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, 4#]

(1) Research Institute of Molecular Pathology (IMP), Campus-Vienna-Biocenter 1, 1030 Vienna, Austria
(2) Laboratory of Bioinformatics and Computational Biology, A. C. Camargo Cancer Center, São Paulo, SP, Brazil
(3) Laboratory of Humoral Immunology, Institut Pasteur, INSERM U1222, Université Paris Cité, 75015 Paris, France
(4) Peter Gorer Department of Immunobiology, King's College London, London SE1 9RT, United Kingdom

* equal contribution

corresponding author (rushad.pavri@kcl.ac.uk)

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.

Introduction

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

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

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

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

98 Here, we address this question by using precision run-on sequencing (PRO-seq) and the
99 related method, PRO-cap, in combination with a customized bioinformatics pipeline to map reads to V
100 regions. By identifying the precise 3' end of nascent RNA within engaged Pol II complexes, PRO-seq
101 provides a map of the location and orientation of actively transcribing Pol II genome-wide at single
102 nucleotide resolution [64,65]. PRO-cap identifies the 5' ends of capped nascent RNAs that are
103 associated with engaged Pol II molecules and hence provides maps of the location and orientation of
104 transcription initiation sites at single nucleotide resolution [64,66]. For these reasons, PRO-seq and PRO-
105 cap are amongst the most widely used assays for studying nascent transcriptional profiles genome-
106 wide and they have been particularly important in identifying recurrent sites of Pol II stalling and
107 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,

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

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

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

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

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

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

190

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

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

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

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

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

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

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

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

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

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

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

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

274

275

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

465 There was no visible trend of higher PRO-seq density in groups with higher mutation
466 frequency although statistical analyses showed that the PRO-seq densities in group 1 were
467 significantly lower than most, but not all, other groups, in line with the fact that mutation rates are
468 higher in more transcribed genes [17,24] (Fig. S8B). However, the difference in mean PRO-seq values
469 between group 1 and other groups ranged between 1.2-1.9-fold. Based on our results from E$\mu^{-/-}$ cells
470 where >2-fold differences did not have a major impact on SHM (Fig. 2), we infer that the reason for
471 the absence of mutation in group 1 is not solely due to lower transcription.

472 Group 2 showed a significant difference in PRO-seq density only with sub-group 8 within
473 group 3. Moreover, the PRO-seq densities in group 3, sub-group 10, the most mutated group, were
474 actually lower than in many other groups and even showed statistically significant differences in some

475 cases (Fig. S8B). This suggests that, within AID target genes, highly mutated C:G residues are not
476 generally located in genomic neighborhoods harboring higher levels of transcription and Pol II
477 occupancy.

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

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

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

499 PRO-seq and PRO-cap revealed that the nascent transcriptional profile of the B1-8^{hi}
500 sequence in Ramos$^{B1\text{-}8hi}$ cells (Fig. 6C) shared similarities with the B1-8^{hi} sequence in the murine *Igh*
501 context (Fig. 5A). For example, similarities were noted in CDR3 and flanking sequences, for instance,
502 the presence of the antisense initiation site (PRO-cap track) and shared spikes of nascent
503 transcriptional activity (PRO-seq) (Fig. 6C). More importantly, MutPE-seq following AID (JP8Bdel)

504 infection revealed a very similar SHM pattern in Ramos$^{B1\text{-}8hi}$ cells compared to the murine B1-8^{hi}
505 context, especially with primary murine B cells (Fig. 6D). Of note, the dominant, palindromic AGCT
506 hotspot in CDR3 was the most mutated residue exactly as in B1-8^{hi} murine primary cells and GCBs
507 (Fig. 6D). The other mutated C:G residues in the Ramos$^{B1\text{-}8hi}$ sequence were largely the same as
508 those mutated in the murine context albeit the relative mutation rates within each V region varied (Fig.
509 6D, compare with Fig. 5 B and C). Stalled Pol II zones within the Ramos$^{B1\text{-}8hi}$ V region bore no
510 significant relationship with mutation frequency compared to non-stalled zones (Fig. S3A-B; P = 0.83).
511 Analysis of PRO-seq density in the neighborhood of mutated and unmutated C:G pairs revealed a
512 lack of statistical significance between mutation frequency and local transcriptional strength (Fig.
513 S3C-F). Thus, as in the case of the other V regions analyzed here, there is no significant relationship
514 between Pol II stalling, Pol II occupancy and local SHM at the Ramos$^{B1\text{-}8hi}$ V region.

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

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

526

527 **Discussion**

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

530 Our analysis of the epigenetic landscape showed that major decreases in activating
531 chromatin marks and nascent transcription, via deletion of the Eμ enhancer, do not have a significant

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

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

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

557 The consistent lack of correlation between transcriptional features and SHM reported here begs the
558 question of what role transcription plays in SHM. SHM depends on the availability of ssDNA, but the
559 source(s) of accessible ssDNA for AID remains a major unresolved question. It has been proposed
560 that ssDNA within transcribing Pol II complexes could be a substrate for AID, especially when Pol II is

561 paused or stalled, an event that would lead to increased ssDNA exposure [17,27,52-56]. However, we
562 found no significant correlation between SHM frequency and Pol II stalling zones or Pol II
563 accumulation, which prompted us to take a closer look at the structures of paused and elongating Pol
564 II complexes. This analysis showed that of the ~11 nucleotide non-template ssDNA in the structure,
565 about half lies buried within Pol II [88,89] (Fig. 7A). Importantly, the exposed portion (~5 nt) is entirely
566 covered by SPT5 [89] (Fig. 7A). The PAF complex and SPT6 are located further away from this site [88]
567 (Fig. 7A). Under such structural constraints, it is difficult to conceive how AID could access the ~5 nt
568 exposed ssDNA and catalyze deamination even if Pol II were paused or stalled. Thus, it appears
569 unlikely that the transcription bubble within an actively elongating Pol II holoenzyme, including a
570 stalled complex, serves as an optimal substrate for AID.

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

582 Another key insight is the fact that significant mutational asymmetry is observed upon
583 depletion of RNA Exosome subunits or the DNA:RNA helicases, Mtr4, Senataxin and others [103-105],
584 which implies that one of the major species targeted by AID is an RNA:DNA hybrid. Indeed, the V
585 region sense transcript is a substrate of the RNA exosome and the latter is required for optimal
586 mutation of the template strand in the B1-8^{hi} V region, findings that directly implicate the processing of
587 DNA:RNA hybrids, or short transient R loops, in directing the patterns of SHM [103].

588 Nevertheless, because all the above-mentioned features are continuously associated with
589 transcription, they cannot, by themselves, explain the distinct and highly reproducible SHM profiles of

590 V regions, nor can they explain why SHM spectra are restricted to a ~2 kb window starting from the V
591 gene promoter [3,62,63]. To reconcile these observations, we support a model centered on premature
592 transcription termination, which, as proposed below, can define the boundaries of SHM and also
593 provide a source of ssDNA, the latter having been suggested previously in the context of RNA
594 exosome-mediated AID targeting [49,50] as well as SHM on plasmid substrates and in the Ramos V
595 region [47,48] (Fig. 7B).

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

607 Another attractive aspect of premature termination in the context of SHM is that it can
608 potentially provide multiple avenues for AID to access ssDNA [47-50]. Firstly, the speed of a terminating
609 Pol II is expected to be lower than a rapidly elongating one, which would increase the steady-state
610 availability of upstream, negatively supercoiled DNA for AID to access. Secondly, although the
611 structural details of how Pol II terminates and dissociates off DNA have not been resolved, it is
612 conceptually plausible that the collapse of the elongation complex could release ssDNA on the non-
613 template strand while the nascent RNA is still hybridized to the template strand, which provides
614 another source of ssDNA (Fig. 7B, step 4). Finally, the RNA would have to be unwound by DNA:RNA
615 helicases [105] and/or digested with RNaseH followed by degradation from the 5' end by Xrn2 and from
616 the 3' end by the RNA exosome, thereby providing AID access to the template strand [31,103-105,111] (Fig.
617 7B, step 4). Given that IG loci are amongst the most highly expressed in activated B cells, the
618 combined effect of having multiple rounds of transcription per cell cycle coupled with stochastic

619 termination within the first 2-3 kb from V gene promoters can plausibly explain the empirically
620 observed boundary of SHM, that is, ~2 kb from V gene promoters [3,62,63]. The model can also explain
621 why, at some non-IG SHM target loci, we observed mutations overlapping with, or in close proximity
622 to, the transcription initiation zone. Most importantly, such stochastic events will not be detected by
623 standard population-based approaches, which can explain the lack of correlation between local SHM
624 frequency and any transcriptional feature, be it Pol II stalling, internal initiation, convergent
625 transcription or ssDNA patches.

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

641

642 **Methods**

643 **Cell culturing**

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

647 DMEM medium (in-house) containing 10% fetal bovine serum (FBS; Invitrogen), glutamine
648 (Invitrogen), sodium pyruvate (Invitrogen), HEPES (made in-house) and antibiotic/antimycotic
649 (Invitrogen).

650 **Mice**

651 The mice were maintained in a C57BL/6 background and housed in the IMBA-IMP animal facility in
652 standard IVC cages with HEPA filtering. B1-8^{hi} Rosa26 ^{AIDER} mice were generated by crossing B1-8^{hi}
653 mice [80] with *Rosa26^{AIDER}* mice [90] and maintained as a homozygous line for both alleles. All animal
654 experiments were carried out with valid breeding and experimental licenses (GZ: MA58-320337-2019-
655 9, GZ: 925665/2013/20 and GZ: 618046/2018/14) obtained from the Austrian Veterinary Authorities
656 and in compliance with IMP-IMBA animal house regulations.

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

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

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

668 To delete the 583 bp segment containing Eμ, we generated homology repair plasmids having Ef1a
669 promoter-driven floxed GFP and mCherry expression cassettes flanked by homology arms. These
670 plasmids were transfected into AID$^{-/-}$ Ramos cells (clone D3) along with three in vitro synthesized
671 guide RNAs (6 μg each, designed using CRISPOR) and 7.5 μg recombinant Cas9 protein (Vienna
672 Biocenter Core Facilities) using the Neon Transfection System (Thermo Fisher Scientific). A week
673 later, GFP/mCherry double-positive single cells were isolated using a BD FACS Aria III sorter (BD

674 Biosciences). Successful knock-in clones were identified by genotyping with PCR and Sanger
675 sequencing of PCR products. Next, 200 μg recombinant Cre recombinase (Molecular Biology Service,
676 IMP) was added to the culture medium to excise the floxed mCherry/GFP cassettes. GFP/mCherry
677 double-negative clones were isolated one week after electroporation using a BD FACS Aria III sorter
678 and genotyped via PCR and Sanger sequencing of PCR products.

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

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

689 **Lentiviral infections**

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

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

696 Mature, naïve B cells were isolated from spleens of 2-4-month-old B1-8^{hi} Rosa26 ^{AIDER} mice, as per
697 established protocols [27]. B cells were cultured in complete RPMI medium supplemented with 10%
698 fetal bovine serum (FBS) and antibiotics, Interleukin 4 (IL4; made in-house by the Molecular Biology
699 Service, IMP), 25 μg/ml Lipopolysaccharide (Sigma) and RP105 (made in-house by the Molecular
700 Biology Service, IMP) and harvested after 3 or 4 days. For SHM assays from activated B1-8^{hi} *Rosa26*

701 AIDER primary B cells, 2μM 4-hydroxy tamoxifen (4-HT) (Sigma) was added at the time of activation
702 with IL-4, LPS and RP105.

