
http://researchrepository.murdoch.edu.au/11427/

Copyright: © The Authors 2012. Creative Commons Attribution license
Critical analysis of strand-biased somatic mutation signatures in TP53 versus Ig genes, in genome-wide data and the etiology of cancer

Robyn A. Lindley¹ and Edward J. Steele¹,2*

¹ Melville Analytics Pty Ltd, Unit 14, Nuggets Crossing, Jindabyne, NSW 2627, AUSTRALIA
² School of Veterinary and Biomedical Sciences, Murdoch University, South Street, Murdoch, WA 6150, AUSTRALIA

Running head: Strand-Biased Mutations in TP53

Keywords: TP53; Ig somatic hypermutation; cancer genomes; C-to-U DNA deamination; A-to-I RNA editing; reverse transcription; 8oxoG mutations

*Correspondence: Associate Professor Edward J Steele; E-mail: e.j.steele@bigpond.com
ABSTRACT

Previous analyses of rearranged immunoglobulin (Ig) variable genes (VDJs) concluded that the mechanism of Ig somatic hypermutation (SHM) involves the Ig pre-mRNA acting as a copying template resulting in characteristic strand-biased somatic mutation patterns at A:T and G:C base pairs. We have since analysed cancer genome data and found the same mutation strand-biases, *in toto* or in part, in non-lymphoid cancers. Here we have analysed somatic mutations in a single well characterised gene TP53. Our goal is to understand the genesis of the strand-biased mutation patterns in TP53 - and in genome-wide data - that may arise by "endogenous" mechanisms as opposed to adduct-generated DNA-targeted strand-biased mutations caused by well characterised "external" carcinogenic influences in cigarette smoke, UV-light and certain dietary components. The underlying strand-biased mutation signatures in TP53, for many non-lymphoid cancers, bear a striking resemblance to the Ig SHM pattern. A similar pattern can be found in genome-wide somatic mutations in cancer genomes that have also mutated TP53. The analysis implies a role for base-modified RNA template intermediates coupled to reverse transcription in the genesis of many cancers. Thus Ig SHM may be inappropriately activated in many non-lymphoid tissues via hormonal and/or inflammation-related processes leading to cancer.
# TABLE OF CONTENTS

1. BACKGROUND
   - Caveat and writing strategy
   - Somatic Hypermutation in rearranged Ig Genes
   - Strand-Biased Mutation Signatures in Ig Genes
   - Origin of the A-to-G *versus* T-to-C Strand Bias

2. THE TP53 MUTATION DATABASE
   - Method of Data Extraction and Presentation
   - Statistics

3. TP53 STRAND-BIASED MUTATION DATA: BREAST, BLADDER AND LUNG CANCERS

4. TP53 G->T *VERSUS* C->A STRAND BIASES IN OTHER CANCERS

5. DNA REPAIR DEFICIENCIES: COLORECTAL, STOMACH AND SKIN CANCERS

6. TP53 MUTATION PATTERNS IN BRAIN CANCERS

7. CONTEXT OF C-TO-U LESIONS BY APOBEC-FAMILY ENZYMES

8. BREAST CANCER MUTATION PATTERNS IN TP53 COMPARED WITH PATTERNS IN GENOME-WIDE DATA

9. INFLAMMATION AND CARCINOGENESIS
A major goal of this paper is to provide an explanation of the origin of the main strand-biased mutation signatures observed in the TP53 tumor suppressor gene in the many tumors likely to arise by "endogenous" mutation processes: that is to say, those cancers not caused by well-known exogenous mechanisms such as exposure to carcinogens in tobacco smoke (Benzo(a)pyrene, G-to-T), toxins in food contamination (Aflatoxin B1, G-to-T; Aristolochic acid, A-to-T) or UV radiation in sun exposure causing DNA photoproducts such as cyclobutane pyrimidine dimers, C-to-T; reviewed in Soussi [1]. The TP53 mutation pattern in "All Breast Cancers" has been chosen as representative of the TP53 "endogenous pattern" as this mutation pattern appears to arise in a tissue "least accessible to carcinogens in tobacco smoke" or directly exposed to such exogenous carcinogens, see Hainaut and Pfeifer [2]. There is also a large number of TP53 point mutations in this tissue category (>1000), similar to the numbers in "All Lung Cancers", the major comparator in the analysis. Further, this basic "endogenous" pattern is evident in many tumors outside of lung, head, neck and oesophagus. All of these can be considered as "directly accessible to tobacco smoke carcinogens". This choice is made despite the known complexity of breast cancer in both aetiology and the diversity of histological subtypes [1] as a similar pattern is evident in mutated TP53 variants in "All Bladder
Cancers” which are likely to be more directly exposed to tobacco smoke-derived carcinogenic metabolites in urine.

Our analysis shows that the underlying “endogenous” strand-biased mutation signatures in TP53, for many different non-lymphoid cancers, bear a striking resemblance to the Ig SHM pattern. This allows inferences to be drawn about the mechanistic role of TP53-mediated DNA repair regulation and base-modified RNA template intermediates coupled to reverse transcription in the genesis of many cancers. It is also consistent with the view that the normally tightly regulated mutation processes targeting VDJ genes in B lymphocytes may, following further loss of DNA damage response regulation by TP53, be inappropriately turned on in non-lymphoid tissues, for example by hormonal and/or inflammation-related processes, leading to cancer.

**BACKGROUND**

**Caveat and writing strategy:** The writing strategy of this review is to provide as clear an introduction as possible concerning what is currently known of the immune system's somatic mutation mechanism as we believe that this mechanism is relevant to understanding the role of somatic mutations in the pathology of
cancer. This is an unexplored topic for most scientists, particularly in the field of cancer biology although it is very topical now given renewed interest in the regulation of inflammatory responses initiating both somatic mutations and thus cancer (see later). However there is a caveat to this analysis that should be highlighted right at the start: strand-biased mutation spectrums although very informative with strong inferential value with respect to molecular mechanisms provide little information about the initial events that precede malignant transformation as malignant cells grow rapidly and are exposed to strong selection. This means "first causes" cannot be precisely defined. Nevertheless the clear possibility that immunoglobulin somatic hypermutation may be one of these "first causes" promoting somatic mutation, both across the genome and in key gene regulators such as TP53, is worth pursuing in its own right as the implications have wide and interesting ramifications for the future directions of cancer research.

**Somatic Hypermutation in rearranged Ig Genes** : The mechanism of SHM of Ig VDJ genes is now well understood and many molecular steps are known or can be plausibly inferred [3-5]. More recently this knowledge has been applied to the etiology of cancer. What we discovered in a preliminary analysis was that the
characteristic strand-biased mutation signatures of Ig SHM were present, *in toto* or in part, in a number of somatic mutation datasets posted at the Welcome Trust Sanger Institute’s website run by the institute's Cancer Genome Project [6]. Indeed the possibility has often been discussed that dysregulation of SHM driven by activation induced cytidine deaminase (AID) conversion of cytosine to uracil (C-to-U) in DNA - normally confined to antigen-stimulated B lymphocytes in post-antigenic Germinal Centers - could lead to somatic mutations and translocations in non-Ig genes thus contributing to oncogenesis [7-10].

The aim here is to use these insights from the Ig SHM field to help explain the strand-biased mutation signatures in TP53 in human non-lymphoid tumors arising by as yet unknown "endogenous" mechanisms. We have analysed in detail strand-biased TP53 mutation signatures in breast, bladder and lung cancers. These are exemplars of oncogenesis in tissues exposed directly to carcinogens in tobacco smoke (lung) or indirectly exposed to such carcinogenic metabolites (bladder via urine) *versus* cancers arising in tissues such as breast generally considered "as least accessible to tobacco smoke" and other known exogenous agents [1,2,11].

We have also analysed and compared strand-biased mutation signatures and mutation patterns in other cancerous tissues. The analytical reviews by Soussi [1] and Pfeifer and Besaratinia [11]
have proved valuable and we recommend these papers and related literature be read in association with the present analysis.

**Strand-Biased Mutation Signatures in Ig Genes:** Strand biased mutation signatures in Ig VDJ gene loci, particularly at A:T base pairs, have been recognised for over 25 years [12]. More recently published data (1984-2008) on mutated mouse VDJ regions and their 3' JH4-flanks have been analysed [5]. In this study, somatic mutation data from thirty two independent studies are summarised and are available in a meta-analysis in the Supplementary Data. Little significant new data has been published in the SHM field since then that changes the basic patterns shown in Table 1 or their interpretation.

The essence of this analysis established that the Ig SHM mutation pattern free of "PCR recombinant artefacts" reveals several significant strand-biased mutation signatures at A:T and G:C base pairs (Table 1A). The first is at A:T base pairs where mutations of A exceed mutations of T by almost three fold, particularly the dominant strand bias of A-to-G exceeding T-to-C mutations. The second main strand bias is at G:C base pairs where mutations of G exceed mutations of C by at least 1.7 fold. The dominant strand bias here concerns G-to-A exceeding C-to-T mutations.
The distortion in the DNA sequence data contaminated with PCR-recombinant artefacts (Table 1B) has previously masked the clear strand bias at G:C base pairs. This distortion has also made it difficult for the field to develop their understanding of the mutator mechanisms operative on A:T and G:C base pairs during Ig SHM in vivo [5,13].

The significant presence of PCR-recombinants - also referred to as PCR-hybrids or mosaic heteroduplexes at the end of a PCR run is due to Taq or Pfu polymerase denaturation generating incomplete extension products acting as forward and reverse primers during amplification cycles. They arise from PCR runs where amplification is from multiple similar templates using the same set of primers. For this reason, the inevitable presence of such sequence artefacts causes blunting if not complete ablation of strand-biased mutation signatures following cloning in E.coli and then sequencing of PCR inserts. Such problems are avoided by reducing PCR cycle numbers, by sequencing target VDJ genes expressed in hybridoma clones or direct sequencing of amplification products from single VDJ loci expressed in FACS sorted single B lymphocytes [5,13-15].

