Somatic hypermutation patterns in immunoglobulin variable regions are established independently of the local transcriptional landscape

Ursula E. Schoeberl[1]*, Johanna Fitz[1]*, Kimon Froussios[1], Renan Valieris[2], Iordanis Ourailidis[1], Marina Makharova[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 01508-010, 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)

Abstract

Somatic hypermutation (SHM) of immunoglobulin variable (V) regions modulates antibody-antigen affinity is initiated by activation-induced cytidine deaminase (AID) on single-stranded DNA (ssDNA). Transcription is essential for SHM and AID target genes harbor activating chromatin marks and RNA polymerase II (Pol II) stalling, leading to the model that these features favor higher rates of mutagenesis. However, whether such relationships exist within V regions is undetermined. Here, we directly compared SHM and nascent transcription across four V regions and 275 non-immunoglobulin SHM targets at single-nucleotide resolution using precision run-on sequencing (PRO-seq). Although locales of Pol II enrichment and zones of Pol II stalling were detected within V regions, their correlation with SHM was not statistically significant. Moreover, SHM was robust against major reductions of activating epigenetic marks and transcription. This data suggests that SHM patterns and spectra are established independently of specific local nascent transcriptional features.

50 **Introduction**

51 Somatic hypermutation (SHM) is the molecular basis for the diversification of antibodies in response
52 to pathogens and vaccines and is hence indispensable for robust long-term immunity [1-3]. SHM occurs
53 in activated B cells within microanatomical structures called germinal centers in secondary lymphoid
54 tissue upon engagement with antigens and helper T cells [4]. Mutations are generated in the V regions
55 of the immunoglobulin (IG) heavy (*IGH*) and light chain (*IGK*, *IGL*) genes by activation-induced
56 cytidine deaminase (AID) [5,6] which acts co-transcriptionally on single-stranded DNA (ssDNA) [7-10]. AID
57 preferentially deaminates cytosine residues within WRCH motifs (where W = A or T, R = A or G and H
58 = A, C or T) [11]. AID also acts at the *IGH* switch regions triggering a reaction cascade leading to class
59 switch recombination (CSR) which yields the various antibody isotypes essential for effector functions
60 following antigen binding [12-15]. In addition, SHM occurs at several transcribed non-IG genes and has
61 been directly linked with B cell genome instability and cancer [16-21].

62 A longstanding question in the field is the relationship of SHM with the nascent transcriptional
63 and epigenetic landscapes. AID is recruited to the transcriptional machinery via specific enhancers [22-
64 25] and can subsequently associate with Pol II and the generic transcriptional machinery, including
65 elongation factors [26-30] and RNA processing proteins [31,32]. SHM and CSR have also been correlated
66 with the presence of activating histone modifications typical of transcribed genes [29,33-41]. Given that
67 AID is a very weak enzyme [42,43], it has been suggested that higher rates of mutation may be favored
68 under specific conditions that cause increased ssDNA exposure. For instance, AID can mutate both
69 DNA strands *in vitro* when the substrate is negatively supercoiled but is far less active on relaxed
70 substrates, leading to the idea that negative supercoils in the wake of transcribing Pol II can provide a
71 transient source of ssDNA [44,45]. SHM has also been correlated with features where the speed of Pol II
72 may be reduced, resulting in longer steady-state exposure of ssDNA. These features include
73 transcription initiation [46], transcription termination [47-50], convergent transcription [51], and most
74 prominently, promoter-proximal paused Pol II and slowly elongating Pol II in gene bodies, which we
75 refer to henceforth as Pol II stalling [17,27,52-56]. Moreover, it has been suggested that ssDNA patches
76 detected in V regions and non-IG SHM targets may correspond to ssDNA within the Pol II
77 holoenzyme or from negative supercoils upstream of stalled Pol II [54,57-59].

78 The idea that stalled Pol II may play a role in SHM was first proposed nearly three decades
79 ago [60]. Subsequently, Pol II was shown to accumulate in IG switch regions during CSR [52,53] and the
80 Pol II stalling factor, SPT5, was found to be required for efficient AID targeting to IG switch regions
81 and non-IG AID target genes [27]. Pol II accumulation in V regions was also found to be higher in SHM-
82 competent germinal center B cells but not in SHM-incompetent primary B cells [54]. Recently, studies
83 using model reporter substrates have found that SHM driven by mutation enhancers correlated with
84 an increase in the frequency of ssDNA patches along with increased SPT5 and Pol II stalling [55].
85 Indeed, high levels of Pol II and SPT5 were found to be sufficient to identify new AID target genes via
86 a machine learning approach [17]. Collectively, this leads to a model suggesting that the likelihood of
87 SHM is higher in the vicinity of zones of stalling and/or convergent transcription due to increased time
88 and access to ssDNA, relative to the neighboring DNA lacking any such feature [51,55,61-63].

89 Importantly, however, the nascent transcriptional landscape of V regions has not been
90 determined and hence a systematic comparison between transcriptional and SHM patterns is lacking.
91 There is evidence for Pol II stalling in V regions based on the accumulation of Pol II observed with
92 chromatin immunoprecipitation followed by quantitative PCR (ChIP-qPCR) [48,54]. However, due to the
93 very short length of V regions (~500 bp) coupled with the low resolution of ChIP-qPCR and ChIP-
94 sequencing (ChIP-seq), one cannot ascertain if Pol II is stalled across the entire V region or at
95 specific locations, and therefore, whether sites of recurrent stalling correlate with zones of increased
96 mutagenesis.

97 Here, we address this question by using precision run-on sequencing (PRO-seq) and the
98 related method, PRO-cap, in combination with a customized bioinformatics pipeline to map reads to V
99 regions. By identifying the precise 3' end of nascent RNA within engaged Pol II complexes, PRO-seq
100 provides a map of the location and orientation of actively transcribing Pol II genome-wide at single
101 nucleotide resolution [64,65]. PRO-cap identifies the 5' ends of capped nascent RNAs that are
102 associated with engaged Pol II molecules and hence provides maps of the location and orientation of
103 transcription initiation sites at single nucleotide resolution [64,66]. For these reasons, PRO-seq and PRO-
104 cap are amongst the most widely used assays for studying nascent transcriptional profiles genome-
105 wide and they have been particularly important in identifying recurrent sites of Pol II stalling and
106 initiation at promoters, enhancers and within gene bodies [66-72]. The single nucleotide resolution of the

assay means that PRO-seq can identify recurrent sites of Pol II stalling within the short V regions. With this approach, we could ask whether and to what extent the discrete SHM patterns in V regions and non-IG SHM targets can be plausibly explained by recurrent stalling, initiation or convergent/antisense transcription. The same approach was used to study the relationship between the levels of activating chromatin marks with SHM within V regions. We found weak and inconsistent correlations between mutation rate and any of these co-transcriptional features or epigenetic marks at all four human and mouse V regions analyzed as well as 275 previously-identified AID target loci in mice [17]. Thus, despite the global correlation between AID targeting and such features, when probed at the single nucleotide level, there does not appear to be a strong relationship between them, raising the possibility that ssDNA associated with these transcriptional features may not be the major source of ssDNA for AID.

Results

V gene locales harbor an atypical chromatin landscape

As a model system of SHM, we chose Ramos, an IgM-positive human Burkitt lymphoma-derived cell line where the single functional, pre-recombined *IGH* V region consists of the V gene, VH4-34, the diversity segment, DH3-10, and the joining segment, JH6. Ramos cells constitutively express AID and undergo low levels of SHM mostly at C:G residues [73,74]. However, SHM can be significantly boosted in these cells upon overexpression of AID variants, and widely used for the study of SHM [25,26,33,34,39,48,55,57,58,73-78].

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, given that *IGH* is highly transcribed, we observed that the V 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 (Supplementary Fig. S1A). As a result,

135 the first nucleosome from the transcription start site has the lowest levels of H3K27ac and H3K4me3
136 which is contrary to what is usually observed at highly active genes where the first nucleosome
137 harbors the highest levels of these marks. In contrast, much higher levels of all marks were observed
138 in the intronic regions flanking the Eμ enhancer with an expected decrease at Eμ itself where
139 nucleosomes are occluded by the presence of transcription factors (Fig. S1A) [79]. Importantly, these
140 atypical profiles were also observed at the murine *Igh* in primary B cells expressing the B1-8^{hi} V
141 region and activated with lipopolysaccharide (LPS), interleukin 4(IL4) and RP105 [80] (Fig. S1B). We
142 conclude that V regions are marked by relatively low levels of activating histone marks and that this
143 landscape is conserved between the human Ramos and mouse B1-8^{hi} V regions.

144 **The Eμ enhancer regulates the levels of chromatin marks in the flanking intronic DNA but not**
145 **within the V region**

146 The fact that the highest enrichments of chromatin marks were seen on either side of the Eμ
147 enhancer suggested that this element may be responsible for the deposition of these marks. This
148 raised the question of whether deletion of Eμ (Eμ$^{-/-}$) would alter the levels of these marks and affect
149 SHM. SHM results in the loss of WRCH motifs (hotspot saturation) as well as a gain of new hotspot
150 motifs [81], hence continuous SHM in AID-sufficient Ramos cells can confound subsequent analysis of
151 SHM profiles. To ensure that all cells have the same V gene sequence prior to SHM, we ablated
152 *AICDA*, encoding AID, in Ramos cells (AID$^{-/-}$) and subsequently deleted Eμ to obtain AID$^{-/-}$ Eμ$^{-/-}$ cell
153 lines (Fig. S1C). Specifically, we used CRISPR-mediated editing to delete a 583 bp region
154 corresponding to the peak of chromatin accessibility measured by assay for transposase-accessible
155 chromatin (ATAC-seq) at Eμ in AID$^{-/-}$ cells (AID$^{-/-}$ Eμ$^{-/-}$) (Fig. S1C). Based on PRO-cap analysis of
156 transcription start sites (see below), this deletion eliminates the Iμ promoter that encodes the sterile
157 germline *IGHM* transcript and antisense transcription start sites residing in Eμ (Fig. S1C). We chose
158 two independent clones, AID$^{-/-}$ Eμ$^{-/-}$ c1 and AID$^{-/-}$ Eμ$^{-/-}$ c2, for subsequent experiments (Fig. S1D).

159 To comprehensively visualize profiles of these histone marks across the V region and flanking
160 sequences, we performed ChIP followed by deep sequencing (ChIP-seq) for H3K27ac, H3K4me3 and
161 H3K36me3. We generated a custom chromosome spanning the Ramos V region and promoter.
162 Importantly, the Ramos V region is highly mappable and hence there is minimal loss of reads due to

163 multimapping with other V gene families (see later section). The results in AID$^{-/-}$ Ramos cells were
164 consistent with ChIP-qPCR in that the levels of all marks were lowest in the promoter-proximal region
165 and gradually increased into the V region, peaking in the intronic sequences flanking Eμ (Fig. 1A and
166 B). This analysis also revealed the absence of bimodal distributions of H3K27ac and H3K4me3 at the
167 VH4-34 promoter, which is commonly observed at highly active divergently transcribed genes. In
168 AID$^{-/-}$ Eμ$^{-/-}$ cells, all three marks were decreased (Fig. 1 A and B), a finding that we confirmed by
169 ChIP-qPCR, and which was not due to changes in nucleosome occupancy as determined by histone
170 H3 measurements (Fig. 1C). Using antibodies against pan-H3 and pan-H4 acetylation, we observed a
171 significant decrease in histone acetylation across the locus in AID$^{-/-}$ Eμ$^{-/-}$ cells (Fig. 1C). Moreover,
172 another mark of transcription elongation, H3K79me3, was also reduced in AID$^{-/-}$ Eμ$^{-/-}$ cells (Fig. 1C).
173 All marks in AID$^{-/-}$ Eμ$^{-/-}$ cells were significantly reduced up to 4-fold in the intron where they are
174 normally at their peak levels (Fig. 1 A to C). However, within the V region, where these marks are
175 normally at their lowest levels, there was no significant change in AID$^{-/-}$ Eμ$^{-/-}$ cells Fig. 1, A to C).
176 Consequently, AID$^{-/-}$ Eμ$^{-/-}$ cells harbor a chromatin landscape wherein the levels of all marks appear
177 to be comparable between the V region and the intron (Fig. 1C, compare the measurements from
178 amplicons 2, 3 and 5 in AID$^{-/-}$ and AID$^{-/-}$ Eμ$^{-/-}$ cells). We conclude that Eμ is responsible for the higher
179 levels of chromatin marks within the intron but does not influence the deposition of these marks within
180 the V region.

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

189

190 **SHM frequency does not correlate with the levels of nascent transcription or activating histone**
191 **marks in the V region locale**

192 AID$^{-/-}$ Eμ$^{-/-}$ cells provide an ideal system to ask whether changes in the levels of nascent transcription
193 and chromatin marks directly impact on SHM without any contribution from secondary, indirect effects.
194 Hence, we sequenced the V region as well as the JH6 intron immediately downstream (Fig. 2C, upper
195 panel). This analysis allowed us to compare SHM between a region where chromatin marks were
196 normally low and Eμ-independent (V region) and a region where the marks were normally high and
197 strongly Eμ-dependent (JH6 intron). To measure SHM, we infected AID$^{-/-}$ and AID$^{-/-}$ Eμ$^{-/-}$ cells with
198 AID (JP8Bdel), a C-terminal truncation mutant of AID that results in nuclear retention and major
199 increase in mutation rates after 6 days [82]. As a control, we also infected cells with a catalytically
200 inactive AID mutant (E58Q) [82]. As discussed in a later section, similar SHM spectra and profiles were
201 obtained upon infection with a hyperactive mutant of AID (m7.3), which retains the C terminal domain
202 [83]. However, since AID (m7.3) required much longer culturing post-infection (21 days), we used AID
203 (JP8Bdel) for our analyses.

204 In both the V region and JH6 intron, we observed a small (~20%) decrease in SHM frequency
205 in AID$^{-/-}$ Eμ$^{-/-}$ relative to AID$^{-/-}$ cells expressing AID (JP8Bdel) that was not statistically significant (Fig.
206 2C). This result is consistent with mouse studies where loss of Eμ caused only modest decreases in
207 SHM [84,85]. As an alternative readout of SHM activity, we assayed for the loss of surface IgM, which
208 occurs due to AID-induced stop codons or frameshifts [73]. The results showed that the magnitude of
209 IgM loss was similar between the control and Eμ$^{-/-}$ cells suggesting that the minor decreases in SHM
210 do not translate into equivalent changes in surface IgM (Fig. 2 D and E). We conclude that the Eμ
211 enhancer is dispensable for robust SHM in Ramos cells.

212 By comparing the levels of chromatin marks with SHM frequencies in the V region and JH6
213 intron, we can draw three major conclusions: First, we infer that the levels of SHM do not correlate
214 with those of activating histone marks. Indeed, SHM frequency is slightly lower in the intron ($1.6 \times 10^{-}$
215 ^{3}), which harbors the highest level of these marks, than in the V region (1.86×10^{-3}), which has the
216 lowest levels of these modifications (Fig. 2C). Second, as exemplified by the JH6 intron, SHM is
217 resistant to major decreases in these marks. Third, we attribute the small reduction in SHM in
218 AID$^{-/-}$ Eμ$^{-/-}$ cells to the decrease in nascent transcription although, here too, the decrease of SHM

219 (~20%; Fig. 2C) is considerably smaller than that of nascent transcription which is reduced up to 70%
220 in the JH6 intron of AID$^{-/-}$ E$\mu^{-/-}$ cells (Fig. 2A, amplicons 2-5).

221 We conclude that although epigenetic modifications mark loci undergoing SHM and may
222 contribute to SHM, their levels do not correlate with SHM frequencies and their significant reduction
223 does not have a major impact on mutation rates, at least in this system.

224 **Elucidating the nascent transcriptional landscape of the Ramos V region**

225 We next asked whether SHM patterns are dictated by local features of nascent transcription
226 landscape that could transiently lead to increased ssDNA exposure, such as during transcription
227 initiation, stalling and/or convergent/antisense transcription. Since V genes occur as families with
228 varying degrees of sequence identity, mapping of next-generation sequencing (NGS) reads to
229 recombined V regions using standard alignment workflows eliminates reads which map to more than
230 one location in the genome (multimappers) and only align reads mapping uniquely (typically allowing
231 2-3 mismatches). In Ramos cells, visual inspection of PRO-seq tracks on the UCSC genome browser
232 revealed robust coverage within the VH4-34 gene with uniquely-mapping reads indicating that this V
233 gene is highly mappable (not shown). To address this more systematically, we retrieved the normally
234 discarded, multimapping reads and allowed them to re-align to the recombined Ramos V region with
235 the important proviso that they only multimap to annotated V, D or J segments and not elsewhere in
236 the genome (Fig. 3A). Separate tracks were created for uniquely mapping reads and the recalled
237 multimapping reads, which were then combined to obtain the total profile (Fig. 3B).

238 We measured nascent transcriptional activity with PRO-seq and capped initiation sites with
239 PRO-cap[66]. For simplicity, we refer to relative accumulation of PRO-seq signal within the gene body
240 as Pol II stalling or stalled Pol II in order to differentiate it from promoter-proximal pausing which is
241 also reported by PRO-seq as the enrichment of signal typically 40-80 bp from transcription start sites.
242 Most notably, promoter-proximal pausing occurs via the binding of NELF, which inhibits elongation
243 whereas elongating Pol II lacks NELF and contains additional proteins such as SPT6 and the PAF
244 complex [86-89]. Accumulation of Pol II signal in gene bodies may arise due to the slowing down of Pol II
245 as a result of steric barriers such as secondary structures or premature termination signals. However,
246 initiation events within gene bodies could also give rise to a paused Pol II state akin to promoter-

247 proximal pausing and can be identified if PRO-seq accumulation is associated with upstream PRO-
248 cap enrichments.

249 Our mapping pipeline revealed that the Ramos V region is predominantly covered by
250 uniquely-mapping reads with very few multimapping reads (Fig. 3B). Of note, the same pipeline was
251 used for the ChIP-seq analysis of Ramos cells described in Fig. 1 above. Therefore, this approach
252 allowed us to delineate the transcriptional and epigenetic landscapes of IG V regions. In contrast to
253 the lack of a bimodal peak of activating histone marks (Fig. 1A), the Ramos VH4-34 promoter showed
254 clear bidirectional transcription and promoter-proximal pausing, the latter being noticeable as the
255 PRO-seq enrichments shortly downstream of the promoter-associated PRO-cap initiation signals (Fig.
256 3C). Next, we infected cells with AID (JP8Bdel) and analyzed SHM patterns in the V region using an
257 optimized version of mutation analysis with paired-end deep sequencing (MutPE-seq) of a PCR-
258 amplified fragment corresponding to the V region [90] (Fig. 3E). Mutations occurred robustly at C:G
259 residues within WRCH motifs with weak targeting of A:T residues, as expected from Ramos cells.
260 Moreover, the cold spot motif, SYC (where S = C or G and Y = C or T) was poorly mutated (Fig. 3 F-
261 H). A similar mutation profile and spectrum was observed when cells were infected with the
262 hyperactive AID mutant, AID (m7.3) [83], for 21 days, indicating that the absence of the C terminal
263 sequence in AID (JP8Bdel) does not alter its targeting to the V region or its motif preferences (Fig.
264 S2A-D).

265 Visual comparison of the profiles of SHM and nascent transcription at the Ramos V region
266 showed that some residues with relatively high mutation rates were associated with increased PRO-
267 seq densities in the neighborhood, such as in the framework region between CDR2 and CDR3 (Fig.
268 3D). However, other mutations did not occur in the vicinity of higher PRO-seq signals or internal
269 initiation events as measured by PRO-cap. For example, the region of highest PRO-seq read density
270 at the 3' end of the V region is not preferentially mutated over upstream C:G motifs where PRO-seq
271 signal is much weaker (Fig. 3 D and E). Indeed, the most mutated residue is located near the 5' end
272 of the V region where nascent transcription signals are lower than the 3' end (Fig. 3D).

273

274

275 **SHM frequency does not significantly correlate with Pol II stalling or transcriptional strength at**
276 **the Ramos V region**

277 To statistically evaluate the relationship between mutations and Pol II stalling, we mapped stalling
278 sites as per the definition used in previous studies, namely, as positions covered by at least five PRO-
279 seq reads and where the signal is >3 standard deviations greater than the mean of the flanking 200
280 bp region [91] (see Methods). Stalling sites were extended by 3 bp on either side to create 7 bp stalling
281 zones (Fig. S3A). This window size was chosen to reflect the approximate length (6-8 bp) of the
282 transcription bubble in elongating Pol II complexes [88,89]. Next, overlapping stalling zones were merged
283 and only those stalling zones lying within the MutPE-seq amplicon (displayed as black boxes in Fig.
284 3D) were considered for further analysis. All regions within the MutPE-seq amplicon that were not
285 defined as stalling zones were termed non-stalling zones. Finally, the normalized mutation rate
286 (mutations/width of zone) was calculated for each stalling and non-stalling zone (Fig. S3A). Mutation
287 frequencies between stalling and non-stalling zones were compared, and statistical significance
288 determined by the Wilcoxon rank-sum test (Fig. S3B). The results showed no statistically significant
289 difference in mutation frequency between stalling and non-stalling zones (P = 0.56, Fig. S3B).
290 Importantly, similar results were obtained when the analysis was performed with larger stalling zones
291 (up to 21 bp).

292 As an alternative to identifying stalling zones, we asked whether C:G residues with higher
293 mutation frequencies occurred in locales having higher PRO-seq densities, that is, whether mutation
294 frequency corelated with local transcriptional strength. We divided all C:G residues in the Ramos V
295 region into unmutated and mutated residues (Fig. S3C). Next, we measured PRO-seq read counts
296 (RPM) in a 7 bp window centered at each C:G pair, that is, C:G ± 3 bp, and selected only non-
297 overlapping windows for analysis (Fig. S3C and Methods). We further divided the mutated group into
298 ten equal sub-groups (deciles) such that sub-group 1 contained the least mutated residues and sub-
299 group 10 contained the most mutated residues. The results, displayed as box plots, showed that there
300 was no discernible increase in PRO-seq density with increasing mutation frequency (Fig. S3D, first
301 row). A Wilcoxon rank sum test following multiple testing correction revealed a lack of statistical
302 significance (defined as $P < 0.05$) between the unmutated and any mutated group, as well as between
303 any pair of mutated groups.

304 Conversely, we asked whether mutated C:G residues within locales of higher transcription
305 were likely to have higher mutation rates than mutated C:G pairs in locales with lower transcription.
306 To do so, we divided all the 7 bp windows from the mutated group in Fig. S3C into deciles such that
307 sub-group 1 contained the least transcribed windows and sub-group 10 contained the most
308 transcribed windows (Fig. S3E). We then generated boxplots for the mutation frequency of the central
309 C:G residue within the 21 bp windows. The results showed that there was no discernible increase in
310 mutation frequency with increasing PRO-seq density, and no statistical significance ($P < 0.05$ using
311 the Wilcoxon rank sum test) was observed between any pair of sub-groups (Fig. S3F). Collectively,
312 these analyses suggest that mutation frequency is unrelated to local Pol II stalling or local
313 transcription strength.

