, Martha L Bulyk, Ashwini Jambhekar, and Galit Lahav. 2019. 'The multiple mechanisms that regulate p53 activity and cell fate', *Nature Reviews Molecular Cell Biology*, 20: 199-210.

Hajdu, Steven I. 2011. 'A note from history: landmarks in history of cancer, part 1', *Cancer*, 117: 1097-102.

Hamdorf, Matthias, Adam Idica, Dimitrios G Zisoulis, Lindsay Gamelin, Charles Martin, Katie J Sanders, and Irene M Pedersen. 2015. 'miR-128 represses L1 retrotransposition by binding directly to L1 RNA', *Nature structural & molecular biology*, 22: 824-31.

Hammal, Fayrouz, Pierre de Langen, Aurélie Bergon, Fabrice Lopez, and Benoit Ballester. 2022. 'ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments', *Nucleic Acids Research*, 50: D316-D25.

Hanahan, Douglas, and Robert A Weinberg. 2011. 'Hallmarks of cancer: the next generation', *Cell*, 144: 646-74.

Hancks, Dustin C, John L Goodier, Prabhat K Mandal, Ling E Cheung, and Haig H Kazazian Jr. 2011. 'Retrotransposition of marked SVA elements by human L1s in cultured cells', *Human molecular genetics*, 20: 3386-400.

Haoudi, Abdelali, O John Semmes, James M Mason, and Ronald E Cannon. 2004. 'Retrotransposition-competent human LINE-1 induces apoptosis in cancer cells with intact p53', *Journal of Biomedicine and Biotechnology*, 2004: 185-94.

Harper, J Wade, Guy R Adami, Nan Wei, Khandan Keyomarsi, and Stephen J Elledge. 1993. 'The p21 Cdk-interacting protein Cip1 is a potent inhibitor of G1 cyclin-dependent kinases', *Cell*, 75: 805-16.

Harris, CR, A Dewan, A Zupnick, R Normart, A Gabriel, C Prives, AJ Levine, and J Hoh. 2009. 'p53 responsive elements in human retrotransposons', *Oncogene*, 28: 3857-65.

Harris, Reuben S, and Jaquelin P Dudley. 2015. 'APOBECs and virus restriction', *Virology*, 479: 131-45.

Hastie, Trevor, Robert Tibshirani, Jerome Friedman, and James Franklin. 2005. 'The elements of statistical learning: data mining, inference and prediction', *The Mathematical Intelligencer*, 27: 83-85.

Hata, Kikumi, and Yoshiyuki Sakaki. 1997. 'Identification of critical CpG sites for repression of L1 transcription by DNA methylation', *Gene*, 189: 227-34.

Hattori, Hiroyoshi, Rekin's Janky, Wilfried Nietfeld, Stein Aerts, M Madan Babu, and Ashok R Venkitaraman. 2014. 'p53 shapes genome-wide and cell type-specific changes in microRNA expression during the human DNA damage response', *Cell Cycle*, 13: 2572-86.

Haupt, Ygal, Ruth Maya, Anat Kazaz, and Moshe Oren. 1997. 'Mdm2 promotes the rapid degradation of p53', *Nature*, 387: 296-99.

Hayward, Joshua A, Mary Tachedjian, Jie Cui, Adam Z Cheng, Adam Johnson, Michelle L Baker, Reuben S Harris, Lin-Fa Wang, and Gilda Tachedjian. 2018. 'Differential evolution of antiretroviral restriction factors in pteropid bats as revealed by APOBEC3 gene complexity', *Molecular biology and evolution*, 35: 1626-37.

Henderson, Stephen, Ankur Chakravarthy, Xiaoping Su, Chris Boshoff, and Tim Robert Fenton. 2014. 'APOBEC-mediated cytosine deamination links PIK3CA helical domain mutations to human papillomavirus-driven tumor development', *Cell Reports*, 7: 1833-41.

Herrmann, Alexandra, Sabine Wittmann, Dominique Thomas, Caitlin N Shepard, Baek Kim, Nerea Ferreirós, and Thomas Gramberg. 2018. 'The SAMHD1-mediated block of LINE-1 retroelements is regulated by phosphorylation', *Mobile DNA*, 9: 1-17.

Hingorani, Aroon D, Valerie Kuan, Chris Finan, Felix A Kruger, Anna Gaulton, Sandesh Chopade, Reecha Sofat, Raymond J MacAllister, John P Overington, and Harry Hemingway. 2019. 'Improving the odds of drug development success through human genomics: modelling study', *Scientific Reports*, 9: 1-25.

Hirabayashi, Shigeki, Kotaro Shirakawa, Yoshihito Horisawa, Tadahiko Matsumoto, Hiroyuki Matsui, Hiroyuki Yamazaki, Anamaria Daniela Sarca, Yasuhiro Kazuma, Ryosuke Nomura, and Yoshinobu Konishi. 2021. 'APOBEC3B is preferentially expressed at the G2/M phase of cell cycle', *Biochemical and biophysical research communications*, 546: 178-84.

Hirano, K, Jing Min, Toru Funahashi, David A Baunoch, and Nicholas O Davidson. 1997. 'Characterization of the human apobec-1 gene: expression in gastrointestinal tissues determined by alternative splicing with production of a novel truncated peptide', *Journal of lipid research*, 38: 847-59.

Hohjoh, Hirohiko, and Maxine F Singer. 1997. 'Sequence-specific single-strand RNA binding protein encoded by the human LINE-1 retrotransposon', *The EMBO Journal*, 16: 6034-43.

Holtz, Colleen M., Holly A. Sadler, and Louis M. Mansky. 2013. 'APOBEC3G cytosine deamination hotspots are defined by both sequence context and single-stranded DNA secondary structure.', *Nucleic Acids Research*, 41: 6139--48.

Hukezalie, Kyle R, Naresh R Thumati, Helene CF Coˆte, and Judy MY Wong. 2012. 'In vitro and ex vivo inhibition of human telomerase by anti-HIV nucleoside reverse transcriptase inhibitors (NRTIs) but not by non-NRTIs', *PLoS ONE*, 7: e47505.

Isozaki, Hideko, Ramin Sakhtemani, Ammal Abbasi, Naveed Nikpour, Marcello Stanzione, Sunwoo Oh, Adam Langenbucher, Susanna Monroe, Wenjia Su, and Heidie Frisco Cabanos. 2023. 'Therapy-induced APOBEC3A drives evolution of persistent cancer cells', *Nature*, 620: 393-401.

Ito, Jumpei, Robert J Gifford, and Kei Sato. 2020. 'Retroviruses drive the rapid evolution of mammalian APOBEC3 genes', *Proceedings of the National Academy of Sciences*, 117: 610-18.