As a consequence of these analyses the reference strand-biased Ig SHM mutation pattern we will use as a comparator for the TP53 mutation data is shown in Table 1A. The patterns in Table 1A are essentially free of strand-biased blunting PCR hybrid artefacts.
For comparison, Table 1B shows the previously cited pattern in the SHM field where there is no strand-bias evident at G:C base pairs. Note that in Table 1A the magnitude of the ratio of mutations of A versus mutations of T (hereafter indicated as the A>>T ratio) is reduced from 2.8x to 1.9x in Table 1B. Thus there is significant blunting of the established A>>T mutation ratio (as well as the dominant and diagnostic A-to-G versus T-to-C ratio, below) and compete ablation of the lower (1.7x) yet significant G>>C strand biased ratio (compare Tables 1A and 1B).

In summary, we now know that the SHM reference pattern shown in Table 1A is characterised by significant strand biases evident for all Watson-Crick complements: A-to-T v T-to-A; A-to-C v T-to-G; A-to-G v T-to-C and G-to-A v C-to-T; G-to-T v C-to-A; G-to-C v C-to-G.

We first addressed this issue [13] by making the point that "the synthesis of a mutated cDNA copy of the transcribed strand (TS) off the pre-mRNA template, and replacement of the original TS with the cDNA is inevitably strand-biased (see Fig 1)." This was underpinned by the finding that the error-prone Y family DNA polymerase-eta (η), an enzyme shown to be at least involved in translesion DNA repair, reviewed in Goodman [16], was an efficient reverse transcriptase at low enzyme-to-RNA template ratios in vitro [17]. It is now firmly
established that Pol-η is the only DNA polymerase involved in physiological Ig SHM \textit{in vivo} \cite{18,19}.

The SHM field now accepts that Pol-η mutates A:T base pairs, particularly A-sites at certain WA hotspots where the target A is preceded 5’ by A or T (=W). With the analysis of most published experimental data 1984-2008 (see legend Table 1) a unifying explanation can be provided for the central role of base-modified RNA template intermediates and cellular reverse transcription in the generation of \textit{all} the Watson-Crick strand biases displayed in Table 1A.

In the updated version of the reverse transcriptase model (RT model SHM, initially proposed in Steele and Pollard \cite{21}) the two different sets of mutation strand biases at A:T and G:C base pairs can, we believe, be explained by a common core mechanism (Figure 1) : emergence of an error-filled mRNA intermediate followed by reverse transcription via DNA polymerase-η \cite{5}. In the case of the A>>T strand bias the model proposes a combination of adenosine to inosine (A-to-I) pre-mRNA editing by ADAR1 deaminase \cite{24} and A-to-T and A-to-C biases via the RT activity of Pol-η during the cDNA synthesis step. In the case of G>>C strand bias it proposes that RNA mutations (G-to-A, G-to-C) generated by RNA Polymerase II (RNAPII), transcribing a DNA template (TS) carrying AID-lesions (uracils and abasic sites), followed then by the RT activity of Pol-η.
This can be considered a form of "transcriptional mutagenesis" as proposed in Fig 3b in Hanawalt and Spivak [25] but coupled now to DNA fixation by reverse transcription.

According to the RT-model, for the G>>C mutation signatures the G-to-A v C-to-T strand bias results from rA being incorporated into RNA opposite unrepaired dU on the TS and the G-to-C v C-to-G strand bias results from rC being incorporated into RNA opposite an abasic site on the TS [23]. For the G-to-T v C-to-A strand bias this would be the alternative pyrimidine substitution (rU) if rC is not inserted opposite an abasic site (Note : modified G residues in DNA due to reactive oxygen species such as 8oxoG are not known to play a role in physiological SHM in vivo [20] and below).

Recently, we have applied this updated RT-model for DNA diversification to both the etiology of strand-biased mutation patterns in many non-lymphoid cancers [6] and, amongst other hypotheses, to the origin of the established genome-wide strand bias for A-to-G over T-to-C in transcribed regions of the human genome [26]. The RNA modifications we are considering are non-bulky simple changes to base pairing, such as adenosine-to-inosine deamination in RNA [24] or putative 8oxoG modifications at G residues in RNA [6]. For Ig SHM and in our genome-wide diversification analysis, we have strongly argued against conventional explanations of strand-biases (mainly the A-to-G v T-to-C bias) for repair of non-bulky DNA lesions
caused by differential DNA repair of transcribed (TS) as opposed to non-transcribed (NTS) strands during transcription-coupled repair (TCR). Indeed, critical evaluation of the TCR field has so far not provided evidence to support a TCR-mediated mechanism for strand biases arising from non-bulky DNA lesions such as C-to-U and abasic sites [26]. Further, in the case of the repair of 8oxoG lesions in DNA the Bohr group have convincingly shown that there is no transcriptional strand bias in their repair [27].

**Origin of the A-to-G versus T-to-C Strand Bias:** The A-to-G versus T-to-C strand bias is a common strand bias in many somatic and germline mutation data sets. This strand bias is found not only in all SHM data sets in mouse and human VDJ genes, in families of similar human germline IgV segments (in Matsuda et al [28], EJS unpublished analysis) but also in almost all TP53 cancer data sets where A/T mutations have not been significantly suppressed or ablated (presumably by genetic deficiencies affecting the mismatch repair (MMR) machinery as in colorectal, stomach, oesophagus adenocarcinomas, skin, rectum and colon cancers, below). And in all cases examined, A-to-G mutations are enriched at some but not all A-site hotspots where the A target is preceded by a 5' A or T (=W).

Key evidence supporting an RNA template intermediate model for the prominent A-to-G versus T-to-C mutation strand bias derives
from an IgV mRNA-stem loop computational analysis where the RNA substrate for ADAR1 mediated A-to-I deamination was modelled and tested on the somatic mutation data set of the rearranged light chain encoding V\(\kappa\)Ox1 passenger transgene [24]. Thus, in an RNA-based pathway for immunoglobulin SHM, A-to-I RNA editing causes A-to-G transitions since I like G pairs with C. The adenosine deaminases (ADARs) are known to preferentially edit A sites that are preceded by an A or U (W) in double-stranded RNA substrates [29]). We showed that a significant and specific Pearson correlation (\(P<0.002\)) exists between the frequency of WA-to-WG mutations and the number of mRNA hairpins that could potentially form at the mutation site. Indeed the statistical significance of the correlation improved with increased stem length (or stability) of the dsRNA substrate (Fig 11, in ref. [24]) and proximity of the nascent dsRNA to the transcription bubble. It is known that ADAR1 edits pre-mRNAs in the nucleus prior to splicing [30,31]. Indeed ADAR1 seems to act on the WA-site closest to the transcription bubble and explains why the A-stem partner in the target A:U editing site must be previously synthesised [24]. This study strongly implies a role for both RNA editing and reverse transcription during SHM \textit{in vivo} involving ADAR1 and Pol-\(\eta\) acting in its RT mode. For these reasons, we consider the elevated A-to-G \(v\) T-to-C ratio as a diagnostic for mutational strand bias.
caused by modified RNA template intermediates and DNA fixation via reverse transcription.

However key direct experimental evidence supporting an RNA template intermediate model is still lacking. The ideal experiment in the context of Ig SHM would be a conditional genetics approach targeting ADAR1 expression in mature B lymphocytes in antigen-activated Germinal Centers. In a collaboration Cre-lox specific gene targeting techniques were used to inactivate ADAR1 during SHM in vivo. A positive result might involve a clear reduction or complete removal of the A-to-G component of the SHM mutation spectrum (that is, those A-to-G changes which correlate strongly with WA sites in dsRNA stem loops). If however ADAR1 is a more central player in the SHM process it may also result in a total reduction in mutations at A:T base pairs (leaving intact mutations at G:C base pairs). In the recent collaboration ADAR1flox alleles on the C57BL6 mouse background, Wang et al 2004 [32]) were crossed into C57BL6 mice with a ‘knocked-in’ Ig antigen receptor (the SWHEL IgVH10 single-copy heavy chain transgene) which was assayed for somatic hypermutation in the adoptive transfer system described in Paus et al 2006 [33]. An inducible Cre-recombinase gene when activated by tamoxifen should specifically target the ADAR1floxed alleles and delete them from B lymphocytes activated by antigen into the somatic hypermutation pathway. Unfortunately no mature donor B
lymphocytes could be recovered in Germinal Centers suggesting that one or more ADAR1 sensitive developmental steps were necessary leading to Germinal Center B lymphocytes. Therefore with current Cre-lox technology approaches to implementing a successful experiment targeting ADAR1 alleles seem limited (R. Brink, K. Nishikura, G. F. Weiller, E. J. Steele unpublished data 2007-2009).

With respect to carcinogenesis, unregulated ADAR-mediated A-to-I RNA editing is a well described phenomenon [34,35]. In a similar vein, unregulated APOBEC family C-to-U DNA deaminases such as AID, APOBEC3G and APOBEC1 are comparable rogue mutator processes thought to be operative in many cancers [7-10]. Thus for the present analysis we can reasonably assume that unregulated RNA and DNA deamination processes (at rA and dC residues) may well be associated with either the genesis or progression of many non-lymphoid cancers. Here we are concerned with the molecular implications of such processes particularly in relation to understanding the strand-biased mutation signatures at A:T and G:C base pairs in the TP53 gene and wider genome.

**THE TP53 MUTATION DATABASE**

The DNA sequence encoding the human tumor suppressor protein TP53, on chromosome 17 located at 17p13.1, has been cloned and
sequenced as both full length DNA and cDNA in many tumors over the past two decades. Mutated variants of the TP53 germline sequence carrying somatic mutations mainly in the region encoding DNA binding, are found in a wide range of cancers [1,11,36]. Oncogenic TP53 mutations are a biased dataset in that they are partially selected for a competitive binding function focused on the DNA binding region. They are usually missense mutations in one allele spanning many sites in the TP53 coding DNA from about codon 130 to 300 [1]. Many investigators have deposited their sequence data in the database funded by the WHO at Lyon in France. The WHO-IARC public database has now curated around 30,000 somatic mutations in TP53. The data analysed here was extracted from this source [37] (http://www-p53.iarc.fr/, R15, November 2010).

