REVIEW ARTICLE

Review of the mutational role of deaminases and the generation of a cognate molecular model to explain cancer mutation spectra

Author
Robyn A Lindley[1,2]

[1]Department of Clinical Pathology
The Victorian Comprehensive Cancer Centre
Faculty of Medicine, Dentistry & Health Sciences
University of Melbourne,
305 Gratton Street,
Melbourne, VIC 3000, AUSTRALIA
Email: robyn.lindley@unimelb.edu.au

[2]GMDx Genomics Ltd,
Level 3 162 Collins Street,
Melbourne VIC3000, AUSTRALIA
Email: robyn.lindley@gmdxgroup.com

Correspondence: Robyn A Lindley, Department of Clinical Pathology, Faculty of Medicine, Dentistry & Health Sciences, University of Melbourne, 305 Gratton Street, Melbourne VIC 3000 AUSTRALIA
Mobile: +61 (0) 414209132
Email: robyn.lindley@unimelb.edu.au

Abstract

Recent developments in somatic mutation analyses have led to the discovery of codon-context targeted somatic mutation (TSM) signatures in cancer genomes: it is now known that deaminase mutation target sites are far more specific than previously thought. As this research provides novel insights into the deaminase origin of most of the somatic point mutations arising in cancer, a clear understanding of the mechanisms and processes involved will be valuable for molecular scientists as well as oncologists and cancer specialists in the clinic. This review will describe the basic research into the mechanism of antigen-driven somatic hypermutation of immunoglobulin variable genes (Ig SHM) that lead to the discovery of TSM signatures, and it will show that an Ig SHM-like signature is ubiquitous in the cancer exome. Most importantly, the data discussed in this review show that Ig SHM-like cancer-associated signatures are highly targeted to cytosine (C) and adenosine (A) nucleotides in a characteristic codon-context fashion. This review also provides an evidence-based model explaining how deaminases that cause mutations in cytosine and adenosine can gain access to their respective target motifs in genomic DNA (C-sites) and RNA (A-sites). It also highlights the clinical importance of understanding the molecular processes underpinning deaminase targeting for the development of new genomic diagnostics and drug discovery for pre-cancerous and clinically diagnosed cancer patients.

Keywords: deaminase, cancer, mutation signatures, Targeted Somatic Mutation, Somatic Hypermutation.

Content

Abbreviations

Aag alkyladenine DNA glycosylase
ADAR adenosine deaminase that act on RNA
AID activation induced cytidine deaminase, an APOBEC family member (most similar in DNA sequence to APOBEC1), initiating via dC-to-dU lesions in ssDNA of class switch recombination (CSR) and somatic hypermutation (SHM) processes at somatically rearranged Ig V(D)J gene loci, and known to activate cytidine mutagenic deamination during transcription in other somatic tissues, particularly in cancer
AKAP A-kinase anchoring protein family
Alu (Arthrobacter luteus) elements are short stretches of WA-rich (~ 300 bp) repetitive retro-elements in the genome dispersed over evolutionary time
AP (apurinic/apyrimidinic site), also known as an abasic site
apoB apolipoprotein B
APOBEC family generic abbreviation for the deoxyribonucleic acid, or dC-to-dU, deaminase family (APOBEC3 A, B, C, D, F, G, H) similar in DNA sequence to the "apolipoprotein B RNA editor" APOBEC1, and known to activate mutagenic cytidine deamination during transcription in somatic tissues, particularly in cancer
AP an abasic, or apurinic or apyrimidinic site
APE AP endonuclease
A-to-I adenosine-to-inosine RNA editing
BER base excision repair
BLGG lower grade glioma
CDS protein coding regions
CPAS cancer progression associated signatures
CSR immunoglobulin class switch recombination
DBD deaminase binding domains of ADAR and AID/APOBEC enzymes
Inf-DBD inferred DBD
Deaminase zinc-catalytic domain in ADAR and AID/APOBEC enzymes
dA deoxyadenosine
dl deoxyinsosine
dMMR deficient MMR
DSB double strand DNA break
dsDNA double stranded DNA
dsRNA double stranded RNA
EVs extracellular vesicles possibly extruded by M1 and M2 polarized macrophages
FPKM fragments per kilobase million (metric**)**
GC germinal centre
HBV hepatitis B virus
HCV hepatitis C virus
Ig any of a class of proteins present in the serum and cells of the immune system, which function as antibodies
Ig SHM-like response, strand-biased somatic mutation patterns similar to that observed in Ig SHM that occurs in non-Ig genes, and sometimes referred to as 'off target' SHM
ISG interferon stimulated gene path
MAR matrix attachment region
MC mutated codon, referring to nucleotides in MC1, MC2, MC3 frame-reading sites respectively as the first, second and third position in a mutated codon read in the 5' to 3'direction
miRNA micro RNA
MMR mismatch repair

mRNA messenger RNA
Motif 2 to 6 nucleotide (N) sequence defining specificity of deaminase mutation target sites
MSH2-MSH6 MutSalpha heterodimer recognising mispaired bases in DNA duplex
NMD the nonsense-mediated messenger RNA decay pathway
NTS the non-transcribed, or “Top”, strand
NGS next generation sequencing
OMIM the online Mendelian inheritance in man database
pre-mRNA precursor mRNA, is the first transcript from a gene
Pol-eta DNA polymerase-eta
R adenosine (A) or guanine (G), purines
RADAR a rigorously annotated database of A-to-I RNA editing events
R-dsRNA right-handed dsRNA
RNA:DNA a hybrid double stranded substrate of RNA and the complementary DNA strand
RNA Pol II is a multiprotein complex that transcribes DNA into precursors of messenger mRNA and most small nuclear RNAs (including snRNAs and miRNAs)
RT reverse transcription
RT model, reverse transcription linked involving a DNA-mRNA-cDNA information flow
RT-PCR reverse transcription polymerase chain reaction
RT Pol-eta reverse transcriptase activity displayed by pol-etaS, strong base pair involving cytosine (C) or guanine (G)
S strong nucleotides (G or C)
SHM somatic hypermutation
T thymine
SNP single nucleotide polymorphism
snRNA small nuclear RNA
ssDNA single stranded DNA
ssRNA single stranded RNA
TAM tumour associated macrophage
TCGA The National Cancer Genome Atlas (National Cancer Institute, USA)
TDG a BER enzyme thymine DNA glycosylase
TS the transcribed, or “Bottom” strand, in context of a transcription bubble
TSM targeted somatic mutations: the process of deaminases targeting actively transcribed genes that results in a dominant type of mutation caused by a DBD or Inf-DBD targeting nucleotide sites at a particular mutated codon (MC) position 1-3
TSRT target site reverse transcription
U uracil
UNG uracil DNA glycosylase involved in BER at dU sites in DNA resulting in either an abasic site (AP) or APE-mediated ssDNA nicks (above)
UTR untranslated regions in the upstream (5’) and downstream (3’) regulatory regions of protein coding genes
V(D)J generic symbol for a rearranged immunoglobulin (or T cell receptor, TCR) variable region gene in the adaptive immune system
W weak base (A or U/T)
WES whole exome sequencing
X bases C or A
Y pyrimidines bases T/U or C
ZDD zinc deaminase domain
Z-DNA the DNA double helix that has a left-handed, rather than the usual right-handed twist and the sugar phosphate backbone following a zigzag course

1. Introduction

Whatever the source of carbon-based life on earth, we know that most DNA/RNA life forms carry a cargo that includes genes encoding deaminases. From yeast to man, mutagenic deaminases have now been found in the genetic material of most animal species. Deaminases are activated in all cells, mainly by invading pathogens, viruses, bacteria and fungi and they are now known to play a key role in health and diseases such as cancer.[1] In recent years, our growing understanding of the targeted mutational activity of deaminases has therefore resulted in a paradigm shift away from the idea that mutations arise randomly. While some mutations are generated directly by external physical sources such as ionizing radiation and hazardous chemicals, the deaminases are truly endogenous.

In humans, there are around 14 different deaminase proteins that change the structure of DNA/RNA by altering a single nucleotide. New mutations may arise in somatic tissue as a consequence of deaminases targeting genes for deamination if the subsequent mutational change remains uncorrected by DNA or RNA repair mechanisms. The result is a *de novo* somatic mutation in the DNA of the newly translated cell. For example, the result may be the mutation of an ‘A’ (adenosine) to a ‘G’ (guanosine) in a gene.

There are two types of deaminase families: those targeting cytidines (‘Cs’) that result in mutations of Cs, and those targeting adenosines (‘As’). The resulting two key endogenous deamination events are C-to-U (U-uracil) and A-to-I (I-inosine) These are the core biochemical transformations at the centre of all cancers.[2-4] The primary focus of this paper is to review the endogenous molecular origins of C-to-U and A-to-I base modifications that occur at C and A targets within polynucleotide strands of DNA and RNA, and their role in oncogenesis.

2. Cytidine deaminases

The human cytidine deaminases include activation induced deaminase (AID), and apolipoprotein B messenger RNA editing enzyme, catalytic polypeptide-like proteins (APOBECs 1, 3A, 3B, 3C, 3D, 3F, 3G, 3H). These preferentially target single stranded DNA (ssDNA) which usually only occurs in the context of an ‘open ’transcription bubble. That is, the activity of cytidine deaminases is transcription linked.

The human AID/APOBEC proteins form a homologous family of deaminases with similar protein structures. The relative alignments of the human AID/APOBEC sequence structures revealing exon junction similarity is shown in Figure 1. While the number of amino acids forming each protein ranges from 198aa for AID to 386aa for APOBEC3D, the different colours show the sequence similarity. The distributions of their quantitative normal mRNA subcellular expression patterns are also summarised in the righthand column of Figure 1.

The predominant subcellular location is in the cytosol,[5] suggesting that access to immunoglobulin (Ig) somatic hypermutation (SHM) transcription sites in the nucleus via a *regulated* portal (possibly involving the matrix attachment region - MAR?) through the nuclear membrane. Thus, from a clinical viewpoint, it matters more to understand and analyse genomic DNA mutation signatures in both pre-cancer or full-blown cancer clones, and then simply assume molecular access to the genomic DNA as a given.

Biochemical deaminase modification of free nucleoside precursor pools is common in purine and pyrimidine biosynthesis. The AID/APOBEC mediated C-to-U deaminations primarily occur in single stranded regions of DNA or RNA. A zinc-coordinated glutamic acid in the active site of the deamination domain guides the removal of the amino group (released as ammonia), and results in a uracil.[6] The AID/APOBEC family of cytidine deaminases then catalyse the C-to-U hydrolytic deaminations at C-target motifs.

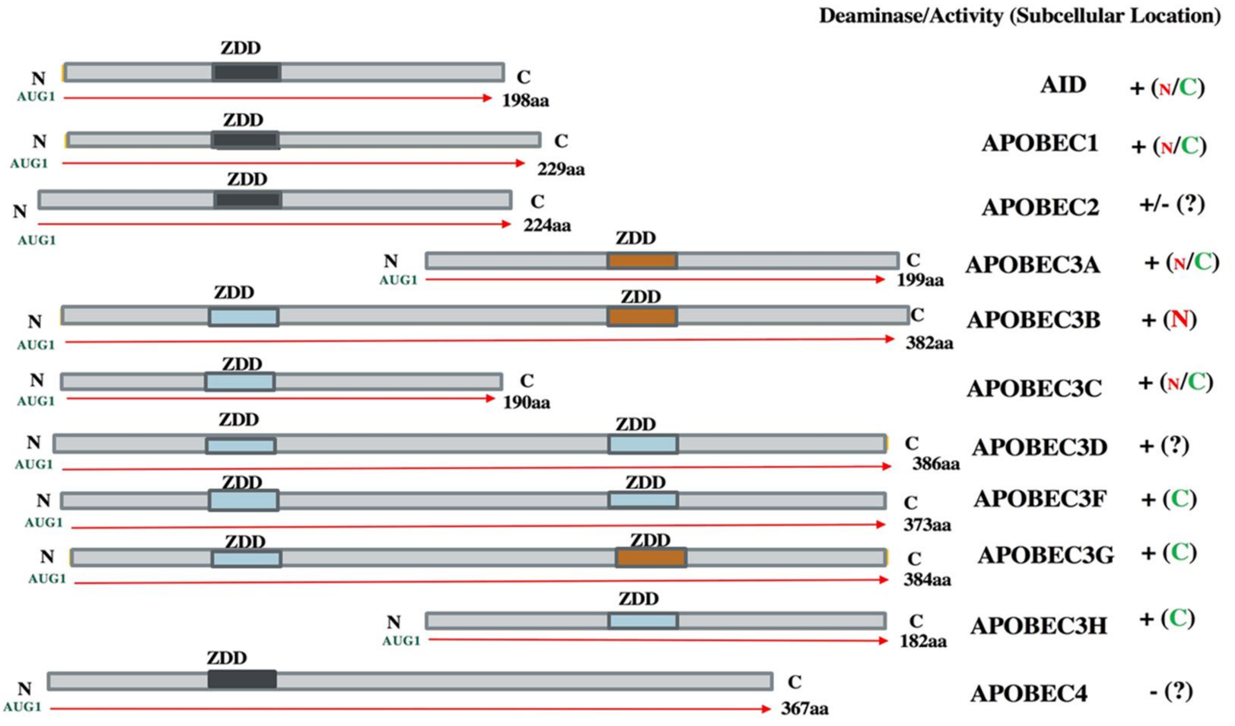


Figure 1. The protein domain structures of the human APOBEC family showing variations in zinc deaminase domains (ZDDs), subcellular localization and deaminase activity. Adopted and assembled from Smith et al (2012), Conticello (2008), Burns et al (2015) and Salter et al (2016).[7-10] The ZDD motifs are shown. The different colours are used to indicate sequence similarity. The number of amino acids is shown at the terminus of each protein. Note that the ZDD for APOBEC4 is significantly divergent from the consensus ZDD. The relative subcellular distribution is on the left with N for nuclear localization, and C for cytoplasmic location. The symbol '?' indicates unknown localization. Known deaminase activity is indicated by + or −, and for trace or uncertain activity +/− for APOBEC2. APOBEC1 and AID are related duplicates on chromosome 12 (at band 12p13.31) and the seven APOBEC3 members via a duplicative tandem expansion (unequal crossing over) on chromosome 22 spread over 150 kb-200 kb (at band 22q13.1), APOBEC2 on chromosome 6 (at band 6p21.1), and APOBEC4 on chromosome 1 (at band 1q25.3).

