Accepted 24 March 2010 *DNA REPAIR*

Somatic mutation patterns in non-lymphoid cancers resemble the strand biased somatic hypermutation spectra of antibody genes

Edward J Steele & Robyn A Lindley

CY O'Connor ERADE Village Foundation,

Canning Vale, Western Australia 6155

AUSTRALIA

Corresponding author :
Edward J Steele
CY O'Connor ERADE Village Foundation
PO Box 5100 , Canning Vale, WA 6155
AUSTRALIA
Tel: 61 (8) 9397 1556
Fax: 61 (8)9397 1559
Emails: ejsteele@cyo.edu.au
and rlindley@cyo.edu.au

Key words: *somatic hypermutation ,human cancer genomes, somatic point mutations, strand biased mutation, RNA intermediate, reverse transcription*

It has been long accepted that many types of B cell cancer (lymphomas, myelomas, plasmacytomas, etc) are derived from the antigen-stimulated B cell Germinal Center (GC) reaction [1-4] i.e. they are aberrant products of the somatic hypermutation mechanism normally targeting rearranged immunoglobulin (Ig) variable genes (so called V[D]J regions). Here we provide evidence that the somatic mutation patterns of some well characterised cancer genomes [5] such as lung carcinomas, breast carcinomas and squamous cell carcinomas, strongly resemble *in toto* or in part the spectrum of somatic point mutations observed in normal physiological somatic hypermutation (SHM) in antibody variable genes [6]. This implies that whilst SHM itself is a tightly regulated and *beneficial* mutational process for B lymphocytes of the immune system, aberrant mutations (or "crises") or inadvertent activation of this complex activation-induced cytidine deaminase (AID)-dependent mechanism in a range of somatic tissue types could result, as often speculated [7], in cancer.

In normal physiological Ig SHM two main groups of strand-biased mutations are known to occur: (i) at A:T base pairs whereby A mutations exceed T mutations by 2-3 fold; and (ii) at G:C base pairs whereby G mutations exceed C mutations by at least 1.7 fold. A critical analysis of the SHM literature in experimental mouse systems of the past 25 years [6] shows that these strand-biased mutation spectra are best understood as occurring first in RNA

molecules which are then copied back into DNA most likely by a cellular reverse transcription (RT) process carried out by the sole error-prone DNA polymerase [6] known to be involved in SHM, DNA polymerase–η (eta) ; which also happens to be a relatively efficient reverse transcriptase [8] being active on dilution at low mole ratios of enzyme-to-template *in vitro* (1:20-1:100). Thus, whilst it has been clearly established that AID deaminase initiates SHM and Ig class switching by direct deamination of C-to-U in ssDNA in the context of transcription [9] the full mutation spectrum of SHM appears to be generated by the synthesis of, modification of and RT-copying of the Ig pre-mRNA template ie. most Ig somatic mutations appear first as "RNA mutations" which are then copied back into B lymphocyte genomic DNA [6]. The elements of this proposal were first advanced by Steele & Pollard in 1987 [10]. Thus the A>>T strand biased mutation pattern is best understood as a combination of adenosine-to-inosine (A-to-I) pre-mRNA editing [11] followed by an error-prone Pol-η dependent reverse transcription step fixing the A-to-G, as well as A-to-T and A-to-C, as strand biased mutations in B cell DNA [6,11]. The G>>C strand biased mutation pattern, only recently recognised, is consistent with the misincorporation signature of RNA polymerase II [6] transcribing template DNA strands carrying AID-mediated lesions generated at C bases viz. uracils and abasic sites [12]. Again a reverse transcription step presumably involving Pol-η would then need to

intervene to fix the RNA mutation pattern in DNA. Possible molecular mechanisms and substrates have been discussed elsewhere [6]. Here we turn our attention to human cancer and question whether these insights can be of use in understanding the genesis of the somatic mutation spectra observed for a range of cancer genomes from different tissue types.

