bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Somatic hypermutation spectra are independent of the local transcriptional and epigenetic landscape

Ursula E. Schoeberl[1]*, Johanna Fitz[1]*, Kimon Froussios[1], Renan Valieris[2], Marina Makharova[1], Iordanis Ourailidis[1], Bernd Bauer[1], Tobias Neumann[1], Eva-Maria Wiedemann[1], Monika Steininger[1], Adriana Cantoran Garcia[1], Marialaura Mastrovito[1], Hugo Mouquet[3], Israel Tojal Da Silva[2] and Rushad Pavri[1#]

(1) Research Institute of Molecular Pathology (IMP), Campus-Vienna-Biocenter 1, 1030 Vienna, Austria
(2) Laboratory of Bioinformatics and Computational Biology, A. C. Camargo Cancer Center, São Paulo, SP, Brazil
(3) Laboratory of Humoral Immunology, Institut Pasteur, INSERM U1222, Université Paris Cité, 75015 Paris, France

* equal contribution

corresponding author (rushad.pavri@imp.ac.at)

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Somatic hypermutation (SHM) of immunoglobulin variable regions in B cells modulates antibody-antigen affinity and is indispensable for adaptive immunity. Mutations are introduced by activation-induced cytidine deaminase (AID) in a co-transcriptional manner resulting in discrete mutation spectra. Current models propose that activating epigenetic marks, transcriptional pausing and convergent transcription are necessary for optimal AID recruitment. However, whether these or other transcriptional features can explain the discrete mutation spectra is unknown. To address this, we compared mutation and nascent transcription at single nucleotide resolution. Surprisingly, with this precision, SHM spectra do not correlate with any transcriptional feature at human and mouse variable regions and non-immunoglobulin AID targets. Moreover, SHM is resistant to up to four-fold reduction of both activating epigenetic marks and transcription. We propose that, following AID recruitment to its target genes, the DNA sequence flanking an AID target motif is the key determinant of mutability rather than the local transcriptional and chromatin landscape.

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

Somatic hypermutation (SHM) is the molecular basis for the diversification of antibodies in response to pathogens and vaccines and is hence indispensable for robust long-term immunity (*1-3*). SHM occurs in activated B cells within microanatomical structures called germinal centers in secondary lymphoid tissue upon engagement with antigens and helper T cells (*4*). Mutations are generated in the variable regions of the immunoglobulin (IG) heavy (*IGH*) and light chain (*IGK*, *IGL*) genes by activation-induced cytidine deaminase (AID) (*5, 6*) which acts co-transcriptionally on single-stranded DNA (ssDNA) (*7-11*). AID also acts at the *IGH* switch regions triggering a reaction cascade leading to class switch recombination (CSR) which yields the various antibody isotypes essential for effector functions following antigen binding (*12-15*). In addition, SHM occurs, albeit with much lower frequency, at several transcribed non-IG genes (*16-20*).

An important and longstanding question in the field is the relationship of SHM with nascent transcription and activating epigenetic marks. AID associates with the RNA polymerase II (Pol II) machinery via interaction with various elongation (*21-25*) and RNA processing factors (*26, 27*). However, since these are generic gene regulatory proteins, such interactions cannot per se explain the specificity of AID targeting. SHM has also been correlated with the epigenetic landscape especially with activating histone marks. In particular, depletion of histone modifying enzymes resulted in decreased SHM and/or CSR accompanied with a decrease in activating histone modifications (*24, 28-36*). However, the precise role of these marks in SHM is unknown. Moreover, these marks are typical of any transcribed gene hence their presence per se cannot confer specificity. Instead, it appears that specificity is conferred by super-enhancers associated with IG and other AID target loci (*37-40*). The murine *Igh* super-enhancer is essential for SHM (*41*) and is itself a target of AID activity (*42*). Importantly, an AID-targeted enhancer when placed in a genomic domain that lacks AID target genes was able to convert these genes into SHM hotspots, thus providing compelling evidence for the critical role of enhancers in targeting AID (*38*).

The final step of SHM, that is, the actual mutagenesis, is least understood. AID preferentially deaminates cytidine residues within WR<u>C</u>H motifs (where W = A or T, R = A or G and H = A, C or T) (*43-45*). However, not all WR<u>C</u>H residues are targeted and the mutation frequencies of those that are

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

mutated can differ substantially (*46-52*). Remarkably, even identical WRCH motifs within a given variable region can differ significantly in mutation frequency (*47-52*). This differential mutability is a ubiquitous feature of SHM and implies the existence of as-yet-unknown mechanisms that determine which residue is mutated and to what extent. More importantly, differential mutability cannot be explained by the association of AID with Pol II complexes, but rather, suggests that access to ssDNA may be biased towards certain sequence contexts. In this regard, features of nascent transcription that can cause transient ssDNA exposure, such as convergent transcription (*53*), transcriptional pausing (*17, 22, 54-57*), initiation (*58*) and transcription termination (*59, 60*), have been correlated with SHM and CSR. Moreover, ssDNA patches are detected in variable regions and suggested to arise via negative supercoiling upstream of transcribing Pol II, although these patches do not correlate with SHM patterns (*56, 57, 60, 61*). Indeed, high occupancy of Pol II and SPT5, a Pol II pausing factor implicated in AID targeting (*22, 56, 57*), was able to identify new AID target genes via a machine learning approach (*17*). Based on these observations, a reasonable hypothesis is that differential mutability of WRCH motifs arises because of local sites of transcription initiation, pausing or convergent transcription which favor AID activity at these sub-regions within the variable region. Since variable regions are very short (<500 bp), answering this question requires high-resolution nascent transcriptional maps of variable regions, which are currently unavailable.

In this study, we aimed to address the precise relationship between SHM, nascent transcriptional features and epigenetic marks using a fresh and direct approach. To do so, we performed precision run-on sequencing (PRO-seq and PRO-cap) (*62*) in Ramos human B cells and murine germinal center B cells to obtain single-nucleotide resolution maps of nascent transcription at multiple human and mouse variable regions as well as at a large collection of non-IG AID target genes (*17*). We combined this with high-resolution mutational profiling to show that patterns of SHM do not consistently correlate with any nascent transcriptional features. Moreover, via deletion of the *IGH* Eμ enhancer in human B cells, we find that SHM is insignificantly affected despite up to four-fold reduction in transcription and active chromatin marks. Importantly, the same variable region sequence, studied in mouse or human regulatory contexts, exhibits a nearly identical SHM pattern. The results strongly suggest that following AID recruitment to the

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

variable gene locale, the SHM spectrum is determined by the DNA sequence flanking the hotspot residues rather than any co-transcriptional or epigenetic feature.

Results

The human and mouse *IGH* variable region locale harbors relatively low levels of histone acetylation and H3K4 trimethylation

As a model system of SHM, we chose the Ramos human B cells. Ramos is an IgM-positive Burkitt lymphoma-derived cell line where the single functional, pre-recombined *IGH* variable region consists of the variable gene, VH4-34, the diversity segment, DH3-10, and the joining segment, JH6. Ramos cells constitutively express AID and undergo very low levels of SHM mostly at C:G residues (*63*).

To determine the chromatin landscape of the Ramos *IGH* locus, we performed chromatin immunoprecipitation followed by quantitative PCR (ChIP-qPCR) for histone modifications associated with active promoters and enhancers (histone H3 acetylated on lysine 27, H3K27ac or trimethylated on lysine 3, H3K4me3) and transcription elongation in gene bodies (H3 trimethylated on lysine 36, H3K36me3). At active genes, H3K27ac and H3K4me3 are enriched at nucleosomes flanking the transcription start site. Surprisingly, we observed that the variable region is characterized by relatively low levels of all marks that, importantly, was not due to decreased nucleosome occupancy as judged by the levels of histone H3 (Fig. S1A). As a result, the first nucleosome from the transcription start site has the lowest levels of H3K27ac and H3K4me3 which is contrary to what is normally observed at highly active genes where the first nucleosome harbors the highest levels of these marks. In contrast, much higher levels of all marks were observed in the intronic regions flanking the Eμ enhancer with an expected decrease at Eμ itself where nucleosomes are occluded by the presence of transcription factors (Fig. S1A) (*64*). Importantly, the same profiles were observed at the murine *Igh* in activated, primary B cells expressing the B1-8^{hi} variable region (*65*) (Fig. S1B). We conclude that variable regions are marked by relatively low levels of activating histone marks and that this landscape is conserved between the human Ramos and mouse B1-8^{hi} variable regions.

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The Eμ enhancer regulates the levels of chromatin marks in the flanking intronic DNA but not within the variable region

The fact that the highest enrichments of chromatin marks were seen on either side of the Eμ enhancer suggested that this element may be responsible for the deposition of these marks. This raised the question of whether the loss of Eμ would alter the levels of these marks and affect SHM. To minimize effects of ongoing SHM on the variable gene sequence, we first ablated AID in Ramos cells ($AID^{-/-}$) and subsequently deleted Eμ in $AID^{-/-}$ cells ($AID^{-/-}$ $E\mu^{-/-}$) (Fig. S1C). Specifically, by CRISPR-mediated editing, we deleted a 583 bp region corresponding to the peak of chromatin accessibility measured by assay for transposase-accessible chromatin (ATAC-seq) signal at Eμ in $AID^{-/-}$ Ramos cells ($AID^{-/-}$ $E\mu^{-/-}$) (Fig. S1C). Based on PRO-cap analysis of transcription start sites (see below), this deletion eliminates the Iμ promoter that encodes the sterile germline *IGHM* transcript and antisense transcription start sites residing in Eμ (Fig. S1C). We chose two independent clones, $AID^{-/-}$ $E\mu^{-/-}$ c1 and $AID^{-/-}$ $E\mu^{-/-}$ c2, for subsequent experiments (Fig. S1D).

To comprehensively visualize profiles of these histone marks across the variable region and flanking sequences, we performed ChIP followed by deep sequencing (ChIP-seq) for H3K27ac, H3K4me3 and H3K36me3. We generated a custom chromosome spanning the Ramos variable region and promoter. Importantly, the Ramos variable region is highly mappable and hence there is minimal loss of reads due to multimapping with other variable gene families (see later section). The results in $AID^{-/-}$ Ramos cells were consistent with ChIP-qPCR in that the levels of all marks were lowest in the promoter-proximal region and gradually increased into the variable region, peaking in the intronic sequences flanking Eμ (Fig. 1A-B). This analysis also revealed the absence of bimodal distributions of H3K27ac and H3K4me3 at the VH4-34 promoter, which is commonly observed at highly active divergently transcribed genes. In $AID^{-/-}$ $E\mu^{-/-}$ cells, all three marks were decreased (Fig. 1A-B), a finding that we confirmed by ChIP-qPCR and which was not due to changes in nucleosome occupancy as determined by histone H3 measurements (Fig. 1C). Using antibodies against pan-H3 and pan-H4 acetylation, we observed a significant decrease in histone acetylation across the locus in $AID^{-/-}$ $E\mu^{-/-}$ cells (Fig. 1C). Moreover,

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

another mark of transcription elongation, H3K79me3, was also reduced in AID$^{-/-}$ Eμ$^{-/-}$ cells (Fig. 1C). All marks in AID$^{-/-}$ Eμ$^{-/-}$ cells were significantly reduced up to 4-fold in the intron where they are normally at their peak levels (Fig. 1A-C). However, within the variable region, where these marks are normally at their lowest levels, there was no significant change in AID$^{-/-}$ Eμ$^{-/-}$ cells Fig. 1A-C). Consequently, AID$^{-/-}$ Eμ$^{-/-}$ cells harbor a chromatin landscape wherein the levels of all marks appear to be comparable between the variable region and the intron (Fig. 1C, compare the measurements from amplicons 2, 3 and 5 in AID$^{-/-}$ and AID$^{-/-}$ Eμ$^{-/-}$ cells). We conclude that Eμ is responsible for the higher levels of chromatin marks within the intron but has no significant influence on the deposition of these marks within the variable region.

These changes in the chromatin landscape in AID$^{-/-}$ Eμ$^{-/-}$ cells were accompanied by a 2-4-fold decrease of nascent transcription in the variable region and intron, as well as a 2-fold decrease of the spliced IgM mRNA (Fig. 2A-B). The decrease in transcription within the variable region and intron implies that Eμ regulates optimal transcription initiation from the variable gene promoter. Nevertheless, despite these decreases in transcription, surface IgM levels were unaffected in Eμ$^{-/-}$ cells suggesting a lack of absolute correlation between IgM mRNA and protein levels (Fig. S1E). We conclude that Eμ positively and significantly regulates nascent transcription in the variable region but has no significant impact on its epigenetic landscape.

SHM frequency does not correlate with the levels of nascent transcription or activating histone marks in the variable region locale

AID$^{-/-}$ Eμ$^{-/-}$ cells provide an ideal system to ask whether changes in the levels of nascent transcription and chromatin marks directly impact on SHM without any contribution from secondary, indirect effects. Hence, we sequenced the variable region as well as the JH6 intron immediately downstream (Fig. 2C, upper panel). This analysis allowed us to compare SHM between a region where chromatin marks were normally low and Eμ-independent (variable region) and a region where the marks were normally high and strongly Eμ-dependent (JH6 intron). To measure SHM, we infected AID$^{-/-}$ and AID$^{-/-}$ Eμ$^{-/-}$ cells with JP8Bdel, a C-terminal truncation mutant of AID that results in nuclear retention and major increase in

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

mutation rates (*66*). As a control, we also infected cells with a catalytically inactive AID mutant (E58Q) (*66*). In both the variable region and JH6 intron, we observed a small (~20%) decrease in SHM frequency in AID$^{-/-}$ Eμ$^{-/-}$ relative to AID$^{-/-}$ JP8Bdel-expressing cells that was not statistically significant (Fig. 2C). This result is consistent with mouse studies where loss of Eμ caused only modest decreases in SHM (*67, 68*). As an alternative readout of SHM activity, we assayed for the loss of surface IgM, which occurs due to AID-induced stop codons or frameshifts (*63*). The results showed that the magnitude of IgM loss was similar between the control and Eμ$^{-/-}$ cells suggesting that the minor decreases in SHM do not translate into equivalent changes in IgM loss (Fig. 2D-E). We conclude that the Eμ enhancer is dispensable for robust SHM in Ramos cells.

By comparing the levels of chromatin marks with SHM frequencies in the variable region and JH6 intron, we can draw three major conclusions: First, we infer that the levels of SHM do not correlate with those of activating histone marks. Indeed, SHM frequency is slightly lower in the intron (1.6×10^{-3}), which harbors the highest level of these marks, than in the variable region (1.86×10^{-3}), which has the lowest levels of these modifications (Fig. 2C). Second, as exemplified by the JH6 intron, SHM is resistant to major decreases in these marks. Third, we attribute the small reduction in SHM in AID$^{-/-}$ Eμ$^{-/-}$ cells to the decrease in nascent transcription although, here too, the decrease of SHM (~20%; Fig. 2C) is considerably smaller than that of transcription which shows up to 70% loss in the JH6 intron of AID$^{-/-}$ Eμ$^{-/-}$ cells (Fig. 2A, amplicons 2-5). We conclude that although our results do not rule out a role for these activating marks in SHM, they do clearly demonstrate that their presence, profiles and even their significant reduction is not a predictor of SHM.

High-resolution transcriptomic profiling of different human variable regions reveals the absence of direct relationships between nascent transcriptional features and SHM spectra

A major, unsolved problem in SHM biology is the differential mutability of WRCH motifs within variable regions which implies the existence of local, sequence-intrinsic mechanisms regulating AID activity post-recruitment (*46-48, 50-52*). We asked whether SHM spectra are dictated by local features of nascent

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

transcription landscape that could transiently lead to increased ssDNA exposure, such as during transcription initiation, pausing and/or convergent/antisense transcription.

Since V genes occur as families with varying degrees of sequence identity, mapping of next-generation sequencing (NGS) reads to recombined variable regions can be problematic because standard alignment workflows eliminate reads that map to more than one location in the genome (multimappers) and only align reads mapping uniquely (typically allowing 2-3 mismatches). In Ramos cells, visual inspection of PRO-seq tracks on the UCSC genome browser revealed robust coverage within the VH4-34 gene with uniquely-mapping reads indicating that this V gene is highly mappable (not shown). Moreover, there were no unique reads mapping to other V gene families suggesting that they are transcriptionally silent in these cells. To address this more systematically, we retrieved the normally discarded, multimapping reads and allowed them to re-align to the recombined Ramos variable region with the important proviso that they only multimap to annotated V, D or J segments and not elsewhere in the genome (Fig. 3A). Separate tracks were created for uniquely mapping reads and the recalled multimapping reads, which were then combined to obtain the total profile (Fig. 3B).

We measured nascent transcriptional activity with PRO-seq, which maps the location and orientation of actively engaged RNA polymerase II (Pol II) at single-nucleotide resolution, and the related method, PRO-cap, which maps transcription initiation sites of capped, nascent RNAs (*62*). Given that variable regions are short (<500 bp), the single-nucleotide resolution of PRO-seq is ideally suited to identify potential sites of pausing, as well as antisense and convergent transcription within variable regions. For simplicity, we use the term “pausing” to refer to accumulation of PRO-seq signal in variable regions although it must be noted this refers to the slower-moving, elongating Pol II molecules and is not equivalent to promoter-proximal pausing which is also reported by PRO-seq. Most notably, promoter-proximal pausing occurs via the binding of NELF which inhibits elongation whereas elongating Pol II lacks NELF and contains additional proteins such as SPT6 and the PAF complex (*69-72*). Accumulation of Pol II signal in gene bodies may arise due to the slowing down of Pol II as a result of steric barriers such as secondary structures or premature termination signals. However, initiation events within gene bodies

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

could give rise to a paused Pol II state akin to promoter-proximal pausing and can be identified if PRO-seq accumulation is associated with upstream PRO-cap enrichments.

Our mapping pipeline revealed that the Ramos variable region is predominantly covered by uniquely-mapping reads with very few multimapping reads (Fig. 3B). Of note, the same pipeline was used for the ChIP-seq analysis of Ramos cells described in Fig. 1 above. Therefore, this approach allowed us, for the first time, to delineate the transcriptional and epigenetic landscapes of IG variable regions. In contrast to the lack of a bimodal peak of activating histone marks (Fig. 1A), the Ramos VH4-34 promoter showed clear bidirectional transcription and promoter-proximal pausing, the latter being noticeable as the PRO-seq enrichments shortly downstream of the promoter-associated PRO-cap initiation signals (Fig. 3C). Increased PRO-seq read density was observed towards the 3' end of the variable region suggestive of slowly-elongating or paused Pol II. However, the rest of the variable region harbored relatively low PRO-seq signals (Fig. 3B and magnified view in Fig. 3C). Next, we infected cells with AID(JP8Bdel) and analyzed SHM patterns in the variable region using an optimized version of mutation analysis with paired-end deep sequencing (MutPE-seq) of a PCR-amplified fragment corresponding to the variable region (*73*) (Fig. 3E). Mutations occurred robustly at C:G residues within WR<u>C</u>H motifs with weak targeting of A:T residues, as expected from Ramos cells. Moreover, the cold spot motif, SY<u>C</u> (where S = C or G and Y = C or T) was poorly mutated (Fig. 3F-H). However, there was no discernible overlap of the mutation profile with that of nascent transcription. In particular, the region of highest PRO-seq read density at the 3' end of the variable region is not preferentially mutated over upstream C:G motifs where PRO-seq signal is much weaker (Fig. 3D-E). Indeed, the most mutated residue is located near the 5' end of the variable region where nascent transcription signals are lower (Fig. 3D). We conclude that there appears to be no discernible feature of the nascent transcriptional landscape that can account for the observed frequency or distribution of mutations.