2.1 Cytidine deaminases target motifs

The cytidine deaminases are known to preferentially target specific motifs that are normally defined by a short sequence consisting of 2-4 nucleotides (nts). That is, deamination is targeted to different C sites in the polynucleotide chain flanked by characteristic 5 prime (5') and 3 prime (3') motif bases. Some previously identified C-target motifs are shown in Table 1.

Table 1. Studies identifying principle motif specificity of the main C-to-U DNA deaminases.

C-to-U deaminase[a] **(eaminase binding domains)** [b]	**ssDNA** **Motif**	**References**
AID [c] (1xDBD, adaptive immunity) [b]	WRC(GYW)	11,12
APOBEC1 [c] (1xDBD)	TCA(TGA)	Confirmed by many studies. Potent DNA deaminase (mutator) in bacterial assay systems.[12]
APOBEC Signature	TCW(WGA)	Mutations often observed as “kataegis” or a “storm” of mutations in cancer clustered C- to-U deaminations.[13]
APOBEC3A [c] (innate immunity and RNA editing)	YTCA(TGAR)	14-16
APOBEC3B (innate immunity)	RTCA(TGAY)	14-15,17
	RTCA(TGAY)	18-19
	RTCG(CGAY)	20
APOBEC3C (innate immunity)	TC=CC > GC > AC	Retroviral restriction. [21]
APOBEC3D (innate immunity)	TCG (or TCT)	Retroviral restriction.[22]
APOBEC3F (innate immunity)	TC(GA)	Retroviral restriction.[23-28]
APOBEC3G (innate immunity)	CC (or TC) (GG, GA)	Retroviral restriction.[12,22-27,29]
APOBEC3H (innate immunity)	TC(GA)	Retroviral restriction.[14,22,28-30]

Table 1 legend. a. Phylogenetics, chromosome location, tissue expression, cellular localization etc. are reviewed in Conticello (2008).[8] b. For further information on AID/APOBEC number and type of zinc coordinating deaminase binding domains (DBDs), and their functional class: RNA editor (e.g. lipid metabolism), innate immunity (retroviral and retroelement restriction), adaptive immunity (immunoglobulin somatic hypermutation and Ig class switch recombination).[6] c. These deaminases are also C-to-U editors of single stranded RNA (ssRNA).[31-35] APOBEC1 has demonstrative site-specific and promiscuous C-to-U RNA editing at 5’ UC 3’ and 5’ AC 3’ motifs.[31,32,35] APOBEC3A has clear C-to-U editing in ssRNA (Cs in unpaired loops) substrates at 5’ UC 3’ motifs.[33,34]

2.2 Differential tissue expression of cytidine deaminases

Generally, the hydrolytic deamination of C residues in DNA occurs at a low rate of around 200 times per mammalian cell per day.[36] The rate of hydrolytic C-to-U reactions driven by the action of endogenous cytidine deaminases of the APOBEC family of DNA and RNA editing enzymes greatly accelerates when acting on unprotected (non-based paired) ssDNA or ssRNA substrates. Deamination by most cytidine deaminases is also tissue or tissue group specific. The differential quantitative mRNA tissue expression levels for AID and APOBECs 1, 3A, 3B, 3C, 3D, 3E, 3F, 3G and 3H in normal healthy tissues is shown in Table 2.

Table 2. Quantitative messenger RNA (mRNA) tissue expression of AID/APOBEC in normal healthy tissues.

TISSUES	AID	A1	A3A	A3B	A3C	A3D	A3F	A3G	A3H
PBMC - Blood[a]	-	-	++++	+/-	+	+	-	+	+
Adipose	-	-	++	+	+	+	+	+	+
Bladder	-	-	+/-	-	+	+	+	+	+
Brain	-	-	-	-	-	-	-	-	-
Cervix	-	-	+	+	+	+	+	+	+
Colon	+/-	+	-	+	+/-	+/-	+	-	+
Esophagus	-	-	+	-	+	-	-	+/-	+
Heart	-	-	+	+	+	+	+	+/-	-
Kidney	-	+	-	+/-	-	-	-	-	-
Liver	-	-	+/-	-	-	-	+	-	-
Lung	-	-	++++	+	+	+	+	+	++
Ovary	-	-	-	-	+	+	++	+	-
Placenta	-	-	+	+	-	-	-	-	+
Prostate	-	-	-	-	-	-	+	+	+
Skeletal Muscle	-	-	-	-	-	-	-	-	-
Small Intestine	+/-	+	-	+	-	-	-	-	-
Spleen	+/-	-	+++	+	+	+	+	++	+
Testes[b]	-	-	-	-	-	-	-	-	-
Thymus	-	-	+/-	-	+	+	+	+	+
Thyroid	-	-	-	-	-	-	+/-	-	-
Trachea	-	-	+	+	-	+/-	+/-	-	-

Table 2 legend. a. For a more detailed breakdown within white cell subsets, lymphocytes, monocytes etc see Koning et al (2009).[37] b. For information on the trace expression for A3C, A3F, A3G see Koning et al (2009).[37] For quantitative reverse transcription polymerase chain reaction (RT-PCR) scaling below refer to Refsland et al (2010) and Burns et al (2013a) where:[19,38]

- zero undetectable mRNA expression = expression level scale 0 to <1.0
+/- trace detectable = expression level scale 1.0
+ = expression level scale >1.0 to 4.0
++ = expression level scale >4.0 to 8.0
+++ = expression level scale 8.0 to 16.0
++++ = expression level scale > 16.0

Note that APOBEC2 (not shown in Table 2)), is expressed in T lymphoblastoid (as CEM cells, a line of lymphoblastic cells originally derived from a child with acute lymphoblastic leukemia); and heart (+) and skeletal muscle (+). APOBEC4 (also not shown in Table 2), shows trace expression in trachea tissue. For other cell line data see Burns et al (2013a) and Koning et al (2009).[19,37] APOBEC1 mediates physiological gene specific C-to-U RNA editing of a single-strand RNA substrate on the nuclear transcript which encodes the intestinal expressed apolipoprotein B (apoB) changing a glutamine CAA codon to a UAA stop codon, generating a truncated protein termed apoB48 thus defining distinct pathways for intestinal and liver lipid transport in mammals.[39] However, C-to-U RNA editing in the wider human transcriptome is an extremely rare event which under normal physiological conditions reveals only a few sites.[40] Note also that APOBEC1 is not expressed in brain, kidney, liver, lung, heart, muscle, and that APOBEC3A is an active C-to-U RNA editor in human monocytes and macrophages.[33] The data also highlights the lack of significant constitutive expression of AID/APOBEC genes in some tissues, particularly in the brain and testes.[37]

Most of the cytidine deaminases included in Table 2 are expressed in lung and spleen tissue. Both are constantly bombarded by external pathogens, and often require a high level of deaminase activity to fight infection as a key part of the innate immune response. In comparison, there is a lack of significant constitutive expression of AID/APOBEC genes, in the brain and testes.[37] These tissues do not have a strong inflammatory immune response when challenged, and they lack lymphatic drainage. Additionally, APOBEC2 (not shown) has been found to have significant expression in T lymphoblastoid (cells and murine), and heart (+) and skeletal muscle (+). There is also trace expression of APOBEC4 found in tracheal tissue. For other cell line data refer to Burns et al (2013a) and Koning et al (2009).[19,37]

Thus, each deaminase targets different motif(s) as shown in Table 1, and as there is a differential constitutive expression of the deaminases across different cell lines and tissues as shown in Table 2, it is expected that the observed mutation patterns will vary across the different tissue groups. This results in divergent cytidine mutation targeting patterns across different cancer types.

3. Adenosine deaminases

The human adenosine deaminases (ADARs) include ADARs 1,2,3 and 4. Only ADARs 1 and 2 are known to be mutagenic. Like the cytidine deaminases, the mutagenic activity of ADARs 1 and 2 is transcription linked, particularly ADAR1. ADAR deaminases (ADAR1, ADAR2), predominantly target W<u>A</u> sites (W = U or A) in double stranded RNA (dsRNA) which occurs when RNA snaps back on itself to form complex double strands and loop structures that are often referred to as ‘hairpins’. The biochemical deamination of adenosine (A) to produce inosine (I) requires a zinc-coordinated glutamic acid molecule to target the active site of the deamination domain guides, and the subsequent removal of an amino group (released as ammonia). The main protein isoforms of the ADAR family of A-to-I editors and their nuclear versus cytosolic subcellular locations are shown in Figure 2.

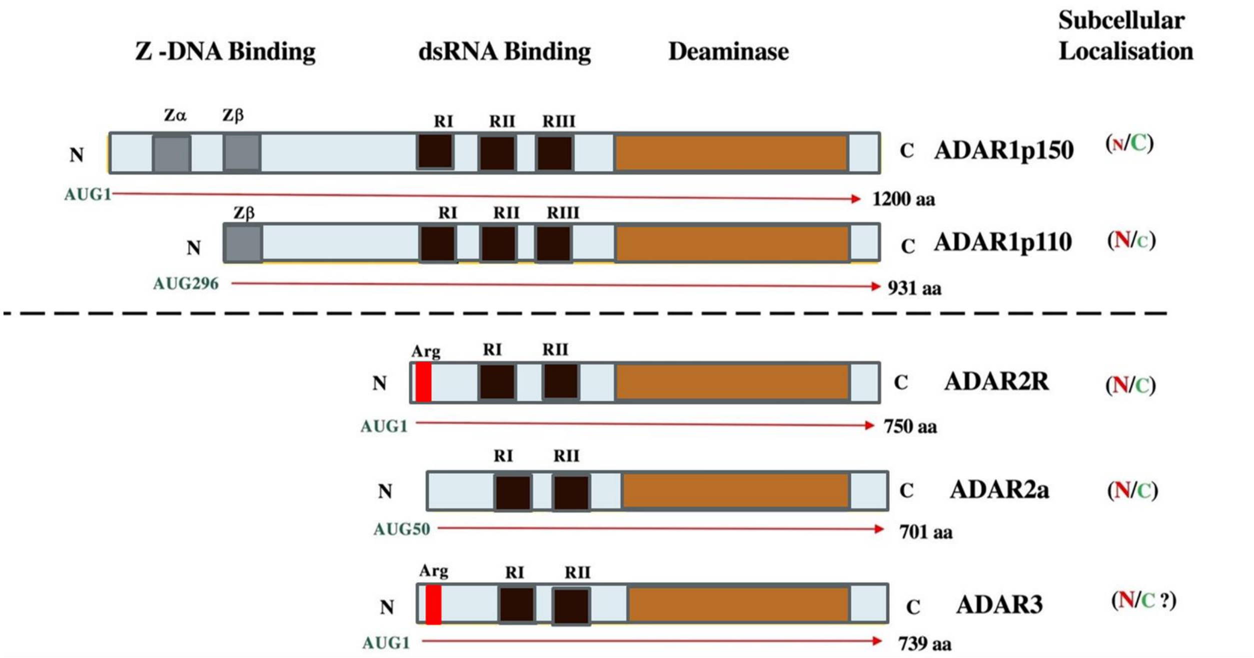


Figure 2. The main protein isoforms of the adenosine deaminase (ADAR) family of A-to-I editors and their nuclear versus cytosolic subcellular location. Assembled from data in Nishikura (2010), Samuel (2011) and Slotkin and Nishikura (2013).[41-43] The dotted line separates ADAR1 from ADAR2 related proteins. The alternatively spliced transcript lengths are shown, as well as the number of amino acids in the protein beginning in the AUG start codon. The different Z-DNA binding domains for ADAR1 are shown, as are the different dsRNA binding domains. Arginine rich domains are shown in red.

In Figure 2, the different Z-DNA binding domains for ADAR1 are shown, as are the different dsRNA binding domains. Note the conservation of each deaminase domain sequence. Motif targeting specificity is likely to be influenced by both the Z-DNA and the right-handed dsRNA (R-dsRNA) binding domains. There are many known ADAR1 polymorphisms associated with disease (not shown), and that ADAR1p150 consists of 1200aa. A catalogue of single nucleotide polymorphisms (SNPs) and disease associations for ADAR1 is provided in Slotkin and Nishikura (2013).[43]

3.1 Differential tissue expression of ADAR family isoforms

Despite molecular similarities, ADAR enzymes are expressed differentially in a range of tissues or tissue groups. Table 3 provides a summary of the qualitative RNA-seq expression of ADAR isoforms for seven normal tissue types.

Table 3. A summary showing the relative quantitative RNA-seq expression of the different ADAR isoforms for seven normal tissue types.