314 PRO-seq in AID$^{-/-}$ Eμ$^{-/-}$ cells revealed that nascent sense transcription was uniformly
315 decreased across the V region and intron (Fig. S4A). Of note, PRO-seq showed that the Eμ enhancer
316 harbors the start sites of an antisense transcript, substantially (~6-7-fold) weaker than the sense
317 transcript (note the Y axis scale in Fig. S4A and quantification in Fig. S4B). This antisense transcript
318 initiates at the 5' boundary of Eμ and terminates within the V region, resulting in convergent
319 transcription within the V region (Fig. 3 C and D). Importantly, the decrease in antisense transcription
320 in AID$^{-/-}$ Eμ$^{-/-}$ cells (6-10-fold) is considerably stronger than sense transcription (~2-fold) suggesting
321 that convergent transcription is strongly reduced in the V region and intron (Fig. S4B). Given the weak
322 SHM phenotype in AID$^{-/-}$ Eμ$^{-/-}$ cells (Fig. 2C), this suggests that convergent transcription may not be
323 a major player in V region SHM, in agreement with a recent study wherein SHM on model reporter
324 substrates occurred without substantial antisense transcription [55].

325 We conclude that although some mutations may result from the presence of recurrent stalled
326 Pol II species in their locales, in general, the SHM patterns bear no significant relationship with zones
327 of Pol II stalling, Pol II accumulation or convergent transcription. In addition, SHM is resistant to major
328 decreases in Pol II levels within the V region and intron in both sense and antisense orientations.

329 **Absence of correlation between Pol II stalling and SHM at two additional human V regions**

330 To determine the transcriptomic profiles of other human V regions, we developed a strategy to
331 replace the endogenous *IGH* V region of Ramos cells with any V region of choice using CRISPR

332 editing (Fig. S5 A and B, and Methods). We chose two V regions corresponding to the unmutated
333 precursors of broadly neutralizing antibodies identified in individuals infected with human
334 immunodeficiency virus 1 (HIV1). One V region consisted of VH4-59*01, DH2-02 and JH6*03 [92] and
335 the other of VH3-30*18, DH2-02 and JH6*02 [93], which we henceforth refer to as VH4-59 and VH3-30,
336 respectively. In both cases, the sequences upstream and downstream of the new V regions, including
337 the VH4-34 promoter, leader peptide and intron are identical to the endogenous Ramos *IGH*
338 sequence. Sub-regions of both VH4-59 and VH3-30 harbored multimapping reads with the most
339 prominent being the framework region 3 preceding CDR3 in VH4-59 which was entirely covered with
340 multimapping reads (Fig. S5 C and D).

341 We analyzed the transcriptional and mutational landscape of these two V regions using PRO-
342 seq, PRO-cap and MutPE-seq (Fig. 4A-F). At VH4-59, we observed an increase in PRO-seq
343 enrichment towards the 3' end of the V region (Fig. 4 A and B), reminiscent of Pol II stalling and akin
344 to that seen at the endogenous Ramos V region (Fig. 3 B and C). Such enrichments were absent
345 within the VH3-30 V region (Fig. 4 D and E). A reasonable explanation is that VH4-59 and VH4-34
346 belong to the same V gene family (VH4) and hence share higher sequence identity than VH3-30
347 which is phylogenetically more distant from the VH4 family [94]. It is also apparent from these analyses
348 that there is no conserved V region transcriptional signature. It is plausible that sequence-intrinsic
349 features may be responsible for the distinct nascent transcription profiles of the different V regions.

350 At VH4-59, the location of some highly mutated residues correlated with increased Pol II
351 stalling in their locales, such as in the framework region between CDR2 and CDR3 (Fig. 4B).
352 However, many other similarly mutated residues within and between CDR1 and CDR2 did not show
353 elevated Pol II stalling in their locales (Fig. 4B). Similarly, at VH3-30, the most highly mutated
354 residues located in CDR1 and CDR3 were not associated with Pol II stalling (Fig. 4E). To analyze this
355 systematically, we defined stalling zones in VH4-59 and VH3-30 and compared mutation frequencies
356 between these and non-stalled zones, as described above. For both V regions, there was no
357 significant difference in mutation load between stalled and non-stalled zones (Fig. S3A-B; $P = 0.6$ for
358 VH4-59 and $P = 0.26$ for VH3-30). We also compared mutation frequencies and PRO-seq densities in
359 the neighborhood of C:G pairs at VH4-59 and VH3-30, which revealed that locales with higher rates of
360 mutations did not show a statistically significant correlation (based on the Wilcoxon rank sum test)

361 with transcriptional strength (Fig. S3C-D). Conversely, locales with higher transcription were not more
362 likely to harbor highly mutated C:G residues (Fig. S3E-F).

363 Of note, when comparing the mutation patterns between all three V regions analyzed here,
364 the strong bias for C:G mutations typical of Ramos cells is seen in all cases (Fig. 3E and Fig. 4 D and
365 F). Moreover, there appears to be some bias of mutation load in the CDRs although several major
366 mutation hotspots are also located within the intervening framework regions (Fig. 3E and Fig. 4 D and
367 F). The coldspot SYC motifs are weakly targeted or untargeted by AID, as expected (Fig. 3 E and Fig.
368 4, D and F). Finally, the differential mutability of WRCH motifs by AID, ranging from highly mutated to
369 unmutated, is seen in all the analyzed V regions (Fig. 3E and Fig. 4 D and F) and is in line with
370 observations in previous studies [81,95].

371 In sum, the results from VH4-34, VH4-59 and VH3-30 analyses lead us to conclude, firstly,
372 that recurrent Pol II stalling is not present across the entire length of V regions but is restricted to
373 specific sub-regions. Secondly, antisense and convergent transcription is much weaker than sense
374 transcription within V regions. Finally, the data suggest that although Pol II stalling may contribute
375 towards mutagenesis in some locales, in general, the discrete SHM profiles of V regions cannot be
376 explained by recurrent patterns of stalling or Pol II accumulation generally.

377 **SHM in germinal center B cells at a murine V region shows no direct relationship with nascent**
378 **transcriptional features**

379 In mice, SHM occurs robustly *in vivo* in germinal center B cells (GCBs) but is inefficient *in vitro* in
380 activated primary B cells [95]. Thus, we asked if such differences could be related to differences in the
381 underlying transcriptional landscape. To address this, we made use of the B1-8^{hi} *Igh* knock-in mouse
382 where the murine B1-8^{hi} V region is knocked-in at the endogenous *Igh* locus [96]. In this system, one
383 can compare SHM and transcriptional features of the B1-8^{hi} V region from both germinal center B
384 cells (GCBs) as well as primary *in vitro* activated B cells. To boost SHM in primary cells, we generated
385 homozygous *Igh*$^{B1\text{-}hi/B1\text{-}8hi}$ *Rosa26*$^{AIDER/AIDER}$ mice where AID fused to the estrogen receptor (AIDER) is
386 expressed constitutively from the *Rosa26* promoter such that upon addition of 4-hydroxytamoxifen (4-
387 HT), AIDER is translocated into the nucleus [90].

388 PRO-seq analysis from homozygous B1-8^{hi} mice showed that the VH1-72 gene used in the
389 B1-8^{hi} recombined V region shares strong homology with other V genes in the murine *Igh* locus
390 resulting in a very high degree of multimapping (Fig. S5E). Importantly, the CDR3 region was covered
391 with uniquely mapping reads since this region is formed via the junction of the V, D and J segments
392 during VDJ recombination resulting in a unique sequence in the genome (Fig. S5E). PRO-seq and
393 PRO-cap showed that there were no striking differences at the B1-8^{hi} V region between GCBs and
394 primary B cells activated with LPS, IL-4 and RP105 (Fig. 5A). Moreover, we observed local zones of
395 Pol II accumulation, internal initiation events and antisense/convergent transcription, but these did not
396 necessarily correlate with zones of mutation (Fig. 5A).

397 We note that the PRO-seq read densities at the B1-8^{hi} V region from GCBs was substantially
398 lower than those seen at the V regions analyzed in Ramos cells. We attribute this to the fact that
399 GCBs are more fragile than primary cells or cell lines resulting in considerable loss of viability and
400 quality during cell sorting. Consequently, the efficiency of the run-on reaction from isolated GCB
401 nuclei, the first step of PRO-seq, is much less efficient resulting in lower numbers of aligned reads in
402 the final libraries. These weak signals precluded a robust identification of stalling sites and hence we
403 were unable to perform the comparative analysis of mutation frequency and stalling zones as we did
404 for the other V regions described earlier.

405 As noted previously [95], SHM profiles of B1-8^{hi} GCBs are not identical to primary B cells
406 because GCBs exhibit 5-10 fold higher frequencies of SHM and substantial mutagenesis of A and T
407 residues arising from error-prone DNA repair of AID-induced mismatches at neighboring C:G residues
408 (Fig. 5 B and C). The palindromic AGCT hotspot in CDR3 is selectively mutated with high frequency
409 in primary cells, in line with a former report showing that this hotspot is the earliest targeted motif both
410 *in vitro* and *in vivo* [95]. Importantly, however, closer inspection showed that even in primary B cells, the
411 same WRCH motif appears to be mutated as in GCBs suggesting that AID targeting specificity is not
412 substantially different between primary B cells and GCBs (Fig. 5 B and C), in line with their
413 comparable transcriptional landscapes (Fig. 5A). AID activity is linked to cell division [97,98] and in our
414 experiments, GCBs were harvested after 10 days of immunization whereas primary cells were
415 harvested after 4 days of 4-HT treatment. Hence, it is plausible that GCBs may have acquired higher

416 rates of mutation, in part, due to more rounds of cell division within germinal centers compared to
417 primary cultures.

418 We conclude that the differences in mutation frequency between primary B cells and GCBs
419 cannot be explained by underlying nascent transcriptional features. Consequently, as in the case of
420 human V regions described above, the nascent transcriptional landscape of the B1-8^{hi} V region is not
421 predictive of the patterns or frequencies of SHM.

422 **Analysis of non-IG AID target loci reveals no correlation between SHM and transcription**
423 **strength**

424 To extend our analysis to non-IG AID targets, we made use of a previously available SHM dataset
425 from murine GCBs where mutation frequencies at 275 AID target genes were identified by deep
426 sequencing the first 500 bp from the annotated transcription start site (TSS) [17]. In this study, mutation
427 analysis was performed in GCBs from mice deficient in base excision and mismatch repair pathways
428 (*Ung*$^{-/-}$*Msh2*$^{-/-}$). In these mice, the processing of AID-induced U:G mismatches is abolished and,
429 consequently, DNA replication over these lesions leads to C to T and G to A transition mutations that
430 represent the direct targets of AID [99]. By comparing these data with PRO-seq and PRO-cap obtained
431 from GCBs, we could compare mutation profiles with those of nascent transcription. Importantly, at
432 many genes, initiation sites or zones defined by PRO-cap are often located at considerable distances
433 from the annotated TSS (defined by the RefSeq database and labeled as TSS in Fig. 5D, Fig. S6 and
434 Fig. S7). Consequently, the 500 bp segment sequenced for mutational analysis [17] often begins further
435 upstream or downstream of the initiation sites newly defined by PRO-cap.

436 The large number of genes in this analysis led to new and unexpected observations regarding
437 the nature of AID targeting. At virtually all genes, the differential mutation of WRCH motifs was
438 evident in that many such motifs were mutated to varying degrees with many others being unmutated.
439 Moreover, several non-WRCH cytidines were as efficiently mutated as WRCH cytidines (Fig. 5D and
440 several additional examples in Fig. S6 and Fig. S7). In V and switch regions, mutations initiate ~100-
441 150 bp after the promoter suggesting that AID may be excluded from initiating or early elongating Pol
442 II [2,3,62,100,101]. However, we identified a set of genes where the highest mutation frequencies were
443 observed within 100-150 bp of the initiation site defined by PRO-cap (*Il4ra* in Fig. 5D and additional
444 examples in Fig. S6A). Indeed, in some of these genes, the most mutated residues were in very close

445 proximity to the initiation site, such as *Mcm7* (Fig. 5G). In some other genes, strongly mutated
446 residues coincided with the strongest peak of initiation (Fig. S6B). These observations indicate that
447 AID can act at the early stages of the transcription cycle at least at some genes, suggesting that AID
448 is not *per se* excluded from initiating or early elongating Pol II.

449 As we observed in V regions, mutation frequencies at individual nucleotides and mutation
450 zones did not necessarily correlate with enrichments of PRO-seq or PRO-cap (Fig. 5D, Fig. S6 and
451 Fig. S7). For example, at *Bcl6*, the major zone of mutation lies in a region of low PRO-seq signal
452 where many mutations are not in WRCH motifs (Fig. 5D). At *Myc*, mutations are observed upstream
453 of the initiation site defined by PRO-cap indicating that mutagenesis can also occur in the antisense
454 orientation (Fig. 5E). Thus, as at V genes, SHM does not correlate with local nascent transcriptional
455 features.

456 To systematically determine if mutation frequency correlated with local transcriptional
457 strength, we divided all C:G residues from Alvarez-Prado et al.[17] into three groups. Group 1 contained
458 C:G residues in unmutated genes. Group 2 contained unmutated C:G residues and group 3 contained
459 the mutated C:G residues within the mutated genes (Fig. S8A). The C:G residues in group 3 were
460 ranked from lowest to highest mutation frequency and further divided into ten equal groups (deciles)
461 where residues in sub-group 1 were the least mutated and those in sub-group 10 were the most
462 mutated. Next, we counted PRO-seq reads in a 7 bp window centered at each C:G residue (C:G ± 3
463 bp) (Fig. S8A), plotted the results as box plots and performed statistical analyses for all pairwise
464 combinations (Fig. S8B).

465 PRO-seq densities around C:G residues in the unmutated genes (group 1) were not
466 significantly different from those at unmutated C:G residues within mutated genes (group 2) (Fig.
467 S8B). However, group 1 residues harbored significantly lower PRO-seq signals in their neighborhood
468 than most, but not all, sub-groups of mutated residues (Fig. S8B). This is in line with observations that
469 mutation rates tend to be higher in transcribed genes [17,24]. However, the difference in mean PRO-seq
470 values between group 1 and other groups ranged between 1.2-1.9-fold. Based on our results from
471 $E\mu^{-/-}$ cells where >2-fold differences did not have a major impact on SHM (Fig. 2), we infer that the
472 reason for the absence of mutation in group 1 is not solely due to lower transcription.

473 There were no major differences in transcriptional strength between the unmutated residues
474 in mutated genes (group 2) and mutated residues in these genes (group 3 subgroups) (Fig. S8B). In
475 fact, the PRO-seq densities in group 3, sub-group 10, which harbors the most mutated C:G residues,
476 were lower than in many other mutated sub-groups and even showed statistically significant
477 differences in some cases (Fig. S8B). The lack of a consistent correlation between mutation
478 frequency and local transcriptional strength suggests that, within AID target genes, highly mutated
479 C:G residues are not generally located in sequence neighborhoods harboring higher levels of
480 transcription and Pol II occupancy than less mutated or unmutated residues.

481 Conversely, we asked whether C:G residues within locales of higher transcription were likely
482 to have higher mutation rates. As was done previously for V regions (Fig. S3E-F), we divided all the 7
483 bp windows from group 3 above into deciles such that sub-group 1 contained the least transcribed
484 and sub-group 10 contained the most transcribed windows, respectively (Fig. S8C). We then
485 generated boxplots for the mutation frequency of the central C:G residue within the 7 bp windows
486 (Fig. S8 C and D). The results showed no trend of higher mutation frequency in groups with higher
487 PRO-seq values with only one pair (group 3 versus group 10) showing a statistically significant
488 difference (Fig. S8D). We conclude that the level of local transcription within mutated genes is not
489 predictive of C:G mutability.

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

491 To determine whether the gene regulatory context affects the patterns of SHM and transcription, we
492 replaced the endogenous V region in Ramos cells with the entire murine B1-8^{hi} V region (Ramos$^{B1\text{-}8hi}$)
493 using the strategy described earlier (Fig. 6A and Fig. S4, A and B). In this scenario, transcription and
494 SHM of the murine B1-8^{hi} V region would be under the control of the Ramos VH4-34 promoter and the
495 human Eμ and 3' *IGH* super-enhancers. This allowed us to ask whether the targeting of SHM to the
496 B1-8^{hi} V region was influenced by differences in the human versus the mouse enhancers and
497 promoters. Importantly, when placed in the context of the human *IGH* locus, the B1-8^{hi} sequence was
498 covered almost exclusively by uniquely mapping reads (Fig. 6B), which is in sharp contrast to the
499 extensive multimapping observed within B1-8^{hi} at the murine *Igh* locus (Fig. S4E). This is because the
500 VH1-72 gene used in the B1-8^{hi} sequence is much less homologous with the human V genes than the
501 mouse V genes.

502 PRO-seq and PRO-cap revealed that the nascent transcriptional profile of the B1-8^{hi}
503 sequence in Ramos^{B1-8hi} cells (Fig. 6C) shared similarities with the B1-8^{hi} sequence in the murine *Igh*
504 context (Fig. 5A). For example, similarities were noted in CDR3 and flanking sequences, for instance,
505 the presence of the antisense initiation site (PRO-cap track) and shared spikes of nascent
506 transcriptional activity (PRO-seq) (Fig. 6C). More importantly, MutPE-seq following AID (JP8Bdel)
507 infection revealed a very similar SHM pattern in Ramos^{B1-8hi} cells compared to the murine B1-8^{hi}
508 context, especially with primary murine B cells (Fig. 6D). Of note, the dominant, palindromic AGCT
509 hotspot in CDR3 was the most mutated residue exactly as in B1-8^{hi} murine primary cells and GCBs
510 (Fig. 6D). The other mutated C:G residues in the Ramos^{B1-8hi} sequence were largely the same as
511 those mutated in the murine context albeit the relative mutation rates within each V region varied (Fig.
512 6D, compare with Fig. 5 B and C). Stalled Pol II zones within the Ramos^{B1-8hi} V region bore no
513 significant relationship with mutation frequency compared to non-stalled zones (Fig. S3A-B; P = 0.83).
514 Analysis of PRO-seq density in the neighborhood of mutated and unmutated C:G pairs revealed a
515 lack of statistical significance between mutation frequency and local transcriptional strength (Fig.
516 S3C-F). Thus, as in the case of the other V regions analyzed here, there is no significant relationship
517 between Pol II stalling, Pol II occupancy and local SHM at the Ramos^{B1-8hi} V region.

518 We note that mutation frequencies in murine GCBs and AID (JP8Bdel)-expressing Ramos
519 cells were comparable (note the Y axes scales in Fig. 6D) with the major difference being the weak
520 A:T mutagenesis typical of Ramos cells. Of note, the sense PRO-cap peak in CDR3 in the murine B1-
521 8^{hi} (Fig. 5A), located just 3 bp upstream of the strong AGCT hotspot, is absent in Ramos^{B1-8hi} (Fig. 6C)
522 implying that this initiation site is not essential for the high mutation frequency of this hotspot motif. In
523 sum, the nascent transcriptional and SHM patterns of the B1-8^{hi} V region in mouse B cells are largely
524 retained in Ramos^{B1-8hi} cells.

525 Collectively, our results lead us to conclude that, following AID recruitment via IG enhancers
526 and its association with Pol II-associated factors, the discrete and reproducible mutation patterns in V
527 regions and non-Ig SHM target genes are apparently “hard-wired” into the DNA sequence rather than
528 the result of the local Pol II stalling, convergent transcription or activating chromatin marks.

529

530

531 **Discussion**

532 In this study, we explored the relationship between transcription and V region SHM by performing a
533 detailed comparative analysis of SHM with nascent transcription and activating chromatin marks.

534 Our analysis of the epigenetic landscape showed that major decreases in activating
535 chromatin marks and nascent transcription, via deletion of the Eμ enhancer, do not have a significant
536 impact on SHM. Although this does not rule out a role for these marks in creating an optimal
537 chromatin environment for SHM, it does suggest that the absolute levels of these marks do not
538 correlate with SHM. Intriguingly, compared to other highly active genes, including non-Ig AID targets
539 [24,102], the V region chromatin locale is unusual for two reasons. First, it does not harbor the typical
540 bimodal distribution of H3K27ac and H3K4me3 at the TSS despite robust bidirectional transcription
541 and normal nucleosome occupancy as measured by the levels of histone H3. Second, the V gene
542 body is marked by relatively low levels of these marks. However, it remains unclear whether this
543 locale contributes to making the V genes among the most favored targets of AID in B cells.

544 The key conclusion of our study is that local patterns and frequencies of SHM, either in V
545 regions or non-Ig AID targets, cannot be predicted from, nor do they appear to be derived from, sites
546 of Pol II stalling, internal initiation or convergent transcription. Our approach clearly detected all these
547 features within all four V regions analyzed here, in line with the global gene-level correlations
548 previously noted [17,27,51]. However, by accurately mapping the precise locations of these features within
549 V regions and non-IG SHM target loci, our study revealed that at the local sequence level, the
550 presence of these features did not generally correlate with increased local SHM. Indeed, several
551 zones of elevated SHM in different V regions and non-Ig AID target loci appeared to be entirely
552 devoid of any discernible transcriptional features. This implies that these features may not be major
553 players in SHM and, therefore, that the ssDNA associated with them may not be the major substrate
554 for AID. Moreover, the nascent transcription landscapes of the four V genes that we report in this
555 study are quite dissimilar, which argues against any conserved V region nascent transcriptional
556 signature. Finally, the murine B1-8^{hi} V region retains its characteristic SHM profile *in vitro* and *in vivo*,
557 and in both murine and human *cis*-regulatory contexts. Altogether, these results suggest that the
558 discrete SHM patterns of individual V regions may be influenced by local sequence context, rather
559 than specific local transcriptional features.

560 **What is the role of transcription in SHM?**

561 The consistent lack of correlation between transcriptional features and SHM reported here begs the
562 question of what role transcription plays in SHM. SHM depends on the availability of ssDNA, but the
563 source(s) of accessible ssDNA for AID remains a major unresolved question. It has been proposed
564 that ssDNA within transcribing Pol II complexes could be a substrate for AID, especially when Pol II is
565 paused or stalled, an event that would lead to increased ssDNA exposure [17,27,52-56]. However, we
566 found no significant correlation between SHM frequency and Pol II stalling zones or Pol II
567 accumulation, which prompted us to take a closer look at the structures of paused and elongating Pol
568 II complexes. This analysis showed that of the ~11 nucleotide non-template ssDNA in the structure,
569 about half lies buried within Pol II [88,89] (Fig. 7A). Importantly, the exposed portion (~5 nt) is entirely
570 covered by SPT5 [89] (Fig. 7A). The PAF complex and SPT6 are located further away from this site [88]
571 (Fig. 7A). Under such structural constraints, it is difficult to conceive how AID could access the ~5 nt
572 exposed ssDNA and catalyze deamination even if Pol II were paused or stalled. Thus, it appears
573 unlikely that the transcription bubble within an actively elongating Pol II holoenzyme, including a
574 stalled complex, serves as an optimal substrate for AID.