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

704 Sheep red blood cells (SRBCs) were washed thrice with PBS of which 0.2x10^9 SRBCs in 100µl PBS
705 (per mouse) were injected into B1-8^{hi} mice followed by another injection of 1x10^9 SRBCs in 100µl PBS
706 five days later. Mice were harvested 10 days after the first immunization. Spleens were harvested and
707 single-cell suspensions were stained with B220 conjugated to fluorescein isothiocyanate (B220-FITC,
708 BD Biosciences, 1:500 dilution), Fas conjugated to PE-cyanine 7 (Fas-PE-Cy7, BD Biosciences,
709 1:1000 dilution) and CD38 conjugated to Allophycocyanin (CD38-APC, ThermoFisher, 1:200 dilution)
710 along with Fc block (BD Biosciences, 1:500 dilution). GCBs (B220$^+$ Fas$^+$ CD38$^-$) were isolated on a
711 BD FACS Aria III sorter (BD Biosciences). Following sorting, nuclei were isolated and used for PRO-
712 seq and PRO-cap (see below). For one experiment, ~10^7 GCBs from fifteen immunized mice were
713 pooled yielding 3-4x10^6 viable nuclei that were used for run-on. For MutPE-seq, genomic DNA was
714 isolated from 15 x10^6 GCBs as described previously [117].

715 [118]previously [64] with several modifications.
716 To isolate nuclei, murine and *Drosphila* S2 cells were resuspended in cold Buffer IA (160 mM
717 Sucrose, 10 mM Tris-Cl pH 8, 3 mM $CaCl_2$, 2 mM $MgAc_2$, 0.5% NP-40, 1 mM DTT added fresh),
718 incubated on ice for 3 min and centrifuged at 700 g for 5 min. The pellet was resuspended in nuclei
719 resuspension buffer NRB (50 mM Tris-Cl pH 8, 40% Glycerol, 5 mM $MgCl_2$, 0.1 mM EDTA). For each
720 run-on, 10^7 nuclei of the sample and 10% *Drosophila* S2 nuclei were combined in a total of 100 µL
721 NRB and incubated at 30°C for 3 min with 100 µL 2x NRO buffer including 5µl of each 1mM Bio-11-
722 NTPs. In some PRO-cap experiments, the run-on reaction was performed in the presence of two
723 biotinylated NTPs, Biotin-11-UTP and Biotin-11-CTP (Perkin-Elmer), and unlabeled ATP and GTP.
724 Subsequent steps were performed as described (Mahat et al 2016), except that 3' and 5' ligations
725 were performed at 16°C overnight and CapClip Pyrophosphatase (Biozym Scientific) was used for
726 5'end decapping. We also used customized adapters (PRO-seq: 3' RNA adapter:
727 5'5Phos/NNNNNNNGAUCGUCGGACUGUAGAACUCUGAAC/3InvdT-3' and 5' RNA adapter: 5'-
728 CCUUGGCACCCGAGAAUUCCANNNN-3'). RNA was reverse transcribed by SuperScript III RT with
729 RP1 Illumina primer to generate cDNA libraries. Libraries were amplified with barcoding Illumina RPI-x

730 primers and the universal forward primer RP1 using KAPA HiFi Real-Time PCR Library Amplification
731 Kit. PRO-cap: 3' linker (DNA oligo): 5rApp/NNNNAGATCGGAAGAGCACACGTCT/3ddC , 5'linker
732 (RNA): ACACUCUUUCCCUACACGACGCUCUUCCGAUCUNNNNNNNNNN, reverse transcription
733 and library amplification as in PRO-seq but using selected TruSeq_IDX_1-48 PCR primers for RT,
734 and the same TruSeq IDX PCR primers together with TruSeq Universal Adapter forward primer for
735 PCR. For both methods amplified libraries were subjected to gel electrophoresis on 2.5% low melting
736 agarose gel and amplicons from 150 to 350 bp were extracted from the gel (carefully separated from
737 possible linker dimers), multiplexed and sequenced on Illumina platforms with SR50 or longer read
738 mode.

739 **RT-qPCR**

740 RT-qPCR assays with externally spiked-in *Drosophila* S2 cells were described in detail in our previous
741 studies [119]. Briefly, *Drosophila* S2 cells were mixed with Ramos cells or B1-8^{hi} B cells at a ratio of 1:4
742 ($4x10^5$ B cells and $1x10^5$ S2 cells) followed by total RNA extraction with TRIzol reagent (Thermo
743 Fisher), DNaseI digestion, and cDNA synthesis with random primers. The 2-ΔΔCt method was used to
744 quantify the data using the *Drosophila Act5c* transcript for normalization.

745 **ChIP-qPCR and ChIP-seq**

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

748 **ATAC-seq**

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

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

751 MutPE-seq was performed following the principles described in two previous reports [90,95] with several
752 adaptations. 80 ng of genomic DNA were amplified by PCR with the Kapa Hifi HS 2x RM (Roche
753 Diagnostics), For the first PCR, we used 20-25 cycles with 0.2µM locus-specific primers fused to a
754 varying number of random nucleotides and to the first part of the Illumina adapter sequences (FW: 5'-
755 CTCTTTCCCTACACGACGCTCTTCCGATCT-(N)x-gene-specific sequence-3'; RV: 5'-

756 CTGGAGTTCAGACGTGTGCTCTTCCGATCT-gene-specific sequence-3'; see Table S1 for a
757 complete list of primer sequences).

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

770 **Primers and sgRNAs**

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

772

773 **Bioinformatics**

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

775 Reads were trimmed for standard adapters and low quality (Q<30) 3'-end bases and filtered for a
776 remaining length of at least 20nt (excluding the UMI, if applicable) using cutadapt [120]. Alignment to the
777 reference genome was done with Bowtie [121]. The NCBI GRCm38.6 assembly was used as the mouse
778 reference. To this we added the sequence of the recombined locus containing the V region as an
779 additional chromosome. Similarly, for human data, the Hg38 assembly was used, with the relevant V
780 region sequence added as a chromosome. Where applicable, the dmr6 assembly of *Drosophila*
781 *melanogaster* from Flybase ([122] release 6.27) was used as the spike-in reference and was added to
782 the alignment index. During alignment, up to three mismatches were allowed. To accommodate

783 mapping to the repetitive *IGH/Igh* locus, a high degree of multimapping was allowed (194 potential V
784 segments annotated for GRCm38.6). In the case of alignment with spike-ins, only reads mapping
785 exclusively to either genome were considered. For PCR deduplication, UMIs were identified and
786 filtered with UMI-tools ([123] v1.0.0). No sequence differences were allowed in UMIs when collapsing the
787 duplicates. Multimapping reads were then filtered to identify those specific to the V region sequence,
788 taking into account the repetitiveness of the reference *IGH/Igh* locus. Reads mapping to the fixed V
789 region sequence were allowed to multimap only to the native *IGH/Igh* locus (mouse
790 chr12:113572929-116009954 or human chr14:105836764-106875071). The native loci were defined
791 as the region between the earliest position belonging to an annotated V segment and the latest
792 position belonging to an annotated J segment. Reads with a mapped location outside these areas
793 were rejected. This filtering was implemented in a custom Python script. The qualifying reads were
794 then classified as reads mapping uniquely to the particular V region sequence and reads mapping to
795 both the V region sequence and to the native *IGH/Igh* region. Genome browser tracks were created
796 by quantifying and scaling read coverage of the V region sequence using bedtools ([124] v2-29.0).
797 Reads were split by strand, and strand labels for PRO-seq were inverted in order to match the strand
798 designations of the respective PRO-cap reads. For both data types, only the first 5' base of the
799 strand-adjusted reads was used for the tracks. Additionally, the 5' end of PRO-seq reads was shifted
800 by 1 nucleotide downstream, to compensate for the fact that the last nucleotide in the RNA was
801 incorporated during the run-on reaction. For normalized tracks, read counts were scaled to RPM
802 (reads per million). Finally, "-" strand coverage tracks were further scaled by -1, for visualization
803 purposes.

804 **Analysis of MutPE-seq**

805 Reads were trimmed for standard adapters with cutadapt [120]. Poor quality (Q<25) 3' bases were
806 trimmed with trimmomatic [125] by averaging over a sliding window of 5nt. Read pairs were then filtered
807 for minimum remaining length (200nt for read 1, 100nt for read 2) using cutadapt. Read mates were
808 merged down to make combined single-end reads with FLASH [126] allowing 10% mismatch between
809 the mates. Obvious erroneous mergers were removed by selecting combined reads with lengths
810 within ±30nt of the amplicon length using cutadapt, The remaining combined reads were aligned with
811 Bowtie2 [127], using the "–very-sensitive-local" alignment mode and only the fixed V region sequence

812 and its immediate vicinity as reference. Samples of V genes VH4-59 and VH4-34 are different only by
813 single nucleotide polymorphisms and were filtered for contamination by the respective second
814 sequence by jointly aligning with the “expected” and the other “contaminating” V region sequence,
815 discarding all aligned contaminating reads. A pile-up was generated with samtools [128] taking into
816 account only bases with quality of at least 30. The pileups were then quantified with a custom Python
817 script and the resulting mutation counts were processed and visualized with custom scripts in R
818 (v3.5.1), with the help of additional R packages (data.table [https://cran.r-
819 project.org/web/packages/data.table/index.html], ggplot2 [https://cran.r-
820 project.org/web/packages/ggplot2/index.html] , ggrepel [https://cran.r-
821 project.org/web/packages/ggrepel/index.html], patchwork [https://cran.r-
822 project.org/web/packages/patchwork/index.html]). Background mutation profiles were controlled for by
823 subtracting the corresponding mutation frequencies in control samples from the frequencies in the
824 samples of interest, at each position and for each substitution type. Annotation of hot and cold spots
825 was created by means of regex search for the corresponding patterns in the reference sequence.

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

827 Stalling sites were identified as described in Fong et al. [91] using the source code provided therein
828 (https://doi.org/10.5281/zenodo.7013905) with no changes. We used the same parameters as Fong et
829 al. to define stalling sites, namely, all sites covered by at least five PRO-seq reads and with a
830 standard deviation three times greater than the surrounding 200 bp window [91]. In addition, we
831 extended these sites by 3 bp on either side to create 7 bp stalling zones. Only stalling zones within
832 the MutPE-seq amplicon were considered and overlapping zones were merged. All other zones within
833 the MutPE-seq amplicon were considered as non-stalling zones. This approach yielded 18-21 stalling
834 zones and 17-19 non-stalling zones depending on the V region analyzed, which provided sufficient
835 datapoints for statistical analyses. Mutations in all zones were counted and divided by the width of the
836 zone to yield the normalized mutation rates which were displayed as box plots.

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

838 To analyze the relationship between PRO-seq read density and mutation frequency at C:G pairs (Fig.
839 S3C-D), we first separated mutated and unmutated C:G residues into two groups. All C:G residues
840 were extended by 3 bp on either side to create 7 bp windows. For the mutated C:G group,

overlapping windows were merged. If a mutated window overlapped an unmutated one, the mutated window was retained and the unmutated window was discarded from the analysis. In this manner, only non-overlapping mutated and unmutated windows were compared. Next, the mutated windows were split into ten equal groups (deciles) ranked from lowest to highest mutation frequency. Finally, box plots were generated comparing the PRO-seq density (normalized for the length of the windows) in the unmutated and mutated groups.

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

Statistical analysis

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

Data availability

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

Author contributions

UES performed all PRO-seq, PRO-cap and MutPE-seq assays and analyzed data. JF generated the Ramos lines with new V regions and performed infections for MutPE-seq. KF and TN developed the bioinformatic pipeline to map NGS reads to V regions. KF developed the bioinformatic pipeline for MutPE-seq and analyzed data. RV and ITS analyzed MutPE-seq and performed the analysis of mutation and Pol II stalling in V regions. Marina M performed ChIP-seq, ChIP-qPCR and RT-qPCR in Ramos cells and analyzed data. IO performed the bioinformatic analysis of non-IG AID target genes. BB performed ChIP-qPCR and RT-qPCRs. EMW and RP generated the E$\mu^{-/-}$ lines. MS conducted GCB PRO-seq with UES. ACG performed ATAC-seq. Marialaura M cloned JP8Bdel expression

868 vectors for MutPE-seq. ACG performed ATAC-seq. HM provided the germline-reverted VH4-59 and
869 VH3-30 sequences and provided feedback on the manuscript. RP conceived the project, analyzed
870 data and wrote the manuscript with critical inputs from UES and JF.