**Method of Data Extraction and Presentation**: The somatic point mutation data presented in the Tables were extracted from the database as follows. On entering the website (http://www-p53.iarc.fr/) "Database Search" is selected allowing entry to http://www-p53.iarc.fr/p53main.html which allows selection of "Search database for data related to SOMATIC MUTATIONS" and selection of "Search (Tumor types)". This allows entry to http://www-p53.iarc.fr/BasicCriteria.asp where the tumor site can be selected at " Select a tumor site" and thus entry to "Mutation pattern" ( at
http://www-p53.iarc.fr/Graph.asp) where selection can be made for the key data sets: "Strand distribution" and "Download data". The "Strand distribution" tables can be downloaded where numbers of all 12 possible base substitutions are displayed including the numbers of C-to-T and G-to-A mutations at CpG islands. The spreadsheet from "Download data" allows construction and analysis of all types of mutations with 5' and 3' flanking sequence context in relation to the unmutilated TP53 exon sequence (and in some cases intronic sequence). This allows development of frequency distributions of various types of mutation (e.g. A-to-G) versus nucleotide (and codon) position across regions of interest such as the DNA binding region eg Figure 2.

**Statistics**: Displayed in each base substitution table are Chi-squared statistical comparisons (1 df) of several types of base substitutions. Thus for A:T base pairs the main comparisons are all mutations of A versus all mutations of T - when strand biased for excessive mutations of A this is symbolized as "A>>T". The common and dominant A-to-G strand bias is represented as "A>G v T>C". Other types of biases are presented and tested similarly. Thus when mutations of G exceed mutations of C this is symbolized as "G>>C"; when G-to-A mutations exceed C-to-T mutations this is symbolized as
The Appendix explains the rationale for detecting strand biases in mutation datasets.

**TP53 STRAND-BIASED MUTATION DATA**

The strand-biased mutation pattern typical of normal Ig SHM (Table 1A) was compared with the somatic point mutation patterns observed in the TP53 coding region for a range of key cancers (Table 2). Chi-squared tests were applied to test the levels of statistical significance of the various strand biases. Attention was focused on mutations of A and G respectively, particularly A-to-G versus T-to-C, G-to-A versus C-to-T and G-to-T versus C-to-A as the analysis of these strand biases has implications for the molecular mechanisms involved.

The main strand-biased mutation patterns observed in TP53 are best represented by the patterns in All Breast, All Bladder and All Lung cancer categories - where each has many (>1000) somatic mutations to analyse (Table 2). We have ranked these tissues in their likely exposure to carcinogens in tobacco smoke as smoking is the leading cause of lung and many other cancers in the world today. The origin of TP53 mutations in breast may be considered the least likely to be caused by direct exposure to tobacco smoke carcinogens and metabolic by-products [2].
In the case of TP53 mutations in bladder cancer we assume the tissues are at least exposed to relatively high levels of carcinogenic metabolic by-products of tobacco smoke in urine.

In the case of TP53 mutations in lung cancer there is direct tissue exposure to tobacco smoke polycyclic hydrocarbons (PAHs) and carcinogenic metabolic derivatives [38]. Here the molecular, biochemical and cellular evidence is overwhelming. Exposure to such smoke-derived carcinogens causes lung cancer. In particular, bulky DNA adducts at certain G sites targeted by carcinogens such as benzo[a]pyrene (B[a]P) is the direct cause of the dominant G-to-T transversion in these cancers in the TP53 gene [39-41] (reviewed in [1,11] ) and throughout the wider lung cancer genome [42].

Some striking similarities are observed for the cancers shown in Table 2.

a. The first is the similarity between the Ig mutation pattern (Table 1A), and the patterns in TP53 for "All Breast " and "All Bladder" cancers. The main difference between the Ig pattern and the TP53 pattern in breast and bladder (and with many other TP53 cancer data sets for that matter) is the ~ 50:50 balance of mutations at A/T and G/C for Ig versus the G/C excess over A/T in TP53. The majority (70-75%) of TP53 mutations occur at G/C sites. This may be partly contributed by the significant excess of G/C (~ 60%) versus A/T (~40%) base composition of the target region in TP53. But the
pattern similarities nevertheless exist within A:T base pairs and G:C base pairs.

b. The second systematic pattern observed in all data sets shown in Table 2 is the A>T and A-to-G versus T-to-C strand bias. They stand out as common strand-biased patterns. To our knowledge of the mainstream TP53 literature this strand bias is rarely highlighted in published discussions. However, this pattern is not observed in colorectal, stomach, skin, and some other cancers as A/T mutations here have been ablated or significantly reduced (below). This pattern is stable across breast, through bladder to lung (and all ovary cancers, Table 3) suggesting a common causal mechanism that may not be associated with exposure to carcinogens in tobacco smoke. For A-to-G hotspots in both breast and lung cancers, the majority are defined by being part of a WA-site, particularly at codons 132, 163, 205, 220, 234, 239; the TAT site in codon 220 is a super hot spot (data not shown).

If the A-to-G spectrum in TP53 at all A-sites in codons 100-300 inclusive for "All Breast Cancer" is compared with the same spectrum in "All Lung Cancer", they are virtually super imposable (Figure 2). The Pearson correlation coefficient (r) is 0.93, which for 129 degrees of freedom gives a P<0.001. Indeed, we think that this repeatable pattern in the two disparate target tissues is consistent with their origin being the result of an "endogenous" process. By comparison
with what we have inferred from Ig SHM (Figure 1) a likely candidate is unregulated ADAR1-mediated A-to-I RNA editing and fixing of the mutated retro-transcripts back into DNA via reverse transcription as envisaged for Ig somatic hypermutation. The cellular reverse transcriptase could be DNA polymerase-$\eta$ (or one of it's Y family relatives iota ($\iota$) and kappa ($\kappa$)) which also possess significant reverse transcriptase activity, [17]).

c. The third common pattern is the excess of G mutations over C mutations (G$\gg$C), particularly the dominant G-to-A over C-to-T strand bias (evident at both CpG and non-CpG sites). The pattern in breast and bladder is very similar again, suggesting common causal mechanisms. All ovary cancers display a similar pattern (Table 3). Again, to our knowledge of the mainstream TP53 literature, this particular and striking strand bias is rarely highlighted in published discussions on the topic.

The likely cause of the G$\gg$C imbalance for lung cancer is known to be due to the binding of bulky tobacco smoke-derived adducts such as B[a]P at certain G-sites (mainly CpG islands although GpG sites can be targeted) in known critical codons such as 154, 157, 158, 245, 248, 249, 273. Such adducted G-sites now mispair with adenosines causing G-to-T transversion mutations if left unrepaired. This grossly imbalances other G mutations at these sites.
leading to a loss of the G-to-A v C-to-T strand bias evident in breast, bladder and ovary cancers.

Direct experiments by Pfeiffer and colleagues have shown that the G-to-T v C-to-A strand bias is caused by the much slower repair of bulky DNA adducts such as B[a]P along the nontranscribed strand compared with the faster repair on the transcribed strand [41]. So the strand biased G-to-T pattern is a direct consequence of a transcription-coupled DNA repair (TCR) process for bulky DNA adducts [25]. This is a DNA-based strand-biased mutation mechanism and thus quite different from the RNA-based mechanisms outlined above for Ig SHM (Figure 1).

Before leaving this section we wish to deal with two further issues. First, we must deal with a relevant point that has now emerged from whole genome sequencing of individual lung cancers such as NCS-H209, Pleasance et al [42]. In a critical section on the origin of the DNA repair pathways that may be responsible for the complex strand biased signatures of tobacco exposure the authors make the following set of statements and assumptions:

"... that bulky adducts on purines are the predominant form of DNA damage induced by tobacco carcinogens and can be sufficiently disruptive to impede RNA polymerase when they occur on the transcribed strand.." and they observed " that guanine and adenine substitutions are generally less frequent
on the transcribed than the non-transcribed strand - confirming that purines seem to be the major target of carcinogens in tobacco smoke."

We accept this explanation for the origin of the G-to-T transversions but our data do not support their conclusions on the origin of the A-to-G strand bias. Indeed we would hardly expect the A-to-G spectra in TP53 to be identical in lung and breast cancers (Figure 2) if this is the case. The fitted curves in Pleasance et al [42] show the effect of gene expression on strand bias mutation rate for the six classes of adenine and guanine mutations in NCI-H209 (Fig2e in [42]). The overall patterns are to be expected from the TP53 lung cancer pattern (Table 2C) which the authors acknowledge in their Supplementary data. However the profiles for G-to-T and A-to-G are quite different. The decline in G-to-T mutation rate with increased transcription level is biphasic suggesting two causal DNA repair mechanisms for G-to-T: one to be expected and rapid depending on increased transcription involving TCR of bulky adducts, the other suggestive of another strand bias process. This could be due to perhaps 8oxoG generation by reactive oxygen species in RNA and an RT step of RNA->DNA fixation.

The curve for A-to-G mutation rate versus transcription level appears monophasic with the difference between the repair of the NTS and TS mutations deepening with increased gene expression (or
transcription). However the slope of the curve is very shallow - a slight decline in mutation rate from expression level 4 to 9 (\(\sim 2.9\) mutation rate/Mb to \(\sim 2.5\) mutation rate/Mb). This suggests a constitutive process of error generation and repair marginally affected by transcription level. One interpretation could be that this is consistent with A-to-I RNA editing, and RNA-\(\rightarrow\)DNA fixation.

The second issue adds to the argument against asymmetrical or strand-differential TCR as an explanation for the excess of G-to-A mutations over C-to-T mutations in both Ig SHM and somatic mutation patterns in TP53 in cancers such as in breast. If TCR was occurring to clear C-to-U lesions arising from the action of unregulated APOBEC family deaminases in cancer cells (or progenitors) we would expect a preferential clearance of C-to-U on the TS and an excess of unrepaired C-to-U lesions, manifest as C-to-T, on the NTS. In fact in Ig SHM and in mutated TP53 genes in breast and other cancers (Table 2) it is the other way around. An excess of G-to-A over C-to-T suggests that C-to-U lesions would need to go unrepaired preferentially on the TS which is not evident in the mutation patterns analysed here.