575 Nevertheless, various studies have provided important insights into potential alternative
576 sources of ssDNA. For instance, supercoiled DNA, but not relaxed DNA, can serve as efficient
577 substrates for AID *in vitro* in biochemical assays and *in vivo* in bacterial SHM assays in the complete
578 absence of transcription [44,45,57]. In fact, untranscribed supercoiled plasmids can be mutated *in vitro* at
579 rates comparable to those seen in V regions *in vivo* [45]. The proposition is that negatively supercoiled
580 DNA, which occurs in the wake of transcribing Pol II and is inherently prone to transient melting, can
581 generate ssDNA patches on either strand as substrates for AID [44,45,57]. Indeed, transcription-
582 dependent ssDNA patches have been detected in SHM model substrates, V regions and other AID
583 target loci [54,55,57,58]. Importantly, however, these studies revealed that the location and frequency of
584 ssDNA patches did not correlate with the patterns and spectra of SHM, implying that the incidence of
585 ssDNA as detected by these assays is not a predictor of SHM patterns [55,57,59].

586 Another key insight is the fact that significant mutational asymmetry is observed upon
587 depletion of RNA Exosome subunits or the DNA:RNA helicases, Mtr4, Senataxin and others [103-105],
588 which implies that one of the major species targeted by AID is an RNA:DNA hybrid. Indeed, the V

589 region sense transcript is a substrate of the RNA exosome and the latter is required for optimal
590 mutation of the template strand in the B1-8^{hi} V region, findings that directly implicate the processing of
591 DNA:RNA hybrids, or short transient R loops, in directing the patterns of SHM [103].

592 Nevertheless, because all the above-mentioned features are continuously associated with
593 transcription, they cannot, by themselves, explain the distinct and highly reproducible SHM profiles of
594 V regions, nor can they explain why SHM spectra are restricted to a ~2 kb window starting from the V
595 gene promoter [3,62,63]. To reconcile these observations, we support a model centered on premature
596 transcription termination, which, as proposed below, can define the boundaries of SHM and also
597 provide a source of ssDNA, the latter having been suggested previously in the context of RNA
598 exosome-mediated AID targeting [49,50] as well as SHM on plasmid substrates and in the Ramos V
599 region [47,48] (Fig. 7B).

600 It is now appreciated that pervasive termination occurs near the 5' ends of all active genes
601 and enhancers resulting from the activity of the Integrator endonuclease complex [106-108]. Specifically,
602 the depletion of Integrator subunits manifests as an increase in Pol II and nascent transcript levels
603 within the first 3-4 kb from TSSs [106-108], which, intriguingly, encompasses the known spectrum of SHM
604 at V regions, that is, ~2 kb from the V gene promoters [3,62,63]. A similar process would be expected to
605 occur at the 5' ends of non-IG AID target genes or at divergently transcribed enhancers that are
606 targets of AID [16,55,109,110]. Importantly, we note that accumulation of Pol II in V regions that we (this
607 study) or others [48,54] have detected cannot be inferred as termination zones because the ChIP and
608 nascent transcript profiling assays used therein do not provide information on the fate of the detected
609 Pol II species. Studies to date reveal no specific locations for premature termination, rather
610 suggesting that such events are stochastic and an inherent feature of early elongation [106-108].

611 Another attractive aspect of premature termination in the context of SHM is that it can
612 potentially provide multiple avenues for AID to access ssDNA [47-50]. Firstly, the speed of a terminating
613 Pol II is expected to be lower than a rapidly elongating one, which would increase the steady-state
614 availability of upstream, negatively supercoiled DNA for AID to access. Secondly, although the
615 structural details of how Pol II terminates and dissociates off DNA have not been resolved, it is
616 conceptually plausible that the collapse of the elongation complex could release ssDNA on the non-
617 template strand while the nascent RNA is still hybridized to the template strand, which provides

618 another source of ssDNA (Fig. 7B, step 4). Finally, the RNA would have to be unwound by DNA:RNA
619 helicases [105] and/or digested with RNaseH followed by degradation from the 5' end by Xrn2 and from
620 the 3' end by the RNA exosome, thereby providing AID access to the template strand [31,103-105,111] (Fig.
621 7B, step 4). Given that IG loci are amongst the most highly expressed in activated B cells, the
622 combined effect of having multiple rounds of transcription per cell cycle coupled with stochastic
623 termination within the first 2-3 kb from V gene promoters can plausibly explain the empirically
624 observed boundary of SHM, that is, ~2 kb from V gene promoters [3,62,63]. The model can also explain
625 why, at some non-IG SHM target loci, we observed mutations overlapping with, or in close proximity
626 to, the transcription initiation zone. Most importantly, such stochastic events will not be detected by
627 standard population-based approaches, which can explain the lack of correlation between local SHM
628 frequency and any transcriptional feature, be it Pol II stalling, internal initiation, convergent
629 transcription or ssDNA patches.

630 Although SHM-specific enhancers could explain why AID does not target all transcribed
631 genes [25,55] and the termination model could explain the spectrum of SHM, they do not explain why
632 WRCH motifs are differentially mutated. Indeed, the same motif can show high variation in mutation
633 frequency depending on its location within a given V region [81,95,112,113], a feature that is also evident in
634 the V regions and non-Ig genes analyzed in this study. This implies that, beyond AID recruitment and
635 ssDNA availability, the DNA sequence context may determine whether ssDNA containing a WRCH
636 motif is mutated and to what extent. This notion is supported by a biochemical study of SHM on
637 reporter cassettes [114], *in vivo* analyses of SHM in V regions from mice [81,95], bioinformatics and
638 computational modeling studies of V regions [112,113,115] and, most recently, by altering the nucleotide
639 context of AGCT hotspot motifs *in vivo* and *in vitro*[116]. In this regard, our study is important in showing
640 that the molecular basis of this sequence context dependency of SHM does not lie at the level of Pol
641 II stalling or any other nascent transcriptional feature, which was one of the initial hypotheses that
642 motivated the present study. In sum, further work is necessary to determine whether sequence
643 context regulates SHM patterns *in vivo* and what the underlying mechanism(s) may be.

644

645

646

647 **Methods**

648 **Cell culturing**

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

655 **Mice**

656 The mice were maintained in a C57BL/6 background and housed in the IMBA-IMP animal facility in
657 standard IVC cages with HEPA filtering. B1-8^{hi} Rosa26 ^{AIDER} mice were generated by crossing B1-8^{hi}
658 mice [80] with *Rosa26^{AIDER}* mice [90] and maintained as a homozygous line for both alleles. All animal
659 experiments were carried out with valid breeding and experimental licenses (GZ: MA58-320337-2019-
660 9, GZ: 925665/2013/20 and GZ: 618046/2018/14) obtained from the Austrian Veterinary Authorities
661 and in compliance with IMP-IMBA animal house regulations.

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

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

672

673

674 **Generation of Eμ$^{-/-}$ Ramos cells**

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

686 **Generation of Ramos cells expressing new, exogenous V regions**

687 AID$^{-/-}$ Ramos cells (clone D3) were electroporated with Cas9-sgRNA ribonucleoprotein complexes
688 (prepared in-house by the Vienna Biocenter Core Facilities - ProTech) and homology repair templates
689 containing a pair of unique sgRNA-target sites to excise the entire V region including the promoter
690 (Fig. S2A). Single, IgM-negative clones were isolated via flow cytometry using a BD Aria III sorter,
691 expanded and genotyped. One validated IgM-negative clone was selected as the parental clone. This
692 clone was then electroporated with Cas9-sgRNA ribonucleoprotein complexes targeting the unique
693 sgRNA sites and homology repair templates containing new V regions under the control of the
694 endogenous Ramos VH4-34 promoter. Single, IgM-positive cells were isolated, expanded and
695 genotyped.

696 **Lentiviral infections**

697 Lentiviral pRRL vectors expressing AID(JP8Bdel) or AID (m7.3) coupled with mCherry were
698 transfected along with ecotropic envelope (Eco-env)-expressing helper plasmid into LentiX cells via
699 the standard calcium phosphate methodology. Lentiviral supernatants were used to infect Ramos
700 cells via spinfection (2350 rpm for 90 min) in the presence of 8 μg/ml Polybrene (Sigma). Flow
701 cytometry (BD Aria III sorter) was used to sort mCherry-positive cells for mutation analysis.

702 **Isolation and activation of murine primary, splenic B cells**

703 Mature, naïve B cells were isolated from spleens of 2-4-month-old B1-8^{hi} Rosa26 ^{AIDER} mice, as per
704 established protocols [27]. B cells were cultured in complete RPMI medium supplemented with 10%
705 fetal bovine serum (FBS) and antibiotics, Interleukin 4 (IL4; made in-house by the Molecular Biology
706 Service, IMP), 25 μg/ml Lipopolysaccharide (Sigma) and RP105 (made in-house by the Molecular
707 Biology Service, IMP) and harvested after 3 or 4 days. For SHM assays from activated B1-8^{hi} *Rosa26*
708 ^{AIDER} primary B cells, 2μM 4-hydroxy tamoxifen (4-HT) (Sigma) was added at the time of activation
709 with IL-4, LPS and RP105.

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

711 Sheep red blood cells (SRBCs) were washed thrice with PBS of which 0.2x10^{9} SRBCs in 100µl PBS
712 (per mouse) were injected into B1-8^{hi} mice followed by another injection of 1x10^{9} SRBCs in 100µl PBS
713 five days later. Mice were harvested 10 days after the first immunization. Spleens were harvested and
714 single-cell suspensions were stained with B220 conjugated to fluorescein isothiocyanate (B220-FITC,
715 BD Biosciences, 1:500 dilution), Fas conjugated to PE-cyanine 7 (Fas-PE-Cy7, BD Biosciences,
716 1:1000 dilution) and CD38 conjugated to Allophycocyanin (CD38-APC, ThermoFisher, 1:200 dilution)
717 along with Fc block (BD Biosciences, 1:500 dilution). GCBs (B220^{+} Fas^{+} CD38^{-}) were isolated on a
718 BD FACS Aria III sorter (BD Biosciences). Following sorting, nuclei were isolated and used for PRO-
719 seq and PRO-cap (see below). For one experiment, ~10^{7} GCBs from fifteen immunized mice were
720 pooled yielding 3-4x10^{6} viable nuclei that were used for run-on. For MutPE-seq, genomic DNA was
721 isolated from 15 x10^{6} GCBs as described previously [117].

722 **PRO-seq and PRO-cap**

723 PRO-seq was performed as described previously [64] with several modifications.
724 To isolate nuclei, murine and *Drosphila* S2 cells were resuspended in cold Buffer IA (160 mM
725 Sucrose, 10 mM Tris-Cl pH 8, 3 mM $CaCl_2$, 2 mM $MgAc_2$, 0.5% NP-40, 1 mM DTT added fresh),
726 incubated on ice for 3 min and centrifuged at 700 g for 5 min. The pellet was resuspended in nuclei
727 resuspension buffer NRB (50 mM Tris-Cl pH 8, 40% Glycerol, 5 mM $MgCl_2$, 0.1 mM EDTA). For each
728 run-on, 10^{7} nuclei of the sample and 10% *Drosophila* S2 nuclei were combined in a total of 100 µL
729 NRB and incubated at 30°C for 3 min with 100 µL 2x NRO buffer including 5µl of each 1mM Bio-11-

730 NTPs. In some PRO-cap experiments, the run-on reaction was performed in the presence of two
731 biotinylated NTPs, Biotin-11-UTP and Biotin-11-CTP (Perkin-Elmer), and unlabeled ATP and GTP.
732 Subsequent steps were performed as described (Mahat et al 2016), except that 3' and 5' ligations
733 were performed at 16°C overnight and CapClip Pyrophosphatase (Biozym Scientific) was used for
734 5'end decapping. We also used customized adapters (PRO-seq: 3' RNA adapter:
735 5'5Phos/NNNNNNNGAUCGUCGGACUGUAGAACUCUGAAC/3InvdT-3' and 5' RNA adapter: 5'-
736 CCUUGGCACCCGAGAAUUCCANNNN-3'). RNA was reverse transcribed by SuperScript III RT with
737 RP1 Illumina primer to generate cDNA libraries. Libraries were amplified with barcoding Illumina RPI-x
738 primers and the universal forward primer RP1 using KAPA HiFi Real-Time PCR Library Amplification
739 Kit. PRO-cap: 3' linker (DNA oligo): 5rApp/NNNNAGATCGGAAGAGCACACGTCT/3ddC , 5'linker
740 (RNA): ACACUCUUUCCCUACACGACGCUCUUCCGAUCUNNNNNNNNNN, reverse transcription
741 and library amplification as in PRO-seq but using selected TruSeq_IDX_1-48 PCR primers for RT,
742 and the same TruSeq IDX PCR primers together with TruSeq Universal Adapter forward primer for
743 PCR. For both methods amplified libraries were subjected to gel electrophoresis on 2.5% low melting
744 agarose gel and amplicons from 150 to 350 bp were extracted from the gel (carefully separated from
745 possible linker dimers), multiplexed and sequenced on Illumina platforms with SR50 or longer read
746 mode.

747 **RT-qPCR**

748 RT-qPCR assays with externally spiked-in *Drosophila* S2 cells were described in detail in our previous
749 studies [118]. Briefly, *Drosophila* S2 cells were mixed with Ramos cells or B1-8^{hi} B cells at a ratio of 1:4
750 ($4x10^5$ B cells and $1x10^5$ S2 cells) followed by total RNA extraction with TRIzol reagent (Thermo
751 Fisher), DNaseI digestion, and cDNA synthesis with random primers. The 2-ΔΔCt method was used to
752 quantify the data using the *Drosophila Act5c* transcript for normalization.

753 **ChIP-qPCR and ChIP-seq**

754 ChIP-seq and ChIP-qPCR were performed as described previously without any modifications [27,117].
755 Antibodies used in this study are listed in Table S1.

756 **ATAC-seq**

757 ATAC-seq was performed exactly as described in detail in our previous work [117].

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

759 MutPE-seq was performed following the principles described in two previous reports [90,95] with several
760 adaptations. 80 ng of genomic DNA were amplified by PCR with the Kapa Hifi HS 2x RM (Roche
761 Diagnostics), For the first PCR, we used 20-25 cycles with 0.2µM locus-specific primers fused to a
762 varying number of random nucleotides and to the first part of the Illumina adapter sequences (FW: 5'-
763 CTCTTTCCCTACACGACGCTCTTCCGATCT-(N)x-gene-specific sequence-3'; RV: 5'-
764 CTGGAGTTCAGACGTGTGCTCTTCCGATCT-gene-specific sequence-3'; see Table S1 for a
765 complete list of primer sequences).

766 Random Ns are used to increase complexity and shift frames of very similar amplicons to improve
767 cluster calling and sample identification. After first PCR, the samples were purified with 0.2x/0.7x
768 SPRI beads (Beckman Coulter), eluted in 10µl water and amplified for 10 cycles with 0.75µM primers
769 containing linker sequences and dual barcoding (FW: 5'-
770 AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGAC-3'; RV: 5'-
771 CAAGCAGAAGACGGCATACGAGATXXXXXXGTGACTGGAGTTCAGACGTGTG-3' where the
772 stretches of X dNTPs (underlined) serve as specific barcodes creating a unique dual barcode
773 combination for each sample). PCR products were purified either by extraction from a 2% low-melt
774 agarose gel or with 0.7x SPRI beads and eluted in 20 µl water. The concentration was determined by
775 Picogreen-based measurements on a Nanodrop machine or by a Fragment Analyzer. Samples were
776 equimolarly pooled for next generation sequencing on an Illumina MiSeq flowcell and sequenced
777 paired-end (PE300).

778 **Primers and sgRNAs**

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

780

781 **Bioinformatics**

782 **Mapping PRO-seq, PRO-cap, ChIP-seq and ATAC-seq reads to IG V regions**

783 Reads were trimmed for standard adapters and low quality (Q<30) 3'-end bases and filtered for a
784 remaining length of at least 20nt (excluding the UMI, if applicable) using cutadapt [119]. Alignment to the

785 reference genome was done with Bowtie [120]. The NCBI GRCm38.6 assembly was used as the mouse
786 reference. To this we added the sequence of the recombined locus containing the V region as an
787 additional chromosome. Similarly, for human data, the Hg38 assembly was used, with the relevant V
788 region sequence added as a chromosome. Where applicable, the dmr6 assembly of *Drosophila*
789 *melanogaster* from Flybase ([121] release 6.27) was used as the spike-in reference and was added to
790 the alignment index. During alignment, up to three mismatches were allowed. To accommodate
791 mapping to the repetitive *IGH/Igh* locus, a high degree of multimapping was allowed (194 potential V
792 segments annotated for GRCm38.6). In the case of alignment with spike-ins, only reads mapping
793 exclusively to either genome were considered. For PCR deduplication, UMIs were identified and
794 filtered with UMI-tools ([122] v1.0.0). No sequence differences were allowed in UMIs when collapsing the
795 duplicates. Multimapping reads were then filtered to identify those specific to the V region sequence,
796 taking into account the repetitiveness of the reference *IGH/Igh* locus. Reads mapping to the fixed V
797 region sequence were allowed to multimap only to the native *IGH/Igh* locus (mouse
798 chr12:113572929-116009954 or human chr14:105836764-106875071). The native loci were defined
799 as the region between the earliest position belonging to an annotated V segment and the latest
800 position belonging to an annotated J segment. Reads with a mapped location outside these areas
801 were rejected. This filtering was implemented in a custom Python script. The qualifying reads were
802 then classified as reads mapping uniquely to the particular V region sequence and reads mapping to
803 both the V region sequence and to the native *IGH/Igh* region. Genome browser tracks were created
804 by quantifying and scaling read coverage of the V region sequence using bedtools ([123] v2-29.0).
805 Reads were split by strand, and strand labels for PRO-seq were inverted in order to match the strand
806 designations of the respective PRO-cap reads. For both data types, only the first 5' base of the
807 strand-adjusted reads was used for the tracks. Additionally, the 5' end of PRO-seq reads was shifted
808 by 1 nucleotide downstream, to compensate for the fact that the last nucleotide in the RNA was
809 incorporated during the run-on reaction. For normalized tracks, read counts were scaled to RPM
810 (reads per million). Finally, "-" strand coverage tracks were further scaled by -1, for visualization
811 purposes.

812

813

814 **Analysis of MutPE-seq**

815 Reads were trimmed for standard adapters with cutadapt [119]. Poor quality (Q<25) 3' bases were
816 trimmed with trimmomatic [124] by averaging over a sliding window of 5nt. Read pairs were then filtered
817 for minimum remaining length (200nt for read 1, 100nt for read 2) using cutadapt. Read mates were
818 merged down to make combined single-end reads with FLASH [125] allowing 10% mismatch between
819 the mates. Obvious erroneous mergers were removed by selecting combined reads with lengths
820 within ±30nt of the amplicon length using cutadapt, The remaining combined reads were aligned with
821 Bowtie2 [126], using the "–very-sensitive-local" alignment mode and only the fixed V region sequence
822 and its immediate vicinity as reference. Samples of V genes VH4-59 and VH4-34 are different only by
823 single nucleotide polymorphisms and were filtered for contamination by the respective second
824 sequence by jointly aligning with the "expected" and the other "contaminating" V region sequence,
825 discarding all aligned contaminating reads. A pile-up was generated with samtools [127] taking into
826 account only bases with quality of at least 30. The pileups were then quantified with a custom Python
827 script and the resulting mutation counts were processed and visualized with custom scripts in R
828 (v3.5.1), with the help of additional R packages (data.table [https://cran.r-
829 project.org/web/packages/data.table/index.html], ggplot2 [https://cran.r-
830 project.org/web/packages/ggplot2/index.html] , ggrepel [https://cran.r-
831 project.org/web/packages/ggrepel/index.html], patchwork [https://cran.r-
832 project.org/web/packages/patchwork/index.html]). Background mutation profiles were controlled for by
833 subtracting the corresponding mutation frequencies in control samples from the frequencies in the
834 samples of interest, at each position and for each substitution type. Annotation of hot and cold spots
835 was created by means of regex search for the corresponding patterns in the reference sequence.

836 **Comparison on Pol II stalling and mutation frequency in V regions**

837 Stalling sites were identified as described in Fong et al. [91] using the source code provided therein
838 (https://doi.org/10.5281/zenodo.7013905) with no changes. We used the same parameters as Fong et
839 al. to define stalling sites, namely, all sites covered by at least five PRO-seq reads and with a
840 standard deviation three times greater than the surrounding 200 bp window [91]. In addition, we
841 extended these sites by 3 bp on either side to create 7 bp stalling zones. Only stalling zones within
842 the MutPE-seq amplicon were considered and overlapping zones were merged. All other zones within

843 the MutPE-seq amplicon were considered as non-stalling zones. This approach yielded 18-21 stalling
844 zones and 17-19 non-stalling zones depending on the V region analyzed, which provided sufficient
845 datapoints for statistical analyses. Mutations in all zones were counted and divided by the width of the
846 zone to yield the normalized mutation rates which were displayed as box plots.

847 **Analysis of mutation frequency and nascent transcription at C:G residues**

848 To analyze the relationship between PRO-seq read density and mutation frequency at C:G pairs (Fig.
849 S3C-D), we first separated mutated and unmutated C:G residues into two groups. All C:G residues
850 were extended by 3 bp on either side to create 7 bp windows. For the mutated C:G group,
851 overlapping windows were merged. If a mutated window overlapped an unmutated one, the mutated
852 window was retained and the unmutated window was discarded from the analysis. In this manner,
853 only non-overlapping mutated and unmutated windows were compared. Next, the mutated windows
854 were split into ten equal groups (deciles) ranked from lowest to highest mutation frequency. Finally,
855 box plots were generated comparing the PRO-seq density (normalized for the length of the windows)
856 in the unmutated and mutated groups.

857 In a separate analysis (Fig. S3E-F), only the 7 bp windows centered on mutated C:G residues
858 were considered and, as above, overlapping windows were merged. Next, the PRO-seq densities in
859 all windows were obtained. Ten equal groups (deciles) were created and were ranked from lowest to
860 highest PRO-seq density. Box plots were generated to compare the mutation frequency of the central
861 mutated C:G between the ten groups.

862 **Statistical analysis**

863 Statistical tests were performed with the Wilcoxon rank sum test following multiple testing correction
864 with the Bonferroni method or the unpaired Students t-test, as indicated in the figure legends.

865 **Data availability**

866 All next-generation sequencing data has been deposited in GEO under accession number
867 GSE202042. Code for the workflow and the custom scripts is available on Github at
868 https://github.com/PavriLab/IgH_VDJ_PROcapseq and https://github.com/PavriLab/IgH_VDJ_MutPE.

869

870 **Author contributions**

871 UES performed all PRO-seq, PRO-cap and MutPE-seq assays and analyzed data. JF generated the
872 Ramos lines with new V regions and performed infections for MutPE-seq. KF and TN developed the
873 bioinformatic pipeline to map NGS reads to V regions. KF developed the bioinformatic pipeline for
874 MutPE-seq and analyzed data. RV and ITS analyzed MutPE-seq and performed the analysis of
875 mutation and Pol II stalling in V regions. Marina M performed ChIP-seq, ChIP-qPCR and RT-qPCR in
876 Ramos cells and analyzed data. IO performed the bioinformatic analysis of non-IG AID target genes.
877 BB performed ChIP-qPCR and RT-qPCRs. EMW and RP generated the $E\mu^{-/-}$ lines. MS conducted
878 GCB PRO-seq with UES. ACG performed ATAC-seq. Marialaura M cloned JP8Bdel expression
879 vectors for MutPE-seq. ACG performed ATAC-seq. HM provided the germline-reverted VH4-59 and
880 VH3-30 sequences and provided feedback on the manuscript. RP conceived the project, analyzed
881 data and wrote the manuscript with critical inputs from UES and JF.