Ivancevic, Atma M, R Daniel Kortschak, Terry Bertozzi, and David L Adelson. 2016. 'LINEs between species: evolutionary dynamics of LINE-1 retrotransposons across the eukaryotic tree of life', *Genome biology and evolution*, 8: 3301-22.

Jackson, Stephen P, and Jiri Bartek. 2009. 'The DNA-damage response in human biology and disease', *Nature*, 461: 1071-78.

Jakobsdottir, G Maria, Daniel S Brewer, Colin Cooper, Catherine Green, and David C Wedge. 2022. 'APOBEC3 mutational signatures are associated with extensive and diverse genomic instability across multiple tumour types', *BMC biology*, 20: 1-12.

Jamal-Hanjani, M., R. A'Hern, N. J. Birkbak, P. Gorman, E. Grnroos, S. Ngang, P. Nicola, L. Rahman, E. Thanopoulou, G. Kelly, P. Ellis, P. Barrett-Lee, S. R. D. Johnston, J. Bliss, R. Roylance, and C. Swanton. 2015. 'Extreme chromosomal instability forecasts improved outcome in ER-negative breast cancer: a prospective validation cohort study from the TACT trial.', *Annals of oncology : official journal of the European Society for Medical Oncology / ESMO*, 26: 1340--6.

Janahi, EM, and MJ McGarvey. 2013. 'The inhibition of hepatitis B virus by APOBEC cytidine deaminases', *Journal of viral hepatitis*, 20: 821-28.

Jeffrey, Philip D, Svetlana Gorina, and Nikola P Pavletich. 1995. 'Crystal structure of the tetramerization domain of the p53 tumor suppressor at 1.7 angstroms', *Science*, 267: 1498-502.

Jones, R Brad, Keith E Garrison, Jessica C Wong, Erick H Duan, Douglas F Nixon, and Mario A Ostrowski. 2008. 'Nucleoside analogue reverse transcriptase inhibitors differentially inhibit human LINE-1 retrotransposition', *PLoS ONE*, 3: e1547.

Jurka, Jerzy. 1997. 'Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons', *Proceedings of the National Academy of Sciences*, 94: 1872-77.

Kanu, Nnennaya, Maria Antonietta Cerone, Gerald Goh, Lykourgos-Panagiotis Zalmas, Jirina Bartkova, Michelle Dietzen, Nicholas McGranahan, Rebecca Rogers, Emily K Law, and Irina Gromova. 2016. 'DNA replication stress mediates APOBEC3 family mutagenesis in breast cancer', *Genome Biology*, 17: 185.

Kastenhuber, Edward R, and Scott W Lowe. 2017. 'Putting p53 in context', *Cell*, 170: 1062-78.

Katz, Lior H, Ying Li, Jiun-Sheng Chen, Nina M Muñoz, Avijit Majumdar, Jian Chen, and Lopa Mishra. 2013. 'Targeting TGF-β signaling in cancer', *Expert opinion on therapeutic targets*, 17: 743-60.

Kemp, Jacqueline R., and Michelle S. Longworth. 2015. 'Crossing the LINE Toward Genomic Instability: LINE-1 Retrotransposition in Cancer.', *Frontiers in chemistry*, 3: 68.

Kimura, Motoo. 1967. 'On the evolutionary adjustment of spontaneous mutation rates', *Genetics Research*, 9: 23-34.

Kinomoto, Masanobu, Takayuki Kanno, Mari Shimura, Yukihito Ishizaka, Asato Kojima, Takeshi Kurata, Tetsutaro Sata, and Kenzo Tokunaga. 2007. 'All APOBEC3 family proteins differentially inhibit LINE-1 retrotransposition.', *Nucleic Acids Research*, 35: 2955--64.

Knudson, Alfred G. 1971. 'Mutation and cancer: statistical study of retinoblastoma', *Proceedings of the National Academy of Sciences*, 68: 820-23.

Kretschmer, Stefanie, Christine Wolf, Nadja König, Wolfgang Staroske, Jochen Guck, Martin Häusler, Hella Luksch, Laura A Nguyen, Baek Kim, and Dimitra Alexopoulou. 2015. 'SAMHD1 prevents autoimmunity by maintaining genome stability', *Annals of the rheumatic diseases*, 74: e17-e17.

Krishnan, Arunkumar, Lakshminarayan M Iyer, Stephen J Holland, Thomas Boehm, and L Aravind. 2018. 'Diversification of AID/APOBEC-like deaminases in metazoa: multiplicity of clades and widespread roles in immunity', *Proceedings of the National Academy of Sciences*, 115: E3201-E10.

Lackey, Lela, Emily K. Law, William L. Brown, and Reuben S. Harris. 2013. 'Subcellular localization of the APOBEC3 proteins during mitosis and implications for genomic DNA deamination', *Cell Cycle*, 12: 762--72.

Lander, Eric S, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, and William FitzHugh. 2001. 'Initial sequencing and analysis of the human genome'.

Lane, David P. 1992. 'Cancer. p53, guardian of the genome', *Nature*, 358: 15-16.

Lane, David P, and Lionel V Crawford. 1979. 'T antigen is bound to a host protein in SY40-transformed cells', *Nature*, 278: 261-63.

Laptenko, O, and C Prives. 2006. 'Transcriptional regulation by p53: one protein, many possibilities', *Cell Death & Differentiation*, 13: 951-61.

Law, Emily K, Anieta M Sieuwerts, Kelly LaPara, Brandon Leonard, Gabriel J Starrett, Amy M Molan, Nuri A Temiz, Rachel Isaksson Vogel, Marion E Meijer-van Gelder, and Fred CGJ Sweep. 2016. 'The DNA cytosine deaminase APOBEC3B promotes tamoxifen resistance in ER-positive breast cancer', *Science advances*, 2: e1601737.

Lee, Daniel D, and H Sebastian Seung. 1999. 'Learning the parts of objects by non-negative matrix factorization', *Nature*, 401: 788-91.

Lee, Hyunsook, Alison H Trainer, Lori S Friedman, Fiona C Thistlethwaite, Martin J Evans, Bruce AJ Ponder, and Ashok R Venkitaraman. 1999. 'Mitotic checkpoint inactivation fosters transformation in cells lacking the breast cancer susceptibility gene, Brca2', *Molecular Cell*, 4: 1-10.

Lee, Yu-Ru, Ming Chen, and Pier Paolo Pandolfi. 2018. 'The functions and regulation of the PTEN tumour suppressor: new modes and prospects', *Nature Reviews Molecular Cell Biology*, 19: 547-62.

