Distribution Agreement

Signature:

_______________________________ _______________
Gordon A. Dale Date

Local and Genome-wide Gene Conversion contributes to secondary somatic diversification of antibody repertoires in Humans and Mice

By

Gordon A. Dale
Doctor of Philosophy

Graduate Division of Biological and Biomedical Science
Immunology and Molecular Pathogenesis

Joshy Jacob, PhD
Advisor

John Altman, PhD
Committee Member

David Steinhauer, PhD
Committee Member

Max Cooper, MD
Committee Member

Ignacio Sanz, MD
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Local and Genome-wide Gene Conversion contributes to secondary somatic diversification of antibody repertoires in Humans and Mice

By

Gordon A. Dale
B.S., University of Miami, 2011

Advisor: Joshy Jacob, PhD

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Graduate Division of Biological and Biomedical Science
Immunology and Molecular Pathogenesis
2019

Abstract

Local and Genome-wide Gene Conversion contributes to secondary somatic diversification of antibody repertoires in Humans and Mice

By Gordon A. Dale

The ability of the humoral immune system to respond to staggering number of antigens is critically dependent on the generation of somatic diversity in B cells. During B lymphopoiesis, discrete gene segments are somatically recombined through the process of V(D)J recombination, establishing the first stage of somatic diversification. In the periphery, antigen binding and cognate T cell help, license B cells to undergo further somatic diversification through the Darwinian microcosm of the germinal center, where antigen specific B cells undergo rounds of proliferation and mutation followed by selection for antigen binding.

Secondary somatic diversification is critically dependent of the master regulator activation-induced cytidine deaminase. The activity of this enzyme results in intrinsic mutability at cytosine residues, with further processing resulting in either the accumulation of untemplated point mutation through the canonical processes of somatic hypermutation or the templated mutations derived from gene conversion. It is widely held that somatic hypermutation results in untemplated point mutations in mice and humans, where as other species such as rabbits and chickens preferentially utilize templated mutations via gene conversion.

Here we demonstrate that somatic mutations observed in sequences at the antibody loci are derived from both local templated mutagenesis (i.e. donors within the antibody loci) as well as global templated mutagenesis (i.e. donors scattered in the genome). We demonstrate in two contexts the contribution of gene conversion to mutation processes during the germinal center reaction. First, we analyze micro-clusters of mutations (≥2 mutations in 8 bp) to demonstrate that mutations in close proximity exhibit significant linkage disequilibrium. We also demonstrate that non-immunoglobulin genes at the antibody locus share templates for micro-clustered mutations as do somatically mutated antibody sequences. Second, we analyzed large templated events at the heavy chain locus and demonstrated that templated mutagenesis results in templating from inter- and intra-chromosomal locations. Together, these studies provide evidence for the role of gene conversion in the somatic diversification of murine and human B cells.

Local and Genome-wide Gene Conversion contributes to secondary somatic diversification of antibody repertoires in Humans and Mice

By

Gordon A. Dale
B.S., University of Miami, 2011

Advisor: Joshy Jacob, PhD

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Graduate Division of Biological and Biomedical Science
Immunology and Molecular Pathogenesis
2019

Acknowledgements

This work is the product of many people and without them none of this would have been possible. I would like to pay special thanks to my parents, friends, and family – all of whom listened to and supported me on many a late night. I would also like to thank my mentor, my lab, and my collaborators for their invaluable insight, skills, and support for the work contained herein. Truly, this work is as much mine as it is all of theirs. Thank you.

Table of Contents

Preface

There are a number of correlates between evolution and the immune system. From clear examples such as the dynamics of host cells during an infection, to more nuanced ones such as the differentiation of hematopoietic stem cells into cells that fill specialized niches, immunology is rife with evolutionary axioms. This, of course, is obvious, as evolutionary processes are not relegated solely to the level of the organism. Indeed, lymphocyte populations are controlled by limited amounts of growth factors (interleukins), leukocytes undergo alteration and differentiation when entering a new environment, and even “predation” can be observed when a macrophage engulfs a foreign pathogen.

Similar to these examples, the work herein also examines an evolutionary axiom: that of mutation and diversification. In Darwinian theory, diversity is paramount, as it serves as the raw substrate upon which an organism better-suited to the environment can be selected. In the context of the immune system this process is undertaken by both T and B lymphocytes. In brief, both T and B lymphocytes possess the ability to generate diversity by stereotyped rearrangement of gene segments, known as VDJ recombination. When these segments combine, a new receptor is generated. During an infection, a few T and B lymphocytes will possess a receptor specific to antigen on a pathogen causing the infection. These specific T and B cells will be “selected” and permitted to further differentiate and proliferate. In doing so, these cells successfully fulfill the evolutionary paradigm of generating diversity and subsequent selection of fit progeny.

Uniquely, B cells undergo a second round of diversification known as somatic hypermutation. In this process, B cells that have generated functional receptors via VDJ recombination and have been “selected” to further differentiate are also permitted to engage in a program of mutagenesis and proliferation, such that daughter B cells accrue mutations in the receptor and from which a daughter cell that produces a receptor that best binds antigen is selected. It is here, during this second round of diversification, where the studies in this dissertation take place.

The corollaries with evolution are particularly important in this work. In part, this is due to the subject matter of mutation and diversification. In another part, this is due to what is known about the evolution of the adaptive immune system. And lastly in part, this is due to the evolution of the scientific body of knowledge. All three of these parts orbit the studies presented here, and provide the necessary context for the studies themselves, how the studies fit in the existing literature, and why these findings had been missed for the greater part of 30 years.

Chapter 1: Introduction

B cell diversification

The ability of the B cell to provide host protection is critically dependent of the ability of these cells to produce high-affinity, specific antibodies to virtually any antigen. In order to do so, B cells undergo two distinct stages of diversification. The first, primary diversification, occurs in the bone marrow during lymphopoiesis. Secondary diversification occurs after B cells have matured and emigrated from the bone marrow.

Primary Diversification and Antibody Structure

The B cell receptor's (BCR's) ability to bind to an exceedingly large repertoire of antigens is primarily due to the combinatorial diversity generated during B cell lymphopoiesis[1,2]. The BCR is a hetero-tetrameric protein that is composed of two identical heavy chains and two identical light chains[3-5]. Each chain is composed of both constant and variable regions[2-5]. The variable regions constitute the antigen binding interface of the antibody, while the constant region functions to link effector functions to antigen binding[2,3]. The variable region itself can be further divided into distinct domains known as complementarity binding regions (CDRs) and framework regions (FWRs)[3]. The CDRs are hypervariable loops that project out from the antibody and serve to directly contact antigen, whereas the FWRs serve as scaffolding to support the CDRs[3,6,7]. Each chain contains four FWR regions and three interspersed CDR regions and for a total of 16 and 12 per BCR, respectively[3,6,7]. As discussed later in depth, the CDRs and FWRs are sites for secondary diversification and accrue mutations to better facilitate antigen binding.

For both the heavy and the light chain, the variable region is the product of a B-cell specific combinatoric rearrangement of discrete gene segments at the IgH locus for the heavy chain and either the IgL or IgK locus for light chains[2]. B cell development is intrinsically linked to the process of generating these heavy and light chains and unsuccessful rearrangement leads to apoptosis.

The IgH locus, located on chromosome 14 in humans and chromosome 12 in mice, is composed of an array of variable (V), diversity (D), and joining (J) gene segments that are rearranged to produce a single variable domain in the heavy chain in a process termed VDJ recombination[8,9]. To produce a variable domain, a single V region, D region, and J region are brought together through the actions of recombination activated gene (RAG) 1 and 2[10-13]. Successful recombination of these gene segments is achieved by flanking regions surrounding each segment, known as recombination signal sequences (RSSs). Each RSS consists of a heptamer sequence (5'-CACAGTG-3'), followed by either a 12 bp or 23 bp spacer, followed by a nonamer sequence (5'-ACAAAAACC-3')[14]. Recombination can only occur between a sequence with an RSS with a 12 bp spacer and a 23 bp spacer[15,16]. At the IgH locus, V gene segments are 3' flanked by an RSS with a 23bp spacer, D segments are flanked 5' and 3' by RSSs with 12bp spacers, and J segments are 5' flanked with 23bp spacers[2]. This organizational property of the segments ensures correct recombination of one V, D, and J segment over the length of the locus. At the light chain loci, there are only V and J segments present. Thus, recombination at the IgK and IgL loci only involve the joining of V to J to produce a light chain variable region[17].

The diversity generated via the recombination of these discrete gene segments, while large, does not entirely account for the BCR's robust breadth of antigen binding. Additional diversity is also generated during the recombination event by the addition of palindromic (P) and random (N) nucleotides between each recombined gene segment. Recognition of the 12 and 23 bp RSSs by RAG-1 and -2 bring the gene segments to be combined proximal to one another.

Endonuclease activity by the RAG complex creates precise single-strand DNA breaks at sites 5’ of each RSS. This break creates a free 3’ hydroxyl group that subsequently attacks the phosphodiester bond on the opposite strand, creating two DNA hairpin structures[18,19]. DNA repair factors KU70 and KU80 bind to the DNA hairpins and recruit the DNA-PK:Artemis complex[20]. This complex opens the DNA hairpins via imprecise nicking[20]. When nicking occurs, complementary nucleotides present in the DNA hairpin are extended out, resulting in P-nucleotides. N-nucleotides are added via the activity of the lymphoid specific enzyme TdT[21,22]. At the free 3’ ends resulting from the cleavage of the DNA hairpins, TdT inserts random nucleotides[21,22]. Once base pairing occurs between the two gene segments following the activity of TdT, further endonuclease activity removes unpaired bases, which allows for filling in of single strand gaps by DNA polymerases and DNA ligase. The addition of P and N nucleotides at the junctions between the V(D)J rearrangement creates junctional diversity that further diversifies the BCR and results in the initial antigen specificities encoded by the naïve B cell repertoire.

Located downstream of each loci’s array of V, (D), and J’s are the coding regions for the antibody constant regions[2,3]. Unlike the antibody gene segments, these regions are not physically recombined to the variable region. At the light chain loci, there is only a single constant region at each locus, however, at the heavy chain locus there are five constant region segments – Cμ, Cδ, Cγ, Cα, and Cε[2,3]. All naïve B cells that have not undergone secondary diversification only express two of these constant regions - Cμ and Cδ - through alternative splicing[3]. The other constant regions will be discussed in greater detail during secondary diversification.

Each V region that is used during V(D)J recombination encodes two CDRs and three framework regions[2]. The third CDR is created by the junction of the VDJ in the heavy chain and the VJ junction in the light chain. As these sites are further diversified by the actions of Artemis and TdT, CDR3 exhibits the greatest diversity between the CDRs, followed by CDR2 and CDR1. At the Ig loci, V genes are organized into families by virtue of sequence similarity[9]. Within families, CDR1 is largely conserved with CDR2 exhibiting more diversity[23]. At the family level, the three encoded frameworks are largely conserved between members[23]. Unlike the other FWRs, the final FWR is encoded by the J region and the expressed constant region following rearrangement.

The V(D)J recombination process described above occurs at distinct stages of B cell development in a stereotyped manner that enforces B cell specificity such that a single B cell will encode a single BCR[24,25]. In the bone marrow, hematopoietic stem cells undergo self-renewal and differentiation into the multipotent progenitor cell, which expresses the receptor FLT3[26-28]. Bone marrow stromal cells express FLT3 ligand and engagement of FLT3:FLT3 ligand allows the multipotent progenitor cell to differentiate into the common lymphoid progenitor (CLP)[27,28]. The CLP expresses IL-7R and is IL-7 produced by the bone marrow stromal cells is required for continued development into an early pro-B cell[29,30]. At this stage of development, rearrangement begins at the heavy chain locus, with the D and J segments undergoing VDJ recombination[31-33]. Following DJ recombination, V is recombined at the late pro-B cell stage[34]. Upon rearrangement, the developing B cell is now a pre-B cell. At this stage the heavy chain is transcribed, translated and associates with the surrogate light chain proteins, VpreB and λ5, forming the pre-BCR[34,35]. The pre-BCR is expressed on the pre-B cell surface and tonic signaling through the receptor enables further development[34,35]. It is important to note that the V(D)J recombination process occurs at both alleles of the heavy chain during development but only one heavy chain is produced despite each locus rearranging independently. This is known as allelic exclusion and is controlled by successful expression of the pre-BCR[36,37]. In this way, if one locus has a non-productive rearrangement, due to junctional diversity, the cell has a second opportunity to produce a productive heavy chain.

Successful expression of the pre-BCR enables the pre-B cell to undergo a limited round of proliferation, increasing the number of cells with a successful heavy chain rearrangement before proceeding to light chain rearrangement[2,38-40]. During light chain rearrangement, the pre-BCR is internalized and the VJ rearrangement at the light chain loci proceeds. Unlike the heavy chain, the light chain is encoded at two loci, IgL and IgK. Rearrangement proceeds at the IgK locus first, and if unsuccessful, proceeds at the IgL locus[2]. Due to this difference in order, most antibodies are encoded by the IgK locus[2]. Successful rearrangement of the light chain allows the pre-B cell to transition to the immature B stage where the BCR is expressed on the cell surface. During the immature B cell stage, only BCRs expressing Cμ (IgM) are expressed. Development in a mature B cell occurs when the immature B cell begins coexpressing Cμ (IgM) and Cδ (IgD). The mature B cell then exits the bone marrow and enters the periphery where circulates throughout the vasculature and lymphatics to survey for antigen.

B cell activation and Secondary Diversification

While some mature B cells circulate, most mature naïve cells reside in the lymph nodes. There, they are awash with lymphatic drainage from distant sites of the body, which during the context of an infection contains draining antigen[41]. Afferent lymphatic vessels carrying antigen enter the lymph node and here antigen comes in contact with B cells in the primary lymphoid follicle[41]. The lymph node is organized such that B cells and T cells are localized to distinct regions. The B-cell rich primary lymphoid follicle is proximal to the draining afferent lymphatic vessel and followed by the paracortical region which primarily contains T cells[41]. This asymmetric division is maintained by chemokines secreted by follicular dendritic cells in the primary follicle and stromal cells in the paracortical area. B cells express the chemokine receptor CXCR5 which allows the B cell to migrate to secreted CXCL13 by the follicular dendritic cells[42,43]. Conversely, T cells express the chemokine receptor CCR7 which attracts the T cells to CCL21 that is produced by stromal cells in the paracortex[42,43].

The secondary diversification in the periphery is only permitted following two distinct steps. First, B cells must encounter a cognate antigen that binds the B cell receptor (antibody) on the cell's surface[44]. This antigen is internalized and shuttled to the lysosome where the antigen is degraded by proteases such as cathepsins. These degraded antigen peptide fragments are then loaded onto major histocompatibility complex class II (MHC-II) and the antigen-fragment-MHC-II complex is shuttled to the B cell surface[45,46]. During this time, the B cell upregulates the chemokine receptor CCR7 while maintaining expression of CXCR5 which allows the B cell to migrate from the primary lymphoid follicle to the border of the paracortical areas[47]. Separately, T cells are activated by interaction with dendritic cells, that like the B cell present engulfed antigen fragments on MHC-II[48,49]. Activation of the T cell results in the upregulation CXCR5 and continued expression of CCR7, which also permits the activated T cell to migrate to the T-B border[50].

At the border between these zones, the activated T cell and activated B cell meet. Recognition of the antigen-MHC complex on the B cell by an effector CD4 T cell through its T-cell receptor (TCR), causes the expression of CD40L on the T cell. CD40L binds to CD40 on the B cell and licenses the B cells[49]. Licensed B cells undergo proliferation at the T-B border and subsequently migrate to either the primary lymphoid follicle, where secondary diversification occurs, or to the medullary cords, a region of the lymph node where terminally differentiated B cells secrete antibody to leave the lymph node via efferent lymphatic vessels. Licensed B cells that migrate to the medullary cords differentiate into short lived antibody secreting cells called plasmablasts. Here, plasmablasts derived from the licensed B cells can secrete antigen-specific IgM antibody rapidly, conferring the first line of humoral defense while the secondary diversification process is underway.

B cells that instead migrate back to the primary lymphoid follicle are further induced to proliferate, forming a structure known as the germinal center. These B cells upregulate the transcription factor BCL-6 which controls the development and maintenance of the germinal center program[51-53]. Rapidly dividing B cells force resting B cells to the periphery of the germinal center creating the mantle zone of resting B cells. Within the germinal center, B cells alternate between a cycle of proliferation and selection[53,54]. Proliferating B cells in the germinal center (centroblasts) are found in the dark zone, whereas cells undergoing selection (centrocytes) are found in the light zone[53,54].

The germinal center functions as a Darwinian microcosm of repeated proliferation, mutation, and selection. This repeated cycle confers the B cell's prolific ability to generate a high-affinity specific antibody to a vast array of antigens. The germinal center cycle begins with antigen complexes that are present on the follicular dendritic cells in the germinal center. Centrocytes compete with one another for binding to limited amounts of these complexes. Those centrocytes that do receive signaling through their BCR, engulf antigen and present them on MHC-II. In the germinal center, selection is dynamic and the strength of positive selection is mediated by centrocyte interaction with T follicular helper cells (T_{FH})[55-57]. The higher the affinity for antigen in a given centrocyte, the more antigen is presented on MHC-II[55]. Thus, antigen-MHC-II complex density is directly related to the affinity of the BCR and density of antigen-

MHC-II correlates with the strength of interaction between the centrocyte and the T_{FH}. Engagement with the T_{FH} provides signals to the centrocyte via CD40L and secreted IL-21[58-60]. This promotes the expression of the chemokine receptor CXCR4 which allows centrocytes to become centroblasts and migrate to the dark zone where stromal cells secrete CXCL12[61].

The dark zone is the site of proliferation and diversification in the germinal center. Entry to the dark zone is concurrent with a mutagenic program known as somatic hypermutation, which is mediated by activity of the mutagenic enzyme activation-induced cytidine deaminase (AID), and DNA repair factors mutS homolog 2 (MSH2), exonuclease 1 (EXO1), uracil DNA glycosylase (UNG), and apurinic endonuclease (APE)[62-65]. These factors, along with others discussed later in greater detail, function to introduce both templated and untemplated to the Ig loci. This results a pool of somatic variants that return to the light zone and begin the selection and mutation process again.

Besides maturation of affinity, the actions of AID and other somatic hypermutation machinery can cause the BCR to switch constant regions, in an eponymous process termed class-switch recombination[62,66]. Recall that all naïve B cells express both IgM and IgD constant regions after maturing in the bone marrow, but other constant regions exist in the IgH locus, namely $C\gamma$, $C\alpha$, and $C\varepsilon$[3,66]. Each constant region serves as the effector for a specific antibody by engaging differential receptors that mediate discrete functions in the immune system[3]. Cytokines, released by T_{FH} cells in the germinal center differentially regulate chromatin accessibility to each of these sites in centrocytes[66,67]. Upon entering the dark zone, if these constant regions are open, they are subject to the activity of AID at specialized sequences upstream of each constant region known as switch regions[66]. Switch regions will be discussed in detail later but are particularly prone to double strand breaks. The formation of breaks at $C\mu$ and another switch region promotes a DNA repair process known as non-homologous end joining (NHEJ) which excises Cμ and all intervening DNA between it and the selected constant region[66,67]. The new constant region is then expressed with all BCRs on the switched B cell.

As mentioned earlier, selection is a dynamic process and centrocyte positive selection is linked to the levels of antigen-MHC-II on the cell surface. TFH cells provide a gradient of signal to centrocytes with the centrocytes expressing the most antigen-MHC-II receiving the strongest signal. Centrocytes that received the strongest signal in the light zone undergo a massive proliferative burst, also known a clonal burst, in the dark zone[55,56,68]. This impacts germinal center dynamics in two ways. First, the clonal burst reduces the number of competing B cells with different VDJ rearranged BCRs (referred to as clones). Thus, clonal bursts restrict the interclonal diversity of the germinal center by means of outcompeting. Clonal bursts occur with some frequency within germinal centers through they do not always lead to purifying selection[68]. Second, clonal bursts ensure that a clonal lineage that possess a BCR which far outclasses other clones in the germinal center is proportionally favored in the subsequent rounds of the diversifying-selection cycle. This maximizes the ability of the germinal center to produce a high affinity antibody while sacrificing clones that are more unlikely to be able to compete in subsequent rounds.

As the germinal center reaction progresses, centrocytes can exit the reaction and differentiate into either memory B cells or plasmablasts. The decision to adopt either of these cell fates is determined by two transcription factors Blimp1, and BACH2[69-72]. During the germinal center reaction, a small subset of centrocytes express Blimp1 which is repressed by the combined actions of BCL6 and BACH2[69,71,73]. Over time those cells who have the highest affinity express higher levels of Blimp1, which serves as the master regulator of plasma cell fate. At high levels

Blimp1 turns off transcription of BCL6 and Pax5, the germinal center, and B cell master regulators, respectively[72]. Centrocytes with slightly lower affinity appear to be selected to entry the memory B cell pool and exit the germinal center, although the mechanisms controlling memory B cell fate are still under investigation[54].

Cells that become plasmablasts will exit the germinal center, where they will adopt one of three fates. A subset of plasmablasts will reside in the medullary cords in the lymph node, where they will continue to secrete antibody for about a week before undergoing programmed cell death[74,75]. The remaining plasmablasts will enter the systemic circulation, where they continue to secrete antibody. A subset of these circulating plasmablasts will home to the bone marrow where they will compete for microenvironments that allow further differentiation into long lived plasma cells, which can secrete antigen-specific antibody for decades following antigen exposure without self-renewal[74-76].

Conversely, cells that undergo memory B cell differentiation are retained in the primary follicle where they await antigen re-exposure[77]. These cells are distinct from their naïve counterparts in two important ways. Primarily, these memory B cells may have undergone the process of somatic hypermutation such that they possess a higher affinity to antigen than other naïve B cells[77]. Upon antigen re-exposure, both naïve and memory B cells can be recruited to the germinal center for somatic hypermutation, however, memory B cells – possessing higher affinity – are likely to out compete naïve B cells via clonal burst[68,77]. Furthermore, memory B cells alter their metabolic profile such that they have decreased time to division after receiving proliferation signals allowing them to further compete against naïve B cells[78].

Mechanisms of Somatic Hypermutation

Activation-induced cytidine deaminase (AID) is an evolutionarily ancient member of the APOBEC family of cytidine deaminases which as a whole function in cellular RNA processing and anti-viral defense[79]. Unlike the other members of the APOBEC family, AID expression is restricted only to activated B cells due to its central role to the somatic hypermutation program[79]. Transcription of the AICDA gene occurs following antigen binding and activation of the B cell in the primary follicle[80,81]. AID mRNA is shuttled out of the nucleus where it is translated by free ribosomes in the cytosol. Once translated AID is subjected to various controls on its activity and localization. The site of AID activity necessary for hypermutation is the cell nucleus and as such entry of AID into the nucleus is tightly regulated. In the cytoplasm, AID binds to eEF1A1, which sequesters it in this compartment[82]. Shuttling of AID to the nucleus is dependent on activity from a class of nuclear importins known as karyopherins as well as the mRNA shuttling protein GANP[83-85]. Once in the nucleus, AID is both able to exert its activity but also is subject to two mechanisms of degradation – through a unknown mechanism of ubiquitination, as well as a ubiquitin-independent pathway facilitated by binding of AID to Reg-γ[86,87]. The sum of these regulatory steps permits ~10% of AID to be in the nucleus at a given time, with the remainder remaining in the cytoplasm[80].

The central role of AID in the somatic hypermutation program is highlighted through its ability to induce both templated and untemplated mutations following its activity[62,88]. As its name indicates, AID functions to deaminate cytosine nucleotides in single stranded DNA (ssDNA)[89]. This permits the activity of AID to be restricted to areas that are actively undergoing transcription, where the transcription bubble allows AID to access its substrate. To act on ssDNA, AID associates with the RNA binding protein ROD1, which guides AID to specific regions in the genome, including the IgH/IgL/IgK loci as well as to other regions of the genome that are “off-target” sites of AID activity such as BCL6, PAX5, CD83 and others[90]. ROD1 itself, is trapped at sites of bi-directional transcription from enhancers and promoters and there it establishes a loading site for AID[90]. Once loaded at the either the IgHV or the Ig(K/L)V, AID exhibits processivity and deaminates multiple cytosines on the cis strand of DNA[91]. Despite the widespread occurrence of cytosines at these genomic sites, AID exhibits preference for the sequence motif 5’-WRC-3’ which are known as AID hotspots[92,93]. Conversely, AID exhibits strong aversion to the motif 5’-SYC-3’, which constitute AID coldspots[93]. Hotspots and coldspots are differentially located within the variable region of the antibody, with hotspots being primarily found at CDRs and coldspots found primarily in the FWRs[94]. This permits AID activity to primarily occur at sites of antigen binding.

The removal of the amine group results in cytosine by AID results in the formation of an uracil[65,91,95]. This creates a base pair mismatch as the cytosine:guanine paired nucleotides become uracil:guanine pairs. As polymerases are insensitive to uracil, which is registered at thymine, U:G mismatches that are unprocessed during replication induces a transversion mutation resulting in a C:G base pair becoming T:A[64,95]. If the U:G mismatches is detected by the heterodimer complex MSH2/6 or the uracil DNA-glycosylase UNG, either faithful regeneration of the deaminated cytosine can occur or a variety of mutagenic outcomes.

First, we will consider the outcomes related to UNG. Activity of the UNG enzyme at the U:G mismatch results in excision of the uracil base leaving an abasic site opposite the guanine[65]. Under normal conditions, UNG carries out a form of DNA repair known as base-excision repair (BER), where it recruits the enzyme APE1 to induce a single strand DNA nick, followed by recruitment of the high-fidelity DNA polymerase polβ to replicate across the abasic site[65]. However, in the germinal center replication over this abasic site recruits specialized polymerases

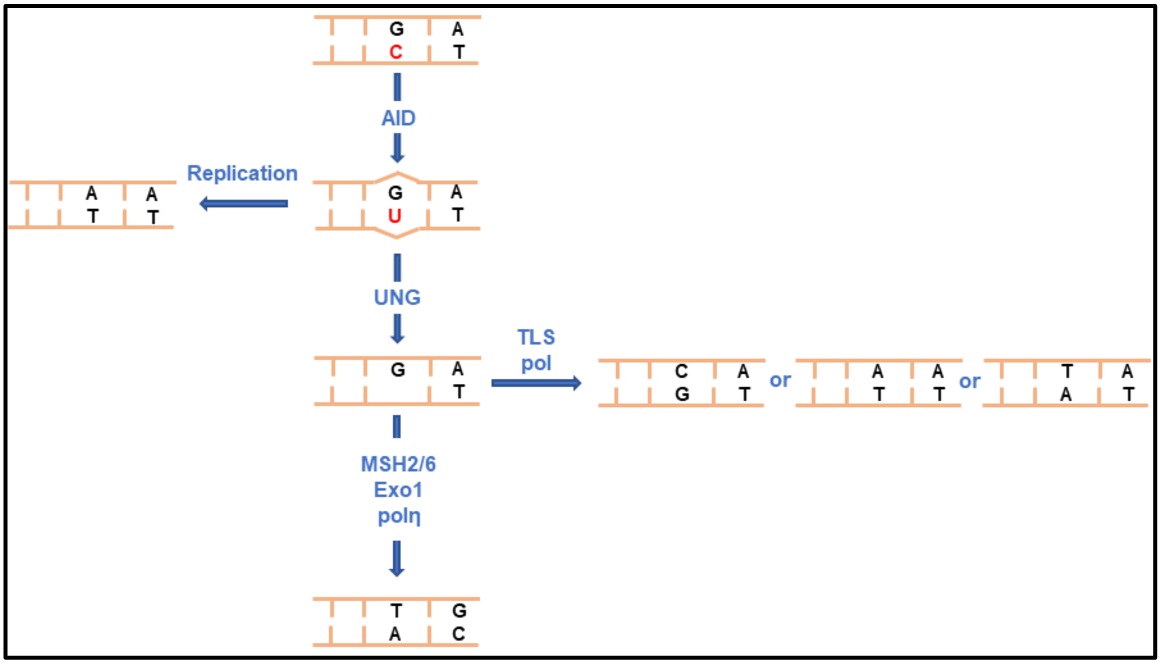


Figure 1:Neuberger Model of SHM

such as REV1 and polη which results in the incorporation of a random nucleotide, causing both transition and transversion mutations at the site of the deamination event[65].

If the U:G mismatch is recognized by MSH2/6 complex, more complex mutational events occur. Recall that AID only acts upon cytosine and thus can only cause mutations at C:G base pairs. However, somatic hypermutation causes mutations at both A:T and G:C pairs in a 2:3 ratio. The mutagenesis at A:T base pairs is mediated by "error-prone" patch synthesis that occurs following recognition of the U:G mismatch by MSH2/6[65,96,97]. This occurs via recruitment of UNG, APE2, and the exonuclease EXO1[65]. As described above, UNG creates an abasic site, which is then acted upon by APE2 creating a single strand nick[65]. The nick is a substrate for EXO1, which then removes nucleotides in a 5'→3' direction creating a region of single stranded DNA. This lesion is repaired by the "error-prone" activities of translesion polymerases polη, polζ, and polι[64,96-98]. These polymerases act over both A:T and G:C base pairs and can introduce transition and transversion mutations along stretch of ssDNA created by EXO1.

As mentioned earlier, AID and associated SHM machinery are responsible for class switch recombination (CSR)[62]. Although the machinery involved in CSR is largely the same as that of SHM, there are important differences that result in DNA rearrangement as opposed to simple mutagenesis. Foremost is the enrichment of a class of AID hotspots that have the motif 5'-WGCW-3'[65,94,99]. These hotspots are unique in that they possess overlapping AID specificities on both strands of DNA. Deamination of both of these overlapping sites results in activity by UNG and APE1 on both strands resulting in formation of a double strand break (DSB)[65]. Generation of DSB at the site upstream of the expressed constant region and one downstream facilitates a form of DNA repair known as canonical non-homologous end joining (cNHEJ)[65]. This allows the downstream constant region to replace the expressed one via deletion of all intervening sequences. It is also possible for another pathway alternative non-homologous end joining (aNHEJ) to facilitate CSR. In this mechanism, DSB are generated that are staggered due to the formation of a DSB from AID, UNG, and APE1 activity at one cytosine followed by AID, MSH2/6, UNG, and EXO1 activity on the opposite strand which results in a staggered DSB with a free 5' overhang. In aNHEJ, microhomology between staggered ends of the expressed constant region and the downstream constant region facilitates recombination through DNA repair factors PARP1 and XRCC1[65].

In addition to SHM and CSR, AID activity can also result in the process of gene conversion[88]. Originally identified as a mechanism of somatic diversification in 1985, gene conversion is the process by which DNA from a source or donor, is copied into a highly homologous recipient sequence such that the recipient sequence is indistinguishable from the donated sequence[100-102]. Gene conversion is thought to occur as a means of primary diversification in birds as they only possess a single V region capable of undergoing VDJ recombination. Thus, B cells in these animals rely on a primary round of gene conversion to form their naïve repertoire[101]. This normally occurs in the gut associated lymphoid tissue of species that utilize gene conversion. In the periphery, these gene conversion experienced B cells may undergo subsequent rounds of gene conversion and somatic hypermutation to generate high affinity antibodies.

Unlike SHM, key mechanisms that drive gene conversion in B cells are unknown, although there are some common dependencies. Foremost, AID is essential for the occurrence of gene conversion[88]. Additionally, the transcription factors BCL6 and Bach2 are required[103,104]. It has also been demonstrated that gene conversion relies on processing of the AID-induced uracil by UNG into an abasic site[105,106]. Recent studies have determined that DSBs in the mutating V region are sufficient to induce gene conversion, even in the absence of AID[107]. Together this suggests that the mechanism of initiating the gene conversion event is DSB-dependent. Following the break, data from multiple groups demonstrate that donor choice for repairing the DSB is dependent on donor proximity, the chromatin state of the donor sequence, transcription and homology between the donor and recipient[108-110]. While the exact mechanism of templated repair is not known, it is believed to rely on RAD51 mediated homologous recombination[111,112]. This is supported by the observation that knockout of the RAD51-loading genes XRCC2 and XRCC3 results in marked decrease in gene conversion events[113]. Interestingly concomitant to the ablation of gene conversion in XRCC2/3 knockout chicken germinal center B cells was the selective loss of A:T mutations as well as focusing of observed somatic mutations to AID hotspots, suggesting a role of gene conversion in generating A:T mutations in these B cells[113].

Mechanisms of Gene Conversion outside of B cells

Gene conversion is a ubiquitous process related to the DNA repair process of homologous recombination (HR). In B cells that undergo gene conversion, the primary initiating event for gene conversion is the presence of a DSB[111]. As alluded to through the mechanism of CSR, a DSB may be repaired using the process of NHEJ, which promotes rapid resolution of the DSB at the cost of loss of nucleotides, or in certain instances forming a carcinogenesis-promoting translocation. Alternatively, the DSB may be repaired via homologous recombination, which depending on the homologous donor used as a template for repair, may or may not induce gene conversion. In the simplest case, a completely homologous sequence is used as a template and the DSB is repaired without introduction of mutations[114-116]. If a homologous, but different, sequence (homeologous sequence) is used to repair the DSB, those differences will be reflected as mutations as the lesion resolves.

Mechanistic studies regarding gene conversion have primarily occurred in yeast, where endonucleases can create precise cuts in DNA. They have demonstrated that HR and gene conversion can occur through three disparate mechanisms: (1) break-induced repair (BIR), (2) strand dependent strand annealing (SDSA), and (3) the double strand break repair pathway (DSBR)[116]. From the DSB, each of these mechanisms occur in three steps: presynapsis, synapsis, and postsynapsis. HR and gene conversion pathway choice occurs at the postsynapsis stage, with the two prior stages being required for each HR pathway to occur[116].

At the presynaptic stage the DSB is resected to yield two free 3’ overhangs. The MRN complex, comprised of Mre11, Rad50, and Xrs2, is recruited to the site of the DSB and conducts initial trimming of either end of the DSB[116]. These ends are further resected 5’ to 3’ by two independent pathways. The first, requires the helicase DNA2 as well as the STR complex, comprised of Sgs1, Top3, and Rmi1[116]. The second pathway relies on activity of EXO1[116]. Both yield ssDNA that is complexed to the single strand binding protein RPA. Next, RPA is replaced by the homology search protein RAD51, which forms a filament along the length of the exposed ssDNA[116]. This is facilitated by proteins BRCA1, RAD52, as well as RAD51 paralogs, including XRCC2 and XRCC3[116]. The complexed RAD51 and exposed 3' ssDNA is the active structure capable of searching the genome for homologous sequences. While the potential number of sequences needed to search is as large as the genome, the RAD51 complex can rapidly survey vast genomic stretches through a combination of 3D sampling (binding and dissociating in three dimension space) and 1D sliding along bound DNA[117,118]. In the RAD51 complex. RAD51 is bound to the phosphate backbone of the ssDNA, leaving the bases exposed for determining homology matches[119].

The synapsis stage begins once a homology match is found. The RAD51 complex promotes strand exchange between the complementary homologous strand of dsDNA and the RAD51-bound 3' ssDNA overhang[116]. This creates a structure known as a displacement-loop (D-loop). Given that the 3' ssDNA is now bound to a homologous sequence, this creates a primer from which the homologous sequence can be copied, extended from the free 3' end. This is done by first freeing the 3'-OH group from the RAD51 complex via the dissociative activity of RAD54[116]. This allows PCNA to be recruited and subsequently recruit either the high-fidelity replicative polymerases polδ and polε or the "error-prone" translesion polymerases such as polη and polκ to extend the broken DSB end[116].

The post-synapsis stage occurs after the D-loop has been extended by polymerases and consists of resolution of the D-loop structure and repair of the DSB. In instances where the DSB is single-ended, as in replication fork collapse, repair occurs via BIR, during which the D-loop continues to undergo polymerization from the template strand[116]. As the DNA is synthesized from the template strand the D-loop, the D-loop migrates along the template strand and concomitantly the resected end of the DSB is synthesized from the newly exposed and synthesized DNA in the D-loop. This results in the phenomenon of half crossover, in which the DNA synthesized past the DSB is identical to that of the donor template in its entirety from the site of RAD51 complex D-loop formation until the end of the DNA sequence.

Alternatively, in SDSA, the D-loop can be displaced by helicases resolving the D-loop structure[116]. If the newly synthesized DNA from the D-loop base pairs with the other 3' overhang in the DSB, the strands will anneal, leaving two ssDNA gaps. These can be repaired using mechanisms of patch repair as described earlier. This leaves a small region of gene conversion at the site of the DSB flanked by newly synthesized DNA derived from the resected ends during the pre-synapsis stage[116]. It is possible that the DNA synthesized from the D-loop is not complementary to base pairs from the 3' overhang, and when this occurs, the RAD51-complex can again survey for homologous sequences, resulting in the formation a second D-loop. This can result in the unusual phenomenon of template switching, which can occur if the second D-loop is formed at a different homologous sequence[120,121]. In this case, resolution of the D-loop and the DSB will leave two overlapping regions of gene conversion flanked by DNA synthesized from overhangs in pre-synapsis. Intriguingly, this phenomenon was observed in an AID null DT40 model – a chicken B cell line that undergoes spontaneous gene conversion at the IgH locus. In these studies, a single inducible DSB resulted in both gene conversion as well as gene conversion with template switch suggesting that SDSA is the primary mechanism of gene conversion in birds[107].

The third mechanism of HR-mediated DSB repair, DSBR, occurs when the displaced strand of the D-loop anneals to the other end of the DSB creating a physical link between the two strands of DNA known as a double holiday junction (dHJ)[116]. Formation of the dHJ allows simultaneous repair of the second end of the DSB as the newly annealed strand can serve as a template to begin polymerization. Resolution of the dHJ occurs via a variety of nucleases, which physically cut the invading DNA – both the DNA that formed the original D-loop, as well as the displaced DNA that now is bound to the second end of the DSB. In cases where a single junction is cut, the resultant structure is a crossover, where the DNA downstream of the DSB is swapped with that of the DNA downstream of the D-loop formation[116]. More commonly, the dHJ is resolved by nuclease activity at both junctions, resulting in a patch of gene conversion at both the DSB and at the donor site[116].

