HHS Public Access

Author manuscript

Genes Immun. Author manuscript; available in PMC 2013 April 01.

Published in final edited form as:

Genes Immun. 2012 October ; 13(7): 523–529. doi:10.1038/gene.2012.28.

Location and length distribution of somatic hypermutation-associated DNA insertions and deletions reveals regions of antibody structural plasticity

Bryan S. Briney[1], **Jordan R. Willis**[3], and **James E. Crowe Jr.**[1,2,*]

[1]Department of Pathology, Microbiology and Immunology, Vanderbilt University Medical Center, Nashville, Tennessee, USA

[2]Department of Pediatrics, Vanderbilt University Medical Center, Nashville, Tennessee, USA

[3]Chemical and Physical Biology Program, Vanderbilt University Medical Center, Nashville, Tennessee, USA

Abstract

Following the initial diversity generated by V(D)J recombination, somatic hypermutation is the principal mechanism for producing further antibody repertoire diversity in antigen-experienced B cells. While somatic hypermutation typically results in single nucleotide substitutions, the infrequent incorporation of genetic insertions and deletions has also been associated with the somatic hypermutation process. We used high throughput antibody sequencing to determine the sequence of thousands of antibody genes containing somatic hypermutation-associated insertions and deletions (SHA indels), which revealed significant differences between the location of SHA indels and somatic mutations. Further, we identified a cluster of insertions and deletions in the antibody framework 3 region which corresponds to the hypervariable region 4 (HV4) in T cell receptors. We propose that this HV4-like region, identified by SHA indel analysis, represents a region of under-appreciated affinity maturation potential. Finally, through analysis of both location and length distribution of SHA indels, we have determined regions of structural plasticity within the antibody protein.

INTRODUCTION

Generation of a diverse antibody repertoire begins with the recombination of variable (V), diversity (D) and joining (J) segments into complete antibody recombinants.[1] Following recombination, diversity is further increased through antigen-driven somatic hypermutation and class-switch recombination.[2–4] The somatic hypermutation process typically results in single nucleotide substitutions, although deletion of germline nucleic acids or insertion of non-germline nucleic acids does occur in association with somatic hypermutation.[5–7] In

*To whom correspondence should be addressed: James E. Crowe, Jr., MD, Vanderbilt Vaccine Center, Vanderbilt University Medical Center, 11475 MRB IV, 2213 Garland Avenue, Nashville, TN 37232-0417, USA, Telephone (615) 343-8064, Fax (615) 343-4456, james.crowe@vanderbilt.edu.

addition, increased frequency of somatic hypermutation-associated (SHA) insertions and deletions has been associated with disease states with B cell abnormalities, including rheumatoid arthritis and several cancers.[8–12] These insertions and deletions are relatively infrequent, with SHA insertions or deletions estimated to be present in 1.3 to 6.5% of circulating B cells.[5–7] Although infrequent, SHA insertion and deletion events add substantially to the diversity of the human antibody repertoire.[13–15]

SHA insertions and deletions also have been shown to play a critical role in the antibody response against viral and bacterial pathogens, including HIV-1, influenza virus, and *Streptococcus pneumoniae*.[16–21] Of particular interest, structural analysis of an SHA insertion in the anti-influenza antibody 2D1 identified a substantial structural alteration induced by the insertion.[17] This insertion, although located in a framework region, caused a large conformational change in a complementarity determining region (CDR), and allowed antibody-antigen interactions that were not possible without the insertion-induced conformational change. In addition to 2D1, the extremely broad and potently neutralizing HIV-1 antibody VRC01 contained a six nucleotide deletion in the CDR1 of the light chain.[18] This SHA deletion shortened the CDR1 loop, thereby removing steric constraints on the CDR2 loop and allowing direct interaction between the HIV antigen and the light chain CDR2 loop of VRC01.[22]

A definitive analysis of the frequency and structural localization of SHA insertions and deletions has been limited in the past by the low frequency of such events. Therefore, we used newly developed high-throughput nucleotide sequence analysis techniques to more thoroughly examine the subset of circulating antibody sequences that contain SHA insertions and deletions. Thorough analysis of the localization of SHA insertions and deletions revealed significant differences from the localization of conventional somatic mutations, suggesting that the structural constraints on SHA insertions and deletions differ from those acting on substitutions. Thus, this in-depth analysis of SHA insertions and deletions reveals regions of structural plasticity within the antibody protein.

RESULTS

Frequency of in-frame insertions and deletions associated with somatic hypermutation

We separately isolated naïve, IgM memory and IgG memory B cells from four healthy individuals using flow cytometric sorting, extracted total RNA and performed RT-PCR to amplify antibody genes from those cells, and subjected the resulting amplicons to high throughput DNA sequencing. After selecting only high-quality, non-redundant antibody sequences, we obtained a total of 294,232 naïve cell sequences, 161,313 IgM memory cell sequences and 94,841 IgG memory cell sequences.

