Published in final edited form as:

Mol Immunol. 2011 July ; 48(12-13): 1477–1483. doi:10.1016/j.molimm.2011.04.002.

Somatic hypermutation targeting is influenced by location within the Immunoglobulin V region

Reuma Magori Cohen¹, Steven H. Kleinstein², and Yoram Louzoun^{1,3}

¹ Math Department and Gonda Brain Research Center, Bar Ilan University, Ramat Gan 52900, Israel

² Interdepartmental Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, USA and Department of Pathology, Yale University School of Medicine, New Haven, Connecticut, USA

Abstract

The observed mutation pattern in Immunoglobulin V genes is influenced by several mechanisms, including AID targeting to DNA motifs (hot spots), negative selection of B cells that accumulate mutations that prevent the expression of the Ig receptor, and positive selection of B cells that carry affinity-increasing mutations. These influences, combined with biased codon usage, produce the well-known pattern of increased replacement mutation frequency in the CDR regions, and decreased replacement frequency in the framework regions. Through the analysis of over 12,000 mutated sequences, we show that the specific location in the V gene also significantly influences mutation accumulation. While this position-specific effect is partially explained by selection, it appears independent of the CDR/FWR structure. To further explore the specific targeting of SHM, we propose a statistical formalism describing the mutation probability of a sequence through the multiplication of independent probabilities. Using this model, we show that C->G (or G->C) mutations are almost as frequent as C->T and G->A mutations, in contrast with C->A (or G->T) mutations, which are not more probable than any other mutation. The proposed statistical framework allows us to precisely quantify the effect of V gene position, substitution type, and micro-sequence specificity on the observed mutation pattern.

Introduction

B cells immunoglobulin (Ig) diversity is obtained in multiple stages. The initial diversity is generated in the bone marrow during the Heavy Chain (HC) and L Chain (LC) V(D)J rearrangement [1], including TdT induced nucleotide addition and exonuclease-mediated deletion at the segment junctions [2]. A second stage of diversification is obtained in the periphery following exposure to antigen through somatic hypermutation (SHM) [3]. In SHM B cells introduce point-mutations into the variable region of their immunoglobulin (Ig) genes [4]. The mutation rates in these genes can reach 1.e-3 mutation per base pair per division [5].

© 2011 Elsevier Ltd. All rights reserved.

³Correspondence and requests should be addressed to Yoram Louzoun. louzouy@math.biu.ac.il, phone- 972-3-5317610.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

SHM occurs in germinal centers (GC) of secondary lymphoid organs [6]. To initiate SHM, activation induced cytidine deaminase (AID) is required. AID initiates hypermutations on both the transcribed and non transcribed strands of Ig variable region genes by deaminating a C to form a U during transcription [7, 8]. The resulting U/G mismatches [3, 9] have several possible fates [10]. First, the mismatch can be replicated over, producing C->T transition mutations. Second, the U can be removed by UNG, leading to C->A/C/T mutations. Finally, the mismatch repair machinery may be engaged leading to excision of neighboring bases and DNA resynthesis primarily by Pol η , but also by other polymerases (ι and ζ) [11]. The repair mechanism is error prone and leads to point mutations at all base positions in the surrounding sequence. AID targets specific locations in the Ig gene [12–14]. SHMs occurs mainly in B cell Ig genes, but have been recently described to occur at a much lower mutation rate in other genes [15].

SHM preferentially targets C bases in the WRCY hot-spot motif (W = A or T, R = G or A, and Y = T or C), or its reverse complementary sequence RGYW [16]. Other variants of this hot-spot motif, such as WRCH (or DGYW) have also been proposed (REF – ROGOZIN and DIAZ). Such sequences are highly over-represented in Ig Complementary Determining Regions (CDRs), which are thus considered to have an increased mutations probability [17, 18], while FWRs have a decreased frequency [19, 20]. This hot-spot motif can at least partially explained by the observed in vitro binding preference for AID to WRC (or GYW) motifs [16]. When analyzing sets of experimentally-observed Ig sequences, the observed increased mutation frequency at these hot-spots is due to factors beyond AID targeting. An important element affecting the observed mutation spectrum in different regions is positive and negative selection of the B cells carrying the resulting mutations. Negative selection is the death or editing (in the case of the L chain) of B cells carrying low affinity, non-functional or self-reactive receptors [2, 21]. It has been estimated that ~50% of replacement mutations in the framework regions are subject to negative selection [22]. Positive selection is the increased expansion rate of B cells with a higher affinity receptor for the introduced antigen [4], most likely due to a decreased death rate [5] [23]. Thus mutations adversely affecting the structure of functionality of the receptors are expected to be purged from the peripheral B cell population, while mutation enhancing antigen binding are expected to be over-represented. This will increase the observed mutation frequency in CDR regions and decrease the observed mutation frequency in structurally important framework (FWR) regions [24]. A final element affecting the observed mutation frequency is the long term evolution of the Ig germline sequence.