Gene conversion and somatic hypermutation during evolution

The use of gene conversion and cytidine deaminases by the adaptive immune system to generate antigen-specific molecules extends to common ancestors that existed ~525 million years ago[79]. Jawless vertebrates, such as lampreys, generate antibody-like molecules known as variable lymphocyte receptors (VLRs) that are composed of multiple leucine rich repeat (LRR) genes[122]. These VLRs must be somatically assembled through a gene conversion-like process to generate functional VLRs[122]. Like that of gene conversion in birds, multiple pseudogene elements are used as donor sequences to replace sequences in the VLR gene, building a somatically diversified gene. Unlike birds, however, VLR assembly in lampreys is necessary for function, whereas VDJ recombination in birds is sufficient to produce a functional antibody. Additionally, lampreys express two cytidine deaminases that have been implicated in VLR assembly, and for which evidence regarding their mutagenic potential is emerging[123].

Gnathostomes, or jawed vertebrates, diverged from jawless fish 525 million years ago and from which fish, amphibians, birds and mammals are derived[79]. The emergence of AID and antibodies is first seen in bony fishes, which can produce antibody and undergo affinity maturation in those antibodies in an antigen specific manner[79,124]. Similarly, all animals who share the common ancestor to bony fishes retain the capacity to use AID and to produce antigen specific antibody[79,125]. Other features of humoral immunity, such as CSR and germinal centers originate later in the evolutionary phylogeny[79,125]. Curiously, while the use of AID and subsequently, development of affinity maturation, CSR, and germinal centers fall sequentially along the evolutionary phylogeny, the use of gene conversion as a mechanism of diversification is sporadic[79,125]. Besides the use of gene conversion in birds, rabbits and artiodactyls (hooved mammals), are known to utilize gene conversion as a means of generating the primary repertoire[125-130]. As in birds, these animals have a limited number of functional V genes to be used during VDJ recombination, generating a limited number of specificities. Additional diversity in these animals is generated in gut associated lymphoid tissue (GALT) where activity by AID in B cells drives gene conversion processes[125]. This activity further diversifies the B cells and completes maturation of the primary repertoire.

Unlike AID activity, affinity maturation, CSR, and germinal center formation, which are all readily and directly observable events, detection of gene conversion relies on inference. Identification of a gene conversion event relies on aligning the germline sequence, a somatically mutated sequence, and a second germline sequence. Should mutations in the somatically mutated sequence match the differences between the germlines in a narrow region, it is assumed to be the product of gene conversion between the two germline sequences. This prompts the question of whether many of the species that are not believed to utilize gene conversion for somatic diversification do so.

Evolution of somatic hypermutation and gene conversion paradigms

To date, nine of 109 Nobel Prizes in Medicine have been granted to persons whose work - whether knowingly or not - pertained to antibodies. Indeed, antibodies have had significant life in biological research. The first Nobel Prize in Medicine was given to Emil von Behring for his work on production of diphtheria antitoxins that were generated by injection of diphtheria toxin into animals. After some time, the animal's serum would be extracted and could be used to treat patients with diphtheria. The action of the serum in patients was due to antigen-specific antibodies that were produced in the animals following inoculation.

The term, antibody, itself was coined by Paul Ehrlich in 1891[131,132]. Later Ehrlich too would go on to be awarded the Nobel for his work on antibodies. In his work, Ehrlich demonstrated that mice fed with small and increasing doses of the toxin ricin would eventually become resistant to ricin[132]. The same phenomenon would occur if the mice were fed a different toxin, abrin[132]. However, ricin-resistant mice would not be protected from abrin and vice-versa[132]. Instead, the resistant mice would be as susceptible as mice who had not been exposed to any toxin. Ehrlich concluded that antibodies were specific to their inducing antigen.

For some time, the basis of this phenomenon went unknown. It wouldn't be until the seminal discoveries of James Watson and Francis Crick regarding DNA and the advent of molecular biology until the answer would become clearer. In the 1960s, there was a fundamental paradox concerning how antibodies could be generated to a plethora of antigens. How was it possible that so many antibody specifies could be generated? By the turn of the next decade, two opposing theories emerged: somatic and germline theory[133]. Somatic theory posited that high rates of somatic mutation in a few germline genes could generate the necessary diversity needed. However opponents to this theory suggested that the number of mutants needed would be exceedingly high. Those opponents held the opposing view that antibody specificities were encoded by many different genes. The flaw in this theory, argued those who subscribed to somatic theory, was that there was not enough DNA to account for the diverse set of antibody specificities that could be generated.

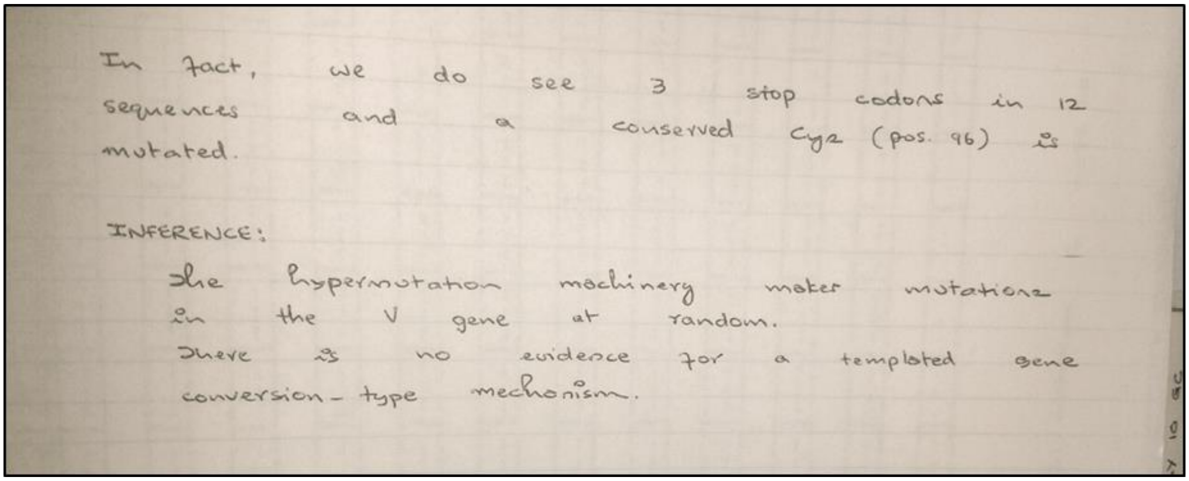
As deftly acknowledged in Patricia Gearhart's review, during this time of intense debate, Francis Crick, Sydney Brenner, and Cesar Milstein worked together in the Laboratory of Molecular Biology in Cambridge[133]. According to Dr. Gearhart, Crick "suggested that Brenner and Milstein collaborate and write a letter to Nature about their ideas on how diversity was generated[133]." That paper, titled "Origin of Antibody Variation," was published in 1966 and posited that a variable segment would be recognized at a specific motif by a restriction enzyme[134]. This would lead a cut in the DNA that upon processing would lead to degradation 3' to 5' and could then be polymerized by an error prone polymerase, generating somatic mutants[134].

Ten years later, in 1976, Hozumi and Tonegawa publish their findings that suggested that combinatorial diversity via somatic rearrangement of multiple gene segments could result in multiple antibody specificities greater than the sum of the gene segments themselves[135]. This finding would later earn Dr. Tonegawa the Nobel prize as well. This finding seemingly settled the debate on how antibodies could generate such diverse responses. However, using technology developed a year earlier in 1975 by Kohler and Milstein (which also earned them a Nobel), four papers emerged in 1981 emerged that demonstrated that somatic mutations accumulated in antigen-specific responses[136-140]. This prompted the field to re-examine Brenner and Milstein's paper and the concept of somatic hypermutation. Later work by Milstein demonstrated that random mutations in CDRs would lead to progressively higher affinity antibodies[141]. But what was the mechanism by which these mutations were introduced?

In 1985, Claude-Agnès Reynaud and Jean-Claude Weill first described that gene conversion between pseudogenes at the chicken light chain locus could produce somatic variants that were diverse but also unselected, seemingly solving the paradox of generating many nonfunctional somatic variants in order to obtain a functional one[100]. The role of gene conversion in the diversification of chicken B cells was confirmed later in 1989 by the same group[102].

At roughly the same time, in 1986, Thereza Imanishi-Kari coauthored a paper with David Baltimore (who was a Nobel recipient for his work on reverse transcriptase)[142]. That paper "Altered repertoire of endogenous immunoglobulin gene expression in transgenic mice containing a rearranged Mu heavy chain gene" ultimately got caught in Former US Representative John Dingell's inquest into – and ultimate failure to prove – scientific misconduct. This resulted in the retraction of the paper, whose chief finding was that delivery of a nonfunctional transgene which encoded a specific idiotype of the BALB/c mouse strain into a mouse of the BL6 strain (that does not produce the idiotype) resulted in steady state appearance of the BALB/c idiotype in the BL6 antibody repertoire[142]. It is worth noting that during the fallout resulting from the "Baltimore affair," Baltimore sent a letter to his collaborators assuring them of the studies findings and the quality of the science[143]. In his letter, Dr. Baltimore suggests that the result of the study may be attributable to "the explanation of trans-switching or gene conversion [which] seemed most likely.[143]" Despite this the careers of both Baltimore and Imanishi-Kari suffered, as well as the vein of research, though no misconduct was found.

In the literature at this time, many studies were conducted demonstrating that gene conversion could not be found in mouse or human B cells, including the seminal paper published during the graduate career of my supervisor[144]. This led to two camps, those who believed that gene conversion was the likely explanation for somatic mutations, and those that subscribed to a



In fact, we do see 3 stop codons in 12 sequences and a conserved Cys (pos. 96) is mutated.

INFERENCE:

The hypermutation machinery makes mutations in the V gene at random.
There is no evidence for a templated gene conversion-type mechanism.

Figure 2: Primary evidence of the state of the field circa 1991. See Jacob et al. (1991)

random mutagenesis model[145]. This would ultimately begin to be resolved in 2000, when Muramatsu and Honjo (who recently won the Nobel in 2018), discovered the cytidine deaminase AID[62]. With this discovery a physical basis for untemplated mutation was discovered. While it was later shown that AID also chiefly regulated gene conversion in chicken B cells, the field settled on untemplated mutagenesis in murine and human B cells in the absence of direct evidence[65,88].

Curiously, in 1994 and 2002, Erik Selsing's group published a pair of papers which placed a nonfunctional hybridoma sequence upstream of a functional hybridoma sequence, both specific to the phenylarsonate (ARS) hapten, at the IgH locus of mice[146,147]. These two sequences differ by 17bp and these papers demonstrated that immunization with the ARS hapten resulted in B cells that expressed sequences in which gene conversion appeared to have occurred between the two sequences. Separately, a group led by Edward Steele, proposed the mechanism of somatic hypermutation to be related to reverse transcriptase activity[148]. They later demonstrated the activity of polη to include reverse transcriptase activity in addition to its role as a translesion synthesis polymerase[149]. However, these papers did not gain traction given the large body of native murine data suggesting gene conversion and templated mutagenesis was not occurring in B cells.

Further work throughout the next two decades saw the characterization of the steps of untemplated somatic hypermutation in mice and humans. It was shown that UNG and MSH2/6 played critical roles in B cell diversification and that polη was the key driver of mutations at A:T base pairs[65,97]. In short, the verdict of whether templated or untemplated mutations occurred in murine and human B cells appeared settled firmly as untemplated.

It is here where this work takes place. The following chapters will address the role of templated mutagenesis in human and murine B cells, demonstrating (1) that somatic mutations exhibit linkage disequilibrium especially when in proximity, (2) that pairs mutations are statistically likely to derive from gene conversion using germline IgHV sequences, even when the hypermutating sequence is not an antibody gene, (3) that mutations can be templated from distant sites of the genome, (4) sites for diversification using gene conversion are centered around the DSB-prone overlapping AID hotspots, and (5) that donor template choice is controlled by epigenetic status, physical distance to the DSB, and homology. Altogether, these studies present compelling evidence that templated mutagenesis is a component of secondary diversification in mice and humans.

Works Cited

1 Elhanati, Y. *et al.* Inferring processes underlying B-cell repertoire diversity. *Philos Trans R Soc Lond B Biol Sci* **370**, doi:10.1098/rstb.2014.0243 (2015).

2 Jung, D., Giallourakis, C., Mostoslavsky, R. & Alt, F. W. Mechanism and control of V(D)J recombination at the immunoglobulin heavy chain locus. *Annu Rev Immunol* **24**, 541-570, doi:10.1146/annurev.immunol.23.021704.115830 (2006).

3 Schroeder, H. W., Jr. & Cavacini, L. Structure and function of immunoglobulins. *J Allergy Clin Immunol* **125**, S41-52, doi:10.1016/j.jaci.2009.09.046 (2010).

4 Edelman, G. M. Antibody structure and molecular immunology. *Scand J Immunol* **34**, 1-22 (1991).

5 Harris, L. J. *et al.* The three-dimensional structure of an intact monoclonal antibody for canine lymphoma. *Nature* **360**, 369-372, doi:10.1038/360369a0 (1992).

6 Decanniere, K., Muyldermans, S. & Wyns, L. Canonical antigen-binding loop structures in immunoglobulins: more structures, more canonical classes? *J Mol Biol* **300**, 83-91, doi:10.1006/jmbi.2000.3839 (2000).

7 Gilliland, L. K. *et al.* Rapid and reliable cloning of antibody variable regions and generation of recombinant single chain antibody fragments. *Tissue Antigens* **47**, 1-20 (1996).

8 Early, P., Huang, H., Davis, M., Calame, K. & Hood, L. An immunoglobulin heavy chain variable region gene is generated from three segments of DNA: VH, D and JH. *Cell* **19**, 981-992, doi:10.1016/0092-8674(80)90089-6 (1980).

9 Matsuda, F. & Honjo, T. Organization of the human immunoglobulin heavy-chain locus. *Adv Immunol* **62**, 1-29 (1996).

10 McBlane, J. F. *et al.* Cleavage at a V(D)J recombination signal requires only RAG1 and RAG2 proteins and occurs in two steps. *Cell* **83**, 387-395, doi:10.1016/0092-8674(95)90116-7 (1995).

11 Mombaerts, P. *et al.* RAG-1-deficient mice have no mature B and T lymphocytes. *Cell* **68**, 869-877, doi:10.1016/0092-8674(92)90030-g (1992).

12 Shinkai, Y. *et al.* RAG-2-deficient mice lack mature lymphocytes owing to inability to initiate V(D)J rearrangement. *Cell* **68**, 855-867, doi:10.1016/0092-8674(92)90029-c (1992).

13 van Gent, D. C. *et al.* Initiation of V(D)J recombination in a cell-free system. *Cell* **81**, 925-934, doi:10.1016/0092-8674(95)90012-8 (1995).

14 Fugmann, S. D., Lee, A. I., Shockett, P. E., Villey, I. J. & Schatz, D. G. The RAG proteins and V(D)J recombination: complexes, ends, and transposition. *Annu Rev Immunol* **18**, 495-527, doi:10.1146/annurev.immunol.18.1.495 (2000).

15 Eastman, Q. M., Leu, T. M. & Schatz, D. G. Initiation of V(D)J recombination in vitro obeying the 12/23 rule. *Nature* **380**, 85-88, doi:10.1038/380085a0 (1996).

16 Sawchuk, D. J. *et al.* V(D)J recombination: modulation of RAG1 and RAG2 cleavage activity on 12/23 substrates by whole cell extract and DNA-bending proteins. *J Exp Med* **185**, 2025-2032, doi:10.1084/jem.185.11.2025 (1997).

17 Sakano, H., Huppi, K., Heinrich, G. & Tonegawa, S. Sequences at the somatic recombination sites of immunoglobulin light-chain genes. *Nature* **280**, 288-294, doi:10.1038/280288a0 (1979).

18 Besmer, E. *et al.* Hairpin coding end opening is mediated by RAG1 and RAG2 proteins. *Mol Cell* **2**, 817-828 (1998).

19 Shockett, P. E. & Schatz, D. G. DNA hairpin opening mediated by the RAG1 and RAG2 proteins. *Mol Cell Biol* **19**, 4159-4166, doi:10.1128/mcb.19.6.4159 (1999).

20 Ma, Y., Pannicke, U., Schwarz, K. & Lieber, M. R. Hairpin opening and overhang processing by an Artemis/DNA-dependent protein kinase complex in nonhomologous end joining and V(D)J recombination. *Cell* **108**, 781-794, doi:10.1016/s0092-8674(02)00671-2 (2002).

21 Gilfillan, S., Dierich, A., Lemeur, M., Benoist, C. & Mathis, D. Mice lacking TdT: mature animals with an immature lymphocyte repertoire. *Science* **261**, 1175-1178, doi:10.1126/science.8356452 (1993).

22 Komori, T., Okada, A., Stewart, V. & Alt, F. W. Lack of N regions in antigen receptor variable region genes of TdT-deficient lymphocytes. *Science* **261**, 1171-1175, doi:10.1126/science.8356451 (1993).

23 Kirkham, P. M. & Schroeder, H. W., Jr. Antibody structure and the evolution of immunoglobulin V gene segments. *Semin Immunol* **6**, 347-360, doi:10.1006/smim.1994.1045 (1994).

24 Alt, F. W., Enea, V., Bothwell, A. L. & Baltimore, D. Activity of multiple light chain genes in murine myeloma cells producing a single, functional light chain. *Cell* **21**, 1-12, doi:10.1016/0092-8674(80)90109-9 (1980).

25 Melchers, F. *et al.* Repertoire selection by pre-B-cell receptors and B-cell receptors, and genetic control of B-cell development from immature to mature B cells. *Immunol Rev* **175**, 33-46 (2000).

26 Busslinger, M. Transcriptional control of early B cell development. *Annu Rev Immunol* **22**, 55-79, doi:10.1146/annurev.immunol.22.012703.104807 (2004).

27 McKenna, H. J. *et al.* Mice lacking flt3 ligand have deficient hematopoiesis affecting hematopoietic progenitor cells, dendritic cells, and natural killer cells. *Blood* **95**, 3489-3497 (2000).

28 Sitnicka, E. *et al.* Key role of flt3 ligand in regulation of the common lymphoid progenitor but not in maintenance of the hematopoietic stem cell pool. *Immunity* **17**, 463-472 (2002).

29 Carvalho, T. L., Mota-Santos, T., Cumano, A., Demengeot, J. & Vieira, P. Arrested B lymphopoiesis and persistence of activated B cells in adult interleukin 7(-/)- mice. *J Exp Med* **194**, 1141-1150, doi:10.1084/jem.194.8.1141 (2001).

30 Miller, J. P. *et al.* The earliest step in B lineage differentiation from common lymphoid progenitors is critically dependent upon interleukin 7. *J Exp Med* **196**, 705-711, doi:10.1084/jem.20020784 (2002).

31 Hardy, R. R., Carmack, C. E., Shinton, S. A., Kemp, J. D. & Hayakawa, K. Resolution and characterization of pro-B and pre-pro-B cell stages in normal mouse bone marrow. *J Exp Med* **173**, 1213-1225, doi:10.1084/jem.173.5.1213 (1991).

32 Li, Y. S., Hayakawa, K. & Hardy, R. R. The regulated expression of B lineage associated genes during B cell differentiation in bone marrow and fetal liver. *J Exp Med* **178**, 951-960, doi:10.1084/jem.178.3.951 (1993).

33 Li, Y. S., Wasserman, R., Hayakawa, K. & Hardy, R. R. Identification of the earliest B lineage stage in mouse bone marrow. *Immunity* **5**, 527-535 (1996).

34 Meffre, E., Casellas, R. & Nussenzweig, M. C. Antibody regulation of B cell development. *Nat Immunol* **1**, 379-385, doi:10.1038/80816 (2000).

35 Melchers, F. Fit for life in the immune system? Surrogate L chain tests H chains that test L chains. *Proc Natl Acad Sci U S A* **96**, 2571-2573, doi:10.1073/pnas.96.6.2571 (1999).

36 Horne, M. C., Roth, P. E. & DeFranco, A. L. Assembly of the truncated immunoglobulin heavy chain D mu into antigen receptor-like complexes in pre-B cells but not in B cells. *Immunity* **4**, 145-158 (1996).

37 Tornberg, U. C., Bergqvist, I., Haury, M. & Holmberg, D. Regulation of B lymphocyte development by the truncated immunoglobulin heavy chain protein Dmu. *J Exp Med* **187**, 703-709, doi:10.1084/jem.187.5.703 (1998).

38 Maki, K., Nagata, K., Kitamura, F., Takemori, T. & Karasuyama, H. Immunoglobulin beta signaling regulates locus accessibility for ordered immunoglobulin gene rearrangements. *J Exp Med* **191**, 1333-1340, doi:10.1084/jem.191.8.1333 (2000).

39 Reth, M., Petrac, E., Wiese, P., Lobel, L. & Alt, F. W. Activation of V kappa gene rearrangement in pre-B cells follows the expression of membrane-bound immunoglobulin heavy chains. *EMBO J* **6**, 3299-3305 (1987).

40 Schlissel, M. S. & Baltimore, D. Activation of immunoglobulin kappa gene rearrangement correlates with induction of germline kappa gene transcription. *Cell* **58**, 1001-1007, doi:10.1016/0092-8674(89)90951-3 (1989).

41 Phan, T. G., Gray, E. E. & Cyster, J. G. The microanatomy of B cell activation. *Curr Opin Immunol* **21**, 258-265, doi:10.1016/j.coi.2009.05.006 (2009).

42 Beyer, T. & Meyer-Hermann, M. Mechanisms of organogenesis of primary lymphoid follicles. *Int Immunol* **20**, 615-623, doi:10.1093/intimm/dxn020 (2008).

43 Cyster, J. G. Chemokines, sphingosine-1-phosphate, and cell migration in secondary lymphoid organs. *Annu Rev Immunol* **23**, 127-159, doi:10.1146/annurev.immunol.23.021704.115628 (2005).

44 Harwood, N. E. & Batista, F. D. Early events in B cell activation. *Annu Rev Immunol* **28**, 185-210, doi:10.1146/annurev-immunol-030409-101216 (2010).

45 Lanzavecchia, A. Antigen-specific interaction between T and B cells. *Nature* **314**, 537-539, doi:10.1038/314537a0 (1985).

46 Rock, K. L., Benacerraf, B. & Abbas, A. K. Antigen presentation by hapten-specific B lymphocytes. I. Role of surface immunoglobulin receptors. *J Exp Med* **160**, 1102-1113, doi:10.1084/jem.160.4.1102 (1984).

47 Okada, T. *et al.* Antigen-engaged B cells undergo chemotaxis toward the T zone and form motile conjugates with helper T cells. *PLoS Biol* **3**, e150, doi:10.1371/journal.pbio.0030150 (2005).

48 Ni, K. & O'Neill, H. C. The role of dendritic cells in T cell activation. *Immunol Cell Biol* **75**, 223-230, doi:10.1038/icb.1997.35 (1997).

49 Parker, D. C. T cell-dependent B cell activation. *Annu Rev Immunol* **11**, 331-360, doi:10.1146/annurev.iy.11.040193.001555 (1993).

50 Schaerli, P. *et al.* CXC chemokine receptor 5 expression defines follicular homing T cells with B cell helper function. *J Exp Med* **192**, 1553-1562, doi:10.1084/jem.192.11.1553 (2000).

51 Basso, K. & Dalla-Favera, R. Roles of BCL6 in normal and transformed germinal center B cells. *Immunol Rev* **247**, 172-183, doi:10.1111/j.1600-065X.2012.01112.x (2012).

52 Toyama, H. *et al.* Memory B cells without somatic hypermutation are generated from Bcl6-deficient B cells. *Immunity* **17**, 329-339 (2002).

53 Victora, G. D. & Nussenzweig, M. C. Germinal centers. *Annu Rev Immunol* **30**, 429-457, doi:10.1146/annurev-immunol-020711-075032 (2012).

54 Mesin, L., Ersching, J. & Victora, G. D. Germinal Center B Cell Dynamics. *Immunity* **45**, 471-482, doi:10.1016/j.immuni.2016.09.001 (2016).

55 Gitlin, A. D. *et al.* HUMORAL IMMUNITY. T cell help controls the speed of the cell cycle in germinal center B cells. *Science* **349**, 643-646, doi:10.1126/science.aac4919 (2015).

56 Gitlin, A. D., Shulman, Z. & Nussenzweig, M. C. Clonal selection in the germinal centre by regulated proliferation and hypermutation. *Nature* **509**, 637-640, doi:10.1038/nature13300 (2014).

57 Ersching, J. *et al.* Germinal Center Selection and Affinity Maturation Require Dynamic Regulation of mTORC1 Kinase. *Immunity* **46**, 1045-1058 e1046, doi:10.1016/j.immuni.2017.06.005 (2017).

58 Han, S. *et al.* Cellular interaction in germinal centers. Roles of CD40 ligand and B7-2 in established germinal centers. *J Immunol* **155**, 556-567 (1995).

59 Linterman, M. A. *et al.* IL-21 acts directly on B cells to regulate Bcl-6 expression and germinal center responses. *J Exp Med* **207**, 353-363, doi:10.1084/jem.20091738 (2010).

60 Zotos, D. *et al.* IL-21 regulates germinal center B cell differentiation and proliferation through a B cell-intrinsic mechanism. *J Exp Med* **207**, 365-378, doi:10.1084/jem.20091777 (2010).

61 Allen, C. D. *et al.* Germinal center dark and light zone organization is mediated by CXCR4 and CXCR5. *Nat Immunol* **5**, 943-952, doi:10.1038/ni1100 (2004).

62 Muramatsu, M. *et al.* Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* **102**, 553-563 (2000).

63 Chahwan, R., Edelmann, W., Scharff, M. D. & Roa, S. AIDing antibody diversity by error-prone mismatch repair. *Semin Immunol* **24**, 293-300, doi:10.1016/j.smim.2012.05.005 (2012).

64 Saribasak, H. & Gearhart, P. J. Does DNA repair occur during somatic hypermutation? *Semin Immunol* **24**, 287-292, doi:10.1016/j.smim.2012.05.002 (2012).

65 Maul, R. W. & Gearhart, P. J. Refining the Neuberger model: Uracil processing by activated B cells. *Eur J Immunol* **44**, 1913-1916, doi:10.1002/eji.201444813 (2014).

66 Xu, Z., Zan, H., Pone, E. J., Mai, T. & Casali, P. Immunoglobulin class-switch DNA recombination: induction, targeting and beyond. *Nat Rev Immunol* **12**, 517-531, doi:10.1038/nri3216 (2012).

67 Pone, E. J. *et al.* Toll-like receptors and B-cell receptors synergize to induce immunoglobulin class-switch DNA recombination: relevance to microbial antibody responses. *Crit Rev Immunol* **30**, 1-29 (2010).

68 Tas, J. M. *et al.* Visualizing antibody affinity maturation in germinal centers. *Science* **351**, 1048-1054, doi:10.1126/science.aad3439 (2016).

69 Huang, C., Geng, H., Boss, I., Wang, L. & Melnick, A. Cooperative transcriptional repression by BCL6 and BACH2 in germinal center B-cell differentiation. *Blood* **123**, 1012-1020, doi:10.1182/blood-2013-07-518605 (2014).

70 Ochiai, K. *et al.* Plasmacytic transcription factor Blimp-1 is repressed by Bach2 in B cells. *J Biol Chem* **281**, 38226-38234, doi:10.1074/jbc.M607592200 (2006).

71 Ochiai, K., Muto, A., Tanaka, H., Takahashi, S. & Igarashi, K. Regulation of the plasma cell transcription factor Blimp-1 gene by Bach2 and Bcl6. *Int Immunol* **20**, 453-460, doi:10.1093/intimm/dxn005 (2008).

72 Shaffer, A. L. *et al.* Blimp-1 orchestrates plasma cell differentiation by extinguishing the mature B cell gene expression program. *Immunity* **17**, 51-62 (2002).

73 Radtke, D. & Bannard, O. Expression of the Plasma Cell Transcriptional Regulator Blimp-1 by Dark Zone Germinal Center B Cells During Periods of Proliferation. *Front Immunol* **9**, 3106, doi:10.3389/fimmu.2018.03106 (2018).

74 Ise, W. & Kurosaki, T. Plasma cell differentiation during the germinal center reaction. *Immunol Rev* **288**, 64-74, doi:10.1111/imr.12751 (2019).

75 Roth, K. *et al.* Tracking plasma cell differentiation and survival. *Cytometry A* **85**, 15-24, doi:10.1002/cyto.a.22355 (2014).

76 Slifka, M. K., Antia, R., Whitmire, J. K. & Ahmed, R. Humoral immunity due to long-lived plasma cells. *Immunity* **8**, 363-372 (1998).

77 Kurosaki, T., Kometani, K. & Ise, W. Memory B cells. *Nat Rev Immunol* **15**, 149-159, doi:10.1038/nri3802 (2015).

78 Tsui, C. *et al.* Protein Kinase C-beta Dictates B Cell Fate by Regulating Mitochondrial Remodeling, Metabolic Reprogramming, and Heme Biosynthesis. *Immunity* **48**, 1144-1159 e1145, doi:10.1016/j.immuni.2018.04.031 (2018).

79 Salter, J. D., Bennett, R. P. & Smith, H. C. The APOBEC Protein Family: United by Structure, Divergent in Function. *Trends Biochem Sci* **41**, 578-594, doi:10.1016/j.tibs.2016.05.001 (2016).

80 Methot, S. P. & Di Noia, J. M. Molecular Mechanisms of Somatic Hypermutation and Class Switch Recombination. *Adv Immunol* **133**, 37-87, doi:10.1016/bs.ai.2016.11.002 (2017).

81 Xu, Z. *et al.* Regulation of aicda expression and AID activity: relevance to somatic hypermutation and class switch DNA recombination. *Crit Rev Immunol* **27**, 367-397 (2007).

82 Methot, S. P. *et al.* Consecutive interactions with HSP90 and eEF1A underlie a functional maturation and storage pathway of AID in the cytoplasm. *J Exp Med* **212**, 581-596, doi:10.1084/jem.20141157 (2015).

83 Hu, Y. *et al.* A combined nuclear and nucleolar localization motif in activation-induced cytidine deaminase (AID) controls immunoglobulin class switching. *J Mol Biol* **425**, 424-443, doi:10.1016/j.jmb.2012.11.026 (2013).

84 Maeda, K. *et al.* GANP-mediated recruitment of activation-induced cytidine deaminase to cell nuclei and to immunoglobulin variable region DNA. *J Biol Chem* **285**, 23945-23953, doi:10.1074/jbc.M110.131441 (2010).

85 Patenaude, A. M. *et al.* Active nuclear import and cytoplasmic retention of activation-induced deaminase. *Nat Struct Mol Biol* **16**, 517-527, doi:10.1038/nsmb.1598 (2009).

86 Aoufouchi, S. *et al.* Proteasomal degradation restricts the nuclear lifespan of AID. *J Exp Med* **205**, 1357-1368, doi:10.1084/jem.20070950 (2008).

87 Uchimura, Y., Barton, L. F., Rada, C. & Neuberger, M. S. REG-gamma associates with and modulates the abundance of nuclear activation-induced deaminase. *J Exp Med* **208**, 2385-2391, doi:10.1084/jem.20110856 (2011).

88 Arakawa, H., Hauschild, J. & Buerstedde, J. M. Requirement of the activation-induced deaminase (AID) gene for immunoglobulin gene conversion. *Science* **295**, 1301-1306, doi:10.1126/science.1067308 (2002).

89 Dickerson, S. K., Market, E., Besmer, E. & Papavasiliou, F. N. AID mediates hypermutation by deaminating single stranded DNA. *J Exp Med* **197**, 1291-1296, doi:10.1084/jem.20030481 (2003).

90 Chen, J. *et al.* The RNA-binding protein ROD1/PTBP3 cotranscriptionally defines AID-loading sites to mediate antibody class switch in mammalian genomes. *Cell Res* **28**, 981-995, doi:10.1038/s41422-018-0076-9 (2018).

91 Pham, P., Bransteitter, R., Petruska, J. & Goodman, M. F. Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. *Nature* **424**, 103-107, doi:10.1038/nature01760 (2003).

92 Rogozin, I. B. & 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* **172**, 3382-3384 (2004).

93 Yaari, G. *et al.* Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. *Front Immunol* **4**, 358, doi:10.3389/fimmu.2013.00358 (2013).

94 Wei, L. *et al.* Overlapping hotspots in CDRs are critical sites for V region diversification. *Proc Natl Acad Sci U S A* **112**, E728-737, doi:10.1073/pnas.1500788112 (2015).

95 Di Noia, J. M. & Neuberger, M. S. Molecular mechanisms of antibody somatic hypermutation. *Annu Rev Biochem* **76**, 1-22, doi:10.1146/annurev.biochem.76.061705.090740 (2007).

96 Kawamoto, T. *et al.* Dual roles for DNA polymerase eta in homologous DNA recombination and translesion DNA synthesis. *Mol Cell* **20**, 793-799, doi:10.1016/j.molcel.2005.10.016 (2005).

97 Delbos, F., Aoufouchi, S., Faili, A., Weill, J. C. & Reynaud, C. A. DNA polymerase eta is the sole contributor of A/T modifications during immunoglobulin gene hypermutation in the mouse. *J Exp Med* **204**, 17-23, doi:10.1084/jem.20062131 (2007).

98 Maul, R. W. *et al.* DNA polymerase iota functions in the generation of tandem mutations during somatic hypermutation of antibody genes. *J Exp Med* **213**, 1675-1683, doi:10.1084/jem.20151227 (2016).

99 Han, L., Masani, S. & Yu, K. Overlapping activation-induced cytidine deaminase hotspot motifs in Ig class-switch recombination. *Proc Natl Acad Sci U S A* **108**, 11584-11589, doi:10.1073/pnas.1018726108 (2011).

100 Reynaud, C. A., Anquez, V., Dahan, A. & Weill, J. C. A single rearrangement event generates most of the chicken immunoglobulin light chain diversity. *Cell* **40**, 283-291 (1985).

101 Reynaud, C. A., Anquez, V., Grimal, H. & Weill, J. C. A hyperconversion mechanism generates the chicken light chain preimmune repertoire. *Cell* **48**, 379-388 (1987).

102 Reynaud, C. A., Dahan, A., Anquez, V. & Weill, J. C. Somatic hyperconversion diversifies the single Vh gene of the chicken with a high incidence in the D region. *Cell* **59**, 171-183 (1989).

103 Williams, A. M., Maman, Y., Alinikula, J. & Schatz, D. G. Bcl6 Is Required for Somatic Hypermutation and Gene Conversion in Chicken DT40 Cells. *PLoS One* **11**, e0149146, doi:10.1371/journal.pone.0149146 (2016).

104 Budzynska, P. M. *et al.* Bach2 regulates AID-mediated immunoglobulin gene conversion and somatic hypermutation in DT40 B cells. *Eur J Immunol* **47**, 993-1001, doi:10.1002/eji.201646895 (2017).

105 Saribasak, H. *et al.* Uracil DNA glycosylase disruption blocks Ig gene conversion and induces transition mutations. *J Immunol* **176**, 365-371, doi:10.4049/jimmunol.176.1.365 (2006).

106 Di Noia, J. M. & Neuberger, M. S. Immunoglobulin gene conversion in chicken DT40 cells largely proceeds through an abasic site intermediate generated by excision of the uracil produced by AID-mediated deoxycytidine deamination. *Eur J Immunol* **34**, 504-508, doi:10.1002/eji.200324631 (2004).

107 Bastianello, G. & Arakawa, H. A double-strand break can trigger immunoglobulin gene conversion. *Nucleic Acids Res* **45**, 231-243, doi:10.1093/nar/gkw887 (2017).

108 Schildkraut, E., Miller, C. A. & Nickoloff, J. A. Transcription of a donor enhances its use during double-strand break-induced gene conversion in human cells. *Mol Cell Biol* **26**, 3098-3105, doi:10.1128/MCB.26.8.3098-3105.2006 (2006).

109 Cummings, W. J. *et al.* Chromatin structure regulates gene conversion. *PLoS Biol* **5**, e246, doi:10.1371/journal.pbio.0050246 (2007).

110 Wang, R. W., Lee, C. S. & Haber, J. E. Position effects influencing intrachromosomal repair of a double-strand break in budding yeast. *PLoS One* **12**, e0180994, doi:10.1371/journal.pone.0180994 (2017).

111 Chen, J. M., Cooper, D. N., Chuzhanova, N., Ferec, C. & Patrinos, G. P. Gene conversion: mechanisms, evolution and human disease. *Nat Rev Genet* **8**, 762-775, doi:10.1038/nrg2193 (2007).

112 Hatanaka, A. *et al.* Similar effects of Brca2 truncation and Rad51 paralog deficiency on immunoglobulin V gene diversification in DT40 cells support an early role for Rad51 paralogs in homologous recombination. *Mol Cell Biol* **25**, 1124-1134, doi:10.1128/MCB.25.3.1124-1134.2005 (2005).

113 Sale, J. E., Calandrini, D. M., Takata, M., Takeda, S. & Neuberger, M. S. Ablation of XRCC2/3 transforms immunoglobulin V gene conversion into somatic hypermutation. *Nature* **412**, 921-926, doi:10.1038/35091100 (2001).

114 Heyer, W. D., Ehmsen, K. T. & Liu, J. Regulation of homologous recombination in eukaryotes. *Annu Rev Genet* **44**, 113-139, doi:10.1146/annurev-genet-051710-150955 (2010).

115 Krejci, L., Altmannova, V., Spirek, M. & Zhao, X. Homologous recombination and its regulation. *Nucleic Acids Res* **40**, 5795-5818, doi:10.1093/nar/gks270 (2012).

116 Sebesta, M. & Krejci, L. in *DNA Replication, Recombination, and Repair: Molecular Mechanisms and Pathology* (eds Fumio Hanaoka & Kaoru Sugasawa) 73-109 (Springer Japan, 2016).