871 **Acknowledgements**

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

883

884 **Conflicts of interest**

885 The authors declare no conflicts of interest.

886

887

888

889

890

891

892

893 **FIGURE LEGENDS**

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

907 **Figure 2: Ablation of the Eμ enhancer significantly decreases nascent transcription but not**
908 **SHM. (A)** RT-qPCR measurements of nascent transcripts at the Ramos V region in AID$^{-/-}$ and AID$^{-/-}$
909 Eμ$^{-/-}$ cells (c1, c2). To account for potential clonal variation, Ramos cells were spiked with *Drosophila*
910 S2 cells prior to RNA extraction. The data was normalized to the levels of the *Drosophila*
911 housekeeping gene, *Act5c*. *GAPDH* mRNA was used as a control. Asterisks indicate $P<0.05$ using
912 the Student's t-test and ns indicates not significant ($P >0.05$). **(B)** RT-qPCR analysis as in A
913 measuring the spliced *IGH* mRNA. **(C)** Table of mutation frequencies at the Ramos V region and the
914 JH6 intron (amplicons shown in the diagram above the table) in AID$^{-/-}$ and AID$^{-/-}$ Eμ$^{-/-}$ cells following
915 infection with AID(JP8Bdel) for 7 days. Statistical analysis was performed with the Student's t-test. **(D)**
916 Flow cytometry analysis of IgM expression in AID$^{-/-}$ and AID$^{-/-}$ Eμ$^{-/-}$ Ramos cells infected with
917 AID(JP8Bdel) for 7 days. **(E)** Bar graph summarizing the flow cytometry analyses from D, showing the
918 percent of IgM loss from three independent experiments.

919 **Figure 3: Comparison of transcriptional and mutational landscapes at the endogenous Ramos**
920 **V region. (A)** Scheme depicting the strategy used to align multimapping reads to the V region. As

921 described in the text and Methods, since upstream, non-recombined V genes are silent in Ramos
922 cells, a read that maps to the recombined V region is retained if it is mapping to any V, D or J
923 segment of the human *IGH* locus but nowhere else in the human genome. The same principle is
924 applied when mapping reads to murine V regions to the mouse genome. **(B)** Integrative Genomics
925 Viewer (IGV) browser snapshot of nascent RNA 3' ends (by PRO-seq) aligned to the Ramos V region.
926 Multimapping (top track) reads are separated from uniquely mapping reads (middle track), and these
927 two tracks are subsequently combined to generate the total profile (bottom track. **(C)** PRO-cap and
928 PRO-seq 5' and 3' end densities, respectively, at the *IGH* locus along with mutation frequencies at the
929 V region displayed on the Integrative Genomics Viewer (IGV) browser. Mapping of the V region
930 transcriptome was done via the bioinformatic pipeline outlined in A and exemplified in B (total profiles
931 are shown). The locations of the antigen-binding complementary determining regions (CDR1-3) are
932 highlighted. Mutation analysis via MutPE-seq was performed following infection of AID$^{-/-}$ cells with
933 AID(JP8Bdel)-expressing lentiviruses for 7 days. **(D)** A magnified view of the V region from C above.
934 The stalling zones are shown in the panel between the PRO-seq and MutPE-seq tracks. **(E)** Details of
935 the mutation patterns at the Ramos V region. Mutated cytidines in AID hotspot motifs (WRCH) are
936 displayed as red bars. All other C:G mutation are shown as black bars and A:T mutations as grey
937 bars. The panel under the graph shows the position of both hotspot (WRCH in black, AGCT in red,
938 upper panel) and coldspot (SYC, bottom panel) motifs. **(F)** Waterfall plot with mutations ordered from
939 highest (left) to lowest (right) frequency following the color code described in E. **(G)** Mutation
940 frequency bar plot showing the percentages (indicated within the bars) of the three mutations classes
941 following the color code described in E. **(H)** Bar graph indicating the percentage of the type of
942 mutation indicated on the X axis. The C to T and G to A transition mutations are the signature of AID
943 activity.

944 **Figure 4: Comparison of transcriptional and mutational landscapes at two different human V**
945 **regions expressed from the Ramos *IGH* locus. (A)** IGV browser snapshots of nascent RNA 5'
946 (PRO-cap) and 3' (PRO-seq) ends and mutation tracks (MutPE-seq) at the VH4-59-DH2-JH6 V region
947 expressed from the Ramos VH4-34 promoter (CDRs 1-3 highlighted). Mutation analysis was
948 performed following infection of AID$^{-/-}$ cells with AID(JP8Bdel)-expressing lentiviruses for 7 days. **(B)**
949 A magnified view of the VH4-59-DH2-JH6 V region from A above. The stalling zones are shown in the
950 panel between the PRO-seq and MutPE-seq tracks. **(C)** Detailed mutational analysis of the VH4-59-

951 DH2-JH6 V region displayed and color-coded as in Fig. 3E. A bar plot, as in Fig. 3G, with the
952 percentage of mutation frequencies is shown on the right. **(D)** Nascent transcriptional and mutational
953 profiles of the VH3-30-DH2-JH6 V region as in A. **(E)** A magnified view of the VH3-30-DH2-JH6 V
954 region from D above. The stalling zones are shown in the panel between the PRO-seq and MutPE-
955 seq tracks. **(F)** Detailed mutational analysis of the VH3-30-DH2-JH6 V region (see C above for
956 details).

957 **Figure 5: Transcriptional and mutational landscapes of the murine B1-8^{hi} V region and non-*Ig***
958 **AID target genes in mice. (A)** Nascent transcriptional analysis at the *Igh* locus in murine B1-8^{hi}
959 primary, splenic B cells stimulated for four days with LPS, IL4 and RP105. IGV browser snapshots
960 show the 5' (PRO-cap) and 3' (PRO-seq) ends of the aligned reads. For MutPE-seq, *Igh*$^{B1\text{-}8hi/\ B1\text{-}8hi}$
961 *Rosa26*$^{AIDER/AIDER}$ primary B cells were activated with LPS, IL4 and RP105 for 4 days with 4-HT. PRO-
962 seq, PRO-cap and MutPE-seq were also performed from sorted, splenic germinal center B cells
963 (GCBs) following immunization with sheep red blood cells for 10 days. **(B)** Analysis of SHM at the B1-
964 8^{hi} V region from B1-8^{hi} *Rosa26*^{AIDER} primary, activated murine splenocytes. The bar graph on the
965 right shows the percentage of each indicated mutation category (see Fig. 3E and 3G for details). **(C)**
966 Analysis of SHM at the B1-8^{hi} V region from splenic GCBs following immunization with sheep red
967 blood cells for 10 days. **(D)** 5' and 3' ends of nascent RNA (PRO-cap, PRO-seq) and mutation profiles
968 at four selected AID target genes. Mutational data are from Alvarez-Prado et al. (2018) wherein the
969 first 500 bp of the genes were sequenced. The region displayed extends from -100 bp from the
970 annotated (RefSeq) TSS up to 50 bp downstream of the sequenced amplicon. The WR<u>C</u>H motifs are
971 indicated as red dots.

972 **Figure 6: Analysis of gene regulatory context on the nascent transcriptional landscape and**
973 **SHM of the murine B1-8^{hi} V region. (A)** Scheme showing the exchange of the endogenous Ramos
974 V region for the murine B1-8^{hi} V region to generate the Ramos$^{B1\text{-}8hi}$ human *IGH* locus following the
975 approach described in Fig. S2A. **(B)** 3' ends of nascent RNAs by PRO-seq at the Ramos$^{B1\text{-}8hi}$ *IGH*
976 locus showing the distribution of multimapping, unique and total signal. **(C)** Nascent RNA 5' and 3'
977 ends of PRO-cap and PRO-seq respectively and MutPE-seq analysis at the Ramos$^{B1\text{-}8hi}$ *IGH* locus.
978 MutPE-seq was performed following infection with AID(JP8Bdel)-expressing lentiviruses for 7 days.
979 The stalling zones are shown in the panel between the PRO-seq and MutPE-seq tracks. **(D)** Details of

980 the mutation patterns of the B1-8^{hi} V region expressed from the Ramos *IGH* locus obtained by MutPE-
981 seq. The bar graphs on the right show the percentage of each indicated mutation category (see Fig.
982 3E and 3G for details). For comparison, the murine B1-8^{hi} mutation profile from primary murine B cells
983 (middle panel) and murine GCBs (bottom panel), exactly as in Fig. 5B-C, are included.

984 **Figure 7: Integrative model for co-transcriptional AID targeting and differential motif mutability**.
985 **(A)** 3D structure of the human RNA Pol II elongation complex visualized with ChimeraX. We
986 superimposed, via their RPB1 subunits, the elongation complex structure (PDB ID 6GMH) onto the
987 transcribing Pol II-DSIF (a complex of SPT5 and SPT4) structure (PDB ID 5OIK). Proteins are shown
988 as surfaces (Pol II, grey; DSIF, salmon; SPT6, brown; PAF complex, yellow) and nucleic acids as
989 cartoons (DNA template strand, blue; DNA non-template strand, cyan; RNA, red). The right panel
990 highlights the trajectory of DNA and RNA buried within Pol II. The exposed non-template strand of the
991 transcription bubble is occluded from interactions with AID due to it being completely covered by
992 SPT5. **(B)** Proposed model for SHM that incorporates various known aspects of SHM biology. The
993 cartoon diagram of the Pol II elongation complex reflects the actual structure described in (A) above
994 and in the main text. AID is recruited to the Pol II complex via super-enhancers in the context of a
995 transcriptional hub (step 1) and is then retained via interactions with elongation factors, SPT5, PAF
996 and SPT6 (step 2). The next step is the availability of ssDNA (step 3). We support the idea that
997 premature transcription termination may be an important source of ssDNA [48]. This event is not
998 associated with stalling but occurs randomly within the first 2-3 kb from the TSSs [106]. A slower-moving
999 terminating Pol II may provide a more stable source of upstream, negative supercoiled DNA favoring
1000 higher AID activity. Upon dissociation of Pol II, the non-template strand is likely available as ssDNA
1001 for some time since the DNA:RNA hybrid would prevent immediate reannealing. The RNA is removed
1002 by RNA helicases and RNaseH followed by degradation of the RNA by the RNA exosome, which is
1003 known to associate with AID and provide access to the template strand. In this manner, AID can
1004 access ssDNA through various means all in the context of termination.

1005 **Figure S1: ChIP-qPCR analysis of the Ramos V region locale and generation of Ramos**
1006 **Eμ$^{-/-}$ cells. (A)** ChIP-qPCR analysis from three independent experiments in AID$^{-/-}$ Ramos cells for the
1007 indicated epigenetic marks as well as histone H3. Amplicons (1-7) used are indicated below the locus
1008 diagram shown above the graphs. The Neg. amplicon corresponds to a gene desert on chromosome

1009 1 and is used as a negative control. **(B)** ChIP-qPCR analysis from five independent experiments in
1010 B1-8^{hi} primary splenic cells for the indicated epigenetic marks as well as histone H3. Amplicons (1-6)
1011 used are indicated below the locus diagram shown above the graphs. The Neg. (negative region)
1012 amplicon corresponds to a gene desert on chromosome 1 and is used as a negative control. **(C)**
1013 Strategy to create the E$\mu^{-/-}$ AID$^{-/-}$ Ramos lines. The deleted region (583 bp) corresponds to the peak
1014 of accessible chromatin detected by ATAC-seq (top panel). This segment was replaced with a floxed
1015 reporter cassette expressing GFP or mCherry using CRISPR. Single clones double-positive for GFP
1016 and mCherry expression were isolated followed by excision of the floxed cassette by Cre
1017 recombinase. Two clones, c1 and c2 were used for all experiments. **(D)** Genotyping PCR analysis to
1018 confirm the loss of Eμ in AID$^{-/-}$ E$\mu^{-/-}$ clones c1 and c2. The location of primers is shown in the
1019 diagram above the gel image. **(E)** Surface IgM expression in E$\mu^{-/-}$ AID$^{-/-}$ Ramos clones relative to the
1020 parental AID$^{-/-}$ line determined by FACS.