Lerner, Taga, F Papavasiliou, and Riccardo Pecori. 2019. 'RNA editors, cofactors, and mRNA targets: an overview of the C-to-U RNA editing machinery and its implication in human disease', *Genes*, 10: 13.

Lewin, Benjamin. 2004. *Genes Viii*.

Lewis, William, Brian J Day, and William C Copeland. 2003. 'Mitochondrial toxicity of NRTI antiviral drugs: an integrated cellular perspective', *Nature reviews Drug discovery*, 2: 812-22.

Li, Peng, Juan Du, John L Goodier, Jingwei Hou, Jian Kang, Haig H Kazazian Jr, Ke Zhao, and Xiao-Fang Yu. 2017. 'Aicardi–Goutières syndrome protein TREX1 suppresses L1 and maintains genome integrity through exonuclease-independent ORF1p depletion', *Nucleic Acids Research*, 45: 4619-31.

Li, Yilong, Nicola D Roberts, Jeremiah A Wala, Ofer Shapira, Steven E Schumacher, Kiran Kumar, Ekta Khurana, Sebastian Waszak, Jan O Korbel, and James E Haber. 2020. 'Patterns of somatic structural variation in human cancer genomes', *Nature*, 578: 112-21.

Lim, Yoong Wearn, Lionel A. Sanz, Xiaoqin Xu, Stella R. Hartono, and Frdric Chdin. 2015. 'Genome-wide DNA hypomethylation and RNA:DNA hybrid accumulation in Aicardi–Gouti \` e res syndrome', *eLife*, 4: e08007.

Liu, Yansheng, Yang Mi, Torsten Mueller, Saskia Kreibich, Evan G Williams, Audrey Van Drogen, Christelle Borel, Max Frank, Pierre-Luc Germain, and Isabell Bludau. 2019. 'Multi-omic measurements of heterogeneity in HeLa cells across laboratories', *Nature biotechnology*, 37: 314-22.

Loewe, Laurence. 2006. 'Quantifying the genomic decay paradox due to Muller's ratchet in human mitochondrial DNA', *Genetics Research*, 87: 133-59.

Loewe, Laurence, and Asher D Cutter. 2008. 'On the potential for extinction by Muller's ratchet in Caenorhabditis elegans', *BMC evolutionary biology*, 8: 125.

Loewe, Laurence, and William G Hill. 2010. "The population genetics of mutations: good, bad and indifferent." In.: The Royal Society.

Loewe, Laurence, and Dunja K Lamatsch. 2008. 'Quantifying the threat of extinction from Muller's ratchet in the diploid Amazon molly (Poecilia formosa)', *BMC evolutionary biology*, 8: 88.

López, Saioa, Emilia L Lim, Stuart Horswell, Kerstin Haase, Ariana Huebner, Michelle Dietzen, Thanos P Mourikis, Thomas BK Watkins, Andrew Rowan, and Sally M Dewhurst. 2020. 'Interplay between whole-genome doubling and the accumulation of deleterious alterations in cancer evolution', *Nature Genetics*, 52: 283-93.

Lynch, Michael, Matthew S Ackerman, Jean-Francois Gout, Hongan Long, Way Sung, W Kelley Thomas, and Patricia L Foster. 2016. 'Genetic drift, selection and the evolution of the mutation rate', *Nature Reviews Genetics*, 17: 704.

Lynch, Michael, Reinhard Bürger, D Butcher, and Wilfried Gabriel. 1993. 'The mutational meltdown in asexual populations', *Journal of Heredity*, 84: 339-44.

Maciver, Sutherland K. 2016. 'Asexual amoebae escape Muller's ratchet through polyploidy', *Trends in parasitology*, 32: 855-62.

Mackenzie, Karen J, Paula Carroll, Carol-Anne Martin, Olga Murina, Adeline Fluteau, Daniel J Simpson, Nelly Olova, Hannah Sutcliffe, Jacqueline K Rainger, Andrea Leitch, Ruby T Osborn, Ann P Wheeler, Marcin Nowotny, Nick Gilbert, Tamir Chandra, Martin A. M. Reijns, and

Andrew P Jackson. 2017. 'cGAS surveillance of micronuclei links genome instability to innate immunity', *Nature*, 548: 461-85.

Mangeat, Bastien, Priscilla Turelli, Gersende Caron, Marc Friedli, Luc Perrin, and Didier Trono. 2003. 'Broad antiretroviral defence by human APOBEC3G through lethal editing of nascent reverse transcripts', *Nature*, 424: 99-103.

Marchetto, Maria C. N., Iigo Narvaiza, Ahmet M. Denli, Christopher Benner, Thomas A. Lazzarini, Jason L. Nathanson, Apu C. M. Paquola, Keval N. Desai, Roberto H. Herai, Matthew D. Weitzman, Gene W. Yeo, Alysson R. Muotri, and Fred H. Gage. 2013. 'Differential L1 regulation in pluripotent stem cells of humans and apes.', *Nature*, 503: 525--9.

Mardis, Elaine R, and Richard K Wilson. 2009. 'Cancer genome sequencing: a review', *Human molecular genetics*, 18: R163-R68.

Martincorena, Iñigo, Keiran M Raine, Moritz Gerstung, Kevin J Dawson, Kerstin Haase, Peter Van Loo, Helen Davies, Michael R Stratton, and Peter J Campbell. 2017. 'Universal patterns of selection in cancer and somatic tissues', *Cell*, 171: 1029-41. e21.

Mathias, Stephen L, Alan F Scott, Haig H Kazazian Jr, Jef D Boeke, and Abram Gabriel. 1991. 'Reverse transcriptase encoded by a human transposable element', *Science*, 254: 1808-10.

Mathur, Radhika. 2018. 'ARID1A loss in cancer: towards a mechanistic understanding', *Pharmacology & therapeutics*, 190: 15-23.

McCann, Joyce, and Bruce N Ames. 1976. 'Detection of carcinogens as mutagens in the Salmonella/microsome test: assay of 300 chemicals: discussion', *Proceedings of the National Academy of Sciences*, 73: 950-54.

McGranahan, Nicholas, Andrew JS Furness, Rachel Rosenthal, Sofie Ramskov, Rikke Lyngaa, Sunil Kumar Saini, Mariam Jamal-Hanjani, Gareth A Wilson, Nicolai J Birkbak, and Crispin T Hiley. 2016. 'Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade', *Science*, 351: 1463-69.

McKerrow, Wilson, Xuya Wang, Carlos Mendez-Dorantes, Paolo Mita, Song Cao, Mark Grivainis, Li Ding, John LaCava, Kathleen H Burns, and Jef D Boeke. 2022. 'LINE-1 expression in cancer correlates with p53 mutation, copy number alteration, and S phase checkpoint', *Proceedings of the National Academy of Sciences*, 119: e2115999119.