882 **Acknowledgements**

883 We thank Almudena Ramiro and Angel Alvarez-Prado (CNIC, Madrid, Spain) for sharing the
884 mutational datasets from $Ung^{-/-}Msh2^{-/-}$ GCBs. We acknowledge the Vienna Biocenter Core Facilities
885 (VBCF) for assistance in generating $E\mu^{-/-}$ lines (VBCF-ProTech) and for next generation sequencing
886 (VBCF-NGS), the IMP/IMBA BioOptics facility for flow cytometry, the IMP/IMBA Molecular biology
887 services for Sanger sequencing and reagents, and the IMP/IMBA animal house. We thank Ryan
888 Sheridan and David Bentley (University of Colorado, USA) for advice on calling Pol II stalling sites.
889 We thank Maximilian von der Linde for uploading NGS tracks to GEO. We thank Carrie Bernecky
890 (ISTA, Klosterneuburg) and Clemens Plaschka (IMP, Vienna) for generating the 3D Pol II elongation
891 complex structural visualizations. This work was funded by Boehringer Ingelheim, The Austrian
892 Industrial Research Promotion Agency (Headquarter Grant FFG-834223), and grants from the
893 Austrian Science Fund to UES (FWF T 795-B30) and RP (FWF P 32043-B).

894

895 **Conflicts of interest**

896 The authors declare no conflicts of interest.

897 **REFERENCES**

898 1 Peled, J. U. *et al.* The Biochemistry of Somatic Hypermutation. *Annu Rev Immunol* **26**, 481-
899 511 (2008).
900 2 Odegard, V. H. & Schatz, D. G. Targeting of somatic hypermutation. *Nat Rev Immunol* **6**, 573-
901 583 (2006).
902 3 Di Noia, J. M. & Neuberger, M. S. Molecular mechanisms of antibody somatic
903 hypermutation. *Annu Rev Biochem* **76**, 1-22 (2007).
904 4 Victora, G. D. & Nussenzweig, M. C. Germinal centers. *Annu Rev Immunol* **30**, 429-457
905 (2012).
906 5 Muramatsu, M. *et al.* Class switch recombination and hypermutation require activation-
907 induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* **102**, 553-563 (2000).
908 6 Revy, P. *et al.* Activation-induced cytidine deaminase (AID) deficiency causes the autosomal
909 recessive form of the Hyper-IgM syndrome (HIGM2). *Cell* **102**, 565-575 (2000).
910 7 Petersen-Mahrt, S. K., Harris, R. S. & Neuberger, M. S. AID mutates E. coli suggesting a DNA
911 deamination mechanism for antibody diversification. *Nature* **418**, 99-103 (2002).
912 8 Ramiro, A. R., Stavropoulos, P., Jankovic, M. & Nussenzweig, M. C. Transcription enhances
913 AID-mediated cytidine deamination by exposing single-stranded DNA on the nontemplate
914 strand. *Nat Immunol* **4**, 452-456 (2003).
915 9 Bransteitter, R., Pham, P., Scharff, M. D. & Goodman, M. F. Activation-induced cytidine
916 deaminase deaminates deoxycytidine on single-stranded DNA but requires the action of
917 RNase. *Proc Natl Acad Sci U S A* **100**, 4102-4107 (2003).
918 10 Pham, P., Bransteitter, R., Petruska, J. & Goodman, M. F. Processive AID-catalysed cytosine
919 deamination on single-stranded DNA simulates somatic hypermutation. *Nature* **424**, 103-107
920 (2003).
921 11 Methot, S. P. & Di Noia, J. M. Molecular Mechanisms of Somatic Hypermutation and Class
922 Switch Recombination. *Adv Immunol* **133**, 37-87 (2017).
923 12 Vaidyanathan, B., Yen, W.-F., Pucella, J. & Chaudhuri, J. AIDing Chromatin and Transcription-
924 Coupled Orchestration of Immunoglobulin Class-Switch Recombination. *Frontiers in*
925 *Immunology* **5** (2014).
926 13 Leeman-Neill, R. J., Lim, J. & Basu, U. The Common Key to Class-Switch Recombination and
927 Somatic Hypermutation: Discovery of AID and Its Role in Antibody Gene Diversification. *The*
928 *Journal of Immunology* **201**, 2527-2529 (2018).
929 14 Saha, T., Sundaravinayagam, D. & Di Virgilio, M. Charting a DNA Repair Roadmap for
930 Immunoglobulin Class Switch Recombination. *Trends Biochem Sci* (2020).
931 15 Chaudhuri, J. & Alt, F. W. Class-switch recombination: interplay of transcription, DNA
932 deamination and DNA repair. *Nat Rev Immunol* **4**, 541-552 (2004).
933 16 Liu, M. *et al.* Two levels of protection for the B cell genome during somatic hypermutation.
934 *Nature* **451**, 841-845 (2008).
935 17 Alvarez-Prado, A. F. *et al.* A broad atlas of somatic hypermutation allows prediction of
936 activation-induced deaminase targets. *J Exp Med* **215**, 761-771 (2018).
937 18 Yamane, A. *et al.* Deep-sequencing identification of the genomic targets of the cytidine
938 deaminase AID and its cofactor RPA in B lymphocytes. *Nat Immunol* (2010).
939 19 Shen, H. M., Peters, A., Baron, B., Zhu, X. & Storb, U. Mutation of BCL-6 gene in normal B
940 cells by the process of somatic hypermutation of Ig genes. *Science* **280**, 1750-1752 (1998).
941 20 Pasqualucci, L. *et al.* Hypermutation of multiple proto-oncogenes in B-cell diffuse large-cell
942 lymphomas. *Nature* **412**, 341-346 (2001).
943 21 Ramiro, A. R. *et al.* AID is required for c-myc/IgH chromosome translocations in vivo. *Cell*
944 **118**, 431-438 (2004).

22 Buerstedde, J. M., Alinikula, J., Arakawa, H., McDonald, J. J. & Schatz, D. G. Targeting of somatic hypermutation by immunoglobulin enhancer and enhancer-like sequences. *PLoS Biol* **12**, e1001831 (2014).

23 Dinesh, R. K. *et al.* Transcription factor binding at Ig enhancers is linked to somatic hypermutation targeting. *Eur J Immunol* **50**, 380-395 (2020).

24 Qian, J. *et al.* B cell super-enhancers and regulatory clusters recruit AID tumorigenic activity. *Cell* **159**, 1524-1537 (2014).

25 Senigl, F. *et al.* Topologically Associated Domains Delineate Susceptibility to Somatic Hypermutation. *Cell Rep* **29**, 3902-3915 e3908 (2019).

26 Methot, S. P. *et al.* A licensing step links AID to transcription elongation for mutagenesis in B cells. *Nat Commun* **9**, 1248 (2018).

27 Pavri, R. *et al.* Activation-induced cytidine deaminase targets DNA at sites of RNA polymerase II stalling by interaction with Spt5. *Cell* **143**, 122-133 (2010).

28 Willmann, K. L. *et al.* A role for the RNA pol II-associated PAF complex in AID-induced immune diversification. *J Exp Med* **209**, 2099-2111 (2012).

29 Begum, N. A., Stanlie, A., Nakata, M., Akiyama, H. & Honjo, T. The histone chaperone Spt6 is required for activation-induced cytidine deaminase target determination through H3K4me3 regulation. *J Biol Chem* **287**, 32415-32429 (2012).

30 Nambu, Y. *et al.* Transcription-coupled events associating with immunoglobulin switch region chromatin. *Science* **302**, 2137-2140 (2003).

31 Basu, U. *et al.* The RNA exosome targets the AID cytidine deaminase to both strands of transcribed duplex DNA substrates. *Cell* **144**, 353-363 (2011).

32 Nowak, U., Matthews, A. J., Zheng, S. & Chaudhuri, J. 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).

33 Duan, Z. *et al.* Role of Dot1L and H3K79 methylation in regulating somatic hypermutation of immunoglobulin genes. *Proc Natl Acad Sci U S A* **118** (2021).

34 Yu, G. *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).

35 Aida, M., Hamad, N., Stanlie, A., Begum, N. A. & Honjo, T. 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).

36 Jeevan-Raj, B. P. *et al.* Epigenetic tethering of AID to the donor switch region during immunoglobulin class switch recombination. *J Exp Med* **208**, 1649-1660 (2011).

37 Stanlie, A., Aida, M., Muramatsu, M., Honjo, T. & Begum, N. A. 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).

38 Bradley, S. P., Kaminski, D. A., Peters, A. H. F. M., Jenuwein, T. & Stavnezer, J. The Histone Methyltransferase Suv39h1 Increases Class Switch Recombination Specifically to IgA. *The Journal of Immunology* **177**, 1179-1188 (2006).

39 Kuang, F. L., Luo, Z. & Scharff, M. D. 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).

40 Daniel, J. A. *et al.* PTIP promotes chromatin changes critical for immunoglobulin class switch recombination. *Science* **329**, 917-923 (2010).

41 Vaidyanathan, B. & Chaudhuri, J. Epigenetic Codes Programing Class Switch Recombination. *Front Immunol* **6**, 405 (2015).

42 Larijani, M. & Martin, A. Single-stranded DNA structure and positional context of the target cytidine determine the enzymatic efficiency of AID. *Mol Cell Biol* **27**, 8038-8048 (2007).

43 Larijani, M. *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).

44 Shen, H. M. & Storb, U. 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).

45 Branton, S. A. *et al.* Activation-induced cytidine deaminase can target multiple topologies of double-stranded DNA in a transcription-independent manner. *FASEB J* **34**, 9245-9268 (2020).

46 Taylor, B. J., Wu, Y. L. & Rada, C. Active RNAP pre-initiation sites are highly mutated by cytidine deaminases in yeast, with AID targeting small RNA genes. *Elife* **3**, e03553 (2014).

47 Kodgire, P., Mukkawar, P., Ratnam, S., Martin, T. E. & Storb, U. Changes in RNA polymerase II progression influence somatic hypermutation of Ig-related genes by AID. *J Exp Med* **210**, 1481-1492 (2013).

48 Wang, X., Fan, M., Kalis, S., Wei, L. & Scharff, M. D. A source of the single-stranded DNA substrate for activation-induced deaminase during somatic hypermutation. *Nat Commun* **5**, 4137 (2014).

49 Pefanis, E. *et al.* Noncoding RNA transcription targets AID to divergently transcribed loci in B cells. *Nature* **514**, 389-393 (2014).

50 Pefanis, E. & Basu, U. RNA Exosome Regulates AID DNA Mutator Activity in the B Cell Genome. *Adv Immunol* **127**, 257-308 (2015).

51 Meng, F. L. *et al.* Convergent transcription at intragenic super-enhancers targets AID-initiated genomic instability. *Cell* **159**, 1538-1548 (2014).

52 Rajagopal, D. *et al.* Immunoglobulin switch mu sequence causes RNA polymerase II accumulation and reduces dA hypermutation. *J Exp Med* **206**, 1237-1244 (2009).

53 Wang, L., Wuerffel, R., Feldman, S., Khamlichi, A. A. & Kenter, A. L. S region sequence, RNA polymerase II, and histone modifications create chromatin accessibility during class switch recombination. *J Exp Med* **206**, 1817-1830 (2009).

54 Maul, R. W. *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).

55 Tarsalainen, A. *et al.* Ig Enhancers Increase RNA Polymerase II Stalling at Somatic Hypermutation Target Sequences. *J Immunol* **208**, 143-154 (2022).

56 Canugovi, C., Samaranayake, M. & Bhagwat, A. S. Transcriptional pausing and stalling causes multiple clustered mutations by human activation-induced deaminase. *The FASEB Journal* **23**, 34-44 (2009).

57 Parsa, J. Y. *et al.* Negative supercoiling creates single-stranded patches of DNA that are substrates for AID-mediated mutagenesis. *PLoS Genet* **8**, e1002518 (2012).

58 Ronai, D. *et al.* Detection of chromatin-associated single-stranded DNA in regions targeted for somatic hypermutation. *J Exp Med* **204**, 181-190 (2007).

59 Yu, G. *et al.* A Bayesian model based computational analysis of the relationship between bisulfite accessible single-stranded DNA in chromatin and somatic hypermutation of immunoglobulin genes. *PLoS Comput Biol* **17**, e1009323 (2021).

60 Peters, A. & Storb, U. Somatic hypermutation of immunoglobulin genes is linked to transcription initiation. *Immunity* **4**, 57-65 (1996).

61 Kenter, A. L. AID targeting is dependent on RNA polymerase II pausing. *Semin Immunol* **24**, 281-286 (2012).

62 Pavri, R. & Nussenzweig, M. C. AID targeting in antibody diversity. *Adv Immunol* **110**, 1-26 (2011).

63 Maul, R. W. & Gearhart, P. J. AID and somatic hypermutation. *Adv Immunol* **105**, 159-191 (2010).

64 Mahat, D. B. *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).

65 Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845-1848 (2008).

66 Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339**, 950-953 (2013).

67 Core, L. J. *et al.* Defining the status of RNA polymerase at promoters. *Cell Rep* **2**, 1025-1035 (2012).

68 Core, L. J. *et al.* Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat Genet* **46**, 1311-1320 (2014).

69 Andersson, R. *et al.* Human Gene Promoters Are Intrinsically Bidirectional. *Mol Cell* **60**, 346-347 (2015).

70 Scruggs, B. S. *et al.* Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. *Mol Cell* **58**, 1101-1112 (2015).

71 Williams, L. H. *et al.* Pausing of RNA polymerase II regulates mammalian developmental potential through control of signaling networks. *Mol Cell* **58**, 311-322 (2015).

72 Henriques, T. *et al.* Widespread transcriptional pausing and elongation control at enhancers. *Genes Dev* **32**, 26-41 (2018).

73 Sale, J. E. & Neuberger, M. S. TdT-accessible breaks are scattered over the immunoglobulin V domain in a constitutively hypermutating B cell line. *Immunity* **9**, 859-869 (1998).

74 Martin, A. *et al.* Activation-induced cytidine deaminase turns on somatic hypermutation in hybridomas. *Nature* **415**, 802-806 (2002).

75 Wu, L. *et al.* HMCES protects immunoglobulin genes specifically from deletions during somatic hypermutation. *Genes Dev* **36**, 433-450 (2022).

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

77 Yabuki, M., Fujii, M. M. & Maizels, N. The MRE11-RAD50-NBS1 complex accelerates somatic hypermutation and gene conversion of immunoglobulin variable regions. *Nat Immunol* **6**, 730-736 (2005).

78 Le, Q. & Maizels, N. Cell Cycle Regulates Nuclear Stability of AID and Determines the Cellular Response to AID. *PLOS Genetics* **11**, e1005411 (2015).

79 Pinaud, E. *et al.* The IgH locus 3' regulatory region: pulling the strings from behind. *Adv Immunol* **110**, 27-70 (2011).

80 Shih, T. A., Roederer, M. & Nussenzweig, M. C. Role of antigen receptor affinity in T cell-independent antibody responses in vivo. *Nat Immunol* **3**, 399-406 (2002).

81 Hwang, J. K. *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).

82 Ito, S. *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).

83 Wang, M., Yang, Z., Rada, C. & Neuberger, M. S. AID upmutants isolated using a high-throughput screen highlight the immunity/cancer balance limiting DNA deaminase activity. *Nat Struct Mol Biol* **16**, 769-776 (2009).

84 Perlot, T., Alt, F. W., Bassing, C. H., Suh, H. & Pinaud, E. Elucidation of IgH intronic enhancer functions via germ-line deletion. *Proc Natl Acad Sci U S A* **102**, 14362-14367 (2005).

85 Li, F., Yan, Y., Pieretti, J., Feldman, D. A. & Eckhardt, L. A. 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).

86 Yamaguchi, Y. *et al.* NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. *Cell* **97**, 41-51 (1999).

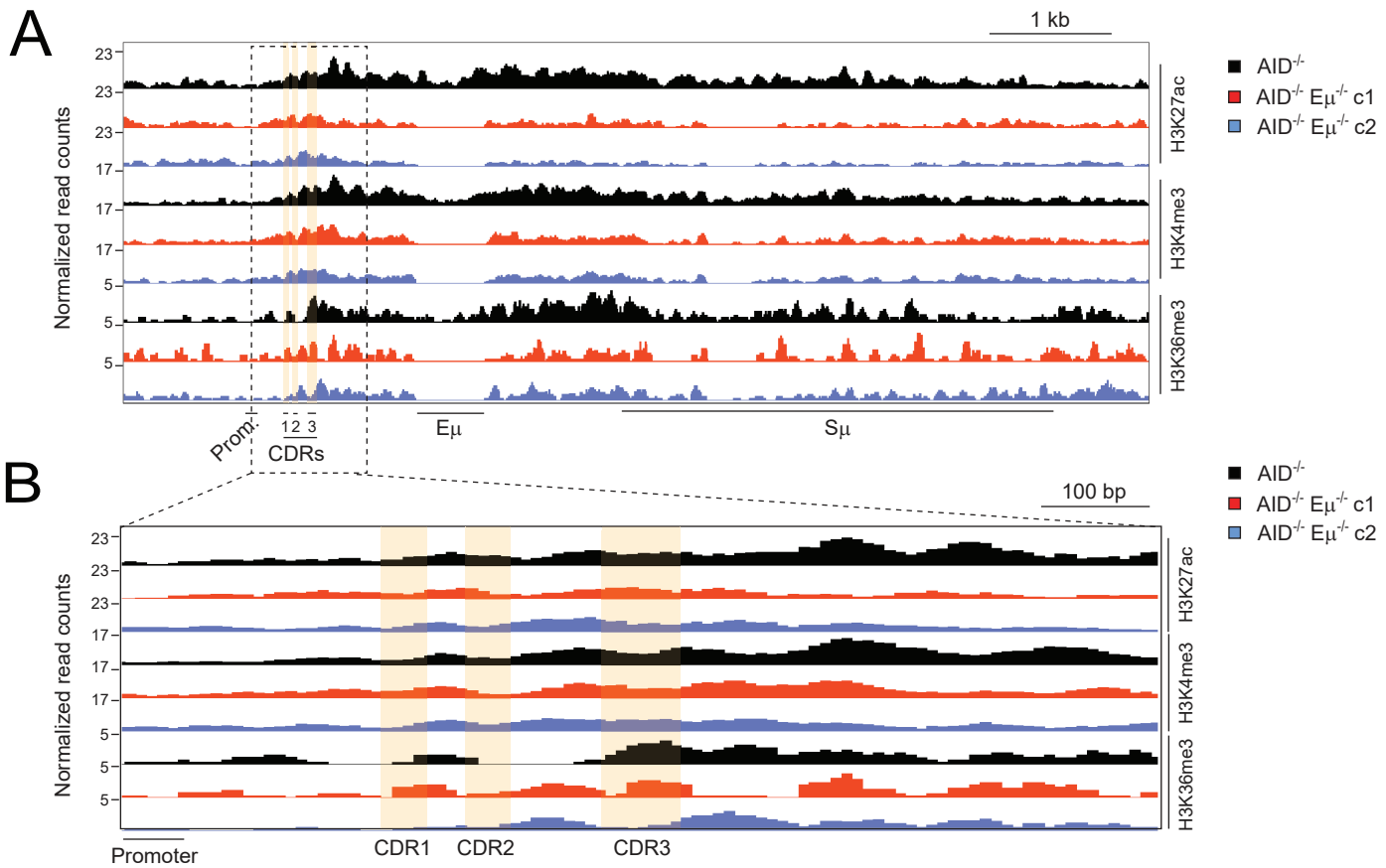
87 Vos, S. M., Farnung, L., Urlaub, H. & Cramer, P. Structure of paused transcription complex Pol II–DSIF–NELF. *Nature* **560**, 601-606 (2018).

88 Vos, S. M. *et al.* Structure of activated transcription complex Pol II–DSIF–PAF–SPT6. *Nature* **560**, 607-612 (2018).

1096 89 Bernecky, C., Plitzko, J. M. & Cramer, P. Structure of a transcribing RNA polymerase II–DSIF
1097 complex reveals a multidentate DNA–RNA clamp. *Nature Structural & Molecular Biology* **24**,
1098 809-815 (2017).
1099 90 Robbiani, D. F. *et al.* Plasmodium Infection Promotes Genomic Instability and AID-Dependent
1100 B Cell Lymphoma. *Cell* **162**, 727-737 (2015).
1101 91 Fong, N., Sheridan, R. M., Ramachandran, S. & Bentley, D. L. The pausing zone and control of
1102 RNA polymerase II elongation by Spt5: Implications for the pause-release model. *Mol Cell* **82**,
1103 3632-3645 e3634 (2022).
1104 92 Mouquet, H. *et al.* Complex-type N-glycan recognition by potent broadly neutralizing HIV
1105 antibodies. *Proc Natl Acad Sci U S A* **109**, E3268-3277 (2012).
1106 93 Lorin, V. *et al.* Epitope convergence of broadly HIV-1 neutralizing IgA and IgG antibody
1107 lineages in a viremic controller. *J Exp Med* **219** (2022).
1108 94 Matsuda, F. *et al.* The complete nucleotide sequence of the human immunoglobulin heavy
1109 chain variable region locus. *J Exp Med* **188**, 2151-2162 (1998).
1110 95 Yeap, L. S. *et al.* Sequence-Intrinsic Mechanisms that Target AID Mutational Outcomes on
1111 Antibody Genes. *Cell* **163**, 1124-1137 (2015).
1112 96 Shih, T. A., Meffre, E., Roederer, M. & Nussenzweig, M. C. Role of BCR affinity in T cell
1113 dependent antibody responses in vivo. *Nat Immunol* **3**, 570-575 (2002).
1114 97 Rush, J. S., Liu, M., Odegard, V. H., Unniraman, S. & Schatz, D. G. Expression of activation-
1115 induced cytidine deaminase is regulated by cell division, providing a mechanistic basis for
1116 division-linked class switch recombination. *Proc Natl Acad Sci U S A* **102**, 13242-13247
1117 (2005).
1118 98 Wang, Q. *et al.* The cell cycle restricts activation-induced cytidine deaminase activity to early
1119 G1. *J Exp Med* **214**, 49-58 (2017).
1120 99 Rada, C., Di Noia, J. M. & Neuberger, M. S. Mismatch recognition and uracil excision provide
1121 complementary paths to both Ig switching and the A/T-focused phase of somatic mutation.
1122 *Mol Cell* **16**, 163-171 (2004).
1123 100 Longerich, S., Tanaka, A., Bozek, G., Nicolae, D. & Storb, U. The very 5' end and the constant
1124 region of Ig genes are spared from somatic mutation because AID does not access these
1125 regions. *J Exp Med* **202**, 1443-1454 (2005).
1126 101 Xue, K., Rada, C. & Neuberger, M. S. The in vivo pattern of AID targeting to immunoglobulin
1127 switch regions deduced from mutation spectra in msh2-/- ung-/- mice. *J Exp Med* **203**, 2085-
1128 2094 (2006).
1129 102 Wang, Q. *et al.* Epigenetic targeting of activation-induced cytidine deaminase. *Proc Natl*
1130 *Acad Sci U S A* **111**, 18667-18672 (2014).
1131 103 Laffleur, B. *et al.* Noncoding RNA processing by DIS3 regulates chromosomal architecture
1132 and somatic hypermutation in B cells. *Nature Genetics* **53**, 230-242 (2021).
1133 104 Lim, J. *et al.* Nuclear Proximity of Mtr4 to RNA Exosome Restricts DNA Mutational
1134 Asymmetry. *Cell* **169**, 523-537 e515 (2017).
1135 105 Kazadi, D. *et al.* Effects of senataxin and RNA exosome on B-cell chromosomal integrity.
1136 *Heliyon* **6**, e03442 (2020).
1137 106 Lykke-Andersen, S. *et al.* Integrator is a genome-wide attenuator of non-productive
1138 transcription. *Mol Cell* **81**, 514-529.e516 (2021).
1139 107 Elrod, N. D. *et al.* The Integrator Complex Attenuates Promoter-Proximal Transcription at
1140 Protein-Coding Genes. *Mol Cell* **76**, 738-752.e737 (2019).
1141 108 Stein, C. B. *et al.* Integrator endonuclease drives promoter-proximal termination at all RNA
1142 polymerase II-transcribed loci. *Mol Cell* **82**, 4232-4245 e4211 (2022).
1143 109 Peron, S. *et al.* AID-driven deletion causes immunoglobulin heavy chain locus suicide
1144 recombination in B cells. *Science* **336**, 931-934 (2012).
1145 110 Bal, E. *et al.* Super-enhancer hypermutation alters oncogene expression in B cell lymphoma.
1146 *Nature* **607**, 808-815 (2022).