PRO-seq in AID$^{-/-}$ E$\mu^{-/-}$ cells revealed that nascent sense transcription was uniformly decreased across the variable region and intron (Fig. S1F-G), consistent with the RT-qPCR results from these cells (Fig. 2A). Of note, PRO-seq showed that the Eμ enhancer harbors the start sites of an antisense transcript, substantially weaker than the sense transcript (note the Y axis scale). This transcript initiates at

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the 5' boundary of Eμ and terminates within the variable region, resulting in convergent transcription within the variable region (Fig. 3C-D). Moreover, the decrease in antisense transcription in AID$^{-/-}$ Eμ$^{-/-}$ cells is stronger than sense transcription resulting in a major decrease in convergent transcription in the variable region and intron (Fig. S1F-G). Given the weak SHM phenotype in AID$^{-/-}$ Eμ$^{-/-}$ cells (Fig. 2C), we conclude that SHM is resistant to major decreases in Pol II levels within the variable region and intron, in both sense and antisense orientations.

To determine the transcriptomic profiles of other human variable regions, we developed a strategy to replace the endogenous *IGH* variable region of Ramos cells with any variable region of choice using CRISPR editing (Fig. S2A-B and Methods). We chose two variable regions corresponding to the unmutated precursors of broadly neutralizing antibodies identified in individuals infected with human immunodeficiency virus 1 (HIV1). One variable region consisted of VH4-59*01, DH2-02 and JH6*03 (termed VH4-59) (*74*) and the other of VH3-30*18, DH2-02 and JH6*02 (*75*). In both cases, the sequences upstream and downstream of the new variable regions, including the VH4-34 promoter, leader peptide and intron are identical to the endogenous Ramos *IGH* sequence. Our alignment pipeline revealed that sub-regions of both VH3-30 and VH4-59 harbored multimapping reads with the most prominent being the framework region 3 preceding CDR3 in VH4-59 which was entirely covered with multimapping reads (Fig. S2C-D).

We analyzed the transcriptional and mutational landscape of these two variable regions using PRO-seq, PRO-cap and MutPE-seq, the latter following infection with AID(JP8Bdel) (Fig. 4). At VH4-59, we observed an increase in PRO-seq enrichment near the end of the variable region (Fig. 4A and magnified in Fig. 4B), reminiscent of Pol II pausing and akin to that seen at the endogenous Ramos variable region (Fig. 3B-C). These enrichments were considerably weaker at the VH3-30 variable region (Fig. 4D-E). The most plausible explanation may lie in the fact that VH4-59 and VH4-34 belong to the same variable gene family (VH4) and hence share higher sequence identity than VH3-30 which is phylogenetically more distant from the VH4 family (*76*). This also suggests that sequence-intrinsic features may be responsible for the observed differences in nascent transcription profiles. Most importantly, and in agreement with the results from the endogenous Ramos variable region, the highest

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

mutation frequencies in either VH4-59 (Fig. 4B) or VH3-30 (Fig. 4E) were often not associated with regions of increased PRO-seq signal.

Of note, when comparing the mutation patterns between all three variable regions analyzed here, the strong bias for C:G mutations typical of Ramos cells is seen in all cases (Fig. 3E and Fig. 4D, F). Moreover, there appears to be some bias of mutation load in the CDRs although several major mutation hotspots are also located within the intervening framework regions (Fig. 3E and Fig. 4D, F). The coldspot SYC motifs are weakly targeted or untargeted by AID, as expected (Fig. 3E and Fig. 4D, F). Finally, the differential mutability of WRCH motifs by AID, ranging from highly mutated to unmutated, is seen in all the analyzed variable regions (Fig. 3E and Fig. 4D, F) and is in line with observations in previous studies (*47, 48*).

Taken together, we conclude that although pausing, antisense transcription and convergent transcription are all observed to varying degrees in different variable regions, none of these features, or any combination thereof, correlates with the spectra of SHM.

SHM in germinal center B cells at a murine variable region shows no direct relationship with nascent transcriptional features

In mice, SHM occurs robustly *in vivo* in germinal center B cells (GCBs) but is inefficient *in vitro* in activated primary B cells (*48*). Thus, we asked if such differences could be related to differences in the underlying transcriptional landscape. To address this, we made use of the B1-8^{hi} *Igh* knock-in mouse where the murine B1-8^{hi} variable region is knocked-in at the endogenous *Igh* locus (*77*). In this system, one can compare SHM and transcriptional features of the B1-8^{hi} variable region from both germinal center B cells (GCBs) as well as primary *in vitro* activated B cells. To boost SHM in primary cells, we generated homozygous B1-8^{hi} *Rosa26*^{AIDER} mice where AID fused to the estrogen receptor (AIDER) is expressed constitutively from the *Rosa26* promoter (*Rosa26*^{AIDER}) such that upon addition of 4-hydroxytamoxifen, AIDER is translocated into the nucleus (*73*).

PRO-seq analysis from homozygous B1-8^{hi} mice showed that the Vh1-72 gene used in the B1-8^{hi} recombined variable region shares strong homology with other V genes in the murine *Igh* locus resulting

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

in a very high degree of multimapping (Fig. S2E). Importantly, the CDR3 region was covered with uniquely mapping reads since this region is formed via the junction of the V, D and J segments during VDJ recombination resulting in a unique sequence in the genome (Fig. S2E). PRO-seq and PRO-cap analysis showed that there were no striking differences at the B1-8^{hi} variable region between primary B cells and GCBs (Fig. 5A). Moreover, we observed local zones of Pol II accumulation, internal initiation events and antisense/convergent transcription, but these did not necessarily correlate with zones of mutation (Fig. 5A). As noted previously (*48*), SHM profiles of B1-8^{hi} GCBs are not identical to primary B cells because GCBs exhibit 5-10 fold higher frequencies of SHM and substantial mutagenesis of A and T residues arising from error-prone DNA repair of AID-induced mismatches at neighboring C:G residues (Fig. 5B-C). The palindromic AGCT hotspot in CDR3 is selectively mutated with high frequency in primary cells, in line with a former report showing that this hotspot is the earliest targeted motif both *in vitro* and *in vivo* (*48*). Importantly, however, closer inspection showed that even in primary B cells, the same WRCH motifs appear to be mutated as in GCBs suggesting that AID targeting specificity is not substantially different between primary B cells and GCBs (Fig. 5B-C), in line with their comparable transcriptional landscapes (Fig. 5A). We suggest that GCBs may have acquired higher rates of mutation, in part, due to many more rounds of cell division within germinal centers compared to primary cultures, although the AGCT hotspot in CDR3 remains enigmatic in this regard. We conclude that the differences in mutation frequency between primary B cells and GCBs cannot be explained by underlying nascent transcriptional features. Consequently, as in the case of human variable regions described above, the nascent transcriptional landscape of the B1-8^{hi} variable region is not predictive of the patterns or frequencies of SHM.

The absence of correlation between SHM and nascent transcription is a global feature of SHM

To extend our analysis to non-IG AID targets, we made use of a previously available SHM dataset from murine GCBs where mutation frequencies at 275 AID target genes were identified by deep sequencing the first 500 bp from the annotated transcription start site (TSS) (*17*). In this study, mutation analysis was performed in GCBs from mice deficient in base excision and mismatch repair pathways (Ung$^{-/-}$Msh2$^{-/-}$). In these mice, the processing of AID-induced U:G mismatches is abolished and, consequently, DNA

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

replication over these lesions leads to C→T and G→A transition mutations that represent the direct targets of AID (*78*). By comparing these data with PRO-seq and PRO-cap obtained from GCBs, we could compare mutation profiles with those of nascent transcription. Importantly, at many genes, initiation sites or zones defined by PRO-cap are often located at considerable distances from the annotated TSS (defined by the RefSeq database and labeled as TSS in Fig. 5D, Fig. S3 and Fig. S4). Consequently, the 500 bp segment sequenced for mutational analysis (*17*) often begins further upstream or downstream of the initiation sites newly defined by PRO-cap.

The large number of genes in this analysis led to new and unexpected observations regarding the nature of AID targeting. At virtually all genes, the differential mutation rates of WRCH motifs are evident with many such motifs being unmutated and many non-WRCH cytidines being as efficiently targeted (Fig. 5D and several additional examples in Figs. S3 and S4). As we observed in variable regions, mutation frequencies at individual nucleotides and mutation zones did not necessarily correlate with enrichments of PRO-seq or PRO-cap (Fig. 5D, S3 and S4). For example, at *Bcl6*, the major zone of mutation lies in a region of low PRO-seq signal where many mutations are not in WRCH motifs (Fig. 5D). At *Myc*, mutations are observed upstream of the initiation site defined by PRO-cap indicating that mutagenesis can occur in the antisense orientation (Fig. 5E).

In variable and switch regions, mutations initiate ~100-150 bp after the promoter leading to the notion that AID is somehow excluded from initiating or early elongating Pol II (*3, 43, 45, 79, 80*). Contrary to this, we identified a set of genes where the highest mutation frequencies were observed within 100-150 bp of the initiation site defined by PRO-cap (*Il4ra* in Fig. 5F and additional examples in Fig. S3A). Indeed, in some of these genes, the most mutated residues were in very close proximity to the initiation site, such as *Mcm7* (Fig. 5G). In some other genes, strongly mutated residues coincided with the strongest peak of initiation (Fig.S3B). These observations indicate that AID can act at the early stages of the transcription cycle at least at some genes, suggesting that AID is not *per se* excluded from initiating or early elongating Pol II.

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Altogether, we conclude that the lack of predictability of mutation spectra from nascent transcription patterns is observed at both IG variable regions and non-IG AID target genes, indicating that the absence of correlation between SHM and nascent transcriptional patterns is a global feature of SHM.

SHM targeting of the murine B1-8^{hi} variable region is retained in the context of the human *IGH* locus

To determine whether the gene regulatory context affects the patterns of SHM and transcription, we replaced the endogenous variable region in Ramos cells with the entire murine B1-8^{hi} variable region (Ramos^{B1-8hi}) using the strategy described earlier (Fig. 6A and Fig. S2A-B). In this scenario, transcription and SHM of the murine B1-8^{hi} variable region would be under the control of the Ramos VH4-34 promoter and the human Eμ and 3' *IGH* super-enhancers. This allowed us to ask whether the targeting of SHM to the B1-8^{hi} variable region was influenced by differences in the human versus the mouse enhancers and promoters.

Importantly, when placed in the context of the human *IGH* locus, the B1-8^{hi} sequence was covered almost exclusively by uniquely mapping reads (Fig. 6B), which is in sharp contrast to the extensive multimapping observed within B1-8^{hi} at the murine *Igh* locus (Fig. S2E). This is because the Vh1-72 gene used in the B1-8^{hi} sequence is much less homologous with the human V genes than the mouse V genes. PRO-seq and PRO-cap revealed that the nascent transcriptional profile of the B1-8^{hi} sequence in Ramos^{B1-8hi} cells (Fig. 6C) shared similarities with the B1-8^{hi} sequence in the murine *Igh* context (Fig. 5A). For example, similarities were noted in CDR3 and flanking sequences, for instance, the presence of the antisense initiation site (PRO-cap track) and shared spikes of nascent transcriptional activity (PRO-seq) (Fig. 6C). More importantly, MutPE-seq following AID(JP8Bdel) infection revealed a very similar SHM spectrum in Ramos^{B1-8hi} cells compared to the murine B1-8^{hi} context, especially with primary murine B cells (Fig. 6D). Of note, the dominant, palindromic AGCT hotspot in CDR3 was the most mutated residue exactly as in B1-8^{hi} murine primary cells and GCBs (Fig. 6D). The other mutated C:G residues in the Ramos^{B1-8hi} sequence were largely the same as those mutated in the murine context albeit the relative mutation rates within each variable region varied (Fig. 6D, compare with Fig. 5B-C). Importantly, mutation frequencies in murine GCBs and AID(JP8Bdel)-expressing Ramos cells were

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

comparable (note the Y axes scales in Fig. 6D) with the major difference being the weak A:T mutagenesis typical of Ramos cells. Of note, the sense PRO-cap peak in CDR3 in the murine B1-8^{hi} (Fig. 5A), located just 3 bp upstream of the strong AGCT hotspot, is absent in Ramos$^{B1\text{-}8hi}$ (Fig. 6C) implying that this initiation site is not essential for the high mutation frequency of this hotspot motif. In sum, the SHM spectrum of the B1-8^{hi} variable region in mouse B cells is largely retained in Ramos$^{B1\text{-}8hi}$ cells despite dissimilarities in their nascent transcriptional profiles.

We conclude that although enhancers are important for recruiting AID to the variable regions and other non-IG genes, the final rules of mutagenesis are apparently “hard-wired” into the DNA sequence rather than specific transcriptional features or chromatin marks.

Discussion

In this study, we aimed to understand the relationship of SHM spectra with local patterns of nascent transcription and epigenetic marks using high-resolution profiling of variable regions. We find that the local SHM patterns and frequencies cannot be predicted from, nor do they appear to be derived from, the underlying transcriptional landscape. Moreover, substantial loss of actively elongating Pol II and nascent transcription in both orientations within variable regions, seen in E$\mu^{-/-}$ cells, does not majorly impact on SHM. These results have important implications for the interpretation of transcription and chromatin modification measurements in the context of SHM or CSR. For example, despite the global correlation of SHM target genes with activating histone marks, our results suggest that they do not play a major role in determining SHM frequency or in defining the SHM spectrum. We suspect that in studies depleting histone modifying enzymes, SHM or CSR defects may result, in part or whole, from secondary effects arising from perturbation of these global gene regulatory factors. In the case of histone modifying enzymes, it is plausible that they function independently of their enzymatic activity or via non-histone substrates, as has been shown for many such enzymes (*81-85*). Similarly, we infer that although AID target genes are often enriched in paused Pol II and convergent transcription, these features do not appear to play a direct, mechanistic role in mutagenesis. Moreover, the fact that SHM remains robust

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

414 despite up to 4-fold reduced *IGH* nascent transcription in E$\mu^{-/-}$ cells implies that although SHM requires
415 transcription, it is not directly correlated with the rate of transcription and, by extension, the level of Pol II
416 occupancy in the variable region. This has important implications for interpreting the meaning of
417 correlations between transcription and SHM observed upon gene knockouts or knockdowns. In such
418 cases, our results suggest that the true cause of the SHM defect may not be due to, or may only partly
419 result from, decreased transcription.

420 During transcription, the non-template strand is released as ssDNA to allow RNA synthesis.
421 Hence, this ssDNA bubble has been thought to serve as a source of ssDNA for AID, especially if Pol II is
422 paused or its elongation rate is slowed down, a situation where ssDNA would be available more stably
423 and for longer periods. However, our data show no consistent correspondence of mutations hotspots with
424 sites of Pol II accumulation. From structural studies of paused and elongating Pol II complexes, it is
425 apparent that of the ~11 nt non-template ssDNA bubble, about half lies buried within Pol II (*71, 72*) (Fig.
426 7A). Importantly, the exposed portion (~5 nt) is entirely covered by SPT5 (*72*) (Fig. 7A). The PAF complex
427 and SPT6 are located further away from this site (*71*) (Fig. 7A). Under such structural constraints, it is
428 difficult to conceive how AID could access the ~5 nt exposed ssDNA and catalyze deamination even if Pol
429 II were paused or slowly elongating. Thus, it appears unlikely that the transcription bubble is a substrate
430 for AID which can explain why we observe no correlation between sites of Pol II accumulation and SHM.
431 Moreover, it appears that SHM at recurrent hotspot residues is not affected by distally located mutations
432 elsewhere in the variable region (*47*). This leads us to infer that sequences directly flanking the target
433 cytidine may be important determinants of mutability.

434 We propose the following four-step model to integrate our data with the literature to date. First,
435 AID is delivered to IG variable regions and non-IG targets via enhancers (*38, 40, 42*) (Fig. 7B. step 1).
436 This is accomplished by chromatin looping between enhancers and target genes, a process that has
437 been suggested to occur via cohesin-mediated loop extrusion (*86*). Such interactions can lead to the
438 formation of enhancer-gene hubs via dynamic, multivalent interactions between Pol II, transcription
439 factors and cofactors like the Mediator complex. This results in the stabilization of transcription factors
440 and recruitment of Pol II and cofactors to activate target genes (*87-90*). Once delivered into the hub, AID

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

associates with Pol II and associated proteins such as SPT5, SPT6 and the PAF complex (Fig. 7B, step 2). We propose that the primary purpose of these interactions is to retain AID in the vicinity of Pol II complexes. The ability of AID to engage in multiple, independent dynamic interactions ensures that it is retained within the transcriptional hub, ready to access ssDNA when it is made available. Indeed, the fact that depletion of any of these proteins leads to AID recruitment defects (*21-24*) suggests that all such interactions are contributing to ensure that AID stays within the transcription hub or that these factors are important for maintaining or forming the hub (*90*).

For successful deamination, ssDNA needs to be exposed and the source of this ssDNA remains enigmatic (Fig. 7B, step 3). As explained above, the transcriptional bubble appears to be inaccessible. Previous work has suggested that DNA upstream of the moving Pol II could provide a source of ssDNA for AID in variable regions since, being negatively supercoiled, it may be more accessible (*60, 91*). Patches of ssDNA have been detected in the variable region and it has been suggested that they arise from negative supercoiling associated with paused Pol II although their presence did not correlate well with mutation spectra (*57, 61, 91*). Indeed, AID was shown to target negatively supercoiled DNA, but not relaxed DNA, on both template and non-template strands (*92*). However, since we find no correlation between mutation and pausing, we infer that these ssDNA patches are likely the result of transcription itself rather than specific sites of pausing.

Although, in principle, negative supercoils could create ssDNA patches that allow mutation on both strands, the fact that significant mutational asymmetry is observed upon depletion of RNA Exosome subunits or the DNA:RNA helicase, Senataxin (*93-95*) argues that the species AID likely targets is one containing an RNA:DNA hybrid. Indeed, the variable region sense transcript is a substrate of the RNA exosome and the latter is required for optimal mutation of the template strand in the B1-8^{hi} variable region, findings that directly implicate the processing of DNA:RNA hybrids in SHM (*93*). In this regard, and in contrast to IG switch regions, variable regions lack R loops (*61, 91*). Given the presence of ssDNA patches, the requirement for RNA processing and the lack of R loops, we propose that transcription termination may provide an important source of ssDNA in variable regions. It is plausible that the collapse of the elongation complex could release DNA with the nascent RNA within the transcription bubble still

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

hybridized (Fig. 7B, step 4). The RNA would have to be unwound by DNA:RNA helicases like Senataxin (*95*) and/or digested with RNaseH followed by degradation from the 5` end by Xrn2 and from the 3` end by the RNA exosome, thereby providing AID access to the template strand (*26, 93-96*) (Fig. 7B, step 4). In this regard, termination has been correlated with SHM (*60, 97*) and degradation of Pol II by ubiquitination has been linked to AID targeting (*98*). We note that if, during multiple rounds of transcription within a cell population, termination occurs stochastically along the variable region and intron, then such events would not be detected by population-based transcriptomic assays like PRO-seq or ChIP-seq.

Interestingly, recent genome-wide analyses have shown that termination is frequent in the first 3 kb from the promoter (*99*) which, in the context of the IG genes, would encompass the entire variable region and, depending on which J segment is used, part or whole of the intronic sequence. Indeed, SHM starts 100-150 bp from the V gene promoter and extends for another 1.5-2 kb (*43-45*). Thus, we suspect that ssDNA patches and DNA:RNA hybrids may arise, at least in part, from premature termination of Pol II within the variable region. A similar process could occur at the 5` ends of non-IG AID target genes or at transcribed enhancers that are also targets of AID. Given that neither we nor others have observed any transcriptional feature correlating with SHM spectra, we favor the idea that deamination occurs on ssDNA made available through co-transcriptional processes that expose ssDNA patches, such as termination, but that it is dissociated from actively elongating Pol II (Fig. 7B, step 4). We note, however, that termination is difficult to measure directly, and how the disassembly of Pol II occurs is poorly understood. Thus, the mechanism and kinetics of resolution of the DNA:RNA hybrid within terminating Pol II are not known.

The final step, that is, the actual mutagenesis, determines the observed SHM spectra (step 4, Fig. 7B). AID has a very high affinity for ssDNA and can remain associated with ssDNA for several minutes which may allow sufficient time for deamination, a reaction which is extremely inefficient and can take up to several minutes (*100, 101*). However, WR<u>C</u>H motifs are not equally mutated, suggesting that the DNA sequence context determines whether ssDNA containing a WR<u>C</u>H motif is mutated and to what extent, a notion which is supported by biochemical studies (*101*). Indeed, analyses of variable regions has suggested that sequence-intrinsic features may regulate the SHM spectra (*47, 50-52*). Given that

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

neither ssDNA patches nor nascent transcriptional features correlate with SHM spectra, we favor the notion that sequence context of a WRCH motif determines its mutability and propose that sequence context regulates the residence time of AID on ssDNA and thereby modulates the catalytic efficiency of AID *in situ* (Fig. 7B, step 4).