Our understanding of the fundamental mechanisms causing somatic mutations in human cancer cells is still relatively rudimentary. However SHM in the vertebrate immune system is one well characterised situation of a highly regulated *beneficial* mutator process where we now have a good understanding of the mechanism albeit incomplete [6]. Thus somatic point mutations are focused on a 1-2 kb region targeting V[D]J genes in GC B lymphocytes and intense antigen-binding selection ensures that mutated B cells bearing surface Ig antigen receptors with similar or better binding affinity for antigen survive, proliferate and become part of the memory B cell pool. It is therefore conceivable, as suggested by Honjo and associates for example [7], that disturbances in the regulation of this system in non-lymphoid somatic cells may unleash an uncontrolled spray or shower of somatic point mutations, and thus contribute to the development of cancer (Figure 1).

For a preliminary analysis to test the feasibility of this idea we choose a subset of the well curated cancer genome data base at

The Welcome Trust Sanger Institute website [5]. The data sampled are at:

http://www.sanger.ac.uk/genetics/CGP/Studies/

The mutation patterns in individual tumour cell lines (or tissue biopsies) included in our analysis were specifically in the "Capillary Screen Data/ Protein Kinase Gene Analysis" at http://www.sanger.ac.uk/perl/genetics/CGP/cgp_viewer?action=study&study_id=34 for Samples BB30-HNC down to PD2543a.

To allow valid comparisons between somatic mutation patterns it was necessary in all previous SHM analyses [6] to establish the most likely somatic mutations that occur *in vivo* during an immune response. Such a pattern would be free of confounding strand-bias blunting effects due to PCR product artefacts (PCR hybrid or recombinant molecules). This is explained in a previous publication [6]. Thus Table 1a shows a true and typical pattern of somatic point mutations observed at rearranged Ig loci in mice undergoing an antibody response (also Table 1a in ref. 6). All mutations are read from the non-transcribed strand, and all mutation frequencies (expressed as %) have been corrected for slight differences in the base composition of the V[D]J target areas assayed for somatic mutations (typically 300-400 bp). It is found that the mutations off A exceed the mutations off T by 2.9 fold, and the mutations off G exceed the mutations off C by 1.7 fold. We interpret these strand bias patterns as being reflective of the

nucleotide sequence errors generated in the Ig mRNA during SHM, and which are then copied back into DNA (see Fig 5 in ref.6).

In the samples of the Cancer Genome Project (CGP) analysed here, it is found that the somatic mutation spectra of lung adenocarcinomas (Table 1b), lung small cell carcinomas (Table 1c), breast ductal carcinomas (Table 1d) and squamous cell carcinomas (Table 1e) in many cases strongly resemble *in toto* or in part the strand biased patterns typical of Ig SHM (Table 1a). Unlike Ig genes where we have been able to correct for base composition we have had to assume base composition "evening-out" effects: this is not an unreasonable assumption given the large number of mutated cancer-associated genes involved. It is also found that different types of cancer show some quite distinct variations in the basic strand bias pattern observed. For example, and as pointed out by Greenman et al [5], some cancers such as skin malignant melanomas have more restricted spectra with mutations highly focused to C:G base pairs with mutations at A:T base pairs suppressed (Table 1f). Such a pattern is typical of the AID deaminase footprint of SHM at the Ig locus established by Neuberger and colleagues for mice lacking uracil DNA glycosylase and functional mismatch repair machinery, MSH2-MSH6 [13,14]. However in contrast to that data we do not see an excess of C-to-T over G-to-A mutations suggestive of AID-mediated deamination preferentially on the displaced non-transcribed strand during

transcription [6]. We have also found that tumors with comparatively large numbers of somatic mutations, such as NCI-H2009 (lung adenocarcinoma, Table 1g) and CP66-MEL (malignant melanoma, Table 1h), display somatic mutation spectra similar to the pooled data for that tumor category (Table 1b and 1f respectively). A comparison of the somatic mutation spectra of the pooled data and the individual data set adds weight to the view that different tissue tumor types can display different somatic mutation spectra as shown by Greenman et al [5]. This is intriguing in the context of the overall resemblance of these tissue-specific somatic mutation spectra to SHM patterns. For example, in another category of human skin cancer Xeroderma Pigmentosum Variant (XPV) patients carry genetic deficiencies in DNA Polymerase-η. Somatic hypermutation analysis on the J-region proximal rearranged VH6 gene in such patients shows a significant reduction in mutations at A:T base pairs which allowed Gearhart and associates [15] to first show a clear role for an error-prone polymerase (Pol-η) involvement in generating strand biased SHM patterns of antibody genes.

