

Figure S1

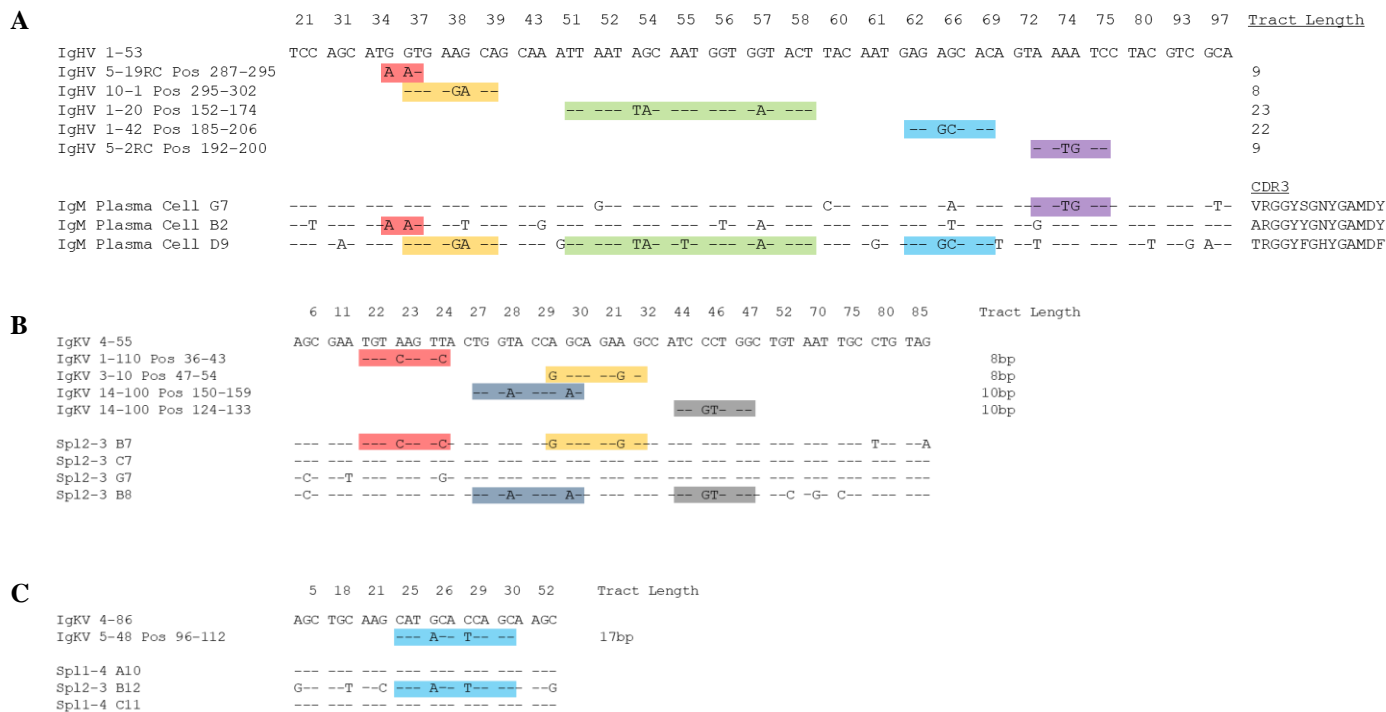
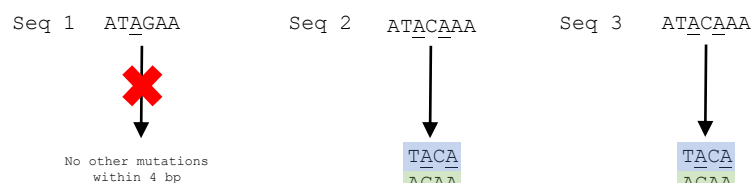


Fig. S1: Templated tracts of mutations are observed in sanger sequenced IgM plasma cells. (A) Nucleotide alignment of somatically-mutated IgM plasma cell sequences to germline IgHV 1-53. **(B-C)** Nucleotide alignment of somatically-mutated IgKV sequences from IgM plasma cells to their respective germline sequence. Data is presented as in Figure 1C.