1021 **Figure S2: MutPE-seq of the Ramos V region following infection of AID$^{-/-}$ cells with AID (m7.3).**
1022 **(A)** Mutation profile of the Ramos V region generated by AID (m7.3) following 21 days of infection.
1023 The type of mutation and location of hotspots is indicated and can be compared directly with Fig. 3E.
1024 **(B—D)** Waterfall plot of mutations ordered by frequency (B), bar plot of mutation frequencies in the
1025 three indicated classes (C) and bar graph of mutation type (D), as explained in the legend of Fig. 3 F
1026 to H).

1027 **Figure S3: Relationship between SHM, Pol II stalling and nascent transcription at different V**
1028 **regions. (A)** Scheme for the analysis of mutation frequency between stalling sites and stalling zones
1029 within V regions. **(B)** Box plots comparing mutation frequency between stalling and non-stalling zones
1030 in the Ramos endogenous (WT) V region, VH4-59, VH3-30 and Ramos$^{B1\text{-}8hi}$. Statistical analyses were
1031 performed with a Wilcoxon rank sum test and indicated within the plots. **(C)** Categorization of
1032 unmutated and mutated C:G pairs from MutPE-seq data, as described in the text. **(D)** Box plot
1033 showing the PRO-seq density in 7 bp windows centered at each C:G residue in the unmutated group
1034 and the ten mutated sub-groups, as described in A. The V gene analyzed is indicated at the top of
1035 each plot. The medians are indicated by black lines inside the boxes. A Wilcoxon rank sum test after
1036 multiple testing correction using the Bonferroni revealed no statistical significance (defined as $P <$
1037 0.05) between any pair of groups within any of the box plots. **(E)** Workflow for the classification of

1038 mutated C:G neighborhoods into 7 bp windows (C:G ±3 bp) based on their PRO-seq levels, as
1039 described in the text. **(F)** Box plots showing the mutation frequencies of the central C:G residue in the
1040 7 bp windows (C:G ±3 bp) within all ten PRO-seq sub-groups. The V gene analyzed is indicated at the
1041 top of each plot. The medians are indicated by black lines inside the boxes. A Wilcoxon rank sum test
1042 with the Bonferroni correction revealed no statistical significance (defined as $P < 0.05$) between any
1043 pair of groups within any of the box plots.

1044 **Figure S4: PRO-seq profiling of the Ramos V region in AID$^{-/-}$ Eμ$^{-/-}$ cells. (A)** Analysis of nascent
1045 transcription 3' ends (PRO-seq) at the *IGH* locus in AID$^{-/-}$ and AID$^{-/-}$ Eμ$^{-/-}$ Ramos cells (c1, c2). The
1046 promoter (Prom.), complementary determining regions (CDR 1-3), intronic enhancer (Eμ) and switch
1047 μ region (Sμ). The lower panel shows a magnified view of the V region locale. **(B)** Quantification of
1048 PRO-seq read counts from the data in A. Signals were quantified separately on the sense and
1049 antisense strands and further divided into regions upstream and downstream of the deleted Eμ
1050 segment. The upstream portion extends from the 5' end of the V region to the 5' end of the deleted Eμ
1051 segment. The downstream portion extends from the 3' end of the deleted Eμ segment to the 5' end of
1052 Sμ.

1053 **Figure S5: Generation of Ramos cell lines expressing exogenous V regions. (A)** Schematic
1054 representation of the workflow for generating Ramos cells expressing new V regions using CRISPR-
1055 based editing. An IgM-negative line was made by excising the endogenous V region in AID$^{-/-}$ cells
1056 (AID$^{-/-}$ΔV) and replacing it with a unique small guide RNA (sgRNA)-targeting sequence (green).
1057 Subsequently, an sgRNA targeting this site is combined with Cas9 and homology repair templates
1058 harboring any new V regions. Correct integration leads to restoration of surface IgM expression which
1059 is used as a readout to identify correctly targeted clones. **(B)** Flow cytometry analysis of surface IgM
1060 expression in Ramos cell clones expressing the indicated human and mouse V regions. Shown are
1061 three clones of human VH4-59-DH2-JH6 and human VH3-30-DH2-JH6 expressing Ramos cells, and
1062 two clones of B1-8^{hi} expressing Ramos cells. **(C)** PRO-seq analysis showing the 3' ends of aligned
1063 reads at the human VH4-59-DH2-JH6 V region expressed from the VH4-34 promoter at the human
1064 *IGH* in Ramos cells. Tracks of Multimapping, uniquely mapping and total are shown (see also Fig. 3A
1065 and 3B for a detailed description). **(D)** PRO-seq analysis as in C of the human VH3-30-DH2-JH6 V

1066 region expressed from the VH4-34 promoter at the human *IGH* in Ramos cells. **(E)** PRO-seq analysis
1067 as in C at the murine B1-8^{hi} V region at the *Igh* locus in mice.

1068 **Figure S6: Analysis of PRO-cap, PRO-seq and SHM at non-Ig AID target genes in murine**
1069 **GCBs. (A)** 5' and 3' ends of nascent RNA (PRO-cap, PRO-seq) and mutation profiles (from Alvarez-
1070 Prado et al. [17]) at selected AID target genes as in Fig. 5D-G. Shown are genes where highly mutated
1071 residues lie within 150 bp of the transcription initiation site defined by the peak of PRO-cap signals.
1072 The WRCH motifs are indicated as red dots. (B) As in A but showing genes where highly mutated
1073 residues lie near the transcription initiation site defined by the peak of PRO-cap signals.

1074 **Figure S7: Additional examples of PRO-cap, PRO-seq and SHM at non-Ig AID target genes in**
1075 **murine GCBs.** See Fig. 5D for detailed legend.

1076 **Figure S8: Relationship between SHM and transcriptional strength at non-IG AID target genes.**
1077 **(A)** Workflow for the classification of mutated and unmutated genes. **(B)** Box plot showing the PRO-
1078 seq density in 7 bp windows centered at each C:G residue in the three groups and sub-groups
1079 described in A. The medians are indicated by black lines inside the boxes. The asterisk indicates $P <$
1080 0.05 based on the Wilcoxon rank sum test after multiple testing correction using the Bonferroni
1081 method. **(C)** Scheme for the classification of mutated C:G neighborhoods based on their PRO-seq
1082 density. **(D)** Box plots showing the mutation frequencies of the central C:G residue in the 7 bp
1083 windows (C:G ±3 bp) within all ten PRO-seq sub-groups. The medians are indicated by black lines
1084 inside the boxes. The asterisk indicates $P < 0.05$ based on the Wilcoxon rank sum test after multiple
1085 testing correction using the Bonferroni method.

1086 **Table S1: List of all primers, oligos, antibodies and sgRNAs used in this study**

1087 **Table S2: Statistical analysis of mutation frequencies and PRO-seq densities in different V**
1088 **regions.** The upper and lower tables show the *P* values for the analyses in Fig. S3D and S3F,
1089 respectively. *P* values are based on the Wilcoxon rank sum test following multiple testing correction
1090 with the Bonferroni method.

1091

1092

1093

1094

1095 **REFERENCES**

1096 1. Peled, J.U., et al. (2008). The Biochemistry of Somatic Hypermutation. Annu Rev Immunol
1097 *26*, 481-511.
1098 2. Odegard, V.H., and Schatz, D.G. (2006). Targeting of somatic hypermutation. Nat Rev
1099 Immunol *6*, 573-83. nri1896 [pii]10.1038/nri1896.
1100 3. Di Noia, J.M., and Neuberger, M.S. (2007). Molecular mechanisms of antibody somatic
1101 hypermutation. Annu Rev Biochem *76*, 1-22.
1102 4. Victora, G.D., and Nussenzweig, M.C. (2012). Germinal centers. Annu Rev Immunol *30*, 429-
1103 57. 10.1146/annurev-immunol-020711-075032.
1104 5. Muramatsu, M., et al. (2000). Class switch recombination and hypermutation require
1105 activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. Cell *102*, 553-
1106 63.
1107 6. Revy, P., et al. (2000). Activation-induced cytidine deaminase (AID) deficiency causes the
1108 autosomal recessive form of the Hyper-IgM syndrome (HIGM2). Cell *102*, 565-75.
1109 7. Petersen-Mahrt, S.K., et al. (2002). AID mutates E. coli suggesting a DNA deamination
1110 mechanism for antibody diversification. Nature *418*, 99-103.
1111 10.1038/nature00862nature00862 [pii].
1112 8. Ramiro, A.R., et al. (2003). Transcription enhances AID-mediated cytidine deamination by
1113 exposing single-stranded DNA on the nontemplate strand. Nat Immunol *4*, 452-6.
1114 10.1038/ni920.
1115 9. Bransteitter, R., et al. (2003). Activation-induced cytidine deaminase deaminates
1116 deoxycytidine on single-stranded DNA but requires the action of RNase. Proc Natl Acad Sci U
1117 S A *100*, 4102-7.
1118 10. Pham, P., et al. (2003). Processive AID-catalysed cytosine deamination on single-stranded
1119 DNA simulates somatic hypermutation. Nature *424*, 103-7. 10.1038/nature01760nature01760
1120 [pii].
1121 11. Methot, S.P., and Di Noia, J.M. (2017). Molecular Mechanisms of Somatic Hypermutation and
1122 Class Switch Recombination. Adv Immunol *133*, 37-87. 10.1016/bs.ai.2016.11.002.
1123 12. Vaidyanathan, B., et al. (2014). AIDing Chromatin and Transcription-Coupled Orchestration of
1124 Immunoglobulin Class-Switch Recombination. Frontiers in Immunology *5*.
1125 10.3389/fimmu.2014.00120.
1126 13. Leeman-Neill, R.J., et al. (2018). The Common Key to Class-Switch Recombination and
1127 Somatic Hypermutation: Discovery of AID and Its Role in Antibody Gene Diversification. The
1128 Journal of Immunology *201*, 2527-2529. 10.4049/jimmunol.1801246.
1129 14. Saha, T., et al. (2020). Charting a DNA Repair Roadmap for Immunoglobulin Class Switch
1130 Recombination. Trends Biochem Sci. 10.1016/j.tibs.2020.10.005.
1131 15. Chaudhuri, J., and Alt, F.W. (2004). Class-switch recombination: interplay of transcription,
1132 DNA deamination and DNA repair. Nat Rev Immunol *4*, 541-52. 10.1038/nri1395nri1395 [pii].
1133 16. Liu, M., et al. (2008). Two levels of protection for the B cell genome during somatic
1134 hypermutation. Nature *451*, 841-5.
1135 17. Alvarez-Prado, A.F., et al. (2018). A broad atlas of somatic hypermutation allows prediction of
1136 activation-induced deaminase targets. J Exp Med *215*, 761-771. 10.1084/jem.20171738.
1137 18. Yamane, A., et al. (2010). Deep-sequencing identification of the genomic targets of the
1138 cytidine deaminase AID and its cofactor RPA in B lymphocytes. Nat Immunol. ni.1964
1139 [pii]10.1038/ni.1964.
1140 19. Shen, H.M., et al. (1998). Mutation of BCL-6 gene in normal B cells by the process of somatic
1141 hypermutation of Ig genes. Science *280*, 1750-2.
1142 20. Pasqualucci, L., et al. (2001). Hypermutation of multiple proto-oncogenes in B-cell diffuse
1143 large-cell lymphomas. Nature *412*, 341-6. 10.1038/3508558835085588 [pii].
1144 21. Ramiro, A.R., et al. (2004). AID is required for c-myc/IgH chromosome translocations in vivo.
1145 Cell *118*, 431-8.