	ADAR1 Protein Isoforms			ADAR2 Protein Isoforms				ADAR 3
	p110-931 aa	p150-1200 aa	p150-1200 aa	701 aa	674 aa	741 aa	714 aa	729 aa
Brain	+++++	+	+	+/-	+/-	+	++	+++
Lung	+++	++++	++	+/-	+	+	++	-
Kidney	+++	+++	+++	-	+/-	+/-	+/-	-
Liver	+++	+	++	-	-	+/-	+/-	-
Heart	+++	+	+	-	+/-	+	+/-	-
Muscle	+	+	+/-	-	+/-	-	+/-	-
Testes	+	+	+	-	-	-	-	+

Table 3 legend. The ADAR2 data is from Agranat et al (2010) data, and the scoring scale is internally consistent and comparable to Picardi et al's (2015) RNA-seq FPKM (fragments per kilobase million) metric.[44,45] Scores for human testes were scaled relative to other tissues, such as brain and is based on data from the GTEx web portal (http://www.gtexportal.org/home/).[46] The other comparisons are a more realistic relative quantitative comparison.[45]

ADAR1p110 and ADAR1p150 are both expressed ubiquitously in the main tissue groups. ADAR2 is constitutively expressed in target tissues or tissue groups. Whilst ADAR1 and 2 are both expressed in many tissues in the body, their main molecular focus of physiological expression targets the A-rich sites of inverted Arthrobacter luteus (Alu) restriction endonuclease repeats located in the intronic precursor mRNA (pre-mRNA) of neuronally expressed synaptic genes in the Brain.[42,43,47] ADAR 3 is found to be expressed only in the brain and testes, and its regulatory role(s) are unknown. Also see Table 4, Paz-Yaacov et al (2010) and Picardi et al (2015). [45,48]

ADAR1p150 (gene at band 1q21.3) like all of the deaminases, is interferon induced as part of the interferon stimulated gene (ISG) path.[49,50] Thus, it plays an important role in an innate immune response to viruses and other pathogens.[42] It is detected in cytoplasm and in the nucleus. ADAR1p150 has two Z-DNA binding domains for binding pre-mRNA by RNA Polymerase II, three dsRNA binding domains, and a conserved deaminase domain.[51,52] While ADAR1p110 (gene at band 1q21.3) is also ubiquitously expressed in many tissues, it is almost exclusively located in the nucleus. It has one Z-DNA binding domain, three dsRNA binding domains, and a conserved deaminase domain.

ADAR2 (gene at band 21q22.3) is also expressed in many tissues and is found in variable length isoforms.[45] It has two dsRNA binding domains, and a conserved deaminase domain. Its nuclear import is controlled, and like ADAR1p110 it may accumulate in the nucleus. ADAR2 also undergoes alternative splicing that results in a diverse range of isoforms.[42] Consequently, both ADARs 1 and 2 are known to display differences in specificity of dsRNA substrate interactions.[53]

ADAR3 (gene at band 10q15.3) is constitutively only found in the brain. Its conservation area and deaminase domains are similar to ADAR2 deaminase domains, and the arginine (R) rich domain allows binding to ssDNA. *In vitro* it is found to block A-to-I editing, yet its possible roles in transcription regulation and control of nuclear import functions are still not fully understood.[54] Recent work on glioblastomas indeed suggests such a regulatory role.[55] It has also been found that the amount of regulatory RNA encoded in the genome and the extent of RNA editing by the post-transcriptional deamination of adenosine to inosine (A-to-I) may be an important factor in the cognitive evolution of animals. Mladenova et al (2018) have shown that mice lacking exon 3 of ADAR3 (which encodes two double stranded RNA binding domains) have increased levels of anxiety and deficits in hippocampus-dependent short- and long-term memory formation.[56] Collectively, these results suggest that ADAR3 contributes to cognitive function in mammals.

Thus, generally, the deamination of adenosine (A) to inosine (I) is a widespread co- and post-transcriptional mechanism mediated by ADAR enzymes acting predominantly on dsRNA, and thus greatly expanding the nucleotide diversity of RNA sequences. More than 90% of the resulting single nucleotide variants are targeted to Alu repeat retro-elements. The recent Inosinome Atlas documents 3 million A-to-I events in the normal human transcriptome across major organs (see Table 4 in Picardi et al 2015).[45]

It is also important to note that so-called 'spontaneous hydrolytic events' such as deoxyadenosine (dA) in DNA polynucleotide strands resulting in deoxyinsosine (dI) in DNA are very rare, and that these may be considered to be potentially mutagenic events. Such events are estimated to occur at a rate of 4-6 times per mammalian cell per day, which is about 200 times less frequent than the estimated number of spontaneous hydrolytic deoxycytidine (dC-to-dU) deamination events in genomic DNA.[36,57] In part, this is because the DNA base excision repair (BER) machinery via alkyladenine DNA glycosylase (Aag) efficiently removes dI from DNA. This is discussed further in Section 7. Such lesions are potentially mutagenic because dI (and I) form a more stable base pair with C rather than T (or U) bases, thus leading to A-to-G mutations in replicated unrepaired DNA.[57]

3.2 Direct A-to-I editing of DNA

The studies summarised to this point have focused on RNA as the deaminated substrate, however recent evidence suggests DNA is also a target for ADAR deamination.[58] It is now well known that RNA:DNA structures are found in abundance in many organisms, and that RNA:DNA hybridisation is known to be a crucial step in Ig class switch recombination in activated B cells.[59] In a landmark study, Zheng et al (2017) used RNA:DNA hybrid substrates *in vitro* to show that both WA-sites in the RNA and DNA moieties of the heteroduplexes are edited at lower, yet reasonable efficiency compared to A-to-I deamination in dsRNA substrates.[58] Under the assay conditions used, A-to-I RNA editing by ADAR2 of 24mer oligonucleotide dsRNA substrates reaches 100% editing in about 10 min, at a time when editing of RNA:DNA hybrids has reached approximately 5% editing. However, after 2 hours, RNA:DNA hybrids are RNA edited about 55%, and DNA edited about 35%, with no detectable direct DNA A-to-I editing of dsDNA duplex substrates. This work, demonstrating that ADARs can directly edit DNA, provides new insight into the molecular processes involved in Ig SHM, and it contributes to our understanding of how somatic mutation profiles of cancer genomes arise.

The direct functional consequence of ADAR-mediated DNA deamination is *de novo* mutations of adenosine that occur in the absence of subsequent base excision repair (BER) via alkyladenine DNA glycosylase (Aag) of the transcribed strand (TS), and/or the presence of faulty mismatch repair (MMR) via MutSalpha heterodimer recognising mispaired bases in DNA duplex (MSH2-MSH6) heterodimers (see Figure 3). Such events will predominantly result in T-to-C mutations.[60] Thus, direct A-to-I DNA editing on the TS can result in T-to-C mutations when read in the normal 5' to 3' polarity of transcription or 5' to 3' on the sequence of the non-transcribed strand (NTS). When detected on the NTS, the T-to-C transitions can now logically be considered to provide presumptive evidence (a proxy signature) of direct A-to-I editing of the DNA on the TS. Thus, direct A-to-I editing of the DNA on the TS may be responsible for the increased number of T-to-C mutations observed during SHM in Aag deficient mice.[61]

3.3 Multiple ADAR target sites

Although the principle sites targeted for deamination by ADAR deaminases (ADAR1, ADAR2) are WA sites (where W = U or A) in dsRNA, i*n silico* modelling suggests that different deaminase isoforms, polymorphs and combinatorial hetero-multimers can emerge in response to cellular stress, such as during late-stage tumour progression.[62] This is particularly the case for the numerous possible polymorphisms of ADAR1, and the different isoforms evident in ADAR2 deaminases where alternative splicing and exon skipping events are very common.[43] Given that ADARs form homodimers, this potential post-transcriptional isoform diversity coupled with the fact that ADAR2 proteins can also auto edit their own RNA adenosines, suggests that heterodimer protein formation could also take place. This may account for the wide spectrum of A/T mutations observed in the mutation patterns of cancers. Additional processes such as error-prone reverse transcriptase copying of inosines by DNA polymerase-eta (pol-eta), and ADAR polymorphisms, isoforms, homodimers, and heterodimers may also result in some additional transversion mutations such as A-to-C and A-to-T that are observed in the absence of complete MMR.

3.4 Inosine in RNA and DNA is potentially oncogenic

Dysregulation of both ADAR1 and ADAR2 expression have been linked to cancer phenotypes. A relative decrease in RNA editing by ADARs 1 and 2 is associated with cancer progression, and possibly indicating that progressive auto-editing amongst the ADARs is leading to their inactivation.[63,64] ADAR1 has also been identified as a tumour promotor, and the gene for ADAR2 as a tumour suppressor in both liver and gastric cancers.[65,66]

Thus, our understanding of how ADAR editing patterns are regulated, and how these alter mutation profiles will be important for advancing our knowledge of the role of ADARs during oncogenesis, even before tumour development.[67] It is also important to understand that A-to-I deamination events are normal and essential modifications introduced by the specific ADAR deaminases which act co-transcriptionally in the nucleus, and later post-transcriptionally, in both nucleoplasm and the cytosol on pre-mRNA, mature mRNA and tRNA molecules.[43,47,68]

The widespread nature of the occurrence of A-to-I events in the normal human transcriptome is exemplified by the fact that more than 2.5 million sites with up to 95% residing in Alu repetitive retro-elements have been recorded in the rigorously annotated database of A-to-I RNA editing (RADAR).[69] Over 1 million of these sites could be classed as 'hyperedited' and most A-to- I events occur at intronic Alu repeats. The estimated breakdown by genomic location is as follows: 5' untranslated regions (UTRs) 10%, protein coding regions (CDS) 0.05%, introns 73%, 3'UTRs 3%, noncoding RNAs 0.15% and intergenic regions 14%. It should also be noted that the expression levels of various ADAR isoforms far exceeds the number of known C-to-U RNA editing events, which in comparison with A-to-I RNA editing is extremely rare with only a few sites so far discovered under normal physiological conditions.[40]

4. Deaminase target site access is directed by epigenetic markers

Much current research on the origins of mutations is now focused on understanding the role of deaminases in normal somatic cells during disease. New molecular evidence also suggests a far more general role for deaminase-based mutator mechanisms targeting non-Ig genes than was originally discovered for SHM in response to invading pathogens: this evidence supports the idea that environment-driven genetic and epigenetic changes, which when combined, target new sites for non-random mutation in many other genes across our genome. It is this genetic-epigenetic coupling that provides a molecular basis for understanding why some regions of a gene allow deamination, while others do not.

It has been known for over three decades that a number of processes write additional "regulatory" information onto the surface of genetic DNA without altering the nucleotide sequence. This process is termed "epigenetic" or "soft" re-wiring of the genome.[70] These epigenetic chemical alterations on sections of the gene make up a part, or all of the genetic regions that can potentially be targets for deamination. Conversely, it makes sense that those regions that are chemically protected from deamination are conserved regions where DNA fidelity needs to be maintained for survival and the proper functioning of an organism. It is also known that some of the epigenetic changes that are triggered by the environment in the non-conserved genomic regions can be stably inherited by offspring for several generations.[71-73]

Thus, environmentally triggered epigenetic change can directly mark DNA sites for possible new mutations, as well as to protect those genes and regions of genes that are to be conserved from potential deamination by.[73,74] In a study by Guo et al (2011), it was found that the TET1 gene and APOBEC1 are actively involved in region-specific neuronal activity-induced DNA methylation changes.[75] That is, environmentally driven epigenetic changes are used to mark regions of DNA as potential deaminase target sites where DNA diversity might be beneficial for an organism. Scourzic et al (2015) have reviewed many of the potential molecular steps implicated in these environmentally triggered epigenetic-somatic mutation paths.[74]

So, the deaminases alone are not directional drivers of mutations in SHM-like processes associated with cancer progression. They can be considered as merely the 'tools' with which the biome relies upon to introduce and regulate the introduction of *de novo* mutations giving rise to new genomic variants.

5. In-frame targeting by deaminases in protein coding regions

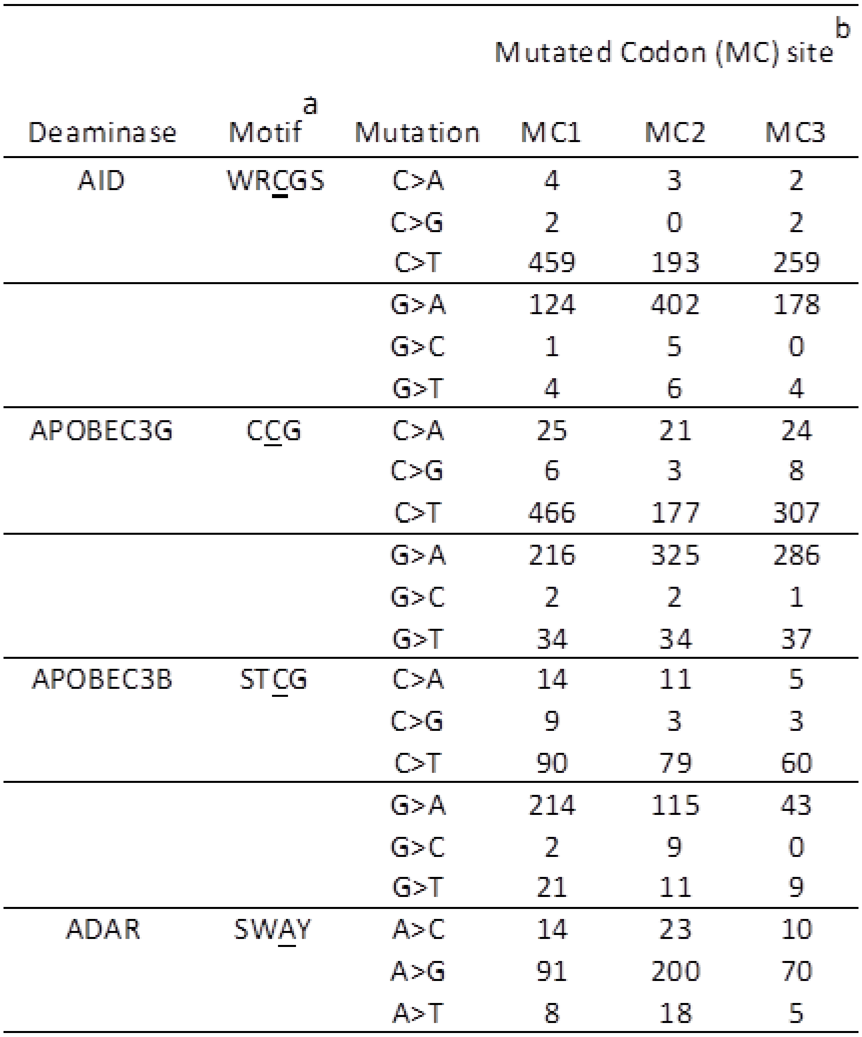
While it is well known that each deaminase predominantly targets motifs defined by a short nucleotide sequence, the discovery that they also preferentially target codon-reading frame sites, and that these sites are associated with a dominant type of mutation as an outcome, is relatively new. The process of in-frame deaminase targeting of protein coding regions is referred to as 'targeted somatic mutation' (TSM). For example, a well-known TSM signature for the cytidine deaminase AID is defined by:

i. the dominant type of mutation (C-to-T);
ii. the target motif WR<u>C</u> (W = A/T, R = A/G); and,
iii. the codon-reading frame site that is preferentially targeted for deamination in the mutated codon MC1 (referring to the first nucleotide in the mutated codon read by convention in the 5' to 3' direction).