We first analyzed the variable gene regions of each sequence for the presence of insertions and deletions that did not shift the reading frame. The frequency of non-frameshift insertions (1.8% and 1.9% for IgM memory and IgG memory, respectively; Figure 1A) and deletions (2.0% and 2.6%; Figure 1B) was similar in both memory cell subsets. The frequency of both insertions and deletions was reduced significantly in the naïve subset when compared to either IgM or IgG memory subsets. This finding is consistent with previous data suggesting

that non-frameshift insertions and deletions within the variable gene are associated with the somatic hypermutation process.[5–7]

Biased variable gene use in sequences containing somatic hypermutation-associated insertions and deletions

We next examined the sequences containing somatic hypermutation-associated insertions and deletions (hereafter designated SHA indels) for evidence of biased variable gene use. The V_H4 variable gene family was much more common in the population of sequences containing insertions (57%; Figure 1C) than in the total antibody repertoire (24%), while the V_H1 and V_H3 families were observed less frequently in the insertion population (7.6% and 28%, respectively) than in the total repertoire (19% and 43%). In the population of sequences containing non-frameshift deletions, both V_H3 and V_H4 families (51% and 38%; Figure 1D) were more frequent than in the total repertoire (43% and 24%). The population of sequences with deletions also displayed reduced use of the V_H1 family (4.8%) and V_H5 family (1.3%) compared to the total repertoire (19% and 6.8%, respectively).

Antibody sequences containing SHA indels were highly mutated

Since SHA indels are only rarely induced by the somatic hypermutation process, we hypothesized that antibody sequences containing SHA indels would display evidence of increased affinity maturation. We examined sequences containing SHA indels from both IgM memory (Figure 2A) and IgG memory (Figure 2B) subsets for evidence of increased affinity maturation. The total IgM memory subset displayed a mean mutation frequency of 12.6 mutations per sequence. Significantly higher mutation frequencies were seen in sequences from the IgM memory subset containing either SHA insertions (17.8; $p = 0.0017$) or SHA deletions (16.1; $p = 0.0022$). Sequences from the total IgG memory subset contained, on average, 14.9 mutations per antibody sequence. Much like the IgM memory subset, significantly higher mutation frequencies were seen in IgG memory sequences containing either SHA insertions (19.0; $p = 0.0056$) or SHA deletions (20.2; $p = 0.0015$).

Duplication of flanking sequence was observed in most non-frameshift SHA insertions

For the population of sequences containing non-frameshift SHA insertions, the sequence immediately adjacent to the insertion (on either the 5′ or 3′ side of the insertion, hereafter referred to as the “flanking region”) was analyzed for homology to the sequence of the insertion. Since sequences containing insertions are highly mutated (Figure 2A–B), it is possible that additional mutations in the insertion sequence or the flanking region accumulated following the insertion event. Therefore, flanking regions that identically matched the insertion sequence or flanking regions that contained a single mismatch were considered likely duplications. Although only 20% of insertion sequences identically matched flanking regions, 82% of sequences with insertions either identically matched or contained a single mismatch (Figure 2C), suggesting that sequence duplication was the primary mechanism of SHA insertions.

Mutations and SHA indels are differentially localized in framework and complementarity determining regions

Although the somatic hypermutation process, which typically results in point mutations, and SHA indels have been shown to be linked,[5–7] it is unclear whether the location of SHA indels is driven primarily by frequency of somatic hypermutation, or whether there are additional structural constraints that apply to SHA indels, but not substitutions. The somatic hypermutation process is known to preferentially target complementarity determining regions (CDRs) over framework regions (FRs) for a variety of reasons, including the increased presence of genetically encoded mutation hotspots. We analyzed the position of mutations and SHA indels (Figure 2D) and observed a significant increase in the fraction of SHA indels found in FRs and a decrease in the fraction of SHA indels found in CDRs when compared to mutations. 16% of mutations were found in FRs, while 24% of observed SHA indels were found in FRs ($p = 0.0075$). Conversely, 85% of mutations were found in CDRs, while only 76% of SHA indels were found in CDRs ($p = 0.0075$).

SHA indels revealed a hypervariable region 4 (HV4)-like region within FR3

Sequences containing non-frameshift SHA insertions or deletions were analyzed for the position of the insertion or deletion. Insertions and deletions were grouped by codon position, and the frequency of insertions (Figure 3A) or deletions (Figure 3B) at each codon position was determined. Non-frameshift insertions and deletions were both concentrated in CDRs and the portion of FRs in close proximity to CDRs. The most common codon position for insertions was codon 35, which is in CDR1. The most common codon position for deletions was codon 57, which is in CDR2. Surprisingly, there was a cluster of codons in FR3 (codons 81–87) that contained a high frequency of deletions. This cluster of deletions was located in the middle of framework region 3a (FR3a, codon positions 78–93), which corresponds to hypervariable region 4 (HV4, also sometimes referred to as CDR4) in T cell receptors. A less prominent cluster of insertions also was seen in a similar location in FR3a.

We next performed a comparative analysis of the relative frequency of insertions and deletions located in the sequences encoding the two CDRs and three FRs that constitute the heavy chain variable (V_H) gene (Figure 3E). The fraction of insertions observed in CDR1 was significantly higher than the fraction of deletions (47% of insertions were found in CDR1 while 29% of deletions were found in CDR1; $p = 0.008$), with a similar pattern seen in FR2 (13% of insertions and 7% of deletions; $p = 0.007$). In contrast, the fraction of deletions found in CDR2 was significantly higher than the fraction of insertions (50% of deletions and 26% of insertions; $p = 0.006$). There was no statistically distinguishable difference between the fraction of insertions and deletions in either FR1 or FR3.