111 Nair, L., Chung, H. & Basu, U. Regulation of long non-coding RNAs and genome dynamics by the RNA surveillance machinery. *Nat Rev Mol Cell Biol* **21**, 123-136 (2020).
112 Zhou, J. Q. & Kleinstein, S. H. Position-Dependent Differential Targeting of Somatic Hypermutation. *J Immunol* **205**, 3468-3479 (2020).
113 Spisak, N., Walczak, A. M. & Mora, T. Learning the heterogeneous hypermutation landscape of immunoglobulins from high-throughput repertoire data. *Nucleic Acids Research* **48**, 10702-10712 (2020).
114 Pham, P., Calabrese, P., Park, S. J. & Goodman, M. F. 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).
115 Tang, C., Krantsevich, A. & MacCarthy, T. Deep learning model of somatic hypermutation reveals importance of sequence context beyond hotspot targeting. *iScience* **25**, 103668 (2022).
116 Wang, Y. *et al.* Mesoscale DNA feature in antibody-coding sequence facilitates somatic hypermutation. *Cell* **186**, 2193-2207 e2119 (2023).
117 Fitz, J. *et al.* Spt5-mediated enhancer transcription directly couples enhancer activation with physical promoter interaction. *Nat Genet* **52**, 505-515 (2020).
118 Fitz, J., Neumann, T. & Pavri, R. Regulation of RNA polymerase II processivity by Spt5 is restricted to a narrow window during elongation. *EMBO J* **37** (2018).
119 Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011* **17**, 3 (2011).
120 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).
121 Larkin, A. *et al.* FlyBase: updates to the Drosophila melanogaster knowledge base. *Nucleic Acids Research* **49**, D899-D907 (2020).
122 Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* **27**, 491-499 (2017).
123 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
124 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
125 Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957-2963 (2011).
126 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357-359 (2012).
127 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

Figure 1


A
1 kb
Normalized read counts
23
23
23
17
17
17
5
5
5
H3K27ac
H3K4me3
H3K36me3
AID-/-
AID-/- Eμ-/- c1
AID-/- Eμ-/- c2
Prom.
1 2 3
CDRs
Eμ
Sμ
B
100 bp
Promoter
CDR1
CDR2
CDR3

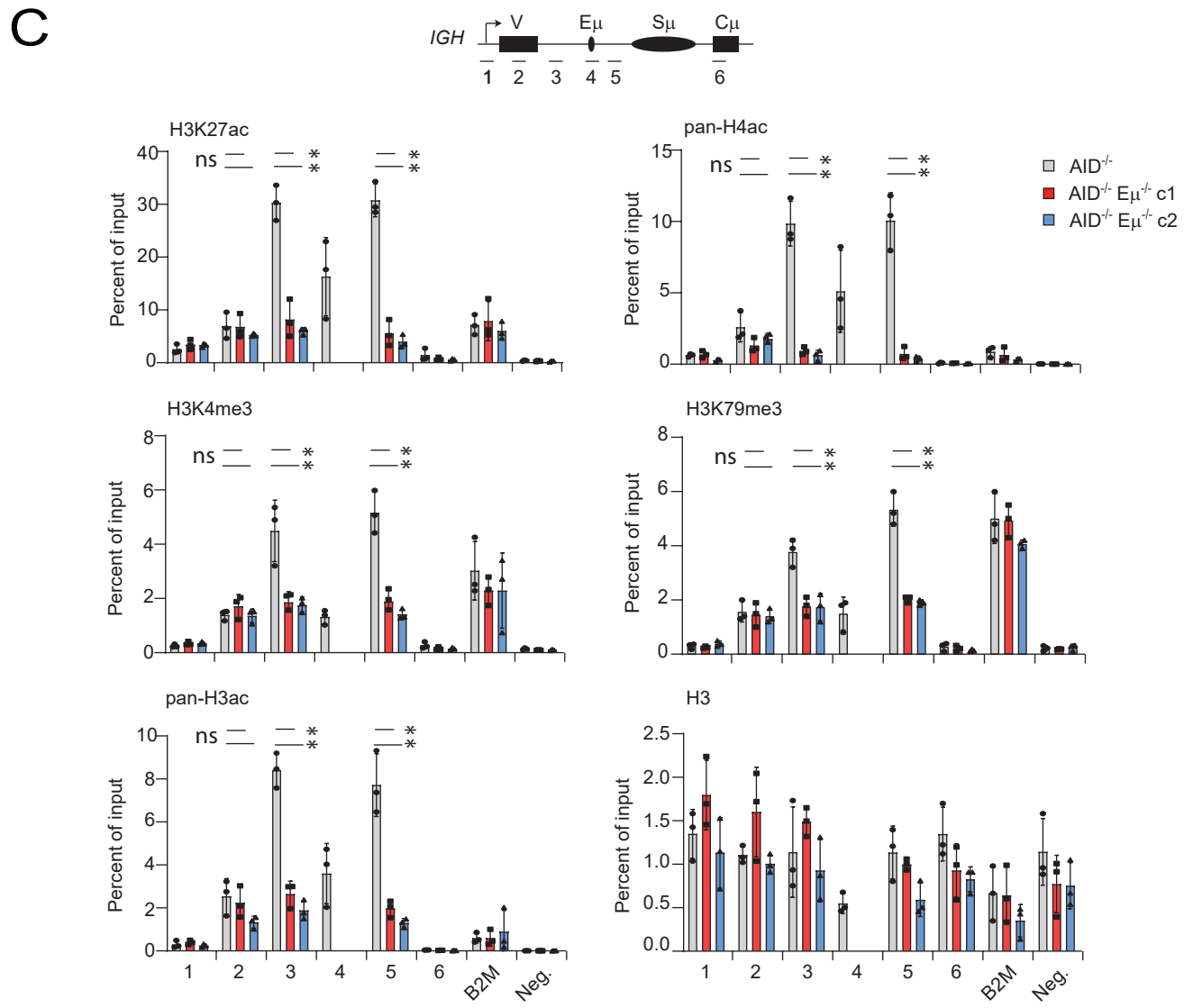
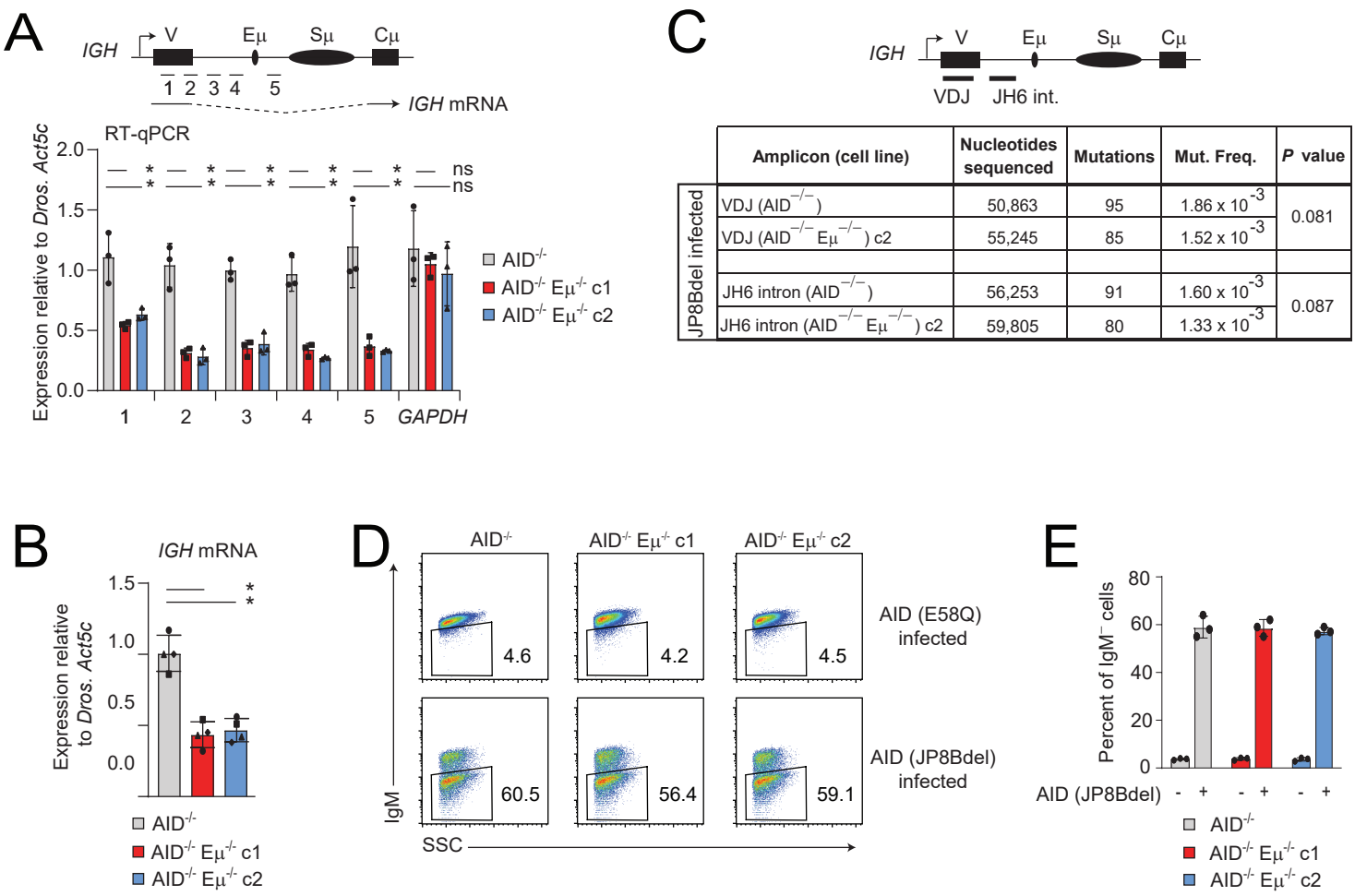

C
IGH
V
Eμ
Sμ
Cμ
1 2 3 4 5 6
H3K27ac
pan-H4ac
H3K4me3
H3K79me3
pan-H3ac
H3
Percent of input
ns
*
*
AID-/-
AID-/- Eμ-/- c1
AID-/- Eμ-/- c2
1 2 3 4 5 6 B2M Neg.

Figure 1: The Eμ enhancer regulates the chromatin landscape in the intron but not in the V 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 segment 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



	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}	

Figure 2: Ablation of the Eμ enhancer significantly decreases nascent transcription but not SHM. (A) RT-qPCR measurements of nascent transcripts at the Ramos V region in AID$^{-/-}$ and AID$^{-/-}$ Eμ$^{-/-}$ cells (clones c1 and 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 *IGH* mRNA. **(C)** Table of mutation frequencies at the Ramos V 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 7 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. **(E)** Bar graph summarizing the flow cytometry analyses from D, showing the percent of IgM loss from three independent experiments.

Figure 3

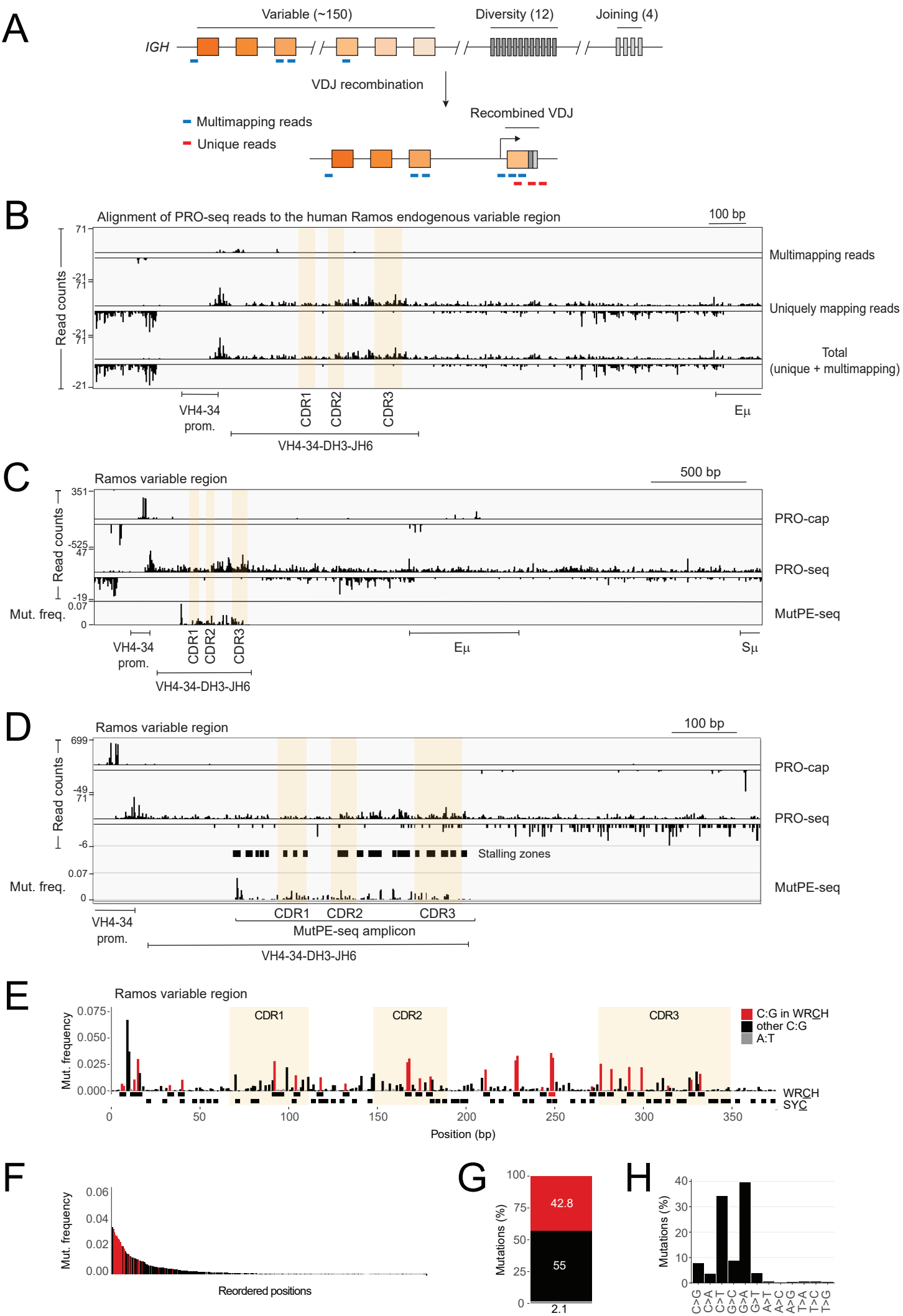

A
Variable (~150)
Diversity (12)
Joining (4)
IGH
VDJ recombination
Recombined VDJ
Multimapping reads
Unique reads
B
Alignment of PRO-seq reads to the human Ramos endogenous variable region
100 bp
Read counts
Multimapping reads
Uniquely mapping reads
Total
(unique + multimapping)
VH4-34
prom.
CDR1
CDR2
CDR3
Eμ
VH4-34-DH3-JH6
C
Ramos variable region
500 bp
Read counts
Mut. freq.
PRO-cap
PRO-seq
MutPE-seq
VH4-34
prom.
CDR1
CDR2
CDR3
Eμ
Sμ
VH4-34-DH3-JH6
D
Ramos variable region
100 bp
Read counts
Mut. freq.
PRO-cap
PRO-seq
Stalling zones
MutPE-seq
VH4-34
prom.
CDR1
CDR2
CDR3
MutPE-seq amplicon
VH4-34-DH3-JH6
E
Ramos variable region
Mut. frequency
CDR1
CDR2
CDR3
C:G in WRCH
other C:G
A:T
WRCH
SYC
Position (bp)
F
Mut. frequency
Reordered positions
G
Mutations (%)
42.8
55
2.1
H
Mutations (%)
C>G
C>A
C>T
G>C
G>A
G>T
A>T
A>C
A>G
T>A
T>C
T>G

Figure 3: Comparison of transcriptional and mutational landscapes at the endogenous Ramos V region. (A) Scheme depicting the strategy used to align multimapping reads to the V 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 V 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 V regions to the mouse genome. **(B)** Integrative Genomics Viewer (IGV) browser snapshot of nascent RNA 3' ends (by PRO-seq) aligned to the Ramos V 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 V region displayed on the Integrative Genomics Viewer (IGV) browser. Mapping of the V 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 7 days. **(D)** A magnified view of the V region from C above. The stalling zones are shown in the panel between the PRO-seq and MutPE-seq tracks. **(E)** Details of the mutation patterns at the Ramos V region. Mutated cytidines in AID hotspot motifs (WRCH) 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 (WRCH in black, AGCT in red, upper panel) and coldspot (SYC, 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 to T and G to A transition mutations are the signature of AID activity.