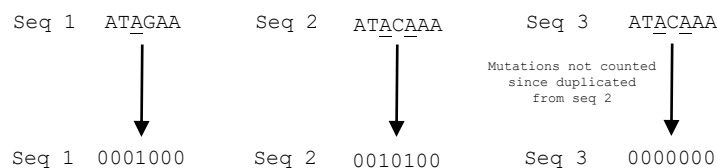
Alignment of Mutated Sequences with Germline

G.L. ³ ⁴ ⁷
 Seq 1 AT-AGAA
 Seq 2 ATACAAA
 Seq 3 ATACAAA

Motif Generation (unaligned)



Mutation Position (aligned)



Reference Query

Ref 1 ATACAC Ref 1 ATACAC
 Ref 2 CTAAAC Ref 2 CTAAAC
 Ref 3 AGACAA Ref 3 AGACAA

Scoring (aligned)

Seq 1 0000000 Seq 2 0111110 Seq 3 0111110

↓

Seq 1 0000000 Seq 2 0111110 Seq 3 0111110

Mutations Represented in Reference

Compared for "1" in each cell

Seq 1 0
 Seq 2 2
 Seq 3 0

Calculation of Gene Conversion Coverage

Mutations represented divided by total mutations analyzed

100%

Only finite values counted (i.e. 0/0 is excluded)

Mutations Analyzed

Must have >1 mutation in 4bp

Seq 1 0001000
 Seq 2 0010100
 Seq 3 0000000

↓

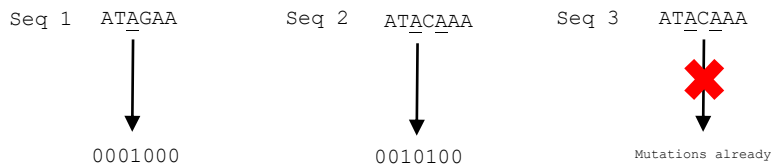
0
 2
 0

Fig. S2: 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.

Alignment of Mutated Sequences with Germline

G.L. ATGCGAA
 Seq 1 AT-AGAA
 Seq 2 ATACAAA
 Seq 3 ATACAAA

Determine Mutations to Change



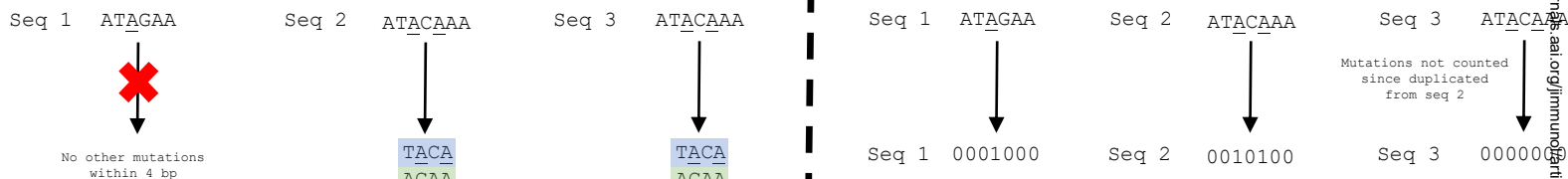
Assign new mutations to model SHM

G.L. ATGCGAA
 Seq 1 AT-NGAA
 Seq 2 ATNCNAA
 Seq 3 ATACAAA

"N" is determined by base pair substitution matrix (example at right)

To \ From	A	C	G	T
A	21%	49%	30%	
C	20%	16%	64%	
G	58%	25%		17%
T	28%	56%	17%	

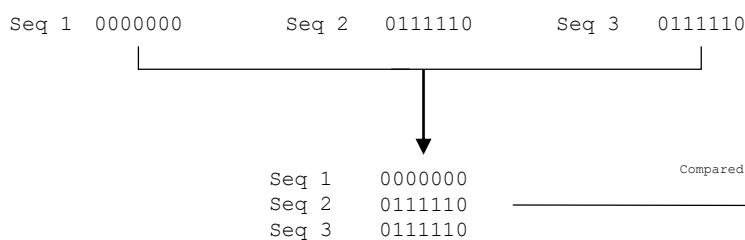
Motif Generation (unaligned)



Reference Query

Ref 1 ATACAC
 Ref 2 CTAAAC
 Ref 3 AGACAA

Scoring (aligned)