Our preliminary analysis therefore raises many more interesting questions and possibilities for further investigation. Further work should establish the sequence context of the A:T focused mutations in those tumor types which appear to show them (Tables 1b,c,d,e) as it is known that both Pol-η in its DNA-based

copying mode as well as the transcription-coupled pre-mRNA editor ADAR1 both display selectivity for mutating or deaminating WA-sites in DNA or RNA (where the 5' W = A or T/U) [11]. Further work is also required to establish that the subset of tumor samples chosen here is indeed consistent and representative in a wider analysis.

Whilst there is a clear resemblance between SHM and CGP somatic mutation spectra, the resemblance is not complete. Thus in SHM, mutations occur more or less equally at A:T and C:G base pairs. In comparison, the sample data here (Table1b,c,d,e) shows an approximate decrease of mutations at A:T by about 50% and a corresponding enrichment of mutations at C:G base pairs. With respect to C:G base pair targets it is also evident from Greenman et al [5] that an AID or AID-like deaminase not dependent on CpG may play a role in most malignant melanoma C:G>T:A mutations and in about half of C:G targeted mutations in lung, breast and other tumors. It is also evident that G>>C strand bias is not evident in the samples of breast ductal carcinomas analysed here. Lastly in many of the lung carcinomas there is a prominent strand bias signature of G-to-T >> C-to-A , which is far more notable than in the SHM spectrum (Table 1a) or non-lung cancer tumors (Table 1d) suggestive maybe of oxidative damage causing 8oxoG mutations in RNA [16] which can now base pair with A and which

are then fixed in DNA by reverse transcription [6] as G-to-T mutations.

We conclude that overall there is a striking resemblance between the patterns of Ig somatic mutations produced in Germinal Center derived hypermutated B lymphocytes, and the various cancer samples analysed here. This allows the qualified conclusion that mutated or base-modified RNA template intermediates coupled to error-prone reverse transcription (via Pol-η) could be responsible for the somatic mutation spectrum in cancer genomes. A schematic outline of the proposed causal relationship between aberrant Ig SHM and the genesis of somatic point mutations in cancer is shown in Figure 1. The observed quantitative differences in somatic mutation spectra among the cancer samples analysed suggests the possible involvement of an RNA intermediate as a part of aberrant SHM activity. This in turn suggests potential targets involved in SHM such as AID, Pol-η, ADAR1 and modulators of the RNA Pol II transcription-coupled repair (TCR) apparatus should be considered as possible drug targets in the development of future cancer therapies. Additional data analyses are being conducted to characterise further the relationships found here. The results also have some strong implications for genetic information transfer in somatic tissues.

References

1. M.G. Weigert, I.M. Cesari, S.J. Yonkovich, M. Cohn, Variability in the Lambda light chain sequences of mouse antibody, Nature 228 (1970) 1045-1047.

2. T.T. Wu, E.A. Kabat, An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity, J. Exp. Med. 132 (1970) 211- 250.

3. R.D. Morin, N.A. Johnson, T.M. Severson, A.J. Mungal et al, Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin, Nat. Genet. 42 (2010) 181-185.

4. N.S. Aguilera, A. Auerbach, C.L. Barekman, J. Lichy, S.L. Abbondanzo, Activation-induced cytidine deaminase expression in diffuse large B-cell lymphoma with a paracortical growth pattern: a lymphoma of possible interfollicular large B-cell origin, Arch. Pathol. Lab. Med. 134 (2010) 449-456.

5. C. Greenman, P. Stephens, R. Smith, G.L. Dalgliesh, et al Patterns of somatic mutation in human cancer genomes Nature 446 (2007) 153-158.