117 Renkawitz, J., Lademann, C. A., Kalocsay, M. & Jentsch, S. Monitoring homology search during DNA double-strand break repair in vivo. *Mol Cell* **50**, 261-272, doi:10.1016/j.molcel.2013.02.020 (2013).

118 Qi, Z. *et al.* DNA sequence alignment by microhomology sampling during homologous recombination. *Cell* **160**, 856-869, doi:10.1016/j.cell.2015.01.029 (2015).

119 Short, J. M. *et al.* High-resolution structure of the presynaptic RAD51 filament on single-stranded DNA by electron cryo-microscopy. *Nucleic Acids Res* **44**, 9017-9030, doi:10.1093/nar/gkw783 (2016).

120 Anand, R. P. *et al.* Chromosome rearrangements via template switching between diverged repeated sequences. *Genes Dev* **28**, 2394-2406, doi:10.1101/gad.250258.114 (2014).

121 Tsaponina, O. & Haber, J. E. Frequent Interchromosomal Template Switches during Gene Conversion in S. cerevisiae. *Mol Cell* **55**, 615-625, doi:10.1016/j.molcel.2014.06.025 (2014).

122 Boehm, T. *et al.* VLR-based adaptive immunity. *Annu Rev Immunol* **30**, 203-220, doi:10.1146/annurev-immunol-020711-075038 (2012).

123 Holland, S. J. *et al.* Expansions, diversification, and interindividual copy number variations of AID/APOBEC family cytidine deaminase genes in lampreys. *Proc Natl Acad Sci U S A* **115**, E3211-E3220, doi:10.1073/pnas.1720871115 (2018).

124 Magor, B. G. Antibody Affinity Maturation in Fishes-Our Current Understanding. *Biology (Basel)* **4**, 512-524, doi:10.3390/biology4030512 (2015).

125 Parra, D., Takizawa, F. & Sunyer, J. O. Evolution of B cell immunity. *Annu Rev Anim Biosci* **1**, 65-97, doi:10.1146/annurev-animal-031412-103651 (2013).

126 Becker, R. S. & Knight, K. L. Somatic diversification of immunoglobulin heavy chain VDJ genes: evidence for somatic gene conversion in rabbits. *Cell* **63**, 987-997 (1990).

127 Meyer, A., Parng, C. L., Hansal, S. A., Osborne, B. A. & Goldsby, R. A. Immunoglobulin gene diversification in cattle. *Int Rev Immunol* **15**, 165-183 (1997).

128 Butler, J. E. Immunoglobulin diversity, B-cell and antibody repertoire development in large farm animals. *Rev Sci Tech* **17**, 43-70 (1998).

129 Schiaffella, E., Sehgal, D., Anderson, A. O. & Mage, R. G. Gene conversion and hypermutation during diversification of VH sequences in developing splenic germinal centers of immunized rabbits. *J Immunol* **162**, 3984-3995 (1999).

130 Winstead, C. R., Zhai, S. K., Sethupathi, P. & Knight, K. L. Antigen-induced somatic diversification of rabbit IgH genes: gene conversion and point mutation. *J Immunol* **162**, 6602-6612 (1999).

131 Lindenmann, J. Origin of the terms 'antibody' and 'antigen'. *Scand J Immunol* **19**, 281-285 (1984).

132 Ehrlich, P. R. & Himmelweit, F. *The collected papers of Paul Ehrlich*. Vol. 1 (Pergamon, 1956).

133 Gearhart, P. J. Antibody wars: extreme diversity. *J Immunol* **177**, 4235-4236, doi:10.4049/jimmunol.177.7.4235 (2006).

134 Brenner, S. & Milstein, C. Origin of antibody variation. *Nature* **211**, 242-243, doi:10.1038/211242a0 (1966).

135 Hozumi, N. & Tonegawa, S. Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions. *Proc Natl Acad Sci U S A* **73**, 3628-3632, doi:10.1073/pnas.73.10.3628 (1976).

136 Bothwell, A. L. *et al.* Heavy chain variable region contribution to the NPb family of antibodies: somatic mutation evident in a gamma 2a variable region. *Cell* **24**, 625-637, doi:10.1016/0092-8674(81)90089-1 (1981).

137 Crews, S., Griffin, J., Huang, H., Calame, K. & Hood, L. A single VH gene segment encodes the immune response to phosphorylcholine: somatic mutation is correlated with the class of the antibody. *Cell* **25**, 59-66, doi:10.1016/0092-8674(81)90231-2 (1981).

138 Kim, S., Davis, M., Sinn, E., Patten, P. & Hood, L. Antibody diversity: somatic hypermutation of rearranged VH genes. *Cell* **27**, 573-581, doi:10.1016/0092-8674(81)90399-8 (1981).

139 Selsing, E. & Storb, U. Somatic mutation of immunoglobulin light-chain variable-region genes. *Cell* **25**, 47-58, doi:10.1016/0092-8674(81)90230-0 (1981).

140 Kohler, G. & Milstein, C. Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature* **256**, 495-497, doi:10.1038/256495a0 (1975).

141 Berek, C. & Milstein, C. Mutation drift and repertoire shift in the maturation of the immune response. *Immunol Rev* **96**, 23-41 (1987).

142 Imanishi-Kari, T., Reis, M. H., Weaver, D. & Baltimore, D. Altered repertoire of endogenous immunoglobulin gene expression in transgenic mice containing a rearranged mu heavy chain gene. *Cell* **55**, 541, doi:10.1016/0092-8674(88)90210-3 (1988).

143 Baltimore, D. *Dear Colleague*, <http://www-gateway.vpr.drexel.edu/files/Gateway_Project_Moshe_Kam/Resource/DBCre/colleague.html> (1988).

144 Jacob, J., Kelsoe, G., Rajewsky, K. & Weiss, U. Intraclonal generation of antibody mutants in germinal centres. *Nature* **354**, 389-392, doi:10.1038/354389a0 (1991).

145 Maizels, N. Might gene conversion be the mechanism of somatic hypermutation of mammalian immunoglobulin genes? *Trends Genet* **5**, 4-8 (1989).

146 Xu, B. & Selsing, E. Analysis of sequence transfers resembling gene conversion in a mouse antibody transgene. *Science* **265**, 1590-1593 (1994).

147 D'Avirro, N., Truong, D., Luong, M., Kanaar, R. & Selsing, E. Gene conversion-like sequence transfers between transgenic antibody V genes are independent of RAD54. *J Immunol* **169**, 3069-3075 (2002).

148 Steele, E. J. Reverse Transcriptase Mechanism of Somatic Hypermutation: 60 Years of Clonal Selection Theory. *Front Immunol* **8**, 1611, doi:10.3389/fimmu.2017.01611 (2017).

149 Franklin, A., Milburn, P. J., Blanden, R. V. & Steele, E. J. Human DNA polymerase-eta, an A-T mutator in somatic hypermutation of rearranged immunoglobulin genes, is a reverse transcriptase. *Immunol Cell Biol* **82**, 219-225, doi:10.1046/j.0818-9641.2004.01221.x (2004).

Chapter 2

Clustered mutations at the murine and human IgH locus exhibit significant linkage consistent with templated mutagenesis

In press at *The Journal of Immunology, 2019*.

Gordon Dale performed all the experiments in this manuscript, except as follows:

Daniel Wilkins primarily wrote PolyMotifFinder and RandomCheck

Trevor Bedford conducted linkage disequilibrium analyses (Figure 3, j-p)

Gordon Dale and Joshy Jacob wrote the manuscript.

ABSTRACT

Somatic hypermutation generates a myriad of antibody mutants in antigen-specific B cells, from which high-affinity mutants are selected. Chickens, sheep, and rabbits use non-templated point mutations and templated mutations via gene conversion to diversify their expressed Ig loci, while mice and humans rely solely on untemplated somatic point mutations. Here, we demonstrate that in addition to untemplated point mutations, templated mutagenesis readily occurs at the murine and human Ig loci. We provide two distinct lines of evidence that are not explained by the Neuberger model of somatic hypermutation: (1) Across multiple data sets there is significant linkage disequilibrium between individual mutations, especially among close mutations. (2) Among those mutations, those less than 8bp apart are significantly more likely to match micro-homologous regions in the IgHV repertoire than predicted by the mutation profiles of somatic hypermutation. Together, this supports the role of templated mutagenesis during somatic diversification of antigen-activated B cells.

INTRODUCTION

The ability of the humoral immune system to respond with high affinity to a vast assortment of antigens is critically dependent on the linked processes of somatic hypermutation and affinity-driven selection. The former generates somatically-mutated antibody substrates, from which clones with higher affinity than their peers can be selected to undergo further somatic mutation. This process repeats itself until there is an emergence of a high affinity clone, specific for an antigen.

Evolutionarily, this strategy of somatic diversification and selection, emerged approximately 500 million years ago in vertebrates [1]. In these earliest humoral responses, diversification occurred using a combination of a cytidine deaminase in tandem with gene conversion [2]. Gene conversion is a mechanism of DNA repair wherein a highly homologous sequence is used as a template to repair a damaged region resulting in the copying of the template sequence into the damaged region. Further along the vertebrate evolutionary tree, the strategy of utilizing a cytidine deaminase to generate diversity is preserved throughout Mammalia (mammals) and Aves (birds). Chickens are known to rely heavily on activation induced cytidine deaminase (AID) and gene conversion. Conversely, mammals are canonically split, with animals such as rabbits [3], cattle [4], and sheep [5] being known to utilize gene conversion, whereas mice and humans are thought to primarily rely on AID and other nontemplated mutations generated from downstream processing of AID activity [6].

Here, we present evidence that somatic hypermutation in murine and human B cells utilizes a gene conversion-like mechanism, referred to hereafter as templated mutagenesis, to generate somatic variants. We observe templated mutagenesis in murine germinal center B cells as well as in terminally differentiated plasma cells. This observation is shared in human

IgM$^+$/IgG$^+$/IgA$^+$ CD138$^+$ plasmablasts. We also observe linkage disequilibrium between mutations that is inversely related to genetic distance between those mutations at lengths <100bp. Templated mutagenesis utilizes donors from variable segments 5' to the rearranged VDJ in both mice and humans as well as from those on the other allele. Lastly, we find that non-immunoglobulin sequences placed at the IgH locus, by transgenes in mice and insertions in humans, mutate such that they share microhomology with tracts from the IgHV repertoire. Taken together, our studies demonstrate a role for templated mutagenesis during somatic hypermutation of murine and human B cells.

RESULTS

Templated mutations occur in murine IgM$^+$ plasma cells

We have recently demonstrated the existence of antigen-specific, long-lived, somatically-mutated IgM plasma cells that have a high ratio of framework (FRW) mutations to that of the complementarity-determining regions (CDR) mutations [7]. It is widely held that framework mutations are selected against since they are likely to disrupt the architecture of the antibody. Based on the current model of mutagenesis by point mutations, it is expected that FRW mutations occur more frequently than those in the CDRs and that these deleterious mutations are quickly selected against [8,9]. We posit that the FWR mutations in IgM plasma cells were non-deleterious and introduced via templated mutagenesis from other IgH FWR presumably through a gene conversion-like mechanism. To determine whether somatic mutation we observed in IgM$^+$ plasma cells was consistent with gene conversion, we intraperitoneally immunized cohorts of mice with 50µg of 4-hydroxy-3-nitrophenylacetyl labelled chicken-γ-globulin (NPCGG). We have previously shown these mutations to be dependent on the enzyme activation-induced cytidine deaminase (AID) [7], which is critical to both somatic hypermutation and gene conversion [10,11]. These mutations also result in the expected ~2:1 ratio of transitions and transversions, which is characteristic of somatic hypermutation [7]. As previously reported, we observe a high frequency of replacement mutations in the framework region of IgM plasma cells as compared to those in the CDRs (Figure 1A). Further analysis of the ratio of replacement mutations to silent mutations in CDRs and FWRs revealed an enrichment of replacement mutations in framework regions 2 and 3 (Figure 1B).

We next aligned somatically-mutated IgHV 1-72 (V186.2) sequences, which canonically encode antibodies specific to the NP hapten [12-14], against the germline IgHV 1-72 - as well as the closely related sequences IgHV 1-55 and 1-53. Sequences shown in Figure 1C are representative of four individual mice. We find that many mutations within and between individual mice can be explained by gene conversion tracts (A…AAC…A) and (A…A…G) from the heavy chain variable genes segments IgHV 1-55 and 1-53, respectively (Figure 1C). IgHV 1-55 and -53 tracts occur in 5% and 20% of IgHV 1-72 sequences, respectively (Figure 1D). Together both tracts accounted for 25% of the mutation load observed in IgHV 1-72 (Figure 1E). Surprisingly, these tracts were observed in different clones within individuals and in different animals, strongly suggesting either concerted selective pressure in the germinal center or templated mutagenesis.

To examine whether these tracts were the result of selective pressure within the germinal center microenvironment, we analyzed linkage disequilibrium between silent and replacement mutations. Linkage disequilibrium (LD) is a measure that is used in genetics to test whether mutations are inherited together. In tracts from IgHV 1-53 (A…A…G), the first is a replacement mutation while the second and third are silent. As silent mutations are not selected for within the germinal center, it would follow that associations between silent mutations or between silent and replacement mutations would not be expected. A LD analysis found significant associations between silent mutations in tracts donated from IgHV1-53 ($p<0.001$). Additionally, the LD was significant between the replacement mutation and both silent sites in the A…A…G tract from IgHV 1-53 ($p<0.001$). All mutations analyzed in tracts donated by IgHV1-55 (A…AAC…A) were replacement mutations and therefore could only be analyzed for linkage disequilibrium between replacements. However, analysis did reveal significant associations between the set of replacement mutations ($p<0.001$). Taken together, this suggests that IgM+ plasma cells do undergo somatic mutation consistent with templated mutagenesis.

Though unlikely, it is formally possible that the mutations could be due to PCR on bulk B cell population [15]. To rule this out we analyzed the IgHV 1-72 sequences as reported by Tas et al. who sequenced sorted, single cells [16]. We found that these single-cell derived IgHV 1-72 sequences possessed the replacement “A” and silent “A” mutations of the IgHV 1-53 tract in two distinct clones (Figure 1F). We also observed a multitude of clusters of mutations that share regions of varying microhomology with other IgHV genes, as opposed to the macro-homology observed between the IgHV 1-72 sequence and the germline IgHV 1-55 and -53 donors.

To further validate this, we sorted single plasma cells into 96 well plates and carried out PCR amplification for the gene rearrangements from Ig heavy and light chains. We obtained a total of 489 IgHV sequences but did not find any IgHV 1-72 sequences (n=6) with the IgHV 1-55 or IgHV 1-53 mutation tract in this limited cohort of IgHV 1-72 sequences. However, we did obtain somatically-mutated IgHV 1-53 sequences as well as somatically-mutated IgKV sequences (Figure S1). As observed in our analysis of the sequences reported in Tas et al., we found that clusters of mutations shared microhomology with germline IgHV and IgKV genes. As we find regions of microhomology accounting for mutations in both data sets, this suggests that the mutations observed in IgM plasma cells is not due to bulk PCR error. In addition, this suggests that clusters of mutations may be due to templated mutagenesis.

Templated mutations occur in murine germinal center B cells

IgM plasma cells appeared to have mutations consistent with templated mutagenesis, but we reasoned that tracts would be difficult to identify in IgG plasma cells since they accrue a high mutation burden. Hence, we analyzed developing (day 12) germinal center B cells as they have just begun to initiate somatic mutation. As hypothesized, we observed long (9-68bp) tracts from IgHV genes 1-55 and 1-53 in each of the three mice in both IgM^{+} or IgG^{+} germinal center B cells (Figure 2 A-B). We also observed the expected profile of somatic hypermutation, with transition mutations occurring twice as frequently than transversion mutations (Figure 2C-D). Long templated tracts of mutations occurred more frequently and accounted for more mutations in IgM germinal center B cells than in those that have switched to IgG (Figure 2E-G). To ensure that this pattern of mutation was not limited to the NPCGG response, we also analyzed Peyer's patch germinal center B cells which respond to gut flora antigens (Figure 1G) and identified tracts demonstrating that templated mutagenesis is not restricted to the hapten-carrier model.

It is possible that the clusters of mutations could have been introduced as independent point mutations as opposed to tracts. To distinguish between the two, we analyzed linkage disequilibrium of IgHV 1-72 sequences from IgM^{+} and IgG^{+} germinal center B cells from day 12 along with all mutated sequences (n=11,456) from day 10, 12, and 14 NPCGG-induced splenic germinal centers. The data for all IgHV genes was then plotted as a function of distance between mutations (Figure 2H-J). Strikingly, we find high levels of linkage at lengths <50bp with multiple instances of complete linkage at these lengths as measured by the correlation coefficient

r^2, where r^2=LD. Furthermore, a LOESS regression demonstrates that linkage decreases with increasing distance between mutations. To further validate this result, we reanalyzed data from Kuraoka et al. [17] who sequenced individual antigen-specific germinal center B cells from mice immunized 8 or 16 days earlier with either recombinant *Bacillus anthracis* protective antigen (rPA) or influenza hemagglutinin (rHA). Ig sequences from rPA at day 8 and day 16 as well as rHA at day 16 demonstrated the same LD phenomenon, suggesting that increasing LD at lengths <100bp was independent of the type of antigen used (Figure 2K-M). We also analyzed the CGG-specific sequences obtained from Tas et al. [16], which also demonstrates the same pattern observed in the Kuraoka sequences as well as those presented here (Figure 2N). Lastly, we analyzed somatically-mutated rabbit heavy and chicken light chain sequences [18-22], which undergo gene conversion predominantly and untemplated point mutation to a lesser extent and we found increasing LD with decreasing distance between mutations (Figure 2O-P).

It is known that AID preferentially targets specific motifs for hypermutation, thus we analyzed whether the observed increase in linkage was due to linked activity at nearby AID hotspots. We compared mutations at the canonical AID hotspot WRC/GYW to determine if there is an associated increase in linkage at mutations in close proximity. We analyzed the data sets from Tas et al., Kuraoka et al., as well as mutated rabbit and chicken sequences [16-22]. We found that there were few WRC/GYW pairs that were mutated across data sets and that there was no consistent pattern to linkage between mutations at WRC/GYW sites (data not presented). An in-depth analysis of somatically mutated IgHV 1-72 sequences from splenic germinal centers from CB6F1/J mice following immunization with NPCGG (d12 p.i.) considered the recently reported hotspots (CRCY/RGYG and ATCT/AGAT) in addition to the canonical (WRC/GYW) revealed an increase in linkage equilibrium among mutations <100bp consistent with data presented in

Figure 2E-M. However, the linkage between these sites represented a small fraction of the overall observed LD at distances <100bp, suggesting a minor contribution of linked AID deamination to observed LD between mutations at these sites (data not presented).

Although AID activity may be responsible for some of the observed LD that we observe across our data and others, we find the bulk of the LD occurs outside of regions that are AID hotspots suggesting that the paradigm of multiple AID-induced point mutations is not responsible for our LD observations. Though this data is not consistent with an AID-mediated pattern, it is consistent with gene conversion-like templated mutagenesis, as multiple mutations may be carried over together by the same templated event as evidenced from our analysis of somatically-mutated rabbit and chicken sequences.

Templated donor tracts primarily originate from 5' upstream V gene segments but can also originate from the trans allele

In Fig.1 we showed that IgHV 1-72 rearrangements exhibit tracts from IgHV 1-55 and IgHV 1-53 gene segments. Interestingly, VDJ rearrangement of the IgHV 1-72 gene segment results in cis-deletion of the IgHV 1-55 and IgHV 1-53. Hence, we hypothesized that the only available template for these genes would be located on the trans allele. To test the contribution of the trans allele, we used a first filial generation BALB/cJ x C57BL/6J cross (CB6F1/J) for our studies. These mice express two different IgHV loci that express strain-specific alleles enabling identification of donor tracts from either allele into the expressed V_H gene.

To quickly identify tracts between IgHV genes from either the BALB/c or BL/6 locus we developed a computational script, PolyMotifFinder, that allows us to find small tracts of two or more mutations within 8bp donated from a set of reference sequences (Figure S1). We chose this strategy for PolyMotifFinder as we observed many mutations in close proximity that exhibited significant linkage disequilibrium, suggesting that these mutations may be due to a templated tract. We tested the robustness of the script to detect gene conversion tracts on the rabbit immunoglobulin sequences of Sehgal et al. [23] and found that this script detects >96% of mutations reported as gene conversion mutations (data not presented). PolyMotifFinder is supplemented with another script, RandomCheck (Figure S2), that is made to simulate the base pair substitution pattern of somatic hypermutation. Within each analysis, PolyMotifFinder generates a “Gene Conversion (GC) coverage” value for each sequence. That value is stored and RandomCheck assigns new mutations at the same position based on the profile of somatic hypermutation, followed by another GC coverage value calculation. The RandomCheck process is iterated either 100 or 1000 times to build a population, which the initial GC coverage is compared against, generating a Z score. The population of Z scores than then be combined into a single analysis using Stouffer’s Z method.

We immunized five CB6F1/J mice with NPCGG as above, sorted, and sequenced splenic germinal center B cells. The NP response differs in BALB/c and BL/6 mice, with different heavy chains encoding the antigen specific response, termed the NP^a and NP^b response and utilizes heavy chains IgHV 14-3 and IgHV 1-72, respectively [24]. To determine whether tracts could originate from either cis or trans alleles, we analyzed IgHV 1-72 (NP^b) and IgHV 14-3 (NP^a) sequences (Figure 3A). In each of these sequence sets we observed increasing LD at decreasing genomic distance between mutations, as before (Figure 3B-C). Sequences were then analyzed with PolyMotifFinder and RandomCheck using the base pair substitution profiles generated from Maul et al. [25] (Figure 3D). Each set of sequences was compared to the germline IgHV gene segments in its entirety, the BALB/c specific IgHVs, the BL/6 IgHVs, 8-mer motifs specific to only the BALB/c or BL/6 IgHV repertoire, as well as the set of 8-mers not represented in any IgHV (negative control). All germline IgHV sequences were obtained via IMGT[26]. In both IgHV 1-72 and IgHV 14-3, all groups were significant against the negative control. We also find an average Stouffer's Z value of approximately five in both sets when the sequences are compared to the entire IgHV repertoire, which suggests that templated mutagenesis is occurring over the background of modeled somatic hypermutation (Figure 3E-F). Both the BL/6 and BALB/c specific motifs had Stouffer's Z scores around zero, suggesting that there was not significant matching to either of these sequence sets over what is predicted by somatic hypermutation. Conversely both the BL/6 and BALB/c IgHV sets produced average Stouffer's Z scores comparable to that of the entire IgHV repertoire, suggesting that templated mutagenesis is utilizing motifs shared between the BL/6 and BALB/c IgHV gene segments but did not resolve whether mutations were occurring in *cis* or in *trans*.

Since F1 mice possess a single copy of the BL/6 IgH locus, templated mutation that occurs in cis is predicted to be directional, as IgHV genes lost during VDJ recombination are not preserved on the opposite locus. Therefore, for any given IgHV rearrangement, we could analyze the donors located 5' as occurring in cis, and any matching to downstream donors is inferred to occur in trans with homologous motifs located on the BALB/c IgH locus. Although this analysis is possible with the BL/6 IgH locus, as the locus map is known, it was not possible to do this analysis on the BALB/c locus as there remains some ambiguity on its organization[27]. As such, we limited our analysis to the BL/6 locus alone. We performed deep sequencing of the antibody repertoire present in the germinal centers of F1 mice immunized with NPCGG. For each mouse, each BL/6 IgHV gene is shown in the order in which the locus is organized. Across mice, 5' donors are preferentially utilized (Figure 3G) (p<0.0001). However, analysis of the 3' donors reveals that there is significant enrichment of 3' donor matching as compared to the somatic hypermutation null, suggesting that templated mutation can also occur in trans, albeit with a much lower frequency (Figure 3H). Interestingly, we observed that IgHV genes located towards the 5' end of the IgH locus displayed a consistent enrichment of 3' matching, suggesting the lack of available templates forces templated mutation to occur in trans, as observed in IgHV 1-72 sequences earlier.

Templated mutation occurs in human plasmablasts

We next sought to investigate if somatic hypermutation in human B cells also occurred due to templated mutagenesis, as we have observed in mice. To do so, we analyzed circulating plasmablasts isolated from the blood of a healthy human donor. Sequences from these cells were generated using ultra-high throughput single cell sequencing as described in DeKosky et al. [28]. As shown in other sequence sets, linkage disequilibrium increases at mutation distances less than 100 bp (Figure 4A). To determine whether this observation was due templated mutagenesis, we analyzed each antibody rearrangement in our sample and compared whether there was significant matching to germline human IgHV gene segments as defined by IMGT [26], or to other 8-mers not present in the germline IgHV gene segments using the human SHM base pair substitution profiles derived from Longo et al [29] (Figure 4B). There is significant matching to the IgHV germline across all IgHVs sequenced (Stouffer's Z Trend: 4.33, $p<0.0001$) as compared to matching motifs not present in the IgHV germline gene segments (Stouffer's Z Trend: -35.62, $p=1$) (Figure 4C). When analyzed across isotypes (IgM/IgG/IgA) we found a similar pattern in each group (Stouffer's Z trend: 1.5235, 1.7664, 1.5922, respectively) compared to the negative control of 8-mers not present in the human IgHV repertoire (Stouffer's Z trend: -17.6049, -

12.3690, -16.1749, respectively) (Figure 4D-F). Lastly, we hypothesized that templated mutagenesis would occur primarily using 5’ donors as demonstrated in mice. Analysis via PolyMotifFinder and RandomCheck demonstrated statistically significant matching to upstream 5’ donors in human plasmablasts as compared to those downstream (Stouffer’s Z trend: 4.04 and 1.84, respectively, p<0.05) (Figure 4G). Together, these results suggest a similar pattern of somatic hypermutation, consistent with templated mutagenesis, in diversification of antigen-activated human B cells.

Templated mutations occur in non-Ig genes inserted into the IgH locus

It has been shown that transgenes inserted into the murine Ig loci as well as insertion of LAIR gene into the human Ig loci undergo somatic mutation [30,31]. Thus, we next sought to examine the patterns of mutation of non-immunoglobulin genes that undergo somatic hypermutation at the IgH locus in mice and humans. To do this, we compared the somatically-mutated LAIR1 insert of the V-LAIR1-DJ antibodies reported in Tan et al. [30] (Figure 5A) against the human IgHV repertoire using PolyMotifFinder and RandomCheck scripts. We also utilized the transgenes □-globin and GPT placed at the passenger IgH allele of mice, where they are not subject to selective pressure, against the murine IgHV repertoire [32] (Figure 5B). As before, we also compared these sequences to 8-mers not present in the respective IgHV repertoires (negative control). In humans, somatically-mutated LAIR1 preferentially matches the IgHV repertoire as compared to the negative control (p=0.004) (Figure 5D). With regards to mice, Alt and colleagues published an elegant paper [32] wherein they generated genetically engineered mice which carried passenger transgenes, β-globin and GPT introduced into the Ig loci. These transgenes cannot undergo selection as they were engineered to be transcribed but not translated.

Interestingly, the passenger transgenes β-globin and GPT displayed the same pattern as LAIR1 and are statistically significant in matching motifs present in the murine IgHV repertoire (p=0.018 and p=0.015, respectively) (Figure 5D-F). For LAIR1, β-globin, and GPT, the negative control was not significant (p= 0.972, p= 0.995, p= 0.998, respectively). This suggests that patterns of mutations in these three datasets (one human and two murine) are consistent with the templated mutagenesis from the IgHV gene segments as the pattern of somatic hypermutation statistically results in motifs matching those in the IgHV repertoire.

To demonstrate that this effect is directly due to templated mutagenesis acting on both IgHV and these non-immunoglobin sequences, we conducted a second analysis in which we first analyze somatically-mutated IgHV sequences and extract those motifs with two or more mutations that match an 8-mer in the IgHV repertoire. These matched motifs are then used as the reference against which somatically-mutated non-immunoglobulin sequences are compared (Figure 5C). We reasoned that the templates (motfis) used for templated mutagenesis of the IgHV gene segments would be the same as those used for the non-immunoglobulin sequences if templated mutagenesis were occurring. Thus, we hypothesized that we would observe a stronger effect when analyzing the somatically-mutated non-immunoglobulin sequences against these enriched motifs, since we only selected the motifs used during templated mutagenesis and excluded those that were not. As hypothesized, Stouffer's Z trend generated from the enriched pool rises for each sequence set when compared against the IgHV references (p=0.023) (Figure 5D-F). Further analysis has also shown that the enrichment effect on Stouffer's Z trend is not limited to enriching for motifs from somatically-mutated IgHV genes with comparison to mutated non-immunoglobulin sequences but also occurs if the enrichment occurs from somatically-mutated non-immunoglobulin genes and mutated IgHVs are compared to that enriched pool (Figure S3). These results demonstrate that the pattern of templated mutation is consistent between somatically-mutated IgHV genes and exogenous genes inserted into the IgH locus and that such effect is not due to the intrinsic activity of canonical somatic hypermutation.

Lastly, we sought to quantify the contribution of templated mutagenesis to somatic hypermutation. In order to do so we conducted a conservative and limited-in-scope analysis that (1) considered mutations within 8bp of at least on other mutation, (2) only considered one instance of a given pair of mutations (to eliminate double counting within a dataset), and (3) did not factor clonality so as to remove any bias from clonal dynamics. This approach allows us to determine the likelihood that a given set of mutations in close proximity has a template in the IgHV segment repertoire. We analyzed human and murine sequence sets from Tas et al. [16], human plasmablasts, the LAIR1 gene segments [30,31], as well as the passenger transgenes (β-globin and gpt) from Yeap et al. [32]. We find that approximately 50-65% of unique mutations fulfill two conditions: 1) proximity to at least one other mutation within 8bp, and 2) there exists a template for those mutations in the IgHV germline gene segments (Figure 5I). Strikingly, this effect extends into the somatically mutated non-Ig sequences, LAIR1, gpt and β-globin, despite the lack of overt homology between these genes and the IgHV repertoire. Extrapolation of this data suggests that 3 out of every 5 mutations at the IgH locus are consistent with templated mutagenesis, whether the gene being mutated is an antibody gene segment or a non-immunoglobulin gene. This is especially significant for the passenger transgenes β-globin and gpt which are not subjected to selection and display the full spectrum of mutations as they occur. Together, this data suggests that templated mutation contributes heavily to mutation clusters, specifically those that are within 8bp of one another.

DISCUSSION

In the present study we provide evidence of templated mutagenesis occurring during antigen-driven somatic diversification of human and murine B cells. We find that this is an intrinsic property of somatic hypermutation, at least as it occurs at the IgH locus in both species, as the non-immunoglobulin sequences examined are enriched for motifs that are exceedingly unlikely to have occurred by the effects of canonical somatic hypermutation. We suspect but cannot confirm within the scope of the current study, that the mechanism of templated mutagenesis is gene conversion. The repeated occurrence of IgHV repertoire motifs within somatically-mutated sequences is akin to that seen in other species such as the chicken, which relies heavily on acquisition of pseudogene motifs into the productive antibody rearrangement [33]. However, unlike chickens, the pattern of templated mutagenesis in mice and humans is largely based on IgHV microhomology, as small fragments around pairs of somatic mutations disproportionately match germline IgHV motifs. This contrasts with typical reports of gene conversion in the literature which rely on large alignments of potential donors and recipient sequences to demonstrate that the mutations are templated. Further, reports of gene conversion in the literature rely on demonstrating an arbitrary number of mutations matching to a pseudogene sequence to establish a given gene conversion tract.

In this work, we address these two limitations by applying a microhomology and statistical approach to identifying potential gene conversion tracts that queries whether it was possible to use fragments of IgHV genes to reconstruct highly-mutated antibody sequences. This process was further extended in our script, which identifies if “neomotifs” – those generated after introduction of somatic mutations – are actually represented in the genetic repertoire of the IgHV loci and does so without regard to large alignments or total number of templated mutations in a tract. This has allowed us to identify potential gene conversion tracts that would have otherwise been excluded in a conventional analysis. Interestingly, gene conversion has been previously suggested to occur in both mice [34-37] and humans [38,39], but has generally been regarded as an infrequent event or even thought to not occur at all [40].

Here, we argue in favor of the alternative, where gene conversion or a mechanism akin to it, is a frequent contributor of somatic mutations along the length of the IgHV gene. This is most evident by the ability of "neomotifs" present in non-selected, passenger transgenes *GPT* and β-globin [32] to (1) match murine IgHV reference sequences and (2) for those motifs that do match the IgHV repertoire, produce significant results for somatically-mutated IgHV genes (Figure S3). This demonstrates that the process of templated mutagenesis is a significant contributor to somatic hypermutation, else such an effect would not occur. Indeed, our conservative estimates of the contribution of templated mutagenesis to somatic hypermutation suggest that between 50-65% of unique mutations occur 1) in close proximity to another mutation and 2) have a putative donor sequence in the IgHV germline repertoire that explains both, or more, mutations in that cluster. That this frequency is observed even within the non-immunoglobulin passenger transgenes in Yeap et al [32] is striking because despite a lack of overt homology with the IgHV repertoire, and absence of selective pressure (Figure 5B), they remarkably have clusters of mutations that match templates in the IgHV repertoire. Coupled with the finding that templates used for diversifying the IgHV genes also serve as templates for diversifying these non-immunoglobulin genes, these findings strongly support the notion of a templated mechanism for the production of local micro-clusters (≤8bp) of mutations.

Direct additional evidence for this templated pattern of mutation, while rare, can be found in published literature. To illustrate, in the classic Nature paper, Weiss and colleagues [12] microdissected individual germinal center B cells, sequenced them and demonstrated that somatic mutants are generated intra-clonally within germinal centers. In Figure 2 of this paper, sequences GC24.I6 and GC24.I12 both contain a replacement G→A substitution at codon 9, and a silent G→A substitution at codon 10. This pair of substation mutations are seen in the IgHV 1-53 tracts presented in our manuscript. Examples of this tract in our manuscript can be seen across multiple clones in both IgM plasma cells, and in single cell sequenced germinal center B cells from the Victora Lab [16], depicted in Figures 1C and D respectively. The IgHV 1-55 tract appears less frequently in the literature but we have observed it in Jacob et al. [41], Figure 6. Sequence B17-2 possesses the three-nucleotide replacement GCA→ AAC in codon 34-35 that we observe in the IgM plasma cells sequences (Fig.1C; current manuscript).

As to why such tracts may seem rare in the literature – when Kelsoe and colleagues conducted these studies in 1991, there was a primary concern for PCR cross-over artifacts. At the time, sequencing was best done by successful manual dissection and there was no obvious control for successful isolation of a single cell. Thus, when these studies were done, “PCR hybrids” that were found were eliminated from any further analysis. Interestingly, the occurrence of PCR hybrid generation was investigated by Kelsoe and colleagues [13], Figure 8. They found a steady increase in the formation of PCR hybrid products as the germinal center reaction progressed. At the time of the study, this effect was attributed to either DNA nicking or apoptotic DNA fragmentation – neither of which could be a possible explanation for our observations in the current manuscript since the IgM plasma cells are (1) successful emigrants from the germinal center reaction and are expected to have repaired any existing DNA lesions, (2) the IgM plasma cells were not undergoing apoptotic processes as sorted cells were stained with a live/dead stain (Annexin V) and were gated on live cells during sorting, and (3) our sequence data was generated from cDNA libraries and reflect the mRNA transcripts present in the IgM plasma cell population, thus excluding DNA damage as a potential explanation. These findings, however, are consistent with templated mutagenesis and the observations presented here.

One of the primary arguments against the occurrence of gene conversion during murine somatic mutagenesis was a series of studies by Bross et al. [42]. These studies sought to elucidate the role of homologous recombination (gene conversion) during somatic hypermutation by analyzing mice deficient in RAD54. The primary function of RAD54 is to facilitate homologous recombination through branch migration [43]. They found that the frequency and pattern of somatic hypermutation was unaltered despite the absence of RAD54 and concluded that there was no contribution of RAD54 to somatic hypermutation. In contrast, another study published around the same time by D'Avirro et al. [44] demonstrated a gene conversion-like phenomenon in IgH knock-in mice that were RAD54$^{-/-}$. They concluded that the gene conversion-like phenomenon was independent of RAD54, suggesting that templated mutagenesis could still readily occur in the absence of RAD54. This is further supported by studies in *S. cerevisiae* that demonstrate that disruption of the RAD54 gene produces a mild but not critical defect in mating-type switch, a process entirely dependent on gene conversion [45]. Based on the results from these studies, it remains plausible that gene conversion-like events occur and can do so independently of RAD54.

The current paradigm of somatic hypermutation is the Neuberger model [46], which describes the many processing events that occur downstream of AID-mediated deamination of cytosine. While the Neuberger model accounts for many observations made regarding somatic hypermutation, the data presented here is not adequately explained under this paradigm. Foremost, we observe a replicable increase in linkage disequilibrium of somatic mutations as the distance between those mutations decreases. Secondly, we find that these pairs of somatic mutations in close proximity, with the highest linkage, disproportionately match the IgHV repertoire, regardless of whether the mutated sequence is an IgHV gene segment or a non-immunoglobulin gene. Together, both lines of evidence suggest that templated mutation is an actively-occurring process during somatic hypermutation. Such a mechanism carries significant implications for the generation of diversity during the humoral immune response and further work should be aimed at elucidating the functional impact such a mechanism has on antibody affinity maturation.

METHODS AND MATERIALS

Experimental Model and Subject Details

Mice

C57BL/6 mice and CB6F1/J (C57BL/6 x BALB/cJ F1 hybrid) mice were purchased from The Jackson Laboratory. All mice were maintained in a specific pathogen-free facility in accordance with the institutional guidelines of The Animal Care and Use Committee at Emory University.

Method Details

Immunizations

Cohorts of female C57BL/6 or CB6F1/J mice were immunized intraperitoneally with 50µg hydroxyl-3-nitrophenylacetyl-chicken-γ-globulin ($NP_{22}CGG$) (Biosearch Technologies) with 50µL alum in PBS for a total volume of 200µL. Spleens and/or bone marrow were collected at time points described post-immunization.