So the argument goes full circle posing the question why then should G-to-A mutations exceed C-to-T mutations? This conundrum may explain why the G-to-A >> C-to-T strand bias has gone relatively unreported in the literature. The simplest explanation of
the G-to-A strand bias is that C-to-U goes unrepaird on the TS prior to RNA Pol II copying this into G-to-A in the mRNA [23] which in turn would be manifest in a strand-biased manner in the DNA by reverse transcription first as a C-to-T in TS DNA and then following replication as G-to-A on the NTS [5] (and see Figure 1). This is the simplest explanation of the G-to-A versus C-to-T strand bias involving AID/APOBEC3G/APOBEC1-deaminations in DNA. The same argument for G-to-C over C-to-G in SHM and TP53 mutation patterns in breast and other cancers with similar mutation patterns (eg bladder Table 2B) is used, and the lesser strand bias of G-to-T over C-to-A. In the TP53 mutation pattern in breast cancers the relative increase in the G-to-T over C-to-A ratio compared to that in Ig SHM (Table 1A) - in "a tissue less exposed to smoke" - suggests it might be contributed by another mechanism. One possibility is the formation of 8oxoG in the TP53 mRNA (as suggested in Steele and Lindley [6]) because it is known that 8oxoG DNA lesions are unlikely to display strand differential TCR biases on DNA repair [27]. Further research on the genetic and biochemical consequences of 8oxoG formation in RNA is clearly required.

**TP53 G->T VERSUS C->A STRAND BIASES IN OTHER CANCERS**
B[a]P and similar DNA-binding tobacco-derived carcinogens are known to form bulky G-site adducts causing G to base pair like T with adenosines. These bulky adducts are preferentially cleared from the transcribed strand during transcription-couple repair and this causes the excess of G-to-T mutations on the upper non-transcribed strand. In a similar vein, in hepatocellular liver cancer the G-to-T versus C-to-A strand bias is caused by adduct formation at G-sites by aflatoxin dietary contaminants, diagnostically at the third position of codon 249 in TP53 [36]. This G-site (GpG) is also a G-to-T hotspot in many lung cancers.

Other geographically and ethnically localised strand biases via dietary contamination for G-to-T over C-to-A, and for A-to-T over T-to-A are reviewed in [1,36]. Thus G-to-T transversions in the third base of TP53 codon 249 correlates strongly in tumors from HBV carriers and exposure to dietary aflatoxinB1; and A-to-T strand bias (codons 131,209,280) has been linked with crops contaminated with Aristolochia sp seeds (in certain Balkan communities in southeastern Europe). Aristolochic acids (AA) is the identified carcinogenic agent, and DNA adducts involving AA have been detected in patients suffering from Balkan endemic nephropathy (BEN). The strand bias suggests preferential TCR of bulky DNA lesions along the transcribed DNA strand as concluded already for G-to-T strand biases in many lung cancers.
DNA REPAIR DEFICIENCIES: COLORECTAL, STOMACH AND SKIN CANCERS

There are also some other more complex TP53 mutation patterns not conforming to the simple strand biases just discussed and worthy of further comment here. Before we do this, it is informative to consider what is known about DNA repair deficiencies and distortions of the Ig SHM mutation pattern (Table 1 and Table 4). Much of this work has been done using single and double knockout mice targeting key base excision repair (BER) and mismatch repair (MMR) genes encoding proteins that have been co-opted to now function aberrantly (a form of "subverted DNA repair" as put by Martomo and Gearhart [52]), see summary in Table 4 and review in Steele [5]. The various protein components of the normally tightly regulated and targeted Ig mutator now act to encourage error-prone DNA synthesis during the somatic hypermutation of rearranged antibody variable genes in Germinal Center B lymphocytes following antigenic challenge. The key proteins and DNA repair enzymes are AID deaminase which initiates the SHM process (and Ig class switch recombination, CSR) by deaminating C-to-U in VDJ DNA; this is then followed by attempts to remove the base in a base excision repair (BER) manner via uracil DNA glycosylase (UNG), followed by other "subverted" DNA repair
enzymes such as translesion DNA polymerase η, and the mismatch repair heterodimer MSH2-MSH6.

Table 4 lists the main consequences on the Ig SHM mutation spectrum of genetic deficiency in uracil DNA-glycosylase (UNG), in the mismatch repair heterodimer (MSH2-MSH6), and deficiencies in Y family DNA polymerases η (eta), i (iota) and κ (kappa) and combinations thereof. Additional deficiencies are shown: in alkyladenine DNA glycosylase (Aag) which removes hypoxanthine (deaminated adenine) from DNA generating an abasic site and 8-hydroxyguanaine-DNA glycosylase (Ogg1) which removes oxidised guanine from DNA. The affect on the SHM spectrum of inactivating TP53 is also shown.

In Ig SHM a failure to remove uracils from DNA as a consequence of dC-to-dU AID deaminase action (UNG-/-, Table 4) has a slight affect on overall A/T mutations and no effect on A>>T or A-to-G strand biases. The main effect is on the focusing of mutations to G:C base pairs with a reduction in transversion mutations.

In UNG-/--MSH-/- double deficient mice mutations at A:T base pairs are virtually eliminated as are transversions at G:C base pairs leaving what is considered the "AID deamination footprint" (Table 4, [44]) which now manifests itself as a strong strand bias of C-to-T exceeding G-to-A by at least 1.5 fold [5]. The simplest interpretation is that since AID-deaminase converts C-to-U in the single stranded
DNA regions of the transcription bubble, this is likely to happen more often on the NTS than the TS, thus the unfettered C-to-T over G-to-A strand bias in such mutant mice. The same strand bias is revealed in double deficient Polη-/-/MSH2-/- mice [18]. These data are consistent with SHM models whereby MSH2-MSH6 heterodimers engage G:U DNA mispairs and recruit Pol η necessary for full blown mutagenesis at A:T base pairs [53]. In the complete absence of Pol η, Pol-κ can step in to affect A/T mutagenesis [19] and probably also the RT step [17].

With respect to the general over-arching molecular mechanism of Ig SHM the deficiencies in Aag, Ogg1 and TP53 are informative. First, the failure to remove 8oxoG from DNA has no affect on the Ig SHM spectrum [20] indicating that the G-to-T/C-to-A component of the SHM spectrum does not involve 8oxoG residues in DNA. This leaves open the possibility that 8oxoG sites in RNA may contribute in other somatic mutation scenarios such as in TP53 as envisaged in Figure 1. Second, direct deamination of adenines in DNA to hypoxanthine (and thus potential A-to-G miscoding) seems to play no role in the generation of the A-to-G spectrum [48] once again leaving open the possibility of adenosine deamination to inosine at the RNA level contributing significantly to the observed A-to-G strand bias as envisaged in Figure 1. The investigators also observed a borderline statistically significant (P~0.03) increase in T-to-C in Aag (-/-) mice -
in our view this variation in T-to-C frequency is well within the range of variation for these PCR based SHM assays as outlined in Steele [5] and supplementary data therein. Third, and most importantly for the present analysis, Storb and associates have clearly shown an affect of TP53 inactivation on the Ig SHM spectrum - there is a striking increase in A-to-G frequency in such mice and a corresponding increase in both A-to-G versus T-to-C and A->T strand bias [49]. This result suggests that TP53 inactivation can profoundly affect the A-to-G component of the SHM spectrum thus implies that the global DNA damage surveillance function of TP53 extends, as expected [49], to Ig SHM. In the context of the present analysis the result suggests that TP53 may well regulate the imprint of A-to-I RNA editing on the DNA somatic mutation pattern as predicted by the RT model of Ig SHM (Figure 1). This in turn has implications for the magnitude of the A->T and A-to-G versus T-to-C strand biases across the human cancers bearing inactivated TP53 alleles such as for example, bladder (Table 2B) and ovary cancers (Table 3) and the wider cancer genome.

With respect to the other major affects on A/T and G/C mutations can we find parallel, or similar, patterns in other cancers with TP53 somatic mutation data? It is well known that colorectal and other aggressive gastrointestinal cancers are typified by the known high incidence of defects in mismatch DNA repair machinery, Bellizzi
and Frankel 2009 [54]. We would therefore expect, if subverted DNA repair components of the Ig SHM process are operative in such tumors, that the signature of ablated or suppressed A/T mutagenesis should be revealed in the TP53 patterns in such cancers. This expectation is partly satisfied by the TP53 mutation data on colorectal and stomach cancers (Table 5) where A/T mutations have been reduced (but not eliminated) in colorectal and stomach cancers. In some cases the A-to-G strand bias has been retained (stomach Table 5B) and in other cases it is lost (colorectal cancers, Table 5A). Reductions in A/T mutations are also noted in esophageus adenocarcinomas (not shown). Whilst there is a relative excess of G-to-A and C-to-T mutations at presumed methylated CpG sites in colorectal cancers, the strand bias here is also evident at non-CpG sites (not shown). In contrast, the TP53 mutation patterns in stomach cancers lack significant strand biases apart from those involving A-to-G v T-to-C and G-to-T v C-to-A (Table 5B). It is conceivable that unrepaired excessive C-to-U deaminations on the NTS (relative to TS) are blunting intrinsic strand biases of G-to-A v C-to-T in the same way such blunting (or strand bias reversal) occurs in Ig SHM in UNG-/MSH-/- and Pol η-/MSH-/- mice (Table 4, and see extended discussion on this point in Steele [5]).

TP53 mutation patterns in skin cancers are shown in Table 5C. Here there is both suppression of A/T mutagenesis and a reversal of
the strand bias at G/C sites, namely, C-to-T mutations clearly exceed G-to-A mutations by almost 2 fold. This is similar to data for Ig SHM in mice displaying the "AID deamination footprint" in UNG-/-/MSH-/- and Pol η-/-/MSH-/- mice (Table 4, [18,44]).

In summary, known DNA repair deficiencies in previously characterised Ig SHM model systems are displayed in toto or in part in TP53-bearing cancer mutation patterns. The results are consistent with the hypothesis that components of "subverted DNA repair" play a similar role in both SHM and non-lymphoid cancers bearing mutated TP53 derivatives. In addition inactivation of TP53 may increase the magnitude of A>>T and A-to-G versus T-to-C strand biases in the tumors that harbor mutated TP53 derivatives [49].

