Mutat Res Fund Mol Mech Mutagen 810 (2018) 33–38

Contents lists available at ScienceDirect

Mutat Res Fund Mol Mech Mutagen

journal homepage: www.elsevier.com/locate/mut

APOBEC and ADAR deaminases may cause many single nucleotide polymorphisms curated in the OMIM database

Robyn A. Lindley[a,b,*], Nathan E. Hall[a]

[a] GMDxCo Pty Ltd, Melbourne Vic, Australia
[b] Department of Clinical Pathology, Faculty of Medicine, Dentistry & Health Sciences, University of Melbourne Vic, Australia

ARTICLE INFO

Keywords:
Single nucleotide polymorphisms (SNPs)
Targeted somatic mutation (TSM)
Codon-context mutations
AID/APOBEC-deaminase patterns
ADAR deaminase A-to-I RNA and DNA editing

ABSTRACT

Cytosine and adenosine deamination events (DNA, RNA substrates) account for most codon-context Targeted Somatic Mutation (TSM) patterns observed in immunoglobulin (Ig) somatic hypermutation (SHM), and in cancer exomes following Ig-SHM-like responses. TSM refers to the process of somatic mutagenesis involving deamination events that results on a dominant type of mutation (e.g., C-to-T), and co-incident at a particular motif (e.g., WRC), and preferentially targeting the first, second or third nucleotide position within the mutated codon (e.g. MC1, MC2 or MC3, read 5-prime to 3-prime). It is now widely accepted that if left uncorrected, the accumulation of uncorrected TSMs involving the deaminases, may lead to a diagnosis of cancer or other degenerative disease. Our hypothesis is that many missense, nonsense and synonymous single nucleotide polymorphisms (SNPs) associated with clinically significant diseases may have arisen in the population by similar highly targeted deamination events. The OMIM database was searched for disease-associated SNPs on the X chromosome, and for all chromosomes. The nucleotide substitution patterns for disease-associated SNPs were analyzed by the TSM method to identify the likely deaminase source for C-to-U (C-to-T/G-to-A) and A-to-I (A-to-G/T-to-C) derived gene mutations preferentially targeting known sequence motifs associated with the deaminases: AID, APOBEC3G, APOBEC3B and ADAR 1/2. Of the 789 OMIM SNPs analysed. In both data sets, the mutation targeting preferences within the mutated codon reveal a statistically significant bias ($p < 0.001$). The results imply that a deamination of C-site and A-site targets are written into the human germline for the chromosome wide exomic SNPs analysed. This is consistent with previously observed mutation patterns arising in cancer genomes and hypermutated Ig genes during SHM. The results imply that similar types of deaminase-mediated molecular processes that occur in somatic hypermutation and cancer, may be contributing causative drivers of human exomic SNPs.

1. Introduction

Genetic and biochemical investigations in immunoglobulin somatic hypermutation and exome wide cancer systems have established two main types of deamination event in DNA and RNA sequences, cytosine to uracil, adenosine to inosine (C-to-U, A-to-I). These transitions

Abbreviations: A > > T, SNPs or mutations of A exceed SNPs or mutations of T; ADAR, Adenosine Deaminase that acts on RNA. There are two main isoformsADAR1 and ADAR2 that have similar WA targeting specificity; AID, activation induced cytidine deaminase, an APOBEC family member, initiating via C-to-U 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; 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; A-to-I, adenosine-to-inosine RNA editing; Deaminase, catalytic domain in ADAR and AID/APOBEC enzymes; Inf-DBD, inferred deaminase binding domain; G > > C, SNPs or mutations of G exceed SNPs or mutations of C; I, inosine; Ig-SHM-like, response strand-biased somatic mutation patterns similar to that observed in Ig SHM; MMR, mismatch repair; Motif, 2 to 6 nucleotide (N) sequence defining the specificity of deaminase targeting domains; MSH2-MSH6, MutSα heterodimer recognising mispaired bases in DNA duplex; NTS, the non-transcribed, or "Top", 5' to 3' DNA strand; Pol-η, or DNA polymerase-η (eta); RNA, Pol II RNA Polymerase II; RT, reverse transcriptase; RT-Pol-η, reverse transcriptase activity displayed by Pol-η; SNP, single nucleotide polymorphism; SHM, somatic hypermutation; T, Thymine; TCR, Transcription-Coupled Repair; TDG, thymine DNA glycosylase, targets TG mispairs that arise from deamination of 5 methyl C in CpG (CG) motifs; TS, the transcribed, or "Bottom", 3' to 5' strand, in the context of an open Transcription Bubble; TSM, targeted somatic mutations tthe process of targeting C and A nucleotides for deamination in actively transcribed genes that results in a dominant type of mutation caused by a deaminase binding domain (DBD) or an Inferred-DBD at a particular codon position; U, uracil; 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 pair involving A or U/T; Y, pyrimidines T/U or C