Figure 4

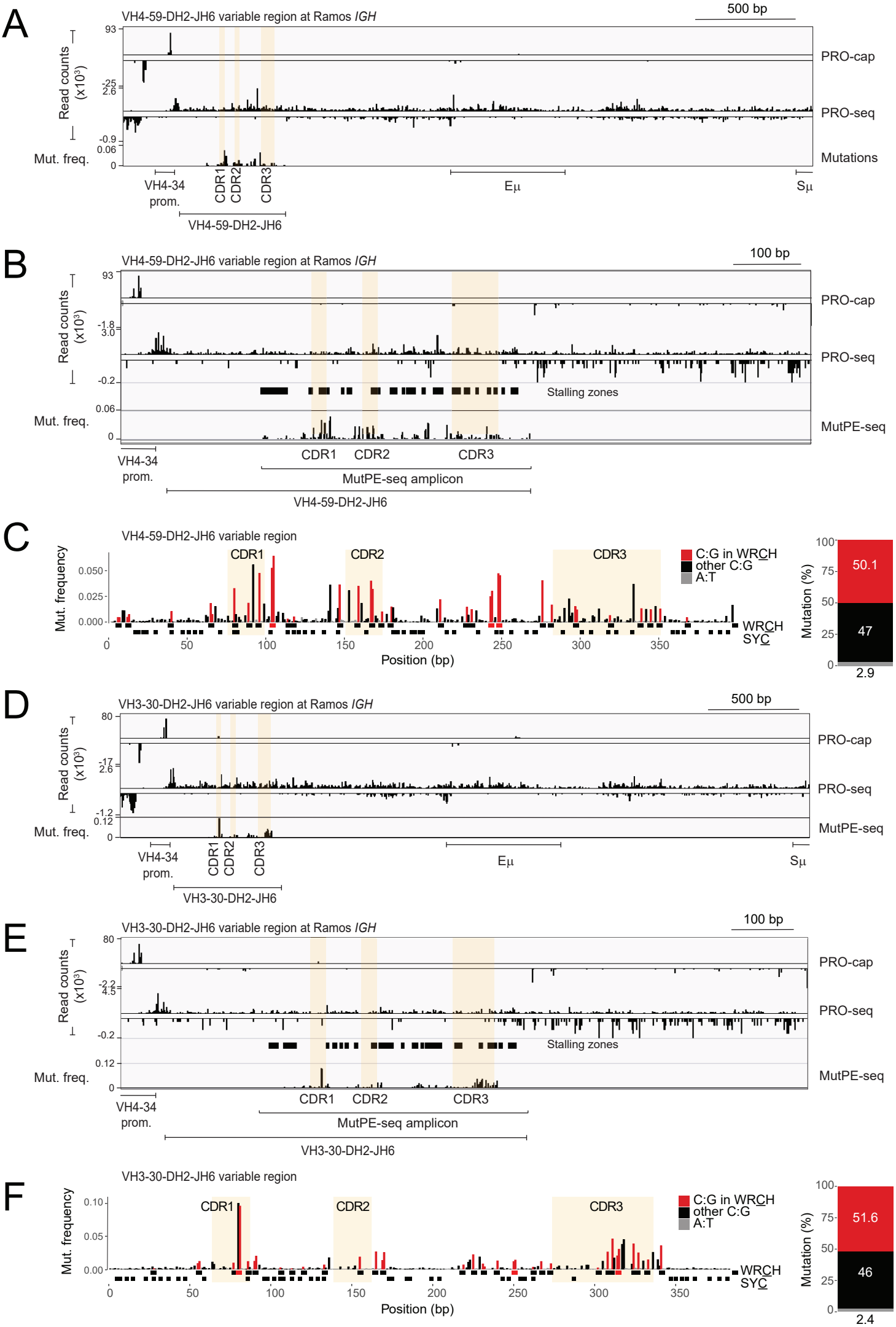
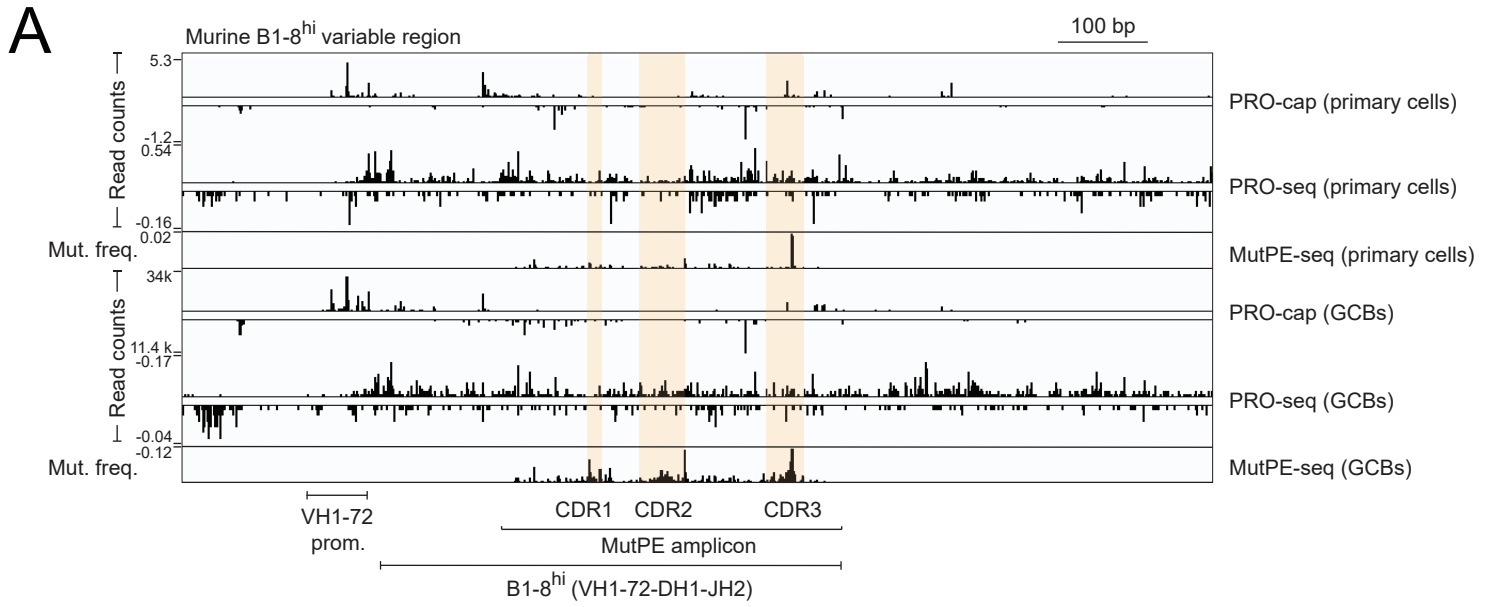
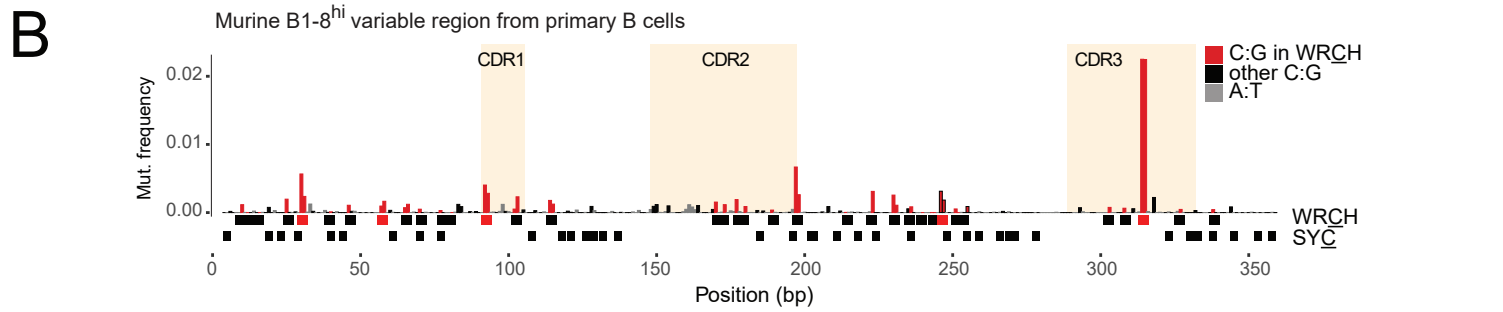
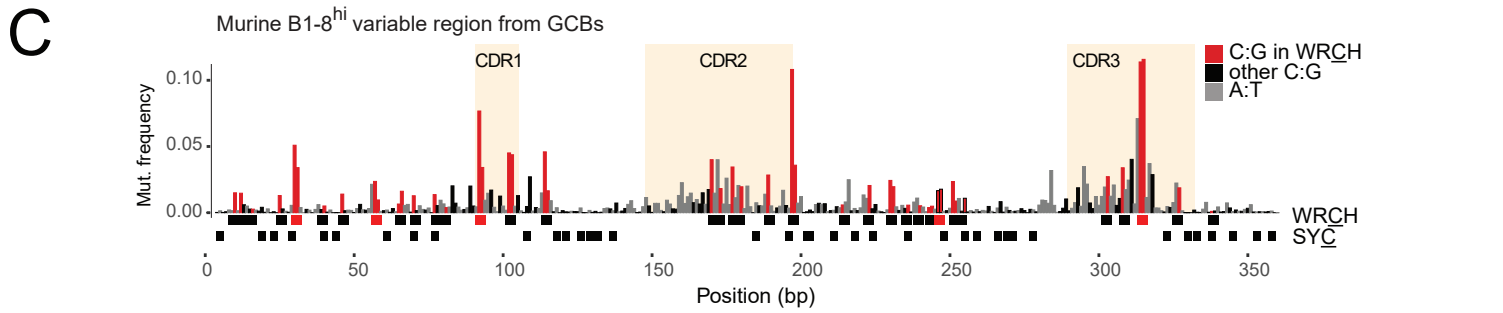

A
VH4-59-DH2-JH6 variable region at Ramos IGH
500 bp
Read counts (x10³)
PRO-cap
PRO-seq
Mut. freq.
Mutations
VH4-34 prom.
CDR1
CDR2
CDR3
Eμ
Sμ
VH4-59-DH2-JH6
B
VH4-59-DH2-JH6 variable region at Ramos IGH
100 bp
Stalling zones
MutPE-seq
MutPE-seq amplicon
C
VH4-59-DH2-JH6 variable region
Mut. frequency
C:G in WRCH
other C:G
A:T
WRCH
SYC
Position (bp)
Mutation (%)
50.1
47
2.9
D
VH3-30-DH2-JH6 variable region at Ramos IGH
VH3-30-DH2-JH6
E
VH3-30-DH2-JH6 variable region at Ramos IGH
F
VH3-30-DH2-JH6 variable region
51.6
46
2.4

Figure 4: Comparison of transcriptional and mutational landscapes at two different human V 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 V 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 V region from A above. The stalling zones are shown in the panel between the PRO-seq and MutPE-seq tracks. **(C)** Detailed mutational analysis of the VH4-59-DH2-JH6 V 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 and mutational profiles of the VH3-30-DH2-JH6 V region as in A. **(E)** A magnified view of the VH3-30-DH2-JH6 V region from D above. The stalling zones are shown in the panel between the PRO-seq and MutPE-seq tracks. **(F)** Detailed mutational analysis of the VH3-30-DH2-JH6 V region (see C above for details).

Figure 5


A
Murine B1-8hi variable region
100 bp
Read counts
Mut. freq.
PRO-cap (primary cells)
PRO-seq (primary cells)
MutPE-seq (primary cells)
PRO-cap (GCBs)
PRO-seq (GCBs)
MutPE-seq (GCBs)
VH1-72 prom.
CDR1
CDR2
CDR3
MutPE amplicon
B1-8hi (VH1-72-DH1-JH2)


B
Murine B1-8hi variable region from primary B cells
Mut. frequency
CDR1
CDR2
CDR3
C:G in WRCH
other C:G
A:T
WRCH
SYC
Position (bp)


C
Murine B1-8hi variable region from GCBs
Mut. frequency
CDR1
CDR2
CDR3
C:G in WRCH
other C:G
A:T
WRCH
SYC
Position (bp)

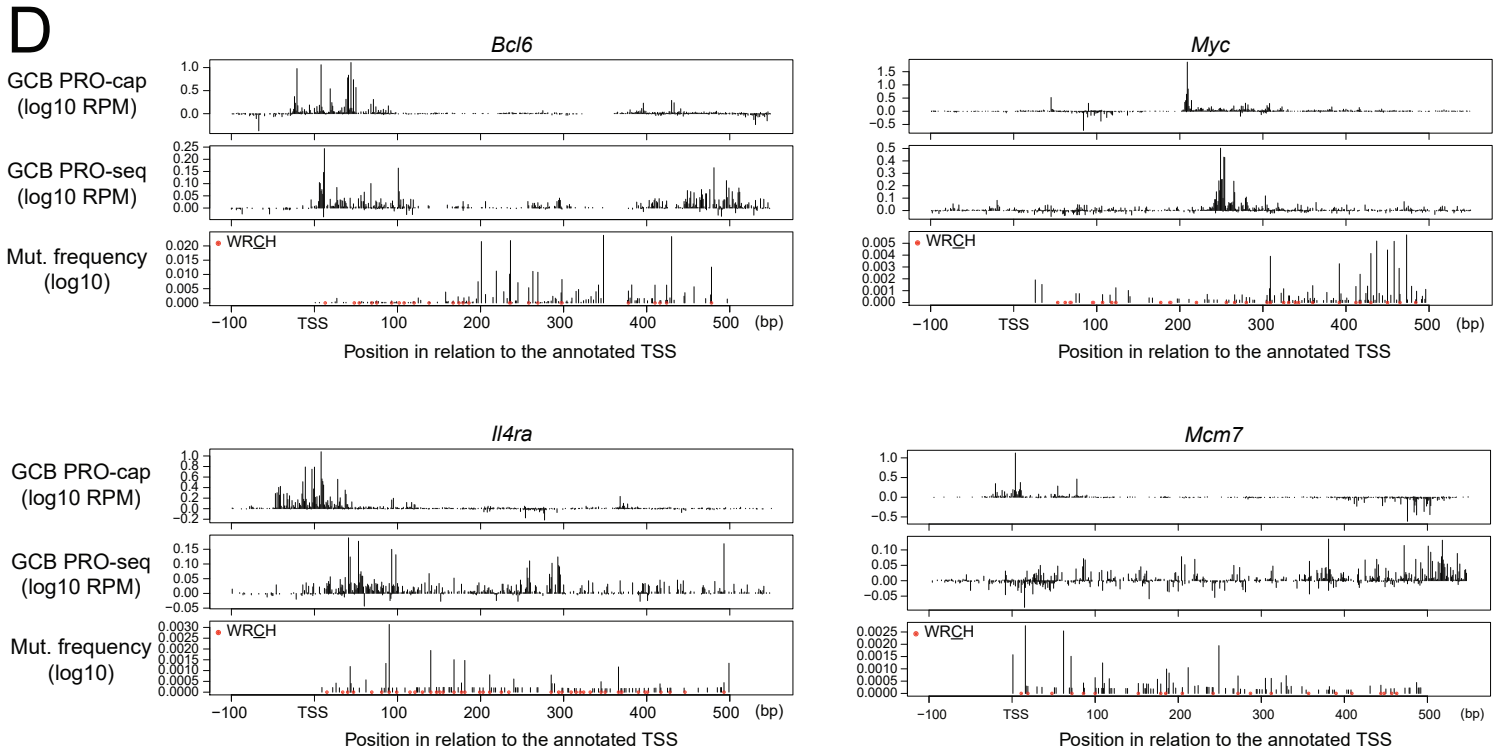

D
Bcl6
Myc
Il4ra
Mcm7
GCB PRO-cap (log10 RPM)
GCB PRO-seq (log10 RPM)
Mut. frequency (log10)
WRCH
TSS
(bp)
Position in relation to the annotated TSS

Figure 5: Transcriptional and mutational landscapes of the murine B1-8^{hi} V 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, *Igh*$^{B1\text{-}8hi/\ B1\text{-}8hi}$ *Rosa26*$^{AIDER/AIDER}$ primary B cells were activated with LPS, IL4 and RP105 for 4 days with 4-HT. 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 10 days. **(B)** Analysis of SHM at the B1-8^{hi} V 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 at the B1-8^{hi} V region from splenic GCBs following immunization with sheep red blood cells for 10 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 bp from the annotated (RefSeq) TSS up to 50 bp downstream of the sequenced amplicon. The WR<u>C</u>H motifs are indicated as red dots.

Figure 6

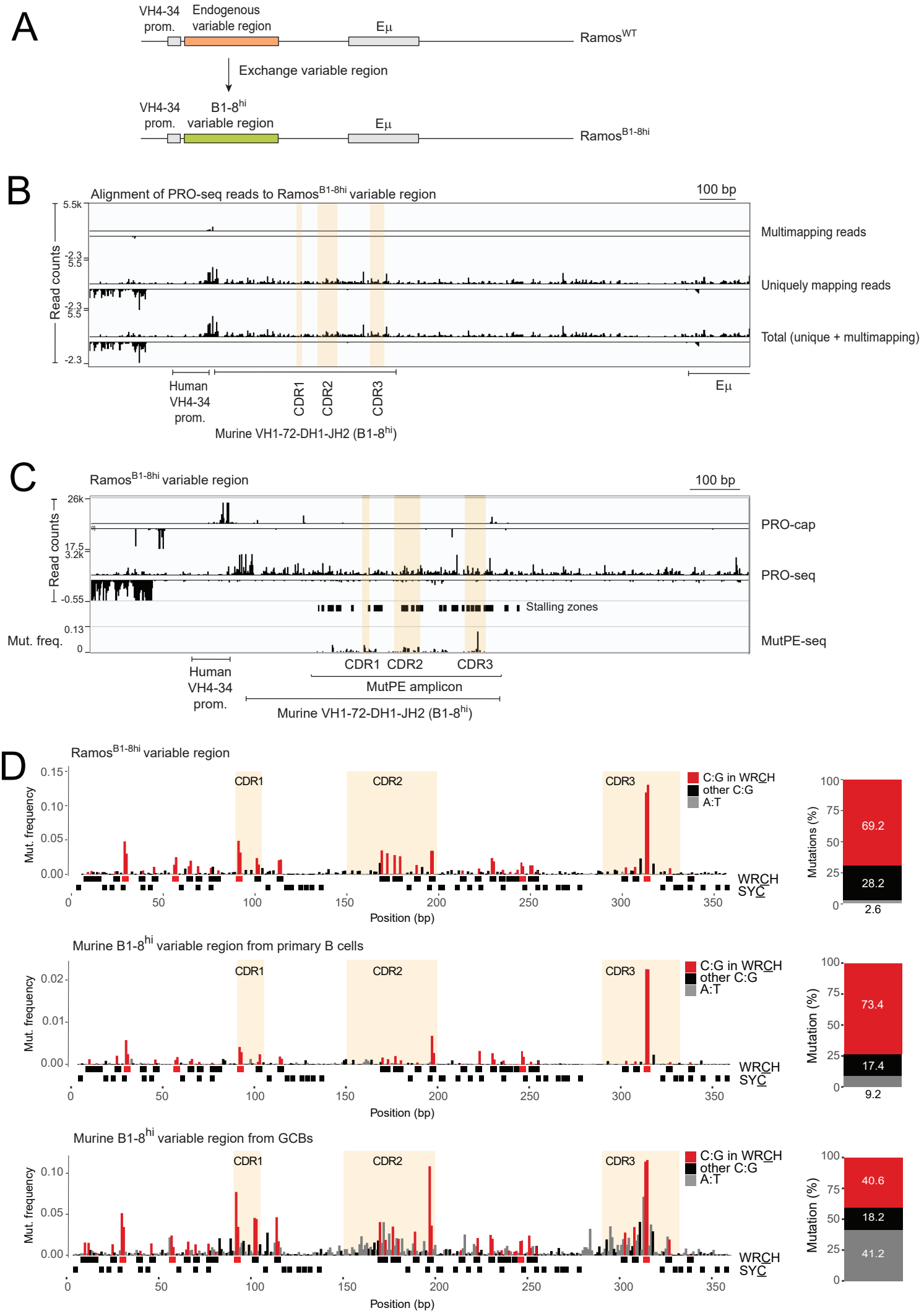
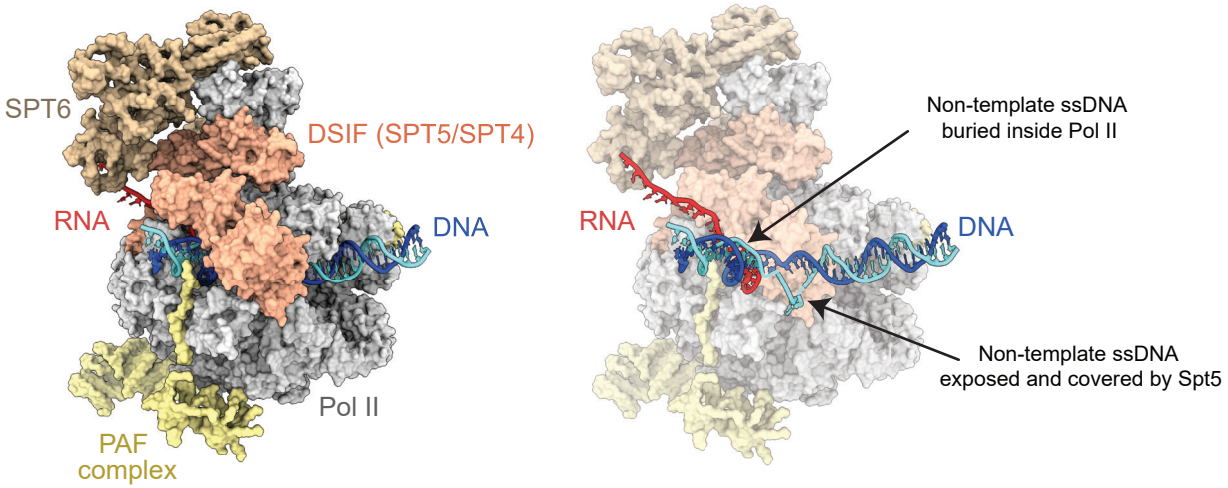

A
VH4-34 prom.
Endogenous variable region
Eμ
Ramos^WT
Exchange variable region
VH4-34 prom.
B1-8^hi variable region
Eμ
Ramos^B1-8hi
B
Alignment of PRO-seq reads to Ramos^B1-8hi variable region
100 bp
Read counts
Multimapping reads
Uniquely mapping reads
Total (unique + multimapping)
Human VH4-34 prom.
CDR1
CDR2
CDR3
Eμ
Murine VH1-72-DH1-JH2 (B1-8^hi)
C
Ramos^B1-8hi variable region
100 bp
Read counts
PRO-cap
PRO-seq
Stalling zones
Mut. freq.
MutPE-seq
Human VH4-34 prom.
CDR1 CDR2 CDR3
MutPE amplicon
Murine VH1-72-DH1-JH2 (B1-8^hi)
D
Ramos^B1-8hi variable region
Mut. frequency
CDR1
CDR2
CDR3
C:G in WRCH
other C:G
A:T
WRCH
SYC
Position (bp)
Mutations (%)
69.2
28.2
2.6
Murine B1-8^hi variable region from primary B cells
Mutation (%)
73.4
17.4
9.2
Murine B1-8^hi variable region from GCBs
40.6
18.2
41.2

Figure 6: Analysis of gene regulatory context on the nascent transcriptional landscape and SHM of the murine B1-8^{hi} V region. (A) Scheme showing the exchange of the endogenous Ramos V region for the murine B1-8^{hi} V 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. The stalling zones are shown in the panel between the PRO-seq and MutPE-seq tracks. **(D)** Details of the mutation patterns of the B1-8^{hi} V 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 profile from primary murine B cells (middle panel) and murine GCBs (bottom panel), exactly as in Fig. 5B-C, are included.

Figure 7

A


SPT6
DSIF (SPT5/SPT4)
RNA
DNA
Pol II
PAF complex
Non-template ssDNA buried inside Pol II
RNA
DNA
Non-template ssDNA exposed and covered by Spt5

B

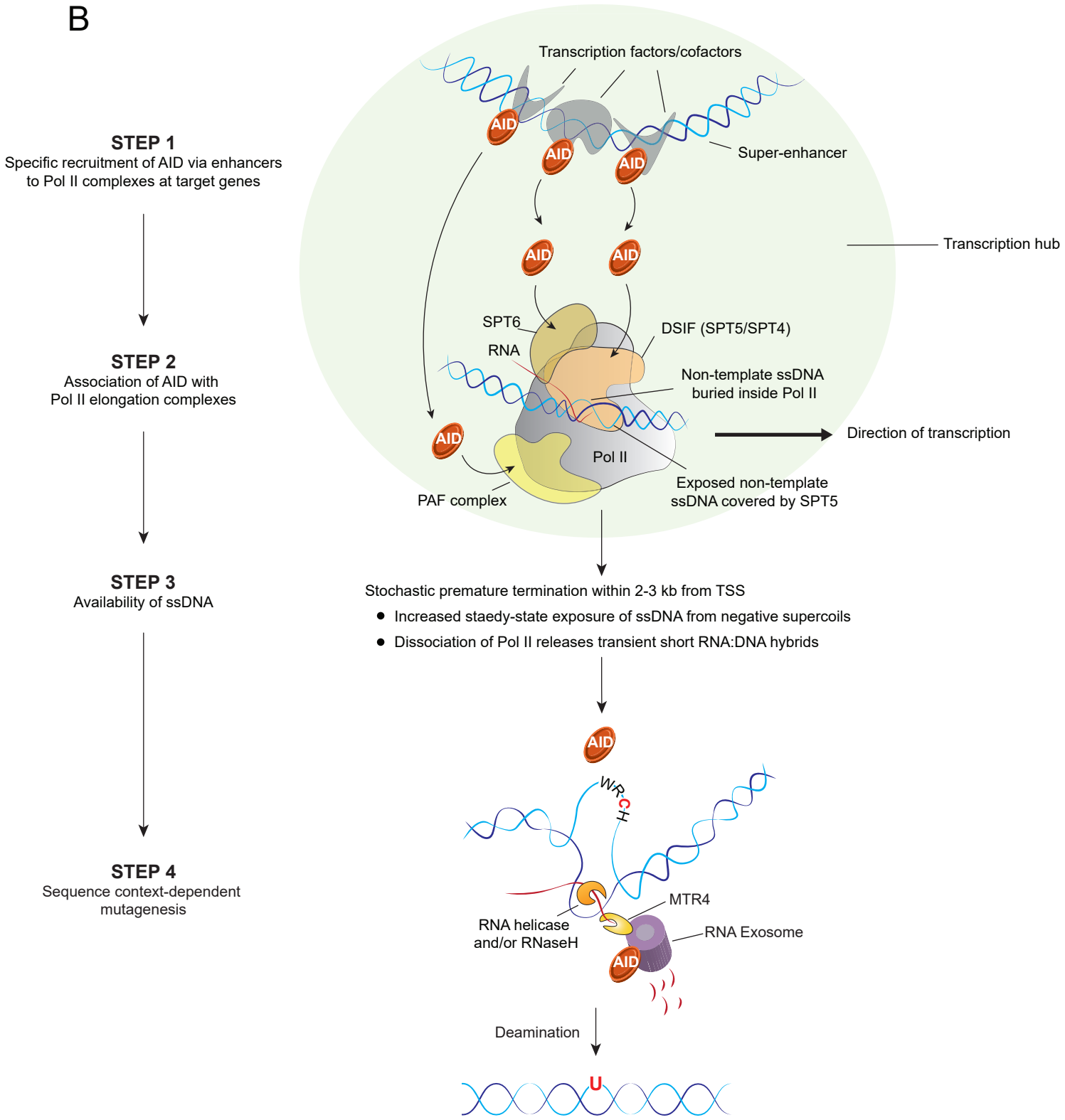

STEP 1
Specific recruitment of AID via enhancers to Pol II complexes at target genes
STEP 2
Association of AID with Pol II elongation complexes
STEP 3
Availability of ssDNA
STEP 4
Sequence context-dependent mutagenesis
Transcription factors/cofactors
AID
Super-enhancer
Transcription hub
SPT6
DSIF (SPT5/SPT4)
RNA
Non-template ssDNA buried inside Pol II
Direction of transcription
Pol II
PAF complex
Exposed non-template ssDNA covered by SPT5
Stochastic premature termination within 2-3 kb from TSS
Increased staedy-state exposure of ssDNA from negative supercoils
Dissociation of Pol II releases transient short RNA:DNA hybrids
WRCH
MTR4
RNA helicase and/or RNaseH
RNA Exosome
Deamination
U

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)** Proposed model for SHM that incorporates various known aspects of SHM biology. 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 (step 1) and is then retained via interactions with elongation factors, SPT5, PAF and SPT6 (step 2). The next step is the availability of ssDNA (step 3). We support the idea that premature transcription termination may be an important source of ssDNA [48]. This event is not associated with stalling but occurs randomly within the first 2-3 kb from the TSSs [106]. A slower-moving terminating Pol II may provide a more stable source of upstream, negative supercoiled DNA favoring higher AID activity. Upon dissociation of Pol II, the non-template strand is likely available as ssDNA for some time 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. In this manner, AID can access ssDNA through various means all in the context of termination.