1146 22. Buerstedde, J.M., et al. (2014). Targeting of somatic hypermutation by immunoglobulin
1147 enhancer and enhancer-like sequences. PLoS Biol *12*, e1001831.
1148 10.1371/journal.pbio.1001831.
1149 23. Dinesh, R.K., et al. (2020). Transcription factor binding at Ig enhancers is linked to somatic
1150 hypermutation targeting. Eur J Immunol *50*, 380-395. 10.1002/eji.201948357.
1151 24. Qian, J., et al. (2014). B cell super-enhancers and regulatory clusters recruit AID tumorigenic
1152 activity. Cell *159*, 1524-37. 10.1016/j.cell.2014.11.013.
1153 25. Senigl, F., et al. (2019). Topologically Associated Domains Delineate Susceptibility to Somatic
1154 Hypermutation. Cell Rep *29*, 3902-3915 e8. 10.1016/j.celrep.2019.11.039.
1155 26. Methot, S.P., et al. (2018). A licensing step links AID to transcription elongation for
1156 mutagenesis in B cells. Nat Commun *9*, 1248. 10.1038/s41467-018-03387-6.
1157 27. Pavri, R., et al. (2010). Activation-induced cytidine deaminase targets DNA at sites of RNA
1158 polymerase II stalling by interaction with Spt5. Cell *143*, 122-33. S0092-8674(10)01065-2
1159 [pii]10.1016/j.cell.2010.09.017.
1160 28. Willmann, K.L., et al. (2012). A role for the RNA pol II-associated PAF complex in AID-
1161 induced immune diversification. J Exp Med *209*, 2099-111. 10.1084/jem.20112145.
1162 29. Begum, N.A., et al. (2012). The histone chaperone Spt6 is required for activation-induced
1163 cytidine deaminase target determination through H3K4me3 regulation. J Biol Chem *287*,
1164 32415-29. M112.351569 [pii]10.1074/jbc.M112.351569.
1165 30. Nambu, Y., et al. (2003). Transcription-coupled events associating with immunoglobulin
1166 switch region chromatin. Science *302*, 2137-40. 10.1126/science.1092481302/5653/2137
1167 [pii].
1168 31. Basu, U., et al. (2011). The RNA exosome targets the AID cytidine deaminase to both strands
1169 of transcribed duplex DNA substrates. Cell *144*, 353-63. S0092-8674(11)00002-X
1170 [pii]10.1016/j.cell.2011.01.001.
1171 32. Nowak, U., et al. (2011). The splicing regulator PTBP2 interacts with the cytidine deaminase
1172 AID and promotes binding of AID to switch-region DNA. Nat Immunol *12*, 160-6. ni.1977
1173 [pii]10.1038/ni.1977.
1174 33. Duan, Z., et al. (2021). Role of Dot1L and H3K79 methylation in regulating somatic
1175 hypermutation of immunoglobulin genes. Proc Natl Acad Sci U S A *118*.
1176 10.1073/pnas.2104013118.
1177 34. Yu, G., et al. (2021). The role of HIRA-dependent H3.3 deposition and its modifications in the
1178 somatic hypermutation of immunoglobulin variable regions. Proc Natl Acad Sci U S A *118*.
1179 10.1073/pnas.2114743118.
1180 35. Aida, M., et al. (2013). Accumulation of the FACT complex, as well as histone H3.3, serves as
1181 a target marker for somatic hypermutation. Proc Natl Acad Sci U S A *110*, 7784-9.
1182 1305859110 [pii]10.1073/pnas.1305859110.
1183 36. Jeevan-Raj, B.P., et al. (2011). Epigenetic tethering of AID to the donor switch region during
1184 immunoglobulin class switch recombination. J Exp Med *208*, 1649-60. jem.20110118
1185 [pii]10.1084/jem.20110118.
1186 37. Stanlie, A., et al. (2010). Histone3 lysine4 trimethylation regulated by the facilitates chromatin
1187 transcription complex is critical for DNA cleavage in class switch recombination. Proc Natl
1188 Acad Sci U S A *107*, 22190-5. 1016923108 [pii]10.1073/pnas.1016923108.
1189 38. Bradley, S.P., et al. (2006). The Histone Methyltransferase Suv39h1 Increases Class Switch
1190 Recombination Specifically to IgA. The Journal of Immunology *177*, 1179-1188.
1191 10.4049/jimmunol.177.2.1179.
1192 39. Kuang, F.L., et al. (2009). H3 trimethyl K9 and H3 acetyl K9 chromatin modifications are
1193 associated with class switch recombination. Proceedings of the National Academy of
1194 Sciences *106*, 5288-5293. 10.1073/pnas.0901368106.
1195 40. Daniel, J.A., et al. (2010). PTIP promotes chromatin changes critical for immunoglobulin class
1196 switch recombination. Science *329*, 917-23. 10.1126/science.1187942.
1197 41. Vaidyanathan, B., and Chaudhuri, J. (2015). Epigenetic Codes Programing Class Switch
1198 Recombination. Front Immunol *6*, 405. 10.3389/fimmu.2015.00405.
1199 42. Larijani, M., and Martin, A. (2007). Single-stranded DNA structure and positional context of
1200 the target cytidine determine the enzymatic efficiency of AID. Mol Cell Biol *27*, 8038-48.
1201 MCB.01046-07 [pii]10.1128/MCB.01046-07.
1202 43. Larijani, M., et al. (2007). AID associates with single-stranded DNA with high affinity and a
1203 long complex half-life in a sequence-independent manner. Mol Cell Biol *27*, 20-30.
1204 MCB.00824-06 [pii]10.1128/MCB.00824-06.

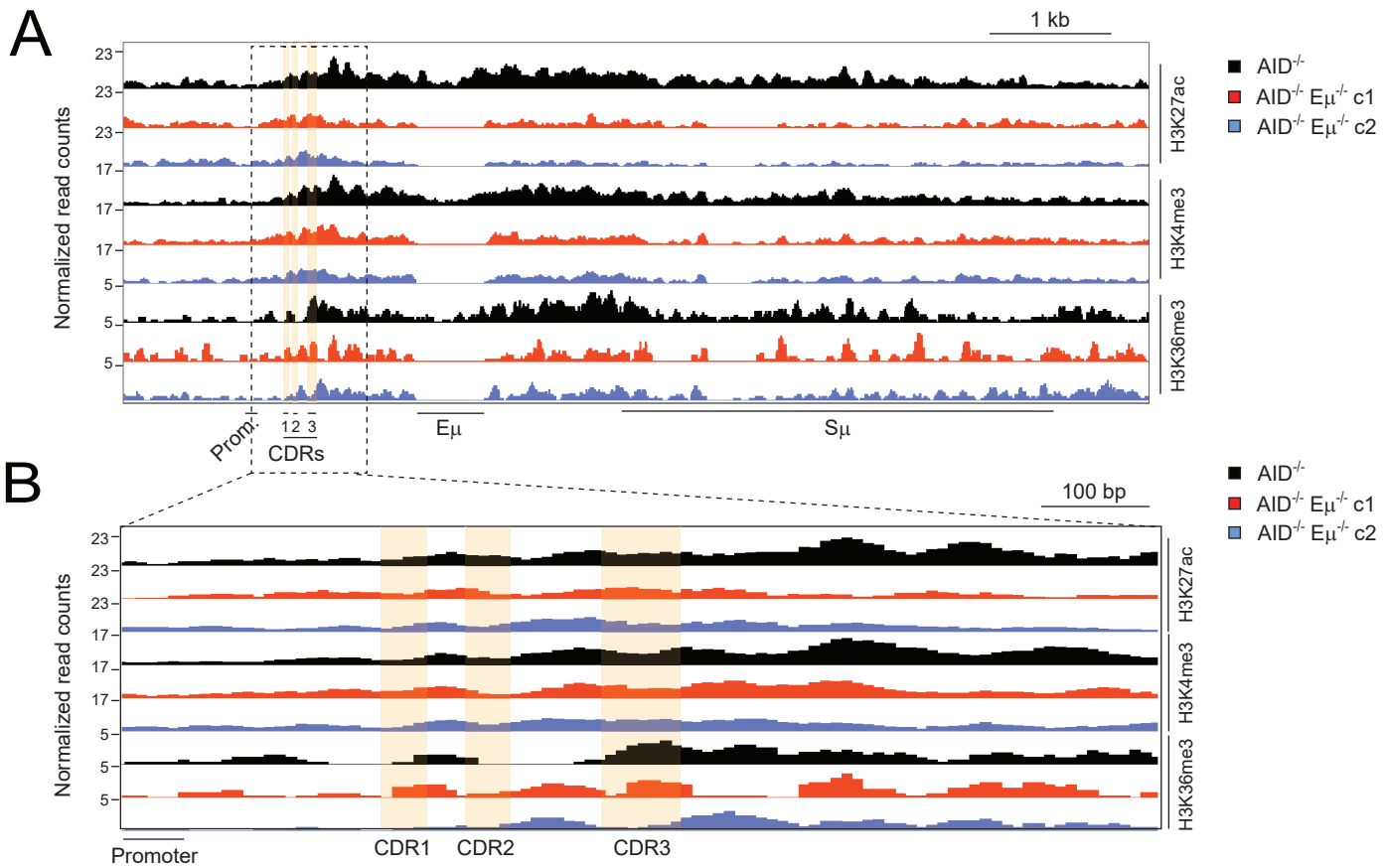
44. Shen, H.M., and Storb, U. (2004). Activation-induced cytidine deaminase (AID) can target both DNA strands when the DNA is supercoiled. Proc Natl Acad Sci U S A *101*, 12997-3002. 10.1073/pnas.0404974101.
45. Branton, S.A., et al. (2020). Activation-induced cytidine deaminase can target multiple topologies of double-stranded DNA in a transcription-independent manner. FASEB J *34*, 9245-9268. 10.1096/fj.201903036RR.
46. Taylor, B.J., et al. (2014). Active RNAP pre-initiation sites are highly mutated by cytidine deaminases in yeast, with AID targeting small RNA genes. Elife *3*, e03553. 10.7554/eLife.03553.
47. Kodgire, P., et al. (2013). Changes in RNA polymerase II progression influence somatic hypermutation of Ig-related genes by AID. J Exp Med *210*, 1481-92. 10.1084/jem.20121523.
48. Wang, X., et al. (2014). A source of the single-stranded DNA substrate for activation-induced deaminase during somatic hypermutation. Nat Commun *5*, 4137. 10.1038/ncomms5137.
49. Pefanis, E., et al. (2014). Noncoding RNA transcription targets AID to divergently transcribed loci in B cells. Nature *514*, 389-393. 10.1038/nature13580.
50. Pefanis, E., and Basu, U. (2015). RNA Exosome Regulates AID DNA Mutator Activity in the B Cell Genome. Adv Immunol *127*, 257-308. 10.1016/bs.ai.2015.04.002.
51. Meng, F.L., et al. (2014). Convergent transcription at intragenic super-enhancers targets AID-initiated genomic instability. Cell *159*, 1538-48. 10.1016/j.cell.2014.11.014.
52. Rajagopal, D., et al. (2009). Immunoglobulin switch mu sequence causes RNA polymerase II accumulation and reduces dA hypermutation. J Exp Med *206*, 1237-44. jem.20082514 [pii]10.1084/jem.20082514.
53. Wang, L., et al. (2009). S region sequence, RNA polymerase II, and histone modifications create chromatin accessibility during class switch recombination. J Exp Med *206*, 1817-30. jem.20081678 [pii]10.1084/jem.20081678.
54. Maul, R.W., et al. (2014). Spt5 accumulation at variable genes distinguishes somatic hypermutation in germinal center B cells from ex vivo-activated cells. J Exp Med *211*, 2297-306. 10.1084/jem.20131512.
55. Tarsalainen, A., et al. (2022). Ig Enhancers Increase RNA Polymerase II Stalling at Somatic Hypermutation Target Sequences. J Immunol *208*, 143-154. 10.4049/jimmunol.2100923.
56. Canugovi, C., et al. (2009). Transcriptional pausing and stalling causes multiple clustered mutations by human activation-induced deaminase. The FASEB Journal *23*, 34-44. https://doi.org/10.1096/fj.08-115352.
57. Parsa, J.Y., et al. (2012). Negative supercoiling creates single-stranded patches of DNA that are substrates for AID-mediated mutagenesis. PLoS Genet *8*, e1002518. 10.1371/journal.pgen.1002518.
58. Ronai, D., et al. (2007). Detection of chromatin-associated single-stranded DNA in regions targeted for somatic hypermutation. J Exp Med *204*, 181-90. jem.20062032 [pii]10.1084/jem.20062032.
59. Yu, G., et al. (2021). 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. 10.1371/journal.pcbi.1009323.
60. Peters, A., and Storb, U. (1996). Somatic hypermutation of immunoglobulin genes is linked to transcription initiation. Immunity *4*, 57-65. S1074-7613(00)80298-8 [pii].
61. Kenter, A.L. (2012). AID targeting is dependent on RNA polymerase II pausing. Semin Immunol *24*, 281-6. 10.1016/j.smim.2012.06.001.
62. Pavri, R., and Nussenzweig, M.C. (2011). AID targeting in antibody diversity. Adv Immunol *110*, 1-26. B978-0-12-387663-8.00005-3 [pii]10.1016/B978-0-12-387663-8.00005-3.
63. Maul, R.W., and Gearhart, P.J. (2010). AID and somatic hypermutation. Adv Immunol *105*, 159-91. 10.1016/s0065-2776(10)05006-6.
64. Mahat, D.B., et al. (2016). Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). Nat Protoc *11*, 1455-76. 10.1038/nprot.2016.086.
65. Core, L.J., et al. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845-8. 10.1126/science.1162228.
66. Kwak, H., et al. (2013). Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science *339*, 950-3. 10.1126/science.1229386.
67. Core, L.J., et al. (2012). Defining the status of RNA polymerase at promoters. Cell Rep *2*, 1025-35. 10.1016/j.celrep.2012.08.034.