Similar localization of long insertions and deletions

We again clustered non-frameshift insertions and deletions by codon position and calculated the mean insertion length (Figure 3C) for each codon position. As seen with insertion frequency, long insertions tended to concentrate in CDRs and in the portions of FRs that are immediately proximal to CDRs. An additional region containing a high concentration of long insertions and deletions was observed between codons 82 and 97 in FR3. Analysis of the mean insertion length of the three FRs (Figure 3F) revealed a trend toward longer

insertions in FR3 when compared to FR1 ($p = 0.13$) and a significant increase in insertion length in FR3 when compared to FR2 ($p < 0.01$). Analysis of the mean insertion length of the two CDRs revealed a significant increase in insertion length in CDR2 when compared to CDR1.

Analysis of deletion length at each codon position (Figure 3D) produced results that were similar to the insertion length distribution, with increased deletion lengths found in CDR1, CDR2 and FR3. A region between codons 82–97 contained extremely long deletion events, with codon 76 displaying a mean deletion length of 54 nucleotides and codon 78 displaying a mean deletion length of 45 nucleotides. Interestingly, the location of the region of long FR3 deletions corresponds to the location of increased FR3 deletion frequency. While the distribution of long insertions and deletions was largely similar in pattern, there was a short region between codons 51 and 55 in FR2 that contained very long deletions, and there was no corresponding region within FR2 for which long insertions were observed. Analysis of the mean deletion length of the three FRs (Figure 3G) revealed significantly longer deletions in FR2 and FR3 when compared to FR1 ($p < 0.05$). We also observed a small but significant increase in the deletion length in CDR1 when compared to CDR2. As with insertions, the CDR with lower alteration frequency (CDR2 for insertions, CDR1 for deletions) contained a significantly longer mean insertion or deletion length.

Structural display of insertion and deletion frequency and length distribution revealed regions of antibody structural plasticity

To gain a better understanding of the location of insertions and deletions in the context of a fully folded antibody protein, we mapped the frequency and length distribution of both insertions and deletions onto a space-filling model of a representative antibody (Figure 3). The model we used was derived from crystallographic structural data for the human influenza virus specific monoclonal antibody (mAb) 2D1 that we had isolated in our laboratory and previously reported.[17] Insertion or deletion frequency was determined by calculating the $\log_{10}$ of the frequency for each codon position and represented as a blue_white_red gradient on the surface of the mAb 2D1 structure. Insertion and deletion frequency hotspots were observed at the top of the protein, with peak insertion and deletion frequencies appearing near the apex of the CDR1 and CDR2 loops. The side orientation revealed a reduced insertion and deletion frequency in the highly structured framework regions, with the lone framework hotspot occurring in a surface-exposed loop region of FR3.

Insertion and deletion length distribution was determined by calculating the $\log_2$ of the mean insertion length for each codon position and represented as a blue_white_red gradient on the surface of the mAb 2D1 structure. The longest insertions and deletions were focused in FR3, and were isolated to loop and short alpha-helical regions.

Long deletions were less frequent than long insertions and were tolerated poorly in CDRs

We examined the ability of the antibody repertoire to generate and maintain sequences with long insertions and deletions. We grouped both insertions and deletions by length (in nucleotides) and selected lengths for which we had at least 100 representative sequences with that length of insertion or deletion. The frequency of insertions and deletions was

plotted for each length (Figure 5A), which revealed a significantly higher frequency of long insertions when compared to frequency of long deletions.

We next investigated whether or not there was a structural reason for the greater tolerance of long insertions over long deletions. We again grouped all insertions and deletions by length (in nucleotides) and plotted the frequency of insertions (Figure 5B) or deletions (Figure 5C) by location in either FR or CDR. We found that both long and short insertions were concentrated in CDRs, with less than 30% of the longest insertion events occurring in FRs. In contrast, however, long deletions were highly concentrated in FRs, with 89% of the longest deletions occurring in FRs. This strong preference against long deletions in CDRs is likely due to the limited length of the CDR loops. Most CDR1 and CDR2 loops are only 8–9 amino acids long, which likely restricted the ability of these CDR loops to structurally accommodate long deletions.

DISCUSSION

In this report, we performed an extensive analysis of SHA indels in the human antibody repertoire. B cells encoding antibodies with SHA indels are unusual in the peripheral circulation, with less than 2% of antibody sequences containing such insertion or deletion events. Due to their rarity, a comprehensive analysis of SHA indels has been difficult in the past. We used high throughput sequencing to determine the location and length distribution of SHA indels; for the most part, the location of SHA indels was similar to that of conventional somatic mutations. However, we identified substantial differences in SHA indel location that were likely related to structural constraints that apply to SHA indels but do not apply to substitutions. Our analysis analyses revealed regions of antibody structural plasticity, *i.e.*, regions that were able to accommodate addition or subtraction of sequence without compromising structural integrity. With much effort being directed toward rational design of both antigens[18,23–25] and antibodies,[26] it is critical to understand the regions of the antibody molecule that can withstand extensive alteration while maintaining the desired structural conformation.

Analysis of length distribution revealed a preference for long SHA insertions over long SHA deletions, especially in CDRs. This preference was largely due to the virtual absence of deletions of more than four codons in CDRs. The low frequency of long deletions in CDRs can be traced to structural limitations of CDRs themselves. Since most CDRs are 8–9 amino acids in length, it is likely that CDR loops of fewer than three amino acids are structurally compromised.