High-throughput heavy chain sequencing

Total plasma cells (CD138^{+}B220^{-}), germinal center B cells (CD19^{+}GL7^{+}), or bone marrow B cells (B220^{+}CD138^{-}CD19^{+}CD25^{+}IgM^{-}CD43^{-}) were isolated via cell sorting from the spleens and bone marrow, respectively, of mice at 30 days post-immunization with NPCGG. Human plasmablasts (CD19^{+}IgD^{-}CD27^{+}CD38^{+}CD138^{+}) were sorted from peripheral blood of a healthy donor. Lysates were amplified and sequenced as previously reported [7,47]. Raw sequence data was then analyzed for mutations using The International Immunogenetics Information System® HighV-QUEST (IMGT.org) [26]. Descriptive statistics for the sequence data can be found in Suppl. Table I, Part A. Sequences used in the analyses presented here can be found in Suppl. Table I, Part B.

Linkage Disequilibrium Plots

Plots of linkage disequilibrium between mutations were generated either with a custom python script, titled LD-analysis or a Matlab implementation of the correlation based tests described in Zaykin et al. [48].

The python script, LD-analysis, identifies major and minor alleles of polymorphic (mutated) sites from sequences grouped by IgHV gene usage. Only biallelic sites were considered. Major alleles are defined as those that are most common for a given site. Minor alleles are defined as the second most frequent alleles that also occur at frequency of ≥ 10% of observed sequences. From the pool of alleles for each position, haplotypes are generated. A chi-squared value is calculated for each haplotype and is subsequently used to generate a squared correlation coefficient (r^2) value that is reported for each haplotype in each LD plot.

The correlation-based tests were utilized for data sets where the restriction of data analysis to only biallelic sites was prohibitive with respect to number of haplotypes observed. In these cases, we report the maximum r^2 for a given haplotype pair.

Data presented is summed from all observed IgHV genes in each experiment, unless otherwise stated.

PolyMotifFinder and RandomCheck

To identify the number of potential templated mutagenesis events that could have contributed to observed somatic mutations, we generated a script with three objectives: 1) to identify mutations in sequences, 2) create motifs that include the mutated site, and 3) query this against the reference sequences. PolyMotifFinder is a script developed in Matlab (v.R2017b) for the purpose of identifying k-mer matches between raw sequence and reference sequence datasets and studying their distribution and frequency (Figure S1). As inputs, PolyMotifFinder uses FASTA formatted files of the raw sequence data, an alignment of the raw sequence data to the germline sequence, and a series of unaligned reference sequences, in this case generated from the IMGT database [26]. The script first identifies the positions where the aligned raw sequence and germline sequence exhibit mismatches, which is stored as a matrix of mutation positions. Then, the script identifies k-mer substrings from the unaligned raw sequence that incorporate two or more polymorphic positions that do not match the unaligned IMGT germline sequence (i.e. are mutations). Pulling the k-mers from unaligned sequences prevents gaps from indicating false mismatch polymorphisms when comparing to the IMGT sequences. The k-mer substrings are then quired against the IMGT reference sequences and if there is a matching IMGT reference sequence, the k-mer's coordinates are annotated in a matrix to indicate the length of the match. This matrix is compared to the matrix of mutation positions for each respective sequence. We tally the number of mutation positions that have a corresponding k-mer match and divide that by the number of mutations for a gene conversion (GC) coverage value for each sequence. This value is utilized below in tandem with the RandomCheck script.

The Matlab script RandomCheck was developed to produce a baseline for motif matching to compare the GC coverage results of PolyMotifFinder against. As input, the script takes the same data set as a PolyMotifFinder run. It identifies two or more mutations within one k-mer and randomizes the mutations, either keeping the polymorphism the same or changing it to another non-germline base. The randomization process is done such that it conforms to the base pair substitution profiles extensively reported to be characteristic of somatic hypermutation (Figures 3G and 4B) and is based on data reported from Longo et al. (human) and Maul et al. (mice) [25,29]. For example, to analyze a murine data set, a T→A mutation at position 50 of a given sequence is given a 27.8% chance of remaining a T→A mutation, a 55.6% chance to change to a T→C mutation, and a 16.7% chance to change to a T→G mutation. This new sequence is run through the same process as PolyMotifFinder, identifying how often the motifs match with somatic hypermutation-modeled combinations at the same positions as a real dataset. The GC coverage for each of these sequences is calculated and stored. This process is repeated with new somatic hypermutation modeled mutations 100 or 1000 times with the same k for direct comparison of results to the associated PolyMotifFinder run. For each sequence, the PolyMotifFinder GC-coverage is compared to the respective sequence's GC-coverage based on somatic hypermutation modeling, generating a Z-score.

Quantification and Statistical Analysis

Linkage disequilibrium analyses used for IgM plasma cells were conducted using correlation based tests described in Zaykin, Pudovkin, and Weir [48]. Nucleotide positions that correspond to a putative gene conversion event were identified and used to calculate a R^2 value between all nucleotide positions that belong to a candidate gene conversion tract. For example, in Figure 1A, we analyzed nucleotide positions 25, 33, and 36 – which correspond to the A-----A--G tract from 1-53. Each pair of positions was tested for linkage. In this case, the statistic would compute the correlation between position 25 and 33, 25 and 36, and 33 and 36. These R^2 values would then be assessed for significance by permutating the alleles of position 25, 33, and 36 and subsequently generating an R^2 value for each permutation. The original R^2 values are compared to the population of permutational R^2 values to generate a p-value. Exact (permutational) p-values are reported and are based on the value and distribution of R^2. 19,999 permutations were conducted for each analysis. Software was obtained from http://www.niehs.nih.gov/research/ resources/software/ biostatistics/rxc/index.cfm.

Stouffer's Z-method was used to determine statistical significance of gene conversion (GC) coverage as determined by PolyMotifFinder. Stouffer's Z is defined by:

$$Z_s = \frac{\sum_{i=1}^{n} Z_i}{\sqrt{n}}$$

where Z_s is Stouffer's Z value, n is the number of Z scores in the analysis, i is the i-th Z score out of a total of n Z scores. Unaltered somatically-mutated sequences were used to identify GC coverage and were compared to each sequence's respective population of GC coverage of altered somatically-mutated sequences generated by RandomCheck. Altered sequences had the identity of mutations changed to any nucleotide that was not germline but retained the mutations in their respective positions for each sequence as described above. GC coverage populations of altered sequences were determined for 100-1,000 iterations of mutation changes. The mean and standard deviation of each altered GC-coverage population was used to compute a Z-score for the respective unaltered GC-coverage for that sequence. Stouffer's Z method of meta-analysis was used to determine if the GC-coverage of the sequence sets is significantly different from that of modeled somatic hypermutation.

Stouffer's Z was combined between IgHVs of murine or human datasets via a weighted Stouffer's Z trend, defined as

$$Z_T = \frac{\sum_{S=1}^{n} w_S Z_S}{\sqrt{\sum_{S=1}^{n} w_S^2}}$$

where Z_T is Stouffer's Z trend, n is the number of Stouffer's Z scores used in the analysis, w is the weight of each Stouffer's Z score (defined as the number of Z scores for the corresponding Stouffer's Z divided by the total number of Z scores across all Stouffer's Z scores used in the analysis), S is the S-th Stouffer's Z score out of the maximum of n [49]. Calculation of Stouffer's Z trend allows us to test whether there is an effect across different IgHV datasets and is weighted according to the number of Z-scores generated for each IgHV.

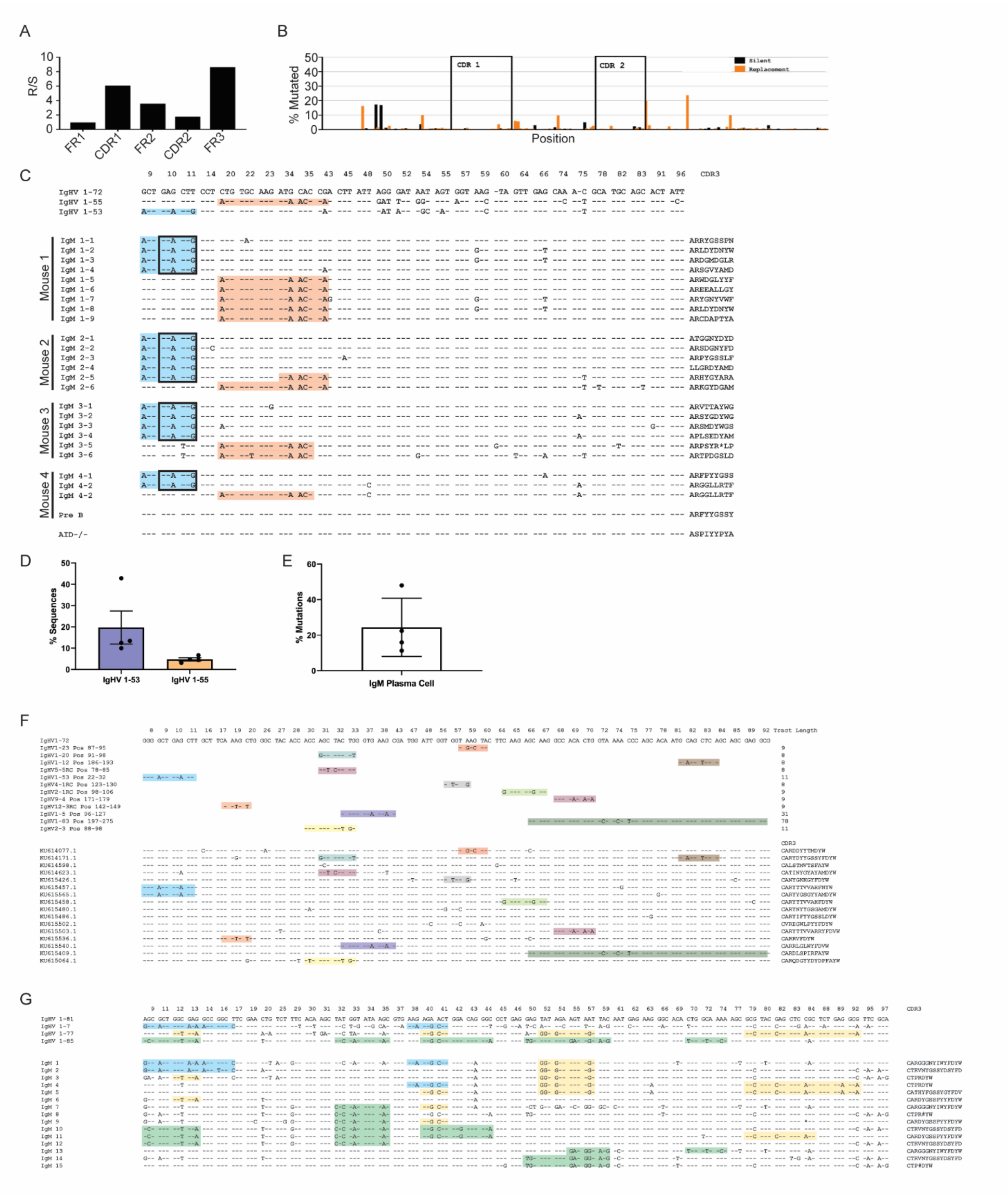


Fig. 1: Somatic hypermutation motifs in IgM plasma cells are shared between individual clones in individual mice. (**A**) Overview of somatic hypermutation in IgM plasma cells. Shown

are percent replacement and silent mutations per position along the IgHV 1-72 sequence. CDR 1 and 2 are highlighted with a box. Framework 1, 2, and 3 are intervening regions. (**B**) Replacement-to-silent mutation ratios for IgHV domains for IgM plasma cell sequences. (**C**) Nucleotide alignment of somatically-mutated IgM plasma cell sequences to germline IgHV 1-72, as well as putative donor germlines IgHV 1-55 and 1-53. Sequences are grouped according source mouse. Sequences shown in Fig. 1C are representative of four individual mice. CDR3 sequences are shown as amino acids. Only codons that differ in the alignment are shown. A representative pre-B cell sequence is shown as a sequencing control. Gene conversion tracts from IgHV 1-53 are colored blue, and IgHV 1-55 are colored orange. Boxed codons represent silent mutations. (**D**) Percent of IgM plasma cell sequences in Figure 1C that possess either the IgHV 1-53 gene conversion tract (blue) or the IgHV 1-55 gene conversion tract (orange). (**E**) Shown are the percent of total mutations in the IgM plasma cell data set attributed to gene conversion with the IgHV 1-55 and IgHV 1-53 tracts. (**F**) Nucleotide alignment of somatically-mutated IgHV 1-72 sequences from Tas et al (2016). Sequences are presented as in **C**. Coordinates for donor IgHV fragments are shown on the left. RC denotes the noncoding strand. Tract length in bp is shown on the right for each tract. (**G**) Shown are somatically-mutated IgHV 1-81 sequences from germinal center B cells isolated from Peyer's patches of C57BL/6 mice. Data is presented as in **C**.

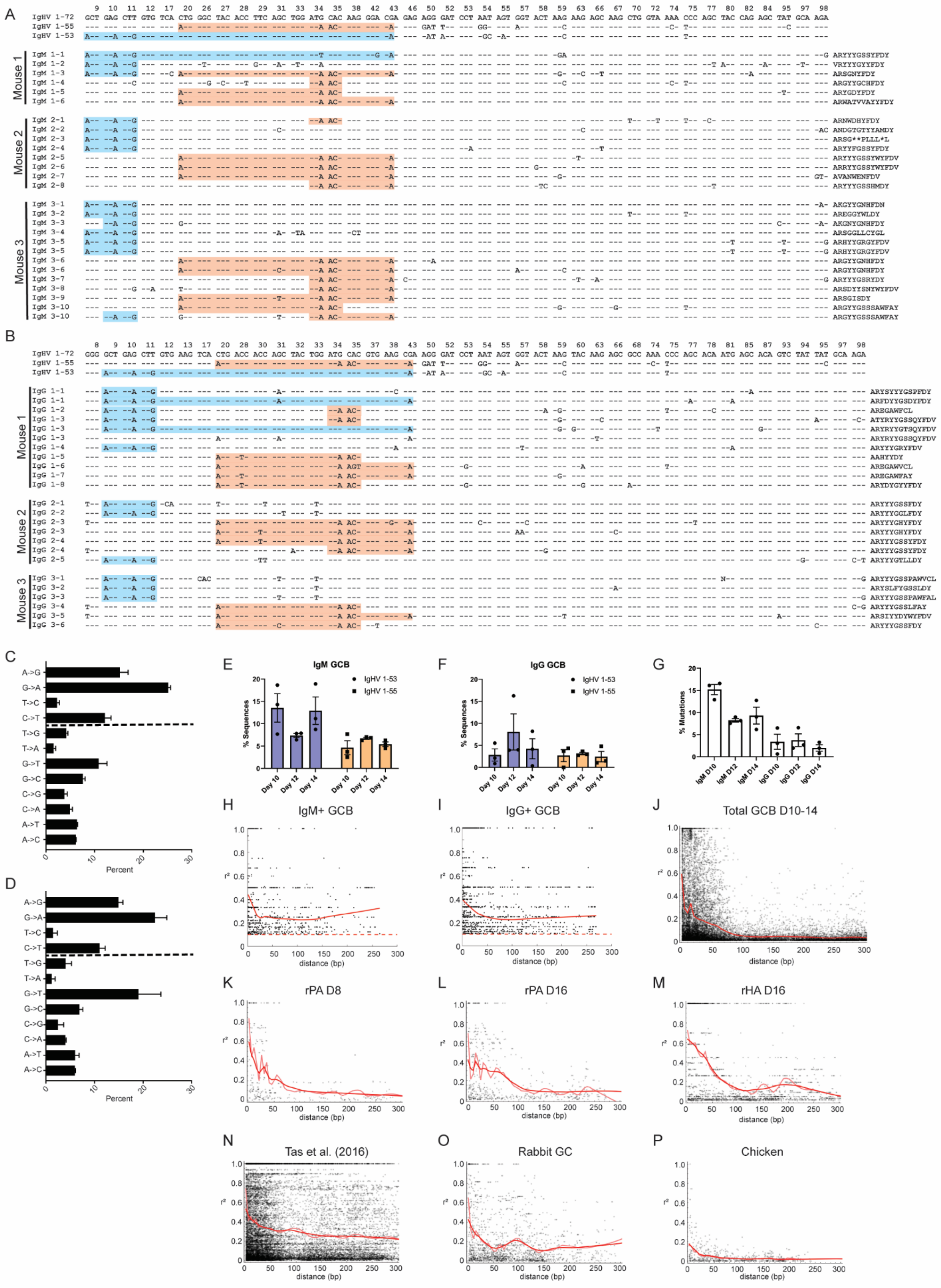
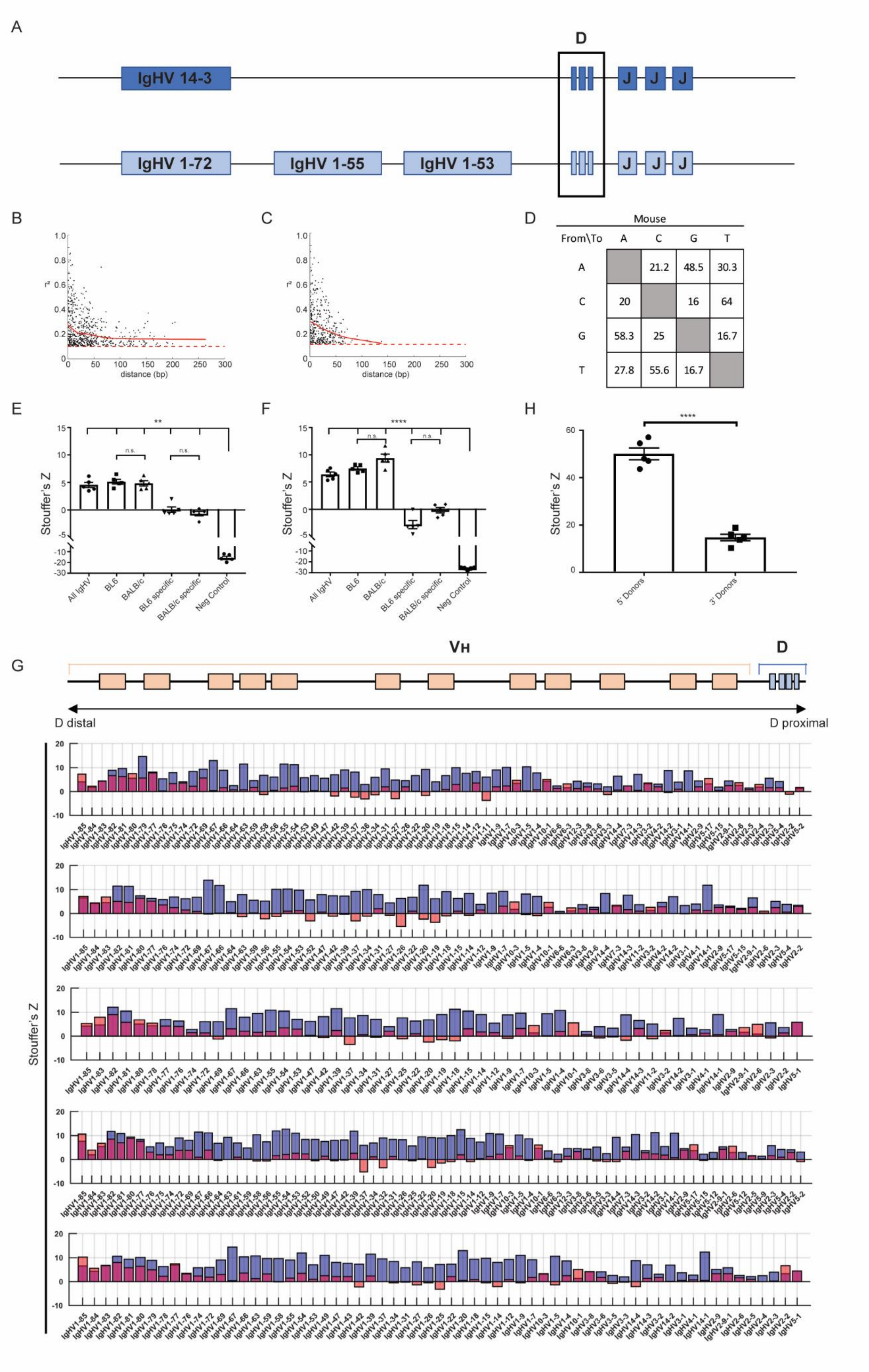

A
IgHV 1-72
IgHV 1-55
IgHV 1-53
Mouse 1
Mouse 2
Mouse 3
B
C
A->G
G->A
T->C
C->T
T->G
T->A
G->T
G->C
C->G
C->A
A->T
A->C
Percent
D
E
IgM GCB
IgHV 1-53
IgHV 1-55
% Sequences
Day 10
Day 12
Day 14
F
IgG GCB
G
% Mutations
IgM D10
IgM D12
IgM D14
IgG D10
IgG D12
IgG D14
H
IgM+ GCB
I
IgG+ GCB
J
Total GCB D10-14
K
rPA D8
L
rPA D16
M
rHA D16
N
Tas et al. (2016)
O
Rabbit GC
P
Chicken
r²
distance (bp)

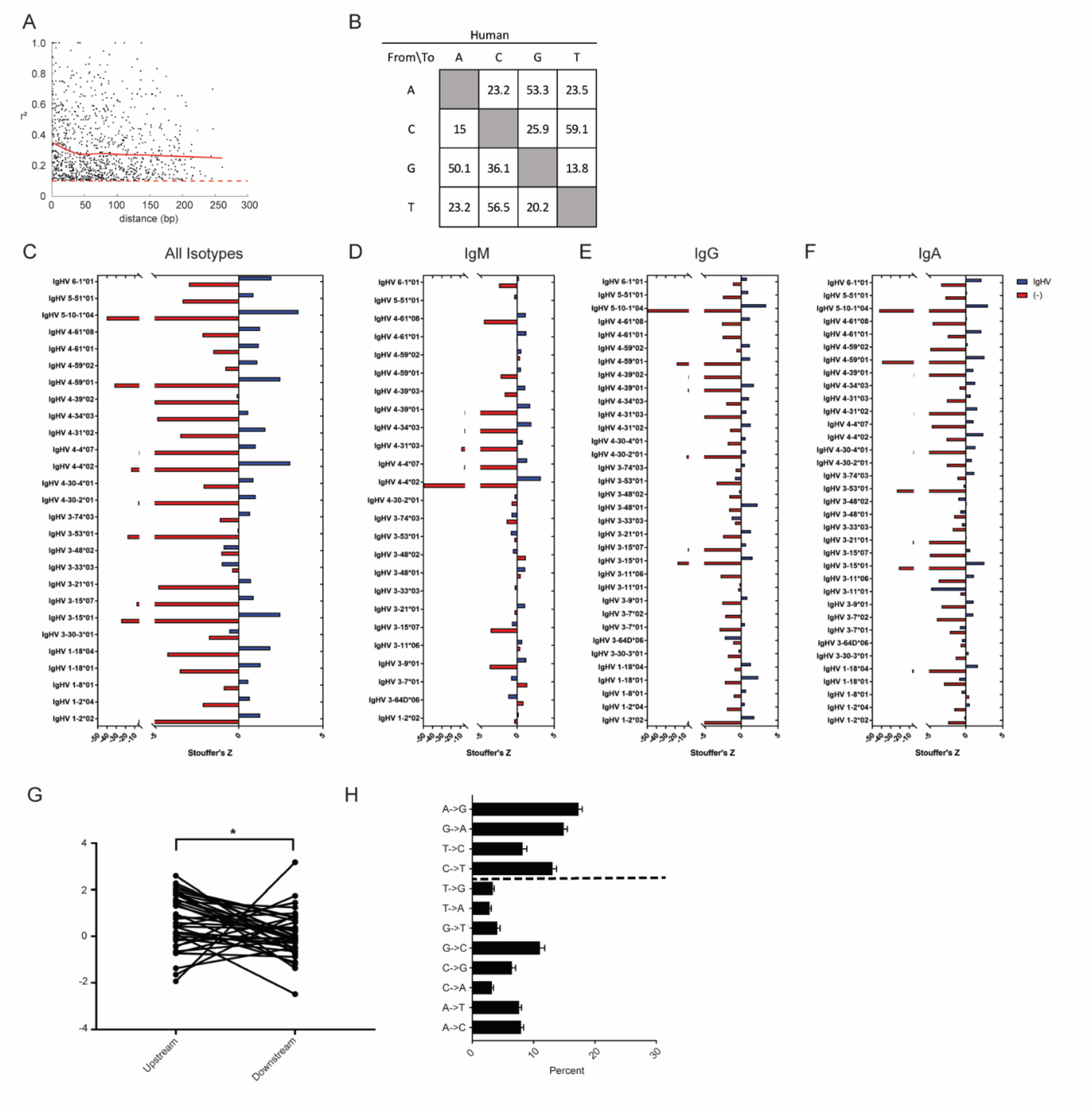
Fig. 2: IgM$^+$ and IgG$^+$ germinal center B cells exhibit gene conversion tracts during the germinal center reaction. (**A**) Nucleotide alignments of IgM$^+$ germinal center B cells. Data is depicted as in Figure 1C. Sequences are named such that the first number corresponds to the source animal and the second corresponds to unique clones. (**B**) Nucleotide alignments of IgG$^+$ germinal center B cells. Data is depicted as in (**A**). (**C-D**) Percent transition and transversion mutations in IgM (**C**) and IgG (**D**) IgHV 1-72 sequences from day 12 germinal center B cells. Transition mutations are shown above the dashed line, whereas transversions are shown below. (**E**) Shown are the percent of IgHV 1-72 IgM germinal center B sequences in Figure 2A that possess either the IgHV 1-53 tract or IgHV 1-55 tract. (**F**) Shown are the percentage of IgHV 1-72 IgG germinal center B cell sequences in Figure 2B that possess gene conversion tracts as in (**E**). (**G**) Shown are the percent of IgHV 1-72 mutations attributable to gene conversion. (**H-P**) Plot of linkage disequilibrium between all pairs of mutations per sequence per IgHV gene. Data is shown as a function of genetic distance between mutations and calculated squared correlation coefficient (r^2) of haplotype pairs. The red line represents a LOESS linear regression of the data points. Shown are linkage disequilibrium plots of IgM+ (**H**) and IgG+ (**I**) germinal center B cells with IgHV 1-72 rearrangement; multiple IgHV rearrangements from germinal center B cells at days 10, 12, 14 post-immunization with NPCGG (**J**); day 8 rPA specific germinal center B cells as reported in Kuraoka et al. (2016)(**K**); day 16 rPA specific germinal center B cells (Kuraoka et al., 2016) (**L**); day 16 rHA specific germinal center B cells (Kuraoka et al., 2016) (**M**); Sanger sequenced antibody sequences from multiple IgHV rearrangements as reported by Tas et al. (2016) (n=2150) (**N**);rabbit somatically-mutated IgHV genes (Schiaffella et al., 1999; Sehgal et al., 2002; Winstead et al., 1999) (**O**); and somatically-mutated chicken IgLV segments (Arakawa et al., 2002b; Mansikka et al., 1990) (**P**). For panel H and I, r^2 values are not shown if less than 0.1.



D

	Mouse			
From\To	A	C	G	T
A		21.2	48.5	30.3
C	20		16	64
G	58.3	25		16.7
T	27.8	55.6	16.7	

Fig. 3: Germinal center B cells from CB6F1/J mice show templated mutagenesis primarily occurs in cis. (**A**) Graphic depicting the haplotypes of CB6F1/J mice. Relevant VH exons are shown (**B**) Shown are Stouffer's Z values for PolyMotifFinder and RandomCheck comparisons of somatically-mutated IgHV 1-72 sequences obtained from day 12 germinal centers to different reference sequence sets. All IgHV is the set of all IgHV sequences regardless of strain. BL6 and BALBc refer to IgHV sequences that are specific to each strain, respectively. BL6-specific and BALBc-specific refer to 8-mer motifs that are only present in either BL6 or BALBc mice, respectively. Negative control refers to all 8-mer motifs that are not present in the IgHV repertoire of either strain. (**C**) Data shown as in (**B**) but for somatically-mutated IgHV 14-3 sequences. (**D**) Shown are base pair substitution matrices used for RandomCheck analysis. Tables were obtained from Maul et al. (2016). The table were then transformed from percent of total observed mutations, as reported, to the percent of observed mutations with a given germline nucleotide, such that each row tallies to 100 percent and indicates the probability of a given base to mutate into another. (**E-F**) Shown are representative r^2 plots of somatically-mutated IgHV 1-72 sequences (**E**) and IgHV 14-3 (**F**). Data is depicted as in Fig. 2. (**G**) A visual schematic of the C57BL/6 IgH locus is shown below which are the Stouffer's Z results of C57BL/6 IgHV rearrangements, grouped by individual CB6F1/J mouse. Each somatically-mutated IgHV is compared against preserved IgHV genes located 5' from the rearranged VDJ or lost IgHV genes 3' to the VH segment that underwent rearrangement. For each IgHV gene, Stouffer's Z against the 5' donors (blue) is overlaid with that of the 3' donors (red). IgHV genes are depicted in the order in which they occur along the IgH locus, with the most D_H-proximal on the right, and the most D_H-distal on the left. (**H**) Stouffer's Z trend is reported for 5' and 3' donors from each mouse. ** p<0.01, ****p<0.0001, n.s. not significant.



From\To	A	C	G	T
A		23.2	53.3	23.5
C	15		25.9	59.1
G	50.1	36.1		13.8
T	23.2	56.5	20.2	

Fig. 4: Human plasmablast sequences demonstrate templated mutagenesis and preferential use of 5' donors. (**A**) Shown is a r^2 plot of somatically-mutated IgHV 1-2 sequences. (**B**) Base pair substitution matrices used for RandomCheck analysis. Tables were obtained from Longo et al. (2009). Data is presented as in Figure 3G. (**C-F**) Shown are Stouffer's Z scores following PolyMotifFinder/ RandomCheck analysis of somatically-mutated human IgHV genes against

either the human IgHV repertoire (blue) or the 8-mer motifs not present in the IgHV repertoire (red) for IgHV genes shared between all isotypes (**C**) or those present of the IgM (**D**), IgG (**E**), or IgA (**F**) isotype. (**G**) Paired dot plot of Stouffer's Z for 5' (upstream) or 3' (downstream) donors for somatically-mutated IgHV sequences. (**H**) Percent transitions and transversions for somatically-mutated IgHV sequences. Data is shown as in Figure 2. *p<0.05

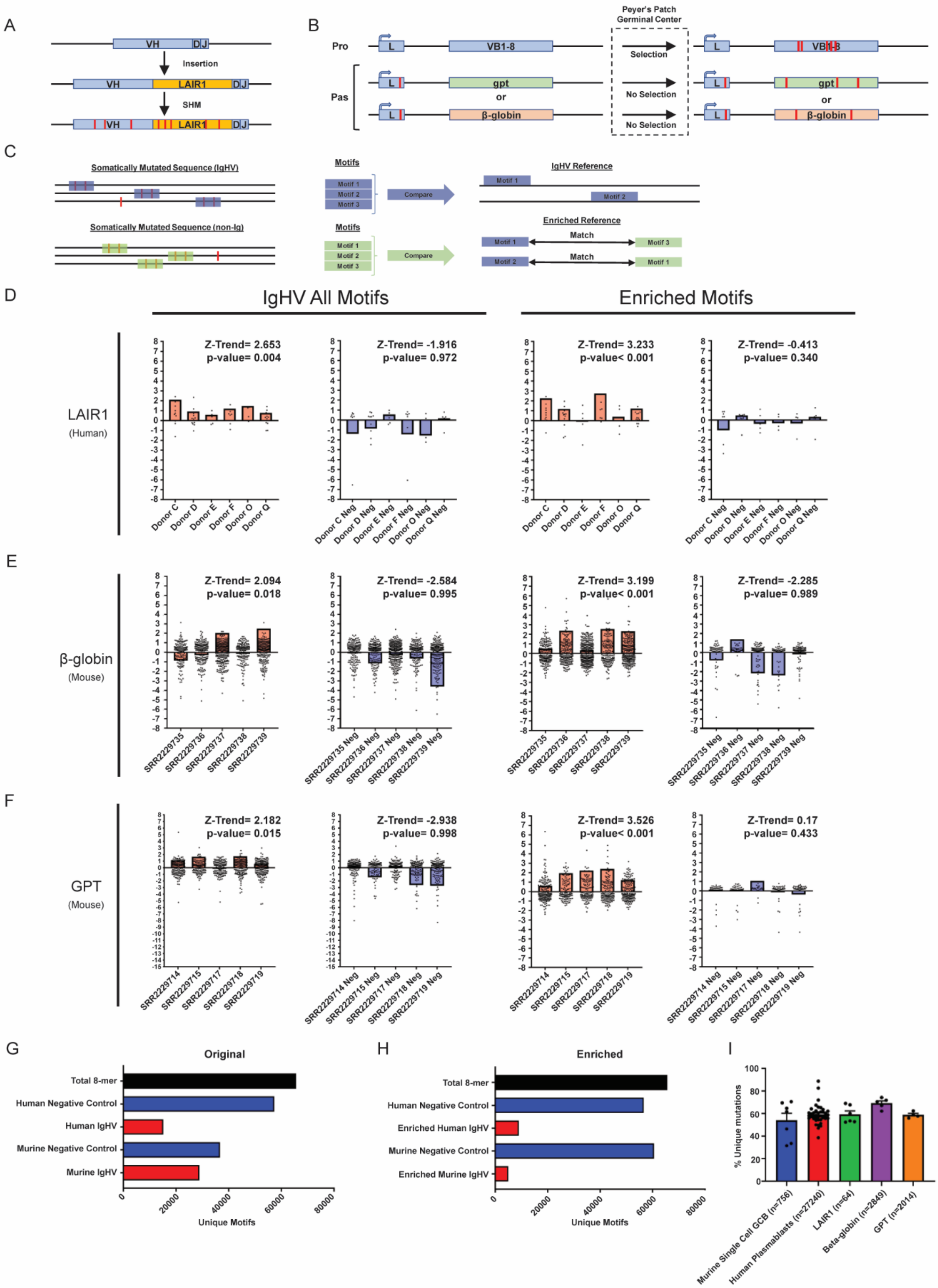

A
VH
DJ
Insertion
VH
LAIR1
DJ
SHM
VH
LAIR1
DJ
B
Peyer's Patch Germinal Center
Pro
L
VB1-8
Selection
L
VB1-8
Pas
L
gpt
or
L
β-globin
No Selection
No Selection
L
gpt
or
L
β-globin
C
Somatically Mutated Sequence (IgHV)
Somatically Mutated Sequence (non-Ig)
Motifs
Motif 1
Motif 2
Motif 3
Compare
Motifs
Motif 1
Motif 2
Motif 3
Compare
IgHV Reference
Motif 1
Motif 2
Enriched Reference
Match
Match
Motif 1
Motif 3
Motif 2
Motif 1
D
IgHV All Motifs
Enriched Motifs
LAIR1
(Human)
Z-Trend= 2.653
p-value= 0.004
Z-Trend= -1.916
p-value= 0.972
Z-Trend= 3.233
p-value< 0.001
Z-Trend= -0.413
p-value= 0.340
Donor C
Donor D
Donor E
Donor F
Donor O
Donor Q
Donor C Neg
Donor D Neg
Donor E Neg
Donor F Neg
Donor O Neg
Donor Q Neg
E
β-globin
(Mouse)
Z-Trend= 2.094
p-value= 0.018
Z-Trend= -2.584
p-value= 0.995
Z-Trend= 3.199
p-value< 0.001
Z-Trend= -2.285
p-value= 0.989
SRR2229735
SRR2229736
SRR2229737
SRR2229738
SRR2229739
SRR2229735 Neg
SRR2229736 Neg
SRR2229737 Neg
SRR2229738 Neg
SRR2229739 Neg
F
GPT
(Mouse)
Z-Trend= 2.182
p-value= 0.015
Z-Trend= -2.938
p-value= 0.998
Z-Trend= 3.526
p-value< 0.001
Z-Trend= 0.17
p-value= 0.433
SRR2229714
SRR2229715
SRR2229717
SRR2229718
SRR2229719
SRR2229714 Neg
SRR2229715 Neg
SRR2229717 Neg
SRR2229718 Neg
SRR2229719 Neg
G
Original
Total 8-mer
Human Negative Control
Human IgHV
Murine Negative Control
Murine IgHV
0
20000
40000
60000
80000
Unique Motifs
H
Enriched
Total 8-mer
Human Negative Control
Enriched Human IgHV
Murine Negative Control
Enriched Murine IgHV
0
20000
40000
60000
80000
Unique Motifs
I
% Unique mutations
100
80
60
40
20
0
Murine Single Cell GCB (n=756)
Human Plasmablasts (n=27240)
LAIR1 (n=64)
Beta-globin (n=2849)
GPT (n=2014)

Fig. 5: Non-immunoglobulin sequences exhibit templated mutagenesis and preferentially utilize a limited number of IgHV-specific motifs as donors. (**A**) Schematic depicting the LAIR1 antibodies described in Tan et al. Somatic mutations are depicted as red bars along the length of the sequence. Only the somatically-mutated LAIR1 segments were used in subsequent analyses (**B**) Schematic depicting the passenger allele transgene system as described in Yeap et al. Leader sequences (L) were intact at the productive allele (pro) but were mutated at the passenger allele (pas) to terminate translation. Only unselected, somatically-mutated gpt or β-globin sequences were used in subsequent analyses. (**C**) Schematic depicting the strategy for enriching motifs used in panels D-F. Somatically-mutated IgHV genes were matched through PolyMotifFinder to IgHV references for each species. In mice, somatically-mutated IgHV 1-72 sequences from CB6F1/J mice were used. In humans, somatic mutations from all IgHVs were used. Motifs produced from two or more mutations within 8bp that matched the IgHV reference were used as a reference for somatically-mutated non-immunoglobulin sequences to be matched to via PolyMotifFinder/RandomCheck. (**D-F**) Non-immunoglobulin sequence sets were compared via PolyMotifFinder/RandomCheck to either the IgHV repertoire (IgHV all motifs) or to an enriched set of motifs from the IgHV repertoire that were found to be donors somatically-mutated IgHV genes (red). In both analyses, each sequence set is compared to the corresponding number of motifs not in the IgHV set or the enriched set, respectively (blue). Stouffer's Z score is shown as a bar for each analyzed data set and dots represent individual Z scores. For each analysis, Stouffer's Z trend is shown along with the corresponding p-value. Sequence sets shown are LAIR1 (**D**), β-Globin (**E**), and GPT (**F**). (**G**) Shown are the number of motifs used for each analysis in comparison to the total number of unique 8-mers. (**H**) Shown are the number of motifs used in enriched analyses in comparison to the total number of unique 8-mers. (**I**) Shown are the percent of unique mutations

that fulfill two conditions: (1) proximity to another mutation within 8bp, and (2) there is a corresponding template present in the IgHV repertoire. GCB data was obtained from the Victora laboratory (Tas et al. *Science* 2016), LAIR1 data was obtained from the Lanzavecchia laboratory (Tan et al. *Nature* 2016; Pieper et al. *Nature* 2017), Beta-globin and GPT data was obtained from the Alt laboratory (Yeap et al. *Cell* 2015). Error bars denote mean ± SEM.