68. Core, L.J., et al. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nat Genet *46*, 1311-20. 10.1038/ng.3142.
69. Andersson, R., et al. (2015). Human Gene Promoters Are Intrinsically Bidirectional. Mol Cell *60*, 346-7. 10.1016/j.molcel.2015.10.015.
70. Scruggs, B.S., et al. (2015). Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. Mol Cell *58*, 1101-12. 10.1016/j.molcel.2015.04.006.
71. Williams, L.H., et al. (2015). Pausing of RNA polymerase II regulates mammalian developmental potential through control of signaling networks. Mol Cell *58*, 311-22. 10.1016/j.molcel.2015.02.003.
72. Henriques, T., et al. (2018). Widespread transcriptional pausing and elongation control at enhancers. Genes Dev *32*, 26-41. 10.1101/gad.309351.117.
73. Sale, J.E., and Neuberger, M.S. (1998). TdT-accessible breaks are scattered over the immunoglobulin V domain in a constitutively hypermutating B cell line. Immunity *9*, 859-69.
74. Martin, A., et al. (2002). Activation-induced cytidine deaminase turns on somatic hypermutation in hybridomas. Nature *415*, 802-6. 10.1038/nature714.
75. Wu, L., et al. (2022). HMCES protects immunoglobulin genes specifically from deletions during somatic hypermutation. Genes Dev *36*, 433-450. 10.1101/gad.349438.122.
76. Wei, L., et al. (2015). Overlapping hotspots in CDRs are critical sites for V region diversification. Proc Natl Acad Sci U S A *112*, E728-37. 10.1073/pnas.1500788112.
77. Yabuki, M., et al. (2005). The MRE11-RAD50-NBS1 complex accelerates somatic hypermutation and gene conversion of immunoglobulin variable regions. Nat Immunol *6*, 730-6. 10.1038/ni1215.
78. Le, Q., and Maizels, N. (2015). Cell Cycle Regulates Nuclear Stability of AID and Determines the Cellular Response to AID. PLOS Genetics *11*, e1005411. 10.1371/journal.pgen.1005411.
79. Pinaud, E., et al. (2011). The IgH locus 3' regulatory region: pulling the strings from behind. Adv Immunol *110*, 27-70. 10.1016/B978-0-12-387663-8.00002-8.
80. Shih, T.A., et al. (2002). Role of antigen receptor affinity in T cell-independent antibody responses in vivo. Nat Immunol *3*, 399-406. 10.1038/ni776ni776 [pii].
81. Hwang, J.K., et al. (2017). 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. 10.1073/pnas.1709203114.
82. Ito, S., et al. (2004). 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*, 1975-80. 10.1073/pnas.0307335101.
83. Wang, M., et al. (2009). AID upmutants isolated using a high-throughput screen highlight the immunity/cancer balance limiting DNA deaminase activity. Nat Struct Mol Biol *16*, 769-76. 10.1038/nsmb.1623.
84. Perlot, T., et al. (2005). Elucidation of IgH intronic enhancer functions via germ-line deletion. Proc Natl Acad Sci U S A *102*, 14362-7. 0507090102 [pii]10.1073/pnas.0507090102.
85. Li, F., et al. (2010). Comparison of identical and functional Igh alleles reveals a nonessential role for Emu in somatic hypermutation and class-switch recombination. J Immunol *185*, 6049-57. 10.4049/jimmunol.0902992.
86. Yamaguchi, Y., et al. (1999). NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. Cell *97*, 41-51. S0092-8674(00)80713-8 [pii].
87. Vos, S.M., et al. (2018). Structure of paused transcription complex Pol II–DSIF–NELF. Nature *560*, 601-606. 10.1038/s41586-018-0442-2.
88. Vos, S.M., et al. (2018). Structure of activated transcription complex Pol II–DSIF–PAF–SPT6. Nature *560*, 607-612. 10.1038/s41586-018-0440-4.
89. Bernecky, C., et al. (2017). Structure of a transcribing RNA polymerase II–DSIF complex reveals a multidentate DNA–RNA clamp. Nature Structural & Molecular Biology *24*, 809-815. 10.1038/nsmb.3465.
90. Robbiani, D.F., et al. (2015). Plasmodium Infection Promotes Genomic Instability and AID-Dependent B Cell Lymphoma. Cell *162*, 727-37. 10.1016/j.cell.2015.07.019.
91. Fong, N., et al. (2022). The pausing zone and control of RNA polymerase II elongation by Spt5: Implications for the pause-release model. Mol Cell *82*, 3632-3645 e4. 10.1016/j.molcel.2022.09.001.
92. Mouquet, H., et al. (2012). Complex-type N-glycan recognition by potent broadly neutralizing HIV antibodies. Proc Natl Acad Sci U S A *109*, E3268-77. 10.1073/pnas.1217207109.

93. Lorin, V., et al. (2022). Epitope convergence of broadly HIV-1 neutralizing IgA and IgG antibody lineages in a viremic controller. J Exp Med *219*. 10.1084/jem.20212045.
94. Matsuda, F., et al. (1998). The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. J Exp Med *188*, 2151-62. 10.1084/jem.188.11.2151.
95. Yeap, L.S., et al. (2015). Sequence-Intrinsic Mechanisms that Target AID Mutational Outcomes on Antibody Genes. Cell *163*, 1124-37. 10.1016/j.cell.2015.10.042.
96. Shih, T.A., et al. (2002). Role of BCR affinity in T cell dependent antibody responses in vivo. Nat Immunol *3*, 570-5. 10.1038/ni803ni803 [pii].
97. Rush, J.S., et al. (2005). Expression of activation-induced cytidine deaminase is regulated by cell division, providing a mechanistic basis for division-linked class switch recombination. Proc Natl Acad Sci U S A *102*, 13242-7. 10.1073/pnas.0502779102.
98. Wang, Q., et al. (2017). The cell cycle restricts activation-induced cytidine deaminase activity to early G1. J Exp Med *214*, 49-58. 10.1084/jem.20161649.
99. Rada, C., et al. (2004). Mismatch recognition and uracil excision provide complementary paths to both Ig switching and the A/T-focused phase of somatic mutation. Mol Cell *16*, 163-71. 10.1016/j.molcel.2004.10.011.
100. Longerich, S., et al. (2005). The very 5' end and the constant region of Ig genes are spared from somatic mutation because AID does not access these regions. J Exp Med *202*, 1443-54. jem.20051604 [pii]10.1084/jem.20051604.
101. Xue, K., et al. (2006). The in vivo pattern of AID targeting to immunoglobulin switch regions deduced from mutation spectra in msh2-/- ung-/- mice. J Exp Med *203*, 2085-94. jem.20061067 [pii]10.1084/jem.20061067.
102. Wang, Q., et al. (2014). Epigenetic targeting of activation-induced cytidine deaminase. Proc Natl Acad Sci U S A *111*, 18667-72. 10.1073/pnas.1420575111.
103. Laffleur, B., et al. (2021). Noncoding RNA processing by DIS3 regulates chromosomal architecture and somatic hypermutation in B cells. Nature Genetics *53*, 230-242. 10.1038/s41588-020-00772-0.
104. Lim, J., et al. (2017). Nuclear Proximity of Mtr4 to RNA Exosome Restricts DNA Mutational Asymmetry. Cell *169*, 523-537 e15. 10.1016/j.cell.2017.03.043.
105. Kazadi, D., et al. (2020). Effects of senataxin and RNA exosome on B-cell chromosomal integrity. Heliyon *6*, e03442. 10.1016/j.heliyon.2020.e03442.
106. Lykke-Andersen, S., et al. (2021). Integrator is a genome-wide attenuator of non-productive transcription. Mol Cell *81*, 514-529.e6. 10.1016/j.molcel.2020.12.014.
107. Elrod, N.D., et al. (2019). The Integrator Complex Attenuates Promoter-Proximal Transcription at Protein-Coding Genes. Mol Cell *76*, 738-752.e7. 10.1016/j.molcel.2019.10.034.
108. Stein, C.B., et al. (2022). Integrator endonuclease drives promoter-proximal termination at all RNA polymerase II-transcribed loci. Mol Cell *82*, 4232-4245 e11. 10.1016/j.molcel.2022.10.004.
109. Peron, S., et al. (2012). AID-driven deletion causes immunoglobulin heavy chain locus suicide recombination in B cells. Science *336*, 931-4. 10.1126/science.1218692.
110. Bal, E., et al. (2022). Super-enhancer hypermutation alters oncogene expression in B cell lymphoma. Nature *607*, 808-815. 10.1038/s41586-022-04906-8.
111. Nair, L., et al. (2020). Regulation of long non-coding RNAs and genome dynamics by the RNA surveillance machinery. Nat Rev Mol Cell Biol *21*, 123-136. 10.1038/s41580-019-0209-0.
112. Zhou, J.Q., and Kleinstein, S.H. (2020). Position-Dependent Differential Targeting of Somatic Hypermutation. J Immunol *205*, 3468-3479. 10.4049/jimmunol.2000496.
113. Spisak, N., et al. (2020). Learning the heterogeneous hypermutation landscape of immunoglobulins from high-throughput repertoire data. Nucleic Acids Research *48*, 10702-10712. 10.1093/nar/gkaa825.
114. Pham, P., et al. (2011). 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. https://doi.org/10.1074/jbc.M111.241208.
115. Tang, C., et al. (2022). Deep learning model of somatic hypermutation reveals importance of sequence context beyond hotspot targeting. iScience *25*, 103668. 10.1016/j.isci.2021.103668.
116. Wang, Y., et al. (2023). Mesoscale DNA feature in antibody-coding sequence facilitates somatic hypermutation. Cell *186*, 2193-2207 e19. 10.1016/j.cell.2023.03.030.