Menendez, Laura, Benedict B Benigno, and John F McDonald. 2004. 'L1 and HERV-W retrotransposons are hypomethylated in human ovarian carcinomas', *Molecular cancer*, 3: 1-5.

Mi, Huaiyu, Anushya Muruganujan, John T Casagrande, and Paul D Thomas. 2013. 'Large-scale gene function analysis with the PANTHER classification system', *Nature protocols*, 8: 1551-66.

Miyashita, Toshiyuki, Stanislaw Krajewski, Maryla Krajewska, Hong Gang Wang, HK Lin, Dan A Liebermann, Barbara Hoffman, and John C Reed. 1994. 'Tumor suppressor p53 is a regulator of bcl-2 and bax gene expression in vitro and in vivo', *Oncogene*, 9: 1799-805.

Moore, Luiza, Alex Cagan, Tim HH Coorens, Matthew DC Neville, Rashesh Sanghvi, Mathijs A Sanders, Thomas RW Oliver, Daniel Leongamornlert, Peter Ellis, and Ayesha Noorani. 2021. 'The mutational landscape of human somatic and germline cells', *Nature*, 597: 381-86.

Morganella, Sandro, Ludmil B Alexandrov, Dominik Glodzik, Xueqing Zou, Helen Davies, Johan Staaf, Anieta M Sieuwerts, Arie B Brinkman, Sancha

Martin, and Manasa Ramakrishna. 2016. 'The topography of mutational processes in breast cancer genomes', *Nature Communications*, 7: 11383.

Muckenfuss, Heide, Matthias Hamdorf, Ulrike Held, Mario Perkovic, Johannes Lwer, Klaus Cichutek, Egbert Flory, Gerald G. Schumann, and Carsten Mnk. 2006. 'APOBEC3 proteins inhibit human LINE-1 retrotransposition.', *The Journal of Biological Chemistry*, 281: 22161--72.

Muller, Hermann Joseph. 1964. 'The relation of recombination to mutational advance', *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 1: 2-9.

Müller, Johannes Peter. 1838. *Ueber den feinern Bau und die Formen der krankhaften Geschwülste, von Dr. Johannes Müller* (g. reimer).

Münk, Carsten, Anouk Willemsen, and Ignacio G Bravo. 2012. 'An ancient history of gene duplications, fusions and losses in the evolution of APOBEC3 mutators in mammals', *BMC evolutionary biology*, 12: 1-16.

Muramatsu, Masamichi, Kazuo Kinoshita, Sidonia Fagarasan, Shuichi Yamada, Yoichi Shinkai, and Tasuku Honjo. 2000. 'Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme', *Cell*, 102: 553-63.

NCI. 2024. 'HIV+ Tumor Molecular Characterization Project: Lung Cancer'. https://gdc.cancer.gov/content/cgci-htmcp-lc-publications-summary.

Negrini, Simona, Vassilis G Gorgoulis, and Thanos D Halazonetis. 2010. 'Genomic instability—an evolving hallmark of cancer', *Nature Reviews Molecular Cell Biology*, 11: 220-28.

Network, Cancer Genome Atlas Research. 2011. 'Integrated genomic analyses of ovarian carcinoma', *Nature*, 474: 609.

Ng, Joseph CF, Jelmar Quist, Anita Grigoriadis, Michael H Malim, and Franca Fraternali. 2019. 'Pan-cancer transcriptomic analysis dissects immune and proliferative functions of APOBEC3 cytidine deaminases', *Nucleic Acids Research*, 47: 1178-94.

Ní Leathlobhair, Máire, and Richard E Lenski. 2022. 'Population genetics of clonally transmissible cancers', *Nature Ecology & Evolution*, 6: 1077-89.

Nik-Zainal, Serena, Ludmil B Alexandrov, David C Wedge, Peter Van Loo, Christopher D Greenman, Keiran Raine, David Jones, Jonathan Hinton, John Marshall, and Lucy A Stebbings. 2012. 'Mutational processes molding the genomes of 21 breast cancers', *Cell*, 149: 979-93.

Nik-Zainal, Serena, David C. Wedge, Ludmil B. Alexandrov, Mia Petljak, Adam P. Butler, Niccolo Bolli, Helen R. Davies, Stian Knappskog, Sancha Martin, Elli Papaemmanuil, Manasa Ramakrishna, Adam Shlien, Ingrid Simonic, Yali Xue, Chris Tyler-Smith, Peter J. Campbell, and Michael R. Stratton. 2014. 'Association of a germline copy number polymorphism of APOBEC3A and APOBEC3B with burden of putative APOBEC-dependent mutations in breast cancer.', *Nature Genetics*, 46: 487--91.

Niknafs, Noushin, Archana Balan, Christopher Cherry, Karlijn Hummelink, Kim Monkhorst, Xiaoshan M Shao, Zineb Belcaid, Kristen A Marrone, Joseph Murray, and Kellie N Smith. 2023. 'Persistent mutation burden drives sustained anti-tumor immune responses', *Nature Medicine*, 29: 440-49.

Nordling, CO. 1953. 'A new theory on the cancer-inducing mechanism', *British journal of cancer*, 7: 68.

Nowell, Peter C. 1976. 'The clonal evolution of tumor cell populations', *Science*, 194: 23-28.

Nowell, Peter C, and David A Hungerford. 1960. 'Chromosome studies on normal and leukemic human leukocytes', *Journal of the National Cancer Institute*, 25: 85-109.

Oh, Sunwoo, Elodie Bournique, Danae Bowen, Pégah Jalili, Ambrocio Sanchez, Ian Ward, Alexandra Dananberg, Lavanya Manjunath, Genevieve P Tran, and Bert L Semler. 2021. 'Genotoxic stress and viral infection induce transient expression of APOBEC3A and pro-inflammatory genes through two distinct pathways', *Nature Communications*, 12: 4917.

Ohms, Stephen, and Danny Rangasamy. 2014. 'Silencing of LINE-1 retrotransposons contributes to variation in small noncoding RNA expression in human cancer cells', *Oncotarget*, 5: 4103.

Oki, Shinya, Tazro Ohta, Go Shioi, Hideki Hatanaka, Osamu Ogasawara, Yoshihiro Okuda, Hideya Kawaji, Ryo Nakaki, Jun Sese, and Chikara Meno. 2018. 'ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data', *EMBO reports*, 19.

Olson, Margaret E, Reuben S Harris, and Daniel A Harki. 2018. 'APOBEC enzymes as targets for virus and cancer therapy', *Cell chemical biology*, 25: 36-49.

Ong, Chin-Tong, and Victor G Corces. 2014. 'CTCF: an architectural protein bridging genome topology and function', *Nature Reviews Genetics*, 15: 234.