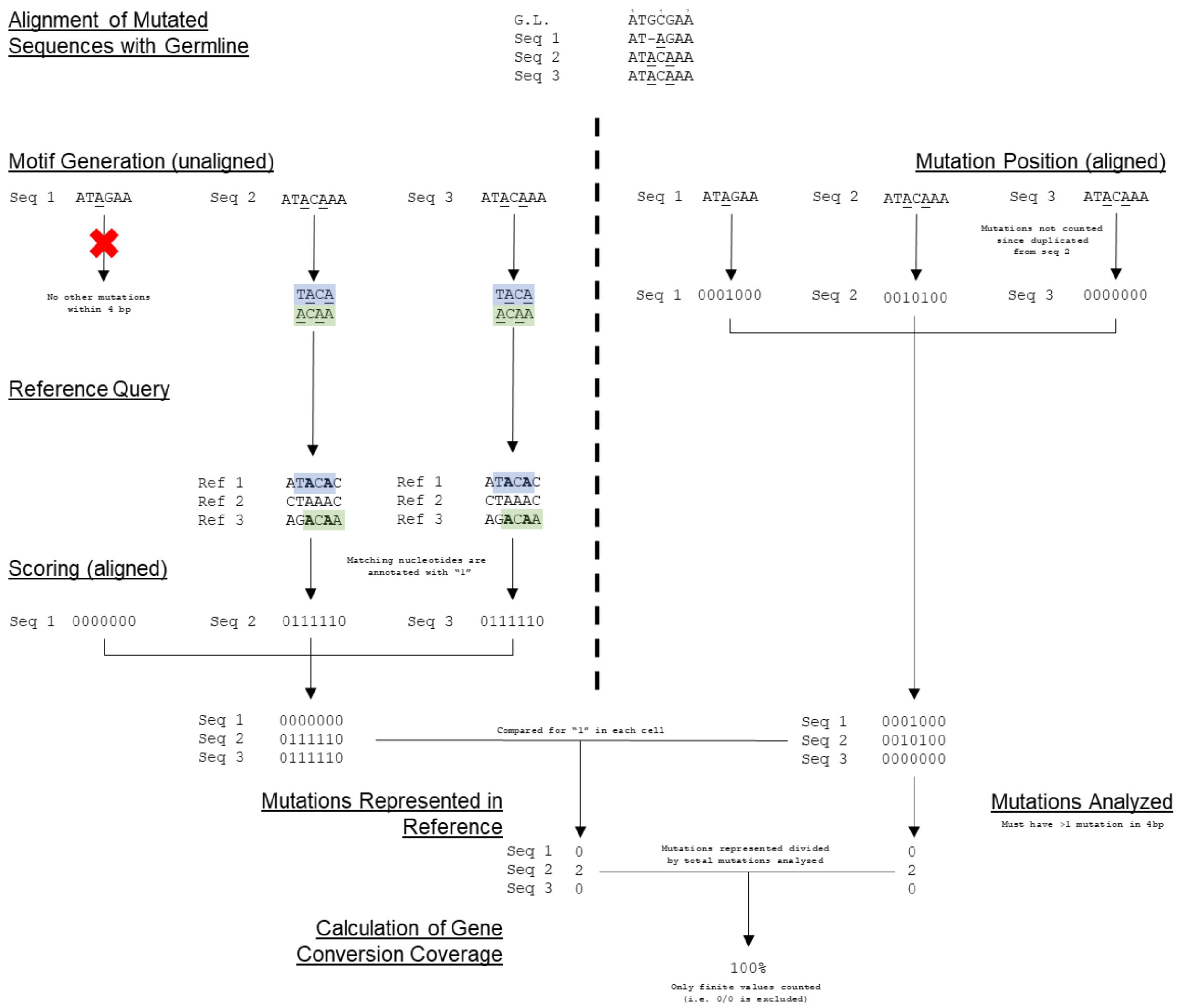

Alignment of Mutated Sequences with Germline
G.L. ATGCGAA
Seq 1 AT-AGAA
Seq 2 ATACAAA
Seq 3 ATACAAA
Motif Generation (unaligned)
Seq 1 ATAGAA
Seq 2 ATACAAA
Seq 3 ATACAAA
No other mutations within 4 bp
TACA
ACAA
TACA
ACAA
Reference Query
Ref 1 ATACAC
Ref 2 CTAAAC
Ref 3 AGACAA
Ref 1 ATACAC
Ref 2 CTAAAC
Ref 3 AGACAA
Scoring (aligned)
Matching nucleotides are annotated with "1"
Seq 1 0000000
Seq 2 0111110
Seq 3 0111110
Mutation Position (aligned)
Seq 1 ATAGAA
Seq 2 ATACAAA
Seq 3 ATACAAA
Mutations not counted since duplicated from seq 2
Seq 1 0001000
Seq 2 0010100
Seq 3 0000000
Seq 1 0000000
Seq 2 0111110
Seq 3 0111110
Compared for "1" in each cell
Seq 1 0001000
Seq 2 0010100
Seq 3 0000000
Mutations Represented in Reference
Mutations Analyzed
Must have >1 mutation in 4bp
Seq 1 0
Seq 2 2
Seq 3 0
Mutations represented divided by total mutations analyzed
0
2
0
Calculation of Gene Conversion Coverage
100%
Only finite values counted (i.e. 0/0 is excluded)

Fig. S1: Schematic of sequence analysis by PolyMotifFinder. Depicted are three somatically-mutated sequences aligned to a reference gene. Initially, mutation positions are defined by comparing the identity of nucleotides within the sequence to the reference sequence. Once positions of mutations are found, PolyMotifFinder creates a numeric array whose length is equal to the length of sequences analyzed and height is equal to the number of sequences analyzed. This array is filled with either "0" to denote a position in which the sequence has the germline nucleotide at a given position, or "1" should the sequence differ from germline at that position. Importantly, if a pair of mutations matches the position and identity of another pair of mutations already marked in the array, these mutations will be marked with "0", such that identical pairs of mutations are

excluded from analysis. Next, PolyMotifFinder will generate motifs of k-mer length. Here k=4, whereas in our analyses k=8. These motifs must contain two or more mutations over their length. The generated motifs are then compared to a reference set of sequences for matches. If a match is found, another numeric array is annotated with "1" over the length of the motif that matched a reference, otherwise the array is annotated with "0". Each row of the mutation position array is compared to the respective row of the motif matched array. Corresponding cells are compared for matches in which both cells contain "1", denoting a mutation that was part of a motif that matched the reference sequence. These mutation matches are tallied and divided by the number of mutations within that sequence to generate a gene conversion (GC) coverage value.

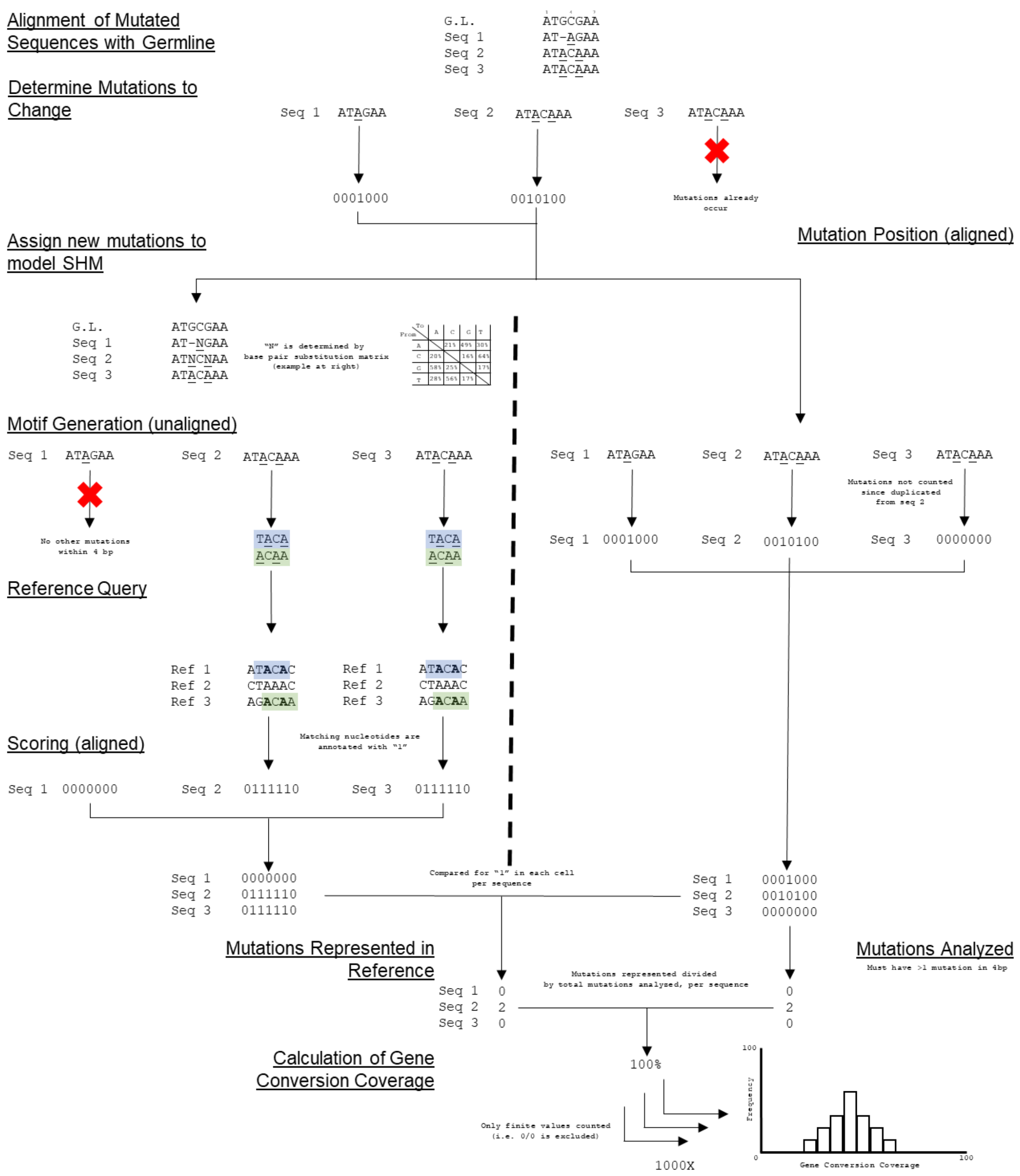


Fig. S2: Schematic of sequence analysis via RandomCheck. Depicted are three somatically-mutated sequences aligned to a reference gene. As in PolyMotifFinder, the positions of mutations are generated in a numeric array with duplicate mutation pairs being removed from the final array.

Next, RandomCheck simulates the effect of canonical somatic hypermutation by changing mutations to any of the three other non-germline nucleotides with the probability of any given change being determined by the base pair substitution profiles reported in Maul et al. (2016) and Longo et al. (2009), for murine and human sequences, respectively. This is followed by the generation of motifs, and matching to references, as done by PolyMotifFinder. The motif matched array is then compared to the mutation position array by row for cells that both contain “1” indicating a mutation that also matched a reference sequence. The GC coverage is then determined for this sequence. This process is then iterated 100 to 1000 times to generate a background population for each sequence based on the activity of canonical somatic hypermutation. The results from PolyMotifFinder for that respective sequence is then compared to its population to generate a Z-score. Application of Stouffer’s method to the set of Z-scores for each sequence set generates Stouffer’s Z value.

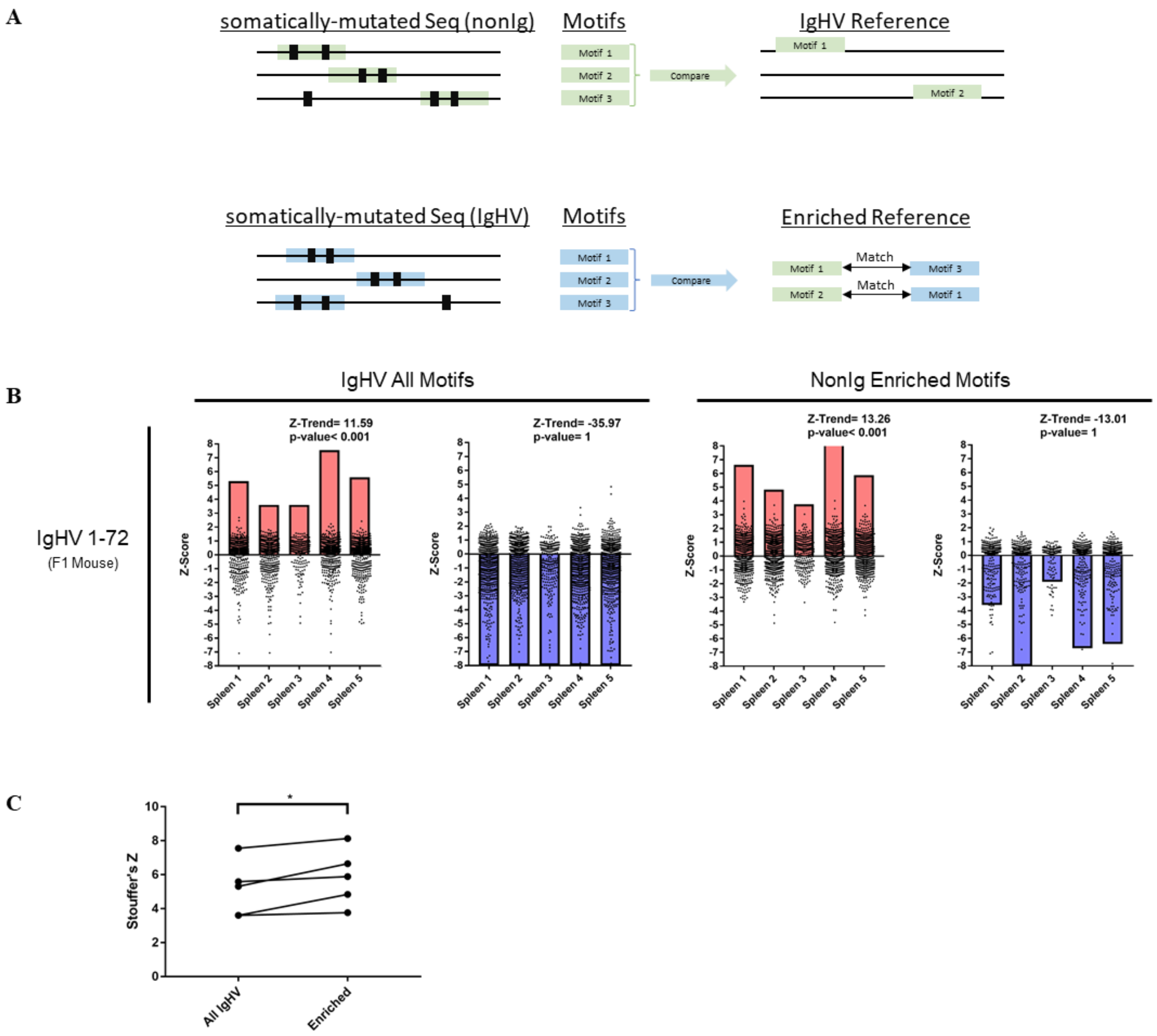


Fig. S3: IgHV genes from F1 germinal centers are enriched for motifs occurring in the IgHV repertoire and the subset of the repertoire that somatically-mutated GPT and β-globin match to. (**A**) Schematic depicting the strategy for enriching motifs used in (**B**). somatically-mutated GPT and β-globin were matched to the murine IgHV repertoire via PolyMotifFinder. Motifs that matched the repertoire were then used as a reference for somatically-mutated IgHV 1-72 sequences isolated from the day 12 germinal center in CB6F1/J mice. (**B**) somatically-mutated IgHV 1-72 sequences from CB6F1/J mice were compared via PolyMotifFinder/RandomCheck to either the IgHV repertoire or the enriched set of motifs defined in (**A**). Data is presented as in Figure 5B-D.

(**C**) Scatter plot depicts the effect of enrichment on Stouffer’s Z score shown in (**B**). *$p<0.05$, Paired t-test.

Works Cited

1 Cooper, M. D. & Herrin, B. R. How did our complex immune system evolve? *Nat Rev Immunol* **10**, 2-3, doi:10.1038/nri2686 (2010).

2 Boehm, T. *et al.* VLR-based adaptive immunity. *Annu Rev Immunol* **30**, 203-220, doi:10.1146/annurev-immunol-020711-075038 (2012).

3 Becker, R. S. & Knight, K. L. Somatic diversification of immunoglobulin heavy chain VDJ genes: evidence for somatic gene conversion in rabbits. *Cell* **63**, 987-997 (1990).

4 Meyer, A., Parng, C. L., Hansal, S. A., Osborne, B. A. & Goldsby, R. A. Immunoglobulin gene diversification in cattle. *Int Rev Immunol* **15**, 165-183 (1997).

5 Butler, J. E. Immunoglobulin diversity, B-cell and antibody repertoire development in large farm animals. *Rev Sci Tech* **17**, 43-70 (1998).

6 Di Noia, J. M. & Neuberger, M. S. Molecular mechanisms of antibody somatic hypermutation. *Annu Rev Biochem* **76**, 1-22, doi:10.1146/annurev.biochem.76.061705.090740 (2007).

7 Bohannon, C. *et al.* Long-lived antigen-induced IgM plasma cells demonstrate somatic mutations and contribute to long-term protection. *Nat Commun* **7**, 11826, doi:10.1038/ncomms11826 (2016).

8 Liu, Y. J. *et al.* Mechanism of antigen-driven selection in germinal centres. *Nature* **342**, 929-931, doi:10.1038/342929a0 (1989).

9 Shlomchik, M. J., Watts, P., Weigert, M. G. & Litwin, S. Clone: a Monte-Carlo computer simulation of B cell clonal expansion, somatic mutation, and antigen-driven selection. *Curr Top Microbiol Immunol* **229**, 173-197 (1998).

10 Arakawa, H., Hauschild, J. & Buerstedde, J. M. Requirement of the activation-induced deaminase (AID) gene for immunoglobulin gene conversion. *Science* **295**, 1301-1306, doi:10.1126/science.1067308 (2002).

11 Muramatsu, M. *et al.* Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* **102**, 553-563 (2000).

12 Jacob, J., Kelsoe, G., Rajewsky, K. & Weiss, U. Intraclonal generation of antibody mutants in germinal centres. *Nature* **354**, 389-392, doi:10.1038/354389a0 (1991).

13 Jacob, J., Przylepa, J., Miller, C. & Kelsoe, G. In situ studies of the primary immune response to (4-hydroxy-3-nitrophenyl)acetyl. III. The kinetics of V region mutation and selection in germinal center B cells. *J Exp Med* **178**, 1293-1307 (1993).

14 Toellner, K. M. *et al.* Low-level hypermutation in T cell-independent germinal centers compared with high mutation rates associated with T cell-dependent germinal centers. *J Exp Med* **195**, 383-389 (2002).

15 De Semir, D. & Aran, J. M. Misleading gene conversion frequencies due to a PCR artifact using small fragment homologous replacement. *Oligonucleotides* **13**, 261-269, doi:10.1089/154545703322460630 (2003).

16 Tas, J. M. *et al.* Visualizing antibody affinity maturation in germinal centers. *Science* **351**, 1048-1054, doi:10.1126/science.aad3439 (2016).

17 Kuraoka, M. *et al.* Complex Antigens Drive Permissive Clonal Selection in Germinal Centers. *Immunity* **44**, 542-552, doi:10.1016/j.immuni.2016.02.010 (2016).

18 Schiaffella, E., Sehgal, D., Anderson, A. O. & Mage, R. G. Gene conversion and hypermutation during diversification of VH sequences in developing splenic germinal centers of immunized rabbits. *J Immunol* **162**, 3984-3995 (1999).

19 Sehgal, D., Obiakor, H. & Mage, R. G. Distinct clonal Ig diversification patterns in young appendix compared to antigen-specific splenic clones. *J Immunol* **168**, 5424-5433 (2002).

20 Winstead, C. R., Zhai, S. K., Sethupathi, P. & Knight, K. L. Antigen-induced somatic diversification of rabbit IgH genes: gene conversion and point mutation. *J Immunol* **162**, 6602-6612 (1999).

21 Arakawa, H. *et al.* Effect of environmental antigens on the Ig diversification and the selection of productive V-J joints in the bursa. *J Immunol* **169**, 818-828 (2002).

22 Mansikka, A., Sandberg, M., Lassila, O. & Toivanen, P. Rearrangement of immunoglobulin light chain genes in the chicken occurs prior to colonization of the embryonic bursa of Fabricius. *Proc Natl Acad Sci U S A* **87**, 9416-9420 (1990).

23 Sehgal, D., Mage, R. G. & Schiaffella, E. VH mutant rabbits lacking the VH1a2 gene develop a2+ B cells in the appendix by gene conversion-like alteration of a rearranged VH4 gene. *J Immunol* **160**, 1246-1255 (1998).

24 Loh, D. Y., Bothwell, A. L., White-Scharf, M. E., Imanishi-Kari, T. & Baltimore, D. Molecular basis of a mouse strain-specific anti-hapten response. *Cell* **33**, 85-93 (1983).

25 Maul, R. W. *et al.* DNA polymerase iota functions in the generation of tandem mutations during somatic hypermutation of antibody genes. *J Exp Med* **213**, 1675-1683, doi:10.1084/jem.20151227 (2016).

26 Lefranc, M. P. *et al.* IMGT(R), the international ImMunoGeneTics information system(R) 25 years on. *Nucleic Acids Res* **43**, D413-422, doi:10.1093/nar/gku1056 (2015).

27 Retter, I. *et al.* Sequence and characterization of the Ig heavy chain constant and partial variable region of the mouse strain 129S1. *J Immunol* **179**, 2419-2427, doi:10.4049/jimmunol.179.4.2419 (2007).

28 DeKosky, B. J. *et al.* In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nat Med* **21**, 86-91, doi:10.1038/nm.3743 (2015).

29 Longo, N. S. *et al.* Analysis of somatic hypermutation in X-linked hyper-IgM syndrome shows specific deficiencies in mutational targeting. *Blood* **113**, 3706-3715, doi:10.1182/blood-2008-10-183632 (2009).

30 Tan, J. *et al.* A LAIR1 insertion generates broadly reactive antibodies against malaria variant antigens. *Nature* **529**, 105-109, doi:10.1038/nature16450 (2016).

31 Pieper, K. *et al.* Public antibodies to malaria antigens generated by two LAIR1 insertion modalities. *Nature* **548**, 597-601, doi:10.1038/nature23670 (2017).

32 Yeap, L. S. *et al.* Sequence-Intrinsic Mechanisms that Target AID Mutational Outcomes on Antibody Genes. *Cell* **163**, 1124-1137, doi:10.1016/j.cell.2015.10.042 (2015).

33 Reynaud, C. A., Anquez, V., Grimal, H. & Weill, J. C. A hyperconversion mechanism generates the chicken light chain preimmune repertoire. *Cell* **48**, 379-388 (1987).

34 Cumano, A. & Rajewsky, K. Clonal recruitment and somatic mutation in the generation of immunological memory to the hapten NP. *EMBO J* **5**, 2459-2468 (1986).

35 David, V., Folk, N. L. & Maizels, N. Germ line variable regions that match hypermutated sequences in genes encoding murine anti-hapten antibodies. *Genetics* **132**, 799-811 (1992).

36 D'Avirro, N., Truong, D., Xu, B. & Selsing, E. Sequence transfers between variable regions in a mouse antibody transgene can occur by gene conversion. *J Immunol* **175**, 8133-8137 (2005).

37 Xu, B. & Selsing, E. Analysis of sequence transfers resembling gene conversion in a mouse antibody transgene. *Science* **265**, 1590-1593 (1994).

38 Darlow, J. M. & Stott, D. I. Gene conversion in human rearranged immunoglobulin genes. *Immunogenetics* **58**, 511-522, doi:10.1007/s00251-006-0113-6 (2006).

39 Sanz, I. Multiple mechanisms participate in the generation of diversity of human H chain CDR3 regions. *J Immunol* **147**, 1720-1729 (1991).

40 Lavinder, J. J., Hoi, K. H., Reddy, S. T., Wine, Y. & Georgiou, G. Systematic characterization and comparative analysis of the rabbit immunoglobulin repertoire. *PLoS One* **9**, e101322, doi:10.1371/journal.pone.0101322 (2014).

41 Jacob, J. & Kelsoe, G. In situ studies of the primary immune response to (4-hydroxy-3-nitrophenyl)acetyl. II. A common clonal origin for periarteriolar lymphoid sheath-associated foci and germinal centers. *J Exp Med* **176**, 679-687 (1992).

42 Bross, L., Wesoly, J., Buerstedde, J. M., Kanaar, R. & Jacobs, H. Somatic hypermutation does not require Rad54 and Rad54B-mediated homologous recombination. *Eur J Immunol* **33**, 352-357, doi:10.1002/immu.200310009 (2003).

43 Mazin, A. V., Mazina, O. M., Bugreev, D. V. & Rossi, M. J. Rad54, the motor of homologous recombination. *DNA Repair (Amst)* **9**, 286-302, doi:10.1016/j.dnarep.2009.12.006 (2010).

44 D'Avirro, N., Truong, D., Luong, M., Kanaar, R. & Selsing, E. Gene conversion-like sequence transfers between transgenic antibody V genes are independent of RAD54. *J Immunol* **169**, 3069-3075 (2002).

45 Schmuckli-Maurer, J. & Heyer, W. D. The Saccharomyces cerevisiae RAD54 gene is important but not essential for natural homothallic mating-type switching. *Mol Gen Genet* **260**, 551-558 (1999).

46 Maul, R. W. & Gearhart, P. J. Refining the Neuberger model: Uracil processing by activated B cells. *Eur J Immunol* **44**, 1913-1916, doi:10.1002/eji.201444813 (2014).

47 Tiller, T., Busse, C. E. & Wardemann, H. Cloning and expression of murine Ig genes from single B cells. *J Immunol Methods* **350**, 183-193, doi:10.1016/j.jim.2009.08.009 (2009).

48 Zaykin, D. V., Pudovkin, A. & Weir, B. S. Correlation-based inference for linkage disequilibrium with multiple alleles. *Genetics* **180**, 533-545, doi:10.1534/genetics.108.089409 (2008).

49 Zaykin, D. V. Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. *J Evol Biol* **24**, 1836-1841, doi:10.1111/j.1420-9101.2011.02297.x (2011).

Chapter 3

Somatic diversification of rearranged antibody gene segments by genome-wide templated mutagenesis

In review at *Cell, 2019*.

Gordon Dale performed all the experiments in this manuscript, except as follows:

> Daniel Wilkins wrote core components of TRACE
>
> Michael Rowley conducted HI-C data analysis
>
> Chistopher Tipton and Jennifer Hom sorted human cell populations, amplified IgH transcripts and sequenced amplicons

Gordon Dale and Joshy Jacob wrote the manuscript.

ABSTRACT

The ability of the humoral immune system to generate high affinity antibodies is critically dependent on the somatic hypermutation program and its ability to introduce mutations. In humans, somatic hypermutation is widely believed to solely be the result of untemplated point mutations. Here, we demonstrate detection of large-scale templated events in humans that are derived intrachromosomally and interchromosomally from IgHV genes/pseudogenes as well as exogenous genes. In addition, we find that exogenous sequences placed at the IgH locus, undergo templated mutagenesis and that homology is the major determinant for donor choice. Further, we find that donor tracts originate from areas in proximity with open chromatin, are transcriptionally active, and are in physical proximity with the IgH locus. These donors are inserted into the Ig genes proximal to overlapping AID hotspots in CDR1 and FWR3. These studies suggest an unrecognized role for templated mutagenesis during maturation of the humoral immune response.

RESULTS

We have previously demonstrated that small clusters of mutations (≥2 mutations over 8bp) in somatically-mutated sequences at the IgH locus are derived from templated events [1]. Our analyses had also demonstrated that the somatic mutations clusters present in the broadly-neutralizing anti-malarial LAIR1-containing antibodies reported by Pieper *et al.* [2] and Tan *et al.* [3] (Fig. 1A) are derived by templated events. These LAIR1-containing antibodies are atypical antibodies that result from a templated insertion event of a segment of the LAIR1 gene on chromosome 19 into the CDR3 of an antibody rearrangement on chromosome 14 and are subsequently mutated to gain broad anti-malarial binding capacity [3]. Interestingly, we observed that, while many clusters of mutations in somatically-mutated LAIR1 appeared to have templates corresponding to IgHV genes, there were multiple instances of heavily-mutated regions that did not have a corresponding IgHV template. We performed BLASTn [4] searches on these subsequences and found that these sequences had matches to distant genomic regions (Fig 1B). Local alignment of resultant matches with the somatically-mutated LAIR1 segment and the LAIR1 germline subsequence revealed that these matches account for several mutations found in the mutated subsequence and retain general homology to the LAIR1 segment (Figure 1B). This raised the possibility that these clustered somatic mutations may derive from distant genomic regions but did not answer whether such matches were significant. Indeed, it remained a possibility that somatic mutations themselves could produce spurious alignments given the size and scale of the human genome.

To address this problem, we generated a custom MatLab script called TRACE (Template Recognition via monte Carlo Experiments). This script was designed to determine if clusters of mutations of a given density are unlikely to have occurred by a mechanism other than templated mutagenesis. In brief, TRACE operates through a nested system of Monte Carlo simulations that identifies clustered mutations (defined as at least 8 mutations over 38bp), and subsequently iterates and analyzes outputs of the BLASTn algorithm [4] (Fig. S1). The script conducts simulations to determine (1) if the identity of the mutations within the cluster is important to produce a distant alignment, (2) if the local position of the mutations relative to one another is important to produce the distant alignment, and (3) if the resultant alignments are statistically different than if the germline sequence were randomly mutated. Only results that pass all three tests are considered for further analysis.

Application of the TRACE pipeline to the LAIR1 and IgHV sequences present in Pieper *et al.* [2] and Tan *et al.* [3] revealed significant donor templates scattered around the genome (Fig. 2A-B). TRACE analysis of the LAIR1 inserts present in five individual human donors revealed 15 putative templates that accounted for approximately 2-15% of total mutations present in the somatically-mutated LAIR1 segments (Fig. 2C). Similarly, analysis of the upstream IgHV segments from LAIR1-containing rearrangements revealed 33 putative donors that accounted for 3-15% of the total somatic mutations (Fig. 2C). On average, templates identified by TRACE accounted for two (IgHV) or four (LAIR1) mutations each (Fig. 2D). Donor templates detected by TRACE ranged between 21 and 38bp (Fig. 2E). Interestingly, donor templates clustered by chromosome with 22 out of 33 (66.7%) donors for the IgHV mutations originating intrachromosomally from chromosome 14 (Z=19.9, p<0.001) and seven out of 15 (46.7%) donors for the LAIR1 mutations originating from chromosome 19 (Z=13.2, p<0.001) (Fig. 2F). Further analysis revealed that donor tracts not only clustered at the chromosomal level, but also at the nucleotide level, with ~40% of all donor templates for LAIR1 derived from independent samples exhibiting significant clustering within 1kb of one another (Z=189.7, p<0.001) (Fig. 2G).These clustered regions of donors were strikingly different depending on whether the recipient sequence was the LAIR1 insert or the IgHV gene segment, suggesting that sequence homology was the main determinant of donor sequences (Fig. 2H). That these donor templates cluster within a narrow genomic region and account for mutations across multiple samples suggests that there are preferred sites that serve as donor sequences. Alignment of TRACE-identified donor templates to the somatically-mutated LAIR1 and IgHV gene segment of donor F reveals that identified donor tracts do explain clusters of mutations though they can retain some mismatches, suggesting that if such sequences are used as templates, they likely undergo further processing and or are subject to additional rounds of mutagenesis (Fig. 2I).

Given that the IgHV gene segment of the LAIR1-containing antibodies appeared to acquire templated tracts of mutations, we next investigated the somatically-mutated antibody repertoire of class-switched memory B cells from four human donors. For each human donor, we isolated CD19^{+}IgD^{-}CD27^{+} B cells from PBMCs and sequenced the heavy chain repertoire of these cells. We find that donor templates, identified by TRACE, occur across the genome in each of the four donors (Fig. 3A-D). Templates were primarily found intrachromosomally on chromosome 14 as well as interchromosomally on chromosomes 15 and 16 and were statistically enriched for donor tracts on these chromosomes as determined by permutational analysis (Z_s=85.488, p<0.001; Z_s =7.10, p<0.001; Z_s =15.17, p<0.001) (Fig. 3E). Interestingly, we also observed an absence of templates in each of the four donors from chromosome 9. As we observed in our analysis of LAIR1-containing antibodies, the percent of mutations accounted for by these templates ranged from 1-20%, depending on the rearranged gene segment (Fig. 3F). Between the same IgHV rearrangements isolated from different donors, there appears to be a varying contribution of TRACE-identified donor sequences to the total mutation load of the B cells expressing a gene segment. As in the TRACE analysis of LAIR1-containing antibodies, identified templated tracts were between 20-40bp and explained 2-8 mutations on average per tract (Fig. 3G-H). Further, we also observed significant clustering between TRACE-identified templates between donors, with 65% of all identified TRACE donor tracts within 1kb of one another (Z=669.43, p<0.001) (Fig. 3I; Fig. S2). Further analysis revealed that there were 28 sites in which each human donor had a TRACE-identified donor template (Fig. 3J). These sites were primarily IgHV pseudogenes located on chromosomes 14, 15, and 16 but we also observed donor templates from an intergenic region of chromosome 16 as well as the gene *TRPM2*. Between all human donors, 12 of 14 interchromosomal IgHV pseudogenes were found to contribute mutations. A total of 18 and 16 template donor regions were shared between 3 of the 4 and 2 of the 4 human donors, respectively. Similar to sites that were shared among all human donors, these sites were primarily IgHV pseudogenes.

As the presumptive mechanism for these templated mutations is gene conversion, and gene conversion donor choice is influenced by spatial proximity [5], chromatin accessibility [6], and transcription [7], we sought to elucidate whether these sites identified by TRACE are suggestive of those used for gene conversion. To do so, we first analyzed published Hi-C chromosome conformation capture data from human germinal center B cells and naïve B cells as reported by Bunting *et al.* [8]. Although the predominant source of TRACE-identified templates originates on chromosome 14, the redundancy of the IgH locus did not allow us to resolve intrachromosomal interactions. Instead, we analyzed interchromosomal contacts between the IgH locus and the rest of the genome. For each human donor, sites of TRACE-identified donor templates were analyzed as a function of distance from sites of interchromosomal interactions with the IgH locus as identified from the Hi-C data. We observed a consistent and significant pattern in which germinal center B cells were most enriched for TRACE-identified donor sites at the site of interchromosomal interactions over naïve B cells and random background (GCB vs NB p=2.175e-4, 1.17e-3, 9.65e-4, 1.23e-6; GCB vs random p=1.469e-5, 1.448e-5, 4.47e-6, 3.833e-5; NB vs random p= 1.653e-4, 9.113e-7, 4.163e-10, 3.883e-6) (Fig. 4A). This pattern also continued from the site of interaction to 1Mb away from sites of interchromosomal interaction. Given that interchromosomal IgHV pseudogenes were common donor sequences, we segregated TRACE identified donor sites into either those that were IgHV pseudogenes or those that were not. Hi-C analysis of these groups in relation to naïve and germinal center B cells revealed a significant enrichment of IgHV pseudogenes at sites of interaction as compared to naïve B cells (p=7.32e-9). Further, we analyzed whether the fraction of reads associated with Hi-C contacts in germinal center B cells was associated with distance from the TRACE donor sites. We find that for each human donor, 85-95% of reads within 10Mb of the TRACE donor sites are located within 2Mb of the TRACE donor sites suggesting that there is significant enrichment for IgH contacting sites near interchromosomal TRACE donor sequences (Figure S3). Altogether this highlight that both pseudogene and non-pseudogene TRACE-identified donor templates are physically close to the IgH locus once B cells enter the germinal center program.

Next, we investigated whether TRACE-identified templates were located within proximity of open chromatin. We assessed genome-wide chromatin accessibility by analyzing ATAC-seq data on subsets of B lymphocytes [9] and queried if TRACE donor templates were located in close proximity (within ≤1kb) of open chromatin peaks. For each human donor, we found that TRACE donor templates were significantly enriched within 1kb of open chromatin peaks as compared to a simulated null distribution of an equal number of randomly-placed donor template locations (701: Z=8.18, 702: Z=7.72, 730: Z=9.30, 752: Z=8.95) (Fig. 4C). Given that a large fraction of TRACE-identified donor sequences are IgHV gene segments, and that multiple rearrangements exist in a population of B cells resulting in open chromatin signal at multiple IgHV sites, this is a likely confounder in this analysis. Thus, we further analyzed whether interchromosomal donor templates were within 1kb of open chromatin peaks. We found that interchromosomal TRACE-identified donor templates were also significantly enriched within 1kb of open chromatin peaks (701: Z=5.33, 702: Z=3.03, 730: Z=6.53, 752: Z=4.99) (Fig. 4D). This suggests that donor templates used for mutagenesis at the IgH locus are associated with regions of the genome that are accessible.