5.1 TSM signatures in cancers

Table 4 provides an example of a TSM table for a cohort of 44 lower grade glioma (BLGG) cancers. In this example, the in-frame targeting preferences by deaminases is apparent. The data shown in Table 4 was downloaded as a vcf file from The Cancer Genome Atlas (TCGA), which is a collaboration between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI). TCGA PanCancer Atlas genomic data is stored and maintained by the US National Institute of Health (NIH) Genomic Data Commons (https://gdc.cancer.gov/access-data/data-access-processes-and-tools) and was accessed and visualized via the cBioPortal for Cancer Genomics (https://www.cbioportal.org/).[76,77] The data was processed and tabulated into a TSM format using GMDx's data processor (codon reference information system - CRISv.4.1.0). The source data file for Table 4 is appended as Supplementary data S.1.

Table 4. An example of a targeted somatic mutation (TSM) table for a cohort of 44 lower grade glioma (BLGG) cancers for which the progression free survival time was known.



			Mutated Codon (MC) site[b]		
Deaminase	Motif[a]	Mutation	MC1	MC2	MC3
AID	WRCGS	C>A	4	3	2
		C>G	2	0	2
		C>T	459	193	259
		G>A	124	402	178
		G>C	1	5	0
		G>T	4	6	4
APOBEC3G	CCG	C>A	25	21	24
		C>G	6	3	8
		C>T	466	177	307
		G>A	216	325	286
		G>C	2	2	1
		G>T	34	34	37
APOBEC3B	STCG	C>A	14	11	5
		C>G	9	3	3
		C>T	90	79	60
		G>A	214	115	43
		G>C	2	9	0
		G>T	21	11	9
ADAR	SWAY	A>C	14	23	10
		A>G	91	200	70
		A>T	8	18	5

Table 4 legend. Data was sourced from the Cancer Genome Atlas (TCGA). TCGA PanCancer Atlas genomic data is stored and maintained by the US National Institute of Health (NIH) Genomic Data Commons (https://gdc.cancer.gov/access-data/data-access-processes-and-tools) and was accessed and visualized via the cBioPortal for Cancer Genomics (https://www.cbioportal.org/).[76,77] a. The target motifs used in this TSM table are known to be associated with the deaminase indicated. b. The mutated codon (MC) site refers to the location of the mutated nucleotide (underlined in each motif) within the 3 nt structure of a codon and, by convention read in the 5-prime to 3-prime direction from DNA sequence on the non-transcribed strand (NTS), or 'top' strand.

Table 4 reveals the preferential targeting preferences for each of the 3-5 nt motifs known to be associated with the mutational activity of a deaminase. Background mutation levels may (at least in part) be the result of sequencing technology and alignment errors. In addition, in cancer cells there may be competition among AID/APOBEC family polymorphic members for deaminase binding sites. Mixed expression of AID/APOBEC polymorphic family members in heterozygotes would allow further heterogeneity in the specificities of their deaminase binding domains (DBDs) via alternative splicing of pre-mRNAs, or even interaction at the protein level involving alterations in DBD specificities as a consequence of potential hetero-multimer formation as discussed in Lindley et al (2016) and Mamrot et al (2019).[30,62,78] In addition, other studies have reported that different APOBEC family DBD isoforms play an important role in modulating deamination activity.[79.80] Yet despite these caveats, it is evident that the deaminase target site preferences for each binding domain is biochemically quite precise, and that these can be used to reveal mutation profile variations in cancer genome mutation patterns for individual patients, different tissue types or for different cancers.

The earliest use of a TSM table for cancer genomes was published as Table 2 in Lindley (2013).[81] It showed the occurrence of codon-context TSM signatures in pooled TP53 breast cancer data. Given that TP53 missense mutations mainly accrue in the DNA binding region of this central tumour suppressor gene (codons 100 to 300) the data set was presumably already selected, and the mutation data supported the expectation of selection, given that there are fewer mutations occurring at MC3 sites. However, sample size and/or selection bias *per se* cannot explain the significant non-random mutation patterns in the data. When mutation patterns are analysed using a TSM approach, it was immediately evident that the different AID/APOBEC family members display differential codon-biased mutation spectra. The TSM pattern also showed for the first time that the distribution pattern for adenosine mutations of ssRNA's were observed in DNA mutation patterns, and that they appeared to be influenced by codon reading frame structure. Another unexpected finding was that the molecular mechanisms involved rely upon the codon reading-frame structure at the level of ssDNA during transcription: as is also seen in Table 4, the differential MC1-MC2 targeting of cytidines appear to distinguish between cytidines on the "top" NTS (recorded as a mutation of 'C') from its Watson and Crick complement on the "bottom" or TS (recorded as mutations of 'G'). That is for example, in Table 4 the AID motif WR<u>C</u>GS shows that the C-to-T mutations (mutations of 'C' on the NTS) preferentially target MC1 sites, whilst the G-to-A mutations (mutations of 'C' on the TS) preferentially target MC2 sites. This transcription linked MC1-2 deaminase targeting 'toggle' of cytosines suggests that the transcription linked molecular structures differentiate between the TS and the NTS in the context of an 'open' transcription bubble. It is a feature of TSM patterns that has since been observed in all cancers to date, including the further published examples briefly described below.

In a later study, ovarian cancer patient whole exome mutation data reveals a similar TSM pattern to that shown in Table 4 (see Table 1, Lindley et al 2016).[78] It also shows that there are far more transitions than transversions. This study is also important as it showed how the targeting specificity for G-to-A transitions occurring in 194 high-grade serous ovarian adenocarcinoma samples is increased as the number of nucleotides defining the target motif is incrementally increased from 3 nts through to 6 nts (see Figure 1 in Lindley et al 2016).[78] A Kaplan–Meier plot predicting progression-free survival times for high grade serous ovarian cancer (HGS-OvCa) samples with a positive test result (based on TSM metrics), and compared to the cohort with a negative test result, revealed that the difference between the two cohorts is highly significant (see Figure 2 in Lindley et al 2016).[78] This was the first published example showing that the putative mutation targeting preferences of deaminases involve changes that can be used prognostically to predict progression (see Figure 3 in, Lindley et al 2016).[78] The Cox P-value is 1.57E-05, and the Log-Rank P value is 7.86E-07. The sensitivity measure is 95%, and the specificity is 90%.

5.2 TSM signatures in viral genomes

There have been many studies on the role of deaminases as direct viral restriction factors. The deaminases APOBEC3B and 3G in particular have been studied for two decades, and they are now colloquially known as ‘virus smashers’ due to their well characterised mechanism of action.[82,83] A study of the codon-contexted mutation patterns arising from infection with Zika virus, acute or chronic HCV and HBV, showed that codon-context also influences the preferred target sites for *de novo* viral RNA mutations at C-sites and A-sites at known APOBEC and ADAR motifs.[84] Analysing virus strains for mutations that characterise the host-parasite relationship during the innate immune response phase of infection revealed distinct TSM patterns: in the acute phase of infection, an innate immune response by the host involves the expression of deaminases to directly attack the invading virus at the most vulnerable phase(s) of its life-cycle. This acts to suppress or eliminate the virus before an effective adaptive immune response is mounted. This study showed that for each of the virus genomes analysed, the primary RNA-targeting APOBEC (APOBEC1, APOBEC3A) and ADAR deaminase signatures account for the majority of observed C-site and A-site transition mutations in viral genomes, primarily accruing at the MC3 position, or 3rd base in a codon read in the 5’ to 3’ direction. This study also showed that viral RNA genomes contain a number of additional layers of information that impact viral potency and function.

5.3 Many SNP signatures reveal a possible deaminase origin

We also tested our hypothesis that many SNPs curated at the online Mendelian inheritance in man (OMIM) database (potentially causative SNPs associated with clinically significant Mendelian inherited diseases) may have arisen by similar highly targeted deamination events.[85] Table 3 in Lindley and Hall (2018) shows the TSM profile for the set of pooled OMIM SNPs for genes on all chromosomes (obtained from the Clinvar database). In this study, it was shown that disease-associated SNPs on both the X chromosome, and for all chromosomes analysed by the TSM method, displayed SNP signatures preferentially targeting sequence motifs associated with known mutagenic deaminases as previously described in cancer genomes viz. AID, APOBEC3G, APOBEC3B and ADARs1/2. The results imply that over evolutionary time, the deamination of C-site and A-site targets appear to be ‘written’ into the human germline. It was concluded that similar types of deaminase-mediated molecular processes that occur in Ig SHM and cancer, may be contributing causative drivers of germline human SNPs.

5.4 TSM signatures predict cancer progression

Another key question asked by us and others was, can we identify mutation signatures predicting cancer progression. Using the TSM approach, a new and diverse range of cytidine and adenosine TSM signatures predicting cancer progression after surgical excision were first identified in data sourced from The TCGA cohort of 194 high grade serous ovarian carcinoma exomes.[78] The study used an *in silico* approach to identify nucleotide sequence changes of the target motifs, or inferred deaminase binding domains (Inf-DBDs) of key deaminases (AID, APOBEC3G, APOBEC3B, ADARs,) during oncogenesis. Examples of cancer progression associated signatures (CPAS) identified in High Grade are: a) ATCS C>T MC3, b) TCGA C>T MC1, c) GCGGC C>T MC1, and, d) TWTY T>C MC3 (see Table 3, Lindley et al 2016).[78]

Several other studies investigating genomic changes associated with cancer progression, and using a range of approaches, have been reported. These include Dieci et al (2016) who found that the PYGM gene is dramatically under-expressed in common cancers as compared to normal tissues and that low expression in tumours is correlated with poor relapse-free survival.[86] A major pan-cancer study by Li et at (2017) has concluded that the highly (or low) expressed genes in advanced cancers are likely to have higher (or lower) expression levels in cancers than in normal tissue, indicating that common gene expression perturbations drive cancer initiation and cancer progression.[87] In a study by Kjällquist et al (2018), it was found that paired metastatic lesions in breast cancers revealed some metastasis-enriched mutations in the A-kinase anchoring protein family (AKAPs) that predict progression.[88]

While there are now several genomic studies linking genomic changes to cancer progression, the underlying molecular processes are not fully understood. One theory is that changes in the polarization of tumour associated macrophages (TAMs) are accompanied by changes in the expression of deaminases with a new and diverse range of DBDs, and thus accounting for the generation of new somatic mutation signatures.[62] Although further work is required, it is hypothesised that M2 polarized macrophages extrude extracellular vesicles (EVs) loaded with deaminase proteins or deaminase-specific transcription/translation regulatory factors and that these may directly trigger deaminase diversification within cancer cells, and thus account for the many new somatic mutation signatures that are indicative of late stage cancer progression. The mechanisms proposed are molecularly reminiscent of combinatorial association of heavy (H) and light (L) protein chains following V(D)J recombination of Ig molecules required for pathogen antigen recognition by B-cells and T-cells respectively. This hypothesis now has a plausible indirect evidentiary base, and it is worth direct testing in future investigations.

6. Deaminase expression and innate immunity

As the molecular processes responsible for TSM patterns are found in tumour-normal mutation data, virus strain diversity mutation data, and in SNP databases associated with clinically significant Mendelian inherited diseases, it seems logical to ask: How are deaminase expression and immunity linked?