Analysis of SHA indel location led to the identification of a cluster of SHA indels in a region of FR3 that corresponds to the HV4 region of TCRs. Recent crystallographic work on the anti-influenza antibody CR6261 has shown that a similarly positioned HV4-like region of FR3 directly contributed to antigen binding.[27] In the case of the anti-HIV antibody 21c, the HV4-like region uniquely contributes to binding of the antigen in complex with the primary host receptor protein.[28] Finally, the FR3a region of antibody heavy chains of the VH3 family interacts with Staphylococcal protein A, a known superantigen.[29] While the HV4-like region in FR3 identified in this report did not contain the same frequency of SHA

indels as CDR1 and CDR2, the presence a substantial cluster of SHA indel events in the HV4-like region suggested the existence of a region that has a heretofore under-appreciated ability to accommodate affinity maturation modifications.

MATERIALS AND METHODS

Sample Preparation and Sorting

Peripheral blood was obtained from healthy adult donors following informed consent, under a protocol approved by the Vanderbilt Institutional Review Board. Mononuclear cells from the blood of four donors were isolated by density gradient centrifugation with Histopaque 1077 (Sigma). Prior to staining, B cells were enriched by paramagnetic separation using microbeads conjugated with antibodies to CD19 (Miltenyi Biotec). Cells from particular B cell subsets were sorted as separate populations on a high speed sorting cytometer (FACSAria III; Becton Dickinson) using the following phenotypic markers, naïve B cells: $CD19^+/CD27^-/IgM^+/IgG^-/CD14^-/CD3^-$, IgM memory B cells: $CD19^+/CD27^+/IgM^+/IgG^-/CD14^-/CD3^-$ and IgG memory B cells: $CD19^+/CD27^+/IgM^-/IgG^+/CD14^-/CD3^-$. Total RNA was isolated from each sorted cell subset using a commercial RNA extraction kit (RNeasy; Qiagen) and stored at −80°C until analysis.

cDNA Synthesis and PCR Amplification of Antibody Genes

100 ng of each total RNA sample and 10 pmol of each RT-PCR primer (primers available upon request) were used in duplicate 50 μl RT-PCR reactions using the OneStep RT-PCR system (Qiagen). Thermal cycling was performed in a BioRad DNA Engine PTC-0200 thermal cycler using the following protocol: 50°C for 30:00, 95°C for 15:00, 35 cycles of (94°C for 0:45, 58°C for 0:45, 72°C for 2:00), 72°C for 10:00. 5 μl of each pooled RT-PCR reaction, 20 pmol of 454-adapter primers and 0.25 units of AmpliTaq Gold polymerase (Applied Biosystems) were used for each 454-Adapter PCR reaction, performed in quadruplicate. Thermal cycling was done as before, but for 10 cycles.

Amplicon Purification and Quantification

Amplicons were purified from the pooled 454-adapter PCR reactions using the Agencourt AMPure XP system (Beckman Coulter Genomics). Purified amplicons were quantified using a Qubit fluorometer (Invitrogen).

Amplicon Nucleotide Sequence Analysis

Quality control of the amplicon libraries and emulsion-based clonal amplification and sequencing on the 454 Genome Sequencer FLX Titanium system were performed by the W. M. Keck Center for Comparative and Functional Genomics at the University of Illinois at Urbana-Champaign, according to the manufacturer's instructions (454 Life Sciences). Signal processing and base calling were performed using the bundled 454 Data Analysis Software version 2.5.3 for amplicons.

Antibody Sequence Analysis

For germline gene assignments and initial analysis, the FASTA files resulting from 454 sequencing were submitted to the IMGT HighV-Quest webserver (IMGT, the international ImMunoGeneTics information system; www.imgt.org; founder and director: Marie-Paule LeFranc, Montpellier, France). Antibody sequences returned from IMGT were considered to be “high-quality” sequences if they met the following requirements: sequence length of at least 300 nt; identified variable and joining genes; an intact, in-frame recombination; and absence of stop codons or ambiguous nucleotide calls within the reading frame.

Data Analysis

All statistical analyses were performed with Graphpad Prism software. Three-dimensional antibody structural models were colored using MacPyMol and custom scripts.

Acknowledgments

This work was supported by NIH U01 AI 78407 and NIAID Contract HHSN272200900047C, and supported in part by the Vanderbilt CTSA grant UL1 RR024975-01 from NCRR/NIH. BSB was supported by NIH T32 HL069765, and JRW by NIH T32 AI060571. The authors thank all patients for participating in the study. We would like to especially thank Chris L. Wright and Alvaro G. Hernandez at the W. M. Keck Center for Comparative and Functional Genomics at the University of Illinois at Urbana-Champaign for performing the 454 sequencing. We are grateful to the IMGT team for its helpful collaboration and the analysis of nucleotide sequences on the IMGT/HighV-QUEST web portal.