Mutations Represented in Reference

Seq 1 0
 Seq 2 2
 Seq 3 0

Mutations Analyzed
 Must have >1 mutation 4bp

Seq 1 0001000
 Seq 2 0010100
 Seq 3 0000000

Calculation of Gene Conversion Coverage

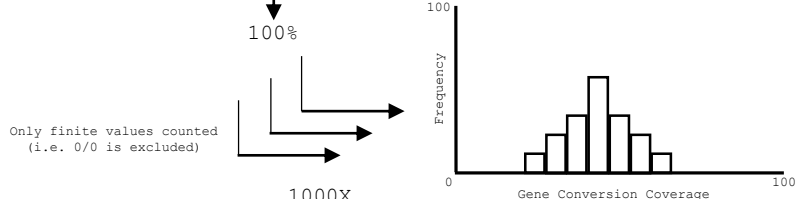
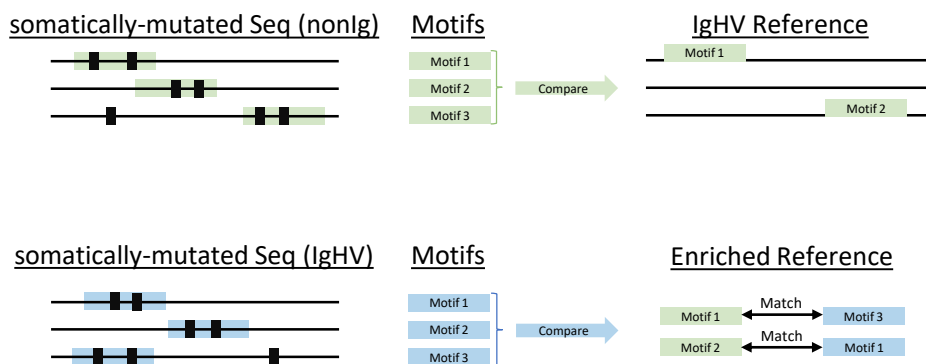


Fig. S3: Schematic of sequence analysis via RandomCheck. Depicted are three somatically-mutated sequences aligned to a reference gene. As in PolyMotifFinder, the positions of mutations are generated in a numeric array with duplicate mutation pairs being removed from the final array. Next, RandomCheck 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.

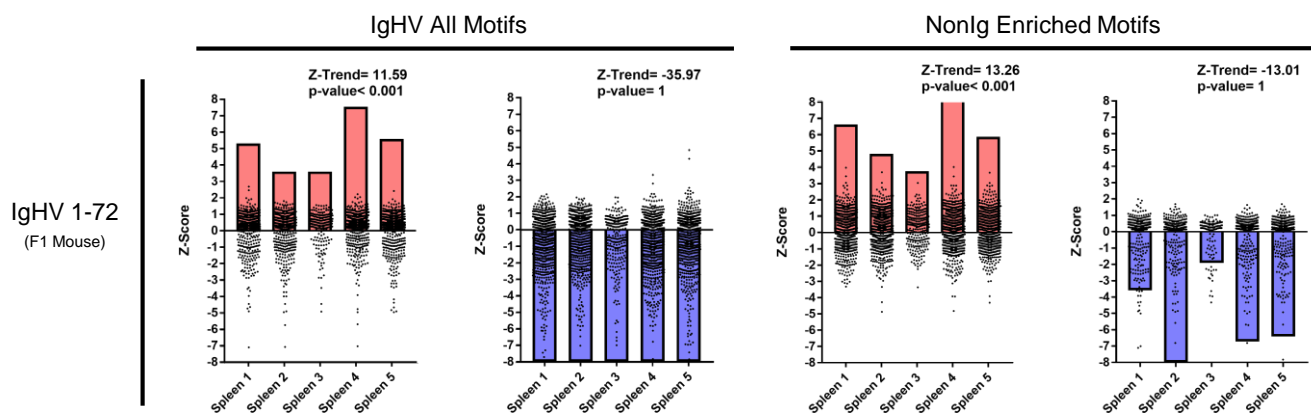
Downloaded from http://journals.sagepub.com/journalInfo/journals/immunology/article-supplement/107497/pdf/ji_1801615_supplemental_figures_11 by guest on 17 Jul 2024

Figure S4

A



B



C

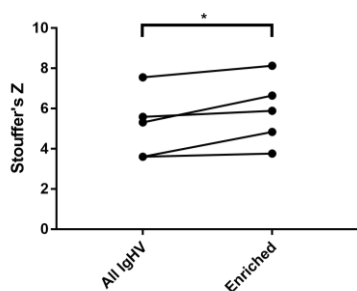


Fig. S4: IgHV genes from F1 germinal centers are enriched for motifs occurring in the IgHV repertoire and the subset of the repertoire that somatically-mutated GPT and β -globin match to.

(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.