6. E.J. Steele, Mechanism of somatic hypermutation : Critical analysis of strand biased mutation signatures at A:T and G:C base pairs Molec Immunol 46 (2009) 305-320.
7. H. Nagaoka, T.H. Tran, M. Kobayashi, M. Aida, T. Honjo, Preventing AID, a physiological mutator, from deleterious activation: regulation of the genomic instability that is associated with antibody diversity, Int. Immunol. In Press (2010).
8. A. Franklin, P.J. Milburn, R.V. Blanden, E.J. Steele, Human DNA polymerase–η, an A-T mutator in somatic hypermutation of rearranged immunoglobulin genes, is a reverse transcriptase, Immunol. Cell Biol. 82 (2004) 219-225.
9. J. M. Di Noia, M.S. Neuberger, Molecular mechanisms of somatic hypermutation Annu.Rev. Biochem. 76 (2007) 1-22
10. E.J. Steele, J.W. Pollard, Hypothesis: Somatic hypermutation by gene conversion via the error prone DNA-to-RNA-to DNA information loop, Mol. Immunol. 24 (1987) 667-673.
11. E.J. Steele, R.A. Lindley, J. Wen, G.F. Weiller, Computational analyses show A-to-G mutations correlate with nascent mRNA hairpins at somatic hypermutation hotspots, DNA Repair 5 (2006) 1346-1363
12. I. Kuraoka,M. Endou, Y. Yamaguchi, Y. Wada, H. Handa, K. Tanaka, Effects of endogenous DNA base lesions on

transcription elongation by mammalian RNA polymerase II, J. Biol. Chem. 278 (2003) 7294-7299.

13. C. Rada, J.M. Di Noia, M.S. Neuberger, Mismatch recognition and uracil excision provide complementary paths to both Ig switching and the A/T-focused phase of somatic mutation, Mol. Cell 16 (2004) 163-171.
14. K. Xue, C. Rada, M.S. Neuberger, The *in vivo* pattern of AID targeting to immunoglobulin switch regions deduced from mutation spectra in msh2-/-ung-/- mice, J. Exp. Med. 203 (2006) 2085-2094
15. X. Zeng, D.B. Winter, C. Kasmer, K.H. Kraemer, A.R. Lehmann, P.J. Gearhart, DNA polymerase η is an A-T mutator in somatic hypermutation of immunoglobulin variable genes, Nat. Immunol. 2 (2001) 537-541.
16. J. Wu, Z. Li , Human polynucleotide phosphorylase reduces oxidative RNA damage and protects HeLa cell against oxidative stress, Biochem Biophys Res Comm 372 (2008) 288-292

Acknowledgement: We thank Professor Roger L Dawkins and the CY O'Connor ERADE Village Foundation for financial support. EJS is currently an AL & M Dawkins fellow supported by the AL & M Dawkins Foundation. The authors have no competing interests.

Footnotes to Table 1: Patterns of somatic point mutations observed at rearranged Ig loci in mice and various cancer genome samples. The right ascending diagonal in each base substitution table shows the transitions in bold.

a. Adapted from Table 1a in Steele 2009 (*2*). Each frequency is the mean percentage of twelve studies involving active immunisation and antigen-selection of immunoglobulin VDJ genes expressed in B cell clones (hybridomas, single B cells, or targeted VH186.2DJ transcripts or genes by nested PCR). PCR hybrid formation has been assessed to be non-existent or minimal in these SHM data sets. Not shown are the standard errors for each estimate which can be found in the original table. **b.** Frequencies of somatic mutations (as %) in pooled data of 13 lung adenocarcinomas involving 456 point mutations and involving ~ 495 genes (NCI-H1395, NCI-H1437, NCI-H2009, NCI-H2087, NCI-H2122, NCI-H2126, PD0277a, PD1342a, PD1351a, PD1352a, PD1353a, PD1414a, PD1418a). The level of statistical significance (Chi square,1 df) for deviation from 1.0 of A over T and G over C mutation ratios are $P<0.05$ and $P<0.01$ respectively. **c.** Frequencies of somatic mutations (as %) in pooled data of 4 lung small cell carcinomas involving 175 point mutations and involving ~ 190 genes (LB 647-SCLC, NCI-H128, NCI-H209, NCI-H2171). The level of statistical significance (Chi square,1 df) for deviation from 1.0 of A over T and G over C mutation ratios are $P<0.05$ and $P>0.05$ (NS) respectively. **d.**