* Corresponding author at: Department of Clinical Pathology, The Victorian Comprehensive Cancer Centre (VCCC) Faculty of Medicine, Dentistry & Health Sciences, The University of Melbourne, Parkville, Victoria, 3010, Australia.
E-mail address: robyn.lindley@unimelb.edu.au (R.A. Lindley).

https://doi.org/10.1016/j.mrfmmm.2018.03.008
Received 7 November 2017; Received in revised form 11 January 2018; Accepted 28 March 2018
Available online 22 June 2018

dominate observed datasets of somatic mutations. Members of the AID/APOBEC deaminase enzyme family cause C-to-U transitions in non-base paired, or single stranded (ss) stem-loop regions of RNA and transcriptionally displaced ssDNA in both viral [1,2], transcriptome [3–7] and genomic sequences [1,2,8–14]. Following failures in base excision and/or mismatch repair (MMR) these mutations after replication are scored in DNA sequences as C-to-T (and reverse complements being G-to-A). Transversion mutations occurring at such C-sites appear at lower frequency and are the likely consequence of error-prone mismatch DNA and gap repair processes (post transcription and post replication). Members of the ADAR deaminase enzyme family cause A-to-I deaminations in both double stranded (ds) RNA [15,16] as well as in both nucleic moieties in RNA:DNA hybrids [17] most likely in vivo at Transcription Bubbles [18]. Since I base pairs like G, this leads to A-to-G transitions in both viral [19] and cellular genomes [20–22]. Following transcription, A-to-I events can be scored in genomic DNA as A-to-G transitions (reverse complement being T-to-C). These RNA-to-DNA error-prone copying events are a possible explanation for the occurrence of targeted somatic mutations, as well as strand biased mutation signatures which accrue on the transcribed strand (TS) viz. mutations of A sites exceeding those of T (A > > T) and mutations of G sites exceeding those of C (G > > C) observed in both rearranged immunoglobulin (Ig) loci (V[D]J) in B lymphocyte genomes and at non-Ig loci across cancer exomes [18,23,24].

It has since been discovered by one of us that deaminases also recognise the open reading frame of a gene [25]. In doing so, each deaminase preferentially targets either the first, second or third nucleotide (read 5-prime to 3-prime) within the mutated codon. The question is often raised: Why analyze the codon context of somatic mutations associated with deamination by known deaminases? The simple answer is that once it was discovered that a deaminase can recognize the open reading frame of a gene, the natural conclusion was that deaminase targeting is far more specific than previously thought [25], and thus provides a unique genetic 'footprint' associated with each mutation. The Targeted Somatic Mutation (TSM) footprint or signature, associated with previous deaminase activity can therefore be more precisely identified by analysing the mutation type, the target motif, and the site targeted within the 3-nucleotide structure of a codon (read 5-prime to 3-prime). The initial discovery of TSM signature specificity arose out of an analysis of TP53 mutations identified in breast cancer genomes [25]. It later became clear that the application of this method allowed the identification of the inferred deaminase binding domains at C-site and A-site substrate motifs in all actively transcribed exomic regions across the genome. Such codon-context patterns have also been identified using a 9 nt region encompassing the motif in cancer exomes during cancer progression [24]. These studies allow informative identification of the likely source of the endogenous specific deaminase causing the observed somatic mutation.