**TP53 MUTATION PATTERNS IN BRAIN CANCERS**

This particular pattern is discussed at length in Soussi [1] and displayed in Table 6. Note that a trend to A>>T strand bias, and a significant A-to-G v T-to-C strand bias is evident. There is also a specific strand bias of G-to-T transversions over C-to-A. However the global G>>C strand bias has been ablated, particularly the dominance of mutations of G-to-A v C-to-T evident in many other cancers harboring mutated TP53 variants. As Soussi [1] discusses, there are similar numbers of excessive G-to-A and C-to-T mutations
at CpG sites suggesting excessive deaminations of 5-Methylcytosine-to-T on both DNA strands (Table 6). The latter may be affected by AID-deaminase [9]. Once again however, as with the analysis of mutation patterns in Ig SHM, competing strand-biased mutational processes at G:C base pairs may accentuate, blunt, ablate or even reverse the strand biased signature presented by a particular tumor.

**CONTEXT OF C-TO-U LESIONS BY APOBEC-FAMILY ENZYMES**

This topic has been extensively studied by the Neuberger group [8]. Here we summarise the main findings and relate them to the sequence context of TP53 mutations in lung and breast cancer. For mutations off G most if not all occur at one of the known 5'NÇN3' motifs on the opposite deamination strand. Thus APOBEC1 targets 5'-TÇA-3' on the deamination strand (or 5'-TGA-3' on the NTS); APOBEC3G targets 5'-CÇG-3' (or 5'-CGG-3 on the NTS) and AID variously targets in descending order 5'-AÇA-3', 5'-GÇA-3', 5'-ACG-3' and 5'-GÇG-3' (or 5'-TGT-3', 5'-TGC-3', 5'-CGT-3' and 5'-CGC-3' on the NTS). We also include the possibility that 5-MeC-to-T deaminations are affected by such enzymes [9].
When this information is applied to the G-site mutations of the "endogenous pattern" represented by the All Breast Cancer data it reveals that all major and minor G-site mutation hot spots can be classed as the direct result of either AID or APOBEC3G C-to-U deaminations targeting the transcribed TP53 DNA strand (Table 7). The dominant likely C-to-U deaminase is AID suggesting that dysregulated SHM initiation via AID activation in non-lymphoid tissue may be the primary cause of the mutagenesis leading to cancer.

BREAST CANCER MUTATION PATTERNS IN TP53 COMPARED WITH PATTERNS IN GENOME-WIDE DATA

The somatic mutation patterns in TP53 is often a very good correlate to genome-wide point mutation patterns eg. lung cancer [42]. However this is not always the case probably because TP53 is not inactivated in all cancers of a given category e.g. about 25% of breast and bladder cancers, 48% of ovary cancers and 38% of lung cancers have an inactivated TP53 allele (see IARC TP53 database). As pointed out in Pfeifer and Besaratinia [11] large-scale genome sequencing of cancer genomes has revealed some interesting results not evident in TP53 patterns [55,56]. Thus there is a quantitative difference evident between strand-biased mutation patterns in TP53 in breast cancers (Table 2A) compared with available data from
genome-wide exome sequencing of breast cancer genomes [55,56]. In Table 8 are displayed data illustrating this difference [55] from the sequencing of exons of close to 20,000 protein coding genes in eleven breast cancer genomes. In the majority of these breast cancers (10/11) TP53 is mutated (see supp data in [55]).

Note first, the strand biased pattern at A:T base pairs in TP53 (Table 2A) and in genome-wide data (Table 8) are similar. The systematic A>>T and prominent A-to-G strand biases are evident. However in this data set (~1445 point mutations) the strand biases at G:C for G>>C is systematic and just significant at the $P<0.05$ level. A key difference pointed out by Pfeifer and Besaratinia [11] is the higher load of mutations for G-to-C/C-to-G. Pfeifer and Besaratinia postulate:

"These data suggest that breast cancers are caused by an etiological agent that induces this particular type of mutation. There are few known mutagens that specifically induce G/C to C/G transversions, let alone selectively at a particular dinucleotide sequence."

They then go on to point out that a significant fraction of these G to C transversions occur at the 5' GpA dinucleotide motif which is 5' TpC on the other strand.

From our perspective it is interesting that this happens to be the favoured APOBEC1 motif if this DNA deaminase enzyme, now
unregulated, deaminated cytosines at such sites ([8] and see previous section, Table 7). Inspection of the TP53 "All Breast Cancer" data reveals there are indeed several 5'GpA sites which account for about a third of the load of G-to-C mutations in the strand-biased G-to-C pattern in TP53 in cancers of the breast (Table 2A). These G-to-C hotspots are at codons 196, 280 and 281 (Table 7) and constitute 25 of 92 (27%) of all G-to-C mutations over codons 150-300 inclusive. Additionally, ten G-to-C mutations in this region occur at a CGC site in codon 156 (a motif favoured by AID deamination on the opposite strand).

Recently the Cancer Genome Project (CGP) at The Welcome Trust Sanger Institute have reported on the exomic mutation spectrum of category-selected sets of 21 breast cancer genomes, Nik-Zainal et al [57]. Few if any significant strand biases are reported in this genome-wide data. Of real interest is the fact that only a minor fraction (4/21) carry an exomic mutation in TP53 and most sample sizes for mutations are statistically small (N values in the hundreds) - except tumor PD4120a which carries 1931 exomic mutations which are predominantly focused on G/C with ≤5% mutations at A/T base pairs. Two of the tumors bearing a TP53 mutation, PD4109a and PD4199a, display the early trends of the significant strand biases at A:T and G:C base pairs (Table 9) evident in the earlier Woods et al 2007 exome study [55].
Collectively these genome-wide data sets suggest that mutations in TP53 accentuate strand-biased mutation patterns across the cancer genome implying that inactivated TP53 and dysregulated SHM contribute to such patterns. This conclusion is underlined by the documented functional interaction between TP53 and the Ig SHM machinery shown in mice by the Storb group, particularly in relation to the accentuated strand-biased mutation pattern of A-to-G versus T-to-C [49].

**INFLAMMATION AND CARCINOGENESIS**

Whilst there are some exceptions and qualifications, we conclude that there is a strong statistically significant similarity between the strand-biased mutation signatures of TP53 in many tumor types and the now well established Ig SHM pattern, particularly in relation to the strand biases of A-to-G over T-to-C and the G-to-A over C-to-T. Previous work on Ig SHM suggest that the A-to-G over T-to-C stand bias correlates strongly with A-to-I RNA editing coupled to reverse transcription to fix the A-to-G pattern in the cellular DNA [24].

The G-to-A over C-to-T pattern is found to be a dominant strand bias in all those cancers arising in tissues "least accessible to tobacco smoke" suggesting that this strand biased pattern (as well as A-to-G over T-to-C) arises from endogenous mutation processes in
most non-lymphoid cancers. We have previously concluded that the G-to-A over C-to-T strand bias is consistent with RNA mutations initiated at C-sites by activation-induced cytidine deaminase (AID)-mediated C-to-U deamination on the transcribed strand (TS) resulting in G-to-A transitions in the mRNA which are fixed as G-to-A mutations on the non-transcribed strand (NTS) following reverse transcription [5] and see Figure 1.

The present analyses therefore confirm and extends our earlier conclusions in a preliminary study of genome-wide somatic mutation data curated by the Cancer Genome Project (CGP) at The Welcome Trust Sanger Institute, Hixton, UK: [6]. The special features of the strand biases at A:T and G:C base pairs in tumors bearing mutated derivatives of TP53 imply a role for base-modified mRNA template intermediates and reverse transcription in somatic mutagenesis leading to or initiating cancer. In addition our analysis and conclusions are consistent with the view that inflammatory infiltrates, or in situ inflammatory episodes, in non-lymphoid tissues may contribute to dysregulated Ig SHM and thus oncogenesis, first by mutating components of the Ig SHM machinery and then by affecting mutations in TP53. This could occur via "a bystander effect" of various liberated cytokines inappropriately activating gene expression pathways in nearby non-lymphoid cells. It is common clinical
knowledge that most tumors have associated inflammatory infiltrates and are part and parcel of tumor growth.

This analysis also identifies several potential new drug targets for cancer therapy: in addition to AID and APOBEC family deaminases we can include Pol-η, ADAR1 and yet to be identified factors that modulate the apparatus of RNA Pol II transcription-coupled repair. Further, identifying the interacting proteins/genes mediating the functional TP53-Ig SHM interaction must also be considered a top priority in drug development and targeting.

Consistent with these conclusions is the large and now rapidly growing literature on chronic inflammation preceding cancer in many tissues [58]. Whilst it may not be possible to control all those induced somatic genetic factors leading to cancer, strategies to dampen and avoid chronic or transient inflammatory episodes in life may depress the chance of triggering "endogenous" mutagenic events, via dysregulated Ig SHM machinery, being turned on in non-lymphoid tissues [6]. This is particularly important in breast and ovarian tissues as estrogen can directly elevate AID expression as demonstrated by Petersen-Mahrt and colleagues [59-61]. Indeed the Pauklin et al data [60] show that estrogen induces AID transcription in these non-lymphoid tissues suggesting that the TP53 G-site mutation hotspots in breast cancers may be directly caused by AID and other APOBEC-family deaminases targeting such sites, as the data in Table 7 imply.
Collectively, these findings and the present analyses are shedding a new light on how we might view oncogenesis. Our work points to the unregulated Ig SHM mechanism as playing a key role in the progression of the main non-lymphoid cancer groups. This provides us with a fundamentally new molecular model with which to view the process of oncogenesis, and ways to develop new strategies for treating (and perhaps preventing) the development of certain cancer groups.

**Acknowledgements**

We thank John A Millman, Brent ("Charlie") J Stewart, Pat Carnegie, Susan Lester, Joseph F Williamson and Roger L Dawkins for early discussions and the AL & M Dawkins Foundation and C Y O'Connor ERADE Village Foundation for support in the early stages of the project. We thank Thierry Soussi for comments on an earlier draft of the manuscript, and Selena Nik-Zainal and Bert Vogelstein with helpful patience and assistance in accessing and analysing the supplementary data in references [55] and [57].
References


E.J. Steele, A. Franklin and R.V. Blanden, "Genesis of the strand biased signature in somatic hypermutation of rearranged..."


