ELSEVIER

Cancer Genetics 206 (2013) 222–226

Cancer Genetics

The importance of codon context for understanding the Ig-like somatic hypermutation strand-biased patterns in *TP53* mutations in breast cancer

Robyn A. Lindley [a,b,*]

[a] *Melville Analytics Pty Ltd, Jindabyne, Australia;* [b] *C Y O'Connor ERADE Village Foundation, Piara Waters, Australia*

Evidence already exists that the activation-induced deaminase (AID)/APOBEC family constitutes a set of differentially expressed enzymes capable of deaminating cytosines (C to U) in single-stranded DNA (ssDNA) and that they are potentially powerful mutagens. The mutagenic processes involved are believed to be activated in many nonlymphoid tissue types—for example, initiating some cancers and/or leading to further somatic mutagenesis. To investigate the extent that codon context might be important in influencing the likely location of *TP53* mutations in breast cancer, the codon-bias patterns resulting from the ssDNA target specificities of cytidine deaminases of the AID/APOBEC family were analyzed. The data indicate that codon context strongly influences the likely location of mutations at motifs for AID/APOBEC1/APOBEC3G, and at WA sites. An unexpected finding is a highly significant preference for transitions of cytosine to occur at the first nucleotide position and for transitions of guanosine to occur at the second nucleotide position in the mutated codon (read 3′ to 5′). Thus, the mechanisms involved appear to be sensitive to codon reading frames and to have an intrinsic ability to differentiate between the cytosines on the nontranscribed strand and those on the transcribed strand in the context of an open "transcription bubble."

Keywords *TP53*, codon-bias, breast cancer, somatic hypermutation, mutagenesis

In an analysis using genome-wide protein kinase gene mutation data provided by the Wellcome Trust Sanger Institute's Cancer Genome Project (CGP), many nonlymphoid cancers were found to display the same dominant strand-biased mutation patterns at Watson–Crick base pairs that are typical of rearranged immunoglobulin (Ig) variable genes after somatic hypermutation (SHM) (1–2). The main strand-bias mutation pattern is characterized by dominant C-to-T and G-to-A transitions, with the number of G-to-A mutations exceeding the number of C-to-T mutations. The second systematic strand-bias pattern observed in the CGP genome-wide data, and in the *TP53* data from the International Agency for Research on Cancer (IARC) for a range of nonlymphoid cancers, targets A:T base pairs, predominantly at WA-hotspot motifs. This common strand-bias pattern is characterized by the number of mutations of A-to-G exceeding the number of mutations of T-to-C (3). Previously, We have also shown that the endogenous pattern represented by the IARC *TP53* "All Breast Cancers" mutation data reveals that all major and minor G-site mutation hotspots can be classified as the direct result of either activation-induced cytidine deaminase (AID) or apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3G (APOBEC3G) C-to-U deaminations targeting the *TP53* DNA strands (3).

The unifying explanation for the common strand-bias mutation patterns found in both genome-wide and *TP53* data is that dysregulated SHM initiation via AID activation in nonlymphoid tissue may be one of the primary causes of the mutagenesis leading to cancer (3). The initial stage involves the conversion of cytosine to uracil (C-to-U) in DNA and the initiation of SHM-like mutagenic processes that contribute to oncogenesis in nonlymphoid somatic tissue. Several other studies have also discussed this possibility (4–6). As a result, a new model of oncogenesis is emerging; it implies

Received February 25, 2013; received in revised form May 10, 2013; accepted May 10, 2013.

* Corresponding author.
E-mail address: admin@melvilleanalytics.com

2210-7762/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.cancergen.2013.05.016

a key role for the AID/APOBEC family of cytidine deaminases and the processes known to generate the characteristic strand-biased mutation signatures following the SHM of Ig genes (3).

The aim of this study is to use these insights to analyze the *TP53* somatic mutation data for "All Breast Cancers" to test the hypothesis that codon context is important in influencing the likely location of mutations at key motifs believed to be associated with somatic mutagenesis. The possibility that the transcription-linked mechanisms involved might be able to read the underlying codon structure of single-stranded DNA (ssDNA) *in-frame* during transcription has been overlooked until now, and the recent findings may help to advance our understanding of the origin of the established genome-wide and *TP53* gene strand-bias patterns (2–3).

Materials and methods

Databases of somatic *TP53* mutation prevalence in human cancer are available through the IARC *TP53* database, R15 November 2010 (p53.iarc.fr/). This resource compiles approximately 30,000 *TP53* somatic variations reported in human cancers, with annotations on mutation type, codon number, and the tumor topography and morphology associated with the disease (7).