We next queried whether TRACE donor templates are enriched for genes upregulated in the germinal center B cells, using RNA-sequencing data from Bunting *et al*.[8] If true, this would suggest that templates that are used for mutagenesis are transcriptionally active during the mutagenesis program. We found a total of 5650 genes were upregulated in germinal center B cells as compared to naïve and 2781 genes were downregulated. Between all four human donors, we found 220 non-IgHV TRACE-identified donor template containing genes. Of those, 163 genes were present in the Naïve B and Germinal Center B RNA-sequencing data. TRACE-identified donor genes that were not present in the data set from Bunting *et al*. included non-coding RNAs and T-cell receptor V regions which is expected, as the studies in Bunting *et al*. only examine polyadenylated transcripts. Of the 163 genes present, 66 were present in the upregulated fraction and 23 were contained in the downregulated fraction, and the remaining 74 were in genes whose expression did not change between naïve and germinal center B (Table S1). By permutational analysis, TRACE overlap with the upregulated genes are significantly enriched above background (Z=4.07, $p<0.001$) as compared to the downregulated genes (Z=0.45, p=0.33). Further, we find that regions of TRACE donor sequences that cluster between human donors overlap with regions that undergo transcription (Fig. S2). Together these suggests that donor templates preferentially use actively-transcribed genes.

Having observed multiple biological correlates of TRACE-identified donor sequences, we next sought to analyze the templated mutagenesis recipient sites. We hypothesized that recipient sites would be proximal to sites of double strand DNA breaks (DSB), as it was shown that a DSB is sufficient to induce gene conversion in a AID$^{-/-}$ DT40 cell line [10]. Furthermore, previous studies have shown that the palindromic AID hotspots WG$\underline{C}$W are enriched in the switch region and allow for the necessary DSB required for class switching [11,12]. We investigated whether such sequences exist within IgHV gene segments and mapped sites that were identified by TRACE as recipients of templated mutagenesis. We found that multiple WG$\underline{C}$W sites exist in IgHV gene segments and that, in general, TRACE events map preferentially to regions proximal to these WG$\underline{C}$W sites (IgHV 1-69: Z_s =-19.32, p<0.001; IgHV 3-15: Z_s =19.51, p=1; IgHV 5-51: Z_s =-88.71, p<0.001), although effects of selection may alter this, as in IgHV 3-15 (Fig. 4E-H, S4-5). Strikingly, we find that TRACE events are primarily clustered in CDR1 and FWR3 regions, across multiple IgHV rearrangements isolated from different human donors, suggesting that diversification of these regions is associated with templated mutagenesis and double strand breaks. To rule out any intrinsic effects of the IgHV sequence and germinal center selection biases, we analyzed the LAIR1 insert [2,3] (Fig. S5), which is a non-IgHV sequence at the IgH locus, as well as the murine passenger transgenes [13] from the Alt laboratory that are both free from selective pressure in addition to bearing no overt homology to the IgHV genes (Fig. S7). In these sequences, as well, we find a highly-significant association between TRACE events and the presence of the WG$\underline{C}$W motif (LAIR1: Z_s =-42.81, p<0.001; GPT: Z_s =-21.58, p<0.001; β-globin: Z_s =-39.59, $p<0.001$) (Fig. 4H). Together, this suggests a critical role for this motif for templated mutagenesis and rules out any selection and/or IgHV-specific effects.

Here, we demonstrate that TRACE is effective in identifying templated mutations and localizing them to distinct regions of the genome. Further, we demonstrate that a subset of TRACE-identified donor sites occur across individuals, suggesting preferential utilization of certain templates. Lastly, we show that TRACE analysis of mutation data is correlated with germinal center B cell biology and is predictive of upregulated genes, chromatin accessibility, and even chromosome organization in the nucleus. Taken together, these studies suggest that somatic mutations acquired during the germinal center reaction are in part templated and that such templates can originate from a variety of ectopic sites across the genome.

Although these studies do correlate with B cell biology, the exact mechanism responsible for templated mutagenesis is elusive. Homology and proximity appear to be critical mediators, although we have found locations throughout the genome that can serve as templates. Studies by Pieper *et al*. [2] suggest that B cells are largely unrestricted in accessing templates at distant locations despite preference being given to more proximal regions. Given that TRACE events occur around DSBs, it remains possible that RNA may facilitate the process of templating mutations as the emerging role of RNA in DSB repair becomes clearer [14-16]. Indeed, the predominant A/T mutator, Polη has recently been shown to possess reverse transcriptase activity in a cellular context [17,18]. Though many questions remain, this work highlights a surprising contribution of the larger genome to the generation of diversity in germinal center B cells. This process has significant implications on antibody maturation and further work should be directed at understanding both the mechanism and the functional contribution of templated mutations during an affinity-matured response.

METHODS

TRACE Script

TRACE (Template Recognition via monte Carlo Experiments) is a custom script written in Matlab (v.2018a) and uses nested, iterative BLASTn to identify donor templates for somatically-mutated sequences at the genome scale. FASTA files containing a germline reference and somatically-mutated sequences are parsed for user-defined mutation clusters (herein ≥8 mutations over 38bp). Subsequences containing the mutation cluster are split into 38bp windows that each contain ≥8 mutations. Each of these windows are passed into BLASTn (word size:11, max hsps: 1, maximum target sequences: 1) against either the human genome (GRCh38) or the mouse genome (GRCm38). The window with the greatest bitscore is stored and passed into two sequential Monte Carlo analyses, with entry into the second contingent on the results in the first. In the first Monte Carlo analysis, the effect of the mutation's identity in the window is assayed by randomizing the identity of the mutated bases, generating a window with different combinations of mutations. 1000 simulated windows are passed into BLASTn and their respective bitscores are used to build a population to which the original stored window is compared. If the Z-score of the original window is ≥ 1.645 then the window is passed into the second Monte Carlo analysis, wherein the effect of the location of mutations is assayed. Here, the number of mutations in the original window are randomly shuffled over the length of the window. As before 1000 simulated windows are generated and passed into BLASTn to generate a second population of bitscores. If the Z-score of the original window is ≥ 1.645 as compared to this second population, the original BLAST hit is stored as a TRACE hit. This process is iterated through all sequences until all mutation clusters have been analyzed.

Alongside the analysis of the input data set, a series of 10 modeled data sets are generated in which the original FASTA is analyzed for mutation clusters as above. The number of mutation clusters and the corresponding number of mutations per cluster per sequence is then randomized such that the locations of the mutation clusters are randomly placed along the length of the sequence. Each of these modeled sets undergoes the same core analysis above involving cluster identification, subsequent Monte Carlo analyses, and recording of TRACE hits that pass both Monte Carlo analyses.

Upon completion of analysis of the modeled data and the original data set, TRACE hits are passed through BLAST and the corresponding template is locally aligned. The number of mutations accounted for by the top BLAST hit is calculated and stored for each of the TRACE hits in the modeled and original data sets. Other data is also gathered at this time, including the percent identity between the TRACE hit and its corresponding template, the strand to which the hit localizes, the gene name (if any), and whether the identified template is an exon or intron.

Finally, TRACE hits from the original data is compared to that of the modeled by placing the data in bins composed of the combination of length, the number of mutations explained by the TRACE identified template, and the percent identity of the mutation cluster window to the identified template. The frequency of hits in each bin is counted for the original and the modeled data sets. Any bins unique to the original data is defined as true hits. For bins in both data sets, original data TRACE hits are only retained if the frequency of the original bin is greater than the frequency of the modeled bin plus two standard deviations of all the frequencies of bins in the modeled data set. Reported data from TRACE are cleaned such that overlapping TRACE hits that map back to the same template are removed.

Subjects

3 healthy subjects vaccinated with trivalent influenza vaccine (701, 702, 752), and 1 SLE patient experiencing an acute flare (730) were enrolled in this study at Emory University between 2013 and 2015. Healthy subjects received the influenza vaccine as part of routine medical care. SLE patient recruitment outside the annual influenza season and patient history were used to determine absence of recent immunization or likely natural exposure to influenza. PBMC were isolated on days 6-9 for vaccination subjects. All studies were approved by the Institutional Review Boards at Emory University School of Medicine. The SLE patient fulfilled ≥ 4 criteria of the modified ACR classification and was routinely evaluated by expert rheumatologists at the Emory Lupus Clinic. The SLE patient was classified as having a moderate-severe flare according to the SELENA-SLEDAI flare index and were on minimal immunosuppression at the time of flare (only hydroxychloroquine and/or <10 mg/day of prednisone or equivalent glucocorticoid).

Multi-color Flow Cytometry and Sorting

Mononuclear cells were isolated from peripheral blood using ficoll density gradient centrifugation and stained with the following anti-human antibody staining reagents: IgD-FITC, CD3-Pacific Orange, CD14-Pacific Orange, CD24-PE-A610 (Invitrogen, Camarillo, CA); CD19-APC-Cy7, CD38-Pacific Blue, CD23-PE-Cy7, CD21-PE-Cy5, CD27-PE (BD Pharmingen, San Diego, CA); CD138-APC (Miltenyi Biotec, Aubrun, CA). Approximately 30,000 cells were collected for each population using a BD FACS Aria II (BD biosciences, San Jose, CA) and sorted directly into RLT lysis buffer (Qiagen, Valencia, CA).

Next Generation Sequencing (NGS) of the IgH repertoire

Total cellular RNA was isolated from each sample using the RNeasy Micro kit by following the manufacturer's protocol (Qiagen, Valencia, CA). Approximately 2ng of RNA was subjected to reverse transcription using the iScript cDNA synthesis kit (BioRad, Hercules, CA). Aliquots of the resulting single-stranded cDNA products were mixed with 50nM of VH1-VH7 FR1 specific primers and 250nM Cα, Cμ, and Cγ specific primers preceded by the respective Illumina Nextera sequencing tag (sequences listed below) in a 25μl PCR reaction (using 4μl template cDNA) using Invitrogen's High-Fidelity Platinum PCR Supermix (Invitrogen, Camarillo, CA). Amplification was performed with a Bio-Rad C1000 Thermal Cycler (Bio-Rad, Hercules, CA) with the following conditions:

PCR1:
95°C for 5 min;
35 cycles of:
- 95°C for 30 sec,
- 55°C for 30 sec,
- 72°C for 30 sec;

72°C for 5 min.

A second PCR was used to add Nextera indices with the following conditions:

PCR2:
72°C for 3 minutes,
98°C for 30 sec,
5 cycles of:
- 98°C for 10 sec,
- 63°C for 30 sec, and
- 72°C for 3 min.

Ampure XP beads (Beckman Coulter Genomics, Danvers, MA) were used to purify the products and they were subsequently pooled and denatured. Single strand products were sequenced on a MiSeq (Illumina, San Diego, CA) using the 300bp x2 v3 kit. Primers for PCR1:

Forward:
VH1a: CAGGTKCAGCTGGTGCAG,

VH1b: SAGGTCCAGCTGGTACAG,
VH1c: CARATGCAGCTGGTGCAG,
VH2: CAGGTCACCTTGARGGAG,
VH3: GGTCCCTGAGACTCTCCTGT,
VH4: ACCCTGTCCCTCACCTGC,
VH5: GCAGCTGGTGCAGTCTGGAG,
VH6: CAGGACTGGTGAAGCCCTCG,
VH7: CAGGTGCAGCTGGTGCAA)

Reverse:
Cμ: CAGGAGACGAGGGGGAAAAGG
Cγ: CCGATGGGCCCTTGGTGGA
Cα: GAAGACCTTGGGGCTGGTCG)

F tag: TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG
R tag: GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG

Bioinformatics analysis of Next Generation Sequencing Data

An in-house developed informatics pipeline was used for initial quality filtering and clonal clustering analysis of sequencing data. After paired-end reads were joined, sequences were filtered based on a length and quality threshold. Sequences less than 200 bp and sequences with poor overlaps (>8% difference in linked region) and/or high number of bp below a threshold score (sequences containing more than 15 bp with less than Q30, 10bp with less than Q20, or any bp with less than Q10 scores) were excluded from further analysis. Isotypes were then determined by analysis of the constant region segment of each sequence and then sequences were aligned using the data provided by IMGT/HighV-quest (http://www.imgt.org/HighV-QUEST/) [19]. See Tipton et al, 2015 for further reasoning and analysis [20].

Overlapping AID hotspot analysis

Analysis of overlapping AID hotspots was performed by identifying locations that contain the WGCW motif, where the mutated base is underlined. As these sites are palindromic, locations for these hotspots were counted twice, once for each strand. To determine whether TRACE recipient sites were located proximally to overlapping hotspots, the average shortest distance between a given TRACE recipient site and any given overlapping AID hotspot was calculated. To determine significance, each data set had the location of TRACE recipient sites randomized across the length of the sequence and the average shortest distance was calculated over 1000 iterations. Z-scores for each individual data set were determined in comparison to each sequence set's randomized set. Z-scores were combined into a single statistic using Stouffer's Z method. [21]

Clustering Analyses

Clustering was performed either on the chromosomal level or on the base pair level and done using permutational or Monte Carlo approach, respectively. For permutational approaches, equal numbers of TRACE hits were randomly assigned to a chromosome with weights to adjust for differences in chromosome size. The frequency of TRACE hits per chromosome bin were tallied and recorded over 1000 iterations and the frequency of TRACE hits at their original positions are compared to the permuted pool for each chromosome. Z-scores were determined by comparing the original TRACE hits' frequency to that of each chromosome in the permutated pool. Z-scores were combined into a single statistic using Stouffer's Z method.

For Monte Carlo approaches, equal number of TRACE hits were randomly assigned to specific locations in the genome, with weights given to account for differences in chromosome size, as above. Counts were determined for TRACE hits that occurred within one kilobase of each other over 1000 iterations of randomly assigned TRACE hits. Counts were also determined for original TRACE hits and the original count was compared to the population of counts from the randomly-assigned pool.

Germinal Center Transcript Analysis

Germinal center B and Naïve B RNA-seq data was obtained from GSE84022 [8] and RPKM values were averaged between replicates. Germinal center B upregulated transcripts were defined by having a RPKM value greater than that of Naïve B plus two standard deviations. Downregulated transcripts were similarly defined, except that the RPKM value for transcripts in germinal center B cells was less than that of naïve B minus two standard deviations. Permutational analysis was done by randomly selecting an equal number of genes and assaying how many genes in this set were present in either fraction for 1000 iterations. Original values were compared to those generated during the permutation process to yield Z scores.

Hi-C Analysis

Germinal center B and Naïve B Hi-C data was obtained from GSE84022 [8] and inter-chromosomal interactions were kept if one anchor overlapped the IgH locus and was supported by at least two reads. Cumulative plots of inter-chromosomal interactions were obtained by taking each examined locus and the closest distance to an anchor that was found to interact with the IgH locus. These distances in kilobases were Log_{10} transformed and plotted with an empirical cumulative distribution function. Kolmogorov-Smirnov tests were used to test the significance of the observed differences in the distributions between samples.

Statistics

Tests for significance used herein include the one sample Z-test, Stouffer's Z method, and Kolmogorov-Smirnov test. One sample Z-tests were used in either permutational or Monte Carlo analyses. For all tests, significance was set at a Z-score of 1.645, which corresponds to a one-tailed alpha of 0.05. In cases where multiple Z-scores were combined for an aggregate statistic, only samples belonging to the same group were combined and Stouffer's Z method was used. As in the one sample Z-test, a cutoff Z-score of ±1.645 was used to determine significance, with direction being chosen depending on the analysis. Kolmogorov-Smirnov tests were performed with a two-tailed alpha of 0.05.

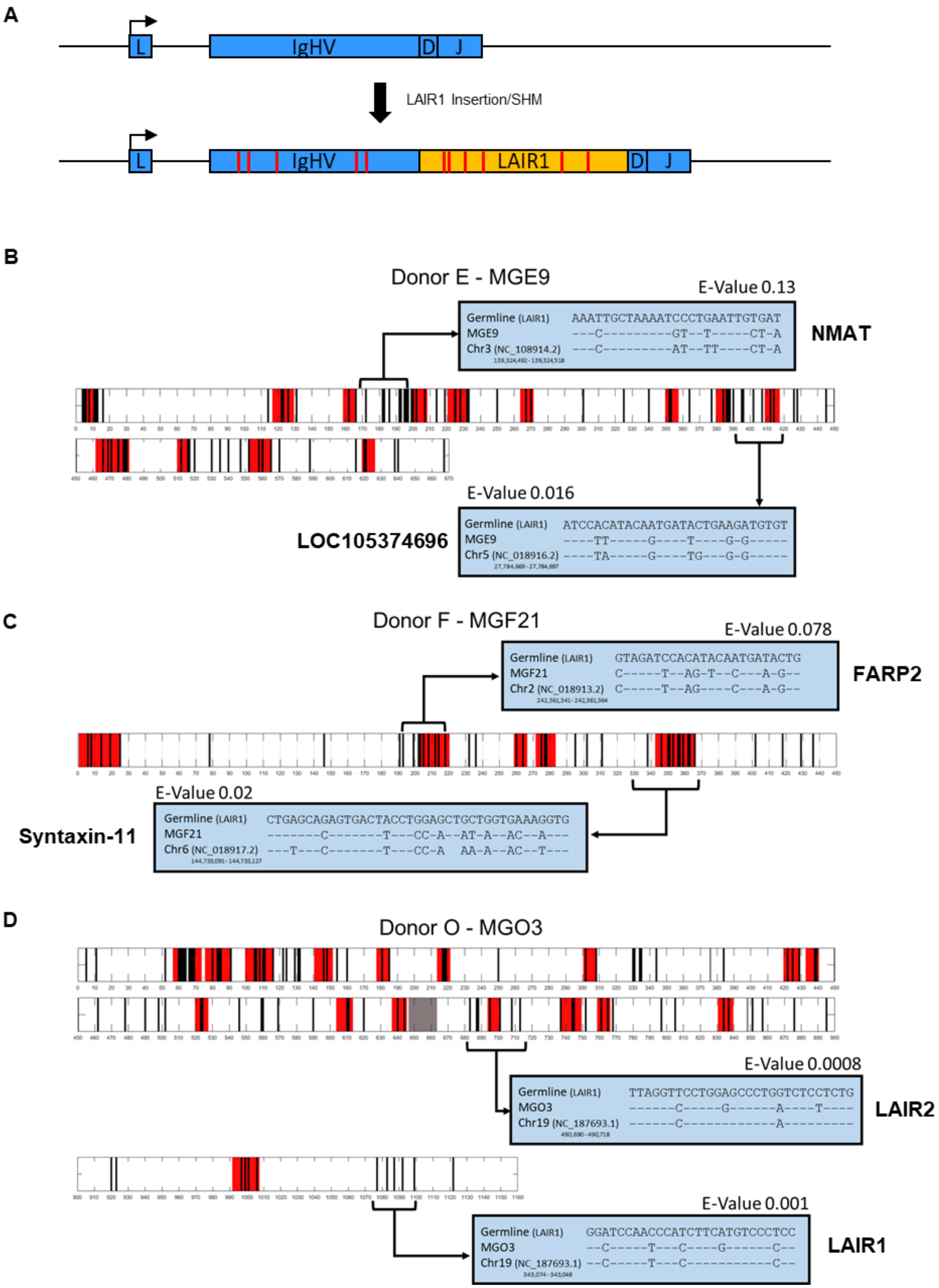
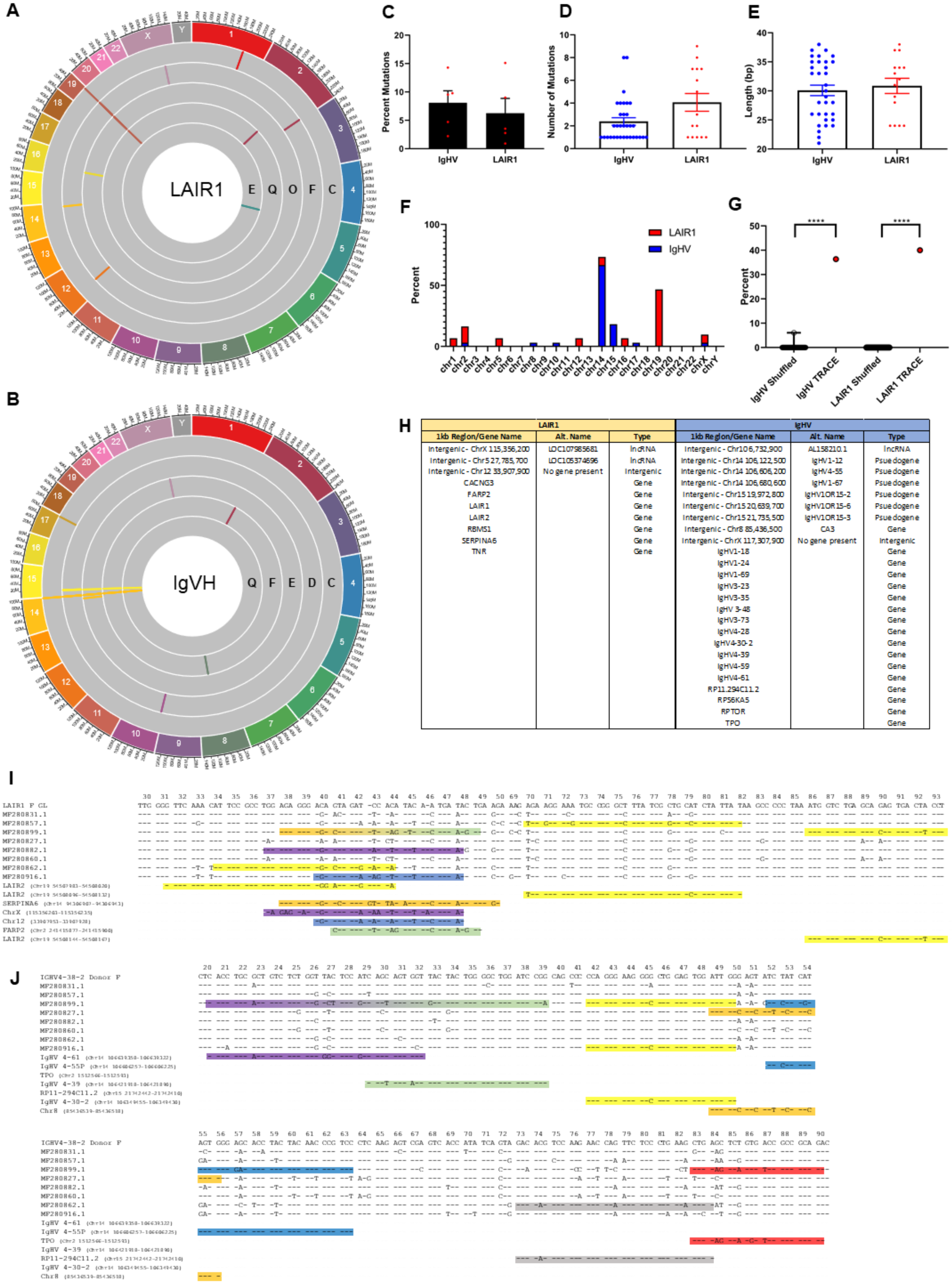
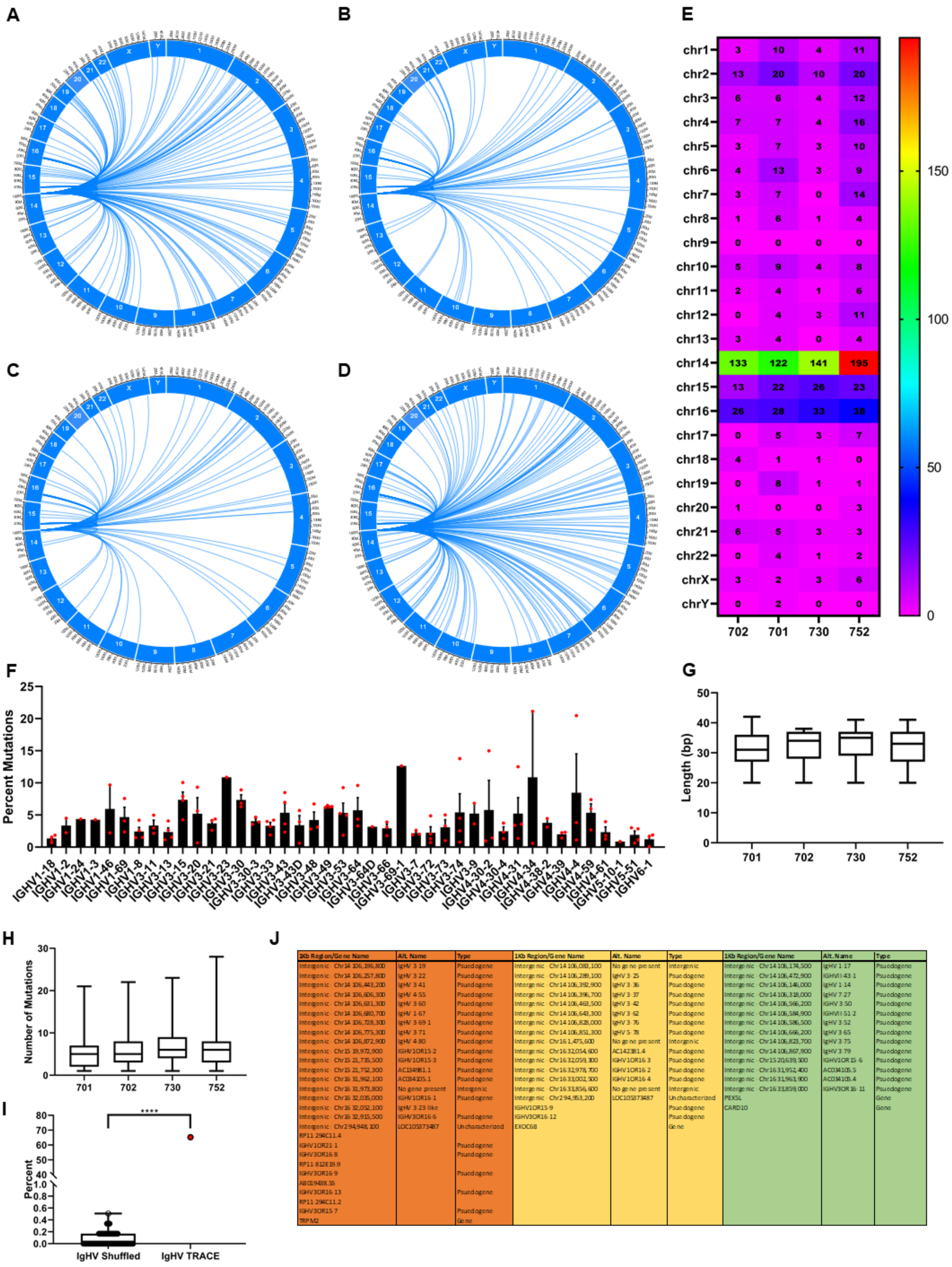

A
L
IgHV
D
J
LAIR1 Insertion/SHM
L
IgHV
LAIR1
D
J
B
Donor E - MGE9
E-Value 0.13
Germline (LAIR1) AAATTGCTAAAATCCCTGAATTGTGAT
MGE9 ---C----------GT--T----CT-A
Chr3 (NC_108914.2) ---C----------AT--TT---CT-A
139,324,492 - 139,324,518
NMAT
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250 260 270 280 290 300 310 320 330 340 350 360 370 380 390 400 410 420 430 440 450
450 460 470 480 490 500 510 520 530 540 550 560 570 580 590 600 610 620 630 640 650 660 670
E-Value 0.016
LOC105374696
Germline (LAIR1) ATCCACATACAATGATACTGAAGATGTGT
MGE9 ----TT-----G----T----G-G-----
Chr5 (NC_018916.2) ----TA-----G----TG---G-G-----
27,784,669 - 27,784,697
C
Donor F - MGF21
E-Value 0.078
Germline (LAIR1) GTAGATCCACATACAATGATACTG
MGF21 C-----T--AG-T--C---A-G--
Chr2 (NC_018913.2) C-----T--AG----C---A-G--
242,361,341 - 242,361,364
FARP2
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250 260 270 280 290 300 310 320 330 340 350 360 370 380 390 400 410 420 430 440 450
E-Value 0.02
Syntaxin-11
Germline (LAIR1) CTGAGCAGAGTGACTACCTGGAGCTGCTGGTGAAAGGTG
MGF21 -------C-------T---CC-A--AT-A--AC--A---
Chr6 (NC_018917.2) ---T---C-------T---CC-A AA-A--AC--T---
144,733,091 - 144,733,127
D
Donor O - MGO3
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250 260 270 280 290 300 310 320 330 340 350 360 370 380 390 400 410 420 430 440 450
450 460 470 480 490 500 510 520 530 540 550 560 570 580 590 600 610 620 630 640 650 660 670 680 690 700 710 720 730 740 750 760 770 780 790 800 810 820 830 840 850 860 870 880 890 900
E-Value 0.0008
Germline (LAIR1) TTAGGTTCCTGGAGCCCTGGTCTCCTCTG
MGO3 ------C-----G------A----T----
Chr19 (NC_187693.1) ------C------------A---------
490,690 - 490,718
LAIR2
900 910 920 930 940 950 960 970 980 990 1000 1010 1020 1030 1040 1050 1060 1070 1080 1090 1100 1110 1120 1130 1140 1150 1160
E-Value 0.001
Germline (LAIR1) GGATCCAACCCATCTTCATGTCCCTCC
MGO3 --C-----T---C----G------C--
Chr19 (NC_187693.1) --C-----T---C-----------C--
343,074 - 343,048
LAIR1

Fig. 1: Somatically-mutated LAIR1 inserts have regions of clustered mutations that match distant genomic regions. (**A**) A schematic of the generation of LAIR1 containing antibodies described by Tan et al. and Pieper et al. Germline VDJ rearrangements acquire LAIR1 insertions in CDR3 that are somatically diversified and confer antigen binding. (**B-D**) Somatically-mutated sequences obtained from different human donors were analyzed for small clusters of templated mutagenesis events as in Dale *et al*. and are depicted as a highlighter plot. In each panel, a single sequence is shown, black bars indicate mutations at a given position, whereas red regions indicate regions that match short subsequences in the IgHV germline repertoire. In panel **D** the gray region indicates the presence of a gap in the alignment to germline LAIR1. Selected subsequences containing regions of clustered mutations (indicated by brackets) were run through BLASTn. Alignments of germline LAIR1, somatically-mutated LAIR1, and the top BLASTn hit identified for that subsequence are shown. Gene names and E-values for searches conducted are shown for corresponding BLASTn hits.



LAIR1			IgHV		
1kb Region/Gene Name	Alt. Name	Type	1kb Region/Gene Name	Alt. Name	Type
Intergenic - ChrX 115,356,200	LOC107985681	lncRNA	Intergenic - Chr106,732,900	AL158210.1	lncRNA
Intergenic - Chr5 27,785,700	LOC105374696	lncRNA	Intergenic - Chr14 106,122,500	IgHV1-12	Psuedogene
Intergenic - Chr12 33,907,900	No gene present	Intergenic	Intergenic - Chr14 106,606,200	IgHV4-55	Psuedogene
CACNG3		Gene	Intergenic - Chr14 106,680,600	IgHV1-67	Psuedogene
FARP2		Gene	Intergenic - Chr15 19,972,800	IgHV1OR15-2	Psuedogene
LAIR1		Gene	Intergenic - Chr15 20,639,700	IgHV1OR15-6	Psuedogene
LAIR2		Gene	Intergenic - Chr15 21,735,500	IgHV1OR15-3	Psuedogene
RBMS1		Gene	Intergenic - Chr8 85,436,500	CA3	Gene
SERPINA6		Gene	Intergenic - ChrX 117,307,900	No gene present	Intergenic
TNR		Gene	IgHV1-18		Gene
			IgHV1-24		Gene
			IgHV1-69		Gene
			IgHV3-23		Gene
			IgHV3-35		Gene
			IgHV 3-48		Gene
			IgHV3-73		Gene
			IgHV4-28		Gene
			IgHV4-30-2		Gene
			IgHV4-39		Gene
			IgHV4-59		Gene
			IgHV4-61		Gene
			RP11.294C11.2		Gene
			RPS6KA5		Gene
			RPTOR		Gene
			TPO		Gene

Fig. 2: TRACE identified regions of the genome contribute to somatic mutagenesis of IgHV/LAIR1 sequences. (**A-B**) Circos plots for (**A**) LAIR1 and (**B**) corresponding IgHV of multiple human donors (gray concentric circles). Human donors are annotated for each concentric circle and donor names correspond to those in Tan *et al*. and Pieper *et al*. (**C**) the percent of mutations that derive from a TRACE identified donor sequence. (**D-E**) the number of mutations accounted for per individual TRACE hit (**D**) as well as the overall length of the identified TRACE hit (**E**). (**F**) A stacked bar graph of the percent of TRACE hits per chromosome. (**G**) The percent of an equal number of randomly scattered TRACE hits (modeled, n=1000) and of the data generated from the LAIR1 rearrangements that cluster within 1Kb of one another. (**H**) A table of the genomic regions that possessed clusters of TRACE hits. (**I-J**) Alignments of the (**I**) LAIR1 and (**J**) IgHV somatic variants identified in the LAIR1-containing antibodies isolated from Donor F in addition to the identified TRACE hits for each of these regions. **** p<0.0001



	702	701	730	752
chr1	3	10	4	11
chr2	13	20	10	20
chr3	6	6	4	12
chr4	7	7	4	16
chr5	3	7	3	10
chr6	4	13	3	9
chr7	3	7	0	14
chr8	1	6	1	4
chr9	0	0	0	0
chr10	5	9	4	8
chr11	2	4	1	6
chr12	0	4	3	11
chr13	3	4	0	4
chr14	133	122	141	195
chr15	13	22	26	23
chr16	26	28	33	38
chr17	0	5	3	7
chr18	4	1	1	0
chr19	0	8	1	1
chr20	1	0	0	3
chr21	6	5	3	3
chr22	0	4	1	2
chrX	3	2	3	6
chrY	0	2	0	0

1Kb Region/Gene Name	Alt. Name	Type	1Kb Region/Gene Name	Alt. Name	Type	1Kb Region/Gene Name	Alt. Name	Type
Intergenic Chr14 106,196,800	IgHV 3 19	Psuedogene	Intergenic Chr14 106,083,100	No gene present	Intergenic	Intergenic Chr14 106,174,500	IgHV 1 17	Psuedogene
Intergenic Chr14 106,257,800	IgHV 3 22	Psuedogene	Intergenic Chr14 106,289,100	IgHV 3 25	Psuedogene	Intergenic Chr14 106,472,900	IGHVII 43 1	Psuedogene
Intergenic Chr14 106,443,200	IgHV 3 41	Psuedogene	Intergenic Chr14 106,392,900	IgHV 3 36	Psuedogene	Intergenic Chr14 106,146,000	IgHV 1 14	Psuedogene
Intergenic Chr14 106,606,300	IgHV 4 55	Psuedogene	Intergenic Chr14 106,396,700	IgHV 3 37	Psuedogene	Intergenic Chr14 106,318,000	IgHV 7 27	Psuedogene
Intergenic Chr14 106,631,300	IgHV 3 60	Psuedogene	Intergenic Chr14 106,463,500	IgHV 3 42	Psuedogene	Intergenic Chr14 106,566,200	IGHV 3 50	Psuedogene
Intergenic Chr14 106,680,700	IgHV 1 67	Psuedogene	Intergenic Chr14 106,643,300	IgHV 3 62	Psuedogene	Intergenic Chr14 106,584,900	IGHVII 51 2	Psuedogene
Intergenic Chr14 106,728,300	IgHV 3 69 1	Psuedogene	Intergenic Chr14 106,828,000	IgHV 3 76	Psuedogene	Intergenic Chr14 106,586,500	IgHV 3 52	Psuedogene
Intergenic Chr14 106,775,300	IgHV 3 71	Psuedogene	Intergenic Chr14 106,851,300	IgHV 5 78	Psuedogene	Intergenic Chr14 106,666,200	IgHV 3 65	Psuedogene
Intergenic Chr14 106,872,900	IgHV 4 80	Psuedogene	Intergenic Chr16 1,475,600	No gene present	Intergenic	Intergenic Chr14 106,823,700	IgHV 3 75	Psuedogene
Intergenic Chr15 19,972,900	IGHV1OR15 2	Psuedogene	Intergenic Chr16 32,054,600	AC142381.4	Psuedogene	Intergenic Chr14 106,867,900	IgHV 3 79	Psuedogene
Intergenic Chr15 21,735,500	IGHV1OR15 3	Psuedogene	Intergenic Chr16 32,059,300	IGHV1OR16 3	Psuedogene	Intergenic Chr15 20,639,300	IGHV1OR 15 6	Psuedogene
Intergenic Chr15 21,752,300	AC134981.1	Psuedogene	Intergenic Chr16 32,978,700	IGHV1OR16 2	Psuedogene	Intergenic Chr16 31,952,400	AC034105.5	Psuedogene
Intergenic Chr16 31,962,100	AC034105.1	Psuedogene	Intergenic Chr16 33,002,500	IGHV1OR16 4	Psuedogene	Intergenic Chr16 31,963,900	AC034105.4	Psuedogene
Intergenic Chr16 31,973,800	No gene present	Intergenic	Intergenic Chr16 33,856,600	No gene present	Intergenic	Intergenic Chr16 33,859,000	IGHV3OR 16 11	Psuedogene
Intergenic Chr16 32,035,000	IGHV1OR16 1	Psuedogene	Intergenic Chr2 94,953,200	LOC105373487	Uncharacterized	PEX5L		Gene
Intergenic Chr16 32,052,100	IgHV 3 23 like		IGHV1OR15 9		Psuedogene	CARD10		Gene
Intergenic Chr16 32,915,500	IGHV3OR16 6	Psuedogene	IGHV3OR16 12		Psuedogene			
Intergenic Chr2 94,948,100	LOC105373487	Uncharacterized	EXOC6B		Gene			
RP11 294C11.4								
IGHV1OR21 1		Psuedogene						
IGHV3OR16 8		Psuedogene						
RP11 812E19.9								
IGHV3OR16 9		Psuedogene						
AB019438.55								
IGHV3OR16 13		Psuedogene						
RP11 294C11.2								
IGHV3OR15 7		Psuedogene						
TRPM2		Gene						

Fig. 3: Somatically-mutated populations of switched memory B cells template somatic mutations from intra- and inter-chromosomal regions. (**A-D**) Circos plots of TRACE identified templates that contribute to the somatic mutation profile of four healthy human donors. Circos plots correspond to human donor (**A**) 701, (**B**) 702, (**C**) 730, and (**D**) 752. (**E**) Heat map depicting the number of unique TRACE hits per healthy human donor per chromosome. Numbers within cells indicate the number of unique TRACE hits identified for each chromosome. (**F**) Percent of mutations per rearranged IgHV gene. Bars indicate the average percent between samples, whereas red dots denote individual values for each healthy human donor. (**G-H**) The lengths of TRACE hits identified in each human donor (**G**) as well as the number of mutations accounted for in each TRACE hit (**H**). (**I**) Percent of TRACE hits that cluster within 1kb. Data was modeled as in Fig. 2G (n=1000). (**J**) TRACE hits that clustered within 1kb between all four (orange), three (yellow), or two (green) human donors.