In these studies, the AID/APOBEC and ADAR deaminases have been found to target C-site and A-site residues in a sequence specific manner. Thus, AID favours C-sies primarily in context of WRC motifs (W = A or T; R = A or G), scored in reverse complements as GYW sites (Y = T or C); APOBEC3G deaminates CCG sites (or CGG); APOBEC3B deaminates TCG (or CGA) and ADARs deaminate WA sites. However, the nucleotide flanking sequence context of these primary motifs is important providing subtle modulations to the motif specificity [26]. Genetic polymorphisms and different somatic isoforms (usually generated by alternative splicing), particularly in heterozygotes, are thought to create molecular competition and modulations in the inferred deaminase binding domains (inf-DBD) in motif targeting as defined by variations in 5' and 3' nucleotides [24]. The emergence of different Inf-DBDs is now an accepted concept in cancer progression.

It has been further shown that C-site and A-site deamination events are very specific, dictated by the codon context of the motif [25]. This is termed "Targeted Somatic Mutation" (TSM). TSM refers to the process of somatic mutagenesis involving deamination events that results on a dominant type of mutation (e.g., C-to-T), and co-incident at a particular motif (e.g., WRC, where W = A or T), and preferentially targeting the first, second or third nucleotide position within the mutated codon (e.g. MC1, MC2 or MC3, read 5-prime to 3-prime). The exquisite deamination specificity can be thought of as targeting C and A substrates in ssDNA, dsRNA and RNA: DNA hybrid substrates most plausibly within the context of Transcription Bubbles stalled in codon register [18].

The hypothesis tested here is that these motif-specific TSM deamination profiles, indicative of AID/APOBEC and ADAR mutator activity on transcriptomes and genomic exome regions in Ig somatic hypermutation in B lymphocytes and Ig-SHM like responses across cancer genomes [24,25] should be detected in validated clinically relevant SNP datasets. In this paper, TSM patterns are analysed in the SNP datasets curated by medical relevance, that is the missense and nonsense SNPs curated at the Online Mendelian Inheritance in Man (OMIM) data base for an X chromosome dataset, and a dataset that pools disease-related SNPs on all chromosomes.

2. Materials and methods

2.1. SNPs harvested from the OMIM database

A search was conducted for "X-Linked Recessive Disease" at the Online Mendelian Inheritance in Man database at the NCBI website from 2 Jan 2017 to 7 Mar 2017 (https://www.ncbi.nlm.nih.gov/omim/?term=X+Linked+Recessive+Dieases). The search results are provided in Supplementary File S1, Table 1. The allelic variant information contains DNA sequence modifications associated with X-linked diseases for exomic SNPs (missense, truncating), splice site sequence changes, other intronic changes associated with disease, exome-related indels, whole exon deletions, codon/amino acid expansions and deletions. Related information for each gene from the wider dbSNP site, and the GeneView protein-coding SNP information also curated at NCBI for that gene, was obtained by accessing SNP at NCBI (https://www.ncbi.nlm.nih.gov/pubmed) and going to GeneView to recording the number listed for that gene e.g. for the BTK gene (https://www.ncbi.nlm.nih.gov/snp/?term=BTK). This was done on 14 June 2017. A summary of the base composition, and the types of genetic modifications for these 49 X-chromosome genes curated in OMIM is provided in Supplementary File S1, Tables 3 and 4 respectively.

As the disease-associated X chromosome SNPs are a limited dataset,

Table 1
Mutation patterns summarising mutations by type and showing strand bias for OMIM disease-linked SNP data for: A. all genes on the X chromosome, and B. all genes on all chromosomes extracted from the Clinvar database.

A. Mutation patterns for pooled OMIM SNP data for all genes on the X Chromosome

WT Base	Variant Base				
	A	T	C	G	TOTAL
A		21	22	**68**	111
T	25		**59**	34	118
C	31	**213**		38	282
G	**157**	77	44		278
					789

Transitions 63%, Transversions 37%

B. Mutation patterns for pooled OMIM SNP data for all genes on all chromosomes

WT Base	Variant Base				
	A	T	C	G	TOTAL
A		566	515	**1844**	2925
T	607		**1772**	726	3105
C	755	**4973**		783	6511
G	**4353**	1169	784		6306
					18847

Transitions 69%, Transversions 31%