Figure S1

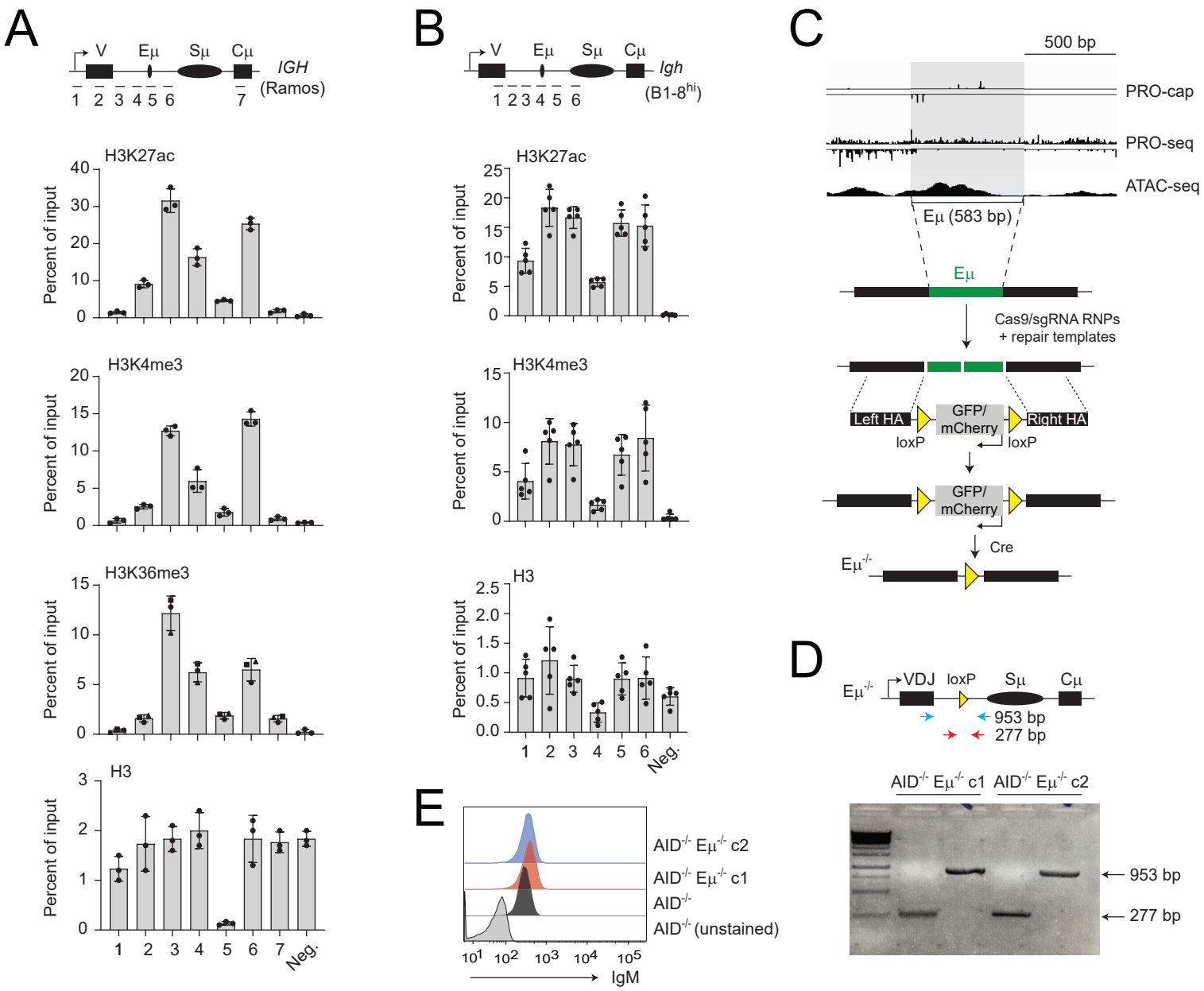


Figure S1: ChIP-qPCR analysis of the Ramos V region locale and generation of Eμ$^{-/-}$ Ramos cells. (A) ChIP-qPCR analysis from three independent experiments in AID$^{-/-}$ Ramos cells for the indicated epigenetic marks as well as histone H3. Amplicons (1-7) used are indicated below the locus diagram 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 five independent experiments in B1-8hi 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 corresponds to a gene desert on chromosome 1 and is used as a negative control. **(C)** Strategy to create the AID$^{-/-}$ Eμ$^{-/-}$ 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 AID$^{-/-}$ Eμ$^{-/-}$ clones c1 and c2. The location of primers is shown in the diagram above the gel image. **(E)** Surface IgM expression in AID$^{-/-}$ Eμ$^{-/-}$ Ramos clones relative to the parental AID$^{-/-}$ line determined by FACS.

Figure S2

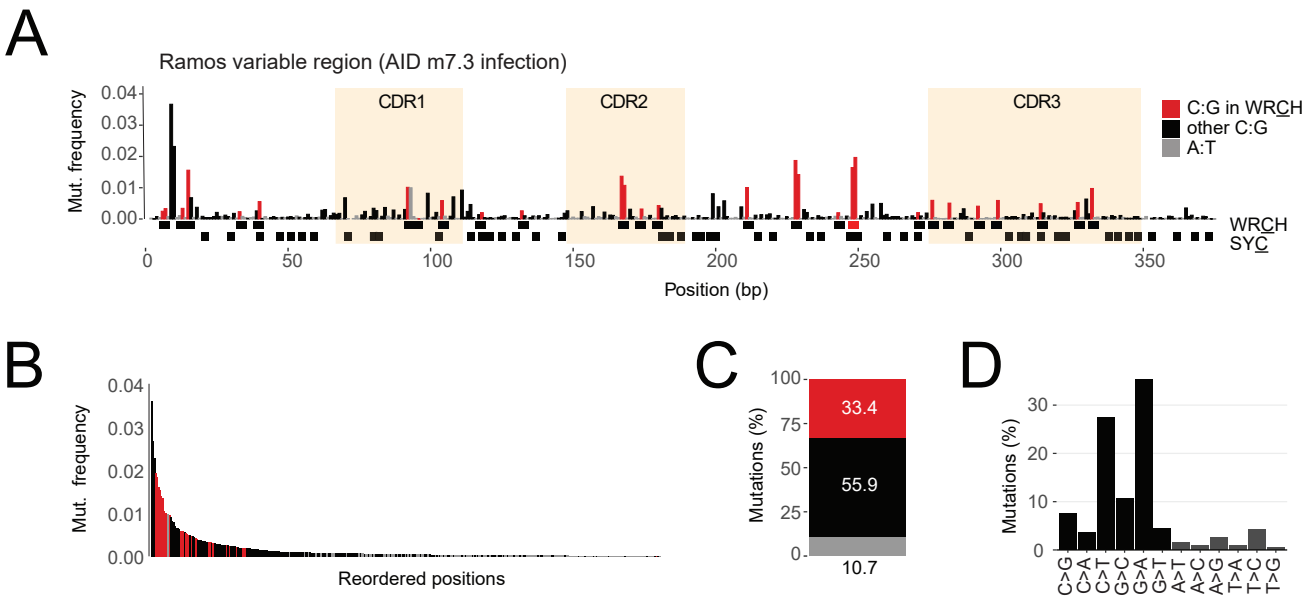


Figure S2: MutPE-seq of the Ramos V region following infection of AID$^{-/-}$ cells with AID m7.3. (A) Mutation profile of the Ramos V region generated by AID (m7.3) following 21 days of infection. The type of mutation and location of hotspots is indicated and can be compared directly with Fig. 3E. **(B—D)** Waterfall plot of mutations ordered by frequency (B), bar plot of mutation frequencies in the three indicated classes (C) and bar graph of mutation type (D) as explained in the legend of Fig. 3 F to H.

Figure S3

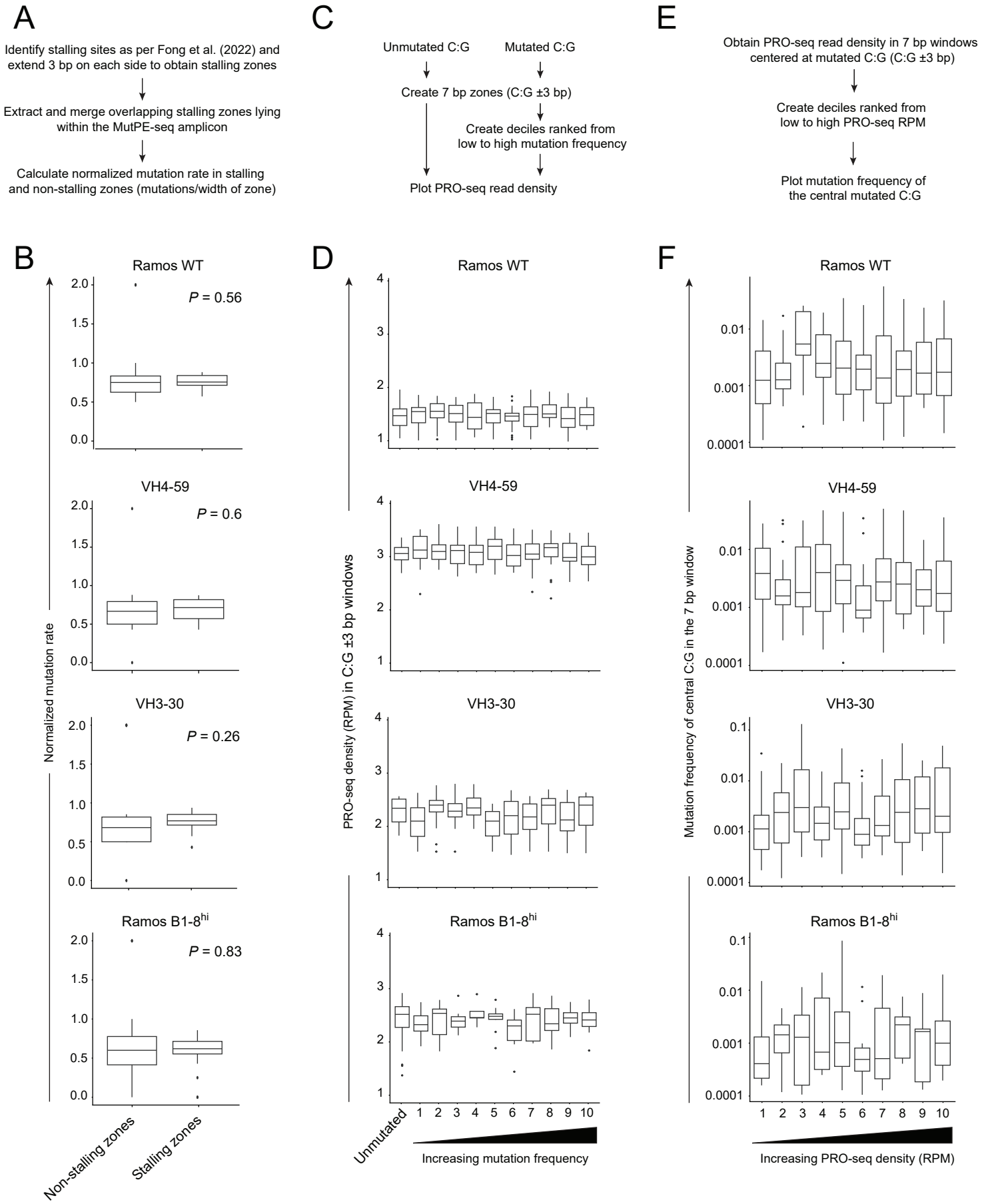

A
Identify stalling sites as per Fong et al. (2022) and extend 3 bp on each side to obtain stalling zones
Extract and merge overlapping stalling zones lying within the MutPE-seq amplicon
Calculate normalized mutation rate in stalling and non-stalling zones (mutations/width of zone)
C
Unmutated C:G
Mutated C:G
Create 7 bp zones (C:G ±3 bp)
Create deciles ranked from low to high mutation frequency
Plot PRO-seq read density
E
Obtain PRO-seq read density in 7 bp windows centered at mutated C:G (C:G ±3 bp)
Create deciles ranked from low to high PRO-seq RPM
Plot mutation frequency of the central mutated C:G
B
Ramos WT
P = 0.56
VH4-59
P = 0.6
VH3-30
P = 0.26
Ramos B1-8hi
P = 0.83
Normalized mutation rate
Non-stalling zones
Stalling zones
D
PRO-seq density (RPM) in C:G ±3 bp windows
Unmutated
Increasing mutation frequency
F
Mutation frequency of central C:G in the 7 bp window
Increasing PRO-seq density (RPM)

Figure S3: Relationship between SHM, Pol II stalling and nascent transcription at different V regions. (A) Scheme for the analysis of mutation frequency between stalling sites and stalling zones within V regions. **(B)** Box plots comparing mutation frequency between stalling and non-stalling zones in the Ramos endogenous (WT) V region, VH4-59, VH3-30 and Ramos$^{B1\text{-}8hi}$. Statistical analyses were performed with a Wilcoxon rank sum test and indicated within the plots. **(C)** Categorization of unmutated and mutated C:G pairs from MutPE-seq data, as described in the text. **(D)** Box plot showing the PRO-seq density in 7 bp windows centered at each C:G residue in the unmutated group and the ten mutated sub-groups, as described in A. The V gene analyzed is indicated at the top of each plot. The medians are indicated by black lines inside the boxes. A Wilcoxon rank sum test after multiple testing correction using the Bonferroni revealed no statistical significance (defined as $P < 0.05$) between any pair of groups within any of the box plots. **(E)** Workflow for the classification of mutated C:G neighborhoods into 7 bp windows (C:G ±3 bp) based on their PRO-seq levels, as described in the text. **(F)** Box plots showing the mutation frequencies of the central C:G residue in the 7 bp windows (C:G ±3 bp) within all ten PRO-seq sub-groups. The V gene analyzed is indicated at the top of each plot. The medians are indicated by black lines inside the boxes. A Wilcoxon rank sum test with the Bonferroni correction revealed no statistical significance (defined as $P < 0.05$) between any pair of groups within any of the box plots.

Figure S4

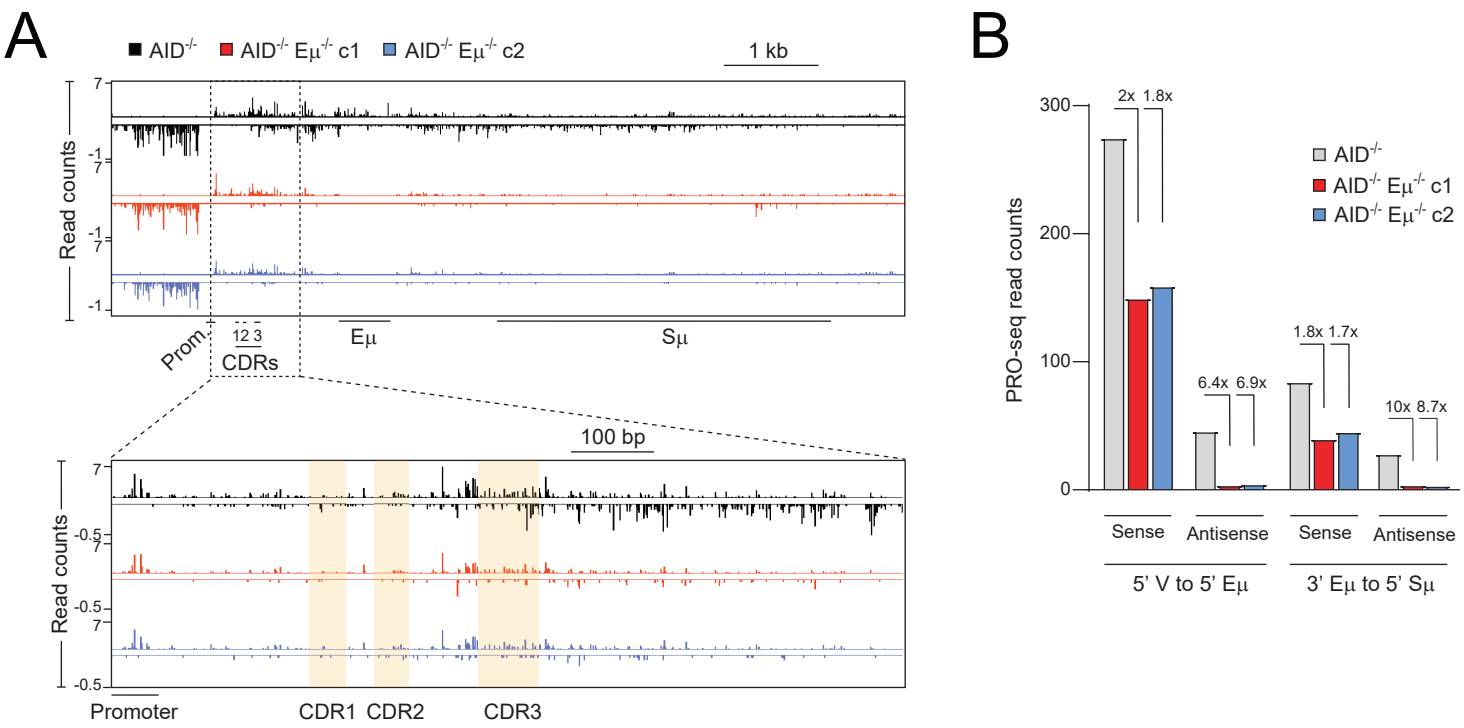
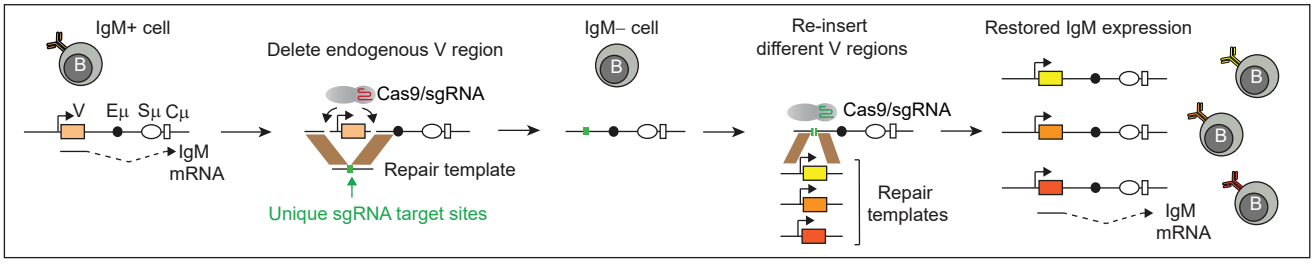


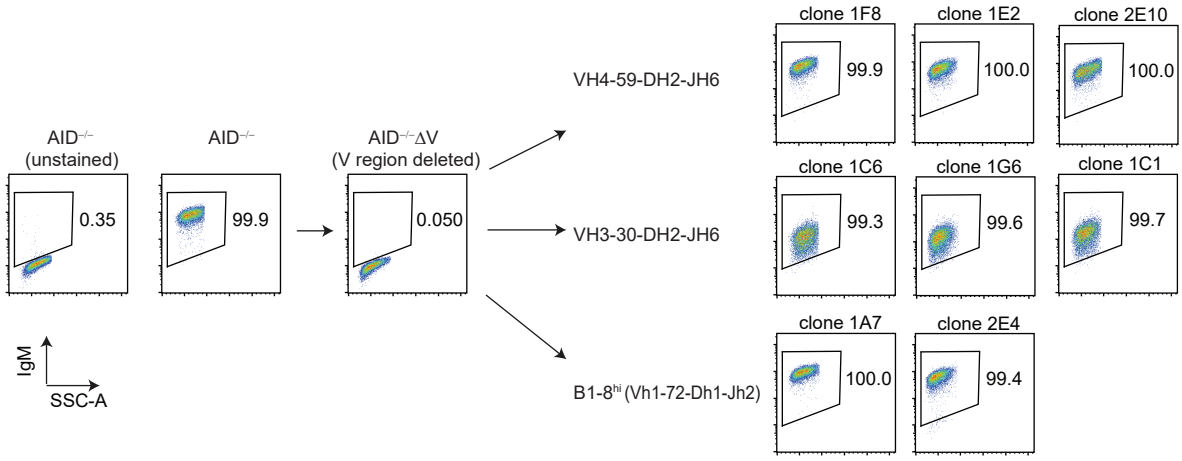
Figure S4. Profiling of the Ramos V region in AID$^{-/-}$ Eμ$^{-/-}$ cells using PRO-seq. (A) Analysis of nascent transcription 3' ends (PRO-seq) at the *IGH* locus in AID$^{-/-}$ and two independent clones of AID$^{-/-}$ Eμ$^{-/-}$ Ramos cells (c1, c2). The promoter (Prom.), complementary determining regions (CDR 1-3), intronic enhancer (Eμ) and switch μ region (Sμ). The lower panel shows a magnified view of the V region locale. **(B)** Quantification of PRO-seq read counts from the data in A. Signals were quantified separately on the sense and antisense strands and further divided into regions upstream and downstream of the deleted Eμ segment. The upstream portion extends from the 5' end of the V region to the 5' end of the deleted Eμ segment. The downstream portion extends from the 3' end of the deleted Eμ segment to the 5' end of Sμ.