The endogenous deamination processes are a part of a programmed *sequelae* of actions triggered and regulated by the numerous ISG pathways. During a normal inflammatory response to infection or wound healing by macrophages and lymphocytes, a complex set of (ISG) pathways are immediately activated as a key part of a healthy innate immune response. Among the hundreds of anti-pathogen gene products co-ordinately expressed during an innate immune response, are the AID/APOBEC family of cytosine (C-site) and ADAR family of adenosine (A-site) deaminases.[1,42,89.90]

The deaminases are the main ISG induced proteins that attack the DNA or RNA of invading pathogens by extensively mutating their genomes, as for example during HCV, HBV or HIV-1 viral infections with C-to-U (T) and A-to-I(G) mutations.[12,42,83,84,91] This is the first line of innate immune defence, and it provides the adaptive immune system with time to mount an effective response which in the case of HBV may take 12 weeks or more. Ultimately, and during ISG induced attacks on foreign pathogens by deaminases, some *de novo* mutations that remain uncorrected will accumulate in the DNA of transcribed non-Ig genes, and possibly lead to a diagnosis of cancer in the infected cells.[92]

Although immunologists have been studying SHM in V-genes for decades, it has only recently been suggested that cancer mutation patterns are the result of the dysregulated SHM-like processes acting on non-Ig genes during transcription. The first study comparing the somatic mutation patterns observed in a range of non-lymphoid cancers with the strand bias SHM spectra of antibody genes was conducted in 2010.[93] It was found that overall, there is a striking resemblance between the patterns of Ig somatic mutations produced in germinal centre (GC) derived hypermutated B lymphocytes, and that of the various cancer samples analysed. This allowed the qualified conclusion that the likely source of mutations responsible for the somatic mutation spectrum in cancer genomes was the result of an ‘SHM-like’ process acting on non-Ig genes.

Later, when the TSM patterns were first discovered in 2012, a major concern raised was that the distinct somatic mutation pattern in cancers arose either as a result of target gene bias, or it was due to mutation selection bias at the level of protein structure.[81] This was further tested by comparing the TSM codon context mutation signatures of AID and ADAR in ovarian cancer genomes, with the mutation pattern targeted to the full-blown Ig hypermutation pattern of a passenger Ig transgene (which is a pattern free of antigen-selection biases due to protein selection). The V-regions of Ig Kappa transgenes are targets for hypermutation in germinal centre B cells. By comparing the AID and ADAR TSM profiles in these data, it was found that the codon-biased TSM spectrum of this population of advanced human cancer exomes is very similar (in *toto* or in part) to the codon-bias mutation spectrum found in the "passenger” (and thus “protein function selection free”) rearranged V kappa-Ox1 Jk5 transgene.[94] It was concluded that the TSM profiles of cancer genomes are the result of SHM-like processes that may occur across the genome during transcription. Thus, from an evolutionary perspective, the ancient deaminase mutational activity potentially targeting any gene during transcription during an innate immune response, has been exquisitely refined to create Ig variable region antigen-binding diversity during an adaptive immune response.

7. DNA and RNA deamination and repair mechanisms

While the previous sections highlight the main features characterising deaminase mutational activity and their role in innate immunity, most DNA deaminations are repaired in normal healthy body cells. Figure 3 shows the general flow of nucleotide sequence information in cancer genomes as a consequence of C-to-U and A-to-I deamination and repair events.

In most cases during transcription, the appearance of C-to-U lesions in growing DNA strands are promptly dealt with by the abundant and ubiquitous action of the BER enzyme, uracil DNA glycosylase (UNG) and resulting in efficient correction of the lesion (Figure 3A). Thus, in normal DNA repair physiology, C-to-U lesions move successively through an abasic (or AP) site which can lead to single stranded (ss) 5' nicks in the DNA as a result ubiquitous AP endonuclease action rapidly initiating a DNA repair process. The 3' OH ends of nicked DNA strands are predicted to be the primers of target site reverse transcription (TSRT) in the RNA/RT-based mechanism at the heart of the Ig SHM process.[96] Thus, dC-to-dU lesions nearby on the complementary DNA strands can result in a staggered double strand (ds) DNA break (DSB) which in turn can attract homologous recombination repair mechanisms. In general a C-to-U lesion in DNA results in a small patch of DNA being repaired by exonuclease digestions from 5' and 3' termini or endonucleases and finally, replicative and repair DNA polymerases and DNA ligases complete the repair by gap filling and ligation.[36,57]

In vivo, C-to-U mutations in DNA most often result in C-to-T mutations as a consequence of unrepaired C-to-U lesions (Figure 3B), or alternatively by an error prone repair synthesis pathway (Figures 3C, 3D). These events are the major contributors to point mutations found in normal B lymphocytes undergoing physiological SHM at rearranged Ig V(D)J loci.[97] They are also the major cause of point mutations in the DNA of cancerous cells affected by both AID and APOBEC aberrant deaminase activity.[8,9,78,98] Most often, the C-to-U and A-to-I lesions attract the MMR MSH2-MSH6 complex recruited by G:U mis-pairings in the DNA duplex (Figure 3C). Both AID and APOBEC1 are also known to actively facilitate direct C-to-T mutations by deaminating 5mCpG sites in DNA.[99] In this case, the ssDNA in the displaced NTS in an 'open' transcription bubble is the primary target for deamination of 5mC to T creating post transcriptional T:G mismatches. T:G mismatches then attract the BER enzyme thymine DNA glycosylase (TDG) which targets the T moiety of the T:G mismatch for repair.[100] Presumably deficiencies in TDG during cellular stress, such as late-stage cancer, could allow such mutations to accumulate.

The molecular DNA repair processes shown in Figure3, were first described in work completed 10-15 years ago by the mainstream immunology research community working on Ig SHM.[97,101] It resulted from the pioneering experiments of Neuberger and his associates in the early 2000s when they showed that AID was not an RNA editor as was originally hypothesised.[102-106] It was shown that AID acts directly on DNA to deaminate cytosines, and thus create highly mutagenic genomic uracils. This pioneering work forms the foundation for the current model of AID-mediated Ig SHM. Further investigation and application of these concepts in the context of disease (e.g. cancer) is currently continuing to build our understanding of the transcription linked mutagenic role of deaminases across the genome in immunology and physiology.

A critical evaluation of the literature identified some additional issues not explained by the previously accepted model of DNA repair.[107] Of particular interest is an RNA templated DNA process at A/T sites that involves the predicted gap filling of AID/APOBEC-mediated lesions by the Y family translesion DNA repair enzyme DNA pol-eta. Pol-eta is the only known error prone DNA polymerase involved in SHM.[108,109] It functions as a reverse transcriptase. and it can readily operate off the locus specific pre-mRNA template via TSRT.[96,110] This forms an important component of our current understanding of Ig SHM-like mutagenesis that occurs in non-Ig genes in cancers during transcription, and permitting the incorporation of RNA modifications into DNA as summarised in Figure 3C. Given the existence of a generic RNA-to-DNA repair process like this in yeast, similar RNA templated DNA repair processes are believed to have originated over a billion years ago at the very outset of metazoan evolution.[95]

To augment the conventional deamination and repair model shown in Figure 3, we have therefore added the information that the gap-filling (error prone) process is performed by DNA pol-eta since this is the translesion DNA repair enzyme recruited to such G:U lesions bound by MSH2-MSH6 complexes.[111] This is important because of DNA pol-eta's known role as an efficient reverse transcriptase (RT). Its human Y family relatives pol-kappa and pol-iota are also RTs.[110,112,113] The RT protein pol-eta is most likely then to produce site-specific integrated cDNA reverse transcripts into the genomic DNA as shown in Figures 3C and 3D.

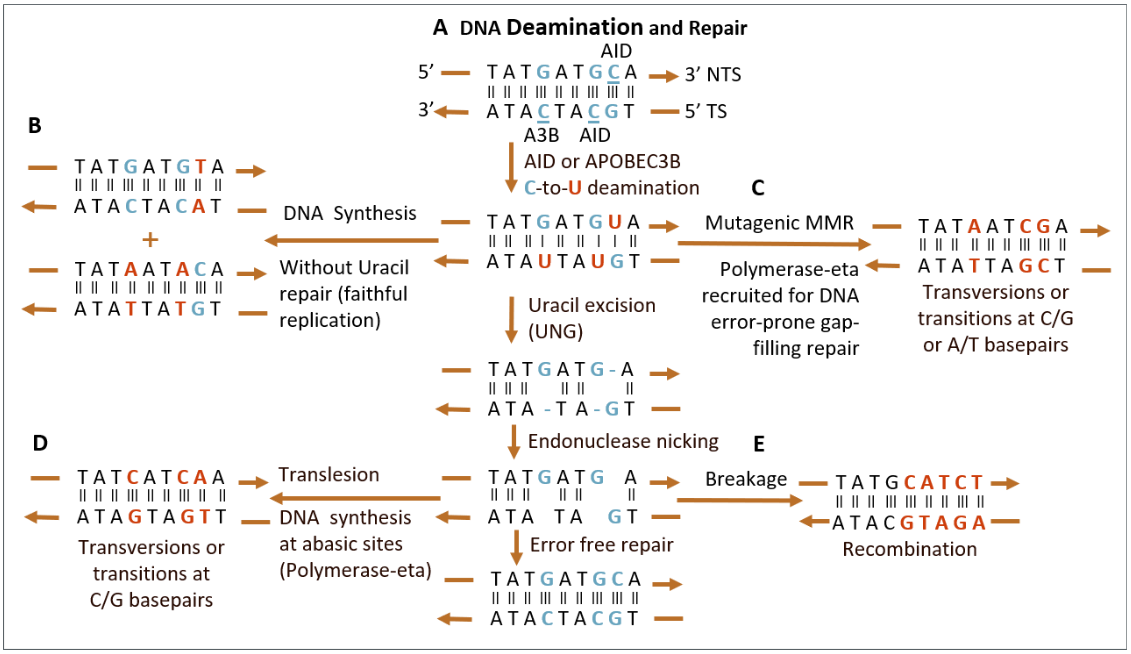


Figure 3. The general flow of nucleotide sequence information in cancer genomes as a consequence of C-to-U and A-to-I deamination and repair events. Adapted and modified from Figure 3 in Burns et al (2015).[9] A critical evaluation of the literature illuminates key issues not shown in the Burns et al rendition of the model, particularly the A/T targeted phase involving the supposed mutagenic gap filling of AID/APOBEC mediated lesions by the Y family translesion DNA repair enzyme DNA polymerase-eta (pol-eta) which is the only known error prone DNA polymerase involved in somatic hypermutation (Zeng et al 2001, Delbos et al 2007).[108,109] Pol-eta has reverse transcriptase activity,[10] and can operate off the locus specific pre-mRNA template via TSRT, as it appears to do so in Ig SHM,[94,,96] and thus incorporating RNA modifications into DNA.

Thus, there is strong evidence that C-to-U deamination of DNA results in C-to-T mutations as a consequence of unrepaired lesions (Figure 3B) or error prone repair synthesis (Figure 3C, 3D) and that these are major contributors to mutations in normal B lymphocytes undergoing physiological SHM at rearranged Ig V(D)J sites.[97,101] These processes are also a major contributor to mutations throughout the DNA of cancer genomes affected by both AID and APOBEC deaminase action.[9,12,78,81]

While single point mutations resulting from the error-prone DNA repair paths shown in Figure 3, such as C-to-T (and G-to-A on the Watson and Crick complementary DNA strand) by themselves, and in isolation, may not initiate any given cancer, it is clear that such mutations in the aggregate and by chance, may disrupt the protein-coding (and exon-intron splice sites, below) of so called "cancer driver" genes.[114] This results in the accumulation of mutations as cancer progresses from a pre-cancerous condition to late-stage cancer.[78,81]

8. Molecular processes underpinning mutagenic deaminase activity

As discussed in the previous sections, our current understanding of deaminases suggests that deaminase mutagenic activity is transcription linked and highly targeted. Deaminases play a crucial role in innate immunity, and the dysregulated SHM-like activity of deaminases is heavily implicated in cancer progression. Yet, despite what we know about the molecular and cellular processes involved, there is not a consensus on how both the cytidine and adenosine deaminases gain access to their respective targets during transcription.

Here, the 'RT model' shown in Figure 4 is proposed and discussed as it is consistent with our current knowledge of the biochemical and molecular processes involved and the 3D geometry of the 'open' transcription bubble. The model incorporates a transcription coupled pathway, and the need to embrace a DNA-mRNA-cDNA information flow. That is, the model implicates the nascent RNA as a copying template intermediary. At this point in time, the idea that the RT process involves an mRNA intermediary has not been embraced by many in the previously disparate fields of SHM or RNA editing, and yet no alternative RT-inclusive molecular model has been proposed to explain the emerging facts about deaminase activity and as described in the previous sections.

Referring to Figure 4, the transcription process is initiated in the 'open' transcription bubble. The resulting ssDNA provides access for AID to bind, which then initiates the SHM-like processes involving deaminases. APOBECs then also gain access to the ssDNA in the open transcription bubble to target their preferred deaminase binding sites. The first thing to note in Figure 4, is that by convention new mutations are called from the target nucleotide in the NTS ('top' DNA strand) to the resulting single base change in the complimentary TS (the newly templated cDNA bottom strand). Note also that pre-mRNA is formed from the TS ('bottom' DNA strand) in the elongated transcription bubble, and forming an estimated 9 nt annealed RNA:DNA hybrid deamination substrate for ADARs to potentially target. Examples of deamination events causing transition mutations targeted at ssDNA, dsRNA and the annealed RNA:DNA duplex substrates during transcription are shown in Figure 4. The target motif (and putative deaminase) and the deamination starting points are indicated (using a rectangle with dashes). By following the mutation path (indicated by colour), the mutation type can be read from the NTS through to the NTS in the newly formed DNA. The target reading-frame sites are also indicated (read 5' to 3'), and these are consistent with the preferred deaminase target sites shown in the TSM table (as shown in Table 4).