References

1. Tonegawa S. Somatic generation of antibody diversity. Nature. 1983 Apr 14; 302(5909):575–581. [PubMed: 6300689]
2. Jackson SM, Wilson PC, James JA, Capra JD. Human B cell subsets. Adv Immunol. 2008; 98:151–224. [PubMed: 18772006]
3. Neuberger MS. Antibody diversification by somatic mutation: from Burnet onwards. Immunol Cell Biol. 2008; 86(2):124–132. [PubMed: 18180793]
4. Schroeder HW, Cavacini L. Structure and function of immunoglobulins. J Allergy Clin Immunol. 2010 Feb; 125(2 Suppl 2):S41–52. [PubMed: 20176268]
5. Wilson PC, de Bouteiller O, Liu Y, Potter K, Banchereau J, Capra JD, et al. Somatic hypermutation introduces insertions and deletions into immunoglobulin genes. J Exp Med. 1998; 187(1):59–70. [PubMed: 9419211]
6. Goossens T, Klein U, Küppers R. Frequent occurrence of deletions and duplications during somatic hypermutation: implications for oncogene translocations and heavy chain disease. Proc Natl Acad Sci USA. 1998 Mar 3; 95(5):2463–2468. [PubMed: 9482908]
7. Bemark M, Neuberger MS. By-products of immunoglobulin somatic hypermutation. Genes Chromosomes Cancer. 2003 Sep; 38(1):32–39. [PubMed: 12874784]
8. Küppers R, Rajewsky K, Zhao M, Simons G, Laumann R, Fischer R, et al. Hodgkin disease: Hodgkin and Reed-Sternberg cells picked from histological sections show clonal immunoglobulin gene rearrangements and appear to be derived from B cells at various stages of development. Proc Natl Acad Sci USA. 1994 Nov 8; 91(23):10962–10966. [PubMed: 7971992]
9. Klein U, Klein G, Ehlin-Henriksson B, Rajewsky K, Küppers R. Burkitt’s lymphoma is a malignancy of mature B cells expressing somatically mutated V region genes. Mol Med. 1995 Jul; 1(5):495–505. [PubMed: 8529116]
10. Kobrin C, Bendandi M, Kwak L. Novel secondary Ig VH gene rearrangement and in-frame Ig heavy chain complementarity-determining region III insertion/deletion variants in de novo follicular lymphoma. J Immunol. 2001 Feb 15; 166(4):2235–2243. [PubMed: 11160277]

11. Miura Y, Chu CC, Dines DM, Asnis SE, Furie RA, Chiorazzi N. Diversification of the Ig variable region gene repertoire of synovial B lymphocytes by nucleotide insertion and deletion. Mol Med. 2003 Apr; 9(5–8):166–174. [PubMed: 14571324]
12. Belessi CJ, Davi FB, Stamatopoulos KE, Degano M, Andreou TM, Moreno C, et al. IGHV gene insertions and deletions in chronic lymphocytic leukemia: “CLL-biased” deletions in a subset of cases with stereotyped receptors. Eur J Immunol. 2006 Jul; 36(7):1963–1974. [PubMed: 16783849]
13. Wilson PC, Liu Y, Banchereau J, Capra JD, Pascual V. Amino acid insertions and deletions contribute to diversify the human Ig repertoire. Immunol Rev. 1998; 162:143–151. [PubMed: 9602360]
14. de Wildt RM, Hoet RM, van Venrooij WJ, Tomlinson IM, Winter G. Analysis of heavy and light chain pairings indicates that receptor editing shapes the human antibody repertoire. J Mol Biol. 1999 Jan 22; 285(3):895–901. [PubMed: 9887257]
15. Reason DC, Zhou J. Codon insertion and deletion functions as a somatic diversification mechanism in human antibody repertoires. Biol Direct. 2006:1. [PubMed: 16542032]
16. Zhou J, Lottenbach KR, Barenkamp SJ, Reason DC. Somatic hypermutation and diverse immunoglobulin gene usage in the human antibody response to the capsular polysaccharide of Streptococcus pneumoniae Type 6B. Infect Immun. 2004 Jun; 72(6):3505–3514. [PubMed: 15155658]
17. Krause JC, Ekiert DC, Tumpey TM, Smith PB, Wilson IA, Crowe JE. An insertion mutation that distorts antibody binding site architecture enhances function of a human antibody. MBio. 2011; 2(1):e00345–10. [PubMed: 21304166]
18. Wu X, Yang Z-Y, Li Y, Hogerkorp C-M, Schief WR, Seaman MS, et al. Rational design of envelope identifies broadly neutralizing human monoclonal antibodies to HIV-1. Science. 2010 Aug 13; 329(5993):856–861. [PubMed: 20616233]
19. Walker LM, Phogat SK, Chan-Hui P-Y, Wagner D, Phung P, Goss JL, et al. Broad and potent neutralizing antibodies from an African donor reveal a new HIV-1 vaccine target. Science. 2009 Oct 9; 326(5950):285–289. [PubMed: 19729618]
20. Walker LM, Huber M, Doores KJ, Falkowska E, Pejchal R, Julien J-P, et al. Broad neutralization coverage of HIV by multiple highly potent antibodies. Nature. 2011 Sep 22; 477(7365):466–470. [PubMed: 21849977]
21. Pejchal R, Doores KJ, Walker LM, Khayat R, Huang P-S, Wang S-K, et al. A Potent and Broad Neutralizing Antibody Recognizes and Penetrates the HIV Glycan Shield. Science. 2011 Oct 13.
22. Zhou T, Georgiev I, Wu X, Yang Z-Y, Dai K, Finzi A, et al. Structural basis for broad and potent neutralization of HIV-1 by antibody VRC01. Science. 2010 Aug 13; 329(5993):811–817. [PubMed: 20616231]
23. Azoitei ML, Correia BE, Ban Y-EA, Carrico C, Kalyuzhniy O, Chen L, et al. Computation-guided backbone grafting of a discontinuous motif onto a protein scaffold. Science. 2011 Oct 21; 334(6054):373–376. [PubMed: 22021856]
24. Ofek G, Guenaga FJ, Schief WR, Skinner J, Baker D, Wyatt RT, et al. Elicitation of structure-specific antibodies by epitope scaffolds. Proc Natl Acad Sci USA. 2010 Oct 19; 107(42):17880–17887. [PubMed: 20876137]
25. Correia BE, Ban Y-EA, Holmes MA, Xu H, Ellingson K, Kraft Z, et al. Computational design of epitope-scaffolds allows induction of antibodies specific for a poorly immunogenic HIV vaccine epitope. Structure. 2010 Sep 8; 18(9):1116–1126. [PubMed: 20826338]
26. Diskin R, Scheid JF, Marcovecchio PM, West AP, Klein F, Gao H, et al. Increasing the Potency and Breadth of an HIV Antibody by Using Structure-Based Rational Design. Science. 2011 Oct 27.
27. Ekiert DC, Bhabha G, Elsliger M-A, Friesen RHE, Jongeneelen M, Throsby M, et al. Antibody recognition of a highly conserved influenza virus epitope. Science. 2009 Apr 10; 324(5924):246–251. [PubMed: 19251591]
28. Diskin R, Marcovecchio PM, Bjorkman PJ. Structure of a clade C HIV-1 gp120 bound to CD4 and CD4-induced antibody reveals anti-CD4 polyreactivity. Nat Struct Mol Biol. 2010 May; 17(5): 608–613. [PubMed: 20357769]