We here analyze a large collection of over 12,000 mutated sequences, carrying over 80,000 mutations and propose a new statistical model to analyze the mutation probability as a function of the surrounding sequence. We show using this model that the SHM targeting pattern is different than previously reported. Most significantly, we show that beyond microsequence-specific targeting, there exists a position-specific effect on the mutation probability.

Methods

Immunoglobulin Sequences

Three sources of immunoglobulin sequences were used: the IMGT database [25], the KABAT database [26] and immunoglobulin sequences published in the NCBI database. An initial quality check was applied to all sequences by checking for length and homology in the V region to a published germline V sequence. Sequences of less than 280 nucleotides, or sequence with less than 80 % homology to any published [25] V germline gene were removed. We then removed all sequences with insertions or deletions when compared with

all similar enough V genes (above 80 %). In the immunoglobulin sequence, we only studied the V region, and ignored the CDR3 region to avoid any bias induced by junctional diversity

Mutation Computation

Mutations were computed by blasting each sequence with all IMGT germline V gene sequences in the appropriate organism. The blast was performed using the NCBI BLAST software [27]. The highest scoring V gene was selected and all the differences between the rearranged sequence and the germline in the first 280 nucleotide were defined as mutations.

Regression model for the sequence-specific mutation probability

We checked whether the appearance probability of a mutation could be explained solely on the base of the DNA sequence surrounding the mutation. We defined the signature of a mutation as the DNA quadruplet surrounding it, with either 2 positions before it and 1 after it or the opposite. We then built a regression model for the mutation probability. The approximated probability was computed as a product of the probability of the preceding di-nucleotides, the probability of the mutation by itself and the probability of a following single nucleotide. A similar inverted model was also tested. For example the expected value of AG(C->A)T was approximated as $P_1(AG) \cdot P_2(C \rightarrow A) \cdot P_3(T)$, and the expected value of A(G->T)GT was approximated by $P_1(A) \cdot P_2(G \rightarrow T) \cdot P_3(GT)$. We developed an independent regression model for the forward (mutation at second position) and backward (mutation at third position) models. The log of the values was used producing $\log_{10}(P(AG(C \rightarrow A)T)) \approx \log_{10}(P_1(AG)) + \log_{10}(P_2(C \rightarrow A)) + \log_{10}(P_3(T))$. Each regression model (forward and backward) had 768 observation (each quadruplet) and 32 variables (16 possible nucleotide couples, 12 possible mutations, and 4 possible single nucleotide probabilities). At a second stage, we simplified the model, where the di-nucleotide was replaced by a multiplication of two independent mutations (e.g. $P_1(AG) = P_{1a}(A)P_{1b}(G)$). The simplified model had 8 less variable producing a model with 24 independent variables.

Classification of mutations based on AID related motifs

An AID hot-spot “signature” is defined as an “RGYW” motif or its reverse complement motifs “WRCY” [28, 29]. We compared the mutation probabilities of all 48 different nucleotide 4-mers that fit these AID signatures to all others 1488 (1536–48) 4-mers. As described above, in our model a signature is composed of a di-nucleotide, a mutation and another nucleotide.