Orecchini, Elisa, Loredana Frassinelli, Silvia Galardi, Silvia Anna Ciafrè, and Alessandro Michienzi. 2018. 'Post-transcriptional regulation of LINE-1 retrotransposition by AID/APOBEC and ADAR deaminases', *Chromosome research*, 26: 45-59.

Papatheodorou, Irene, Nuno A Fonseca, Maria Keays, Y Amy Tang, Elisabet Barrera, Wojciech Bazant, Melissa Burke, Anja Füllgrabe, Alfonso

Muñoz-Pomer Fuentes, and Nancy George. 2017. 'Expression Atlas: gene and protein expression across multiple studies and organisms', *Nucleic Acids Research*, 46: D246-D51.

Park, Peter J. 2009. 'ChIP–seq: advantages and challenges of a maturing technology', *Nature Reviews Genetics*, 10: 669-80.

Parker, Richard C, Harold E Varmus, and J Michael Bishop. 1984. 'Expression of v-src and chicken c-src in rat cells demonstrates qualitative differences between pp60v-src and pp60c-src', *Cell*, 37: 131-39.

Pecori, Riccardo, Salvatore Di Giorgio, J Paulo Lorenzo, and F Nina Papavasiliou. 2022. 'Functions and consequences of AID/APOBEC-mediated DNA and RNA deamination', *Nature Reviews Genetics*, 23: 505-18.

Penzkofer, Tobias, Marten Jäger, Marek Figlerowicz, Richard Badge, Stefan Mundlos, Peter N Robinson, and Tomasz Zemojtel. 2016. 'L1Base 2: more retrotransposition-active LINE-1s, more mammalian genomes', *Nucleic Acids Research*: gkw925.

Pepper, John W, C Scott Findlay, Rees Kassen, Sabrina L Spencer, and Carlo C Maley. 2009. 'Synthesis: cancer research meets evolutionary biology', *Evolutionary applications*, 2: 62-70.

Periyasamy, Manikandan, Anup K Singh, Carolina Gemma, Raed Farzan, Rebecca C Allsopp, Jacqueline A Shaw, Sara Charmsaz, Leonie S Young, Paula Cunnea, and R Charles Coombes. 2021. 'Induction of APOBEC3B expression by chemotherapy drugs is mediated by DNA-PK-directed activation of NF-κB', *Oncogene*, 40: 1077-90.

Periyasamy, Manikandan, Anup K Singh, Carolina Gemma, Christian Kranjec, Raed Farzan, Damien A Leach, Naveenan Navaratnam, Hajnalka

L Pálinkás, Beata G Vértessy, and Tim R Fenton. 2017. 'p53 controls expression of the DNA deaminase APOBEC3B to limit its potential mutagenic activity in cancer cells', *Nucleic Acids Research*, 45: 11056-69.

Perucho, Manuel, Mitchell Goldfarb, Kenji Shimizu, Concepcion Lama, Jorgen Fogh, and Michael Wigler. 1981. 'Human-tumor-derived cell lines contain common and different transforming genes', *Cell*, 27: 467-76.

Petljak, Mia, Ludmil B Alexandrov, Jonathan S Brammeld, Stacey Price, David C Wedge, Sebastian Grossmann, Kevin J Dawson, Young Seok Ju, Francesco Iorio, and Jose MC Tubio. 2019. 'Characterizing mutational signatures in human Cancer cell lines reveals episodic APOBEC mutagenesis', *Cell*, 176: 1282-94. e20.

Petljak, Mia, Alexandra Dananberg, Kevan Chu, Erik N Bergstrom, Josefine Striepen, Patrick von Morgen, Yanyang Chen, Hina Shah, Julian E Sale, and Ludmil B Alexandrov. 2022. 'Mechanisms of APOBEC3 mutagenesis in human cancer cells', *Nature*, 607: 799-807.

Petljak, Mia, and John Maciejowski. 2020. 'Molecular origins of APOBEC-associated mutations in cancer', *DNA Repair*, 94: 102905.

Pettersen, Henrik Sahlin, Anastasia Galashevskaya, Berit Doseth, Mirta ML Sousa, Antonio Sarno, Torkild Visnes, Per Arne Aas, Nina-Beate Liabakk, Geir Slupphaug, and Pål Sætrom. 2015. 'AID expression in B-cell lymphomas causes accumulation of genomic uracil and a distinct AID mutational signature', *DNA Repair*, 25: 60-71.

Pezic, Dubravka, Sergei A Manakov, Ravi Sachidanandam, and Alexei A Aravin. 2014. 'piRNA pathway targets active LINE1 elements to establish the repressive H3K9me3 mark in germ cells', *Genes & development*, 28: 1410-28.

Pizzi, Sara, Sarah Sertic, Simona Orcesi, Cristina Cereda, Marika Bianchi, Andrew P Jackson, Federico Lazzaro, Paolo Plevani, and Marco Muzi-Falconi. 2015. 'Reduction of hRNase H2 activity in Aicardi–Goutieres syndrome cells leads to replication stress and genome instability', *Human molecular genetics*, 24: 649-58.

Pleasance, Erin D, R Keira Cheetham, Philip J Stephens, David J McBride, Sean J Humphray, Chris D Greenman, Ignacio Varela, Meng-Lay Lin, Gonzalo R Ordóñez, and Graham R Bignell. 2010. 'A comprehensive catalogue of somatic mutations from a human cancer genome', *Nature*, 463: 191-96.

Pokatayev, Vladislav, Naushaba Hasin, Hyongi Chon, Susana M Cerritelli, Kiran Sakhuja, Jerrold M Ward, H Douglas Morris, Nan Yan, and Robert J Crouch. 2016. 'RNase H2 catalytic core Aicardi-Goutières syndrome–related mutant invokes cGAS–STING innate immune-sensing pathway in mice', *Journal of Experimental Medicine*, 213: 329-36.

Pomerantz, Jason, Nicole Schreiber-Agus, Nanette J Liégeois, Adam Silverman, Leila Alland, Lynda Chin, Jason Potes, Ken Chen, Irene Orlow, and Han-Woong Lee. 1998. 'The Ink4a tumor suppressor gene product, p19Arf, interacts with MDM2 and neutralizes MDM2's inhibition of p53', *Cell*, 92: 713-23.

Priestley, Peter, Jonathan Baber, Martijn P Lolkema, Neeltje Steeghs, Ewart de Bruijn, Charles Shale, Korneel Duyvesteyn, Susan Haidari, Arne van Hoeck, and Wendy Onstenk. 2019. 'Pan-cancer whole-genome analyses of metastatic solid tumours', *Nature*, 575: 210-16.