29. Potter KN, Li Y, Capra JD. Staphylococcal protein A simultaneously interacts with framework region 1, complementarity-determining region 2, and framework region 3 on human VH3-encoded Igs. J Immunol. 1996 Oct 1; 157(7):2982–2988. [PubMed: 8816406]

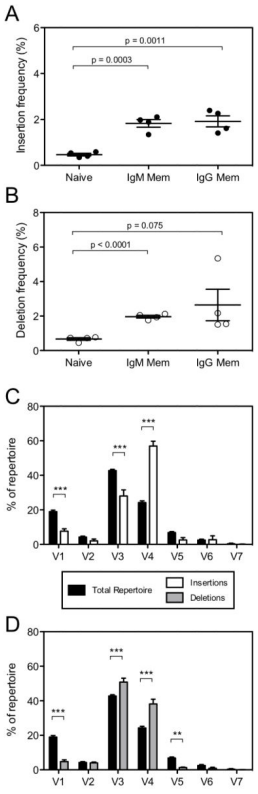

Figure 1. Frequency and variable gene use of sequences containing non-frameshift insertions or deletions

The frequency of (A) insertions or (B) deletions that were codon-length (*i.e.*, did not result in protein reading frame shift) was determined separately for the naïve, IgM memory and IgG memory subsets. Pairwise comparisons were made using a two-tailed Student's T test. The variable gene usage of VDJ gene recombinants containing insertions (C; white bars) or deletions (D; grey bars) was compared to the variable gene usage of the total repertoire (C and D; black bars). The p values for gene use were calculated using a two-way ANOVA with Bonferroni's post-test. ** = $p<0.01$; *** = $p<0.001$

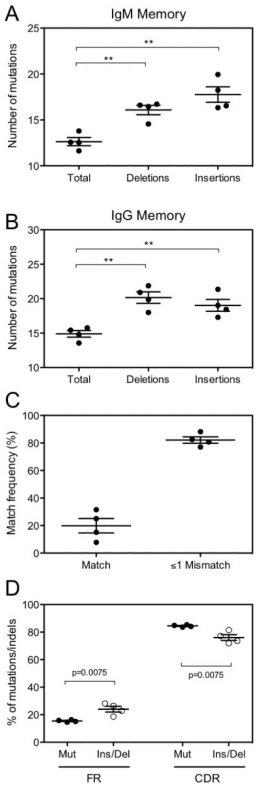

Figure 2. Sequences containing SHA indels are highly mutated

Sequences from the IgM memory (A) or IgG memory (B) subsets were segregated into three groups: the total sequence pool for each subset (total), sequences containing insertions, or sequences containing deletions. The mean number of mutations for each of these groups was calculated for each of four healthy donor sequence pools. The mean value ± SEM for four donors is shown. Pairwise comparisons were made using a two-tailed Student's T test. (C) For each insertion, the 5′ flanking sequence was analyzed for identity to the insertion sequence. The fraction of sequences with insertions that contained flanking regions that were a perfect match to the insertion sequence (match) or those that contained less than two mismatches (≤1 Mismatch) are plotted for each of four healthy donors. (D) Mutations and SHA indels (Ins/Del) were each grouped by localization in either framework (FR) or complementarity determining region (CDR). The p values for pairwise comparisons were determined by using a two-tailed Student's T test. ** = $p<0.01$