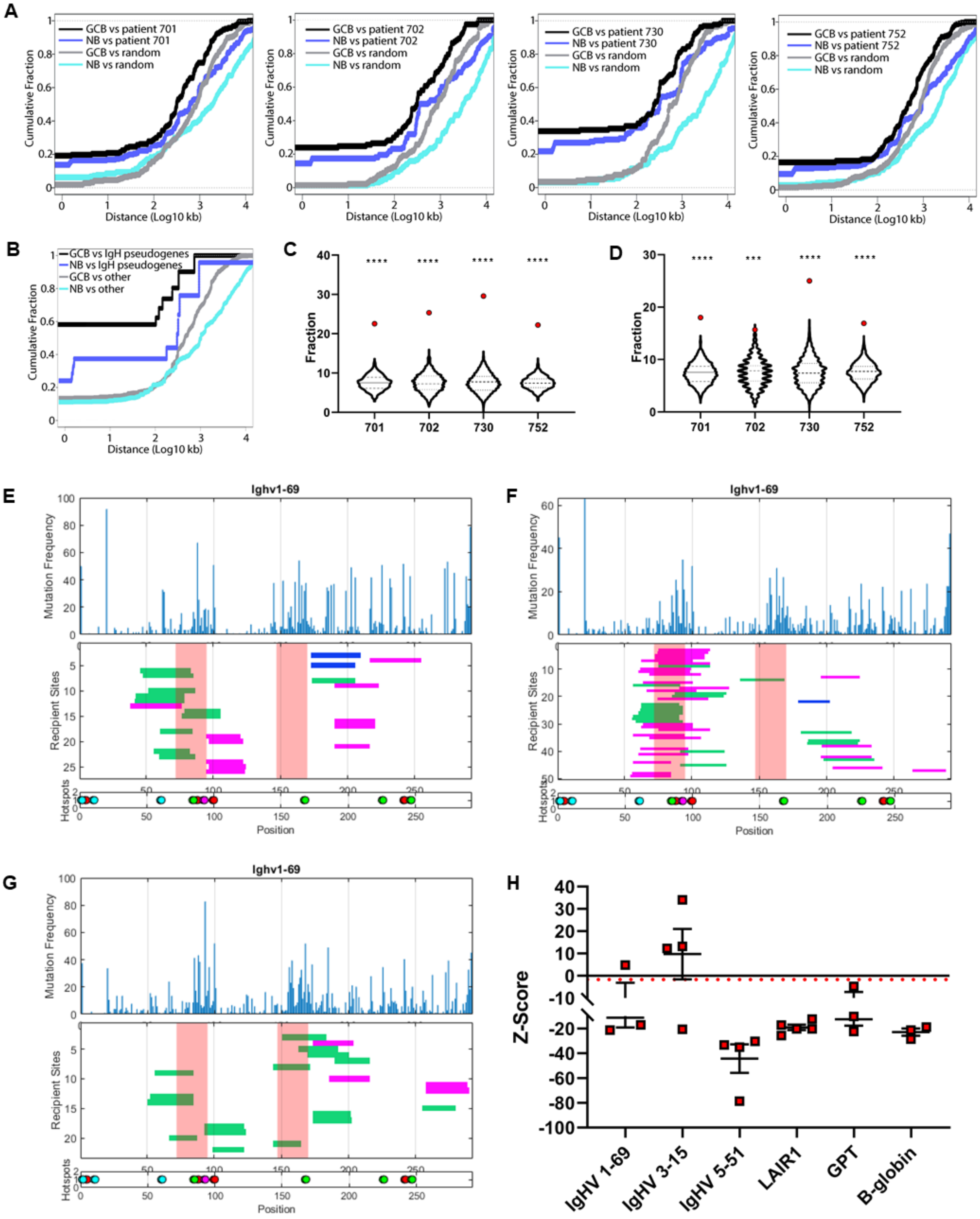

A
GCB vs patient 701
NB vs patient 701
GCB vs random
NB vs random
GCB vs patient 702
NB vs patient 702
GCB vs patient 730
NB vs patient 730
GCB vs patient 752
NB vs patient 752
Cumulative Fraction
Distance (Log10 kb)
B
GCB vs IgH pseudogenes
NB vs IgH pseudogenes
GCB vs other
NB vs other
C
D
Fraction
****
***
701
702
730
752
E
F
G
Ighv1-69
Mutation Frequency
Recipient Sites
Hotspots
Position
H
Z-Score
IgHV 1-69
IgHV 3-15
IgHV 5-51
LAIR1
GPT
B-globin

Fig. 4: TRACE hits correlate with multiple aspects of germinal center B cell biology. (**A**) Cumulative frequency plots of interchromosomal TRACE hits as a function of distance away from interchromosomal Hi-C contact points obtained from Bunting *et al*. For each human donor (patient) four curves are generated depicting the relationship between TRACE hits and germinal center B cell (GCB) Hi-C data, naïve B cell (NB) Hi-C data, and randomized Hi-C data for both. (**B**) Interchromosomal TRACE hits were separated into those that were identified as IgHV pseudogenes versus those that were not and were compared to interchromosomal Hi-C contact points as before. (**C-D**) Fraction of TRACE hits that are within 1kb of open chromatin peaks (red dot). Modeled data (violin plots) represents 1000 iterations of an equal number of randomly-scattered TRACE hits. Heavy dashed line represents the mean, with lighter dashed lines indicating interquartile range. Data shown are total TRACE hits (**C**) as well as interchromosomal TRACE hits (**D**). (**E-G**) Shown are three plots depicting mutation frequency, location of TRACE hits, and location of overlapping AID hotspots for the IgHV 1-69 rearrangement from human donors in Fig. 3. The mutation frequency plot depicts the frequency of mutation at a given position along the IgHV gene segment, independent of clonality. The plot depicting the location of TRACE hits in recipient sequences depicts each TRACE hit as a horizontal bar indicating its length along the IgHV gene segment. Colors used indicate intrachromosomal (blue), interchromosomal (green), and IgHV-derived (magenta). Vertical red shaded regions represent CDR1 and CDR2 respectively. Unshaded areas are FR1, FR2, and FR3 respectively. The plot depicting overlapping AID hotspots depicts the location of the WGCW motif. Each color represents a different member of WGCW: AGCT (red), AGCA (green), TGCA (cyan), and TGCT (magenta). (**H**) Z score statistics for Monte Carlo simulations of average shortest distance between TRACE hits and overlapping AID sites for IgHV 1-69, IgHV 3-15, and IgHV 5-51 as

well as LAIR1 insertions from Tan *et al*. and Pieper *et al*. and unselected passenger transgenes in a mouse model GPT and β-globlin reported in Yeap *et al*. *** p<0.001 **** p<0.0001

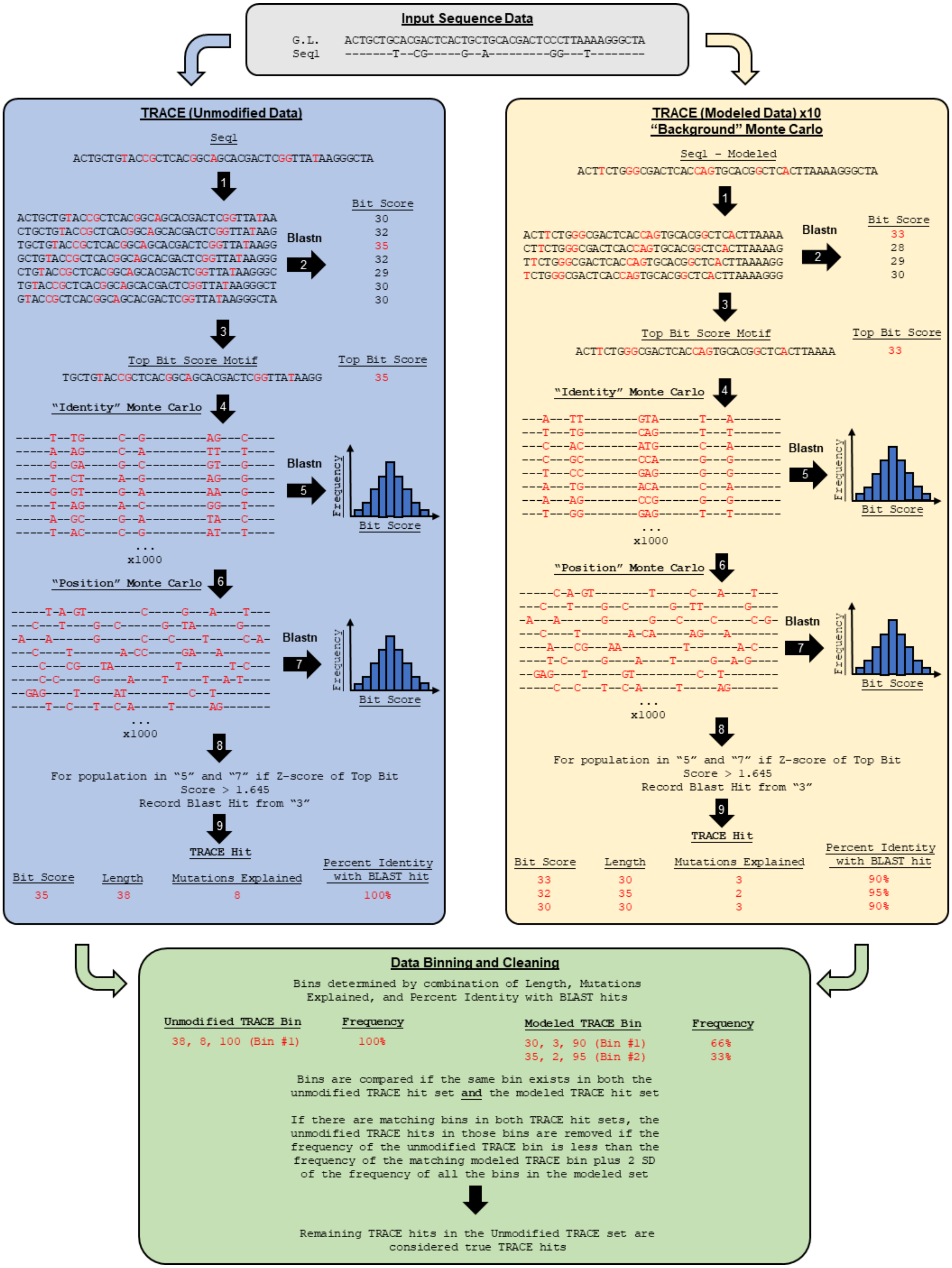

Input Sequence Data
G.L. ACTGCTGCACGACTCACTGCTGCACGACTCCCTTAAAAGGGCTA
Seql --------T--CG-----G--A---------GG---T--------
TRACE (Unmodified Data)
Seql
ACTGCTGTACCGCTCACGGCAGCACGACTCGGTTATAAGGGCTA
1
Bit Score
ACTGCTGTACCGCTCACGGCAGCACGACTCGGTTATAA 30
CTGCTGTACCGCTCACGGCAGCACGACTCGGTTATAAG 32
TGCTGTACCGCTCACGGCAGCACGACTCGGTTATAAGG 35
GCTGTACCGCTCACGGCAGCACGACTCGGTTATAAGGG 32
CTGTACCGCTCACGGCAGCACGACTCGGTTATAAGGGC 29
TGTACCGCTCACGGCAGCACGACTCGGTTATAAGGGCT 30
GTACCGCTCACGGCAGCACGACTCGGTTATAAGGGCTA 30
Blastn
2
3
Top Bit Score Motif
TGCTGTACCGCTCACGGCAGCACGACTCGGTTATAAGG
Top Bit Score
35
"Identity" Monte Carlo
4
-----T--TG-----C--G---------AG---C----
-----A--AG-----C--A---------TT---T----
-----G--GA-----G--C---------GT---G----
-----T--CT-----A--G---------AG---G----
-----G--GT-----G--A---------AA---G----
-----T--AG-----A--C---------GG---T----
-----A--GC-----G--A---------TA---C----
-----T--AC-----C--G---------AT---T----
...
x1000
Blastn
5
Frequency
Bit Score
"Position" Monte Carlo
6
-----T-A-GT--------C-----G---A----T---
---C---T----G--C------G--TA------G----
-A---A------G------C---C----T-----C-A-
---C----T-------A-CC-----GA---A-------
----C---CG---TA---------T-------T-C---
----C-C-----G----A----T-----T--A-T----
--GAG----T-----AT---------C--T--------
-----T--C---T--C-A-----T-----AG-------
...
x1000
Blastn
7
Frequency
Bit Score
8
For population in "5" and "7" if Z-score of Top Bit Score > 1.645
Record Blast Hit from "3"
9
TRACE Hit
Bit Score
Length
Mutations Explained
Percent Identity with BLAST hit
35
38
8
100%
TRACE (Modeled Data) x10
"Background" Monte Carlo
Seql – Modeled
ACTTCTGGGCGACTCACCAGTGCACGGCTCACTTAAAAGGGCTA
1
Bit Score
ACTTCTGGGCGACTCACCAGTGCACGGCTCACTTAAAA 33
CTTCTGGGCGACTCACCAGTGCACGGCTCACTTAAAAG 28
TTCTGGGCGACTCACCAGTGCACGGCTCACTTAAAAGG 29
TCTGGGCGACTCACCAGTGCACGGCTCACTTAAAAGGG 30
Blastn
2
3
Top Bit Score Motif
ACTTCTGGGCGACTCACCAGTGCACGGCTCACTTAAAA
Top Bit Score
33
"Identity" Monte Carlo
4
---A---TT--------GTA------T---A-------
---T---TG--------CAG------T---T-------
---C---AC--------ATG------C---A-------
---C---GC--------CCA------G---G-------
---T---CC--------GAG------G---G-------
---A---TG--------ACA------C---A-------
---A---AG--------CCG------G---G-------
---T---GG--------GAG------T---T-------
...
x1000
Blastn
5
Frequency
Bit Score
"Position" Monte Carlo
6
-----C-A-GT--------T-----C---A----T---
---C---T----G--C------G--TT------G----
-A---A------G------G---C----C-----C-G-
---C----T-------A-CA-----AG---A-------
----A---CG---AA---------T-------A-C---
----T-C-----G----A----T-----G--A-G----
--GAG----T-----GT---------C--T--------
-----C--C---T--C-A-----T-----AG-------
...
x1000
Blastn
7
Frequency
Bit Score
8
For population in "5" and "7" if Z-score of Top Bit Score > 1.645
Record Blast Hit from "3"
9
TRACE Hit
Bit Score
Length
Mutations Explained
Percent Identity with BLAST hit
33 30 3 90%
32 35 2 95%
30 30 3 90%
Data Binning and Cleaning
Bins determined by combination of Length, Mutations Explained, and Percent Identity with BLAST hits
Unmodified TRACE Bin
Frequency
38, 8, 100 (Bin #1)
100%
Modeled TRACE Bin
Frequency
30, 3, 90 (Bin #1) 66%
35, 2, 95 (Bin #2) 33%
Bins are compared if the same bin exists in both the unmodified TRACE hit set and the modeled TRACE hit set
If there are matching bins in both TRACE hit sets, the unmodified TRACE hits in those bins are removed if the frequency of the unmodified TRACE bin is less than the frequency of the matching modeled TRACE bin plus 2 SD of the frequency of all the bins in the modeled set
Remaining TRACE hits in the Unmodified TRACE set are considered true TRACE hits

Fig. S1: Workflow of TRACE. TRACE is a custom Matlab script (v. 2018a) that analyzes input aligned FASTA files that have no gaps in the alignment. TRACE uses the first sequence (germline) to annotate the positions of mutations in all subsequent sequences. The script then scans sequences for regions which fulfill the user-supplied criteria for clustered mutations (herein ≥8 mutations over 38bp). The script generates sliding windows of the clustered region, such that each window is 38bp and the number of mutations contained in each window is ≥8. Each of these windows is sent through a local BLAST search (word size:11, max hsps: 1, maximum target sequences: 1) against either the human genome (GRCh38) or the mouse genome (GRCm38). The maximum bit score for each window is recorded and the highest scoring window is selected for further analysis. The selected window is then passed through two Monte Carlo simulations which check mutation identity and mutation position effects. In the first simulation, identity effects are analyzed by randomizing the mutation identity at the position in which it occurs in the window. This is iterated 1000 times to build a population of simulated windows with random mutation identities. Each simulated window is passed through BLAST as before to build a population of maximum bit scores. If the Z-score of the original window's bit score is ≥1.645 (corresponding to >95% of simulated window bit scores) as compared to the simulated windows, the original window is passed into the second Monte Carlo simulation. In the second simulation, positional effects are analyzed. This is done by randomizing the position of mutations across the original window. As positions are randomized, mutation identities at these new positions are also randomized. As before, 1000 simulated windows are generated and are sent through BLAST. If the Z-score of the original window's bit score is again ≥1.645 as compared to this second set of simulated windows, the window is recorded as a TRACE hit. This is iterated through the all sequences in the original data set. After completion of the original data

set, TRACE generates a user-defined number of simulated datasets, which contain the same number of analyzed clusters as in the original data set, albeit with mutations scrambled and clusters randomly placed across the sequence. This modeled data is analyzed in the same manner as the original data set. Upon completion of the modeled data set, TRACE compares the outputs of the original data and the modeled data. It does so by parameterizing each TRACE hit according to length, number of mutations accounted for, and percent identity with its corresponding BLAST output. Each TRACE hit is binned according to these parameters and the frequency of hits in each bin is determined. Cleaned TRACE hits are then determined by comparing bins between the original and modeled data sets. If a bin occurs in the original data set but does not occur in the modeled, the corresponding data is defined as a true hit. If a bin occurs in both data sets, the frequency of the modeled bin is compared to the frequency of the original bin. If the original bin's frequency is ≥2 standard deviations (calculated from the frequency of all bins in the modeled data) above the modeled bin's frequency, the data in the original bin defined as a true hit.

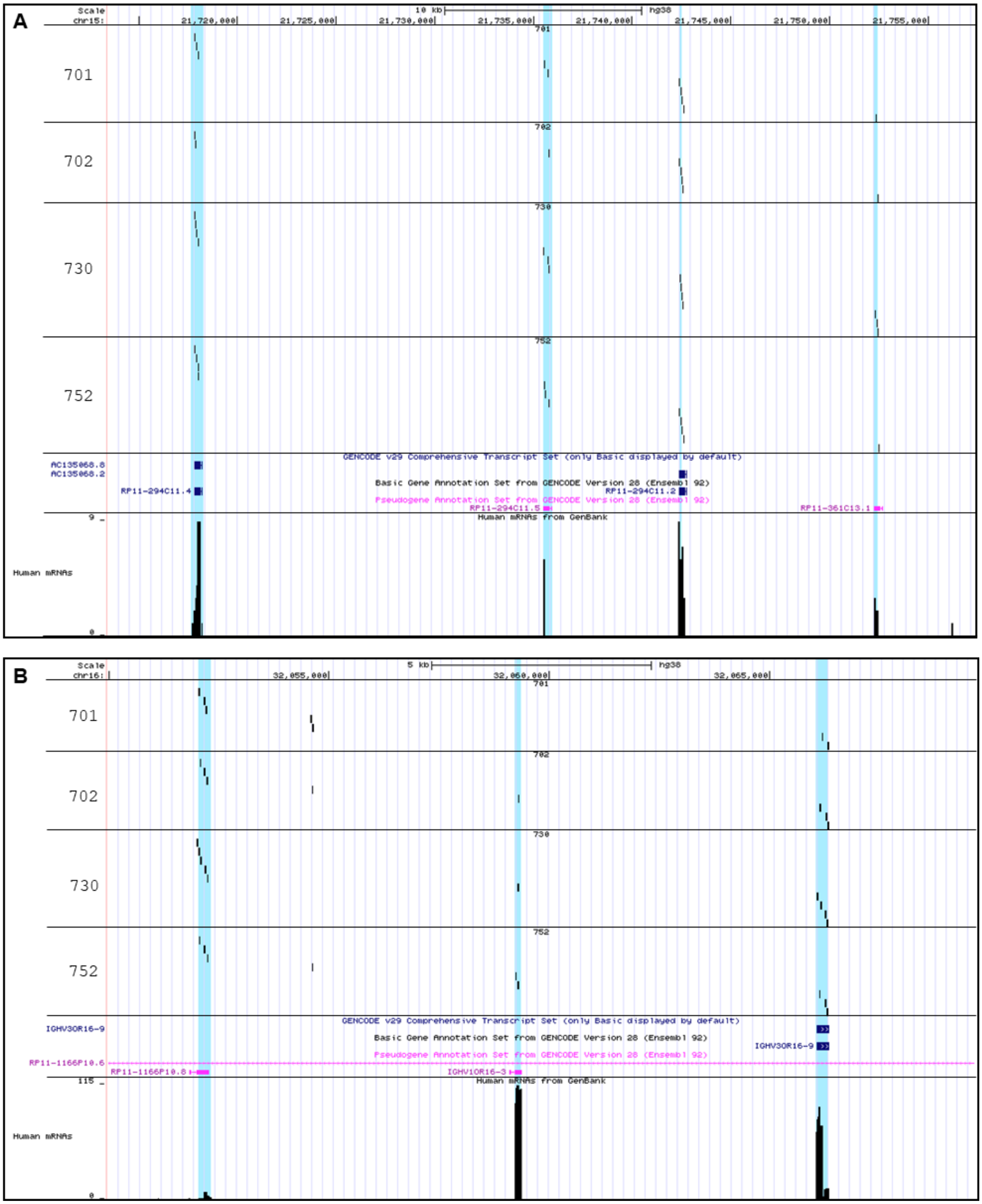


Fig. S2: TRACE donor regions cluster between and within human donor samples and correlate with areas of transcription. (**A-B**) UCSC Genome Browser sessions depicting

clustered mutations on (**A**) chromosome 15 and (**B**) chromosome 16. TRACE-identified donor sequences are depicted per human donor. Each black vertical bar represents a distinct TRACE donor sequence. GENCODE annotations are shown for transcripts and psuedogenes. Human mRNAs from Genbank (curated data by UCSC) are shown as scaled bar graphs (bottom). Blue highlighted regions depict transcribed regions. Scale bars are shown at the top of each panel.

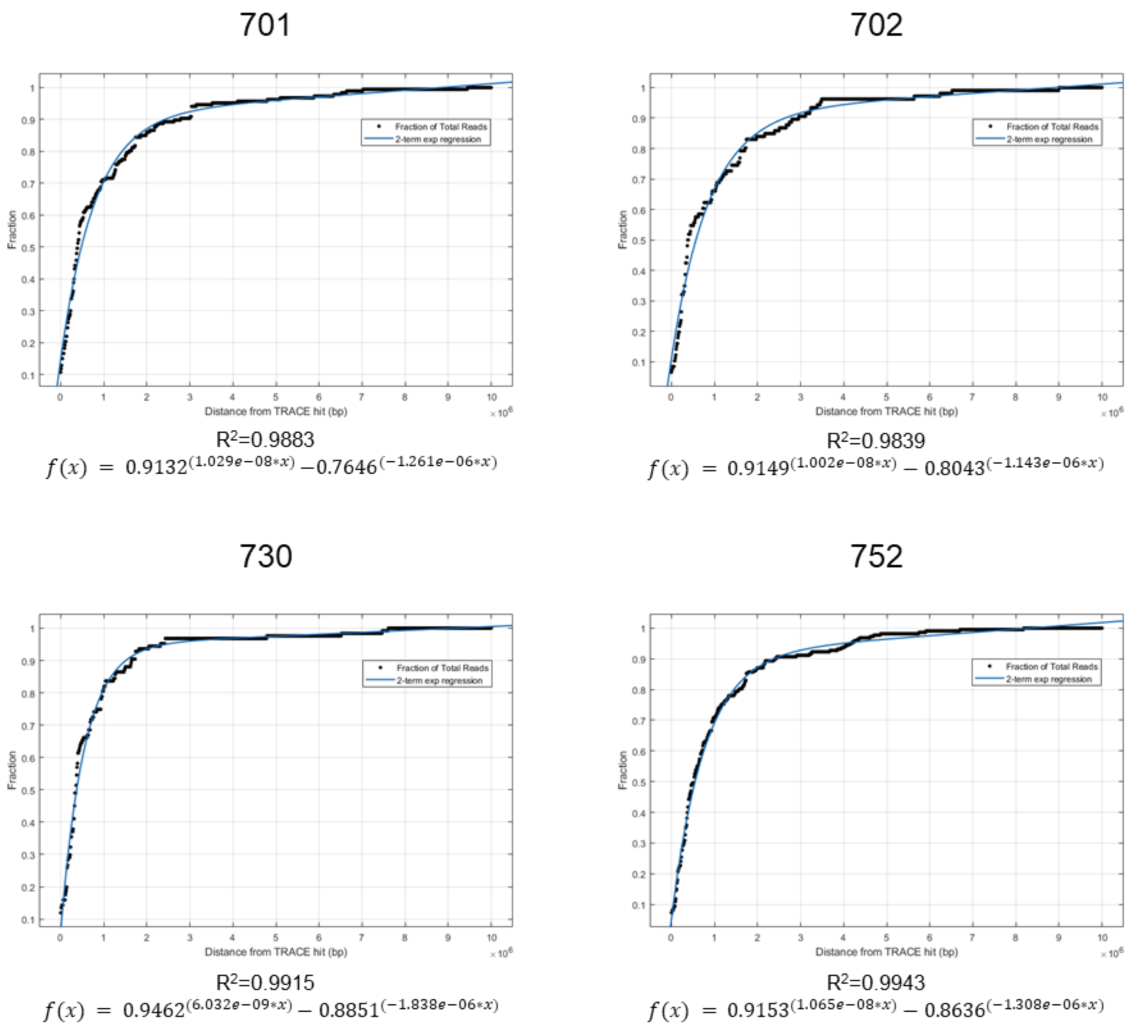


Fig. S3: Read-number of Hi-C data correlates with distance from TRACE-identified number. Data shown is a cumulative frequency distribution of the percent of total Hi-C contact reads with the IgH locus that are contained within 10Mb from TRACE donor sites. Hi-C contacts that are within 2Mb of TRACE hits comprise 85-95% of total reads that are within 10Mb and follows a two-term exponential distribution (blue). Fitted curves were generated using the curve fitting toolbox in Matlab (v. 2018a). Equations and R^2 values for each data set are shown below.

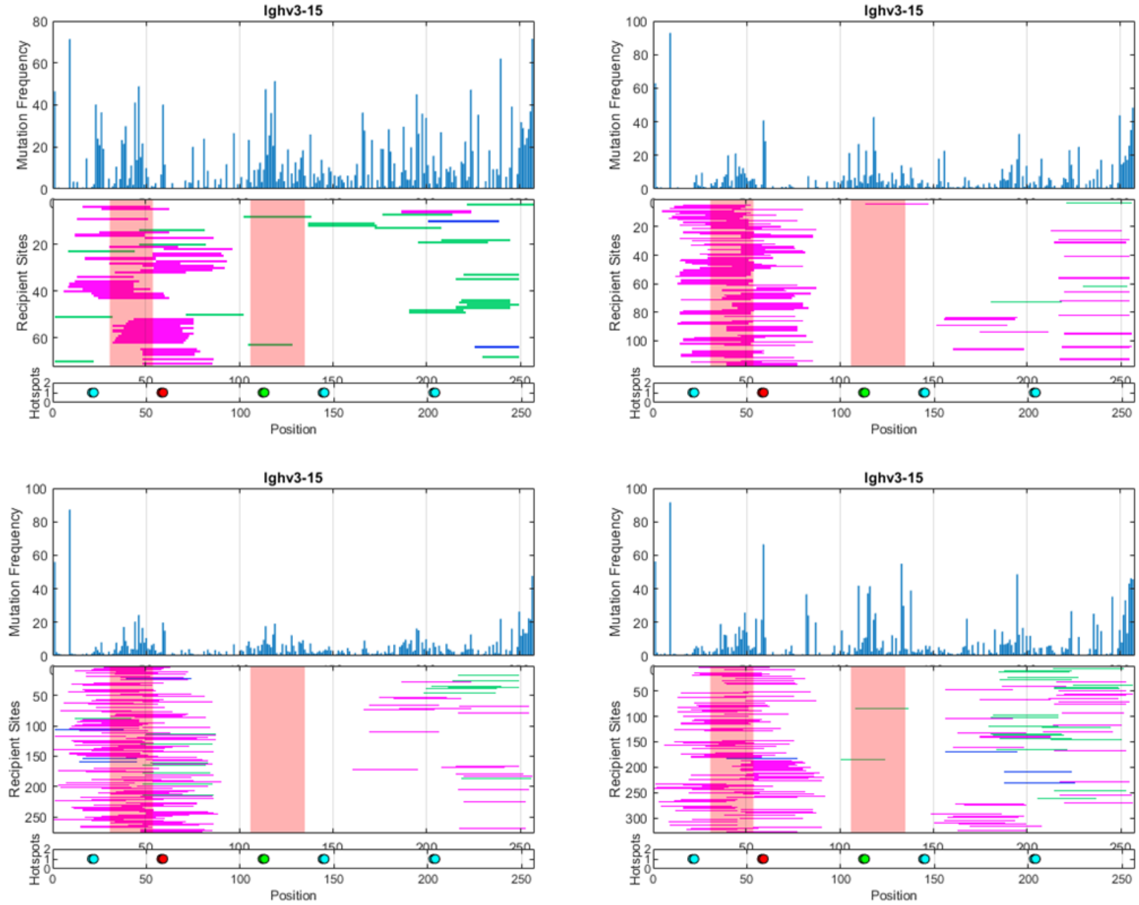


Fig. S4: Location of TRACE hits in IgHV 3-15. Data shown is from human donors 701, 702, 730, and 752 as presented in Figure 4E-G.

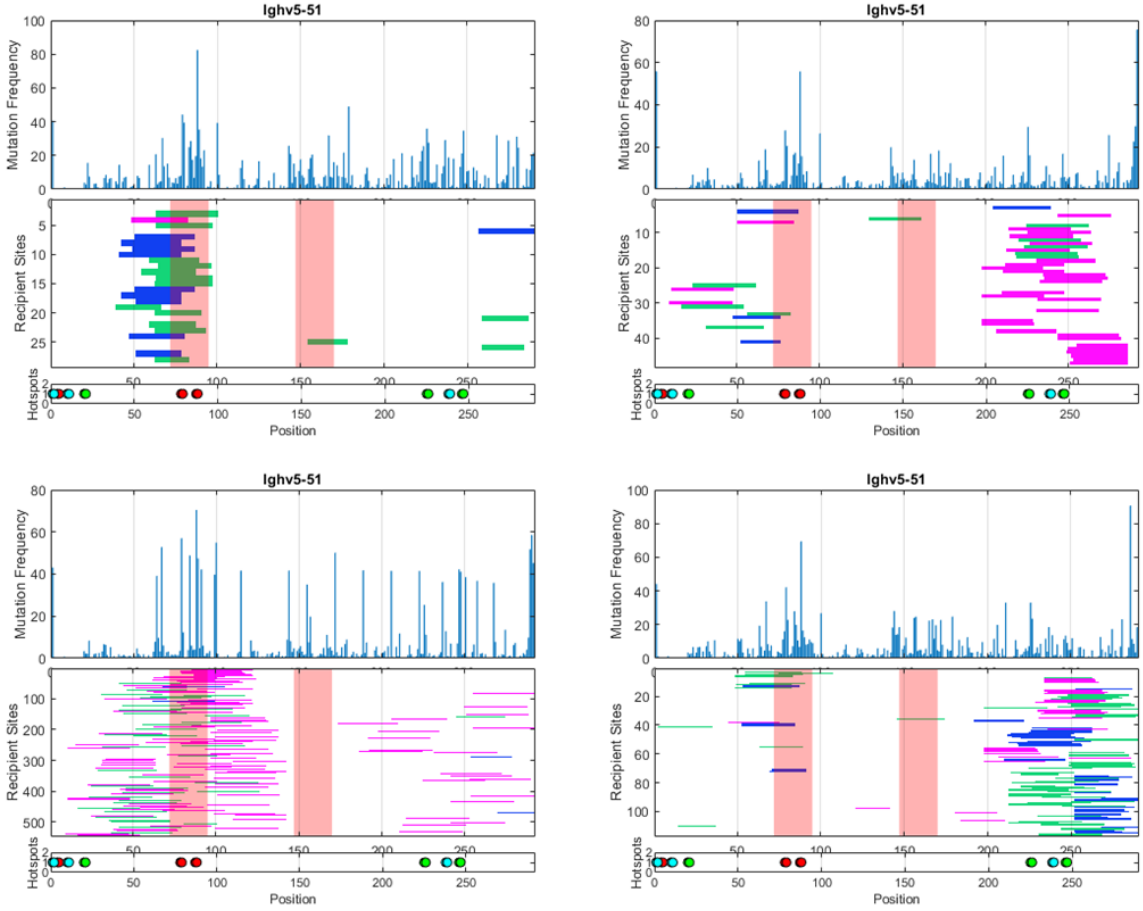


Fig. S5: Location of TRACE hits in IgHV 5-51. Data shown is from human donors 701, 702, 730, and 752 as presented in Figure 4E-G.

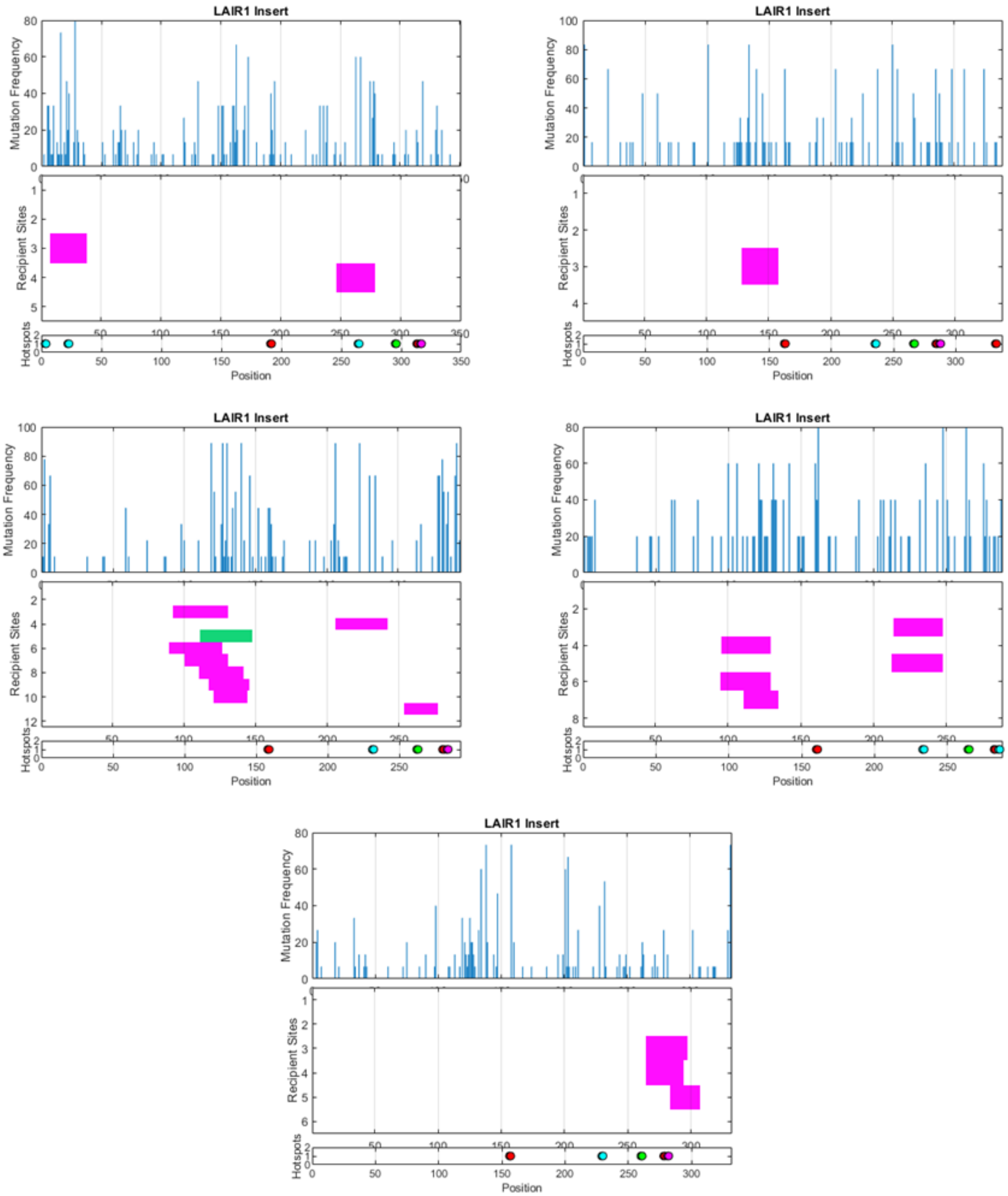


Fig. S6: Location of TRACE hits in LAIR1 insertions. Data shown is from human donors C, D, E, F, and Q as presented in Tan *et al*. and Pieper *et al*.