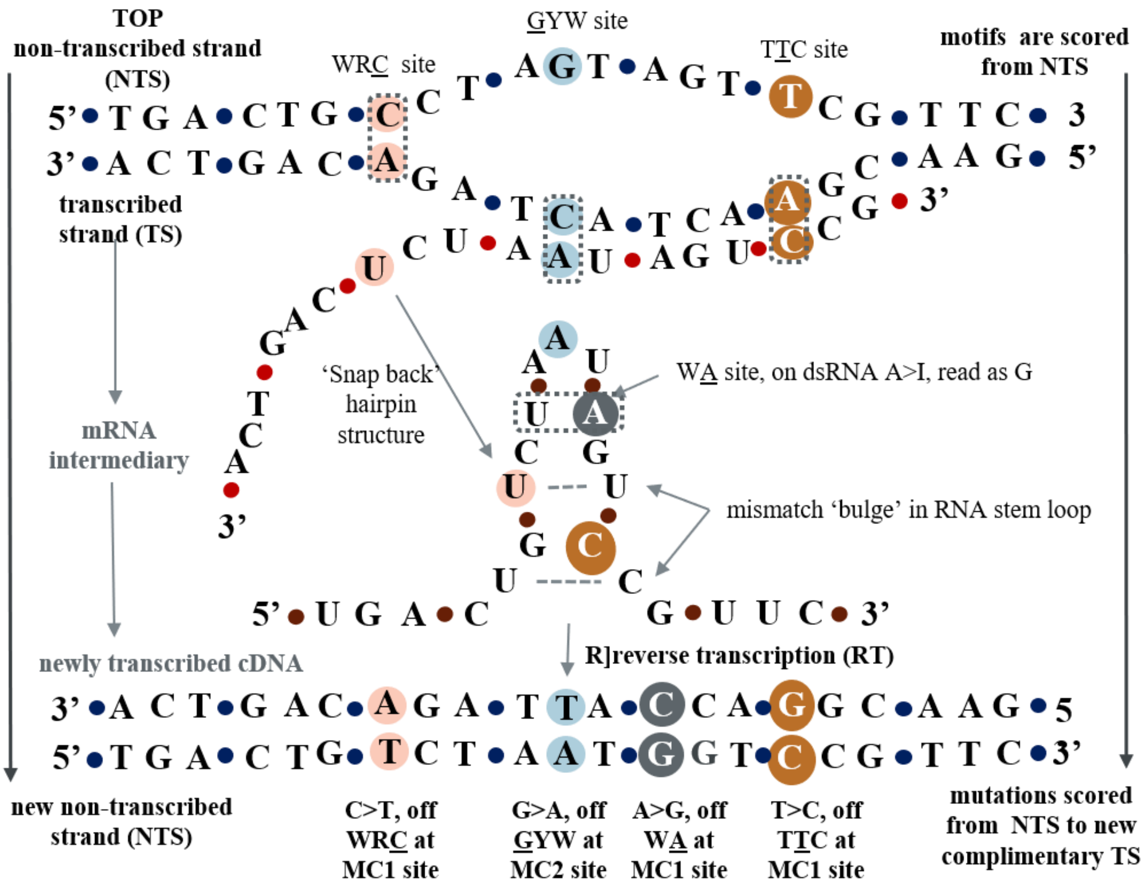


Figure 4. Diagram showing examples of the information flow for key deamination events causing transition mutations at DNA and RNA substrates, and at hybrid RNA:DNA hybrid substrates during transcription. For background on target site reverse transcription (TSRT) see Luan et al (1993).[96] For stalling of transcription elongation see Mooney et al (1998) and Moore and Proudfoot (2009).[115,116] For the normal preference for the displaced NTS strand for AID/APOBEC strand-biased C-to-U deamination, see.[117] For background on A-to-I RNA editing see Bass (2002), and for the action of the RNA exosome revealing ssDNA substrates on the TS, see Basu et al (2011).[68,118] Lower frequency transversions (not shown) may occur at abasic sites G-to-C and as a consequence of oxidative damage in DNA and RNA (G-to-T), and via error prone copying (e.g. DNA Pol-eta) opposite template Inosines, A-to-C, A-to-T.[60,107,119]

The newly templated RNA intermediary is produced from the TS (bottom strand) template, which then forms a snapback hairpin structure that in part forms dsRNA. During RT, it is this intermediary RNA template that is used to produce the new cDNA. Uncorrected mutations may also be generated by adenosine deamination targeting the annealed hybrid RNA:DNA in the transcription bubble, the dsRNA in a stem loop, or by targeting the RNA:DNA substrate during the RT step in which the mRNA is used as a template to form the cDNA.[60]

A key feature of the RT model of deaminase action at transcription bubbles, is that it suggests that most often the deamination events may target the central codon of the 9 nt in-frame register. Further research is required to understand the 3D geometry of the transcription bubble regarding access and substrate specificity by deaminases and other necessary proteins targeting the structure (e.g. RNA exosome). However, the main argument in support of adopting the model shown in Figure 4 is that there are no known deamination mechanisms causing A-to-I editing in DNA. While the caveat is that one cannot ever "prove a negative", DNA repair experts for many years have claimed there are no known deoxyadenosine (dA) deaminases which act directly on adenine in polynucleotide DNA strands.[36,57] This is an important fact. Yet there are now numerous codon-context TSM signatures observed at ADAR-specific WA-sites or WA-site variants in the DNA of cancer genomes (see for example the TSM table shown in Table 4). Logically these mutations could have only arisen by an RT step incorporating the A-to-I change into genomic cDNA as an A-to-G mutation. Such an RNA-to-DNA fixation event can occur either via a generic (metazoan) process associated with the DNA replicative/DNA repair machinery and/or directly in the case of Ig SHM, via cellular DNA repair reverse transcriptases such as the Y family of translesion DNA polymerases, particularly DNA pol-eta acting in its reverse transcriptase mode.[95,110] Whilst A-to-G is the predominant mutation type resulting from ADAR deamination events, the less frequent A-to-C and A-to-T transversions will also occur following likely error-prone reverse transcription and read as DNA mutations opposite RNA template inosines (e.g. as I is also expected to base pair with G or A as RT pol-eta misincorporation events). A-to-I events can also generate or delete splice sites at exon-intron junctions, or create new cryptic splice sites in the WA-rich target sites in Alu inverted repeats embedded in intronic pre-mRNAs, particularly in synaptic receptor genes in the brain (altering specific protein function.[48,98] Detailed DNA mutation and RNA stem-loop correlative analysis has also shown that they potentially contribute significantly to Ig variable region antigen-binding diversity during Ig SHM at V-region WA-hotspots.[107,121] A-to-I editing events can also modify regulatory siRNA and miRNA molecules and binding site structures in 5' and 3' UTRs and in the 3' UTRs of mature mRNA molecules, and, as part of the innate immune response via and ISG path during viral infections modifying adenosines in RNA viral genomes.[43,47,122] Hence, A-to-I RNA editing events can recode specific codons in target genes to alter protein function, and a molecular model extending the function of ADARs to integrate the new mutations into DNA during transcription at WA-motif sites involves an mRNA intermediary and an RT step.[68]

To further our explanations in support of this model, a wider knowledge of the mechanics of RNA Polymerase II elongation, its tempo of about 20 RNA ribonucleotides added per second with frequent pauses or stalling episodes is also required.[115,116,123,124] These pausing events are possibly related to the process of co-transcriptional splicing involving, for example, exon tethering, and the physiological necessity to maintain in-frame surveillance by the nonsense-mediated messenger RNA decay pathway (NMD) machinery which monitors exon-intron boundaries and premature UAG, UAA and UGA stop codons in the nascent pre-mRNA transcript.[125,126] Moreover, the work of Cook and colleagues strongly suggests that genetic loci involved in specialised and related RNA pol II transcriptional activities all occur within the *aegis* of organised proteinaceous structures termed "Transcription Factories.[127-130]

Thus, the model shown in Figure 4 is consistent with all of the facts known at this time, and it provides an opportunity to coherently explain how *de novo* mutations may be introduced into DNA or RNA before a modified nucleotide sequence is copied into newly transcribed DNA via a reverse transcription step: it allows one to logically break down the *sequelae* of events giving rise to DNA single nucleotide changes, and particularly incorporating those changes resulting from ADAR deamination events that are known to target dsRNA.

9. Concluding remarks

While this is a review of the current evidence for TSM processes giving rise to cancer, the TSM approach is different both in concept and utility in comparison with other mutation studies in that it is not specific for any gene or genic region, as in, for example, the identification of cancer driver genes. It is an approach based on the targeted nature of deaminase-associated mutation signatures and implying the likely genesis of many of the identified DNA mutation signatures in all or part of the genome, and it is dependent upon an RT mutation model.

In adopting the TSM approach for mutation analyses, the codon reading-frame biases observed are the result of endogenous mutation processes targeting protein-coding genes anywhere in the genome. i.e. targeted to that important 2% of the genome encoding all transcribed exomes in pre-cancerous or post-cancerous clones now revealed in routine next generation sequencing. As this approach enables one to define the deaminase mutation target sites with greater specificity, and thus identify the inf-DBDs, it also enables us to develop a new generation of genomic cancer diagnostics. Examples of clinical applications derived from TSM genomic metrics include the ability to identify cancer patients with: deficient MMR (dMMR); disrupted A-to-I editing (either too high or too low compared to controls); damaged or no APOBEC3B gene; possible unidentified chronic viral infection that may benefit from antiviral therapy in parallel with cancer treatment (e.g. by quantifying the number of variants inferring elevated APOBEC3G and/or APOBEC3B mutational activity compared to controls); dysregulated deaminase mutation profile metrics (consistent with damaged innate immunity compared to controls); and, CPAS (e.g. to identify those patients who may benefit from closer follow-up or further treatment following resection). While it is not within the scope of this paper to address each of these, and others, understanding the TSM method and the underlying molecular processes involved enables one to recognise how such metrics may be implemented in the clinic. A comprehensive set of TSM metrics (with outliers identified) takes us one step closer to personalising genomic testing and its relationship with the innate immune status of an individual.

Appendix A. Supplementary data S.1

Supplementary material related to the source data for Table 4 will be made freely available by contacting the corresponding author.

Financial Disclosure: This work was supported by GMDx Genomics Limited. The author has a financial interest in and is the Chief Scientific Officer (CSO) of GMDx Group Ltd (Australia), and holds several patent applications and patents granted for the clinical utility of some aspects of TSM phenomena.

Acknowledgements:

The author thanks Jared Mamrot, Nathan Hall and Edward Steele for the critical reading of this manuscript.

References

1. Schoggins JW, Rice CM. Interferon-stimulated genes and their antiviral effector functions *Curr Opin Virol*. 1(6) (2011) 519 - 525. doi: 10.1016/j.coviro.2011.10.008.

2. Lindley RA. The importance of codon context for understanding the Ig-like somatic hypermutation strand-biased patterns in TP53 mutations in breast cancer. Cancer Genet. 2013;206:222-226.
doi: 10.1016/j.cancergen.2013.05.016.

3. Petljak M, Alexandrov LB, Brammeld JS, et al. Characterizing mutational signatures in human cancer cell lines reveals episodic APOBEC mutagenesis. *Cell*. 2019;176:1282-1294. doi: 10.20517/cdr.2019.005.

4. Alexandrov LB, Kim J, Haradhvala NJ, et al. The repertoire of mutational signatures in human cancer. *Nature*. 2020;578:94-101. doi: 10.1038/s41586-020-1943-3.

5. Rada C, Jarvis JM, Milstein C. AID-GFP chimeric protein increases hypermutation of Ig genes with no evidence of nuclear localization. *Proc Natl Acad Sci*. 2002;99:7003-7008. doi: 10.1073/pnas.092160999.

6. Refsland EW, Harris RS. The APOBEC3 family of retroelement restriction factors. *Curr Top Microbiol Immunol*. 2013;371:1-27.
doi: 10.1007/978-3-642-37765-5_1.

7. H.C. Smith HC, Bennett RP, Kizilyer A, et al. Functions and regulation of the APOBEC family of proteins. *Cell Dev Biol*. 2012;23:258 - 268. doi:10.1016/j.semcdb.2011.10.004.

8. Conticello SG.The AID/APOBEC family of nucleic acid mutators. *Genome Biol*. 2008;9:229. doi:10.1186/gb-2008-9-6-229.

9. Burns MB, Leonard B, Harris RS. APOBEC3B: Pathological consequences of an innate immune *DNA Mutator Biomed J*. 2015;38:102-110.
doi: 10.4103/2319-4170.148904

10. Salter JD, Bennett RP, Smith HC. The APOBEC Protein Family: United by Structure, Divergent in Function. *Trends in Biochem Sciences*. 2016;41(7):578-594.
doi: 10.1016/j.tibs.2016.05.001.

11. Rogozin IB, Kolchanov NA. Somatic hypermutagenesis in immunoglobulin genes. II. Influence of neighbouring base sequences on mutagenesis. *Biochimica et biophysica acta*. 1992;1171:11-8. PMC6423074.

12. Beale RCL, Petersen-Mahrt SK, Watt IN, et al. Comparison of the different context-dependence of DNA deamination by APOBEC enzymes: correlation with mutation spectra in vivo. *J Mol Biol*. 2004;337:585-596.
doi: 10.1016/j.jmb.2004.01.046.

13. Roberts SA, Lawrence MS, Klimczak LJ, et al. An APOBEC Cytidine mutagenesis pattern is widespread in human cancers. *Nat Genet*. 2013;45:970-9766.
doi: 10.1038/ng.2702/.

14. Chan K, Roberts RA, Klimczak LJ, et al. An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat Genet*. 2015;47(9):1067-1072. doi: 10.1038/ng.3378.

15. Taylor BJM, Nik-Zainal S, Wu YL, et al, DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. *eLife*. 2013;2: e00534.
doi: 10.7554/eLife.00534.