Results

Mutation frequencies in different hosts and chains

The nucleotides surrounding mutations in human and mouse H and L chains were extracted by blasting each reported immunoglobulin sequence V regions against all IMGT V germline genes [25]. We then detected all mutations and tallied the frequency of each mutation accompanied by four flanking nucleotides (two 5' and two 3'). The classical AID targeting motif (WRCY/RGYW) suggests that the most critical positions influencing targeting are the two 5' bases along with the first 3' base. Thus, for each mutation we focused on these surrounding bases and generated two quadruplets to account for the possibility of targeting on either strand. For example, an observed mutation at the middle position in ACAGA would be analyzed as the di-nucleotide AC, a mutation at A followed by the nucleotide G (or on the opposite strand as C, followed by a mutation at A, then GT). We write this as AC(A->T)G or C(A->T)GT. The frequency of such quadruplets was measured separately for human and mouse H and L chains (Figure 1). Each mutation quadruplet (e.g., AC(A->T)G) was normalized by the frequency of its underlying nucleotide quadruplet (e.g.,

ACAG), which we denote as its base quadruplet. Normalization was applied either at the entire sequence level, or locally per position along the sequence as shall be further discussed. Our model covers 1536 possible quadruplets, 768 ($4 \times 4 \times 12 \times 4$) with a mutation at the third position, which we denote as forward quadruplets and 768 with mutations at the second position, which we denote as reverse quadruplets. Different quadruplets have drastically different observed mutation frequencies (Figure 1 diagonal for histogram in log scale), with a distribution spanning almost 5 orders. This is consistent with the finding of a hierarchy of mutation described by Smith et al [30]. While a clear correlation was observed between the log of the quadruplet frequencies in the four chains tested (human H,L and mouse H,L), some combination had much higher log frequencies in one chain or the other (Figure 1). Furthermore, there is a significant scatter around the diagonal. Most of the difference between the different chains results from the variability in base quadruplets with a low appearance frequency in the germline sequences. The average Pearson correlation between the mutations frequencies in the different chains studied is 0.4. When limiting the analysis to quadruplet whose base quadruplets have at least 1,000 observations (1262 possibilities out of 1536), the average correlation coefficient rises to 0.59. When further limiting the comparison to base quadruplets with over 5,000 observations in each chain (635 possibilities out of 1536), the average Pearson coefficient rises to 0.71. Limiting the selection to even more frequent quadruplets does not increase the correlation. Thus the difference between chains is not fully explained by random fluctuations in rare quadruplets. These differences can reach two orders of magnitude and no bias was observed in the variance for any particular mutation quadruplet. These differences suggest that micro-sequence specificity is not the only element affecting the observed mutation pattern. Such differences were previously described and assigned to differential targeting of AID in different chains [31, 32]. We suggest here that they represent the combination of selection and a position-specific effect.

The signature of AID

AID initiates SHM by deaminating C, with a preference for the hot-spot WRCY/RGYW. We tested whether the observed mutation distribution in our database was consistent with this classical AID signature. We summed the mutations from human and mouse H and L chains and computed the average mutation probability. We then divided all quadruplets into AID-positive quadruplets (i.e. quadruplets that fit the AID hot-spot signature) and AID-negative quadruplets (all others), and computed their frequency distribution (Figure 2). Interestingly, we found that AID-positive quadruplets are actually composed of two populations; one group of motifs has significantly higher frequencies compared with AID-negative motifs, while the other group is no different from this background. This high-frequency subset is enriched for motifs containing C->G or T and G->A or C mutations. In contrast, C->A and G->T mutations are no more frequent than AID-negative motifs. The preference of C->T and G->A mutations is expected by the deamination function of AID [33] and the previously observed preference for transition mutations. However, it is noteworthy that C->G and G->C are almost as frequent. This could be explained by the codon composition of Ig germline sequences [13] that induce selection of mutations beyond C->T and G->A. Note that some AID-positive quadruplets with C->G or T and G->A or C mutations have a lower mutation probability than some AID-negative quadruplets, as can be observed by the right hand tail of the AID-negative distribution in Figure 2.

Probabilistic model for mutation frequency

In order to quantitatively explain the observed mutation frequencies; we developed a regression model of the quadruplet frequency as a combination of the independent probabilities of two nucleotides preceding the mutation, a single nucleotide following the mutation and the mutation itself. For example the frequency of AC(A->T)G was modeled by

the following probabilities product: $P(AC)P(A \rightarrow T)P(G)$. These probabilities were extracted using a linear regression on the log of the frequencies. The regression results show that such a model can explain over half the variability in the log of the quadruplet frequency (Figures 3 and 4).