Protasova, Maria Sergeevna, Tatiana Vladimirovna Andreeva, and Evgeny Ivanovich Rogaev. 2021. 'Factors regulating the activity of LINE1 retrotransposons', *Genes*, 12: 1562.

Pulciani, Simonetta, Eugenio Santos, Anne V Lauver, Linda K Long, Keith C Robbins, and Mariano Barbacid. 1982. 'Oncogenes in human tumor cell lines: molecular cloning of a transforming gene from human bladder carcinoma cells', *Proceedings of the National Academy of Sciences*, 79: 2845-49.

Quaresma, Manuela, Michel P Coleman, and Bernard Rachet. 2015. '40-year trends in an index of survival for all cancers combined and survival adjusted for age and sex for each cancer in England and Wales, 1971–2011: a population-based study', *The lancet*, 385: 1206-18.

Reddy, E Premkumar, Roberta K Reynolds, Eugenio Santos, and Mariano Barbacid. 1982. 'A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene', *Nature*, 300: 149-52.

Refsland, Eric W., Mark D. Stenglein, Keisuke Shindo, John S. Albin, William L. Brown, and Reuben S. Harris. 2010. 'Quantitative profiling of the full APOBEC3 mRNA repertoire in lymphocytes and tissues: implications for HIV-1 restriction.', *Nucleic Acids Research*, 38: 4274--84.

Rice, Gillian I, Candice Meyzer, Naïm Bouazza, Marie Hully, Nathalie Boddaert, Michaela Semeraro, Leo AH Zeef, Flore Rozenberg, Vincent Bondet, and Darragh Duffy. 2018. 'Reverse-transcriptase inhibitors in the Aicardi–Goutières syndrome', *New England journal of medicine*, 379: 2275-77.

Ries, Lynn AG, D Harkins, M Krapcho, Angela Mariotto, BA Miller, Eric J Feuer, Limin X Clegg, MP Eisner, Marie-Josèphe Horner, and Nadia Howlader. 2006. 'SEER cancer statistics review, 1975-2003'.

Roberts, Steven A., Michael S. Lawrence, Leszek J. Klimczak, Sara A. Grimm, David Fargo, Petar Stojanov, Adam Kiezun, Gregory V. Kryukov, Scott L. Carter, Gordon Saksena, Shawn Harris, Ruchir R. Shah, Michael A. Resnick, Gad Getz, and Dmitry A. Gordenin. 2013. 'An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers.', *Nature Genetics*, 45: 970--6.

Rodić, Nemanja, Reema Sharma, Rajni Sharma, John Zampella, Lixin Dai, Martin S Taylor, Ralph H Hruban, Christine A Iacobuzio-Donahue, Anirban Maitra, and Michael S Torbenson. 2014. 'Long interspersed element-1 protein expression is a hallmark of many human cancers', *The American Journal of Pathology*, 184: 1280-86.

Rodriguez-Martin, Bernardo, Eva G Alvarez, Adrian Baez-Ortega, Jorge Zamora, Fran Supek, Jonas Demeulemeester, Martin Santamarina, Young Seok Ju, Javier Temes, and Daniel Garcia-Souto. 2020. 'Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition', *Nature Genetics*, 52: 306-19.

Rosenthal, Rachel, Elizabeth Larose Cadieux, Roberto Salgado, Maise Al Bakir, David A Moore, Crispin T Hiley, Tom Lund, Miljana Tanić, James L Reading, and Kroopa Joshi. 2019. 'Neoantigen-directed immune escape in lung cancer evolution', *Nature*, 567: 479-85.

Rosenthal, Rachel, Nicholas McGranahan, Javier Herrero, Barry S Taylor, and Charles Swanton. 2016. 'DeconstructSigs: delineating mutational

processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution', *Genome Biology*, 17: 31.

Rosner, Bernard. 2015. *Fundamentals of biostatistics* (Nelson Education).

Rous, Peyton. 1910. 'A transmissible avian neoplasm.(sarcoma of the common fowl.)', *The Journal of Experimental Medicine*, 12: 696-705.

Rowley, Janet D. 1973. 'A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining', *Nature*, 243: 290-93.

Sanger, Frederick, Steven Nicklen, and Alan R Coulson. 1977. 'DNA sequencing with chain-terminating inhibitors', *Proceedings of the National Academy of Sciences*, 74: 5463-67.

Sassaman, Donna M, Beth A Dombroski, John V Moran, Michelle L Kimberland, Thierry P Naas, Ralph J DeBerardinis, Abram Gabriel, Gary D Swergold, and Haig H Kazazian. 1997. 'Many human L1 elements are capable of retrotransposition', *Nature Genetics*, 16: 37.

Sawyer, Sara L, Michael Emerman, and Harmit S Malik. 2004. 'Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G', *PLoS biology*, 2: e275.

Sharma, Reema, Nemanja Rodić, Kathleen H Burns, and Martin S Taylor. 2016. 'Immunodetection of human LINE-1 expression in cultured cells and human tissues.' in, *Transposons and Retrotransposons* (Springer).

Shay, Jerry W, Olivia M Pereira-Smith, and Woodring E Wright. 1991. 'A role for both RB and p53 in the regulation of human cellular senescence', *Experimental cell research*, 196: 33-39.

Shih, Chiaho, LC Padhy, Mark Murray, and Robert A Weinberg. 1981. 'Transforming genes of carcinomas and neuroblastomas introduced into mouse fibroblasts', *Nature*, 290: 261-64.

Sondka, Zbyslaw, Sally Bamford, Charlotte G Cole, Sari A Ward, Ian Dunham, and Simon A Forbes. 2018. 'The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers', *Nature Reviews Cancer*, 18: 696-705.

Sookdeo, Akash, Crystal M Hepp, Marcella A McClure, and Stéphane Boissinot. 2013. 'Revisiting the evolution of mouse LINE-1 in the genomic era', *Mobile DNA*, 4: 1-15.

Speicher, Michael R, Stephen Gwyn Ballard, and David C Ward. 1996. 'Karyotyping human chromosomes by combinatorial multi-fluor FISH', *Nature Genetics*, 12: 368-75.

Stavrou, Spyridon, and Susan R Ross. 2015. 'APOBEC3 proteins in viral immunity', *The Journal of Immunology*, 195: 4565-70.

Steele, Christopher D, Ammal Abbasi, SM Ashiqul Islam, Amy L Bowes, Azhar Khandekar, Kerstin Haase, Shadi Hames-Fathi, Dolapo Ajayi, Annelien Verfaillie, and Pawan Dhami. 2022. 'Signatures of copy number alterations in human cancer', *Nature*, 606: 984-91.