Figure S5

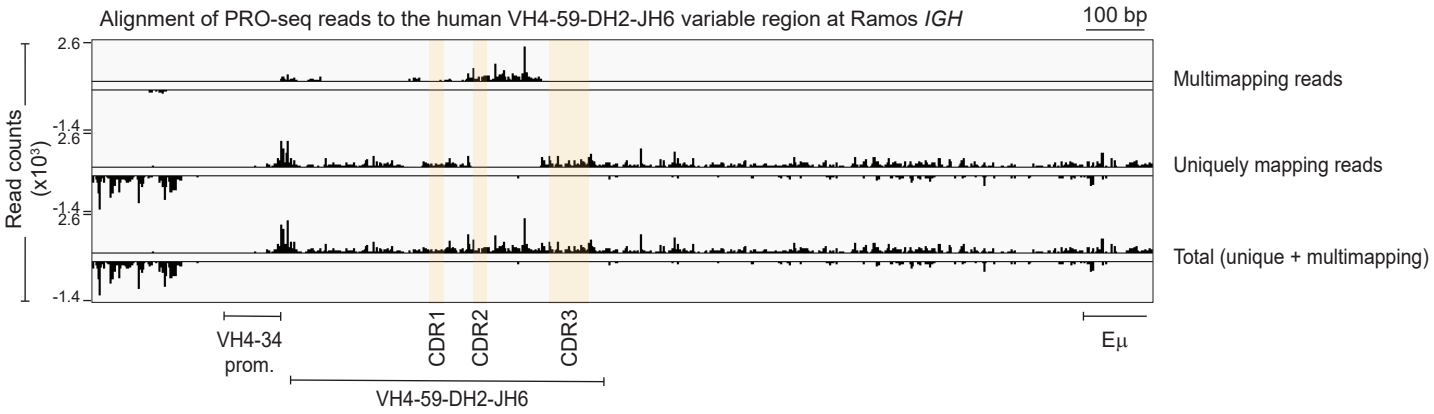
A



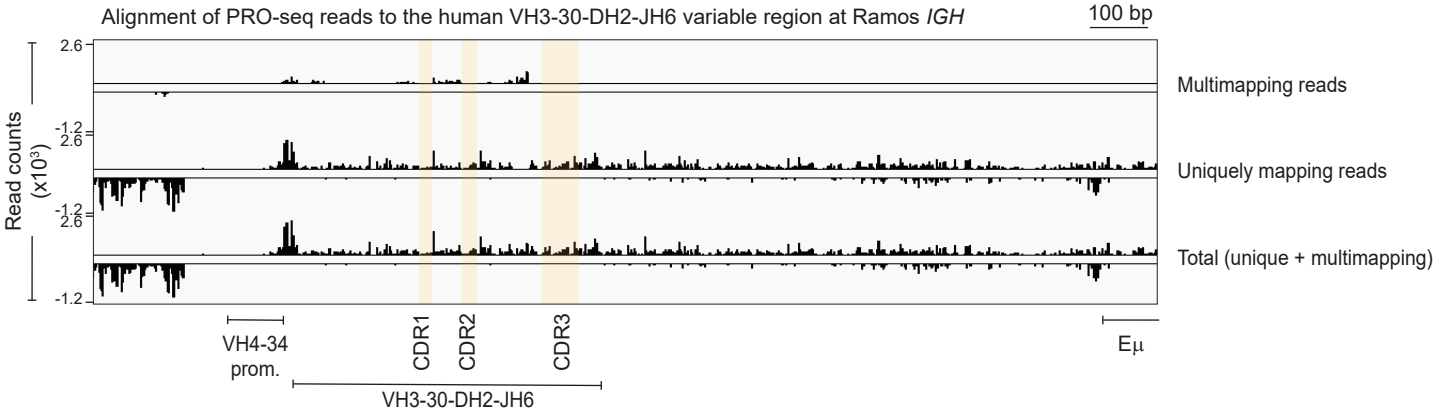
B



C



D



E

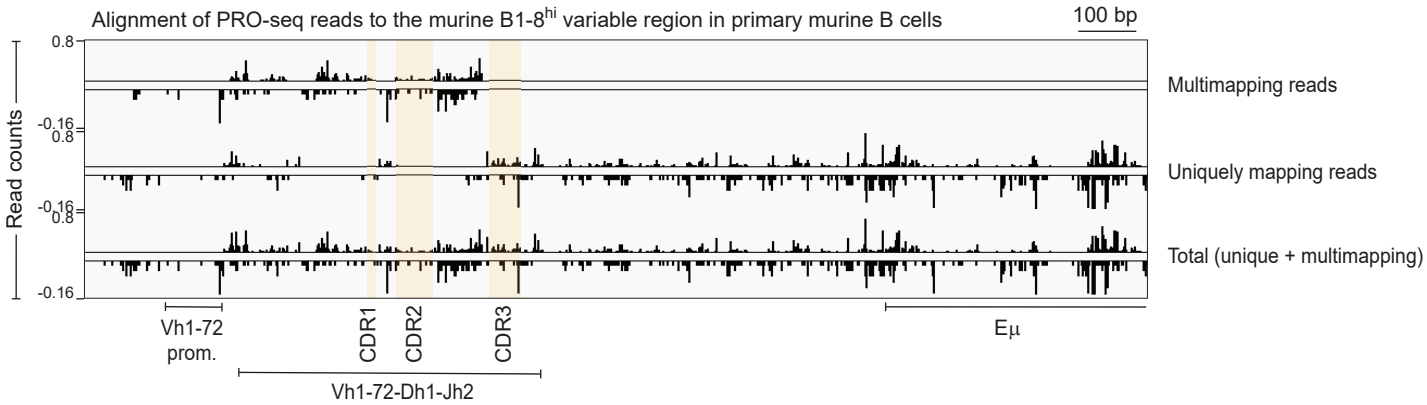
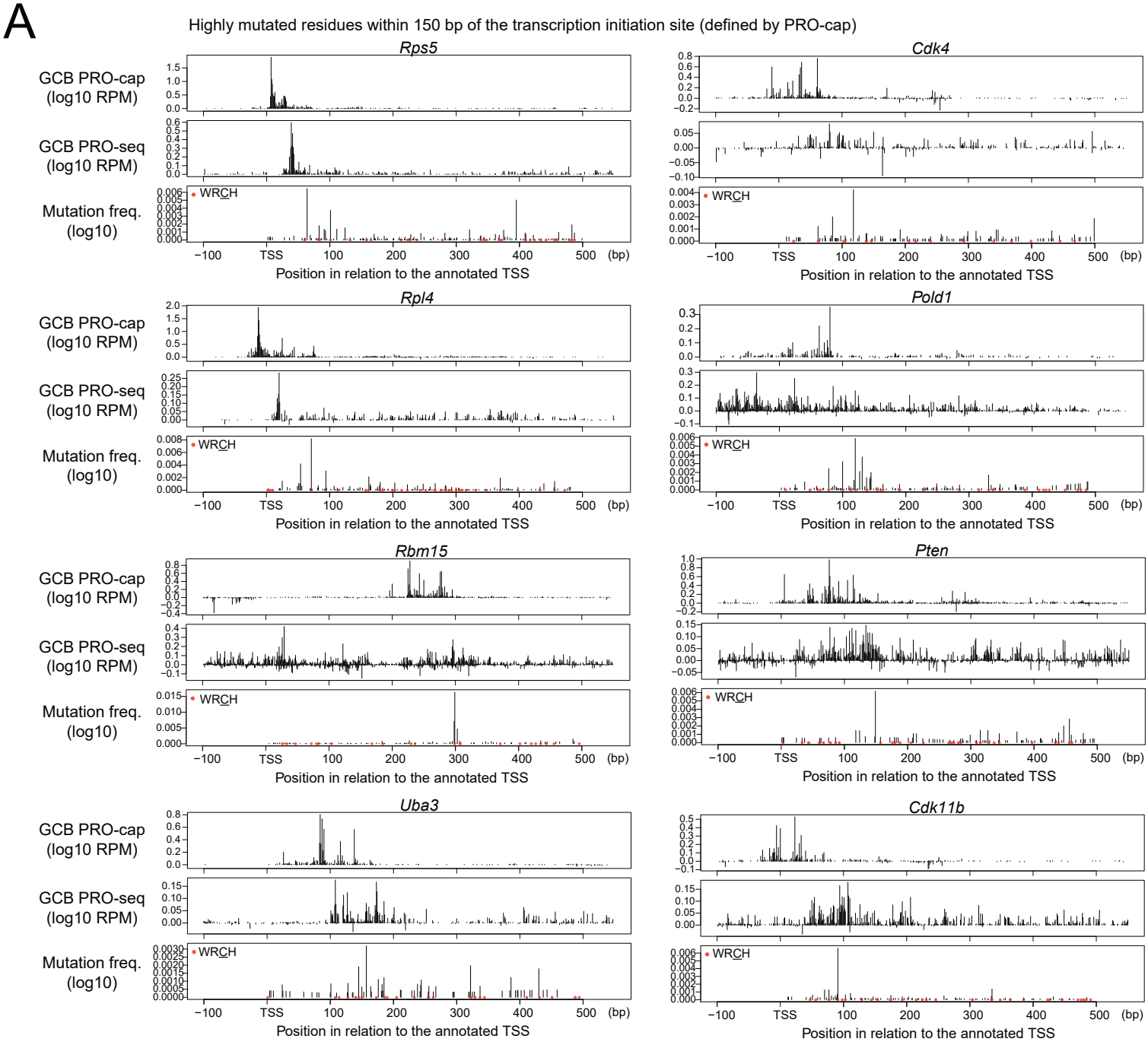


Figure S5: Generation of Ramos cell lines expressing exogenous V regions. (A) Schematic representation of the workflow for generating Ramos cells expressing new V regions using CRISPR-based editing. An IgM-negative line was made by excising the endogenous V region in $AID^{-/-}$ cells ($AID^{-/-}\Delta V$) 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 V regions. Correct integration leads to restoration of surface IgM expression which is used as a readout to identify correctly targeted clones. **(B)** Flow cytometry analysis of surface IgM expression in Ramos cell clones expressing the indicated human and mouse V 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 V region expressed from the VH4-34 promoter at the human *IGH* in Ramos cells. Tracks of Multimapping, uniquely mapping and total are shown (see also Fig. 3A and 3B for a detailed description). **(D)** PRO-seq analysis as in C of the human VH3-30-DH2-JH6 V 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} V region at the *Igh* locus in mice.

Figure S6


A
Highly mutated residues within 150 bp of the transcription initiation site (defined by PRO-cap)
Rps5
Cdk4
Rpl4
Pold1
Rbm15
Pten
Uba3
Cdk11b
GCB PRO-cap (log10 RPM)
GCB PRO-seq (log10 RPM)
Mutation freq. (log10)
WRCH
−100
TSS
100
200
300
400
500
(bp)
Position in relation to the annotated TSS

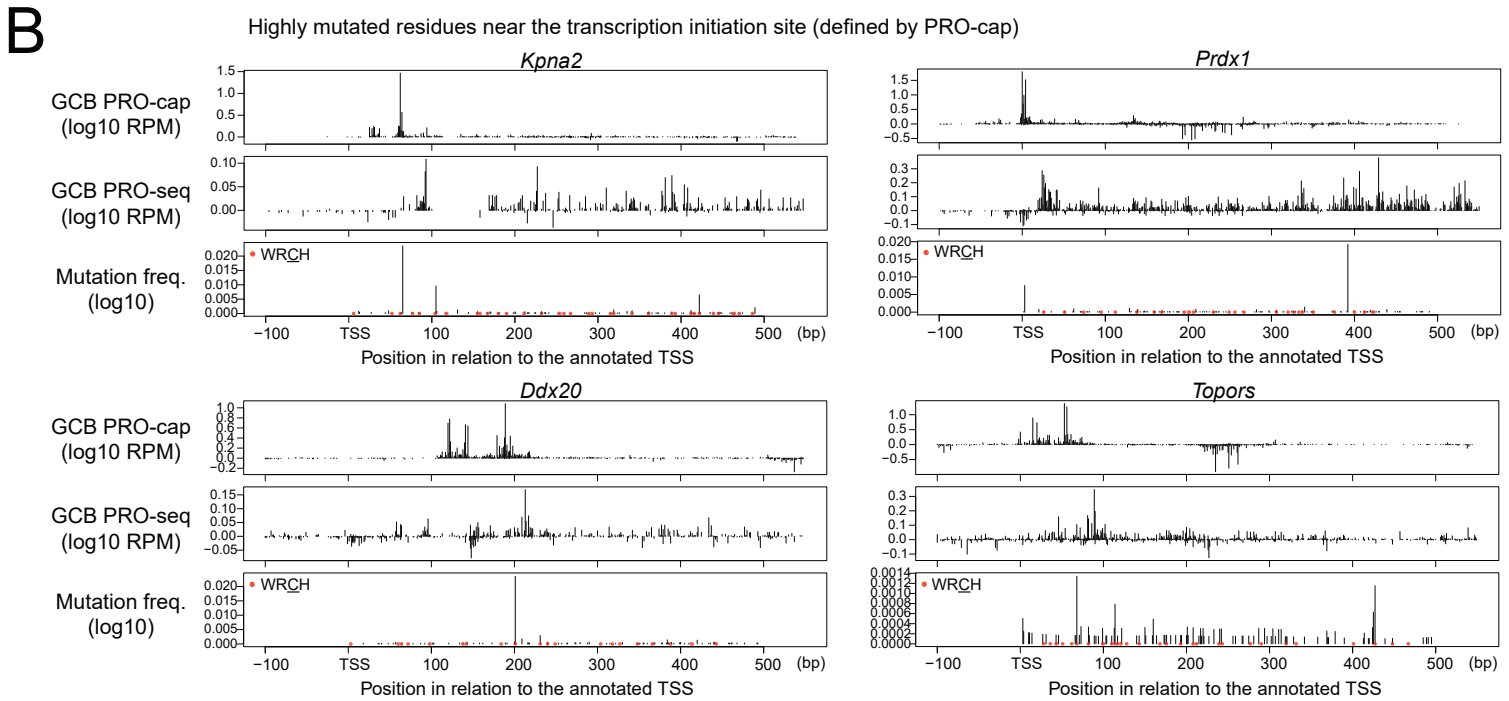

B
Highly mutated residues near the transcription initiation site (defined by PRO-cap)
Kpna2
Prdx1
Ddx20
Topors
GCB PRO-cap (log10 RPM)
GCB PRO-seq (log10 RPM)
Mutation freq. (log10)
WRCH
−100
TSS
100
200
300
400
500
(bp)
Position in relation to the annotated TSS

Figure S6: 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. [17]) 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 S7

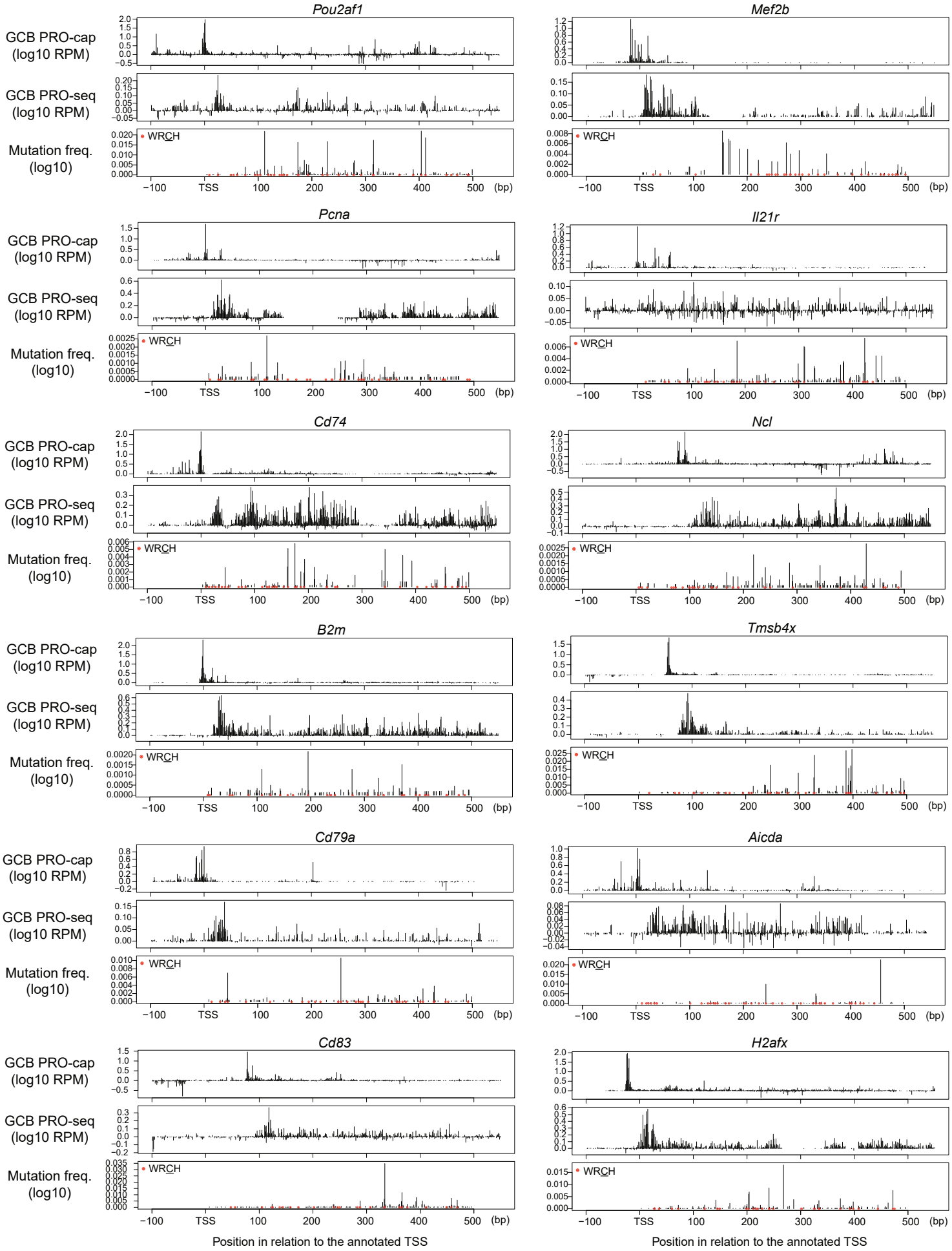

Pou2af1
Mef2b
Pcna
Il21r
Cd74
Ncl
B2m
Tmsb4x
Cd79a
Aicda
Cd83
H2afx
GCB PRO-cap (log10 RPM)
GCB PRO-seq (log10 RPM)
Mutation freq. (log10)
WRCH
−100
TSS
100
200
300
400
500
(bp)
Position in relation to the annotated TSS

Figure S7: Additional examples of PRO-cap, PRO-seq and SHM at non-Ig AID target genes in murine GCBs. See Fig. 5D for detailed legend.

Figure S8

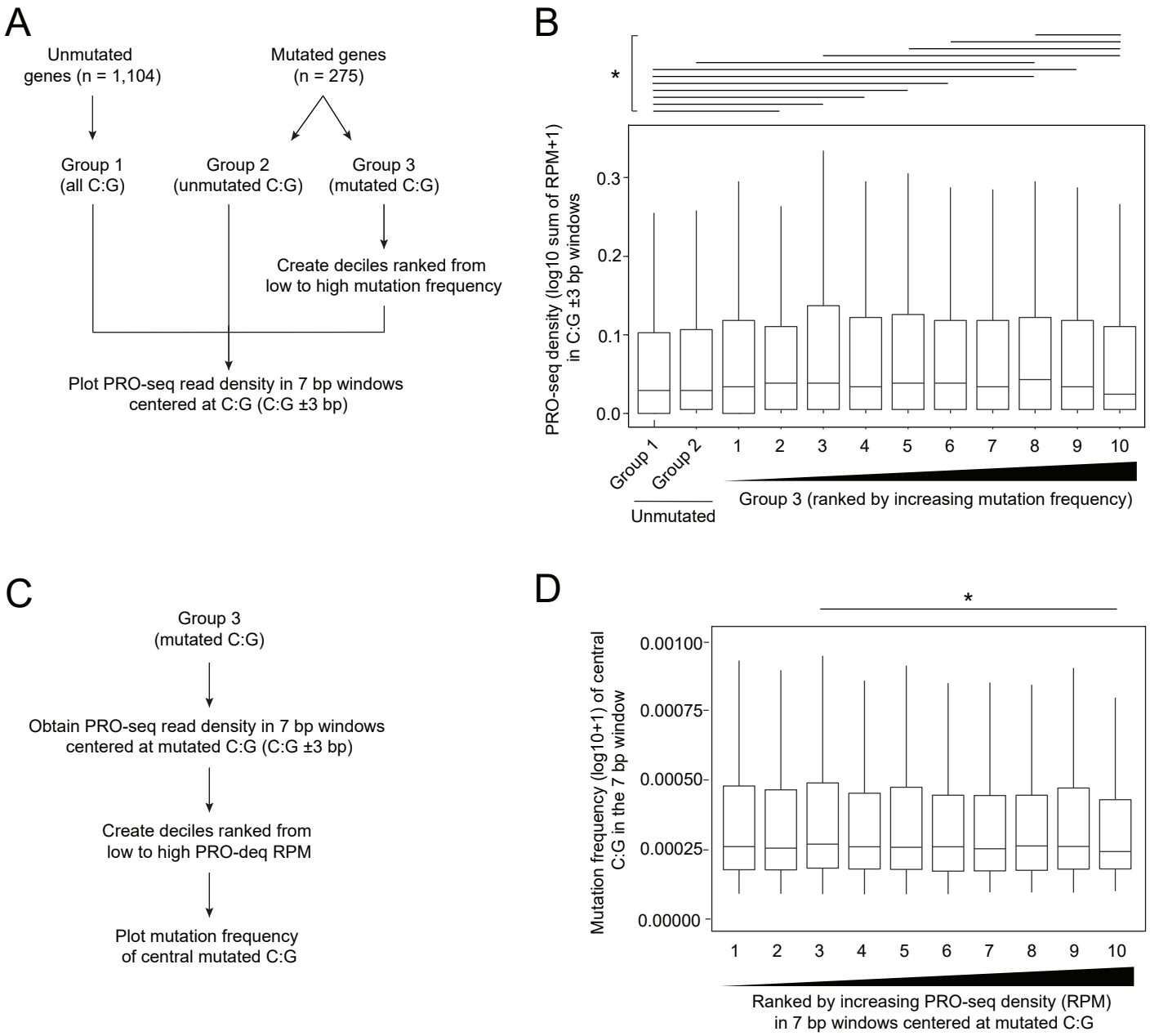


Figure S8. Relationship between SHM and transcriptional strength at non-IG AID target genes. (A) Workflow for the classification of mutated and unmutated genes (from ref. 17). **(B)** Box plot showing the PRO-seq density in 7 bp windows centered at each C:G residue in the three groups and sub-groups described in A. The medians are indicated by black lines inside the boxes. The asterisk indicates $P < 0.05$ based on the Wilcoxon rank sum test after multiple testing correction using the Bonferroni method. **(C)** Scheme for the classification of mutated C:G neighborhoods based on their PRO-seq density. (D) Box plots showing the mutation frequencies of the central C:G residue in the 7 bp windows (C:G ±3 bp) within all ten PRO-seq sub-groups. The medians are indicated by black lines inside the boxes. The asterisk indicates $P < 0.05$ based on the Wilcoxon rank sum test after multiple testing correction using the Bonferroni method.

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

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)