As previously reported [17, 18], the last position in the AID hot-spot motif WRCY (or first position in RGYW) does not contribute very much to the observed mutation frequency (Figure 4 vs. 5). To test whether this was true in our dataset, we performed a quantitative comparison of a model that includes the last position to a model without it. The decrease in the residual variability of the log frequency was less than 10% ($F_{18,736}=27, p < 1.e-20$, (Sup mat Figure 1)). Although this was statistically significant, the effect is much less than other positions. These results suggest that the downstream nucleotide is less important to predicting the mutation probability than others, but cannot be neglected. On the other hand a model with independent probabilities for the two positions before the mutation (Sup mat Figure 1) had a significantly higher residual variance ($F_{18,736}=14, p < 1.e-20$). This suggests that the interaction between the two 5' nucleotides is important for targeting SHM/AID.

It is important to note that a large part of the variability in the proposed mutation model is explained by the specific nucleotide substitution, rather than the DNA motif surrounding the nucleotide position. In other words, the variance in the frequency of different substitutions that occur at a position selected for mutation is as high as the variance in the mutation frequencies across different positions. While mutations from a Cytosine to Thymine are highly over-represented, as expected from the biology of AID, mutations from Cytosine to Guanine or Adenine are actually underrepresented. We compared the frequencies of the 12 possible mutations with the frequencies obtained by Patrick Wilson on a set of 2,000 human H chains and obtained similar results in most of the mutations for the human H chain. The regression coefficient represents the contribution of a given nucleotide or mutation type to the total frequency of the quadruplet. Positive regression coefficients imply a high frequency of quadruplets and vice versa. A comparison of the mutation frequencies from the previous section as well as from Zheng et al. [34] to the our regression coefficients show that a good match exists for 10 out of the 12 possible mutations from one nucleotide to another: A->G, G->A, C->T, T->C, G->C and A->C have high mutation frequencies (in previous publications) and positive regression coefficients (in the current analysis). C->A, T->A, G->T and T->G have low mutation frequencies and negative regression coefficients. Only type two mutation C->G and A->T have a intermediate frequencies with negative coefficients.

The regression model predicts many of the known AID hot- and cold-spots, with the contribution of Adenines around the mutation being much more significant than any other nucleotide or dinucleotide. To summarize, while the AID signature may be in general overrepresented, a super-signature exists, which is significantly more over-represented than any other mutation frequency and it is composed of A(G->A)[CT]T or [AT]A(C->T)T. Moreover, signatures containing a C->A or G->T mutations are not over represented at all.

The effect of location within the V gene

The differences between the H and L chain and between the human and mouse quadruplet mutation frequencies suggests a mutation mechanism that is not directly mediated by the sequence. While selection may play a role as discussed below, we hypothesized that this effect was due at least in part to the position along the V gene. However, any position-specific effect may be confounded by micro-sequence specificity as it is well-know that AID hotspots cluster in the CDRs. To resolve this problem, a two way ANOVA model was built. We split the V region into 10 base windows, and computed the probability of each quadruplet in each window. For example, the frequency of AG(C->A)T in window 5 (where the position is determined by the nucleotide mutating) was approximated as the frequency

AGCT in the same window multiplied by $P_1(AG) * P_2(C \rightarrow A) * P_3(T)$ multiplied by the total mutation probability in this window $P(\text{mutations in window } 5)$.

Taking the log of the two sides yields:

$$\begin{aligned} & \log_{10}(F(AG(C \rightarrow A)T, 5)) - \log_{10}(F(AGCT, 5)) = \\ & = \log_{10}(P_1(AG)) + \log_{10}(P_2(C \rightarrow A)) + \log_{10}(P_3(T)) + \log_{10}(P(\text{mutations}, 5)), \end{aligned}$$