16. Logue FC, Bloch N, Dhuey E, et al. A DNA sequence recognition loop on APOBEC3A controls substrate specificity. *PLoS ONE*. 2014;9(5):e97062.
doi:10.1371/journal.pone.0097062.

17. Chan K, Roberts SA, Klimczak LJ, et al. An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat Genet*. 2015;47(9):1067-1072. doi:10.1038/ng.3378.

18. Leonard B, Hart SN, Burns MB, et al. APOBEC3B upregulation and genomic mutation patterns in serous ovarian carcinoma. *Cancer Res.* 2013;73:7222-7231. doi: 10.1158/0008-5472.CAN-13-1753.

19. Burns MB, Lackey L, Carpenter MA, et al. APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature.* 2013a; 494:366-371. doi: 10.1038/nature11881.

20. Burns MB, Temiz NA, Harris RS. Evidence for APOBEC3B mutagenesis in multiple human cancers. *Nat. Genet*. 2013b;45:1-7. doi: 10.1038/ng.2701.

21. Langlois MA, Beale RC, Conticello SG, et al. Mutational comparison of the single-domained APOBEC3C and double-domained APOBEC3F/G anti-retroviral cytidine deaminases provides insight into their DNA target site specificities. *Nucleic Acids Res*. 2005;33(6):1913-1923. doi: 10.1093/nar/gki343.

22. Dang Y, Wang X, Esselman WJ, et al. Identification of APOBEC3DE as another antiretroviral factor from the Human APOBEC Family. *J. Virol*. 80 (2006) 10522-10533. doi: 10.1128/JVI.01123-06.

23. Yu Q, Chen D, Konig R, et al. APOBEC3B and APOBEC3C are potent inhibitors of simian immunodeficiency virus replication. *J of Biol Chem*. 2004;279:53379-53386. doi: 10.1074/jbc.M408802200.

24. Ehara H, Sekine SI. Architecture of the RNA polymerase II elongation complex: new insights into Spt4/5 and Elf1. *Transcription*. 2018;9(5):286-291. doi: 10.1080/21541264.2018.1454817.

25. Liddament MT, Brown WL, Schumacher AJ et al, APOBEC3F properties and hypermutation preferences indicate activity against HIV-1 in vivo. *Curr Biol*. 2004;14:1385-1391. doi: 10.1016/j.cub.2004.06.050.

26. Bishop KN, Holmes RK, Sheehy AM, et al. Cytidine deamination of retroviral DNA by diverse APOBEC proteins. *Curr Biol*. 2004;14:1392-1396. doi: 10.1016/j.cub.2004.06.057.

27. Hache ́G, Liddament MT, Harris RS. 2005 The retroviral hypermutation specificity of APOBEC3F and APOBEC3G is governed by the C-terminal DNA cytosine deaminase domain. *J Biol Chem*. 2005;280:10920-10924. doi: 10.1074/jbc.M500382200.

28. E. Miyagi E, Brown CR, Opi, S, et al. Stably expressed APOBEC3F has negligible antiviral activity. *J Virol*. 2010;84:11067-11075. doi: 10.1128/JVI.01249-10.

29. Henry M, Guetard D, Suspene R, et al. Genetic editing of HBV DNA by monodomain human APOBEC3 cytidine deaminases and the recombinant nature of APOBEC3G. *PLoS ONE*. 2009;4(1):e4277. doi: 10.1371/journal.pone.0004277.

30. Harari A., Ooms M, Mulder, LC, et al. Polymorphisms and splice variants influence the antiretroviral activity of human APOBEC3H. *J Virol*. 2009;83:295-303. doi: 10.1128/JVI.01665-08.

31. Sowden M, Hamm JK, Smith HC. Overexpression of APOBEC-1 results in mooring sequence-dependent promiscuous RNA editing. *J Biol Chem*. 1996;271:3011-3017. http://www.jbc.org/content/271/6/3011.long.

32. Blanc V, Park E, Schaefer S, et al. Genome-wide identification and functional analysis of Apobec-1-mediated C-to-U RNA editing in mouse small intestine and liver. *Genome Biol.* 2014;15(6):R79. doi: 10.1186/gb-2014-15-6-r79.

33. Sharma S, Patnaik SK, Taggart, RT et al. APOBEC3A cytidine deaminase induces RNA editing in monocytes and macrophages. *Nat Commun*. 2015;6:6881. doi: 10.1038/ncomms7881.

34. Sharma S, Patnaik SK, Kemera Z, et al. 2016 Transient overexpression of exogenous APOBEC3A causes C-to-U RNA editing of thousands of genes. *RNA Biol*. 2016;14:603-610. doi: 10.1080/15476286.2016.1184387.

35. Rosenberg BR, Hamilton CE, Mwangi MM, et al. Transcriptome-wide sequencing reveals numerous APOBEC1 mRNA editing targets in transcript 3′ UTRs. *Nat Struct Mol Biol*. 2011;18:230–236.
doi: 10.1038/nsmb.1975.

36. Alseth I, Dalhus B, Bjøras M. Inosine in DNA and RNA. *Curr Opin Genet Dev.* 2014;26:116-123.
doi: 10.1016/j.gde.2014.07.008.

37. Koning FA, Newman ENC, Kim E-Y, et al. Defining APOBEC3 expression patterns in human tissues and hematopoietic cell subsets *J Virol*. 2009;83:9474-9485.
doi: 10.1128/JVI.01089-09.

38. Refsland EW, Stenglein MD, Shindo K, et al. Quantitative profiling of the full APOBEC3 mRNA repertoire in lymphocytes and tissues: implications for HIV-1 restriction. *Nucl Acids Res*. 2010;38:4274 -4284.
doi: 10.1093/nar/gkq174.

39. Blanc V, Davidson NO. C-to-U RNA Editing: Mechanisms leading to genetic diversity. J. Biol. Chem. 2003;278:1395-1398.
doi: 10.1074/jbc.R200024200.

40. Blanc V, Davidson NO. APOBEC-1-mediated RNA editing. *Wiley Interdiscip Rev Syst Biol Med.* 2010;2:594-602. doi: 10.1002/wsbm.82.

41. Nishikura K, Functions and regulation of RNA editing by ADAR deaminases. *Annu Rev Biochem*. 2010;79:321-349.
doi:10.1146/annurev-biochem-060208-105251.

42. Samuel CE. Adenosine deaminases acting on RNA (ADARs) are both antiviral and proviral. *Virology*. 2011;411:180-193. doi: 10.1016/j.virol.2010.12.004.

43. Slotkin W, Nishikura K. Adenosine-to-inosine RNA editing and human disease *Genome Medicine*. 2013;5:105.
http://genomemedicine.com/content/5/11/105.

44. Agranat L, Sperling J, Sperling R. A novel tissue-specific alternatively spliced form of the A-to-I RNA editing enzyme ADAR2. *RNA Biol*. 2010;7:253-262. PMCID: PMC3062093.

45. Picardi EC, Manzari, C, Mastropasqua F, et al. Profiling RNA editing in human tissues: towards the inosinome Atlas. *Scientific Reports*. 2015;5:14941.
doi: 10.1038/srep14941.

46. Wu DD, Ye L-Q, Li Y, et al. Integrative analyses of RNA editing, alternative splicing, and expression of young genes in human brain transcriptome by deep RNA sequencing. *J Mol Cell Biol*. 2015;7:1–12.
doi: 10.1093/jmcb/mjv043.

47. O'Connell MA, Mannion NM, Keegan LP. The epitranscriptome and innate immunit. *PLoS Genet.* 2015;11 :e1005687.
doi: 10.1371/journal.pgen.100568.

48. Paz-Yaacov Levanon NEY, Nevo E, et al. Adenosine-to-inosine RNA editing shapes transcriptome diversity in primates. *Proc Natl Acad Sci USA.* 2010;107:12174-12179. doi: 10.1073/pnas.1006183107.

49. George CX, Samuel CE. Human RNA-specific adenosine deaminase ADAR1 transcripts possess alternative exon 1 structures that initiate from different promoters, one constitutively active and the other interferon inducible. *Proc Natl Acad Sci USA.* 1999:96:4621-4626.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC16382/pdf/pq004621.pdf.

50. George CX, Wagner MV, Samuel CE. Expression of interferon-inducible RNA adenosine deaminase ADAR1 during pathogen infection and mouse embryo development involves tissue-selective promoter utilization

and alternative splicing. *J Biol Chem*. 2005;280:15020-15028.
doi: 10.1074/jbc.M500476200.

51. Herbert A, Alfken J, Kim Y-G, et al. A Z-binding domain present in the human editing enzyme, double-stranded RNA adenosine deaminase. *Proc Natl Acad Sci USA*. 1997;94:8421-8426.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC22942/pdf/pq008421.pdf.

52. Herbert A, Rich A. The role of binding domains for dsRNA and Z-DNA in the in vivo editing of minimal substrates by ADAR1. *Proc Natl Acad Sci USA*. 2001;98:12132-12137.
doi: 10.1073/pnas.211419898.

53. Thomas JM, Beal PA. How do ADARs bind RNA? New protein-RNA structures illuminate substrate recognition by the RNA editing ADARs. Bioessays: News and *Reviews in Molec, Cellular and Devel Biol*. 2017 Apr;39(4).
doi: 10.1002/bies.201600187.

54. Chen CX, Cho DS, Wang Q, et al. A third member of the RNA-specific adenosine deaminase gene family, ADAR3, contains both single- and double-stranded RNA binding domains. *RNA*. 2000;6:755-767.
http://rnajournal.cshlp.org/content/6/5/755.long.

55. Oakes E, Anderson A, Cohen-Gadol A, et al. Adenosine Deaminase that acts on RNA 3 (ADAR3) binding to glutamate receptor subunit B pre-mRNA inhibits RNA editing in Glioblastoma. *J Biol Chem*. 2017;292(10):4326-4335.
doi: 10.1074/jbc.M117.779868.

56. Mladenova D, Barry G, Konen LM, et al. Adar3 Is Involved in Learning and Memory in Mice. *Front Neurosci*. 2018 Apr 13;12:243.
doi: 10.3389/fnins.2018.00243.

57. Lindahl T. Instability and decay of the primary structure of DNA. *Nature*. 1993;362(6422):709-715.
doi: 10.1038/362709a0.

58. Zheng YC, Lorenzo C, Beal PA, DNA Editing in DNA/RNA hybrids by adenosine deaminases that act on RNA. *Nucleic Acids Res*. 2017;45:3369-3377.
doi: 10.1093/nar/gkx05.

59. Brambatia A, Zardoniab L, Nardinia E, et al. The dark side of RNA:DNA hybrids. *Mutation Research/Reviews in Mutation Research*. 2020;784:108300.
doi: 10.1016/j.mrrev.2020.108300.

60. Steele EJ, Lindley RA. ADAR deaminase A-to-I editing of DNA and RNA moieties of RNA:DNA hybrids has implications for the mechanism of Ig somatic hypermutation. *DNA Repair*. 2017;55:1-6.
doi: 10.1016/j.dnarep.2017.04.004.

61. Longerich S, Meira L, Shah D, et al. Alkyladenine DNA glycosylase (Aag) in somatic hypermutation and class switch recombination. *DNA Repair*. 2007;6:1764-1773. doi: 10.1016/j.dnarep.2007.06.012.

62. Mamrot J, Balachandran S, Steele EJ, et al. Molecular model linking Th2 polarized M2 tumour-associated macrophages with deaminase-mediated cancer progression mutation signatures. *Scand J of Immunol*. 2019;89(5):e12760. doi: 10.1111/sji.12760.

63. Gallo A, Galardi S. A-to-I RNA editing and cancer: from pathology to basic science. *RNA Biol*. 2008;5:135-139. PMID: 18758244.

64. Gallo A, Locatelli F. ADARs: allies or enemies? The importance of A-to-I RNA editing in human diseases: from cancer to HIV1. *Biol Rev Camb Philos Soc*. 2012;87:95-110. doi: 10.1111/j.1469-185X.2011.00186.x.

65. Chan TH, Lin CH, Qi L, et al. 2014. A disrupted RNA editing balance mediated by ADARs (Adenosine Deaminases that act on RNA) in human hepatocellular carcinoma. *Hepatology*. 2014l63:832 -843.
doi: 10.1136/gutjnl-2012-304037.

66. Chan THM, Qamra A, Tan KT, et al. ADAR-Mediated RNA Editing Predicts Progression and Prognosis of Gastric Cancer. *Gastroenterology*. 2016;151:637- 650. doi: 10.1053/j.gastro.2016.06.043.

67. Tomasetti C, Vogelstein B, Parmigiani. G. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumour initiation. *Proc Natl Acad Sci USA*. 2013;110:1999–2004.
doi: 10.1073/pnas.1221068110.

68. Bass BL. RNA editing by adenosine deaminases that act on RNA. *Ann Rev Biochem*. 2002;71:817-846. doi: 10.1146/annurev.biochem.71.110601.135501.

69. Ramaswami G, Li JB. RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Res*. 2014;42:D109–113.
doi: 10.1093/nar/gkt996.

70. Jablonka E. The evolutionary implications of epigenetic inheritance. *Interface Focus*. 2017;7(5):20160135.
doi: 10.1098/rsfs.2016.0135.

71. Lindley RA. How Evolution Occurs: Was Lamarck Also Right? *EdgeScience*. 2011a;8:6-9.

72. Lindley RA. Born to Evolve; How Mutational and Epigenetic Changes Enable Adaptive Evolution. *G.I.T Laboratory J*. 2011b;1 September:3-4.