Stetson, Daniel B., Joan S. Ko, Thierry Heidmann, and Ruslan Medzhitov. 2008. 'Trex1 Prevents Cell-Intrinsic Initiation of Autoimmunity', *Cell*, 134: 587--98.

Stratton, Michael R, Peter J Campbell, and P Andrew Futreal. 2009. 'The cancer genome', *Nature*, 458: 719-24.

Subramanian, Aravind, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy,

Todd R Golub, and Eric S Lander. 2005. 'Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles', *Proceedings of the National Academy of Sciences*, 102: 15545-50.

Supernat, Anna, Oskar Valdimar Vidarsson, Vidar M Steen, and Tomasz Stokowy. 2018. 'Comparison of three variant callers for human whole genome sequencing', *Scientific Reports*, 8: 17851.

Suzuki, David T, and Anthony JF Griffiths. 1976. *An introduction to genetic analysis* (WH Freeman and Company.).

Swergold, Gary D. 1990. 'Identification, characterization, and cell specificity of a human LINE-1 promoter', *Molecular and Cellular Biology*.

Taylor, Benjamin Jm, Serena Nik-Zainal, Yee Ling Wu, Lucy A. Stebbings, Keiran Raine, Peter J. Campbell, Cristina Rada, Michael R. Stratton, and Michael S. Neuberger. 2013. 'DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis.', *eLife*, 2: e00534.

Temko, Daniel, Ian P. M. Tomlinson, Simone Severini, Benjamin Schuster-Bckler, and Trevor A. Graham. 2018. 'The effects of mutational processes and selection on driver mutations across cancer types', *Nature Communications*, 9: 1857.

Teng, BaBie, Charles F Burant, and Nicholas O Davidson. 1993. 'Molecular cloning of an apolipoprotein B messenger RNA editing protein', *Science*, 260: 1816-19.

Thawani, Akanksha, Alfredo Jose Florez Ariza, Eva Nogales, and Kathleen Collins. 2023. 'Template and target site recognition by human LINE-1 in retrotransposition', *Nature*: 1-3.

The, ICGC, TCGA Pan-Cancer Analysis of Whole, and Genomes Consortium. 2020. 'Pan-cancer analysis of whole genomes', *Nature*, 578: 82.

Thielen, Beth K, Kevin C Klein, Lorne W Walker, Mary Rieck, Jane H Buckner, Garrett W Tomblingson, and Jaisri R Lingappa. 2007. 'T cells contain an RNase-insensitive inhibitor of APOBEC3G deaminase activity', *PLoS pathogens*, 3: e135.

Thomas, Charles A, Leon Tejwani, Cleber A Trujillo, Priscilla D Negraes, Roberto H Herai, Pinar Mesci, Angela Macia, Yanick J Crow, and Alysson 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: 319-31. e8.

Treangen, Todd J., and Steven L. Salzberg. 2012. 'Repetitive DNA and next-generation sequencing: computational challenges and solutions.', *Nature reviews. Genetics*, 13: 36--46.

Tubio, Jose M. C., Yilong Li, Young Seok Ju, Inigo Martincorena, Susanna L. Cooke, Marta Tojo, Gunes Gundem, Christodoulos P. Pipinikas, Jorge Zamora, Keiran Raine, Andrew Menzies, Pablo Roman-Garcia, Anthony Fullam, Moritz Gerstung, Adam Shlien, Patrick S. Tarpey, Elli Papaemmanuil, and Stian and Knappskog. 2014. 'Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes.', *Science (New York, N.Y.)*, 345: 1251343.

Tyzzer, Ernest Edward. 1916. 'Tumor immunity', *The Journal of Cancer Research*, 1: 125-56.

Uggenti, Carolina, Alice Lepelley, Marine Depp, Andrew P Badrock, Mathieu P Rodero, Marie-Thérèse El-Daher, Gillian I Rice, Somdutta Dhir, Ann P Wheeler, and Ashish Dhir. 2020. 'cGAS-mediated induction of type I interferon due to inborn errors of histone pre-mRNA processing', *Nature Genetics*, 52: 1364-72.

Uriu, Keiya, Yusuke Kosugi, Narumi Suzuki, Jumpei Ito, and Kei Sato. 2021. 'Elucidation of the complicated scenario of primate APOBEC3 gene evolution', *Journal of Virology*, 95: 10.1128/jvi. 00144-21.

van Middendorp, Joost J, Gonzalo M Sanchez, and Alwyn L Burridge. 2010. 'The Edwin Smith papyrus: a clinical reappraisal of the oldest known document on spinal injuries', *European Spine Journal*, 19: 1815-23.

Vassilev, Lyubomir T, Binh T Vu, Bradford Graves, Daisy Carvajal, Frank Podlaski, Zoran Filipovic, Norman Kong, Ursula Kammlott, Christine Lukacs, and Christian Klein. 2004. 'In vivo activation of the p53 pathway by small-molecule antagonists of MDM2', *Science*, 303: 844-48.

Venkitaraman, Ashok R. 2014. 'Cancer suppression by the chromosome custodians, BRCA1 and BRCA2.', *Science (New York, N.Y.)*, 343: 1470--5.

Vieira, Valdimara C., Brandon Leonard, Elizabeth A. White, Gabriel J. Starrett, Nuri A. Temiz, Laurel D. Lorenz, Denis Lee, Marcelo A. Soares, Paul F. Lambert, Peter M. Howley, and Reuben S. Harris. 2014. 'Human Papillomavirus E6 Triggers Upregulation of the Antiviral and Cancer Genomic DNA Deaminase APOBEC3B', *mBio*, 5: e02234--14.

Virchow, Rudolf. 1858. *Die Cellularpathologie in ihrer Begründung auf physiologische und pathologische Gewebelehre: 20 Vorlesungen, gehalten während d. Monate Febr., März u. April 1858 im Patholog. Inst. zu Berlin* (Hirschwald).

Vitre, Benjamin D, and Don W Cleveland. 2012. 'Centrosomes, chromosome instability (CIN) and aneuploidy', *Current opinion in cell biology*, 24: 809-15.

Voet, Donald, Judith G Voet, and Charlotte W Pratt. 2013. *Fundamentals of biochemistry: life at the molecular level*.

von Hansemann, David T. 1890. 'Ueber asymmetrische Zelltheilung in epithel Krebsen und deren biologische Bedeutung', *Virchow's Arch. Path. Anat*, 119: 299.

Vryer, Regan, and Richard Saffery. 2017. 'What's in a name? Context-dependent significance of 'global'methylation measures in human health and disease', *Clinical epigenetics*, 9: 1-4.