where F are the appropriate frequencies. A similar approach was used for all quadruplets in all windows. The difference between the log of the mutation quadruplet frequency and the nucleotide quadruplet frequency was introduced into a two way ANOVA model. The results of the ANOVA were highly significant for both the quadruplet and the position effect ($p < 1.e-10$). When looking at the coefficient of each position in the V gene a clear and structured order appears with a higher mutation frequency in CDRs and a lower mutation frequency in FWR regions (Figure 6). Such a location effect might be due to the negative selection of mutations in framework regions (or positive selection of mutations in the CDRs). In order to test the possible effect of selection, we repeated the same analysis including only silent mutations, and again found a significant position-specific effect ($p < 1.e-3$) in all chains except for the mouse H chain. In the case of silent mutations there is no clear bias of high mutation frequency in the CDR regions (Sup mat Figure 2). Thus, we conclude that the observed position-specific effect is not fully due to selection. Once the position along the germline V gene is incorporated within the ANOVA framework, the regression model explains over 75% of the observed mutation frequency (F test $F_{28,736}=27$, $p < 1.e-20$). A more limited effect was recently proposed based on the distance between mutations [35].

Discussion

We have here proposed a new statistical formalism to analyze the effect of the surrounding nucleotide sequence on the mutation probability. We have shown that a model composed of multiplication of the probability of preceding di-nucleotides, the probability of each mutation at a given nucleotide and the probability of the following single nucleotide can explain most of the observed variance in the observed mutation frequencies among a large set of sequences. This quadruplet based model explains more than half the differences in mutations rates among quadruplets. The proposed model highlights that the main element affecting the mutation probability is not the flanking region, but the mutation itself, with $C \leftrightarrow T$ and $A \leftrightarrow G$ probability 10 times higher than the average mutation probability. While $C \rightarrow T$ and $G \rightarrow A$ transitions are expected from the AID targeting model (as they are the direct result of replicating over AID-induced mismatches), $A \rightarrow G$ and $T \rightarrow C$ may be the result of selection. Indeed, when only silent mutations are incorporated in the analysis, the contribution of $A \rightarrow G$ and $T \rightarrow C$ to the mutation frequency is much lower ($\sim 3.5\times$) compared with $C \rightarrow T$ and $G \rightarrow A$. Within the context of the AID targeting signature, $C \rightarrow G$ substitutions are almost as probable as $C \rightarrow T$, and similarly $G \rightarrow C$ are almost as probable as $G \rightarrow A$. $C \rightarrow A$ and $G \rightarrow T$, on the other hand, are no more probable than any other mutation even when surrounded by an AID targeting motif. This is true both for a model that only includes mutations at C (and G), and when developing a general model incorporating mutations at all positions (Sup mat Figure 3).

The contribution of the di-nucleotide preceding the mutation is significantly more important than the one nucleotide downstream of the mutation. Within the first di-nucleotide, the model containing the correlation between the two nucleotides is slightly better than the one not including such correlations. However, when one focuses on C mutations in the coding strand (or G mutations in the opposite strand), the correlation between the two first (or last

two) nucleotides becomes significant. For example, the TG di-nucleotide is much more frequent than the TA nucleotide, while AA and AG have similar frequencies. A similar phenomenon was recently observed by Zukerman et al [36].

Our sequence-based model explains more than half, but not all, the differences in the mutation frequencies among quadruplets. This is clearly seen by the large differences observed between the mutation probabilities for the same quadruplet in different chains. We show that half of the remaining variability can be explained by a position effect that is not affected by the sequence distribution at each position. This effect can explain the large difference between mutation frequencies in CDR and FWR regions that is not fully explained by the local sequence context. This position effect is a combination of selection (and its differential impact on CDR and FW regions) and a pure position effect as can be observed in the continued significance of position as a predictor when only considering silent mutations.

This position effect may be affected by structural elements, such as possible different methylation profiles in different regions [37]. Another possibility is long range effect of AID. We have here only studied quadruplet based models of mutation probability. If AID targeting affects the mutation probability more than 2 or 3 positions beyond the initially targeted Cytosine then the entire CDR region may have a higher mutation probability that is somewhat different from what we estimate here.