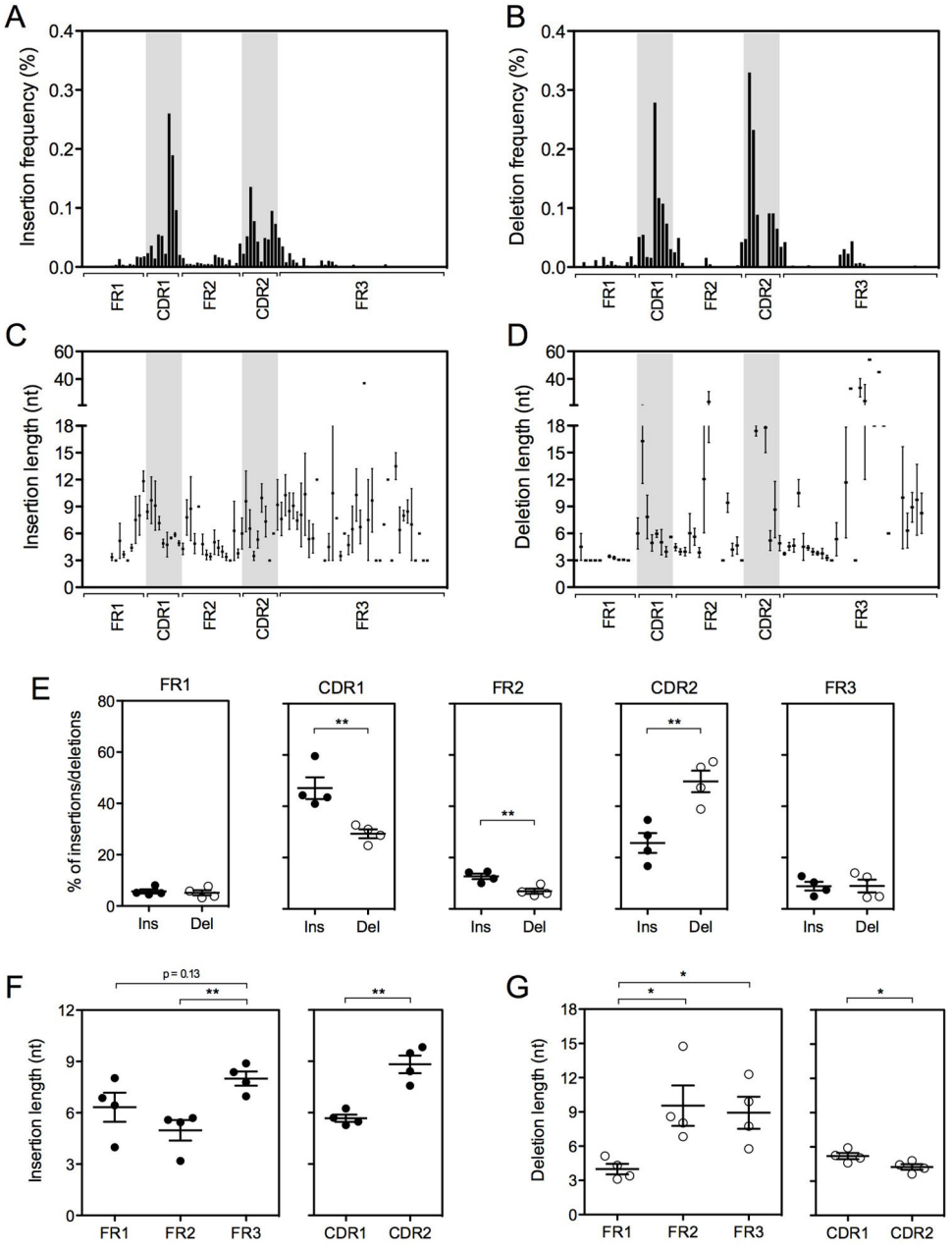

Figure 3. Genetic location and length distribution of non-frameshift insertions and deletions

The frequency of (A) insertions or (B) deletions at each codon of the variable gene reading frame was determined. The framework regions (FR) and complementarity determining regions (CDR) region were identified. The length distribution of insertions (C) or deletions (D) at each codon of the variable gene reading frame is shown. The mean value ± SEM for four donors is shown. (E) Comparison of the fraction of insertions (filled circles) or deletions (open circles) that were localized to each FR or CDR. The mean value ± SEM for four donors is shown. Comparison of the location of insertions (F) or deletions (G) for each FR or CDR is shown. The mean value ± SEM for four donors is shown. The p values for

pairwise comparisons were determined by using a two-tailed Student's T test. * = $p<0.05$; ** = $p<0.01$