The IARC *TP53* "All Breast Cancers" data set was selected as representative of the endogenous somatic mutation pattern because this mutation pattern appears to arise in tissue that is least accessible to carcinogens in tobacco smoke or is directly exposed to other exogenous carcinogens (3,8). Further, a potential link between hormone exposure (estrogen) and the inappropriate expression of AID in breast tissue has been experimentally established (9). The number of single point mutations in this subset is large (n = 2,514), and most of the mutations are single point mutations, which are predominantly focused in the DNA binding region (codons ~130–300) of the *TP53* gene. Although there is an excess of G/C (~60%) versus A/T (~40%) in the base composition of this target region, the dominant strand-bias pattern nevertheless exists *within* A:T and G:C base pairs and is akin to the strand-bias pattern observed in the somatic hypermutation of Ig genes (3). Given that the proportion of A-to-T mutations and G-to-C mutations are similar in the mutation target areas, no corrections were made for base composition.

Figure 1 shows the defined region of interest around the mutated codon and the nomenclature used to identify the relative positions of each nucleotide. The region of interest includes nine nucleotides encompassing the mutated codon, the flanking 5′ codon, and the flanking 3′ codon. The respective positions of the nucleotides in the mutated codon (MC) sequence are annotated as MC-1, MC-2, and MC-3 (read 5′ to 3′). The respective positions of the nucleotides (N) in the flanking 5′ codon are annotated as 5′N1, 5′N2, and 5′N3, respectively (also read 5′ to 3′). Similarly, the positions of the nucleotides in the flanking 3′ codon are annotated as 3′N1, 3′N2, and 3′N3, respectively. In the example shown in Figure 1, a mutation of an A at an MC-1 site on the nontranscribed strand (NTS) is mutated to a C in the replicated nontranscribed strand (NTS′). The mutation of A in the mutated codon is associated with a G in the 5′N3 position. This is annotated as 'S •• A,' where S = G/C and the mutated codon A can be any one of the four nucleotides, A/G/C/T. This annotation is used regardless of the location of a mutation within the mutated codon. The codon structure and nucleotide sequence information in the region of interest for each mutation were added to the downloaded IARC *TP53* somatic mutation data.

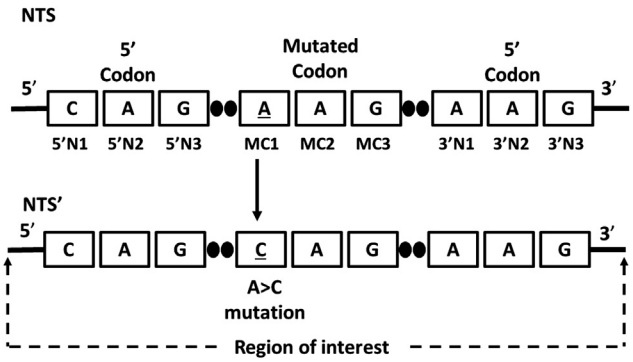

Figure 1 The defined region of interest flanking a point mutation. An example of a single point mutation in the region of interest and the annotations used to identify each of the nucleotides for this analysis is shown. The region of interest surrounding the MC includes the base composition of the flanking 5′ and 3′ codons. The nucleotide sites in the MC are annotated as MC-1, MC-2, and MC-3, respectively (read 5′ to 3′). Similarly, the positions of the nucleotides (N) in the flanking 5′ and 3′ codons are annotated as 5′N1 to 5′N3 and 3′N1 to 3′N3, respectively (also read 5′ to 3′). In the example shown for an A-to-C point mutation (A>C), a mutation of A at an MC-1 site on the NTS is mutated to a C in the replicated NTS (NTS′). The mutation of A in the mutated codon is associated with a G in the 5′N3 position. This is annotated as 'S •• A' (S = G/C). This annotation is used for mutations of A/G/C/T, regardless of the location of a mutation within the mutated codon (MC-1/MC-2/MC-3).

Table 1 shows the tabulated raw count of coincident occurrences for each mutation type against the key variables associated with codon context for the "All Breast Cancers" data set. Mutations with an unknown or missing base (entered as 'X' in the source data) within the region of interest are excluded.

The most studied motifs are associated with the cytosine deaminase activity of AID, APOBEC1, and APOBEC3G at G:C pairs. The motifs for these members of the AID/APOBEC family are identified using a study by Beale et al. that compared the targeting preferences for AID, APOBEC1, and APOBEC3G during transcription (4). Beale et al. showed that a target preference for AID agrees with the previously identified motif GYW/WRC, and this motif is included for analysis. For APOBEC1, the motif TG/CA is included, as 79% of transitions in the presence of APOBEC1 are associated with a 5′ T. For APOBEC3G, the motif CG/CG is selected as APOBEC3G shows a significant preference for transitions with C in the −1 (89%) position.

Another well-known mutation hotspot is associated with the WA motif, and it is included for analysis. These are potential sites for mutations at A:T base pairs (10). The "All Breast Cancers" data set in Table 1 was then used to select