117. Fitz, J., et al. (2020). Spt5-mediated enhancer transcription directly couples enhancer activation with physical promoter interaction. Nat Genet *52*, 505-515. 10.1038/s41588-020-0605-6.
118. Costea, J., et al. (2023). A de novo transcription-dependent TAD boundary underpins critical multiway interactions during antibody class switch recombination. Mol Cell *83*, 681-697 e7. 10.1016/j.molcel.2023.01.014.
119. Fitz, J., et al. (2018). Regulation of RNA polymerase II processivity by Spt5 is restricted to a narrow window during elongation. EMBO J *37*. 10.15252/embj.201797965.
120. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011 *17*, 3. 10.14806/ej.17.1.200.
121. Langmead, B., et al. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25. 10.1186/gb-2009-10-3-r25.
122. Larkin, A., et al. (2020). FlyBase: updates to the Drosophila melanogaster knowledge base. Nucleic Acids Research *49*, D899-D907. 10.1093/nar/gkaa1026.
123. Smith, T., et al. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res *27*, 491-499. 10.1101/gr.209601.116.
124. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842. 10.1093/bioinformatics/btq033.
125. Bolger, A.M., et al. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114-2120. 10.1093/bioinformatics/btu170.
126. Magoč, T., and Salzberg, S.L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics *27*, 2957-63. 10.1093/bioinformatics/btr507.
127. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods *9*, 357-359. 10.1038/nmeth.1923.
128. Li, H., et al. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079. 10.1093/bioinformatics/btp352.

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

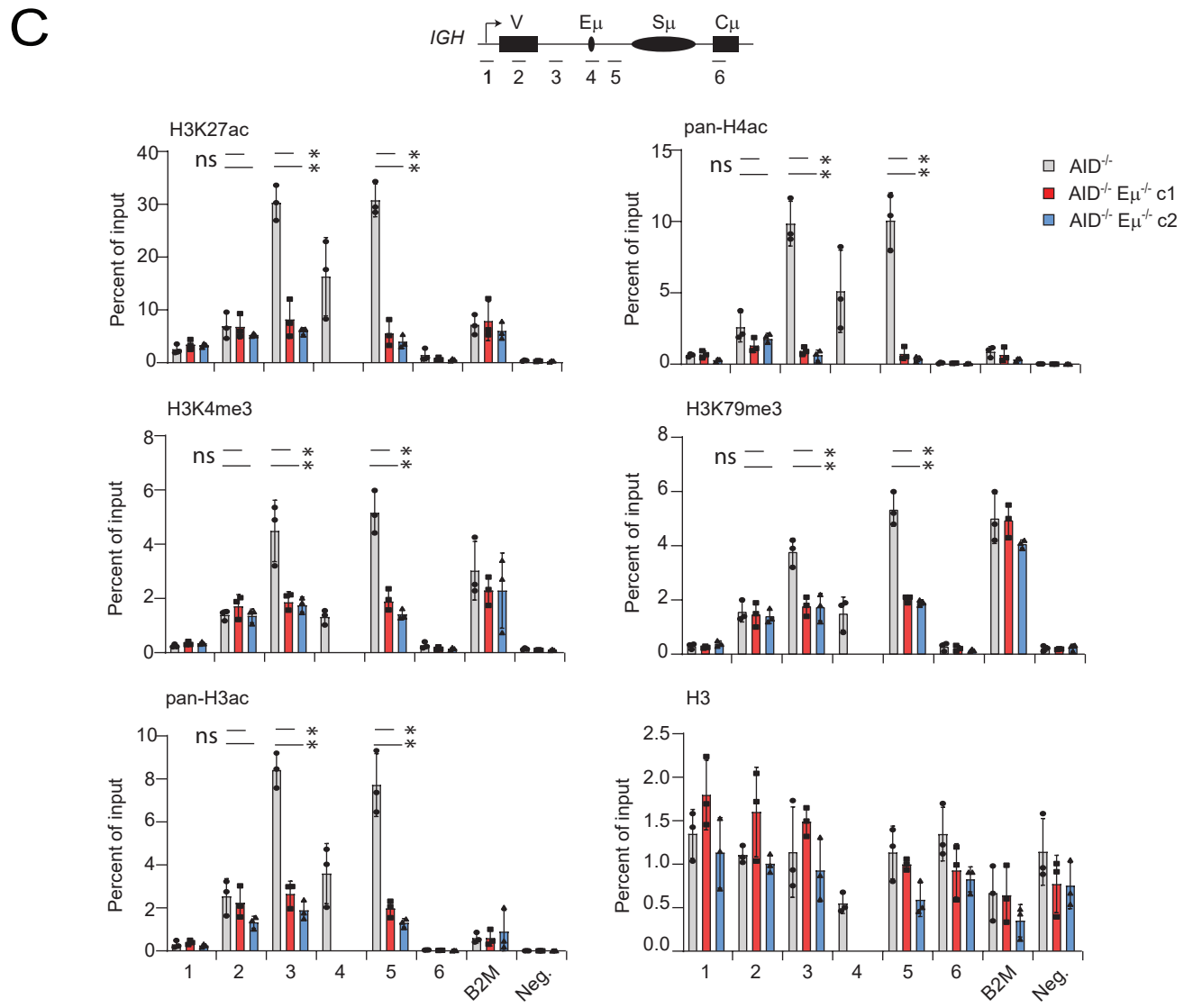
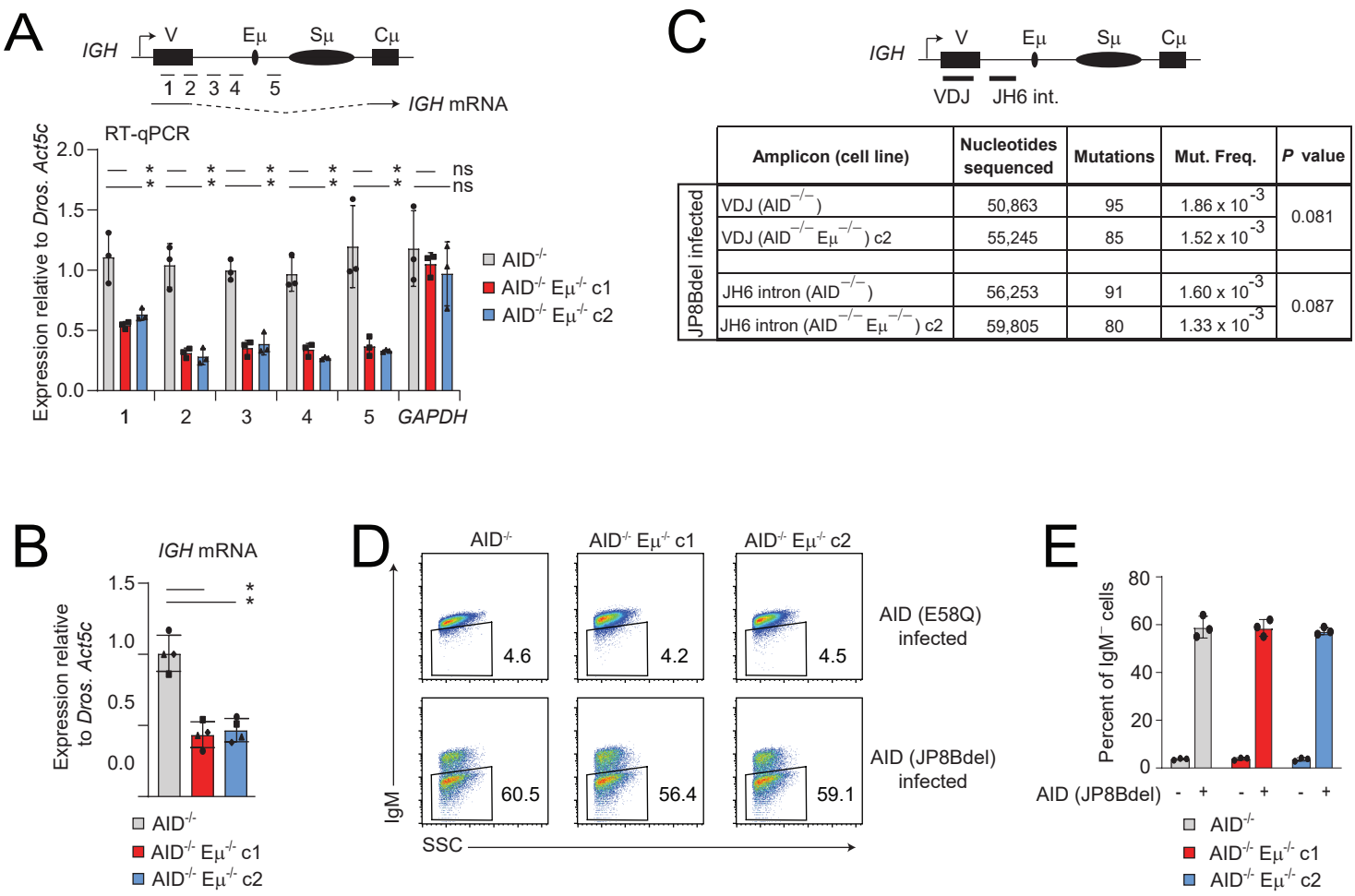

C
IGH
V
Eμ
Sμ
Cμ
1 2 3 4 5 6
H3K27ac
pan-H4ac
H3K4me3
H3K79me3
pan-H3ac
H3
Percent of input
ns
*
*
AID-/-
AID-/- Eμ-/- c1
AID-/- Eμ-/- c2
1 2 3 4 5 6 B2M Neg.

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

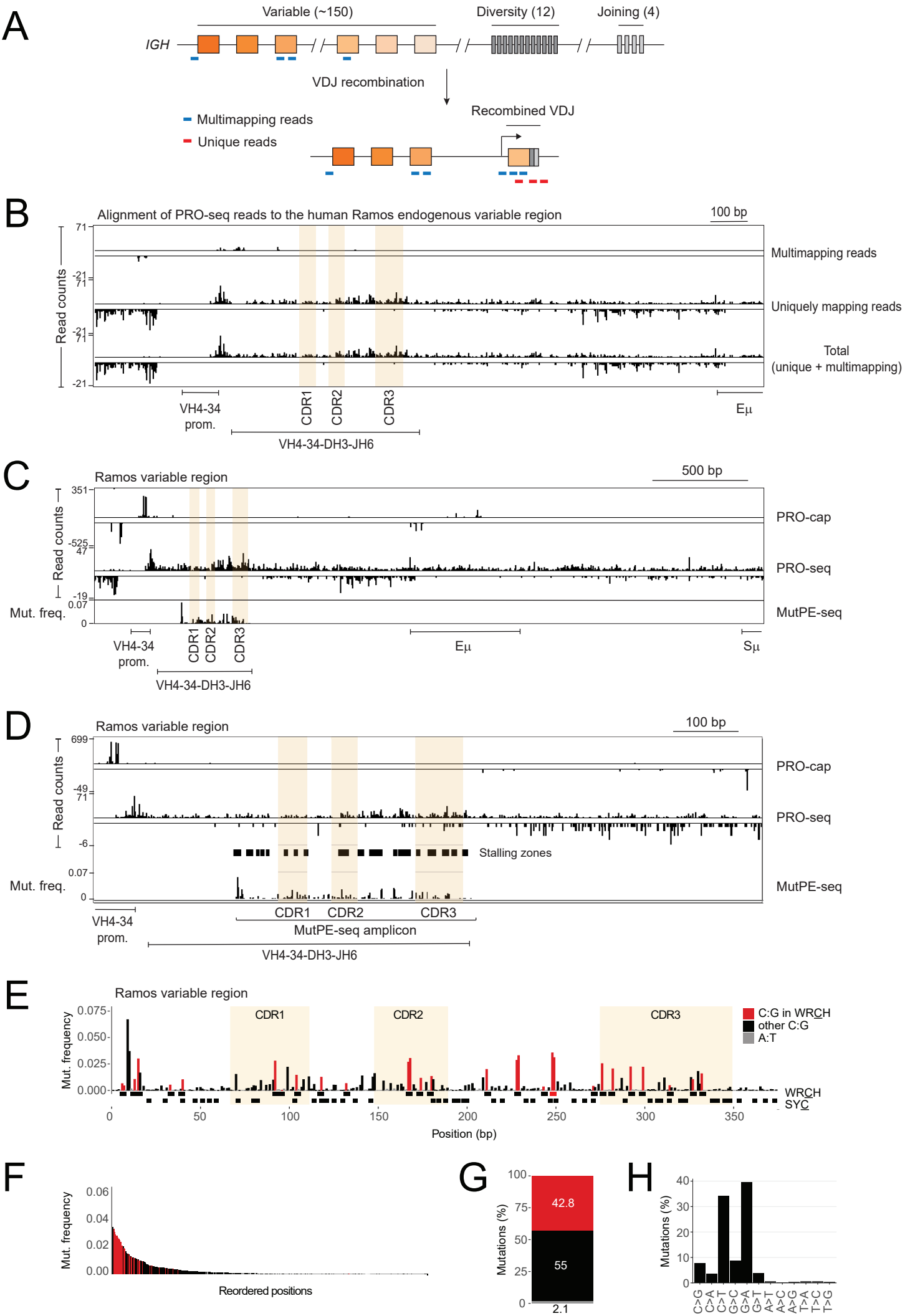

A
Variable (~150)
Diversity (12)
Joining (4)
IGH
VDJ recombination
Recombined VDJ
Multimapping reads
Unique reads
B
Alignment of PRO-seq reads to the human Ramos endogenous variable region
100 bp
Read counts
Multimapping reads
Uniquely mapping reads
Total
(unique + multimapping)
VH4-34
prom.
CDR1
CDR2
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