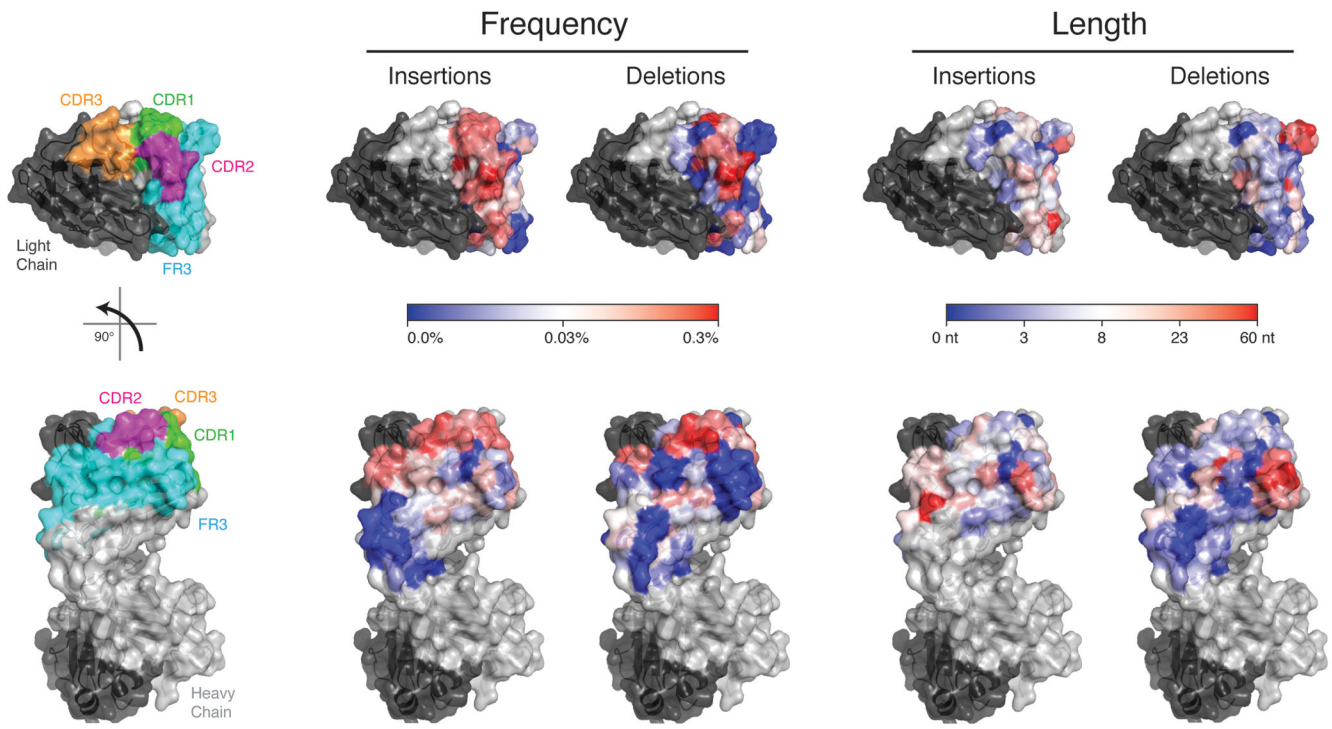

Figure 4. Structural location of non-frameshift insertions and deletions

A space-filling representation of the high resolution structure of the represent native human Fab del2D1 determined by x-ray crystallography (citation) is shown at left. The del2D1 antibody light chain is colored dark grey. The del2D1 antibody heavy chain CDR1 (green), CDR2 (magenta), CDR3 (orange) and FR3 (cyan) regions are indicated, with the remaining heavy chain regions colored light grey. The insertion and deletion frequency were determined for each codon position in the variable gene. In the middle two panels, the surface of del2D1 is colored to indicate insertion or deletion frequency. The mean insertion and deletion length were calculated for each codon position in the variable gene. In the right two panels, the surface of del2D1 is colored to indicate mean insertion or deletion length in nucleotides (nt).

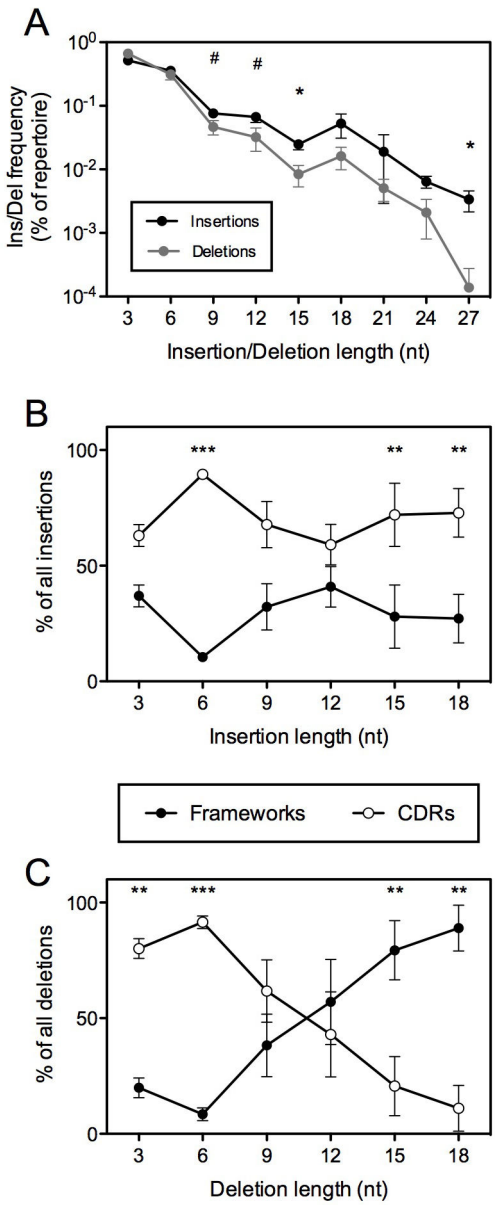

Figure 5. Difference in tolerance of long insertions and deletions in FRs and CDRs

(A) Non-frameshift insertions (black) or deletions (grey) were grouped by length and the frequency of each insertion or deletion length was calculated. The mean value ± SEM for four donors is shown. Non-frameshift insertions (B) or deletions (C) were grouped by length and the location of each insertion or deletion length to FRs (solid circles) or CDRs (open circles) was determined. The mean value ± SEM for four donors is shown. The p values were determined by using a two-way ANOVA with Bonferroni's post-test. # = $p<0.10$, * = $p<0.05$, ** = $p<0.01$, *** = $p<0.001$