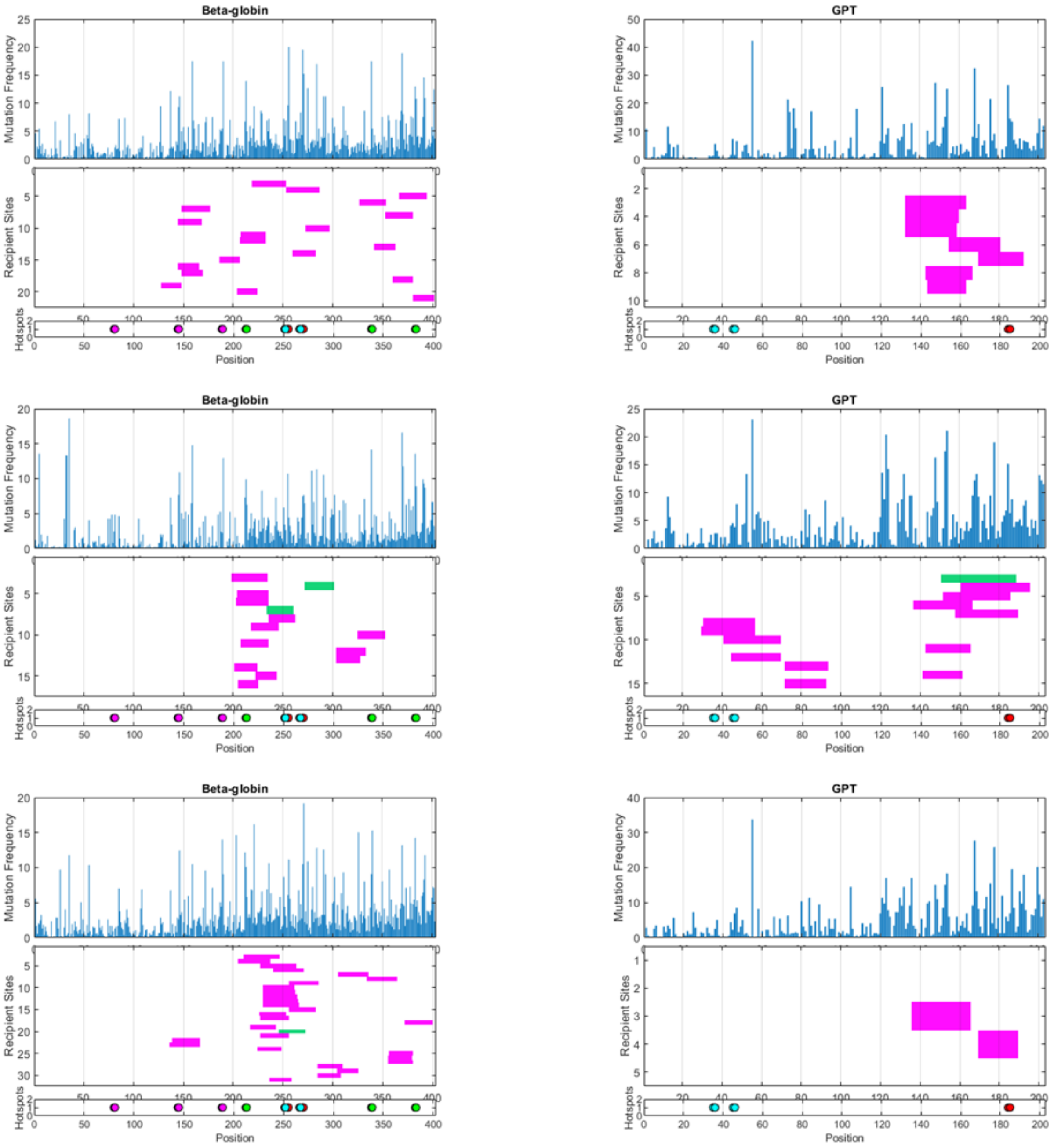
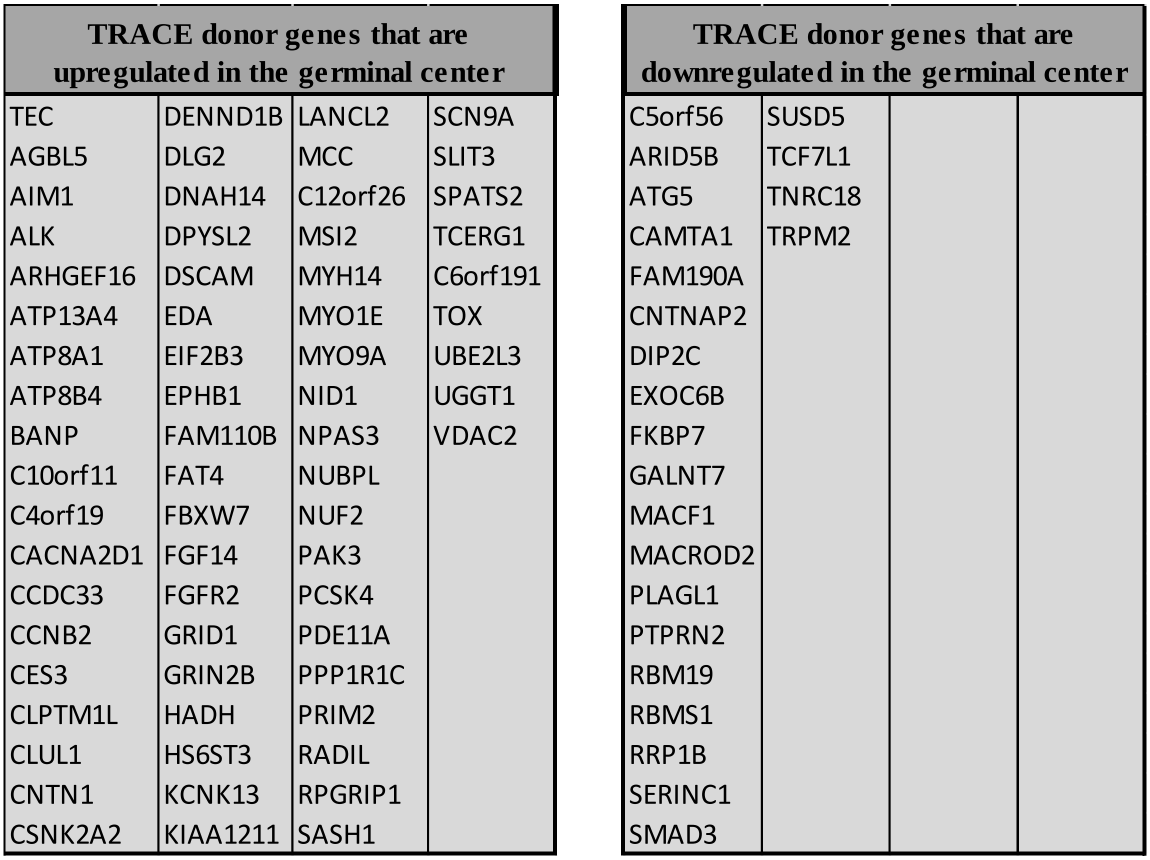


Fig. S7: Location of TRACE hits in unselected passenger transgenes. Data shown is from three murine samples for both GPT and β-globin as presented in Yeap *et al*.



TRACE donor genes that are upregulated in the germinal center			
TEC	DENND1B	LANCL2	SCN9A
AGBL5	DLG2	MCC	SLIT3
AIM1	DNAH14	C12orf26	SPATS2
ALK	DPYSL2	MSI2	TCERG1
ARHGEF16	DSCAM	MYH14	C6orf191
ATP13A4	EDA	MYO1E	TOX
ATP8A1	EIF2B3	MYO9A	UBE2L3
ATP8B4	EPHB1	NID1	UGGT1
BANP	FAM110B	NPAS3	VDAC2
C10orf11	FAT4	NUBPL	
C4orf19	FBXW7	NUF2	
CACNA2D1	FGF14	PAK3	
CCDC33	FGFR2	PCSK4	
CCNB2	GRID1	PDE11A	
CES3	GRIN2B	PPP1R1C	
CLPTM1L	HADH	PRIM2	
CLUL1	HS6ST3	RADIL	
CNTN1	KCNK13	RPGRIP1	
CSNK2A2	KIAA1211	SASH1	

TRACE donor genes that are downregulated in the germinal center			
C5orf56	SUSD5		
ARID5B	TCF7L1		
ATG5	TNRC18		
CAMTA1	TRPM2		
FAM190A			
CNTNAP2			
DIP2C			
EXOC6B			
FKBP7			
GALNT7			
MACF1			
MACROD2			
PLAGL1			
PTPRN2			
RBM19			
RBMS1			
RRP1B			
SERINC1			
SMAD3			

Table S1: TRACE donors in upregulated and downregulated fractions of differentially regulated genes in germinal center B cells.

Works Cited

1 Dale, G. A., Wilkins, D. J., Bohannon, C., Dilernia, D., Hunter, E., Bedford, T., Antia, R., Sanz, I., Jacob, J. Clustered mutations at the murine and human IgH locus exhibit significant linkage consistent with templated mutagenesis. *Journal of Immunology* (2019).

2 Pieper, K. *et al.* Public antibodies to malaria antigens generated by two LAIR1 insertion modalities. *Nature* **548**, 597-601, doi:10.1038/nature23670 (2017).

3 Tan, J. *et al.* A LAIR1 insertion generates broadly reactive antibodies against malaria variant antigens. *Nature* **529**, 105-109, doi:10.1038/nature16450 (2016).

4 Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421, doi:10.1186/1471-2105-10-421 (2009).

5 Wang, R. W., Lee, C. S. & Haber, J. E. Position effects influencing intrachromosomal repair of a double-strand break in budding yeast. *PLoS One* **12**, e0180994, doi:10.1371/journal.pone.0180994 (2017).

6 Cummings, W. J. *et al.* Chromatin structure regulates gene conversion. *PLoS Biol* **5**, e246, doi:10.1371/journal.pbio.0050246 (2007).

7 Schildkraut, E., Miller, C. A. & Nickoloff, J. A. Transcription of a donor enhances its use during double-strand break-induced gene conversion in human cells. *Mol Cell Biol* **26**, 3098-3105, doi:10.1128/MCB.26.8.3098-3105.2006 (2006).

8 Bunting, K. L. *et al.* Multi-tiered Reorganization of the Genome during B Cell Affinity Maturation Anchored by a Germinal Center-Specific Locus Control Region. *Immunity* **45**, 497-512, doi:10.1016/j.immuni.2016.08.012 (2016).

9 Scharer, C. D. *et al.* Epigenetic programming underpins B cell dysfunction in human SLE. *Nat Immunol*, doi:10.1038/s41590-019-0419-9 (2019).

10 Bastianello, G. & Arakawa, H. A double-strand break can trigger immunoglobulin gene conversion. *Nucleic Acids Res* **45**, 231-243, doi:10.1093/nar/gkw887 (2017).

11 Han, L., Masani, S. & Yu, K. Overlapping activation-induced cytidine deaminase hotspot motifs in Ig class-switch recombination. *Proc Natl Acad Sci U S A* **108**, 11584-11589, doi:10.1073/pnas.1018726108 (2011).

12 Zhang, Z. Z. *et al.* The strength of an Ig switch region is determined by its ability to drive R loop formation and its number of WGCW sites. *Cell Rep* **8**, 557-569, doi:10.1016/j.celrep.2014.06.021 (2014).

13 Yeap, L. S. *et al.* Sequence-Intrinsic Mechanisms that Target AID Mutational Outcomes on Antibody Genes. *Cell* **163**, 1124-1137, doi:10.1016/j.cell.2015.10.042 (2015).

14 Ohle, C. *et al.* Transient RNA-DNA Hybrids Are Required for Efficient Double-Strand Break Repair. *Cell* **167**, 1001-1013 e1007, doi:10.1016/j.cell.2016.10.001 (2016).

15 Onozawa, M., Goldberg, L. & Aplan, P. D. Landscape of insertion polymorphisms in the human genome. *Genome Biol Evol* **7**, 960-968, doi:10.1093/gbe/evv043 (2015).

16 Onozawa, M. *et al.* Repair of DNA double-strand breaks by templated nucleotide sequence insertions derived from distant regions of the genome. *Proc Natl Acad Sci U S A* **111**, 7729-7734, doi:10.1073/pnas.1321889111 (2014).

17 Su, Y., Egli, M. & Guengerich, F. P. Human DNA polymerase eta accommodates RNA for strand extension. *J Biol Chem* **292**, 18044-18051, doi:10.1074/jbc.M117.809723 (2017).

18 Su, Y. *et al.* Human DNA polymerase eta has reverse transcriptase activity in cellular environments. *J Biol Chem* **294**, 6073-6081, doi:10.1074/jbc.RA119.007925 (2019).

19 Brochet, X., Lefranc, M. P. & Giudicelli, V. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res* **36**, W503-508, doi:10.1093/nar/gkn316 (2008).

20 Tipton, C. M. *et al.* Diversity, cellular origin and autoreactivity of antibody-secreting cell population expansions in acute systemic lupus erythematosus. *Nat Immunol* **16**, 755-765, doi:10.1038/ni.3175 (2015).

21 Whitlock, M. C. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *J Evol Biol* **18**, 1368-1373, doi:10.1111/j.1420-9101.2005.00917.x (2005).

Chapter 4: Discussion

Gene conversion as an active contributor to somatic hypermutation

Over chapters 2 and 3, we have examined the role of templated mutagenesis (i.e. gene conversion) in two contexts: in chapter 2 we examine gene conversion as a function of microhomology motifs at the IgH locus, whereas in chapter 3 we examine gene conversion in a broader context and determine whether gene conversion utilizes the larger genome for diversification. Both methods are diametrically opposed in their approach, representing complementary views of the same phenomenon. In chapter 2, we qualitatively detect the contribution of gene conversion, being unable to define any given tract as gene conversion-derived but detecting a strong overall bias to germline motifs within multiple sequences. In chapter 3, we quantitively determine the contribution of the genome and large tracts of gene conversion to somatically mutated sequences. These approaches can be loosely conceptualized as defining upper and lower bounds for the contribution of gene conversion to the process of somatic hypermutation.

In chapter 2, we show (1) that tracts resembling gene conversion occur in murine IgM plasma cells following immunization with the hapten NPCGG, (2) the presence of these tracts in somatically-mutating murine IgM and IgG germinal center B cells, (3) a highly significant pattern of increasing linkage disequilibrium between mutations as the distance between them decreases, (4) use of motifs primarily upstream of the rearranged V segment but also the use of motifs on the opposite allele to introduce micro-clusters of mutations, (5) extension of these findings to circulating human plasmablasts, (6) that non-immunoglobulin sequences at the IgH locus mutate using the same motifs used to diversify the native antibodies, even in the absence of selective pressure, and (7) between 50-60% of unique somatic mutations in IgHV and non-immunoglobulin sequences are locally clustered (≥2 mutations over 8 bp), and contain a corresponding template in the germline IgHV repertoire. As a whole, these findings provide evidence against the accepted model of somatic hypermutation in mice and humans, that only untemplated point mutations arise and are selected in the germinal center[1-3]. Alternatively, our data indicate that local clusters of mutations, are derived from germline motifs present in the set of germline IgHV sequences. This model explains both (1) the observed linkage disequilibrium at short distances between pairs of mutations as well as (2) the selective use of certain motifs for introduction of multiple mutations in rearranged IgHV sequences as those in non-immunoglobulin sequences placed at the IgH locus, even in the absence of selective pressure.

In chapter 3, we provide evidence for gene conversion utilizing the genome itself for diversification of the rearranged IgHV segments. We show (1) large gene conversion tracts originate from inter- and intra-chromosomal locations, (2) donor choice is primarily driven by local homology, (3) between 5-10% of all somatic mutations are derived from these large templated events, (4) IgHV pseudogenes are used for templated mutagenesis, (5) donor sequences cluster between individuals suggesting a shared pattern of donor choice and mutagenesis, (6) non-immunoglobulin sequences at the IgH locus also acquire large templated events from discrete regions of the genome, (7) interchromosomal donor sequences are in physical proximity to the IgH locus during the germinal center reaction, (8) donor sequences are preferentially in areas of open chromatin and are in transcriptionally active genes during the germinal center reaction, and (9) sites of gene conversion in the recipient sequence correlate with the DSB-prone overlapping AID hotspot motif 5’-WGCW-3’[4,5]. Together, these findings provide strong evidence for a new paradigm of gene conversion during secondary somatic diversification of B cells in the germinal center. In addition to defining large intrachromosomal tracts of templated mutation using germline IgHV sequences, we show an unappreciated ability of gene conversion to utilize ectopic templates in trans. The expansion of the role of gene conversion is somewhat unsurprising in the larger context of DNA repair as studies have demonstrated the ability of gene conversion to occasionally utilize distant genomic regions for diversity[6,7]. In addition, the studies here may also point to the mechanism of "templated insertions" described in work by Lanzavecchia[8,9]. Nonetheless, the use of interchromosomal templates for *mutations*, is both significant and novel, and expands the role of gene conversion during somatic hypermutation.

With this in mind, how would gene conversion fit in the larger existing literature on somatic hypermutation? Mechanistically, somatic hypermutation is known to rely on AID, UNG, MSH2/6, EXO1, and polη[1,2]. Recall that activity by AID deaminates cytosine to uracil, which can then be processed by MSH2/6 and UNG to create a single strand nick that is extended to a ssDNA gap by the activity of EXO1, which is then filled in with "error-prone" polη. In the context of overlapping AID hotspots, the activity of AID, UNG, MSH2/6, and EXO1 can produce a DSB. Further, the activity of polη is known to be involved in the repair of DSBs[10].

In the DT40 chicken B cells, the presence of a DSB is sufficient to induce gene conversion, suggesting that this is the common pathway by which gene conversion would occur in murine and human B cells[11]. It has been demonstrated that DSB are generated with high frequency at the Ig loci during the somatic hypermutation program[12,13]. Curiously, polη has also been shown by Kawamoto et al. to be required for gene conversion in chicken DT40 B cells[14]. In their studies the loss of polη disrupted DSB-induced homologous recombination as well as Ig gene conversion, and this effect was reversed by complementation with *human* polη[14]. Two pivotal conclusions can be derived from these studies. First, polη is involved in normal gene conversion processes that are underway in chicken B cells. Second, the function of polη in both chickens and humans is conserved, since human polη rescued both the deficiency in DSB-induced HR (for which polη is known to function in human cells) as well as the deficiency in Ig gene conversion (for which the role of polη in humans is unclear). Thus, it is probable that at the Ig loci in human B cells, processing of a DSB, likely generated through the combined activities of AID, UNG, MSH2/6, and EXO1, could result in gene conversion as the key polymerase polη can execute a gene conversion event in a heterologous system. Indeed in chapter 3, we demonstrate that detected gene conversion events occurred primarily in the vicinity of predicted sites for DSBs.

The mutation frequency of motifs in SHM has been of great interest as understanding of what motifs are more or less likely to mutate could inform rational vaccine design. To that end, a large body of literature has been derived to analyze for patterns of somatic hypermutation as it relates to local sequence context. From this body of work, it was derived that AID preferentially targets 5'-WR$\underline{C}$Y-'3 motifs although additional motifs are enriched as well reflecting the activity of polη for which the hotspot is 5'-WA-'3[15-17]. Intriguingly, there are marked differences within these hotspots, with some motifs mutating more frequently than others and even the identical motif more likely to mutate depending on local position in the antibody sequence. Interestingly, further work extended motifs to 5- and 7-mers, yet still this only accounted for up to 80% of the variability in mutation profiles[18,19]. As it relates to the work in chapter 2, such models are unable to account for the observed linkage disequilibrium found between mutation in close proximity and further, the use of similar motifs for introduction of multiple mutations regardless of sequence identity. Thus, it remains a possibility that the variability found in these models may relate to effects from gene conversion donor choice.

Putting these two lines of research together, Wei et al. demonstrated that in human B cells that overlapping AID hotspots (5'-WGCW-3') are critical sites for diversification[5]. In their studies they found that mutation of these hotspots was correlated with the total number of mutations acquired in the V region. Furthermore, replacement of these hotspots with coldspots or neutral spots resulted in decreased numbers of mutations within those motifs as well as a global decrease in the number of mutations in the V region[5]. Thus, the mutability of these overlapping hotspots related to the mutability of the entire V region. As reviewed earlier, overlapping AID hotspots are sites prone to DSBs and as mentioned in chapter 3, templated events follow DSBs predicted to occur in these sites. Thus, the mutability of a singular motif can influence nonlocal mutability that is explained by the process of templated mutagenesis.

A model for gene conversion in SHM

As mentioned in Chapter 1, a theory for reverse transcription as a mechanism for SHM has been proposed throughout the last 30 years and has been championed by Edward Steele[20]. In his proposed model, AID deamination leads to invasion of pre-mRNA of the V-region that is subject to "error-prone" cDNA synthesis via polη. This is followed by resolution of the pre-mRNA intermediate from the DNA duplex and the replication of the error-filled cDNA. While I disagree with the model as presented, Steele makes a case that this model is superior to the Neuberger deamination model of SHM (reviewed in chapter 1) because, among other reasons, it incorporates the reverse transcriptase activity of polη[20]. Steele first reported the *in vitro* reverse transcriptase activity of polη in 2004, and this has been supported by emerging evidence from the Guengerich group[21-23]. Indeed, subsequent work by Su et al. demonstrated that human polη

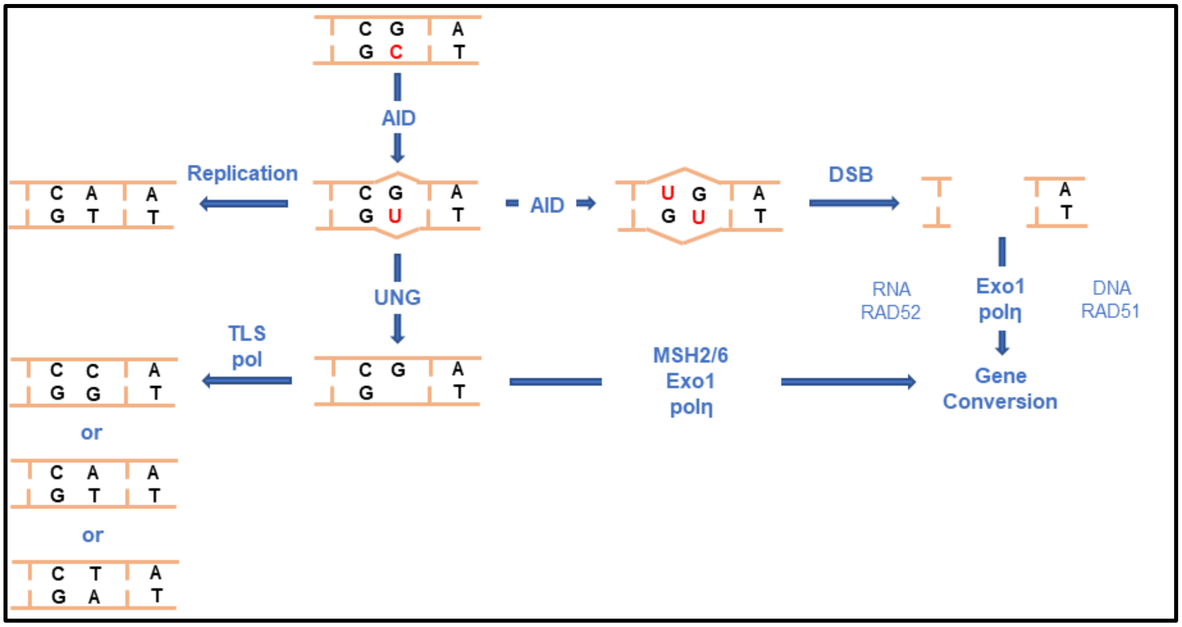


Figure 3: Updated model of somatic hypermutation and gene conversion

retains its reverse transcriptase activity in cellular environments[23]. As Steele argues reverse transcription plays a role in somatic hypermutation, the studies presented in chapter 3 indicate that both transcriptionally active and transcriptionally inactive regions are donor sequences for templated mutagenesis. Below, I will present a new model that marries the well-supported Neuberger model, with the reverse transcriptase activity of polη, and the newly emerging field of RNA-mediated DSB repair[24].

The model that best accommodates the current understanding of SHM-associated proteins and that of gene conversion is a modified Neuberger DNA deamination model[1,2]. In this model, the Neuberger model proceeds until the generation of a DSB. The DSB can either be repaired by RAD51-mediated homology search for homologous/homeologous DNA and be repaired by the actions of polη, or the lesion can be repaired by RAD52-mediated RNA invasion which forms a RNA/DNA hybrid. Experiments from the Guengerich group demonstrate that polη can use DNA as a primer to synthesize DNA from an RNA template through reverse transcription[22,23].

Synthesis of DNA is conducted such that the two ends of the DSB anneal and gaps can be filled by high fidelity polymerases.

Under this model, it is predicted that DSB locations are a hotspot for templated events. Indeed, supporting evidence can be found in chapter 3 as well as in Pieper et al. who demonstrated that templated insertions of B cell transcribed genes occurs preferentially in the switch regions – which readily undergo DSBs. Furthermore, it has been shown that DSBs are readily mutagenic in non-B cells, suggesting their central role in generating diversity.

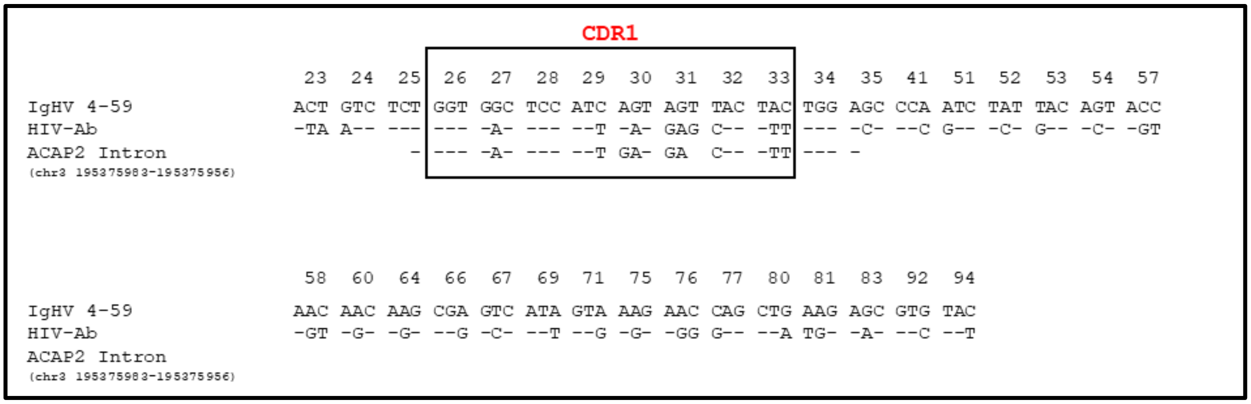
With respect to the role of polη in patch synthesis under the Neuberger model of SHM, it is likely that this activity too is somehow mediated by a templated mechanism. The activity of polη is intrinsically tied to the A:T mutation spectrum[25]. In DT40 cells, ablation of XRCC2 and XRCC3 results in both the loss of gene conversion as well as selective loss of A:T mutations[26]. This suggests that in DT40 B cells, gene conversion is critically dependent for the generation of A:T mutations, a hypothesis that is further supported by the role of polη in facilitating gene conversion[14]. Polη has also been demonstrated to have a role in D-loop extension, suggesting that the "error-prone" mechanism of action is through templated mutagenesis[27]. Further, based on the studies in chapter 2, it is unlikely that all linked mutations occur in the vicinity of the overlapping AID hotspot motifs. Therefore, there must be some mechanism by which A:T mutations in site that are not proximal to overlapping AID hotspots acquired linked mutations via polη, in a probable templated mutagenesis event. At present such a mechanism is elusive.

Gene conversion and potential for vaccine design

The pinnacle of achievement in the field of SHM would likely be harnessing the mechanisms associated with SHM to produce a specific antibody with key desired mutations. Indeed, much work has been done in designing vaccination strategies to elicit the broadly-neutralizing antibodies (bNAbs) in HIV and Influenza[28-30]. Such a feat would revolutionize both immunology and to a larger degree, medicine. The antibody has been fundamental in the humoral immune system as well as in the clinic. While Honjo and Allison won the Nobel prize for their discovery of PD-1 and CTLA-4, in the clinic it isn't the proteins themselves that are responsible for treating cancer. Antibody-based drugs, all of which end in -mab to indicate that they are monoclonal antibodies, are the effector molecules that modify the behaviors of these proteins and facilitates treatment. Thus, harnessing the ability of antibody diversity in a predicable and stereotyped way, would be a zenith.

The work contained in this dissertation helps to understand cryptic mechanisms at work during the course of the germinal center response and the somatic hypermutation program. Indeed, the understanding that mutations in micro-clusters have a preference to match specific germline motifs at the IgH locus, reveals that certain pairs of mutations that may be required for generating a high affinity antigen specific antibody response to a neutralizing epitope may be unfavorable. Likewise, based on the data contained in chapter 3, large scale templated events appear to be largely restricted to CDR1 and FWR3, indicating that patterns of mutation at these regions may be more frequently derived from discrete regions of the genome and in some respects may be predictable. Integration of these findings as it relates to affinity maturation in the germinal center would be largely informative of how these dynamics occur in vivo. Furthermore, based on the work presented in chapter 3, it may be possible to skew germinal center B cell transcriptional profiles to facilitate selection of particular templates for DSB repair and gene conversion.

Additional work should be directed to analysis of antigen-specific antibodies and conducting analysis of these selected antibodies to detect patterns for binding. In unpublished



```
                                                CDR1
                       23  24  25  26  27  28  29  30  31  32  33  34  35  41  51  52  53  54  57
IgHV 4-59             ACT GTC TCT GGT GGC TCC ATC AGT AGT TAC TAC TGG AGC CCA ATC TAT TAC AGT ACC
HIV-Ab                -TA A-- --- --- -A- --- --T -A- GAG C-- -TT --- -C- --C G-- -C- G-- -C- -GT
ACAP2 Intron                  -   --- -A- --- --T GA- GA  C-- -TT --- -
(chr3 195375983-195375956)

                       58  60  64  66  67  69  71  75  76  77  80  81  83  92  94
IgHV 4-59             AAC AAC AAG CGA GTC ATA GTA AAG AAC CAG CTG AAG AGC GTG TAC
HIV-Ab                -GT -G- -G- --G -C- --T --G -G- -GG G-- --A TG- -A- --C --T
ACAP2 Intron
(chr3 195375983-195375956)
```

Figure 4: Ectopic gene conversion in CDR1 of HIV-specific antibody

work, our computational script TRACE was able to identify a highly mutated region of CDR1 in an antibody specific for HIV for which the predicted template was an intron of a gene that had not appeared in the studies in chapter 3. This donor accounted for 8 mutations in this region, and spanned the length of CDR1. If this can occur in terminal, antigen-specific, desirable antibodies, it is paramount to understand if specific genomic regions are key to antigen binding.

The studies presented in this dissertation, represent a first step in applying our understanding of gene conversion in humoral immunity to further human health. Further work will characterize the roles of gene conversion in antigen-specific contexts, and yet further work will seek to understand the principles by which gene conversion can be directed. Such work will fundamentally lead to new horizons in vaccinology and ultimately will lead to firmer understanding on our targeting of the humoral immune system to human pathogens.

Works Cited

1 Maul, R. W. & Gearhart, P. J. Refining the Neuberger model: Uracil processing by activated B cells. *Eur J Immunol* **44**, 1913-1916, doi:10.1002/eji.201444813 (2014).

2 Methot, S. P. & Di Noia, J. M. Molecular Mechanisms of Somatic Hypermutation and Class Switch Recombination. *Adv Immunol* **133**, 37-87, doi:10.1016/bs.ai.2016.11.002 (2017).

3 Mesin, L., Ersching, J. & Victora, G. D. Germinal Center B Cell Dynamics. *Immunity* **45**, 471-482, doi:10.1016/j.immuni.2016.09.001 (2016).

4 Han, L., Masani, S. & Yu, K. Overlapping activation-induced cytidine deaminase hotspot motifs in Ig class-switch recombination. *Proc Natl Acad Sci U S A* **108**, 11584-11589, doi:10.1073/pnas.1018726108 (2011).

5 Wei, L. *et al.* Overlapping hotspots in CDRs are critical sites for V region diversification. *Proc Natl Acad Sci U S A* **112**, E728-737, doi:10.1073/pnas.1500788112 (2015).

6 Anand, R. P. *et al.* Chromosome rearrangements via template switching between diverged repeated sequences. *Genes Dev* **28**, 2394-2406, doi:10.1101/gad.250258.114 (2014).

7 Tsaponina, O. & Haber, J. E. Frequent Interchromosomal Template Switches during Gene Conversion in S. cerevisiae. *Mol Cell* **55**, 615-625, doi:10.1016/j.molcel.2014.06.025 (2014).

8 Pieper, K. *et al.* Public antibodies to malaria antigens generated by two LAIR1 insertion modalities. *Nature* **548**, 597-601, doi:10.1038/nature23670 (2017).

9 Tan, J. *et al.* A LAIR1 insertion generates broadly reactive antibodies against malaria variant antigens. *Nature* **529**, 105-109, doi:10.1038/nature16450 (2016).

10 Sebesta, M. & Krejci, L. in *DNA Replication, Recombination, and Repair: Molecular Mechanisms and Pathology* (eds Fumio Hanaoka & Kaoru Sugasawa) 73-109 (Springer Japan, 2016).

11 Bastianello, G. & Arakawa, H. A double-strand break can trigger immunoglobulin gene conversion. *Nucleic Acids Res* **45**, 231-243, doi:10.1093/nar/gkw887 (2017).

12 Bross, L. *et al.* DNA double-strand breaks in immunoglobulin genes undergoing somatic hypermutation. *Immunity* **13**, 589-597 (2000).

13 Rush, J. S., Fugmann, S. D. & Schatz, D. G. Staggered AID-dependent DNA double strand breaks are the predominant DNA lesions targeted to S mu in Ig class switch recombination. *Int Immunol* **16**, 549-557, doi:10.1093/intimm/dxh057 (2004).

14 Kawamoto, T. *et al.* Dual roles for DNA polymerase eta in homologous DNA recombination and translesion DNA synthesis. *Mol Cell* **20**, 793-799, doi:10.1016/j.molcel.2005.10.016 (2005).

15 Cui, A. *et al.* A Model of Somatic Hypermutation Targeting in Mice Based on High-Throughput Ig Sequencing Data. *J Immunol* **197**, 3566-3574, doi:10.4049/jimmunol.1502263 (2016).

16 Yaari, G. *et al.* Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. *Front Immunol* **4**, 358, doi:10.3389/fimmu.2013.00358 (2013).

17 Rogozin, I. B. & 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* **172**, 3382-3384 (2004).

18 Schramm, C. A. & Douek, D. C. Beyond Hot Spots: Biases in Antibody Somatic Hypermutation and Implications for Vaccine Design. *Front Immunol* **9**, 1876, doi:10.3389/fimmu.2018.01876 (2018).

19 Elhanati, Y. *et al.* Inferring processes underlying B-cell repertoire diversity. *Philos Trans R Soc Lond B Biol Sci* **370**, doi:10.1098/rstb.2014.0243 (2015).

20 Steele, E. J. Reverse Transcriptase Mechanism of Somatic Hypermutation: 60 Years of Clonal Selection Theory. *Front Immunol* **8**, 1611, doi:10.3389/fimmu.2017.01611 (2017).

21 Franklin, A., Milburn, P. J., Blanden, R. V. & Steele, E. J. Human DNA polymerase-eta, an A-T mutator in somatic hypermutation of rearranged immunoglobulin genes, is a reverse transcriptase. *Immunol Cell Biol* **82**, 219-225, doi:10.1046/j.0818-9641.2004.01221.x (2004).

22 Su, Y., Egli, M. & Guengerich, F. P. Human DNA polymerase eta accommodates RNA for strand extension. *J Biol Chem* **292**, 18044-18051, doi:10.1074/jbc.M117.809723 (2017).

23 Su, Y. *et al.* Human DNA polymerase eta has reverse transcriptase activity in cellular environments. *J Biol Chem* **294**, 6073-6081, doi:10.1074/jbc.RA119.007925 (2019).

24 McDevitt, S., Rusanov, T., Kent, T., Chandramouly, G. & Pomerantz, R. T. How RNA transcripts coordinate DNA recombination and repair. *Nat Commun* **9**, 1091, doi:10.1038/s41467-018-03483-7 (2018).

25 Delbos, F., Aoufouchi, S., Faili, A., Weill, J. C. & Reynaud, C. A. DNA polymerase eta is the sole contributor of A/T modifications during immunoglobulin gene hypermutation in the mouse. *J Exp Med* **204**, 17-23, doi:10.1084/jem.20062131 (2007).

26 Sale, J. E., Calandrini, D. M., Takata, M., Takeda, S. & Neuberger, M. S. Ablation of XRCC2/3 transforms immunoglobulin V gene conversion into somatic hypermutation. *Nature* **412**, 921-926, doi:10.1038/35091100 (2001).

27 McVey, M., Khodaverdian, V. Y., Meyer, D., Cerqueira, P. G. & Heyer, W. D. Eukaryotic DNA Polymerases in Homologous Recombination. *Annu Rev Genet* **50**, 393-421, doi:10.1146/annurev-genet-120215-035243 (2016).

28 Nabel, G. J. & Fauci, A. S. Induction of unnatural immunity: prospects for a broadly protective universal influenza vaccine. *Nat Med* **16**, 1389-1391, doi:10.1038/nm1210-1389 (2010).

29 Burton, D. R. & Hangartner, L. Broadly Neutralizing Antibodies to HIV and Their Role in Vaccine Design. *Annu Rev Immunol* **34**, 635-659, doi:10.1146/annurev-immunol-041015-055515 (2016).

30 Haynes, B. F. & Mascola, J. R. The quest for an antibody-based HIV vaccine. *Immunol Rev* **275**, 5-10, doi:10.1111/imr.12517 (2017).