The precise contribution of DNA sequence context towards SHM will be a major focus of investigation for the future with the goal of understanding the rules governing differential mutability. Insights from such studies could be harnessed to further increase the potency and breadth of mature broadly neutralizing antibodies where hotspot saturation can prevent further SHM (*47*). In this regard, our system of efficiently replacing variable regions in Ramos cells offers an ideal platform for asking how changes in the sequence locale affects the targeting of SHM to a given hotspot.

Materials and Methods

Cell culturing

Ramos were cultured in complete RPMI medium (in-house) supplemented with 10% fetal bovine serum (FBS; Invitrogen), glutamine (Invitrogen), sodium pyruvate (Invitrogen), HEPES (made in-house) and antibiotic/antimycotic (Invitrogen). LentiX packaging cells were cultured in complete DMEM medium (in-house) containing 10% fetal bovine serum (FBS; Invitrogen), glutamine (Invitrogen), sodium pyruvate (Invitrogen), HEPES (made in-house) and antibiotic/antimycotic (Invitrogen).

Mice

The mice were maintained in a C57BL/6 background and housed in the IMBA-IMP animal facility in standard IVC cages with HEPA filtering. All animal experiments were carried out with valid breeding and experimental licenses (GZ: MA58-320337-2019-9, GZ: 925665/2013/20 and GZ: 618046/2018/14) obtained from the Austrian Veterinary Authorities and in compliance with IMP-IMBA animal house regulations.

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Generation of AID$^{-/-}$ Ramos cells

We generated Ramos cells expressing EcoR, the receptor for the ecotropic envelope protein, Eco-Env, so as to allow efficient infection of Ramos cells with ecotropic lentiviruses. Ramos cells were infected with a pRRL lentiviral vector expressing EcoR and maintained under Puromycin selection. These cells were transfected with a small guide RNA targeting *AICDA*, the gene expressing AID, and recombinant Cas9 protein via electroporation using the Neon transfection system (Thermo Fisher Scientific). The following day, single cells were sorted into 96-well plates and allowed to expand. Four weeks later, clones harboring frameshifting indels at the Cas9 target site were identified by genotyping and loss of AID in these clones was confirmed by Western blot analysis. One clone (D3) was used for all subsequent analyses and for generation of new lines.

Generation of Eμ$^{-/-}$ Ramos cells

To delete the 583 bp segment containing Eμ, we generated homology repair plasmids having Ef1a promoter-driven floxed GFP and mCherry expression cassettes flanked by homology arms. These plasmids were transfected into AID$^{-/-}$ Ramos cells (clone D3) along with three in vitro synthesized guide RNAs (6 μg each, designed using CRISPOR) and 7.5 μg recombinant Cas9 protein (Vienna Biocenter Core Facilities) using the Neon Transfection System (Thermo Fisher Scientific). A week later, GFP/mCherry double-positive single cells were isolated using a BD FACS Aria III sorter (BD Biosciences). Successful knock-in clones were identified by genotyping with PCR and Sanger sequencing of PCR products. Next, 200 μg recombinant Cre recombinase (Molecular Biology Service, IMP) was added to the culture medium to excise the floxed mCherry/GFP cassettes. GFP/mCherry double-negative clones were isolated one week after electroporation using a BD FACS Aria III sorter and genotyped via PCR and Sanger sequencing of PCR products.

Generation of Ramos cells expressing new, exogenous variable regions

AID$^{-/-}$ Ramos cells (clone D3) were electroporated with Cas9-sgRNA ribonucleoprotein complexes (prepared in-house by the Vienna Biocenter Core Facilities) and homology repair templates containing a pair of unique sgRNA-target sites to excise the entire variable region including the promoter (Fig. S2A).

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Single, IgM-negative clones were isolated via flow cytometry, expanded and genotyped. One validated IgM-negative clone was selected as the parental clone. This clone was then electroporated with Cas9-sgRNA ribonucleoprotein complexes targeting the unique sgRNA sites and homology repair templates containing new variable regions under the control of the endogenous Ramos VH4-34 promoter. Single, IgM-positive cells were isolated, expanded and genotyped.

Lentiviral infections

Lentiviral pRRL vectors expressing AID(JP8Bdel) coupled with mCherry were transfected along with ecotropic envelope (Eco-env)-expressing helper plasmid into LentiX cells via the standard calcium phosphate methodology. Lentiviral supernatants were used to infect Ramos cells via spinfection (2350 rpm for 90 min) in the presence of 8 μg/ml Polybrene (Sigma). Flow cytometry was used to sort mCherry-positive cells for mutation analysis.

Isolation and activation of murine primary, splenic B cells

Mature, naïve B cells were isolated from spleens of 2-4-month-old wild-type C57BL/6, as per established protocols (Pavri et al., 2010). B cells were cultured in complete RPMI medium supplemented with 10% fetal bovine serum (FBS) and antibiotics, Interleukin 4 (IL4; made in-house by the Molecular Biology Service, IMP), 25 μg/ml Lipopolysaccharide (Sigma) and RP105 (made in-house by the Molecular Biology Service, IMP) and harvested after 3 or 4 days. B1-8^{hi} Rosa26 ^{AIDER} mice were generated by crossing B1-8^{hi} mice (*65*) with *Rosa26*^{AIDER} mice (*73*) and maintained as a homozygous line for both alleles. For SHM assays from activated B1-8^{hi} *Rosa26* ^{AIDER} primary B cells, 2μM 4-hydroxy tamoxifen (4-HT) (Sigma) was added at the time of activation with IL-4, LPS and RP105.

Isolation of germinal center B cells (GCBs) from immunized B1-8^{hi} mice

Sheep red blood cells (SRBCs) were washed thrice with PBS of which 0.2x10^{9} SRBCs in 100μl PBS (per mouse) were injected followed by another injection of 1x10^{9} SRBCs in 100μl PBS five days later. Mice were harvested twelve days after the first immunization. Spleens were harvested and single-cell suspensions were stained with B220 conjugated to fluorescein isothiocyanate (B220-FITC, BD

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Biosciences, 1:500 dilution), Fas conjugated to PE-cyanine 7 (Fas-PE-Cy7, BD Biosciences, 1:1000 dilution) and CD38 conjugated to Allophycocyanin (CD38-APC, ThermoFisher, 1:200 dilution) along with Fc block (BD Biosciences, 1:500 dilution). GCBs (B220+ Fas+ CD38^{-}) were isolated on a BD FACS Aria II sorter (BD Biosciences). Following sorting, nuclei were isolated and used for PRO-seq and PRO-cap (see below). For one experiment, ~10^7 GCBs from fifteen immunized mice were pooled yielding 3-4x10^6 viable nuclei that were used for run-on. For MutPE-seq, genomic DNA was isolated from 15 x10^6 GCBs as described previously (*102*).

PRO-seq and PRO-cap

PRO-seq was performed as described previously (*103*) with several modifications.
To isolate nuclei, murine and *Drosphila* S2 cells were resuspended in cold Buffer IA (160 mM Sucrose, 10 mM Tris-Cl pH 8, 3 mM $CaCl_2$, 2 mM $MgAc_2$, 0.5% NP-40, 1 mM DTT added fresh), incubated on ice for 3 min and centrifuged at 700 g for 5 min. The pellet was resuspended in nuclei resuspension buffer NRB (50 mM Tris-Cl pH 8, 40% Glycerol, 5 mM $MgCl_2$, 0.1 mM EDTA). For each run-on, 10^7 nuclei of the sample and 10% *Drosophila* S2 nuclei were combined in a total of 100 µL NRB and incubated at 30°C for 3 min with 100 µL 2x NRO buffer including 5µl of each 1mM Bio-11-NTPs. In some PRO-cap experiments, the run-on reaction was performed in the presence of two biotinylated NTPs, Biotin-11-UTP and Biotin-11-CTP (Perkin-Elmer), and unlabeled ATP and GTP. Subsequent steps were performed as described (Mahat et al 2016), except that 3' and 5' ligations were performed at 16°C overnight and CapClip Pyrophosphatase (Biozym Scientific) was used for 5'end decapping. We also used customized adapters (PRO-seq: 3' RNA adapter: 5'5Phos/NNNNNNNGAUCGUCGGACUGUAGAACUCUGAAC/3InvdT-3' and 5' RNA adapter: 5'-CCUUGGCACCCGAGAAUUCCANNNN-3'). RNA was reverse transcribed by SuperScript III RT with RP1 Illumina primer to generate cDNA libraries. Libraries were amplified with barcoding Illumina RPI-x primers and the universal forward primer RP1 using KAPA HiFi Real-Time PCR Library Amplification Kit. PRO-cap: 3' linker (DNA oligo): 5rApp/NNNNAGATCGGAAGAGCACACGTCT/3ddC , 5'linker (RNA): ACACUCUUUCCCUACACGACGCUCUUCCGAUCUNNNNNNNNNN, reverse transcription and library amplification as in PRO-seq but using selected TruSeq_IDX_1-48 PCR primers for RT, and the same

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

TruSeq_IDX PCR primers together with TruSeq Universal Adapter forward primer for PCR. For both methods amplified libraries were subjected to gel electrophoresis on 2.5% low melting agarose gel and amplicons from 150 to 350 bp were extracted from the gel (carefully separated from possible linker dimers), multiplexed and sequenced on Illumina platforms with SR50 or longer read mode.

RT-qPCR

RT-qPCR assays with externally spiked-in *Drosophila* S2 cells were described in detail in our previous studies (*104*). Briefly, *Drosophila* S2 cells were mixed with Ramos cells or B1-8^{hi} B cells at a ratio of 1:4 ($4x10^5$ B cells and $1x10^5$ S2 cells) followed by total RNA extraction with TRIzol reagent (Thermo Fisher), DNaseI digestion, and cDNA synthesis with random primers. The 2-ΔΔCt method was used to quantify the data using the *Drosophila Act5c* transcript for normalization.

ChIP-qPCR and ChIPseq

ChIP-seq and ChIP-qPCR were performed as described previously without any modifications (*22, 102*). Antibodies used in this study are listed in Table S1.

ATAC-seq

ATAC-seq was performed exactly as described in detail in our previous work (*102*).

Mutational analysis by paired-end deep sequencing (MutPE-seq)

MutPE-seq was performed following the principles described in two previous reports (*48, 73*) with several adaptations. 80 ng of genomic DNA were amplified by PCR with the Kapa Hifi HS 2x RM (Roche Diagnostics), For the first PCR, we used 20-25 cycles with 0.2µM locus-specific primers fused to a varying number of random nucleotides and to the first part of the Illumina adapter sequences (FW: 5'-CTCTTTCCCTACACGACGCTCTTCCGATCT-(N)x-gene-specific sequence-3'; RV: 5'-CTGGAGTTCAGACGTGTGCTCTTCCGATCT-gene-specific sequence-3'; see Table S1 for a complete list of primer sequences). Random Ns are used to increase complexity and shift frames of very similar amplicons to improve cluster calling and sample identification. After first PCR, the samples were purified with 0.2x/0.7x SPRI beads (Beckman Coulter), eluted in 10µl water and amplified for 10 cycles with

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

0.75µM primers containing linker sequences and dual barcoding (FW: 5'-AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGAC-3'; RV: 5'-CAAGCAGAAGACGGCATACGAGATXXXXXXGTGACTGGAGTTCAGACGTGTG-3' where the stretches of X nucleotides serve as barcodes creating a unique dual barcode combination for each sample). PCR products were purified either by extraction from a 2% low-melt agarose gel or with 0.7x SPRI beads and eluted in 20 µl water. The concentration was determined by Picogreen-based measurements on a Nanodrop machine or by a Fragment Analyzer. Samples were equimolarly pooled for next generation sequencing on an Illumina MiSeq flowcell and sequenced paired-end (PE300).

Primers and sgRNAs

All oligos, primers and adapters used in this study are listed in Table S1.

Bioinformatics

Mapping PRO-seq, PRO-cap, ChIP-seq and ATAC-seq reads to IG variable regions

Reads were trimmed for standard adapters and low quality (Q<30) 3'-end bases and filtered for a remaining length of at least 20nt (excluding the UMI, if applicable) using cutadapt (*105*). Alignment to the reference genome was done with Bowtie (*106*). The NCBI GRCm38.6 assembly was used as the mouse reference. To this we added the sequence of the recombined locus containing the variable region as an additional chromosome. Similarly, for human data, the Hg38 assembly was used, with the relevant variable region sequence added as a chromosome. Where applicable, the dmr6 assembly of *Drosophila melanogaster* from Flybase ((*107*) release 6.27) was used as the spike-in reference and was added to the alignment index. During alignment, up to three mismatches were allowed. To accommodate mapping to the repetitive *IGH/Igh* locus, a high degree of multimapping was allowed (194 potential V segments annotated for GRCm38.6). In the case of alignment with spike-ins, only reads mapping exclusively to either genome were considered. For PCR deduplication, UMIs were identified and filtered with UMI-tools ((*108*) v1.0.0). No sequence differences were allowed in UMIs when collapsing the duplicates. Multimapping reads were then filtered to identify those specific to the variable region sequence, taking

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

into account the repetitiveness of the reference *IGH/Igh* locus. Reads mapping to the fixed variable region sequence were allowed to multimap only to the native *IGH/Igh* locus (mouse chr12:113572929-116009954 or human chr14:105836764-106875071). The native loci were defined as the region between the earliest position belonging to an annotated V segment and the latest position belonging to an annotated J segment. Reads with a mapped location outside these areas were rejected. This filtering was implemented in a custom Python script. The qualifying reads were then classified as reads mapping uniquely to the particular variable region sequence and reads mapping to both the variable region sequence and to the native *IGH/Igh* region. Genome browser tracks were created by quantifying and scaling read coverage of the variable region sequence using bedtools ((*109*) v2-29.0). Reads were split by strand, and strand labels for PRO-seq were inverted in order to match the strand designations of the respective PRO-cap reads. For both data types, only the first 5' base of the strand-adjusted reads was used for the tracks. Additionally, the 5' end of PRO-seq reads was shifted by 1 nucleotide downstream, to compensate for the fact that the last nucleotide in the RNA was incorporated during the run-on reaction. For normalized tracks, read counts were scaled to RPM (reads per million). Finally, "-" strand coverage tracks were further scaled by -1, for visualization purposes.

Code for the workflow and the custom scripts is available on Github at https://github.com/PavriLab/IgH_VDJ_PROcapseq.

Analysis of MutPE-seq

Reads were trimmed for standard adapters with cutadapt (*105*). Poor quality (Q<25) 3' bases were trimmed with trimmomatic (*110*) by averaging over a sliding window of 5nt. Read pairs were then filtered for minimum remaining length (200nt for read 1, 100nt for read 2) using cutadapt. Read mates were merged down to make combined single-end reads with FLASH (*111*) allowing 10% mismatch between the mates. Obvious erroneous mergers were removed by selecting combined reads with lengths within ±30nt of the amplicon length using cutadapt, The remaining combined reads were aligned with Bowtie2 (*112*), using the "–very-sensitive-local" alignment mode and only the fixed variable region sequence and

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

its immediate vicinity as reference. Samples of V genes VH4-59 and VH4-34 are different only by single nucleotide polymorphisms and were filtered for contamination by the respective second sequence by jointly aligning with the "expected" and the other "contaminating" variable region sequence, discarding all aligned contaminating reads. A pile-up was generated with samtools (*113*) taking into account only bases with quality of at least 30. The pileups were then quantified with a custom Python script and the resulting mutation counts were processed and visualized with custom scripts in R (v3.5.1), with the help of additional R packages (data.table [https://cran.r-project.org/web/packages/data.table/index.html], ggplot2 [https://cran.r-project.org/web/packages/ggplot2/index.html] , ggrepel [https://cran.r-project.org/web/packages/ggrepel/index.html], patchwork [https://cran.r-project.org/web/packages/patchwork/index.html]). Background mutation profiles were controlled for by subtracting the corresponding mutation frequencies in control samples from the frequencies in the samples of interest, at each position and for each substitution type. Annotation of hot and cold spots was created by means of regex search for the corresponding patterns in the reference sequence.

Code for the workflow and the custom scripts is available on Github at https://github.com/PavriLab/IgH_VDJ_MutPE.

Data availability

All NGS data has been deposited in GEO under accession number GSE202042.

Author contributions

UES performed all PRO-seq, PRO-cap and MutPE-seq assays and analyzed data. JF generated the Ramos lines with new variable regions and performed infections for MutPE-seq. KF and TN developed the bioinformatic pipeline to map reads to variable regions, KF developed the bioinformatic pipeline for MutPE-seq. Marina M performed ChIP-seq, ChIP-qPCR and RT-qPCR in Ramos cells and analyzed data. IO performed the bioinformatic analysis of non-IG AID target genes. BB performed ChIP-qPCR and RT-qPCRs. EMW and RP generated the E$\mu^{-/-}$ lines. MS conducted GCB PRO-seq with UES. ACG performed ATAC-seq. Marialaura M cloned JP8Bdel expression vectors for MutPE-seq. ACG performed ATAC-seq. HM provided the germline-reverted VH4-59 and VH3-30 sequences and provided feedback on the

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

manuscript. ITS performed bioinformatic analysis, provided resources and analyzed data. RP conceived the project, analyzed data and wrote the manuscript with critical inputs from UES and JF.

Acknowledgements

We thank Almudena Ramiro and Angel Alvarez-Prado (CNIC, Madrid, Spain) for sharing the mutational datasets from Ung$^{-/-}$Msh2$^{-/-}$ GCBs. We gratefully acknowledge the Vienna Biocenter Core Facilities (VBCF) for assistance in generating E$\mu^{-/-}$ lines and for next generation sequencing, the IMP/IMBA BioOptics facility for flow cytometry usage, the IMP/IMBA Molecular biology services for Sanger sequencing and reagents, and the IMP/IMBA animal house. We thank Maximilian von der Linde for uploading NGS tracks to GEO. We thank Carrie Bernecky (IST, Vienna) and Clemens Plaschka (IMP) for generating the 3D Pol II elongation complex structural visualizations. This work was funded by Boehringer Ingelheim, The Austrian Industrial Research Promotion Agency (Headquarter Grant FFG-834223), and grants from the Austrian Science Fund to UES (FWF T 795-B30) and RP (FWF P 32043-B).

Conflicts of interest

The authors declare no conflicts of interest.

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure Legends

Figure 1: The Eμ enhancer regulates the chromatin landscape in the intron but not in the variable region in Ramos human B cells. (A) ChIP-seq profiles of H3K27ac, H3K4me3 and H3K36me3 at the *IGH* locus from AID$^{-/-}$ and two independent clones (c1 and c2) of AID$^{-/-}$ Eμ$^{-/-}$ Ramos cells, the latter generated as described in Fig. S1C-D. The sequence annotations indicate the positions of the IGH promoter (Prom), the complementary determining regions (CDR1-3), the intronic enhancer (Eμ) and the switch μ region (Sμ). The bioinformatic approach to include multimapping reads (explained in the text and depicted in Fig. 3A) was applied. **(B)** A magnified view of the locus from A. **(C)** ChIP-qPCR analysis at *IGH* from AID$^{-/-}$ and AID$^{-/-}$ Eμ$^{-/-}$ Ramos cells (c1, c2) to measure the relative levels of H3K27ac, H3K4me3, H3K79me3, pan-acetylation of histones H3 and H4, and histone H3. The amplicons used for PCR are indicated in the schematic diagram at the top. The active *B2M* gene is used as a positive control and a gene desert on chromosome 1 is used as a negative control (Neg.). The data represent three independent experiments. Asterisks indicate $P<0.05$ using the Student's t-test and ns indicates not significant ($P>0.05$).