73. Lindley RA. A new treaty between disease and evolution - are deaminases the “universal mutators” responsible for our own evolution? *EdgeScience*. December 2018;36:16-20.

74. Scourzic L, Mouly E, Bernard OA. TET proteins and the control of Cytosine demethylation. *Cancer Genome Med*. 2015;7(1):9. doi: 10.1186/s13073-015-0134-6.

75. Guo JU1, Su Y, Zhong C, et al. Hydroxylation of 5-MethylCytosine by TET2 Promotes Active DNA Demethylation in the Adult Brain. *Cell*. 2011;145(3):423-434. doi: 10.1016/j.cell.2011.03.022.

76. Cerami E, Gao J, Dogrusoz U, et al.. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*. 2012;2(5):401-404.
doi: 10.1158/2159-8290.CD-12-0326.

77. Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science Signaling*.2013; 6(269):l1.
doi: 10.1126/scisignal.2004088.

78. Lindley RA, Humbert, Larmer PC, et al. Association between Targeted Somatic Mutation (TSM) signatures and HGS-OvCa progression. *Cancer Med*. 2016;5:2629-2640.
doi: 10.1002/cam4.825.

79. Gourraud PA, Karaouni A, Woo JM, et al. APOBEC3H haplotypes and HIV-1 pro-viral vif DNA sequence diversity in early untreated human immunodeficiency virus-1 infection. *Hum. Immunol*. 2011;72:207–212.
doi: 10.1016/j.humimm.2010.12.008.

80. Rathmore A, Carpenter MA, Demir O, et al The local dinucleotide preference of APOBEC3G can be altered from 5'-CC to 5'-TC by a single amino acid substitution. *J Mol Biol*. 2013;425(22):4442-54.
doi: 10.1016/j.jmb.2013.07.040.

81. Lindley RA. The importance of codon context for understanding the Ig-like somatic hypermutation strand-biased patterns in TP53 mutations in breast cancer. *Cancer Genet*. 2013;206:222-226.
doi: 10.1016/j.cancergen.2013.05.016.

82. Suspène R, Guetard D, Henry M, et al. Extensive editing of both hepatitis B virus DNA strands by APOBEC3 Cytidine deaminases in vitro and in vivo. *Proc of the Natl Acad of Science USA*. 2005;102(23):8321–8326.
doi: 10.1073/pnas.0408223102.

83. Vartanian J-P, Henry M, Marchio A, et al. Massive APOBEC3 editing of hepatitis B viral DNA in cirrhosis. *PLoS Pathogens*. 2010;6:e1000928.
doi: 10.1371/journal.ppat.1000928.

84. Lindley RA, Steele EJ. ADAR and APOBEC editing signatures in viral RNA during acute-phase innate immune responses of the host-parasite relationship to Flaviviruses. *Research Reports*. 2018;2:e1-e22.
doi: 10.9777/rr.2018.10325.

85. Lindley RA, Hall NA. APOBEC and ADAR deaminases may cause many single nucleotide polymorphisms curated in the OMIM database, *Mutat Res Fund Mol Mech Mutagen*. 2018;810:33-38.
doi: 10.1016/j.mrfmmm.2018.03.008.

86. Dieci MV, Smutná V, Scott V, et al. 2016 Whole exome sequencing of rare aggressive breast cancer histologies. *Breast cancer research and treatment*. 2016;156:21-32. doi: 10.1007/s10549-016-3718-y.

87. Li M, Sun Q, Wang X. Transcriptional landscape of human cancers. *Oncotarget*. 2017;8:34534-34551.
doi: 10.18632/oncotarget.15837.

88. Kjällquist U, Erlandsson R, Tobin NP, et al. Exome sequencing of primary breast cancers with paired metastatic lesions reveals metastasis-enriched mutations in the A-kinase anchoring protein family (AKAPs). *BMC cancer*. 2018;18:174. doi:1 0.1186/s12885-018-4021-6.

89. Schneider WM, Chevillotte MD, Rice CM. Interferon-stimulated genes: a complex web of host defenses. *Annu Rev Immunol*. 2014;232:513–545.
doi:10.1146/annurev-immunol-032713-120231.

90. Murphy K, Weaver C, Mowat A, et al. Janeway's Immunobiology 9th Edition, G.S. *Garland. Science*, Taylor & Francis Group, New York and London. 2017.

91. Gonzalez MC, Suspène R, Henry M, et al. Human APOBEC1 Cytidine deaminase edits HBV DNA. *Retrovirology*. 2009;6:96. doi: 10.1186/1742-4690-6-96.

92. Roberts SA, Gordenin DA. Hypermutation in human cancer genomes: footprints and mechanisms. *Nature Reviews Cancer*. 2014;14:786-800. doi: 10.1038/nrc3816.

93. Steele EJ, Lindley RA. Somatic mutation patterns in non-lymphoid cancers resemble the strand biased somatic hypermutation spectra of antibody genes. *DNA Repair*. 2010;9:600-603. doi: 10.1016/j.dnarep.2010.03.007.

94. Steele EJ. Somatic hypermutation in immunity and cancer: Critical analysis of strand-biased and codon-context mutation signatures. *DNA Repair*. 2016;45:1-24. doi: 10.1016/j.dnarep.2016.07.001.

95. Storici F, Bebenek K, Kunkel TA, et al. RNA-templated DNA repair. *Nature*. 2007;447:338-341.
doi: 10.1038/nature05720.

96. Luan DD, Korman MH, Jakubczak JL, et al. Reverse transcription of R2B mRNA is primed by a nick at the chromosomal target site; A mechanism for non-LTR retrotransposition. *Cell*. 1993;72:595-605. PMID: 7679954.

97. Di Noia JM, Neuberger MS. Molecular mechanisms of somatic hypermutation, *Annu Rev Biochem*. 2007;76:1-22. doi: 10.1146/annurev.biochem.76.061705.090740.

98. Chan K, Gordenein DA. Clusters of multiple mutations: Incidence and molecular mechanisms *Annu Rev Genet*. 2015;49:243-267. doi: 10.1146/annurev-genet-112414-054714.

99. Morgan HD, Dean W, Coker HA, et al. Activation-induced cytidine deaminase deaminates 5-Methylcytosine in DNA and is expressed in pluripotent tissues: implications for epigenetic reprogramming. *J Biol Chem*. 2004:279:52353-52360.

doi: 10.1074/jbc.M407695200.

100. Nabel CS, Manning SA, Kohli RM. The curious chemical biology of cytosine: deamination, methylation and oxidation as modulators of genomic potential. *ACS Chem Biol*. 2012;7:20-30. doi: 10.1021/cb2002895.

101. Teng G, Papavasiliou FN. Immunoglobulin somatic hypermutation. *Annual Rev Genet*. 2007;41:107-120. doi: 10.1146/annurev.genet.41.110306.130340.

102. Petersen-Mahrt SK, Harris RS, Neuberger MS. AID mutates E. coli suggesting a DNA deamination mechanism for antibody diversification, *Nature*. 2002;418:99-104. doi: 10.1038/nature00862.

103. Di Noia J, Neuberger MS. Altering the pathway of immunoglobulin hypermutation by inhibiting uracil-DNA glycosylase, *Nature*. 2002;419:43-48. doi: 10.1038/nature00981.

104. Harris RS, Petersen-Mahrt SK, Neuberger MS. RNA editing enzyme APOBEC1 and some of its homologs can act as DNA mutators. *Mol Cell*. 2002;10:1247-1253. doi: 10.1016/S1097-2765(02)00742-6.

105. Harris RS, Bishop KN, Sheehy AM, et al. DNA deamination mediates innate immunity to retroviral infection, *Cell*. 2003;113:803–809. doi: 10.1016/S0092-8674(03)00423-9.

106. Muramatsu M, Sankaranandi VS, Anant S, et al. Specific expression of activation-induced cytidine deaminase (AID), a novel member of the RNA-editing deaminase family in germinal centre B Cells. *J Biol Chem*. 1999;274:18470-18476. http://www.jbc.org/content/274/26/18470.long.

107. Lindley RA, Steele EJ. Critical analysis of strand-biased somatic mutation signatures in TP53 versus Ig genes, in genome wide data and the etiology of cancer *ISRN Genomics*. Vol 2013 Article ID 921418, 18 pages. https://www.hindawi.com/journals/isrn/2013/921418/

108. Zeng X, Winter DB, Kasmer C. et al. DNA polymerase-eta as an A-T mutator in somatic hypermutation of immunoglobulin variable genes, *Nat Immunol*. 2001;2:537-541. doi: 10.1038/88740.

109. Delbos F, Aoufouchi S, Faili A, et al. DNA polymerase-eta is the sole contributor of A/T modifications during immunoglobulin gene hypermutation in the mouse, *J Exp Med*. 2007;204:17-23 doi: 10.1084/jem.20062131.

110. Franklin A, Milburn PJ, Blanden RV, et al. Human DNA polymerase-eta an A-T mutator in somatic hypermutation of rearranged immunoglobulin genes, is a reverse transcriptase. *Immunol Cell Biol*. 2004;82:219-225. doi: 10.1046/j.0818-9641.2004.01221.x.

111. Wilson TM, Vaisman A, Martomo SA, et al. MSH2-MSH6 stimulates DNA polymerase eta, suggesting a role for A:T mutations in antibody genes, *J Exp Med*. 2005;201:637-645. doi: 10.1084/jem.20042066.

112. Su Y, Egli M, Guengerich FP. Human DNA polymerase η accommodates RNA for strand extension. *J Biol Chem*. 2017;292:18044-18051.
doi: 10.1074/jbc.M117.809723.

113. Su Y, Ghodke PP, Egli M, et al. Human DNA polymerase η has reverse transcriptase activity in cellular environments. *J Biol Chem*. 2019;294:6073–6081.
doi: 10.1074/jbc.RA119.007925.

114. Vogelstein B, Papadopoulos N, Velculescu VE, et al. Cancer genome landscapes. *Science*. 2013;339:1546-1558. doi: 10.1126/science.1235122.

115. Mooney RA, Artsinovitch I, Landick R. Information processing by RNA polymerase: recognition of regulatory signals during RNA chain elongation. *J Bact*. 1998;180: 3265-3275. http://jb.asm.org/content/180/13/3265.long.

116. Moore MJ, Proudfoot NJ. Pre-mRNA processing reaches back to transcription and ahead to translation. *Cell*. 2009;136:688-700.

doi: 10.1016/j.cell.2009.02.001.

117. Sohail A, Klapacz J, Samaranayake M, et al. Human activation-induced cytidine deaminase causes transcription dependent, strand-biased C to U deaminations. *Nucl Acids Res*. 2003;31:2990-2994. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC162340/.

118. Basu U, Meng FL, Keim C, et al. The RNA Exosome targets the AID cytidine deaminase to both strands of transcribed duplex DNA substrates. *Cell*. 2011;144:353-363. D oi: 10.1016/j.cell.2011.01.001.

119. Kuraoka I, Endou M, Yamaguchi Y. Effects of endogenous DNA base lesions on transcription elongation by mammalian RNA polymerase II. *J Biol Chem*. 2003;278:7294-7299.
doi: 10.1074/jbc.M208102200.

120. Paul MS, Bass BL. Inosine exists in mRNA at tissue-specific levels and is most abundant in brain mRNA. *EMBO J*. 1998;17:1120=1127.
doi: 10.1093/emboj/17.4.1120

121. Steele EJ, Lindley RA, Wen J, Weiler GF. Computational analyses show A-to-G mutations correlate with nascent mRNA hairpins at somatic hypermutation hotspots. *DNA Repair*. 2006;5:1346-1363. doi: 10.1016/j.dnarep.2006.06.002.

122. Samuel CE. Adenosine deaminase acting on RNA (ADAR1), a suppressor of double-stranded RNA-triggered innate immune responses. *J Biol Chem*. 2019;294(5):1710-1720.
doi: 10.1074/jbc.TM118.004166.

123. Brody Y, Neufeld N, Bieberstein N, et al. The in vivo kinetics of RNA Polymerase II elongation during co-transcriptional splicing. *PLoS Biol*. 2011;9(1):e1000573. doi: 10.1371/journal.pbio.1000573.

124. Veloso A, Kirkconnell KS, Magnuson B, et al. Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. *Genome Res*. 2014;24:896-905.
doi: 10.1101/gr.171405.113.

125. Dye MJ, Gromak N, Proudfoot NJ. Exon tethering in transcription by RNA polymerase II. Mol. *Cell*. 2006;21:849–859. doi: 10.1016/j.molcel.2006.01.032.

126. Chang YF, Imam JS, Wilkinson MF. The nonsense-mediated decay RNA surveillance pathway. *Annu Rev of Biochem*. 2007;76:51-74.
doi:10.1146/annurev.biochem.76.050106.093909.

127. Iborra FJ, Pombo A, Jackson DA, Cook PR. Active RNA polymerases are localised within discrete transcription'factories' in human nuclei. *J Cell Sci*. 1996:109:1427-1436. http://jcs.biologists.org/content/109/6/1427.long.

128. Eskiw CH, Rapp A., Carter DRF et al. RNA polymerase II activity is
located on the surface of protein-rich transcription factories. *J Cell Sci*. 121 (2008) 1999-2007.
doi: 10.1242/jcs.027250.

129. Osborne CS, Chakalova L, Mitchell JA, et al. Myc dynamically and preferentially relocates to a transcription factory occupied by Igh. *PLoS Biol*. 2007;5(8):e192.
doi: 10.1371/journal.pbio.0050192.

130. Cook PR. A model for all genomes: the role of transcription factories. *J Mol Biol*. 2010;395:1-10.
doi: 10.1016/j.jmb.2009.10.031.