Frequencies of somatic mutations (as %) in pooled data of 12 breast ductal carcinomas involving 275 point mutations and involving ~ 287 genes (HCC 1148, HCC 1187, HCC 1395, HCC 1937, HCC 1599, HCC 1954, HCC 2157, HCC 2218, HCC 38, PD0118a, PD0119a, PD1233a). The level of statistical significance (Chi square,1 df) for deviation from 1.0 of A over T and G over C mutation ratios are $P<0.01$ and $P>0.05$ (NS) respectively. **e.** Frequencies of somatic mutations (as %) in pooled data of 8 squamous cell carcinomas, mainly lung involving 117 point mutations and involving ~ 103 genes (BB 30-HNC, BB 49-HNC, LB 771-HNC, PD0248a, PD0251a, PD0269a, PD1369a, PD1379a). The level of statistical significance (Chi square,1 df) for deviation from 1.0 of A over T and G over C mutation ratios are $P>0.05$ (NS) and $P<0.05$ respectively. **f.** Frequencies of somatic mutations (as %) in pooled data of 7 skin malignant melanomas, involving 1166 point mutations and involving ~ 1087 genes (Colo-829, CP50-MEL-B, CP66-MEL, LB 2518-MEL, LB 373-MEL-D, MZ7-Mel). **g.** Frequencies of somatic mutations (as %) in NCI-H2009, a lung adenocarcinoma involving 146 point mutations and involving ~ 142 genes. The level of statistical significance (Chi square, 1 df) for deviation from 1.0 of A over T and G over C mutation ratios are $P>0.05$ (NS) and $P>0.05$ (NS) respectively. **h.** Frequencies of somatic mutations (as %) in CP66-MEL, a skin malignant melanoma involving 248 point mutations and involving ~ 225 genes.

Table 1 : Patterns of somatic point mutations at rearranged Ig loci in mice and various cancer genome samples

a.

From \ To	A	T	C	G	Total
A	-	10.6	6.3	**14.6**	31.6
T	3.1	-	**5.3**	2.6	11.0
C	4.3	**13.4**	-	3.6	21.3
G	**20.1**	7.2	8.7	-	36.1

Rearranged Ig loci in mice

b.

From \ To	A	T	C	G	Total
A	-	6.4	2.4	**5.7**	14.4
T	4.3	-	**1.5**	2.0	7.9
C	12.2	**14.0**	-	10.0	36.4
G	**18.4**	24.1	9.6	-	52.1

Lung adenocarcinoma pool

c.

From \ To	A	T	C	G	Total
A	-	6.3	0.6	**12.5**	19.4
T	2.3	-	**2.2**	1.7	6.3
C	15.4	**12.6**	-	4.6	32.6
G	**13.7**	25.1	8.6	-	49.1

Lung small cell carcinoma pool

d.

From \ To	A	T	C	G	Total
A	-	3.6	2.9	**7.3**	13.8
T	0.4	-	**2.9**	1.1	4.4
C	7.3	**19.6**	-	13.8	40.7
G	**18.9**	9.5	12.7	-	41.0

Breast ductal carcinoma pool

e.

From \ To	A	T	C	G	Total
A	-	7.7	0.9	**7.7**	16.2
T	0.0	-	**4.3**	0.9	5.1
C	6.0	**14.5**	-	5.1	25.6
G	**18.8**	16.2	14.5	-	49.6

Squamous cell carcinoma pool

f.

From \ To	A	T	C	G	Total
A	-	0.3	1.3	**1.8**	3.3
T	1.1	-	**1.3**	0.9	3.3
C	1.6	**45.1**	-	0.5	47.3
G	**43.4**	2.0	0.6	-	46.1

Malignant melanoma pool

g.