Figure 2: Ablation of the Eμ enhancer significantly decreases nascent transcription but not SHM. (A) RT-qPCR measurements of nascent transcripts at the Ramos variable region in AID$^{-/-}$ and AID$^{-/-}$ Eμ$^{-/-}$ cells (c1, c2). To account for potential clonal variation, Ramos cells were spiked with *Drosophila* S2 cells prior to RNA extraction. The data was normalized to the levels of the *Drosophila* housekeeping gene, *Act5c*. *GAPDH* mRNA was used as a control. Asterisks indicate $P<0.05$ using the Student's t-test and ns indicates not significant ($P>0.05$). **(B)** RT-qPCR analysis as in A measuring the spliced IgM mRNA. **(C)** Table of mutation frequencies at the Ramos variable region and the JH6 intron (amplicons shown in the diagram above the table) in AID$^{-/-}$ and AID$^{-/-}$ Eμ$^{-/-}$ cells following infection with AID(JP8Bdel) for 6 days. Statistical analysis was performed with the Student's t-test. **(D)** Flow cytometry analysis of IgM expression in AID$^{-/-}$ and AID$^{-/-}$ Eμ$^{-/-}$ Ramos cells infected with AID(JP8Bdel) for 7 days.

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(E) Bar graph summarizing the flow cytometry analyses from D, showing the percent of IgM loss from three independent experiments.

Figure 3: Comparison of transcriptional and mutational landscapes at the endogenous Ramos variable region. (A) Scheme depicting the strategy used to align multimapping reads to the variable region. As described in the text and Methods, since upstream, non-recombined V genes are silent in Ramos cells, a read that maps to the recombined variable region is retained if it is mapping to any V, D or J segment of the human *IGH* locus but nowhere else in the human genome. The same principle is applied when mapping reads to murine variable regions to the mouse genome. **(B)** Integrative Genomics Viewer (IGV) browser snapshot of nascent RNA 3' ends (by PRO-seq) aligned to the Ramos variable region. Multimapping (top track) reads are separated from uniquely mapping reads (middle track), and these two tracks are subsequently combined to generate the total profile (bottom track. **(C)** PRO-cap and PRO-seq 5' and 3' end densities, respectively, at the *IGH* locus along with mutation frequencies at the variable region displayed on the Integrative Genomics Viewer (IGV) browser. Mapping of the variable region transcriptome was done via the bioinformatic pipeline outlined in A and exemplified in B (total profiles are shown). The locations of the antigen-binding complementary determining regions (CDR1-3) are highlighted. Mutation analysis via MutPE-seq was performed following infection of AID$^{-/-}$ cells with AID(JP8Bdel)-expressing lentiviruses for 6 days. **(D)** A magnified view of the variable region from C above. **(E)** Details of the mutation spectrum at the Ramos variable region. Mutated cytidines in AID hotspot motifs (WR<u>C</u>H) are displayed as red bars. All other C:G mutation are shown as black bars and A:T mutations as grey bars. The panel under the graph shows the position of both hotspot (WR<u>C</u>H in black, AGCT in red, upper panel) and coldspot (SY<u>C</u>, bottom panel) motifs. **(F)** Waterfall plot with mutations ordered from highest (left) to lowest (right) frequency following the color code described in E. **(G)** Mutation frequency bar plot showing the percentages (indicated within the bars) of the three mutations classes following the color code described in E. **(H)** Bar graph indicating the percentage of the type of mutation indicated on the X axis. The C→T and G→A transition mutations are the signature of AID activity.

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 4: Comparison of transcriptional and mutational landscapes at two different human variable regions expressed from the Ramos *IGH* locus. (A) IGV browser snapshots of nascent RNA 5' (PRO-cap) and 3' (PRO-seq) ends and mutation tracks (MutPE-seq) at the VH4-59-DH2-JH6 variable region expressed from the Ramos VH4-34 promoter (CDRs 1-3 highlighted). Mutation analysis was performed following infection of AID$^{-/-}$ cells with AID(JP8Bdel)-expressing lentiviruses for 7 days. **(B)** A magnified view of the VH4-59-DH2-JH6 variable region from A above. **(C)** Detailed mutational analysis of the VH4-59-DH2-JH6 variable region displayed and color-coded as in Fig. 3E. A bar plot, as in Fig. 3G, with the percentage of mutation frequencies is shown on the right. **(D)** Nascent transcriptional analysis of the VH3-30-DH2-JH6 variable region as in A. **(E)** A magnified view of the VH3-30-DH2-JH6 variable region from D above. **(F)** Detailed mutational analysis of the VH3-30-DH2-JH6 variable region (see C above for details).

Figure 5: Transcriptional and mutational landscapes of the murine B1-8^{hi} variable region and non-*Ig* AID target genes in mice. (A) Nascent transcriptional analysis at the *Igh* locus in murine B1-8^{hi} primary, splenic B cells stimulated for four days with LPS, IL4 and RP105. IGV browser snapshots show the 5' (PRO-cap) and 3' (PRO-seq) ends of the aligned reads. For MutPE-seq, B1-8^{hi} *Rosa26*^{AIDER} primary B cells (expressing AID fused to the estrogen receptor from the *Rosa26* locus) were activated with LPS, IL4 and RP105 for four days in the presence of 4-hydroxytamoxifen (4-HT) to trigger AIDER nuclear import. PRO-seq, PRO-cap and MutPE-seq were also performed from sorted, splenic germinal center B cells (GCBs) following immunization with sheep red blood cells for 11 days. **(B)** Analysis of SHM spectra of the B1-8^{hi} variable region from B1-8^{hi} *Rosa26*^{AIDER} primary, activated murine splenocytes. The bar graph on the right shows the percentage of each indicated mutation category (see Fig. 3E and 3G for details). **(C)** Analysis of SHM spectra of the B1-8^{hi} variable region from splenic GCBs following immunization with sheep red blood cells for 11 days. **(D)** 5' and 3' ends of nascent RNA (PRO-cap, PRO-seq) and mutation profiles at four selected AID target genes. Mutational data are from Alvarez-Prado et al. (2018) wherein the first 500 bp of the genes were sequenced. The region displayed extends from -100

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

bp from the annotated (RefSeq) TSS up to 50 bp downstream of the sequenced amplicon. The WRCH motifs are indicated as red dots.

Figure 6: Analysis of gene regulatory context on the nascent transcriptional landscape and SHM of the murine B1-8^{hi} variable region. (A) Scheme showing the exchange of the endogenous Ramos variable region for the murine B1-8^{hi} variable region to generate the Ramos^{B1-8hi} human *IGH* locus following the approach described in Fig. S2A. **(B)** 3' ends of nascent RNAs by PRO-seq at the Ramos^{B1-8hi} *IGH* locus showing the distribution of multimapping, unique and total signal. **(C)** Nascent RNA 5' and 3' ends of PRO-cap and PRO-seq respectively and MutPE-seq analysis at the Ramos^{B1-8hi} *IGH* locus. MutPE-seq was performed following infection with AID(JP8Bdel)-expressing lentiviruses for 7 days. **(D)** Details of the mutation spectrum of the B1-8^{hi} variable region expressed from the Ramos *IGH* locus obtained by MutPE-seq. The bar graphs on the right show the percentage of each indicated mutation category (see Fig. 3E and 3G for details). For comparison, the murine B1-8^{hi} mutation spectra from primary murine B cells (middle panel) and murine GCBs (bottom panel), exactly as in Fig. 5B-C, are included.

Figure 7: Integrative model for co-transcriptional AID targeting and differential motif mutability. **(A)** 3D structure of the human RNA Pol II elongation complex visualized with ChimeraX. We superimposed, via their RPB1 subunits, the elongation complex structure (PDB ID 6GMH) onto the transcribing Pol II-DSIF (a complex of SPT5 and SPT4) structure (PDB ID 5OIK). Proteins are shown as surfaces (Pol II, grey; DSIF, salmon; SPT6, brown; PAF complex, yellow) and nucleic acids as cartoons (DNA template strand, blue; DNA non-template strand, cyan; RNA, red). The right panel highlights the trajectory of DNA and RNA buried within Pol II. The exposed non-template strand of the transcription bubble is occluded from interactions with AID due to it being completely covered by SPT5. **(B)** Model for SHM. The cartoon diagram of the Pol II elongation complex reflects the actual structure described in (A) above and in the main text. AID is recruited to the Pol II complex via super-enhancers in the context of a transcriptional hub

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(step 1) and is then retained in the elongation complex via interactions with elongation factors, SPT5, PAF and SPT6 (step 2). We leave open the possibility that multiple AID molecules can be present in the hub via association with different elongation factors which would serve to increase the local concentration of AID. The next step is the availability of ssDNA (step 3). Importantly, since SPT5 covers the exposed ssDNA, AID has no access to this ssDNA. Sources of ssDNA include upstream negative supercoils (not shown) which may allow AID to target both strands. Although R loops do not appear to form in variable regions, the RNA exosome is necessary to ensure normal distribution of mutations on both strands during SHM which implies that RNA processing is important in SHM. Hence, we propose that transcription termination may be an important source of ssDNA. Upon dissociation of Pol II, the non-template strand is available since the DNA:RNA hybrid would prevent immediate reannealing. The RNA is removed by RNA helicases and RNaseH followed by degradation of the RNA by the RNA exosome, which is known to associate with AID and provide access to the template strand. AID can bind ssDNA in a sequence-independent manner with high affinity and remain bound for several minutes (high on-rate, step 4). Deamination would occur occasionally on exposed WRCH motifs. Importantly, we hypothesize that the probability of deamination is strongly influenced by the sequence context of the WRCH motif and that this may be due to differential off-rates of AID in different sequence contexts. Thus, some WRCH motifs are more frequently deaminated to uridines (U) when they are embedded in a favorable nucleotide context that retains AID on ssDNA for longer periods (step 4, right) than other contexts where deamination is inefficient because of the higher dissociation rate of AID (step 4, left).

Figure S1: ChIP-qPCR analysis of the Ramos variable region locale and generation of Ramos Eμ$^{-/-}$ cells. (A) ChIP-qPCR analysis from three independent experiments in AID$^{-/-}$ Ramos cells for the indicated epigenetic marks as well as histone H3. Amplicons (1-7) used are indicated below the locus diagram shown above the graphs. The Neg. amplicon corresponds to a gene desert on chromosome 1 and is used as a negative control. **(B)** ChIP-qPCR analysis from three independent experiments in B1-8^{hi} primary splenic cells for the indicated epigenetic marks as well as histone H3. Amplicons (1-6) used are indicated below the locus diagram shown above the graphs. The Neg. (negative region) amplicon

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

corresponds to a gene desert on chromosome 1 and is used as a negative control. **(C)** Strategy to create the Eμ$^{-/-}$ AID$^{-/-}$ Ramos lines. The deleted region (583 bp) corresponds to the peak of accessible chromatin detected by ATAC-seq (top panel). This segment was replaced with a floxed reporter cassette expressing GFP or mCherry using CRISPR. Single clones double-positive for GFP and mCherry expression were isolated followed by excision of the floxed cassette by Cre recombinase. Two clones, c1 and c2 were used for all experiments. **(D)** Genotyping PCR analysis to confirm the loss of Eμ in AID$^{-/-}$ Eμ$^{-/-}$ clones c1 and c2. The location of primers is shown in the diagram above the gel image. **(E)** Surface IgM expression in Eμ$^{-/-}$ AID$^{-/-}$ Ramos clones relative to the parental AID$^{-/-}$ line determined by FACS. **(F)** Analysis of nascent transcription 3' ends (PRO-seq) at the *IGH* locus in AID$^{-/-}$ and AID$^{-/-}$ Eμ$^{-/-}$ Ramos cells (c1, c2). The promoter (Prom.), complementary determining regions (CDR 1-3), intronic enhancer (Eμ) and switch μ region (Sμ). **(G)** As in F but showing a magnified view of the variable region locale.

Figure S2: Generation of Ramos cell lines expressing exogenous variable region. (A) Schematic representation of the workflow (described in Methods) for generating Ramos cells expressing new variable regions using CRISPR-based editing. An IgM-negative line was made by excising the endogenous variable region and replacing it with a unique small guide RNA (sgRNA)-targeting sequence (green). Subsequently, an sgRNA targeting this site is combined with Cas9 and homology repair templates harboring any new variable regions to restore IgM expression. In essence, this system uses the restoration of IgM expression, which can occur only upon correct integration of the new variable regions, as a rapid and sensitive means to identify correctly targeted clones via flow cytometry (Fig. S4B). **(B)** Flow cytometry analysis of Ramos cell clones expressing new human and mouse variable regions used in this study (see scheme in Fig. 2C). Starting from AID$^{-/-}$ cells, the endogenous variable region was deleted (AID$^{-/-}$ΔV) followed by re-insertion of new variable regions. Shown are three clones of human VH4-59-DH2-JH6 and human VH3-30-DH2-JH6 expressing Ramos cells, and two clones of B1-8^{hi} expressing Ramos cells. **(C)** PRO-seq analysis showing the 3' ends of aligned reads at the human VH4-59-DH2-JH6 variable region expressed from the VH4-34 promoter at the human *IGH* in Ramos cells. Tracks of Multimapping, uniquely mapping and total are shown (see Fig. 3A and 3B for detailed description). **(D)**

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

PRO-seq analysis as in C of the human VH3-30-DH2-JH6 variable region expressed from the VH4-34 promoter at the human *IGH* in Ramos cells. **(E)** PRO-seq analysis as in C at the murine B1-8^{hi} variable region at the *Igh* locus in mice.

Figure S3: Analysis of PRO-cap, PRO-seq and SHM at non-Ig AID target genes in murine GCBs. **(A)** 5' and 3' ends of nascent RNA (PRO-cap, PRO-seq) and mutation profiles (from Alvarez-Prado et al. (2018) at selected AID target genes as in Fig. 5D-G. Shown are genes where highly mutated residues lie within 150 bp of the transcription initiation site defined by the peak of PRO-cap signals. The WRCH motifs are indicated as red dots. (B) As in A but showing genes where highly mutated residues lie near the transcription initiation site defined by the peak of PRO-cap signals.

Figure S4: Additional examples of PRO-cap, PRO-seq and SHM at non-Ig AID target genes in murine GCBs. See Fig. 5D for detailed legend.

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

1. S. P. Methot, J. M. Di Noia, Molecular Mechanisms of Somatic Hypermutation and Class Switch Recombination. *Adv Immunol* **133**, 37-87 (2017).
2. J. U. Peled *et al.*, The Biochemistry of Somatic Hypermutation. *Annu Rev Immunol* **26**, 481-511 (2008).
3. V. H. Odegard, D. G. Schatz, Targeting of somatic hypermutation. *Nat Rev Immunol* **6**, 573-583 (2006).
4. G. D. Victora, M. C. Nussenzweig, Germinal centers. *Annu Rev Immunol* **30**, 429-457 (2012).
5. M. Muramatsu *et al.*, Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* **102**, 553-563 (2000).
6. P. Revy *et al.*, Activation-induced cytidine deaminase (AID) deficiency causes the autosomal recessive form of the Hyper-IgM syndrome (HIGM2). *Cell* **102**, 565-575 (2000).
7. S. K. Petersen-Mahrt, R. S. Harris, M. S. Neuberger, AID mutates E. coli suggesting a DNA deamination mechanism for antibody diversification. *Nature* **418**, 99-103 (2002).
8. A. R. Ramiro, P. Stavropoulos, M. Jankovic, M. C. Nussenzweig, Transcription enhances AID-mediated cytidine deamination by exposing single-stranded DNA on the nontemplate strand. *Nat Immunol* **4**, 452-456 (2003).
9. R. Bransteitter, P. Pham, M. D. Scharff, M. F. Goodman, Activation-induced cytidine deaminase deaminates deoxycytidine on single-stranded DNA but requires the action of RNase. *Proc Natl Acad Sci U S A* **100**, 4102-4107 (2003).
10. P. Pham, R. Bransteitter, J. Petruska, M. F. Goodman, Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. *Nature* **424**, 103-107 (2003).
11. J. Chaudhuri *et al.*, Transcription-targeted DNA deamination by the AID antibody diversification enzyme. *Nature* **422**, 726-730 (2003).
12. B. Vaidyanathan, W.-F. Yen, J. Pucella, J. Chaudhuri, AIDing Chromatin and Transcription-Coupled Orchestration of Immunoglobulin Class-Switch Recombination. *Frontiers in Immunology* **5**, (2014).
13. R. J. Leeman-Neill, J. Lim, U. Basu, The Common Key to Class-Switch Recombination and Somatic Hypermutation: Discovery of AID and Its Role in Antibody Gene Diversification. *The Journal of Immunology* **201**, 2527-2529 (2018).
14. T. Saha, D. Sundaravinayagam, M. Di Virgilio, Charting a DNA Repair Roadmap for Immunoglobulin Class Switch Recombination. *Trends Biochem Sci*, (2020).
15. J. Chaudhuri, F. W. Alt, Class-switch recombination: interplay of transcription, DNA deamination and DNA repair. *Nat Rev Immunol* **4**, 541-552 (2004).
16. M. Liu *et al.*, Two levels of protection for the B cell genome during somatic hypermutation. *Nature* **451**, 841-845 (2008).
17. A. F. Alvarez-Prado *et al.*, A broad atlas of somatic hypermutation allows prediction of activation-induced deaminase targets. *J Exp Med* **215**, 761-771 (2018).
18. A. Yamane *et al.*, Deep-sequencing identification of the genomic targets of the cytidine deaminase AID and its cofactor RPA in B lymphocytes. *Nat Immunol*, (2010).
19. H. M. Shen, A. Peters, B. Baron, X. Zhu, U. Storb, Mutation of BCL-6 gene in normal B cells by the process of somatic hypermutation of Ig genes. *Science* **280**, 1750-1752 (1998).
20. L. Pasqualucci *et al.*, Hypermutation of multiple proto-oncogenes in B-cell diffuse large-cell lymphomas. *Nature* **412**, 341-346 (2001).

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

21. S. P. Methot *et al.*, A licensing step links AID to transcription elongation for mutagenesis in B cells. *Nat Commun* **9**, 1248 (2018).
22. R. Pavri *et al.*, Activation-induced cytidine deaminase targets DNA at sites of RNA polymerase II stalling by interaction with Spt5. *Cell* **143**, 122-133 (2010).
23. K. L. Willmann *et al.*, A role for the RNA pol II-associated PAF complex in AID-induced immune diversification. *J Exp Med* **209**, 2099-2111 (2012).
24. N. A. Begum, A. Stanlie, M. Nakata, H. Akiyama, T. Honjo, The histone chaperone Spt6 is required for activation-induced cytidine deaminase target determination through H3K4me3 regulation. *J Biol Chem* **287**, 32415-32429 (2012).
25. Y. Nambu *et al.*, Transcription-coupled events associating with immunoglobulin switch region chromatin. *Science* **302**, 2137-2140 (2003).
26. U. Basu *et al.*, The RNA exosome targets the AID cytidine deaminase to both strands of transcribed duplex DNA substrates. *Cell* **144**, 353-363 (2011).
27. U. Nowak, A. J. Matthews, S. Zheng, J. Chaudhuri, The splicing regulator PTBP2 interacts with the cytidine deaminase AID and promotes binding of AID to switch-region DNA. *Nat Immunol* **12**, 160-166 (2011).
28. Z. Duan *et al.*, Role of Dot1L and H3K79 methylation in regulating somatic hypermutation of immunoglobulin genes. *Proc Natl Acad Sci U S A* **118**, (2021).
29. G. Yu *et al.*, The role of HIRA-dependent H3.3 deposition and its modifications in the somatic hypermutation of immunoglobulin variable regions. *Proc Natl Acad Sci U S A* **118**, (2021).
30. M. Aida, N. Hamad, A. Stanlie, N. A. Begum, T. Honjo, Accumulation of the FACT complex, as well as histone H3.3, serves as a target marker for somatic hypermutation. *Proc Natl Acad Sci U S A* **110**, 7784-7789 (2013).
31. B. P. Jeevan-Raj *et al.*, Epigenetic tethering of AID to the donor switch region during immunoglobulin class switch recombination. *J Exp Med* **208**, 1649-1660 (2011).
32. A. Stanlie, M. Aida, M. Muramatsu, T. Honjo, N. A. Begum, Histone3 lysine4 trimethylation regulated by the facilitates chromatin transcription complex is critical for DNA cleavage in class switch recombination. *Proc Natl Acad Sci U S A* **107**, 22190-22195 (2010).
33. S. P. Bradley, D. A. Kaminski, A. H. F. M. Peters, T. Jenuwein, J. Stavnezer, The Histone Methyltransferase Suv39h1 Increases Class Switch Recombination Specifically to IgA. *The Journal of Immunology* **177**, 1179-1188 (2006).
34. F. L. Kuang, Z. Luo, M. D. Scharff, H3 trimethyl K9 and H3 acetyl K9 chromatin modifications are associated with class switch recombination. *Proceedings of the National Academy of Sciences* **106**, 5288-5293 (2009).
35. J. A. Daniel *et al.*, PTIP promotes chromatin changes critical for immunoglobulin class switch recombination. *Science* **329**, 917-923 (2010).
36. B. Vaidyanathan, J. Chaudhuri, Epigenetic Codes Programing Class Switch Recombination. *Front Immunol* **6**, 405 (2015).
37. J. M. Buerstedde, J. Alinikula, H. Arakawa, J. J. McDonald, D. G. Schatz, Targeting of somatic hypermutation by immunoglobulin enhancer and enhancer-like sequences. *PLoS Biol* **12**, e1001831 (2014).
38. F. Senigl *et al.*, Topologically Associated Domains Delineate Susceptibility to Somatic Hypermutation. *Cell Rep* **29**, 3902-3915 e3908 (2019).
39. R. K. Dinesh *et al.*, Transcription factor binding at Ig enhancers is linked to somatic hypermutation targeting. *Eur J Immunol* **50**, 380-395 (2020).
40. J. Qian *et al.*, B cell super-enhancers and regulatory clusters recruit AID tumorigenic activity. *Cell* **159**, 1524-1537 (2014).