The current analysis was performed on a large database of sequences currently deposited in the IMGT and NCBI databases as well as sequences from the older KABAT database. These sequences were obtained in many different ways. Thus we do not expect a systemic mutation pattern to emerge from the experimental procedure and sequencing method. However such a possibility must always be taken into account in the explanation of the observed mutation pattern. Another possible caveat of the current analysis is the definition of the germline V genes. If some V gene alleles are not included in the IMGT definition then the differences between these V genes and the most similar V gene will be defined as a consistent mutation patterns. Since no consistent mutations appeared in the sequences at the same position, we assume that even if such alleles exist, their effect on the results is limited.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Bibliography

1. Maizels N. Immunoglobulin gene diversification. *Annu Rev Genet.* 2005; 39:23–46. [PubMed: 16285851]
2. Prak EL, Weigert M. Light chain replacement: a new model for antibody gene rearrangement. *J Exp Med.* 1995; 182:541–548. [PubMed: 7629511]
3. Di Noia JM, Neuberger MS. Molecular mechanisms of antibody somatic hypermutation. *Annu Rev Biochem.* 2007; 76:1–22. [PubMed: 17328676]
4. Betz AG, Rada C, Pannell R, Milstein C, Neuberger MS. Passenger transgenes reveal intrinsic specificity of the antibody hypermutation mechanism: clustering, polarity, and specific hot spots. *Proc Natl Acad Sci U S A.* 1993; 90:2385–2388. [PubMed: 8460148]
5. Kleinstein SH, Louzoun Y, Shlomchik MJ. Estimating hypermutation rates from clonal tree data. *J Immunol.* 2003; 171:4639–4649. [PubMed: 14568938]
6. MacLennan IC. Germinal centers still hold secrets. *Immunity.* 2005; 22:656–657. [PubMed: 15963779]

7. Muramatsu M, Kinoshita K, Fagarasan S, Yamada S, Shinkai Y, Honjo T. Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell*. 2000; 102:553–563. [PubMed: 11007474]
8. Revy P, Muto T, Levy Y, Geissmann F, Plebani A, Sanal O, Catalan N, Forveille M, Dufourcq-Labelouse R, Gennery A, Tezcan I, Ersoy F, Kayserili H, Ugazio AG, Brousse N, Muramatsu M, Notarangelo LD, Kinoshita K, Honjo T, Fischer A, Durandy A. Activation-induced cytidine deaminase (AID) deficiency causes the autosomal recessive form of the Hyper-IgM syndrome (HIGM2). *Cell*. 2000; 102:565–575. [PubMed: 11007475]
9. Peled JU, Kuang FL, Iglesias-Ussel MD, Roa S, Kalis SL, Goodman MF, Scharff MD. The biochemistry of somatic hypermutation. *Annu Rev Immunol*. 2008; 26:481–511. [PubMed: 18304001]
10. Odegard VH, Schatz DG. Targeting of somatic hypermutation. *Nat Rev Immunol*. 2006; 6:573–583. [PubMed: 16868548]
11. Delbos F, Aoufouchi S, Faili A, Weill JC, Reynaud CA. DNA polymerase eta is the sole contributor of A/T modifications during immunoglobulin gene hypermutation in the mouse. *J Exp Med*. 2007; 204:17–23. [PubMed: 17190840]
12. Rogozin IB, Pavlov YI, Bebenek K, Matsuda T, Kunkel TA. Somatic mutation hotspots correlate with DNA polymerase eta error spectrum. *Nat Immunol*. 2001; 2:530–536. [PubMed: 11376340]
13. Wilson TM, Vaisman A, Martomo SA, Sullivan P, Lan L, Hanaoka F, Yasui A, Woodgate R, Gearhart PJ. MSH2-MSH6 stimulates DNA polymerase eta, suggesting a role for A:T mutations in antibody genes. *J Exp Med*. 2005; 201:637–645. [PubMed: 15710654]
14. Zeng X, Winter DB, Kasmer C, Kraemer KH, Lehmann AR, Gearhart PJ. DNA polymerase eta is an A-T mutator in somatic hypermutation of immunoglobulin variable genes. *Nat Immunol*. 2001; 2:537–541. [PubMed: 11376341]
15. Liu M, Duke JL, Richter DJ, Vinuesa CG, Goodnow CC, Kleinstein SH, Schatz DG. Two levels of protection for the B cell genome during somatic hypermutation. *Nature*. 2008; 451:841–845. [PubMed: 18273020]
16. Rogozin IB, Diaz M. Cutting edge: DGYW/WRCH is a better predictor of mutability at G:C bases in Ig hypermutation than the widely accepted RGYW/WRCY motif and probably reflects a two-step activation-induced cytidine deaminase-triggered process. *J Immunol*. 2004; 172:3382–3384. [PubMed: 15004135]
17. Pham P, Bransteitter R, Petruska J, Goodman MF. Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. *Nature*. 2003; 424:103–107. [PubMed: 12819663]
18. Yu K, Huang FT, Lieber MR. DNA substrate length and surrounding sequence affect the activation-induced deaminase activity at cytidine. *J Biol Chem*. 2004; 279:6496–6500. [PubMed: 14645244]
19. Kepler TB. Codon bias and plasticity in immunoglobulins. *Mol Biol Evol*. 1997; 14:637–643. [PubMed: 9190065]
20. Shapiro GS, Wysocki LJ. DNA target motifs of somatic mutagenesis in antibody genes. *Crit Rev Immunol*. 2002; 22:183–200. [PubMed: 12498382]
21. Melamed D, Benschop RJ, Cambier JC, Nemazee D. Developmental regulation of B lymphocyte immune tolerance compartmentalizes clonal selection from receptor selection. *Cell*. 1998; 92:173–182. [PubMed: 9458042]
22. Shlomchik, MJ.; Litwin, S.; Weigert, M. The influence of somatic mutation on clonal expansion. *Prog Immunol; Proceedings of the Seventh International Congress of Immunology*; 1990. p. 415
23. Anderson SM, Khalil A, Uduman M, Hershsberg U, Louzoun Y, Haberman AM, Kleinstein SH, Shlomchik MJ. Taking advantage: high-affinity B cells in the germinal center have lower death rates, but similar rates of division, compared to low-affinity cells. *J Immunol*. 2009; 183:7314–7325. [PubMed: 19917681]
24. Celada F, Seiden PE. Affinity maturation and hypermutation in a simulation of the humoral immune response. *Eur J Immunol*. 1996; 26:1350–1358. [PubMed: 8647216]
25. Lefranc, MP. IMGT, The International ImMunoGeneTics Information System. *Methods Mol Biol*. 2004. p. 27-49.<http://imgt.cines.fr>