19 A. Faili, A. Stary, F. Delbos, S. Weller, S. Aoufouchi, A. Sarasin, J-C. Weill and C-A. Reynaud, "A backup role of DNA polymerase kappa in Ig gene hypermutation only takes place in the


36 M. Olivier, M. Hollstein and P. Hainaut, "TP53 mutations in human cancers: origins, consequences, and clinical use. In Levine AJ, Lane D "Additional Perspectives on the p53 family"


47 J.P. McDonald, E.G. Frank, B.S. Plosky, I.B. Rogozin, C. Masutani, F. Hanaoka, R. Woodgate and P.J. Gearhart," 129-derived strains of mice are deficient in DNA polymerase I and


53 T.M. Wilson, A. Vaisman, S.A. Martomo, P. Sullivan, L. Lan, F. Hanaoka, A. Yasui, R. Woodgate and P.J. Gearhart, "MSH2-MSH6 stimulates DNA polymerase eta, suggesting a role for A:T


Footnotes to Tables

Table 1

Data taken from Table 1 in [5]. Each entry is the mean percentage and standard error of the mean (shown in brackets) for each type of mutation rounded to one decimal place. All mutations scored from the coding or non-transcribed strand (NTS). Data in all studies were corrected for the base composition of the V region or 3' flanks assayed for mutations. In A, data averaged over twelve studies (n=12); in B, data averaged over twenty studies (n=20). Note that one study by Winter et al [20] inadvertently left out of the survey (
see supplementary data in [5]) but its omission does not change the patterns documented in Table 1B.

**Table 2**

Values in A represent the % of the total of 2279 somatic point mutations scored in category "All Breast Cancers" (R15); values in B represent the % of the total of 1215 somatic point mutations scored in category "All Bladder Cancers" (R15); values in C represent the % of the total of 2471 somatic point mutations scored in category "All Lung Cancers" (R15). † The percentage base composition in the TP53 coding region for codons 130-300 inclusive - the region which contains the vast majority of mutations spanning the DNA binding region. The Chi-Squared statistics (significance levels) are essentially unaltered if mutation frequencies are corrected for base composition. The breakdown of G-to-A and C-to-T mutations at CpG and non-CpG sites is as follows: In A. for C-to-T mutations 199 occur at CpG sites and 267 occur at non-CpG sites. For G-to-A 320 occur CpG and 387 occur at non-CpG sites. In B. for C-to-T 104 occur at CpG sites and 126 occur at non-CpG sites. For G-to-A 153 occur at CpG sites and 265 occur at non-CpG sites. In C. for C-to-T 209 occur at CpG and 173 occur at non-CpG sites. For G-to-A 214 occur at CpG and 154 occur at non-CpG sites.
Table 3

Values in the table represent the % of the total of 1575 somatic point mutations scored in category "All Ovary" (R15). † The percentage base composition in the TP53 coding region for codons 130-300 inclusive - the region which contains the vast majority of mutations spanning the DNA binding region. The Chi-Squared statistics (significance levels) are essentially unaltered if mutation frequencies are corrected for base composition. Note for C-to-T mutations 122 are at non-CpG and 183 are at CpG sites; for G-to-A mutations 222 are at non-CpG and 260 are at CpG sites.

Table 4

† All experimental mouse studies except work in cells from human Xeroderma Pigmentosium Variant (XPV) patients with genetic deficiency in DNA polymerase η (eta), Zeng et al [45]. UNG, uracil DNA glycosylase, removes uracil from DNA generating an abasic site; MSH2, MSH2 protein component of MSH2-MSH6 mismatch repair heterodimer; Polη, DNA polymerase eta, Y family polymerase; Polκ, DNA polymerase kappa, Y family polymerase; Poli, DNA polymerase iota, Y family polymerase; "-/-", homozygous deficiency (gene knock-
out mouse line); ¶ Aag, alkyladenine DNA glycosylase which removes hypoxanthine (deaminated adenine) from DNA generating an abasic site; § Ogg1, 8-hydroxyguanine-DNA glycosylase which removes oxidised guanine ("8oxoG") from DNA thus generating an abasic site. Not shown are two additional deficiencies: (a) The affect of Rev1 deficiency which affects G-to-C mutations as Rev1 plays a special and limited role in SHM by inserting dC opposite dU and abasic sites in VDJ DNA carrying AID-induced lesions, Jansen et al [50] and Steele [5] (member of Y family polymerase family) ); and (b) The effect of DNA polymerases zeta (ζ) deficiency where studies show Pol-ζ plays a special role in SHM by generating the observed low frequency tandem mutations in immunoglobulin variable regions, Saribasak et al [51]. C>>G. mutations off C exceed mutations off G by ratio indicated; A>>T, mutations off A exceed mutations off T by ratio indicated.

Table 5

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Value in A represent the % of the total of 1553 somatic point mutations scored in category &quot;All Colorectum Cancers&quot; (R15); values in B represent the % of the total of 844 somatic point mutations scored in category &quot;All Stomach Cancers&quot; (R15); values in C represent the % of the total of 789 somatic point mutations scored</td>
<td></td>
</tr>
</tbody>
</table>
in category "All Skin Cancers" (R15). † The percentage base composition in the TP53 coding region for codons 130-300 inclusive - the region which contains the vast majority of mutations spanning the DNA binding region. The Chi-Squared statistics (significance levels) are essentially unaltered if mutation frequencies are corrected for base composition. The breakdown of G-to-A and C-to-T mutations at CpG and non-CpG sites is as follows: In A. for C-to-T mutations 328 occur at CpG sites and 93 occur at non-CpG sites. For G-to-A 489 occur CpG and 186 occur at non-CpG sites. In B. for C-to-T 145 occur at CpG sites and 87 occur at non-CpG sites. For G-to-A 180 occur at CpG sites and 90 occur at non-CpG sites. In C. for C-to-T 97 occur at CpG and 217 occur at non-CpG sites. For G-to-A 44 occur at CpG and 123 occur at non-CpG sites.

**Table 6**

Values in the table represent the % of the total of 748 somatic point mutations scored in category "Glioblastoma" (R15). † The percentage base composition in the TP53 coding region codons 130-300 inclusive - the region which contains the vast majority of mutations spanning the DNA binding region. The Chi-Squared statistics (significance levels) are essentially unaltered if mutation frequencies are corrected for base composition. Note for C-to-T mutations 57 are at non-CpG
and 174 are at CpG sites; for G-to-A mutations 89 are at non-CpG and 154 are at CpG sites

**Table 7**

a. Only G-sites with $\geq 5$ mutations listed

b. NTS, non-transcribed strand; TS, transcribed strand

c. From Beale et al 2004 Table 2 [8]. ?, indicates that first two bases in TS motif consistent with deamination target in [8].

**Table 8**

Values in the table represent the % of the total of 1445 somatic point mutations scored over the exons of 11 individual breast cancer genomes comprising $\sim 20,000$ protein coding genes Wood et al [48].

The very low number of ambiguous and heterozygote double mutation data not included. The Chi-Squared statistics (significance levels) are indicated as in previous tables.

**Table 9**

From Exome Supplementary Data from Nik-Zainal et al 2012 [57].

The numbers in the table are the number of mutations scored. $N =$ total number of exomic mutations.
Legend to Figure 1.

Explanation for TP53 strand biased mutations in the "Endogenous Patterns" outlined in the text based on the reverse transcriptase model of Ig SHM viz. transcription-coupled DNA and RNA deamination followed by reverse transcription. Adapted and modified from Figure 5 in Steele 2009 [5] following Steele and Lindley 2010 [6]. Shown are some hypothesised DNA and RNA intermediates highlighted for the generation of the main strand-biased mutation signatures involving A-to-G, G-to-A, G-to-T and G-to-C. Black lines are DNA strands, red lines are mRNA, blue lines are cDNA strands copied off mRNA by a cellular reverse transcriptase such as DNA polymerase η. Steps A. through D. show various mutated DNA and RNA intermediates and substrate complexes for both deamination reactions, 8oxoG modifications in RNA, Wu and Z. 2008 [22] and cDNA synthesis (it is not known if 8oxoG sites are preferred in unpaired loops or dsRNA regions). In overview, mutations are first introduced at the DNA level by AID/APOBEC family-mediated C-to-U deaminations and then uracil
DNA glycosylase (UNG)-generated abasic sites in the TS (which can further mature into single strand nicks via the action of AP endonuclease (APE1). These template sites are transcribed into mRNA by RNA Pol II generating G-to-A and G-to-C modifications respectively in the mRNA, Kuraoka et al, 2003 [23] which on reverse transcription, integration and DNA replication result in G-to-A and G-to-C mutations in the NTS. Separately adenosine-to-inosine (A-to-I) RNA editing events at WA targets in the nascent ds mRNA stem loops may be copied back into DNA by reverse transcription via Pol-η, Franklin et al, 2004 [17]. Also shown are 8oxoG modifications in mRNA which on reverse transcription, integration and DNA replication would result in strand-biased G-to-T transversions on the NTS. The strand invasion (?) and integration of newly synthesised cDNA TS (?) are hypothesized necessary steps. In more detail: A. RNA Pol II introduces mutations in mRNA as it copies the AID/APOBEC lesions in TS DNA [22], concurrently A-to-I RNA edited sites appear in RNA stem(-loops) forming in nascent mRNA near the transcription bubble [Steele et al., 2006 [24] or 8oxoG modifications via reactive oxygen species [22]. B. Formation of RT-priming substrate (DNA polymerase-η) by annealing of nicked TS strand with an exposed 3'-OH end. This could arise due to excision at a previous AID-mediated abasic site or an excision introduced by endonuclease activity associated with the MSH2-MSH6 heterodimer engaging a U:G mispaired lesion. C.
Extension of new TS by cDNA synthesis from the 3'-OH end copying the already base modified mRNA template (with I base pairing preferentially, like G, with C; and 8oxoG mispairing with A). **D.** Then an unknown and indeterminant number of steps involving strand invasion(?), heteroduplex formation and/or resolution of heteroduplex (?), full length copying of newly synthesized transcribed strand (?) cDNA.