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

41. P. Rouaud *et al.*, The IgH 3' regulatory region controls somatic hypermutation in germinal center B cells. *J Exp Med* **210**, 1501-1507 (2013).
42. S. Peron *et al.*, AID-driven deletion causes immunoglobulin heavy chain locus suicide recombination in B cells. *Science* **336**, 931-934 (2012).
43. J. M. Di Noia, M. S. Neuberger, Molecular mechanisms of antibody somatic hypermutation. *Annu Rev Biochem* **76**, 1-22 (2007).
44. R. W. Maul, P. J. Gearhart, AID and somatic hypermutation. *Adv Immunol* **105**, 159-191 (2010).
45. R. Pavri, M. C. Nussenzweig, AID targeting in antibody diversity. *Adv Immunol* **110**, 1-26 (2011).
46. A. G. Betz, C. Rada, R. Pannell, C. Milstein, M. S. Neuberger, Passenger transgenes reveal intrinsic specificity of the antibody hypermutation mechanism: clustering, polarity, and specific hot spots. *Proc Natl Acad Sci U S A* **90**, 2385-2388 (1993).
47. J. K. Hwang *et al.*, Sequence intrinsic somatic mutation mechanisms contribute to affinity maturation of VRC01-class HIV-1 broadly neutralizing antibodies. *Proc Natl Acad Sci U S A* **114**, 8614-8619 (2017).
48. L. S. Yeap *et al.*, Sequence-Intrinsic Mechanisms that Target AID Mutational Outcomes on Antibody Genes. *Cell* **163**, 1124-1137 (2015).
49. J. Q. Zhou, S. H. Kleinstein, Position-Dependent Differential Targeting of Somatic Hypermutation. *The Journal of Immunology*, ji2000496 (2020).
50. N. Spisak, A. M. Walczak, T. Mora, Learning the heterogeneous hypermutation landscape of immunoglobulins from high-throughput repertoire data. *Nucleic Acids Research* **48**, 10702-10712 (2020).
51. C. Tang, A. Krantsevich, T. MacCarthy, Deep learning model of somatic hypermutation reveals importance of sequence context beyond hotspot targeting. *iScience* **25**, 103668 (2022).
52. J. Q. Zhou, S. H. Kleinstein, Position-Dependent Differential Targeting of Somatic Hypermutation. *J Immunol* **205**, 3468-3479 (2020).
53. F. L. Meng *et al.*, Convergent transcription at intragenic super-enhancers targets AID-initiated genomic instability. *Cell* **159**, 1538-1548 (2014).
54. D. Rajagopal *et al.*, Immunoglobulin switch mu sequence causes RNA polymerase II accumulation and reduces dA hypermutation. *J Exp Med* **206**, 1237-1244 (2009).
55. L. Wang, R. Wuerffel, S. Feldman, A. A. Khamlichi, A. L. Kenter, S region sequence, RNA polymerase II, and histone modifications create chromatin accessibility during class switch recombination. *J Exp Med* **206**, 1817-1830 (2009).
56. R. W. Maul *et al.*, Spt5 accumulation at variable genes distinguishes somatic hypermutation in germinal center B cells from ex vivo-activated cells. *J Exp Med* **211**, 2297-2306 (2014).
57. A. Tarsalainen *et al.*, Ig Enhancers Increase RNA Polymerase II Stalling at Somatic Hypermutation Target Sequences. *J Immunol* **208**, 143-154 (2022).
58. B. J. Taylor, Y. L. Wu, C. Rada, Active RNAP pre-initiation sites are highly mutated by cytidine deaminases in yeast, with AID targeting small RNA genes. *Elife* **3**, e03553 (2014).
59. X. Wang, M. Fan, S. Kalis, L. Wei, M. D. Scharff, A source of the single-stranded DNA substrate for activation-induced deaminase during somatic hypermutation. *Nature Communications* **5**, 4137 (2014).
60. P. Kodgire, P. Mukkawar, S. Ratnam, T. E. Martin, U. Storb, Changes in RNA polymerase II progression influence somatic hypermutation of Ig-related genes by AID. *J Exp Med* **210**, 1481-1492 (2013).
61. D. Ronai *et al.*, Detection of chromatin-associated single-stranded DNA in regions targeted for somatic hypermutation. *J Exp Med* **204**, 181-190 (2007).
62. H. Kwak, N. J. Fuda, L. J. Core, J. T. Lis, Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339**, 950-953 (2013).

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

63. J. E. Sale, M. S. Neuberger, TdT-accessible breaks are scattered over the immunoglobulin V domain in a constitutively hypermutating B cell line. *Immunity* **9**, 859-869 (1998).
64. E. Pinaud *et al.*, The IgH locus 3' regulatory region: pulling the strings from behind. *Adv Immunol* **110**, 27-70 (2011).
65. T. A. Shih, M. Roederer, M. C. Nussenzweig, Role of antigen receptor affinity in T cell-independent antibody responses in vivo. *Nat Immunol* **3**, 399-406 (2002).
66. S. Ito *et al.*, Activation-induced cytidine deaminase shuttles between nucleus and cytoplasm like apolipoprotein B mRNA editing catalytic polypeptide 1. *Proc Natl Acad Sci U S A* **101**, 1975-1980 (2004).
67. T. Perlot, F. W. Alt, C. H. Bassing, H. Suh, E. Pinaud, Elucidation of IgH intronic enhancer functions via germ-line deletion. *Proc Natl Acad Sci U S A* **102**, 14362-14367 (2005).
68. F. Li, Y. Yan, J. Pieretti, D. A. Feldman, L. A. Eckhardt, Comparison of identical and functional Igh alleles reveals a nonessential role for Emu in somatic hypermutation and class-switch recombination. *J Immunol* **185**, 6049-6057 (2010).
69. Y. Yamaguchi *et al.*, NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. *Cell* **97**, 41-51 (1999).
70. S. M. Vos, L. Farnung, H. Urlaub, P. Cramer, Structure of paused transcription complex Pol II–DSIF–NELF. *Nature* **560**, 601-606 (2018).
71. S. M. Vos *et al.*, Structure of activated transcription complex Pol II–DSIF–PAF–SPT6. *Nature* **560**, 607-612 (2018).
72. C. Bernecky, J. M. Plitzko, P. Cramer, Structure of a transcribing RNA polymerase II–DSIF complex reveals a multidentate DNA–RNA clamp. *Nature Structural & Molecular Biology* **24**, 809-815 (2017).
73. D. F. Robbiani *et al.*, Plasmodium Infection Promotes Genomic Instability and AID-Dependent B Cell Lymphoma. *Cell* **162**, 727-737 (2015).
74. H. Mouquet *et al.*, Complex-type N-glycan recognition by potent broadly neutralizing HIV antibodies. *Proc Natl Acad Sci U S A* **109**, E3268-3277 (2012).
75. V. Lorin *et al.*, Epitope convergence of broadly HIV-1 neutralizing IgA and IgG antibody lineages in a viremic controller. *J Exp Med* **219**, (2022).
76. F. Matsuda *et al.*, The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. *J Exp Med* **188**, 2151-2162 (1998).
77. T. A. Shih, E. Meffre, M. Roederer, M. C. Nussenzweig, Role of BCR affinity in T cell dependent antibody responses in vivo. *Nat Immunol* **3**, 570-575 (2002).
78. C. Rada, J. M. Di Noia, M. S. Neuberger, Mismatch recognition and uracil excision provide complementary paths to both Ig switching and the A/T-focused phase of somatic mutation. *Mol Cell* **16**, 163-171 (2004).
79. S. Longerich, A. Tanaka, G. Bozek, D. Nicolae, U. Storb, The very 5' end and the constant region of Ig genes are spared from somatic mutation because AID does not access these regions. *J Exp Med* **202**, 1443-1454 (2005).
80. K. Xue, C. Rada, M. S. Neuberger, The in vivo pattern of AID targeting to immunoglobulin switch regions deduced from mutation spectra in msh2-/- ung-/- mice. *J Exp Med* **203**, 2085-2094 (2006).
81. K. M. Dorighi *et al.*, Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation. *Mol Cell* **66**, 568-576 e564 (2017).
82. J. Kim *et al.*, Polycomb- and Methylation-Independent Roles of EZH2 as a Transcription Activator. *Cell Rep* **25**, 2808-2820.e2804 (2018).
83. T. Hoshii *et al.*, A Non-catalytic Function of SETD1A Regulates Cyclin K and the DNA Damage Response. *Cell* **172**, 1007-1021.e1017 (2018).

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

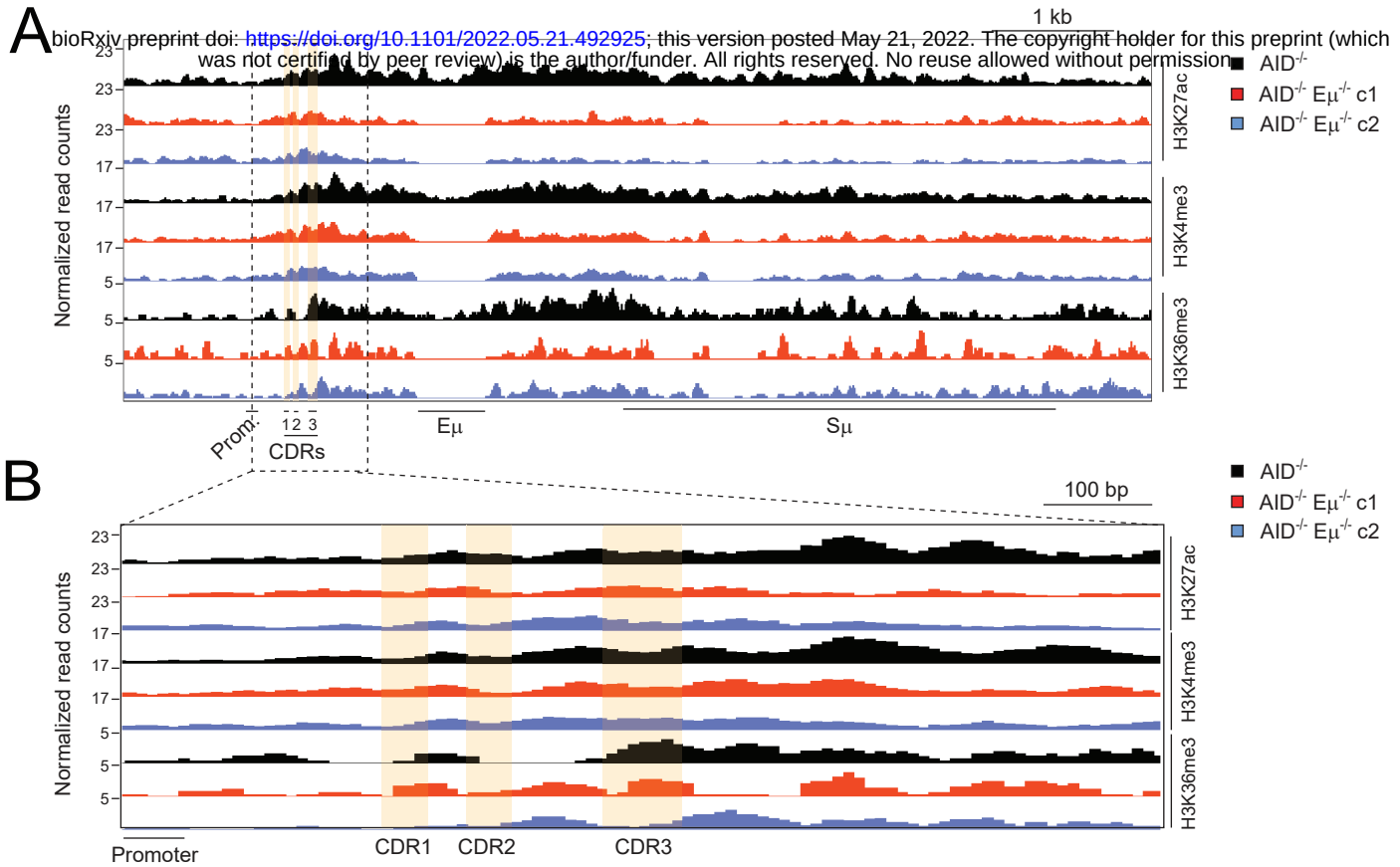
84. C. Wang *et al.*, UTX regulates mesoderm differentiation of embryonic stem cells independent of H3K27 demethylase activity. *Proc Natl Acad Sci U S A* **109**, 15324-15329 (2012).
85. Y. Aubert, S. Egolf, B. C. Capell, The Unexpected Noncatalytic Roles of Histone Modifiers in Development and Disease. *Trends Genet* **35**, 645-657 (2019).
86. X. Zhang *et al.*, Fundamental roles of chromatin loop extrusion in antibody class switching. *Nature* **575**, 385-389 (2019).
87. S. Chong *et al.*, Imaging dynamic and selective low-complexity domain interactions that control gene transcription. *Science* **361**, (2018).
88. X. Darzacq, R. Tjian, Weak multivalent biomolecular interactions: a strength versus numbers tug of war with implications for phase partitioning. *RNA* **28**, 48-51 (2022).
89. J. P. Karr, J. J. Ferrie, R. Tjian, X. Darzacq, The transcription factor activity gradient (TAG) model: contemplating a contact-independent mechanism for enhancer-promoter communication. *Genes Dev* **36**, 7-16 (2022).
90. B. Lim, M. S. Levine, Enhancer-promoter communication: hubs or loops? *Current Opinion in Genetics & Development* **67**, 5-9 (2021).
91. J. Y. Parsa *et al.*, Negative supercoiling creates single-stranded patches of DNA that are substrates for AID-mediated mutagenesis. *PLoS Genet* **8**, e1002518 (2012).
92. H. M. Shen, U. Storb, Activation-induced cytidine deaminase (AID) can target both DNA strands when the DNA is supercoiled. *Proc Natl Acad Sci U S A* **101**, 12997-13002 (2004).
93. B. Laffleur *et al.*, Noncoding RNA processing by DIS3 regulates chromosomal architecture and somatic hypermutation in B cells. *Nature Genetics* **53**, 230-242 (2021).
94. J. Lim *et al.*, Nuclear Proximity of Mtr4 to RNA Exosome Restricts DNA Mutational Asymmetry. *Cell* **169**, 523-537 e515 (2017).
95. D. Kazadi *et al.*, Effects of senataxin and RNA exosome on B-cell chromosomal integrity. *Heliyon* **6**, e03442 (2020).
96. L. Nair, H. Chung, U. Basu, Regulation of long non-coding RNAs and genome dynamics by the RNA surveillance machinery. *Nat Rev Mol Cell Biol* **21**, 123-136 (2020).
97. X. Wang, M. Fan, S. Kalis, L. Wei, M. D. Scharff, A source of the single-stranded DNA substrate for activation-induced deaminase during somatic hypermutation. *Nat Commun* **5**, 4137 (2014).
98. J. Sun *et al.*, E3-ubiquitin ligase Nedd4 determines the fate of AID-associated RNA polymerase II in B cells. *Genes Dev* **27**, 1821-1833 (2013).
99. S. Lykke-Andersen *et al.*, Integrator is a genome-wide attenuator of non-productive transcription. *Mol Cell* **81**, 514-529.e516 (2021).
100. M. Larijani *et al.*, AID associates with single-stranded DNA with high affinity and a long complex half-life in a sequence-independent manner. *Mol Cell Biol* **27**, 20-30 (2007).
101. P. Pham, P. Calabrese, S. J. Park, M. F. Goodman, Analysis of a Single-stranded DNA-scanning Process in Which Activation-induced Deoxycytidine Deaminase (AID) Deaminates C to U Haphazardly and Inefficiently to Ensure Mutational Diversity*◆. *Journal of Biological Chemistry* **286**, 24931-24942 (2011).
102. J. Fitz *et al.*, Spt5-mediated enhancer transcription directly couples enhancer activation with physical promoter interaction. *Nat Genet* **52**, 505-515 (2020).
103. D. B. Mahat *et al.*, Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat Protoc* **11**, 1455-1476 (2016).
104. J. Fitz, T. Neumann, R. Pavri, Regulation of RNA polymerase II processivity by Spt5 is restricted to a narrow window during elongation. *EMBO J* **37**, (2018).
105. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011* **17**, 3 (2011).