26. Kabat EA, Wu TT. Identical V region amino acid sequences and segments of sequences in antibodies of different specificities. Relative contributions of VH and VL genes, minigenes, and complementarity-determining regions to binding of antibody-combining sites. *J Immunol.* 1991; 147:1709–1719. [PubMed: 1908882]
27. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990; 215:403–410. [PubMed: 2231712]
28. Rogozin IB, Kolchanov NA. Somatic hypermutagenesis in immunoglobulin genes. II. Influence of neighbouring base sequences on mutagenesis. *Biochim Biophys Acta.* 1992; 1171:11–18. [PubMed: 1420357]
29. Milstein C, Neuberger MS, Staden R. Both DNA strands of antibody genes are hypermutation targets. *Proc Natl Acad Sci U S A.* 1998; 95:8791–8794. [PubMed: 9671757]
30. Smith DS, Creadon G, Jena PK, Portanova JP, Kotzin BL, Wysocki LJ. Di- and trinucleotide target preferences of somatic mutagenesis in normal and autoreactive B cells. *J Immunol.* 1996; 156:2642–2652. [PubMed: 8786330]
31. Chezari I, Lobel-Lavi L, Steinitz M, Laskov R. Ongoing somatic hypermutation of the rearranged VH but not of the V-lambda gene in EBV-transformed rheumatoid factor-producing lymphoblastoid cell line. *Molecular Immunology.* 2008; 46:80–90. [PubMed: 18718665]
32. Griffiths GM, Berek C, Kaartinen M, Milstein C. Somatic mutation and the maturation of immune response to 2-phenyl oxazolone. *Nature.* 1984; 312:271–275. [PubMed: 6504141]
33. Petersen S, Casellas R, Reina-San-Martin B, Chen HT, Difilippantonio MJ, Wilson PC, Hanitsch L, Celeste A, Muramatsu M, Pilch DR, Redon C, Ried T, Bonner WM, Honjo T, Nussenzweig MC, Nussenzweig A. AID is required to initiate Nbs1/gamma-H2AX focus formation and mutations at sites of class switching. *Nature.* 2001; 414:660–665. [PubMed: 11740565]
34. Zheng NY, Wilson K, Jared M, Wilson PC. Intricate targeting of immunoglobulin somatic hypermutation maximizes the efficiency of affinity maturation. *J Exp Med.* 2005; 201:1467–1478. [PubMed: 15867095]
35. MacCarthy T, Kalis SL, Roa S, Pham P, Goodman MF, Scharff MD, Bergman A. V-region mutation in vitro, in vivo, and in silico reveal the importance of the enzymatic properties of AID and the sequence environment. *Proc Natl Acad Sci U S A.* 2009; 106:8629–8634. [PubMed: 19443686]
36. Zuckerman NS, Hazanov H, Barak M, Edelman H, Hess S, Shcolnik H, Dunn-Walters D, Mehr R. Somatic hypermutation and antigen-driven selection of B cells are altered in autoimmune diseases. *J Autoimmun.* 2010; 35:325–335. [PubMed: 20727711]
37. Larijani M, Frieder D, Sonbuchner TM, Bransteitter R, Goodman MF, Bouhassira EE, Scharff MD, Martin A. Methylation protects cytidines from AID-mediated deamination. *Mol Immunol.* 2005; 42:599–604. [PubMed: 15607819]