**Legend to Figure 2**

**A-to-G spectrum in TP53 for All Lung Cancers versus All Breast Cancers** Covering A-sites in TP53 from codon 100 to 300 inclusive. The Pearson correlation coefficient ($r$) is 0.93, which for 129 degrees of freedom gives a $P<0.001$. 
Appendix: Detection of Strand-Biased Somatic Mutation Signatures.

In a data set containing a large number of somatic mutations strand biased signatures are revealed by comparing the base substitution frequencies of Watson-Crick complements on the same strand. By convention nucleotide substitutions are read from the non-transcribed strand (NTS). However the known direction of transcription in a region of genomic DNA encoding a protein allows identification of the strands. Thus if A-to-G mutations occur with equal frequency on both strands, then its Watson-Crick complement, T-to-C will occur with equivalent frequency when scored off the same strand. However if there is a bias in the mutations favouring the NTS then A-to-G mutations will exceed T-to-C mutations. If there are systematic strand biases involving excessive mutations of A or G (e.g. as seen in many of the data tables presented herein) then the sum total of mutations of A will exceed the sum total of mutations of T (at A:T base pairs where A>>T) and the sum total of mutations of G will exceed the sum total of mutations of C (at G:C base pairs where G>>C).
Table 1 Typical patterns of somatic point mutations observed at rearranged immunoglobulin variable region loci in mice

A. Studies where PCR recombinant artefacts in DNA sequence data absent or minimized

<table>
<thead>
<tr>
<th>Original Base</th>
<th>Mutant Base</th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td>10.6</td>
<td>6.3</td>
<td>14.6</td>
<td>31.6</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>3.1 (0.6)</td>
<td>5.3</td>
<td>2.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>4.3 (0.8)</td>
<td>13.4</td>
<td>3.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>20.1 (1.9)</td>
<td>7.2</td>
<td>8.7</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

B. Studies where PCR recombinant artefacts in DNA sequence data were not minimized

<table>
<thead>
<tr>
<th>Original Base</th>
<th>Mutant Base</th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td>9.9</td>
<td>7.3</td>
<td>16.5</td>
<td>33.7</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>5.4 (0.3)</td>
<td>9.0</td>
<td>3.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>3.7 (0.3)</td>
<td>15.0</td>
<td>5.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>14.4 (0.6)</td>
<td>4.2</td>
<td>6.5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2 Somatic Point Mutation Patterns in the TP53 coding region in "All Breast Cancers", "All Bladder Cancers" and "All Lung Cancers".

A. All Breast Cancers

<table>
<thead>
<tr>
<th>Original Base</th>
<th>Mutant Base</th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
<th>Total</th>
<th>Statistics:</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (22%)†</td>
<td></td>
<td>2.6</td>
<td>1.9</td>
<td>11.0</td>
<td>15.5</td>
<td>A&gt;&gt;T 2.5x</td>
<td>P&lt;0.001</td>
</tr>
<tr>
<td>T (21%)</td>
<td></td>
<td>3.1</td>
<td>-</td>
<td>4.8</td>
<td>3.9</td>
<td>11.8</td>
<td>A&gt;G v T&gt;C 5.8x</td>
</tr>
<tr>
<td>C (29%)</td>
<td></td>
<td>2.8</td>
<td>20.4</td>
<td>-</td>
<td>3.6</td>
<td>26.8</td>
<td>G&gt;&gt;C 2.0x</td>
</tr>
<tr>
<td>G (28%)</td>
<td></td>
<td>31.0</td>
<td>9.3</td>
<td>5.5</td>
<td>-</td>
<td>45.9</td>
<td>G&gt;A v C&gt;T 1.8x</td>
</tr>
</tbody>
</table>

B. All Bladder Cancers

<table>
<thead>
<tr>
<th>Original Base</th>
<th>Mutant Base</th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
<th>Total</th>
<th>Statistics:</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (22%)†</td>
<td></td>
<td>3.1</td>
<td>1.4</td>
<td>9.6</td>
<td>14.1</td>
<td>A&gt;&gt;T 2.5x</td>
<td>P&lt;0.001</td>
</tr>
<tr>
<td>T (21%)</td>
<td></td>
<td>1.8</td>
<td>-</td>
<td>1.7</td>
<td>2.2</td>
<td>5.7</td>
<td>A&gt;G v T&gt;C 5.8x</td>
</tr>
<tr>
<td>C (29%)</td>
<td></td>
<td>2.9</td>
<td>18.9</td>
<td>-</td>
<td>5.2</td>
<td>27.0</td>
<td>G&gt;&gt;C 2.0x</td>
</tr>
<tr>
<td>G (28%)</td>
<td></td>
<td>34.4</td>
<td>9.5</td>
<td>9.3</td>
<td>-</td>
<td>53.3</td>
<td>G&gt;A v C&gt;T 1.8x</td>
</tr>
</tbody>
</table>

C. All Lung Cancers

<table>
<thead>
<tr>
<th>Original Base</th>
<th>Mutant Base</th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
<th>Total</th>
<th>Statistics:</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (22%)†</td>
<td></td>
<td>5.4</td>
<td>1.6</td>
<td>10.3</td>
<td>17.3</td>
<td>A&gt;&gt;T 2.5x</td>
<td>P&lt;0.001</td>
</tr>
<tr>
<td>T (21%)</td>
<td></td>
<td>2.1</td>
<td>-</td>
<td>2.3</td>
<td>2.5</td>
<td>6.9</td>
<td>A&gt;G v T&gt;C 4.5x</td>
</tr>
<tr>
<td>C (29%)</td>
<td></td>
<td>3.1</td>
<td>15.5</td>
<td>-</td>
<td>5.0</td>
<td>23.9</td>
<td>G&gt;&gt;C 2.2x</td>
</tr>
<tr>
<td>G (28%)</td>
<td></td>
<td>14.9</td>
<td>30.4</td>
<td>7.0</td>
<td>-</td>
<td>52.2</td>
<td>G&gt;A v C&gt;T 1.0</td>
</tr>
</tbody>
</table>

G>T v C>A 12.1x | P<0.001 |
G>C v C>G 1.4x | P<0.05 |
Table 3 Somatic Point Mutation Patterns in the TP53 coding region in "All Ovary" cancers

<table>
<thead>
<tr>
<th>Original Base</th>
<th>Mutant Base</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>A (22%)†</td>
<td>2.1</td>
<td>1.6</td>
</tr>
<tr>
<td>T (21%)</td>
<td>2.9</td>
<td>4.1</td>
</tr>
<tr>
<td>C (29%)</td>
<td>2.1</td>
<td>19.4</td>
</tr>
<tr>
<td>G (28%)</td>
<td>30.6</td>
<td>12.3</td>
</tr>
</tbody>
</table>

Statistics:

- A>>T 1.6x *P<0.001*
- A>G v T>C 3.2x *P<0.001*
- G>>C 1.8x *P<0.001*
- G>A v C>T 1.6x *P<0.001*
- G>T v C>A 5.9x *P<0.001*
- G>C v C>G 0.9x NS
### Table 4: Affect of main DNA repair deficiencies in Ig somatic hypermutation studies in mice

<table>
<thead>
<tr>
<th>Gene-protein deficiency†</th>
<th>Affect on frequency of:</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A/T mutations</td>
<td>G/C mutations</td>
</tr>
<tr>
<td>UNG-/-</td>
<td>15% suppression (A&gt;&gt;T strand bias)</td>
<td>95% C-to-T/G-to-A</td>
</tr>
<tr>
<td>UNG-/--MSH2-/-</td>
<td>98% suppression</td>
<td>&gt;99.5% C-to-T/G-to-A (C&gt;&gt;G 1.5x P&lt;0.001)</td>
</tr>
<tr>
<td>Pol η-/-</td>
<td>70% suppression (residual A&gt;&gt;T)</td>
<td>No change</td>
</tr>
<tr>
<td>Pol κ-/-</td>
<td>No change</td>
<td>No change</td>
</tr>
<tr>
<td>Pol η-/-Pol κ-/-</td>
<td>87% suppression (residual A&gt;&gt;T)</td>
<td>No change (enriched in expected proportion)</td>
</tr>
<tr>
<td>MSH2-/-</td>
<td>80% suppression (residual A&gt;&gt;T ?)</td>
<td>&gt;80% C-to-T/G-to-A (C&gt;&gt;G 1.5x P&lt;0.05)</td>
</tr>
<tr>
<td>Pol η-/--MSH2-/-</td>
<td>&gt;99% suppression</td>
<td>&gt;85% C-to-T/G-to-A (C&gt;&gt;G 1.5x P&lt;0.001)</td>
</tr>
<tr>
<td>Ogg1 (-/-)§</td>
<td>No change</td>
<td>No change</td>
</tr>
<tr>
<td>Aag (-/-)¶</td>
<td>No affect on A-to-G mutation frequency. Slight statistically significant increase in T-to-C and thus slight reduction in A&gt;&gt;T strand bias.</td>
<td>No change</td>
</tr>
<tr>
<td>p53 (-/-)</td>
<td>56% increase in A-to-G (increased A&gt;&gt;T strand bias)</td>
<td>Decreased G-to-A. Enhanced G-to-C v C-to-G ratio and thus strand bias.</td>
</tr>
</tbody>
</table>
Table 5 Somatic Point Mutation Patterns in the TP53 coding region in "All Colorectum Cancers", "All Stomach Cancers" and "All Skin Cancers".