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

106. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).
107. A. Larkin *et al.*, FlyBase: updates to the Drosophila melanogaster knowledge base. *Nucleic Acids Research* **49**, D899-D907 (2020).
108. T. Smith, A. Heger, I. Sudbery, UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* **27**, 491-499 (2017).
109. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
110. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
111. T. Magoč, S. L. Salzberg, FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957-2963 (2011).
112. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357-359 (2012).
113. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

Figure 1

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



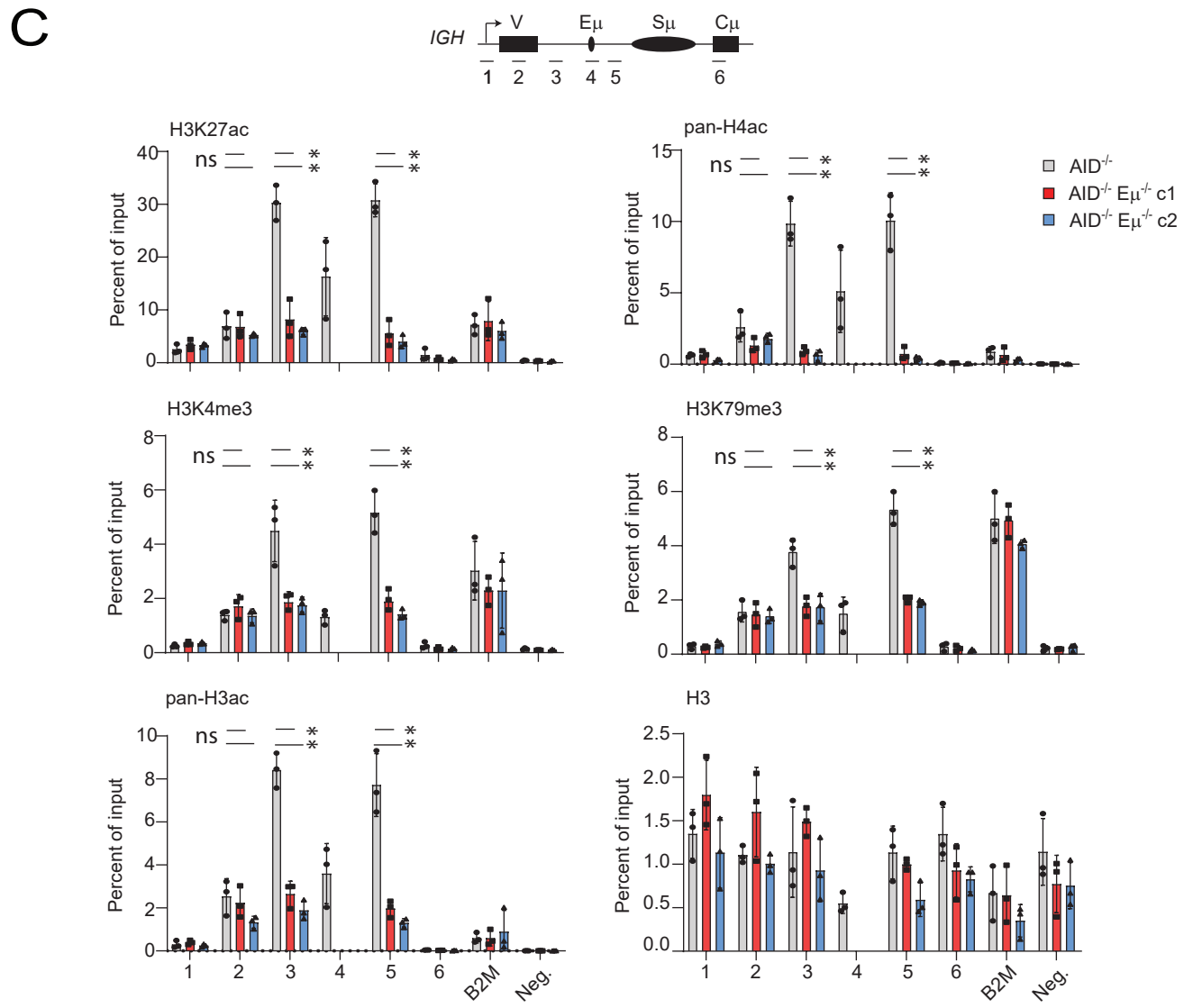
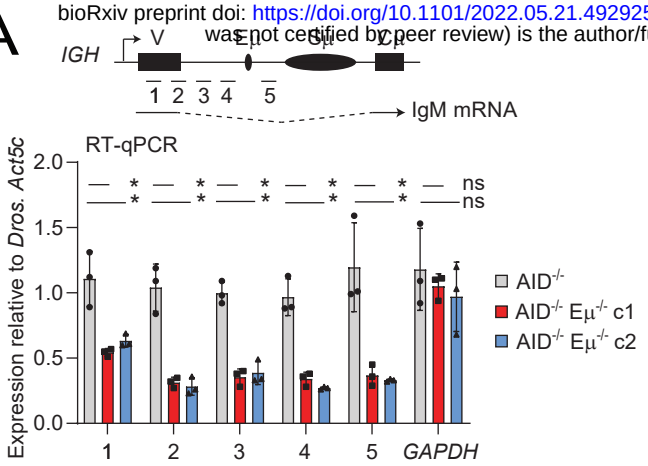


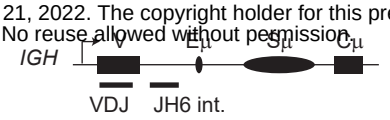
Figure 2

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A

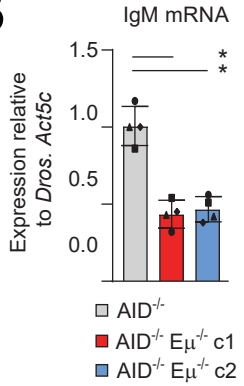


C

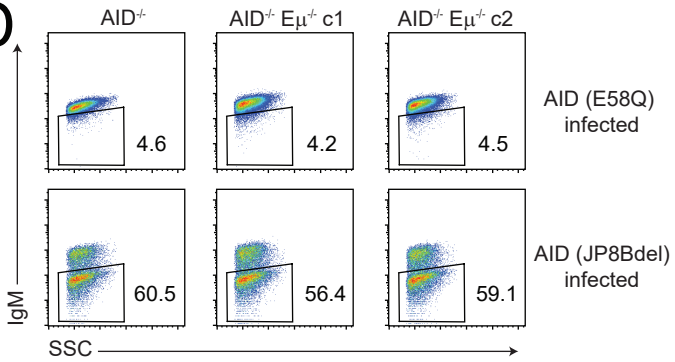


	Amplicon (cell line)	Nucleotides sequenced	Mutations	Mut. Freq.	*P* value
JP8Bdel infected	VDJ (AID$^{-/-}$)	50,863	95	1.86 x 10^{-3}	0.081
	VDJ (AID$^{-/-}$ Eμ$^{-/-}$) c2	55,245	85	1.52 x 10^{-3}	
	JH6 intron (AID$^{-/-}$)	56,253	91	1.60 x 10^{-3}	0.087
	JH6 intron (AID$^{-/-}$ Eμ$^{-/-}$) c2	59,805	80	1.33 x 10^{-3}	

B



D



E

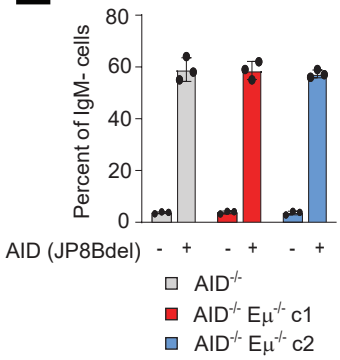


Figure 3

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

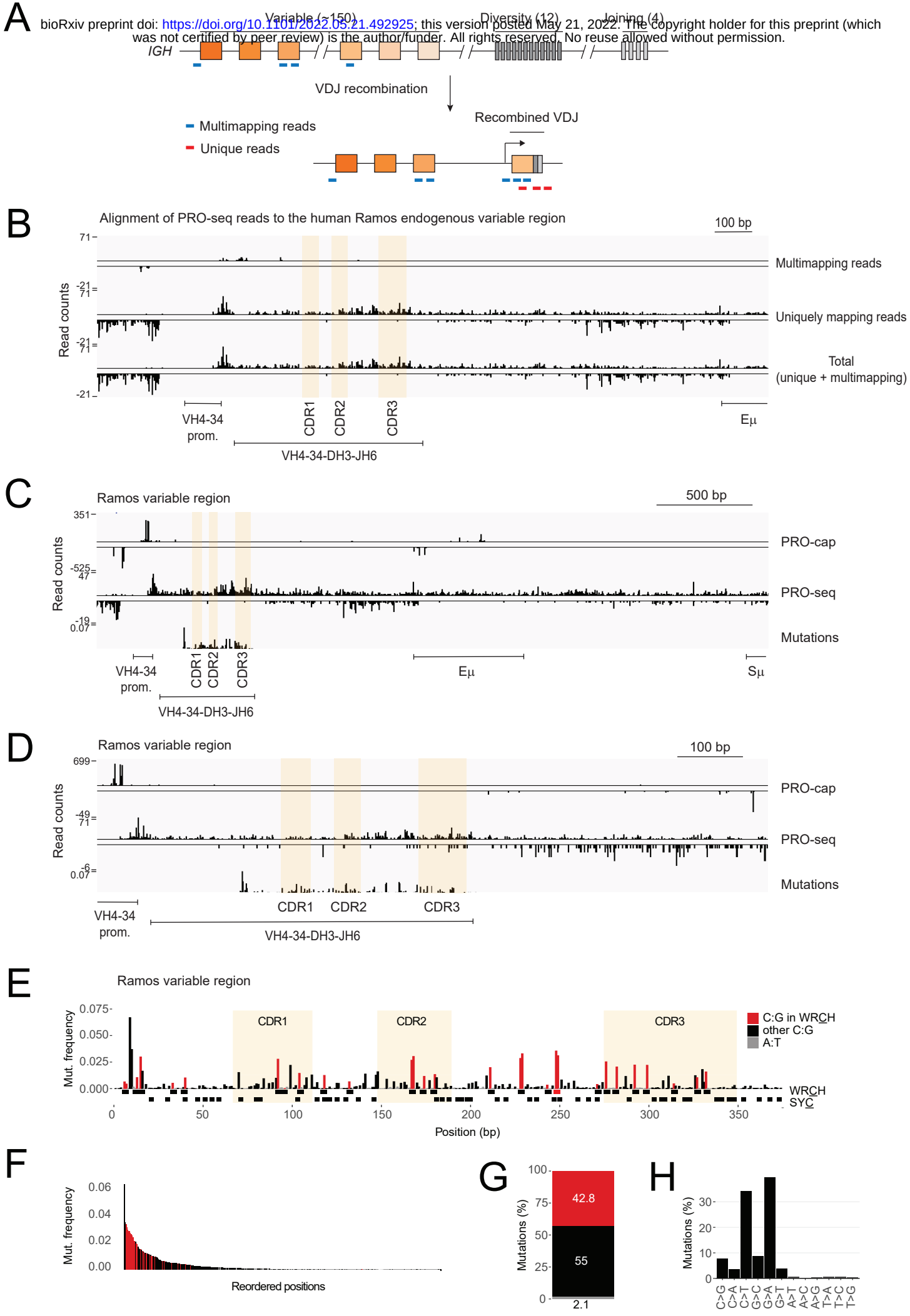


Figure 4

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

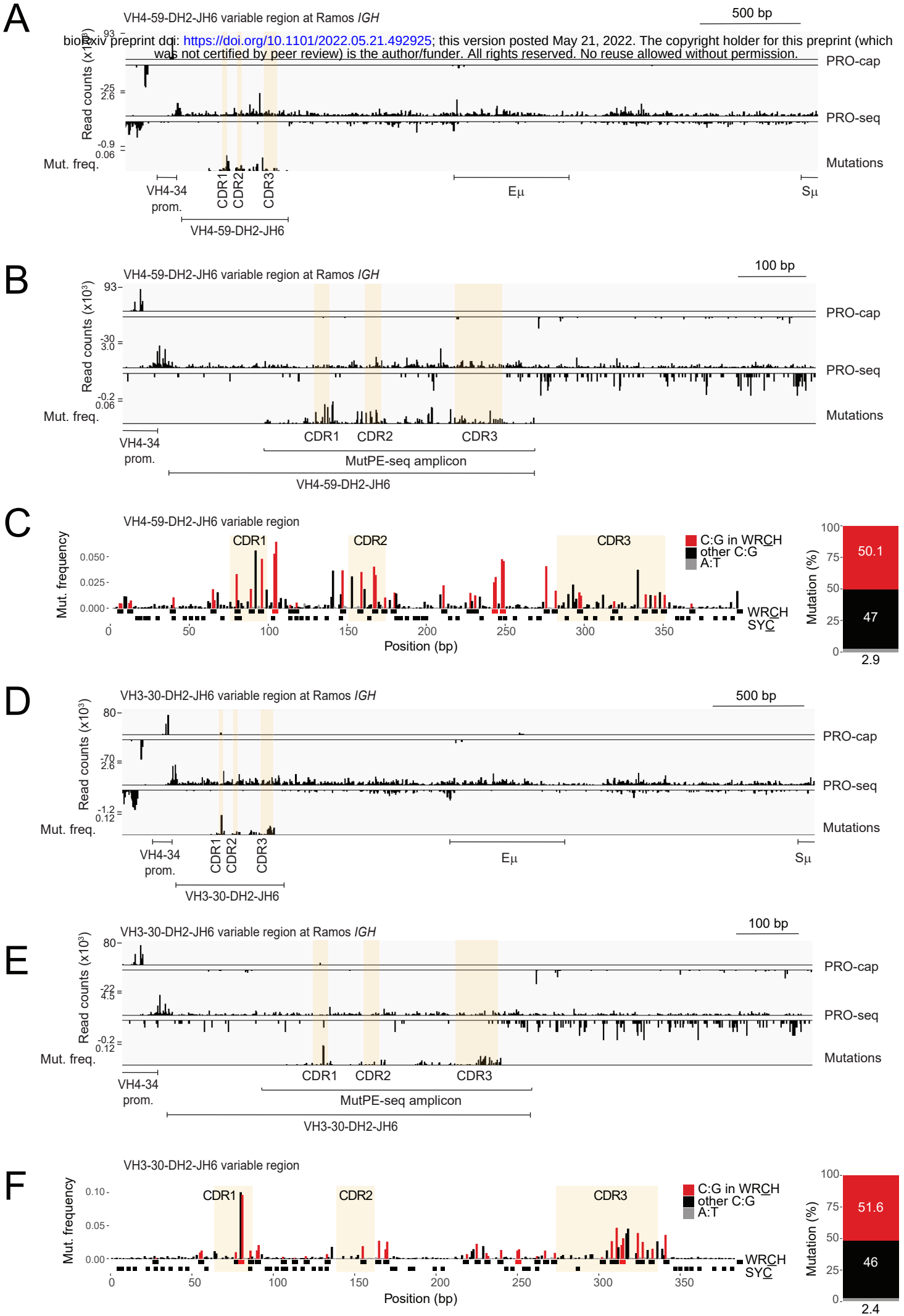
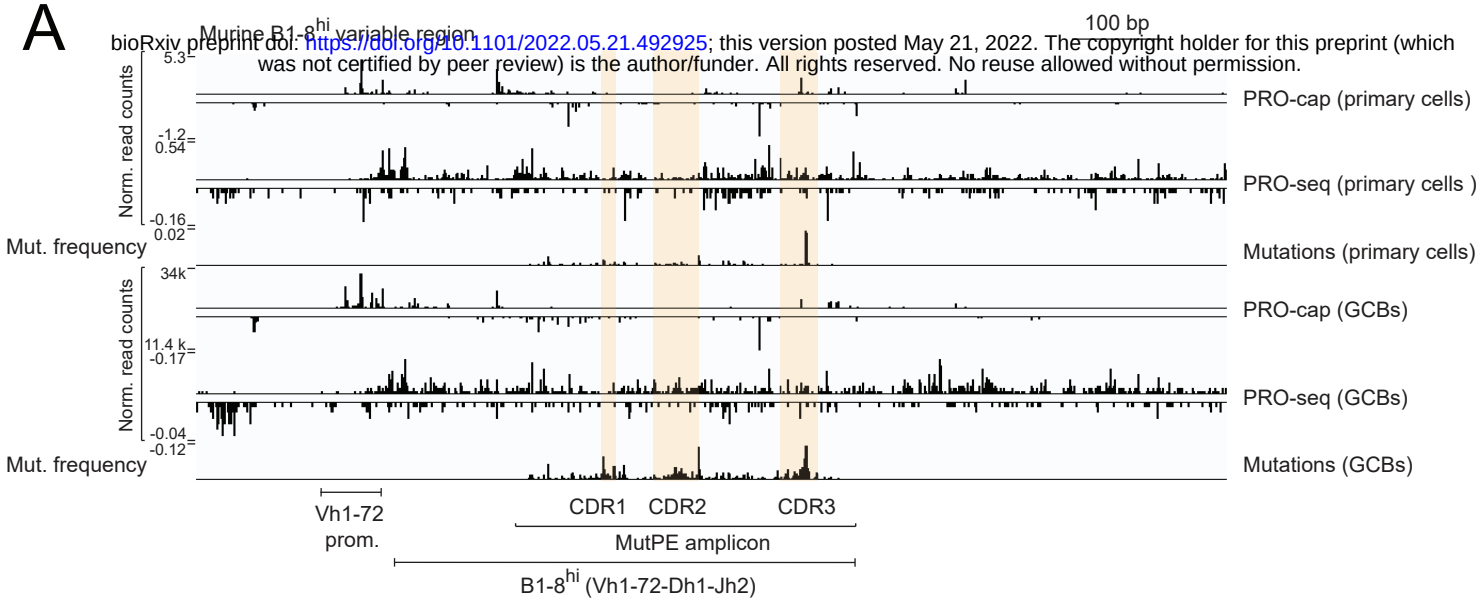