From \ To	A	T	C	G	Total
A	-	6.8	0.0	**8.2**	15.1
T	2.7	-	**2.7**	2.1	7.5
C	6.2	**12.3**	-	15.7	34.2
G	**11.6**	19.1	12.3	-	43.1

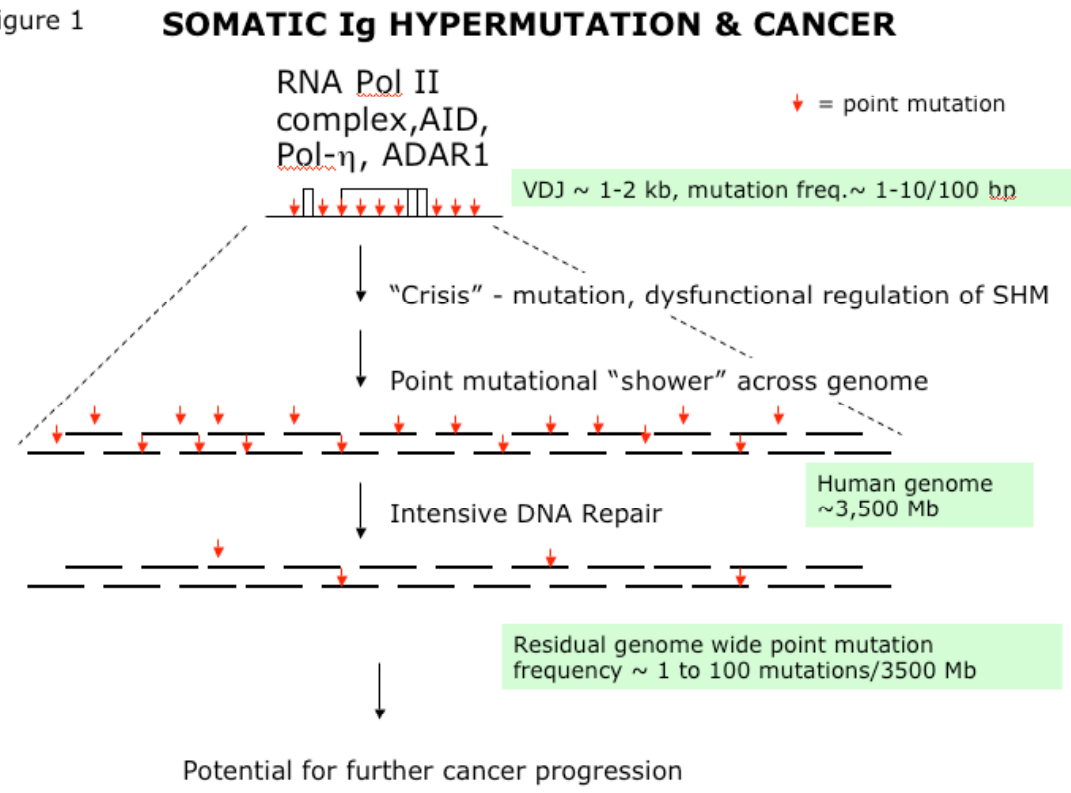
NCI-H2009 lung adenocarcinoma

h.

From \ To	A	T	C	G	Total
A	-	0.0	1.6	**0.8**	2.4
T	1.2	-	**0.8**	0.4	2.4
C	0.0	**47.5**	-	0.4	48.0
G	**46.0**	0.4	0.8	-	47.2

CP66-MEL malignant melanoma

Figure 1


SOMATIC Ig HYPERMUTATION & CANCER
RNA Pol II complex,AID, Pol-η, ADAR1
= point mutation
VDJ ~ 1-2 kb, mutation freq.~ 1-10/100 bp
"Crisis" - mutation, dysfunctional regulation of SHM
Point mutational "shower" across genome
Human genome ~3,500 Mb
Intensive DNA Repair
Residual genome wide point mutation frequency ~ 1 to 100 mutations/3500 Mb
Potential for further cancer progression