Wang, Haidong, Mohsen Naghavi, Christine Allen, Ryan M Barber, Zulfiqar A Bhutta, Austin Carter, Daniel C Casey, Fiona J Charlson, Alan Zian Chen, and Matthew M Coates. 2016. 'Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015', *The lancet*, 388: 1459-544.

Wang, Yichen, Philip S Robinson, Tim HH Coorens, Luiza Moore, Henry Lee-Six, Ayesha Noorani, Mathijs A Sanders, Hyunchul Jung, Riku Katainen, and Robert Heuschkel. 2023. 'APOBEC mutagenesis is a common process in normal human small intestine', *Nature Genetics*, 55: 246-54.

Watson, James D, and Francis HC Crick. 1953. 'Molecular structure of nucleic acids', *Nature*, 171: 737-38.

Weinberg, Robert A. 2007. *The biology of cancer* (Garland Science).

Weinstein, John N, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. 2013. 'The cancer genome atlas pan-cancer analysis project', *Nature Genetics*, 45: 1113-20.

Williams, Ashley B, and Björn Schumacher. 2016. 'p53 in the DNA-damage-repair process', *Cold Spring Harbor perspectives in medicine*, 6.

Wissing, Silke, Mauricio Montano, Jose Luis Garcia-Perez, John V. Moran, and Warner C. Greene. 2011. 'Endogenous APOBEC3B restricts LINE-1 retrotransposition in transformed cells and human embryonic stem cells', *The Journal of Biological Chemistry*, 286: 36427--37.

Wylie, Annika, Amanda E. Jones, Alejandro D'Brot, Wan-Jin Lu, Paula Kurtz, John V. Moran, Dinesh Rakheja, Kenneth S. Chen, Robert E. Hammer, Sarah A. Comerford, James F. Amatruda, and John M. Abrams. 2016. 'p53 genes function to restrain mobile elements.', *Genes \& development*, 30: 64--77.

Yanai, Hideyuki, Hideo Negishi, and Tadatsugu Taniguchi. 2012. 'The IRF family of transcription factors: Inception, impact and implications in oncogenesis', *Oncoimmunology*, 1: 1376-86.

Yang, Bin, Keyang Chen, Chune Zhang, Sophia Huang, and Hui Zhang. 2007. 'Virion-associated uracil DNA glycosylase-2 and apurinic/apyrimidinic endonuclease are involved in the degradation of APOBEC3G-edited nascent HIV-1 DNA', *Journal of Biological Chemistry*, 282: 11667-75.

Yang, Yun-Gui, Tomas Lindahl, and Deborah E Barnes. 2007. 'Trex1 exonuclease degrades ssDNA to prevent chronic checkpoint activation and autoimmune disease', *Cell*, 131: 873-86.

Yu, Xianghui, Yunkai Yu, Bindong Liu, Kun Luo, Wei Kong, Panyong Mao, and Xiao-Fang Yu. 2003. 'Induction of APOBEC3G ubiquitination and degradation by an HIV-1 Vif-Cul5-SCF complex', *Science*, 302: 1056-60.

Zhang, Ao, Beihua Dong, Aurelien J Doucet, John B Moldovan, John V Moran, and Robert H Silverman. 2014. 'RNase L restricts the mobility of engineered retrotransposons in cultured human cells', *Nucleic Acids Research*, 42: 3803-20.

Zhao, Ke, Juan Du, Yanfeng Peng, Peng Li, Shaohua Wang, Yu Wang, Jingwei Hou, Jian Kang, Wenwen Zheng, and Shucheng Hua. 2018. 'LINE1 contributes to autoimmunity through both RIG-I-and MDA5-mediated RNA sensing pathways', *Journal of autoimmunity*, 90: 105-15.

Zhao, Yingdong, Ming-Chung Li, Mariam M Konaté, Li Chen, Biswajit Das, Chris Karlovich, P Mickey Williams, Yvonne A Evrard, James H Doroshow, and Lisa M McShane. 2021. 'TPM, FPKM, or normalized counts? A comparative study of quantification measures for the analysis of RNA-seq data from the NCI patient-derived models repository', *Journal of translational medicine*, 19: 1-15.

Appendices

Appendix I: Publications associated with this thesis

Saif S Ahmad, Karim Ahmed and Ashok R Venkitaraman. 2018. 'Science in Focus: Genomic instability and its implications for clinical cancer care', *Clinical Oncology*, 30: 751-55.

Shawn LW Tan, Saakshi Chadha, Yansheng Liu, Evelina Gabasova, David Perera, Karim Ahmed, Stephanie Constantinou, Xavier Renaudin, MiYoung Lee, Ruedi Aebersold and Ashok R Venkitaraman. 2017. 'A class of environmental and endogenous toxins induces BRCA2 haploinsufficiency and genome instability', *Cell*, 169: 1105-18.

Appendix II: Source data for Figure 3.1.2 and Figure 3.1.4

Data for Figure 3.1.2

HCT116 p53$^{wt/wt}$

Gene	Ct					
Actin	18.1869200	18.1861670	18.2038600	18.2162410	18.1765600	19.1906330
APOBEC3B	26.6181210	26.6370150	26.6245030	26.6083390	26.6049760	27.7152310
ORF1	21.1063140	21.1169680	21.1467510	21.1807040	21.1005630	22.4143990
ORF2	20.0205200	20.1320140	20.2348700	20.2413800	20.1797010	21.9087020

HCT116 p53$^{-/-}$

Gene	Ct					
Actin	19.6457920	19.6334560	19.6435290	19.6154890	19.6394770	20.7322300
APOBEC3B	26.5866430	26.5924870	26.6278490	26.6262300	26.5868570	27.8098110
ORF1	21.3268060	21.3677580	21.2921810	21.3419810	21.3143530	22.7420580
ORF2	20.4913540	20.3979280	20.3844020	20.4353870	20.4230400	21.8979140

Data for Figure 3.1.4

Vehicle-treated

Gene	Ct					
Actin	20.4789238	20.1745796	20.3936863	20.7767448	19.3506718	19.0721626
APOBEC3B	29.3181858	28.9162579	29.2849827	29.3563805	27.8790112	27.9372330
ORF1	19.9865799	19.9351368	20.6324806	20.7063961	19.4255810	19.7062817
ORF2	18.1295357	17.1355591	18.5711403	18.5403938	18.5849495	18.7065392

Nutlin-treated

Gene	Ct					
Actin	20.3917580	20.1976242	20.9781837	20.9432907	31.4161415	30.8860035
APOBEC3B	29.1902618	29.6727200	30.1705303	30.1122952	37.0444832	36.4303093
ORF1	21.3118210	20.9966164	21.2331409	21.0229053	19.5149193	19.0666790
ORF2	19.4672661	19.2933865	19.4826813	19.3169899	16.8296909	16.9100800