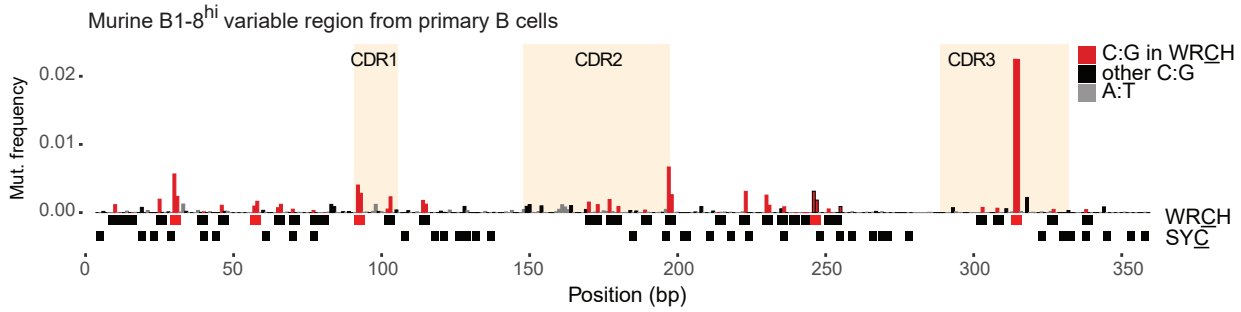
Figure 5

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

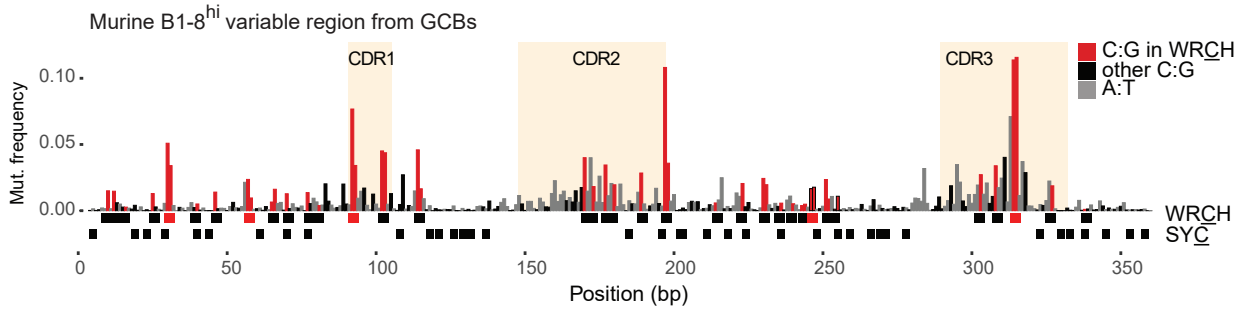
A



B



C



D

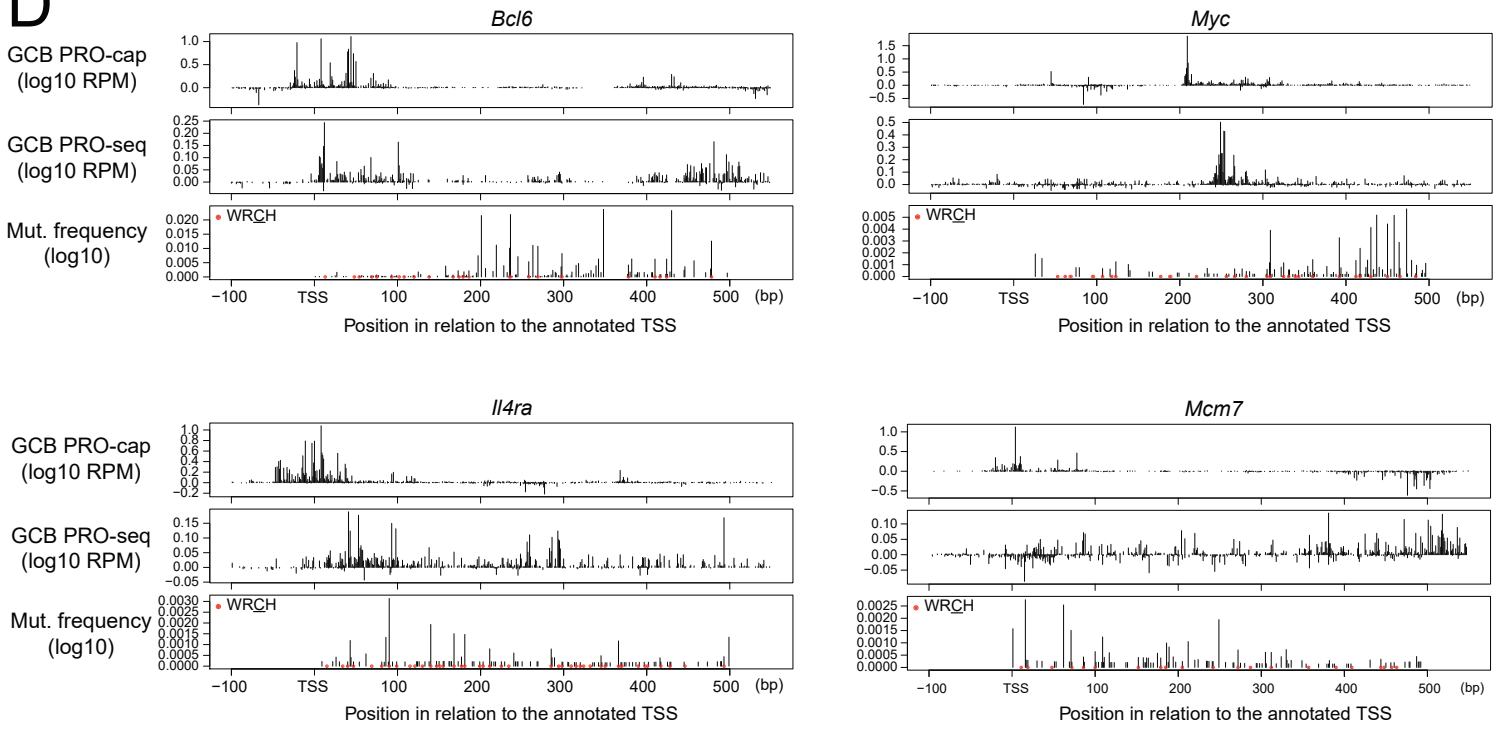


Figure 6

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

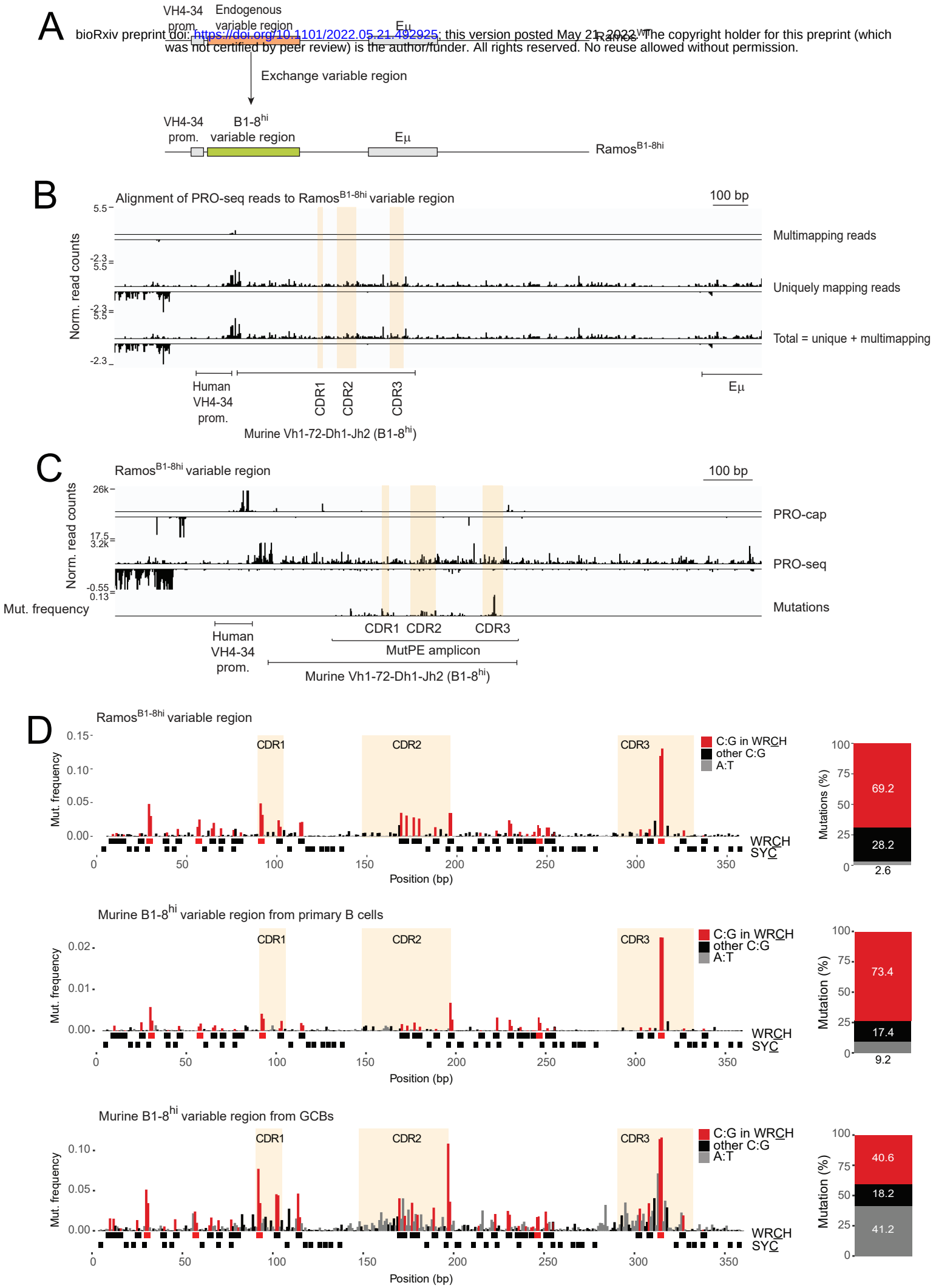
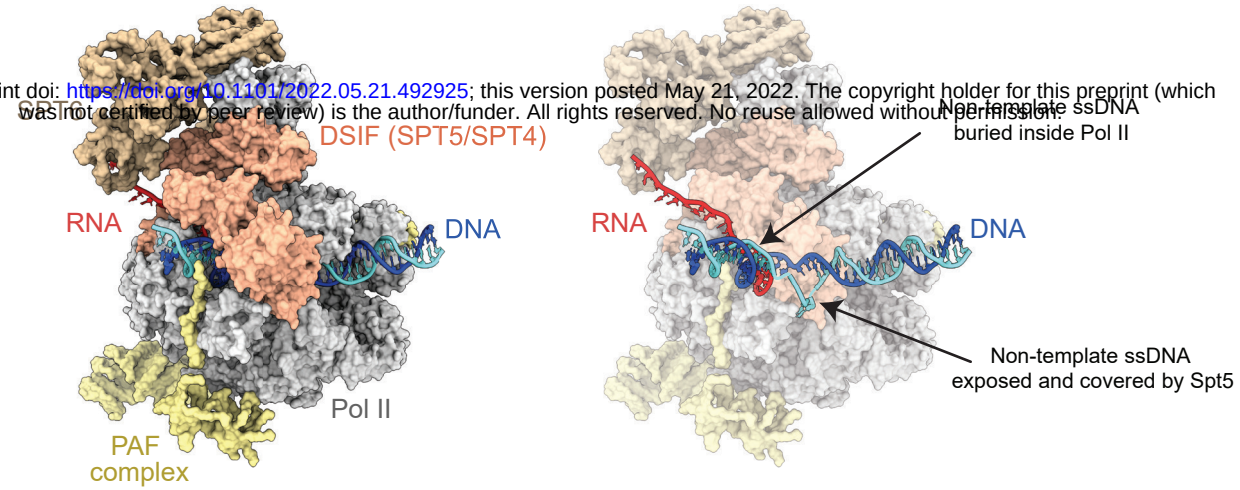


Fig. 7

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A



B

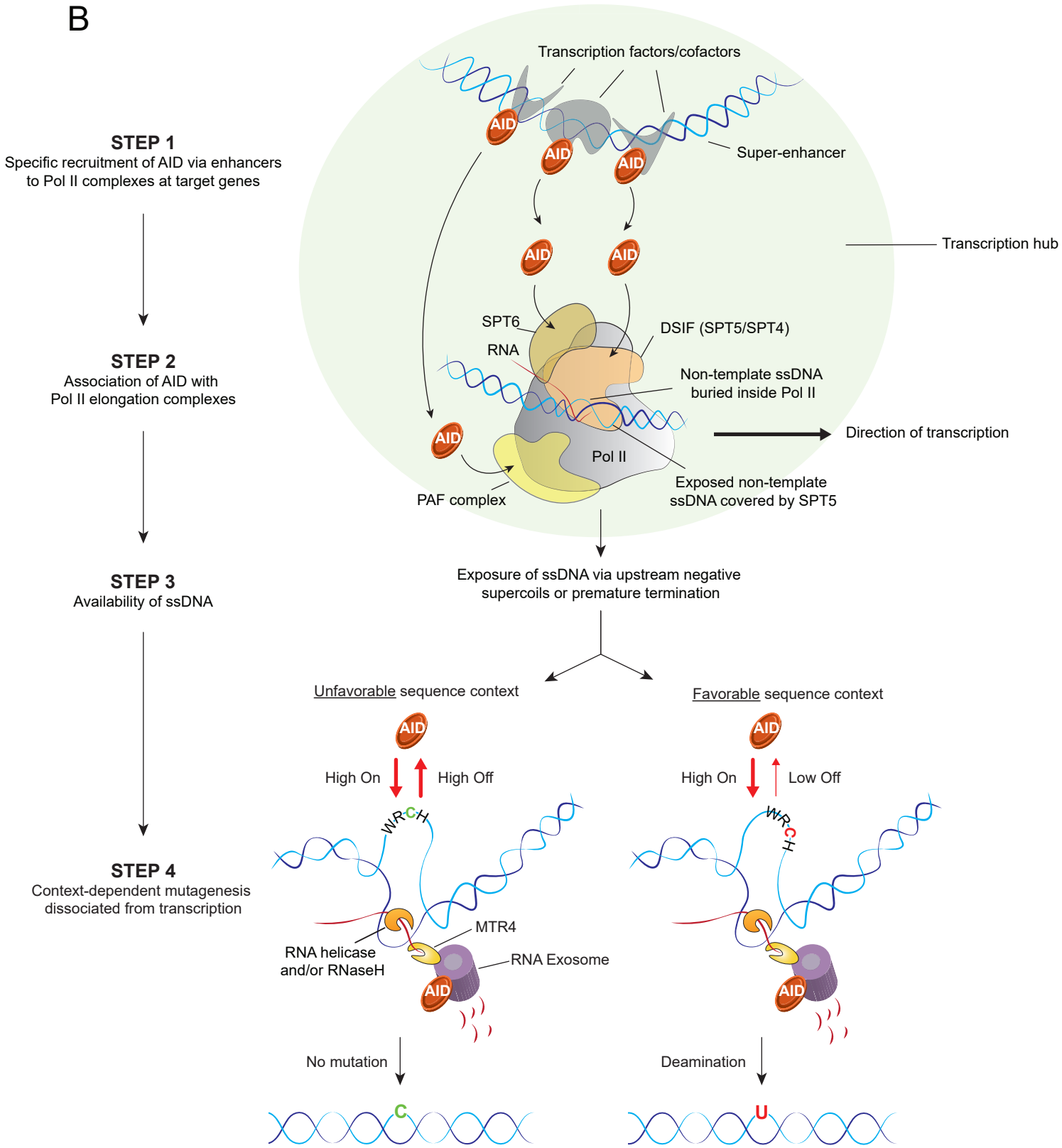


Figure S1

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

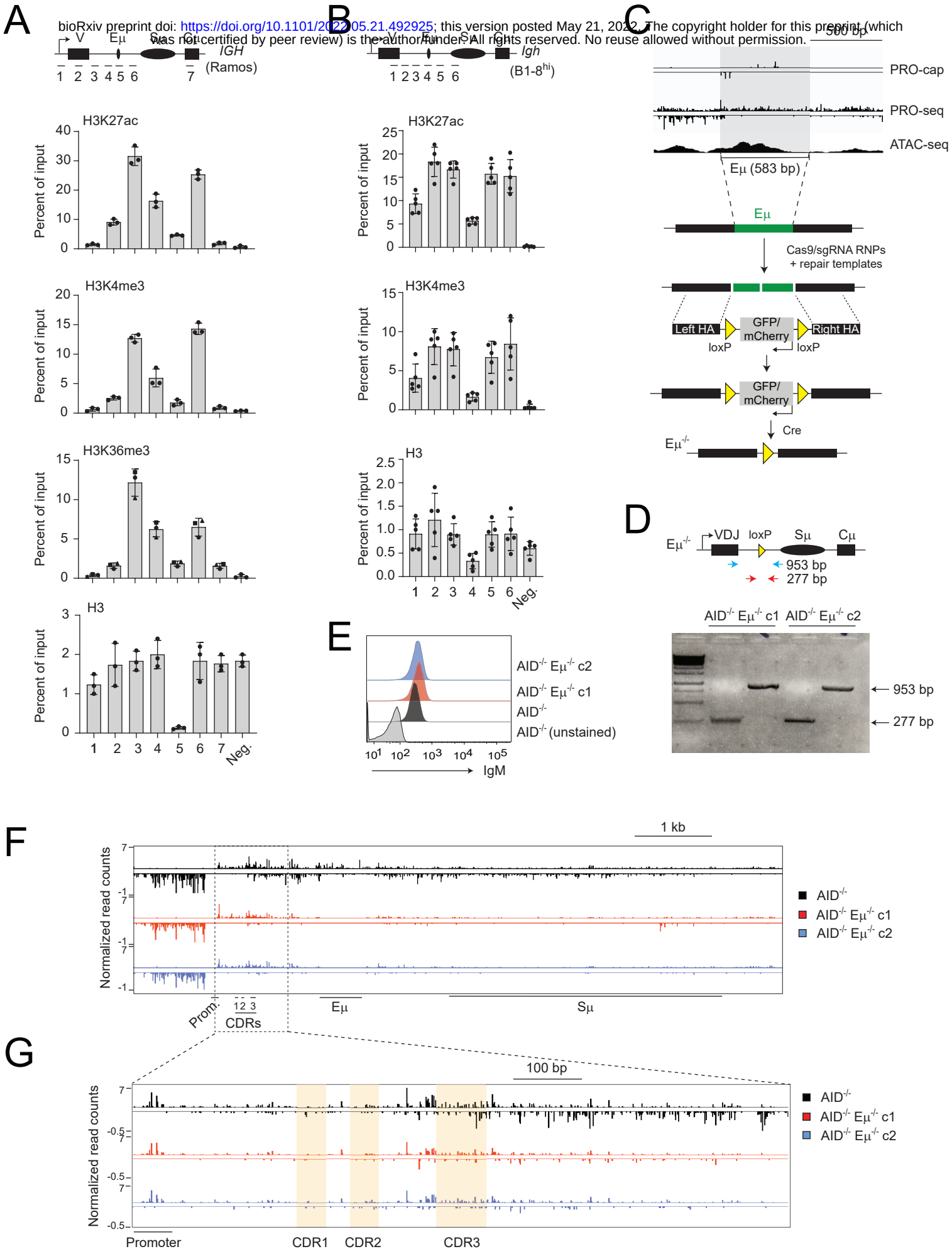
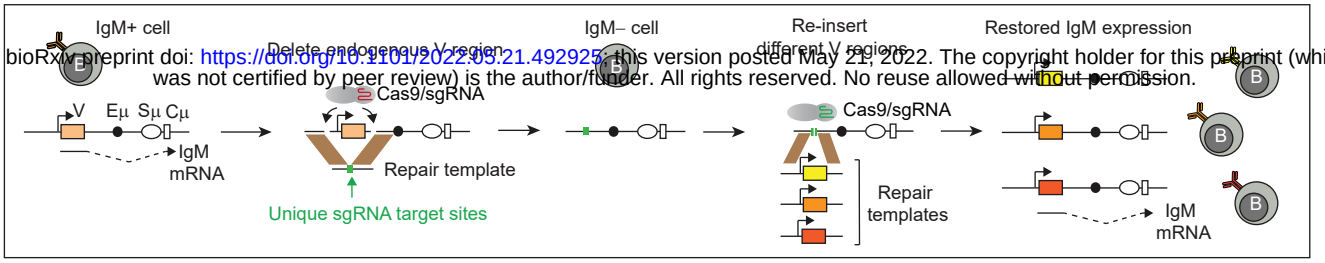


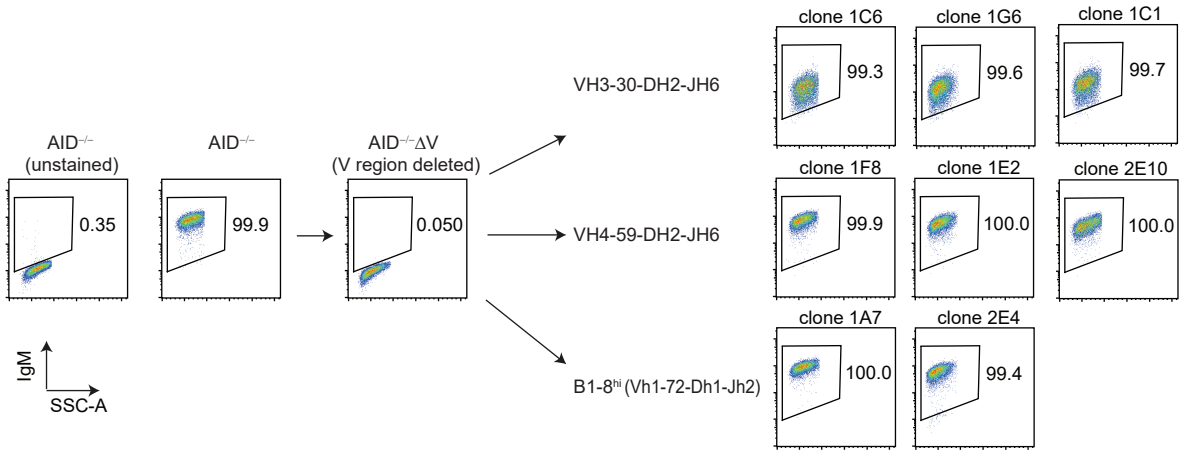
Figure S2

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925 this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