A. All Colorectum Cancers

<table>
<thead>
<tr>
<th>Original Base</th>
<th>Mutant Base</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>A (22%)†</td>
<td></td>
<td>2.0</td>
</tr>
<tr>
<td>T (21%)</td>
<td>1.9</td>
<td></td>
</tr>
<tr>
<td>C (29%)</td>
<td>2.1</td>
<td>27.1</td>
</tr>
<tr>
<td>G (28%)</td>
<td>43.5</td>
<td>7.5</td>
</tr>
</tbody>
</table>

Statistics:
- **A>>T**: 1.0x  *NS*
- **A>G v T>C**: 1.2x  *NS*
- **G>>C**: 1.7x  *P<0.001*
- **G>A v C>T**: 1.6x  *P<0.001*
- **G>T v C>A**: 3.7x  *P<0.001*
- **G>C v C>G**: 1.2x  *NS*

B. All Stomach Cancers

<table>
<thead>
<tr>
<th>Original Base</th>
<th>Mutant Base</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>A (22%)†</td>
<td></td>
<td>2.3</td>
</tr>
<tr>
<td>T (21%)</td>
<td>3.8</td>
<td></td>
</tr>
<tr>
<td>C (29%)</td>
<td>3.6</td>
<td>27.5</td>
</tr>
<tr>
<td>G (28%)</td>
<td>32.5</td>
<td>7.7</td>
</tr>
</tbody>
</table>

Statistics:
- **A>>T**: 1.2x  *NS*
- **A>G v T>C**: 1.8x  *P<0.05*
- **G>>C**: 1.2x  *P<0.05*
- **G>A v C>T**: 1.2x  *NS*
- **G>T v C>A**: 2.2x  *P<0.05*
- **G>C v C>G**: 0.9x  *NS*

C. All Skin Cancers

<table>
<thead>
<tr>
<th>Original Base</th>
<th>Mutant Base</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>A (22%)†</td>
<td></td>
<td>3.3</td>
</tr>
<tr>
<td>T (21%)</td>
<td>2.4</td>
<td></td>
</tr>
<tr>
<td>C (29%)</td>
<td>3.2</td>
<td>39.8</td>
</tr>
<tr>
<td>G (28%)</td>
<td>21.2</td>
<td>7.0</td>
</tr>
</tbody>
</table>

Statistics:
- **A>>T**: 0.9x  *NS*
- **A>G v T>C**: 0.8x  *NS*
- **C>>G**: 1.4x  *P<0.01*
- **C>T v G>A**: 1.9x  *P<0.001*
- **G>T v C>A**: 2.2x  *P<0.05*
- **G>C v C>G**: 1.6x  *NS*
Table 5 Somatic Point Mutation Patterns in the TP53 coding region in "Glioblastoma" cancers of brain

<table>
<thead>
<tr>
<th>Original Base</th>
<th>Mutant Base</th>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>A (22%)†</td>
<td>-</td>
<td>1.5</td>
</tr>
<tr>
<td>T (21%)</td>
<td>1.2</td>
<td>-</td>
</tr>
<tr>
<td>C (29%)</td>
<td>3.2</td>
<td>30.9</td>
</tr>
<tr>
<td>G (28%)</td>
<td>32.5</td>
<td>7.4</td>
</tr>
</tbody>
</table>

G>T v C>A 2.3x P<0.05
G>C v C>G 1.1x NS
Table 7 Hotspots in breast cancers for G-site mutations in TP53 in relation to codon, target motif and likely dC-to-dU deaminase

<table>
<thead>
<tr>
<th>Codon&lt;sup&gt;a&lt;/sup&gt;</th>
<th>&lt;sup&gt;a&lt;/sup&gt;Number mutations:</th>
<th>&lt;sup&gt;b&lt;/sup&gt;Motif on:</th>
<th>&lt;sup&gt;c&lt;/sup&gt;Likely C&gt;U DNA deaminase</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>G&gt;A</td>
<td>G&gt;T</td>
<td>G&gt;C</td>
</tr>
<tr>
<td>156</td>
<td>1</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>157</td>
<td>0</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>158</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>173</td>
<td>7</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>175</td>
<td>128</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>176</td>
<td>6</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>181</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>196</td>
<td>2</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>216</td>
<td>12</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>237</td>
<td>17</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>238</td>
<td>13</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>242</td>
<td>7</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>243</td>
<td>5</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>244</td>
<td>10</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>244</td>
<td>8</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>245</td>
<td>38</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>245</td>
<td>14</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>248</td>
<td>104</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>249</td>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>249</td>
<td>1</td>
<td>14</td>
<td>2</td>
</tr>
<tr>
<td>258</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>266</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>266</td>
<td>10</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>272</td>
<td>10</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>273</td>
<td>93</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>275</td>
<td>8</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>277</td>
<td>2</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>280</td>
<td>7</td>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td>281</td>
<td>5</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>282</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>283</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>285</td>
<td>21</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>286</td>
<td>9</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
| 294              | 0   | 5   | 0   | GGA | TCC | APOBEC1?
Table 8: Somatic point mutation patterns in breast cancer derived from large-scale exomic sequencing of breast cancer genomes

<table>
<thead>
<tr>
<th>Original Base</th>
<th>Mutant Base</th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td>4.3</td>
<td>3.0</td>
<td>6.0</td>
<td>13.4</td>
</tr>
<tr>
<td>T</td>
<td>2.7</td>
<td></td>
<td>3.5</td>
<td>1.7</td>
<td>7.8</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>7.8</td>
<td>16.2</td>
<td></td>
<td>11.9</td>
<td>35.9</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>19.7</td>
<td>9.1</td>
<td>14.1</td>
<td></td>
<td>42.9</td>
<td></td>
</tr>
</tbody>
</table>

Statistics:
- A >> T  1.7x  P<0.01
- A >> G v T >> C  1.7x  P<0.05
- G >> C  1.2x  P<0.05
- G >> A v C >> T  1.2x  NS
- G >> T v C >> A  1.2x  NS
- G >> C v C >> G  1.2x  NS
Table 9 Somatic Point Mutation Patterns in the TP53 mutated Breast Cancers in Nik-Zainal et al exomic data [57].

**PD3890a Brac1 N= 104**

<table>
<thead>
<tr>
<th>Original Base</th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>T</td>
<td>3</td>
<td>-</td>
<td>5</td>
<td>7</td>
<td>15</td>
</tr>
<tr>
<td>C</td>
<td>10</td>
<td>19</td>
<td>-</td>
<td>10</td>
<td>39</td>
</tr>
<tr>
<td>G</td>
<td>13</td>
<td>7</td>
<td>18</td>
<td>-</td>
<td>38</td>
</tr>
</tbody>
</table>

Statistics:
- A $\Rightarrow$T  NS
- A $\Rightarrow$G v T$\Rightarrow$C  NS
- G $\Rightarrow$C  NS
- G$\Rightarrow$A v C$\Rightarrow$T  NS
- G$\Rightarrow$T v C$\Rightarrow$A  NS
- G$\Rightarrow$C v C$\Rightarrow$G  NS

**PD4005a Brac1 N= 125**

<table>
<thead>
<tr>
<th>Original Base</th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-</td>
<td>2</td>
<td>6</td>
<td>5</td>
<td>13</td>
</tr>
<tr>
<td>T</td>
<td>6</td>
<td>-</td>
<td>8</td>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td>C</td>
<td>7</td>
<td>24</td>
<td>-</td>
<td>17</td>
<td>48</td>
</tr>
<tr>
<td>G</td>
<td>22</td>
<td>10</td>
<td>13</td>
<td>-</td>
<td>45</td>
</tr>
</tbody>
</table>

Statistics:
- A $\Rightarrow$T  NS
- A $\Rightarrow$G v T$\Rightarrow$C  NS
- G $\Rightarrow$C  NS
- G$\Rightarrow$A v C$\Rightarrow$T  NS
- G$\Rightarrow$T v C$\Rightarrow$A  NS
- G$\Rightarrow$C v C$\Rightarrow$G  NS

**PD4109a Triple Negative N= 174**

<table>
<thead>
<tr>
<th>Original Base</th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-</td>
<td>11</td>
<td>5</td>
<td>11</td>
<td>27</td>
</tr>
<tr>
<td>T</td>
<td>7</td>
<td>-</td>
<td>7</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td>C</td>
<td>8</td>
<td>25</td>
<td>-</td>
<td>20</td>
<td>53</td>
</tr>
<tr>
<td>G</td>
<td>33</td>
<td>18</td>
<td>23</td>
<td>-</td>
<td>74</td>
</tr>
</tbody>
</table>

Statistics:
- A $\Rightarrow$T  1.35x  NS
- A $\Rightarrow$G v T$\Rightarrow$C  1.57x  NS
- G $\Rightarrow$C  1.4x  NS
- G$\Rightarrow$A v C$\Rightarrow$T  1.32x  NS
- G$\Rightarrow$T v C$\Rightarrow$A  2.25x  NS
- G$\Rightarrow$C v C$\Rightarrow$G  1.15x  NS

**PD4199a ER-ve, HER2 +ve N= 156**

<table>
<thead>
<tr>
<th>Original Base</th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>T</td>
<td>1</td>
<td>-</td>
<td>3</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>C</td>
<td>6</td>
<td>23</td>
<td>-</td>
<td>19</td>
<td>48</td>
</tr>
<tr>
<td>G</td>
<td>56</td>
<td>8</td>
<td>32</td>
<td>-</td>
<td>96</td>
</tr>
</tbody>
</table>

Statistics:
- A $\Rightarrow$T  NS
- A $\Rightarrow$G v T$\Rightarrow$C  NS
- G $\Rightarrow$C  2.0x  P<0.01
- G$\Rightarrow$A v C$\Rightarrow$T  2.43x  P<0.05
- G$\Rightarrow$T v C$\Rightarrow$A  1.33x  NS
- G$\Rightarrow$C v C$\Rightarrow$G  1.68x  NS
Formation RT- priming substrate

Annealing of AID-nicked TS and base modified mRNA copied off the TS (with AID/APOBEC-mediated lesions) and A-to-I edited WΔ sites within nascent dsRNA stems and ROS-mediated 8oxoG sites.

A

5' NTS

3' TS

3' OH

TS

RT-mediatd cDNA synthesis

B

C

C-to-U Abasic site

8oxoG

A-to-I

mRNA

mRNA

G-to-A G-to-C

TS

3' OH

mRNA

5' NTS

TS

G TG A C

G TG A C

A-to-I

A-to-I

8oxoG

I I A C

I I A C

8oxoG

I I A C

8oxoG

Figure 1
Figure 2

A-to-G spectrum in All Lung Cancers versus All Breast Cancers covering A-sites in TP53 from codon 100 to 300 inclusive. The Pearson correlation coefficient ($r$) is 0.93, which for 129 degrees of freedom gives a $P<0.001$. 