Figure 1. Distribution of mutation frequencies in human and mouse H and L chain quadruplets
 Plots along the diagonal represent the overall distribution in each category (Human H, Human L, Mouse H, Mouse L). Scatter plots above the diagonal show the common distribution of quadruplet frequencies in the different categories. For example, each pixel in the second drawing of the first row represents the mutation frequency of a quadruplet in the Human H and Human L chains. The plots below the diagonal represent the same information as a two dimensional heat-map.

Figure 2. Histogram of mutation frequency in quadruplets based on their categorization into two subsets of AID-positive motifs and all other motifs

Each histogram was normalized to a total sum of 1. AID-positive motifs (black lines) have, on average, a mutation rate that is one order of magnitude higher than AID-negative motifs (grey line). Within the set of AID-positive motifs (dashed lines), one can clearly see an increased mutation frequency in motifs containing C->GT or G->AC mutations.

Figure 3. Observed vs. Predicted average log mutation frequency over all Ig chains studied

The predicted mutation frequencies were computed using a regression model, where the frequency of each motif is determined by: (1) the pair of nucleotides before the mutation, the mutation itself and the nucleotide following it (stars), or (2) a single nucleotide, the mutation and the pair following it (circles). A clear correlation between the predicted and observed results can be seen. However, the prediction only explains approximately half the variance. Thus other sources of variation must still exist.

Figure 4. Coefficients of regression model for the prediction of the mutation frequency based on a model where the probability of mutation is determined by the multiplication of the mutation probabilities around the point of mutation

The model was transferred to log space and the coefficients were computed using a linear regression. We have used two models: One where the mutation occurred at the third position (top row) and one where it occurred at the second position (bottom row). For example, the coefficients $P(AC)$, $P(G \rightarrow C)$, $P(T)$ would be on the top row. Each subplot is the contribution of a nucleotide or a mutation to the probability that a given quadruplet would appear. Note that different subplots have different scales. Bars highlighted in black are consistent with the classical AID motif hotspots, while white bars represent cold spots.

Figure 5. Coefficients of regression model for the prediction of the mutation frequency based on a model where the probability of mutation is determined by the multiplications of the mutation probabilities around the point of mutation with only three positions incorporated (instead of four)

The subplots are similar to figure 4, but the single nucleotide that follows the mutated position was removed from the analysis. As can be seen, the results are highly similar to Figure 4 although the model is simpler.

Figure 6. Effect of position on mutation frequency after all other elements were taken into account

The average mutation frequency of each quadruplet was normalized to one, and the mutation frequency was computed in each 10 nucleotide group along the different V genes. One can see that the mutation frequency in the FWR regions is lower than in the CDRs. This difference is not due to a codon bias or to AID motifs since each quadruplet mutation frequency was individually normalized to one.