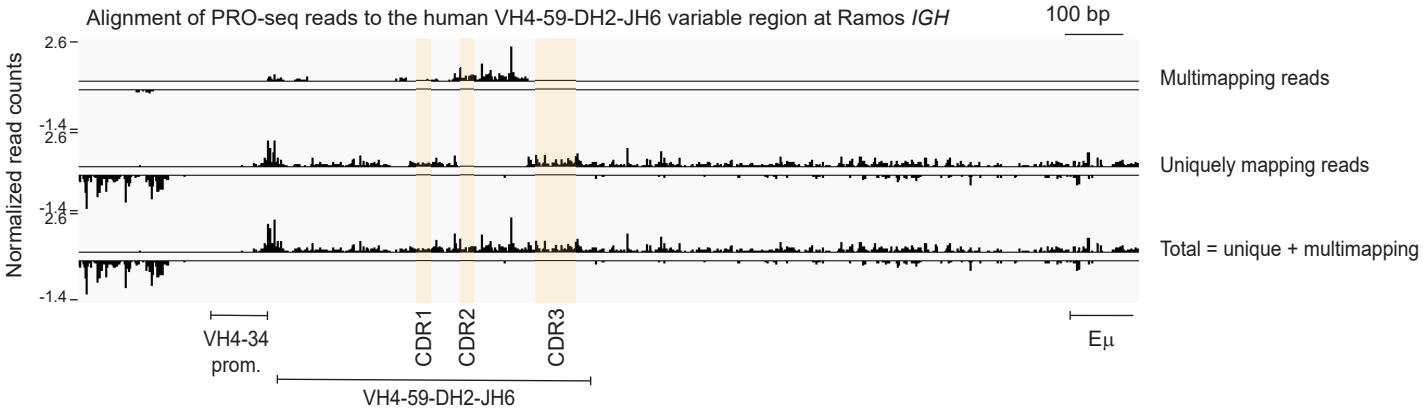
A



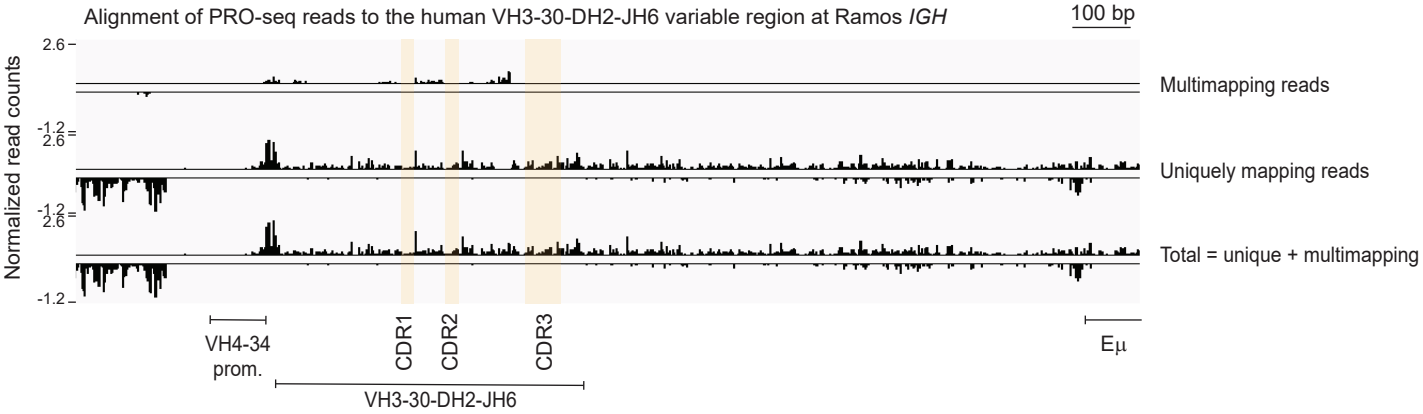
B



C



D



E

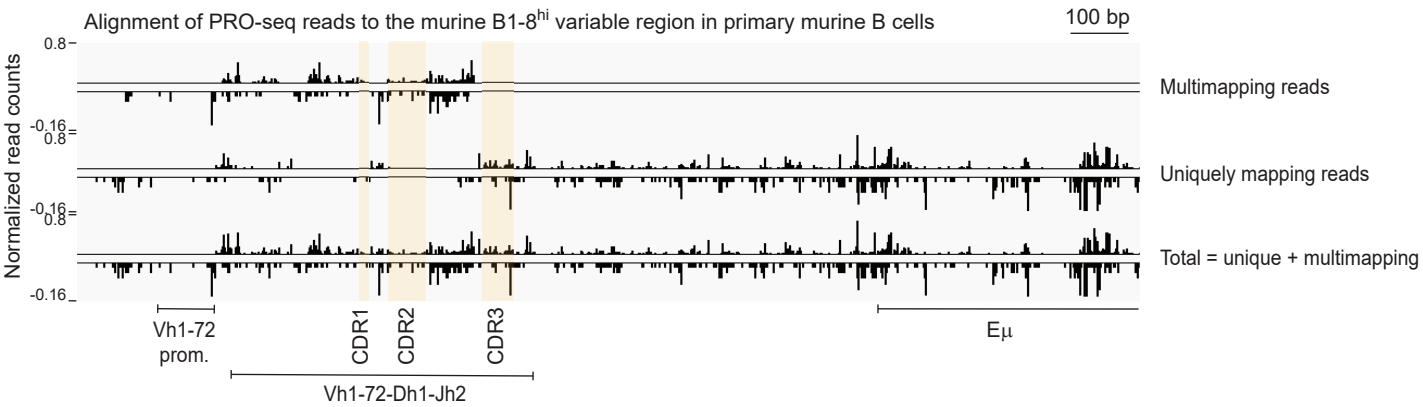
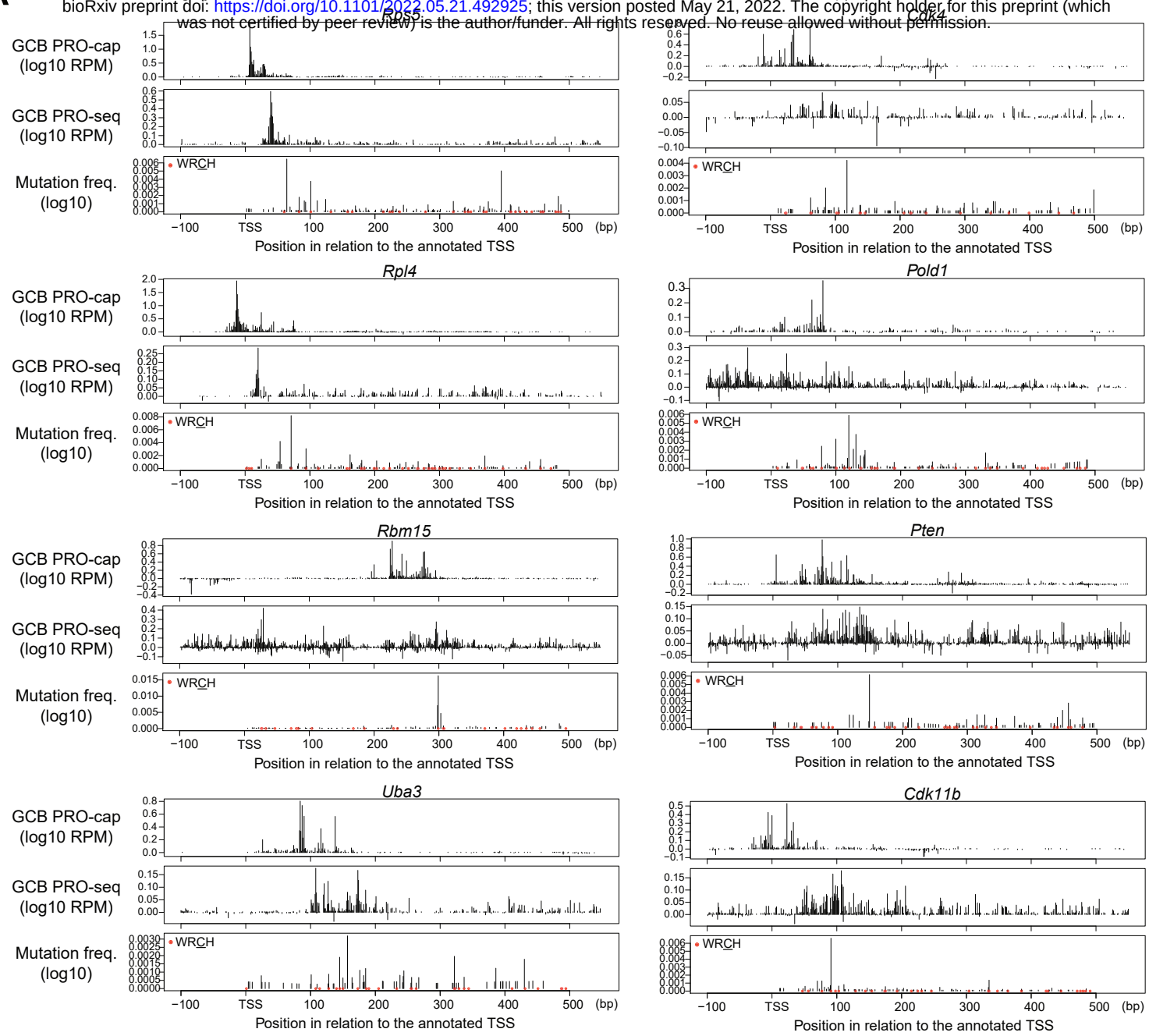


Fig. S3

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A

Highly mutated residues within 150 bp of the transcription initiation site (defined by PRO-cap)



B

Highly mutated residues near the transcription initiation site (defined by PRO-cap)

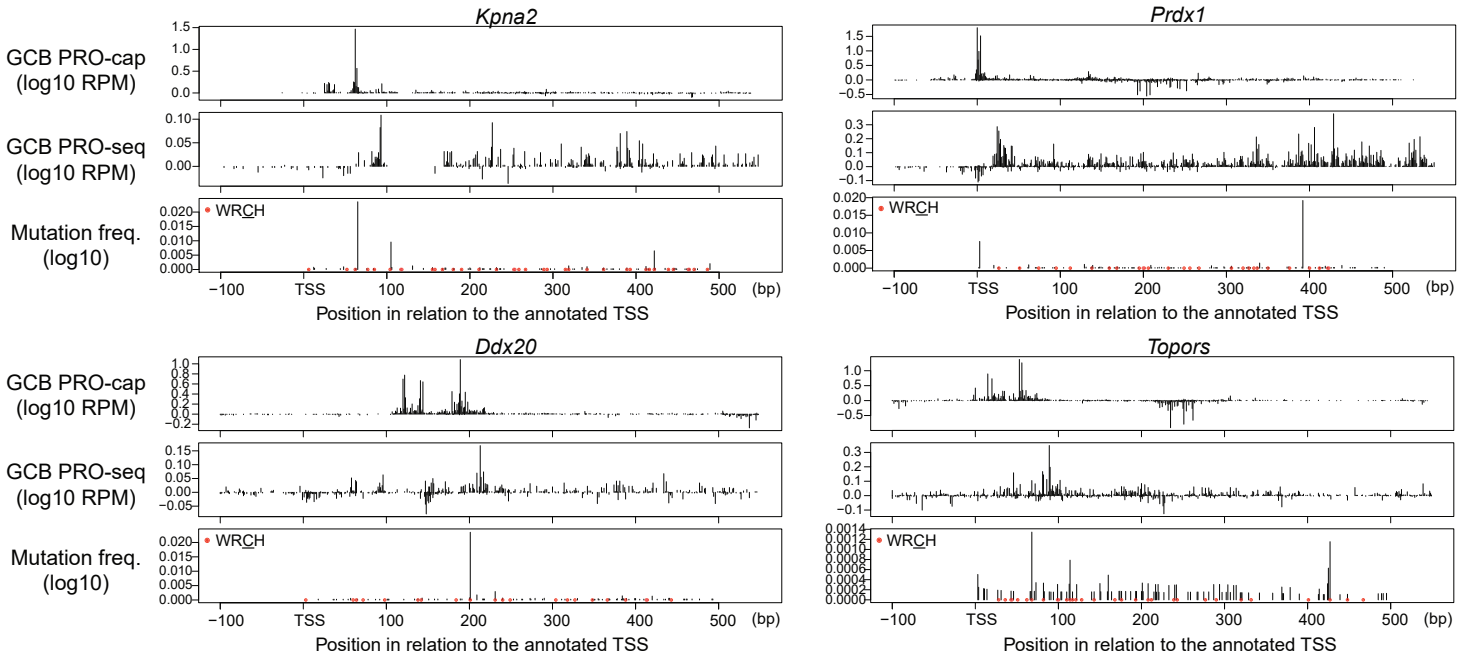


Fig. S4

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

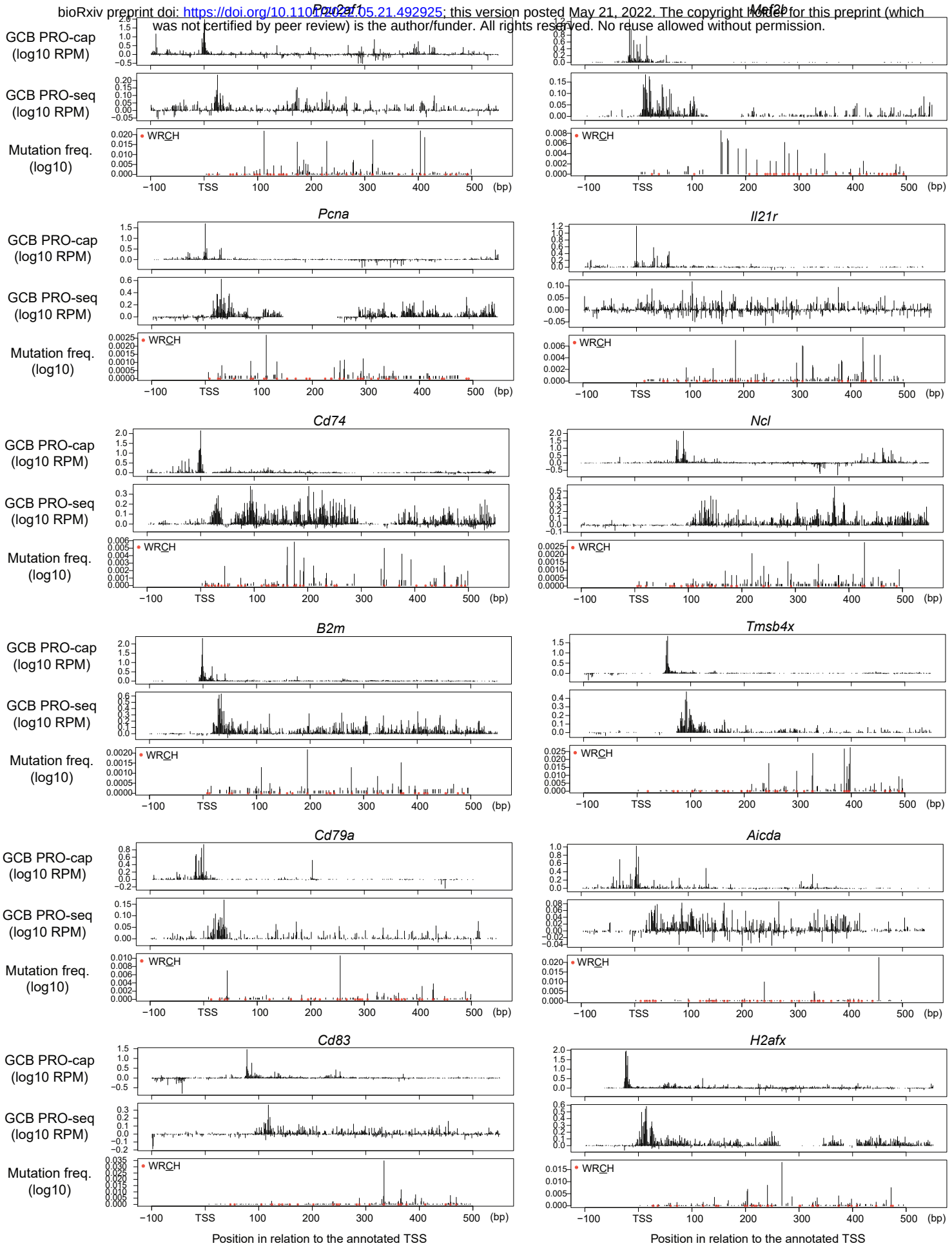


Table S1: List of primers, oligos and sgRNAs

sgRNAs for generating Ramos Eμ$^{-/-}$ cells		
sgRNA_Eμ1	CAGTGCTGTCGGCCCCGATG	
sgRNA_Eμ2	ACTGTGTTCACAATCTTTTT	
sgRNA_Eμ3	GAGGCTGACCGAAACTGAAA	

Genotyping primers for Ramos Eμ$^{-/-}$ cells	Forward	Reverse	Product sizes
gt_Eμ 9+10	ATGTGTCTGGAATTGAGGCCAA	TTTCTCTTATGCCACCAACTAACA	825 bp WT, 277 bp KO
gt_Eμ 11+12	ACAGACGGGAGGTACGGTAT	CCTTAATCACAAAGAGAAAACGAGT	1501 bp WT, 953 bp KO

sgRNAs for generating Ramos variable region knock-in lines		Reference
sgRNA-Ramos VDJ 5'	CCTTCAGCACATTTCCTACC	Voss et al., eLife, 2019
sgRNA-Ramos VDJ 3'	TCCTCGGGGCATGTTCCGAG	Voss et al., eLife, 2019
sgRNA-GFP	GAGCTGGACGGCGACGTAAA	This study

RT-qPCR primers (Ramos)	Forward	Reverse
1. Variable region	CAAGAACATGAAACACCTGTGG	ATCCGCATTCCTGAGACACTC
2. JH6-intron junction	GAGGTACGGTATGGACGTCTG	AGGACCAACCTGCAATGCTC
3. Intron (5` of Eμ)	GGAGCCACATTTGGACGAGA	ACACACAGCGCACCTCATAA
4. Intron (5` of Eμ)	ATGCGGGACTGCGTTTTGA	CATCATCTGCTTCCAGCTTCG
5. Intron (3` of Eμ)	GCGCCCGACATGGTAAGAGA	GACCCAGACAATGGTCACTCAA
IgM mRNA	ACAGACGGGAGGTACGGTAT	CGACGGGGAATTCTCACAGG

ChIP-qPCR primers (Ramos)	Forward	Reverse
1. Promoter	CACAGCCAGCATACACCTCC	CCTGTGGGTGCCTAAGTGAG
2. Variable region	AGGAATGCGGATATGAAGATATGA	CACTGAAGGACCCACCATAGAC
3. Intron (5` of Eμ)	GGAGCCACATTTGGACGAGA	ACACACAGCGCACCTCATAA
4. Intron (5` of Eμ)	ATGCGGGACTGCGTTTTGA	CATCATCTGCTTCCAGCTTCG
5. Eμ	GGTCACCGCGAGAGTCTATTT	TTCGGTCAGCCTCGCCTTAT
6. Intron (3` of Eμ)	GCGCCCGACATGGTAAGAGA	GACCCAGACAATGGTCACTCAA
7. Cμ	CGGTACTTCGCCCACAGCA	ACTTGTCCACGGTCCTCTCGGTGA
B2M	CTGTGCTCGCGCTACTCTC	AACTTGGAGAAGGGAAGTCACG
Neg. (negative control region, chr 1)	GGAAAGGCCCCAGAGTTCAA	CCTCCTGATGTGTAAGGCCC

ChIP-qPCR primers (B1-8^{hi})	Forward	Reverse
1. Jh4-intron junction	TGGACGAGGCCTTGAGTGG	AGACTGTGAGAGTGGTGC
2. Intron	CAGTCTCCTCAGGTGAGTCCT	CCCAATGACCCTTTCTGACT
3. Intron	ATGGTGTTGGTGGAGTCC	AATCTCCAACTACAGCCC
4. Eμ	TGGGGCACTTTCTTTAGATTTG	GACAGCAACTACCCTTTTGAGACC
5. Intron	CTGCAGCAGCTGGCAGG	GGCTGGACAGAGTGTTTCAAAACCAC
6. Intron	GTTGCCTGTTAACCAATAATCATAGAGCTCATGG	GTATAACTGAAGTAGAGACAGCATCAGTACCTCAAC

ChIP antibodies	Vendor	Catalog no.
H3K4me3	Abcam	ab8580
H3K27ac	Abcam	ab4729
H3K36me3	Abcam	ab9050
H3 pan-Ac	Active Motif	61937
H4 pan-Ac	Active Motif	39026
H3K79me3	Abcam	ab2621
H3	Abcam	ab1791
IgG	Invitrogen	10500C

FACS antibody	Vendor	Catalog no.	Dilution
IgM-APC (MHM-88)	BioLegend	314510	1/200
B220-FITC	BD Biosciences	553088	1/500
Fas-PE-Cy7	BD Biosciences	557653	1/1000
CD38-APC	ThermoFisher	17-0389-42	1/200

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.21.492925; this version posted May 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

MutPE-seq primers		
Round 1_FWD (B1-8hi)	CTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNTCCACAGGTGTCCACTCCCAG	N is A,G,C,T randomly entered during primer synthesis
Round 1_FWD (VH4-34, VH4-59, VH3-30)	CTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNTGTTCACAGGGGTCCTGTCC	N is A,G,C,T randomly entered during primer synthesis
Round 1_REV	CTGGAGTTCAGACGTGTGCTCTTCCGATCTCCTAGAGTGGCCATTCTTACC	
Round 2_FWD	AATGATACGGCGACCACCGAGATCTACACNNNNNNNNACACTCTTTCCCTACACGAC	N represents 8 positions of individual sample barcode (IdxE501-E5xx)
Round 2_REV	CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTGGAGTTCAGACGTGTG	N represents 6 positions of individual sample barcode (revIDX1-xx)