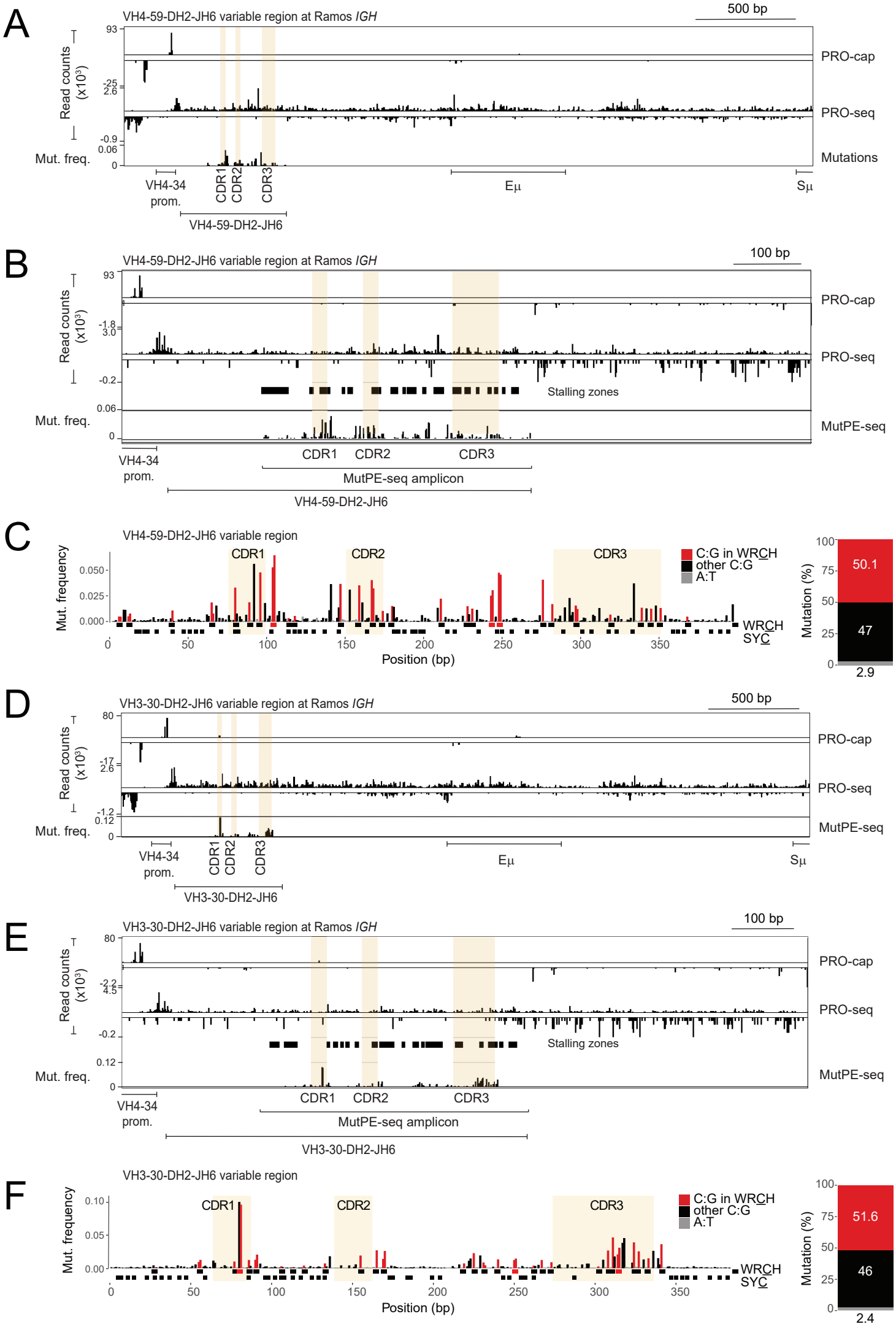
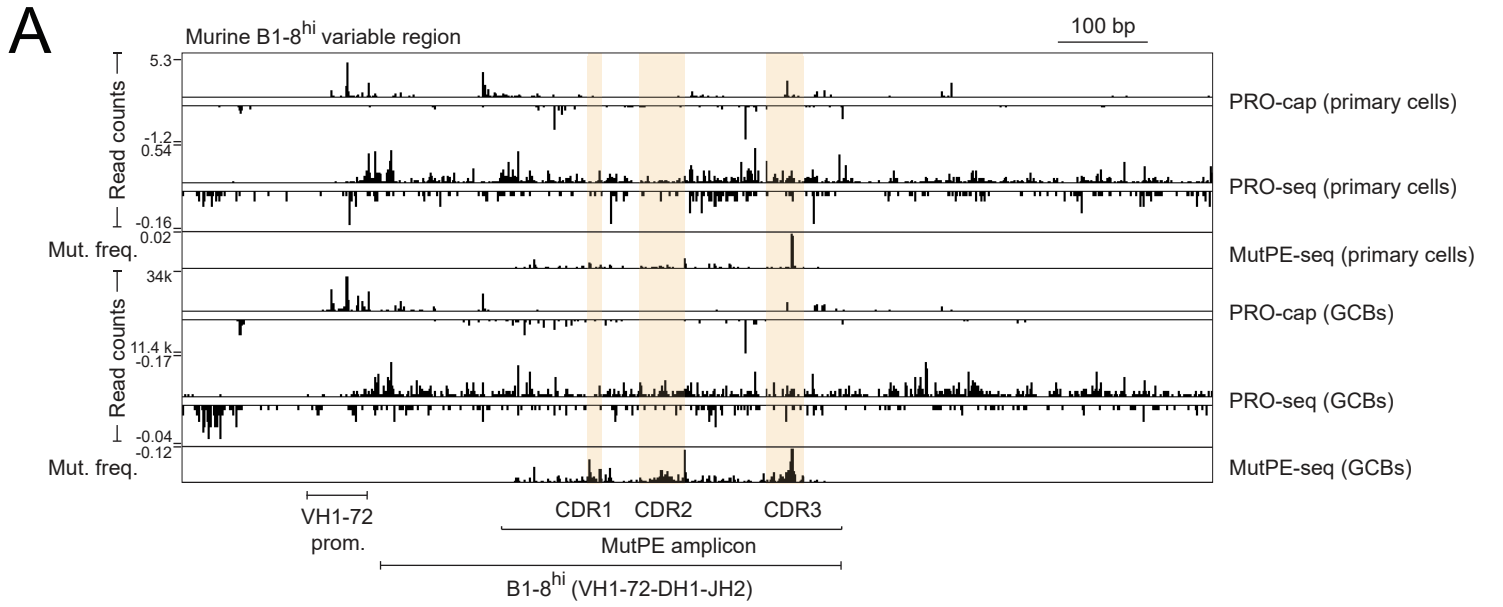
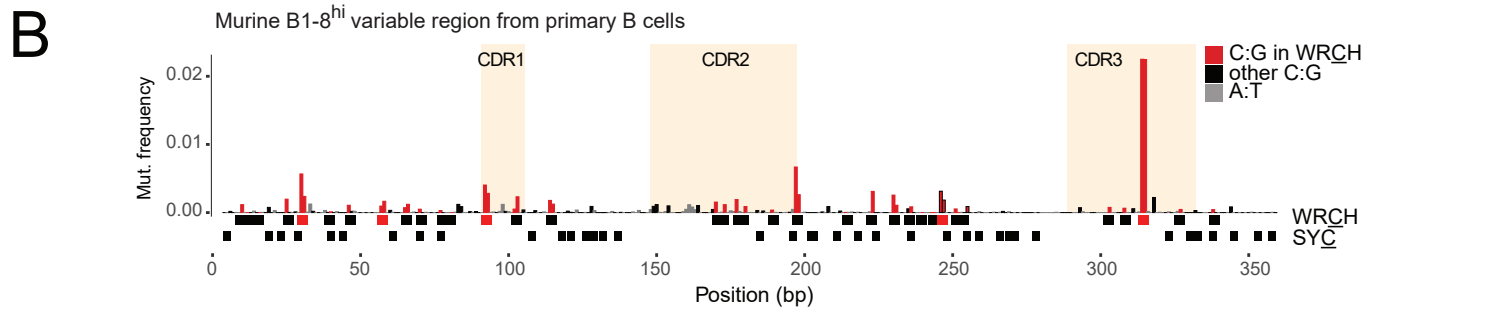
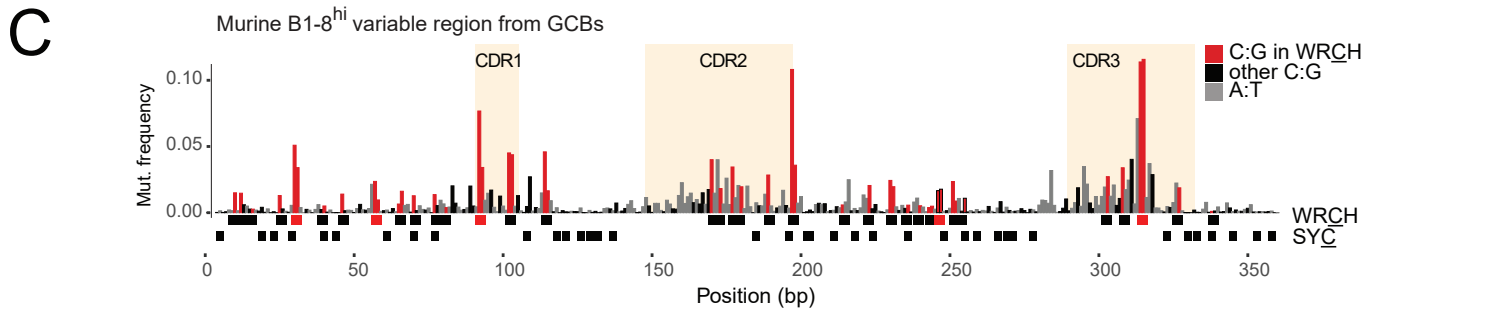

A
VH4-59-DH2-JH6 variable region at Ramos IGH
500 bp
Read counts (x10³)
PRO-cap
PRO-seq
Mut. freq.
Mutations
VH4-34 prom.
CDR1
CDR2
CDR3
Eμ
Sμ
VH4-59-DH2-JH6
B
VH4-59-DH2-JH6 variable region at Ramos IGH
100 bp
Stalling zones
MutPE-seq
MutPE-seq amplicon
C
VH4-59-DH2-JH6 variable region
Mut. frequency
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)

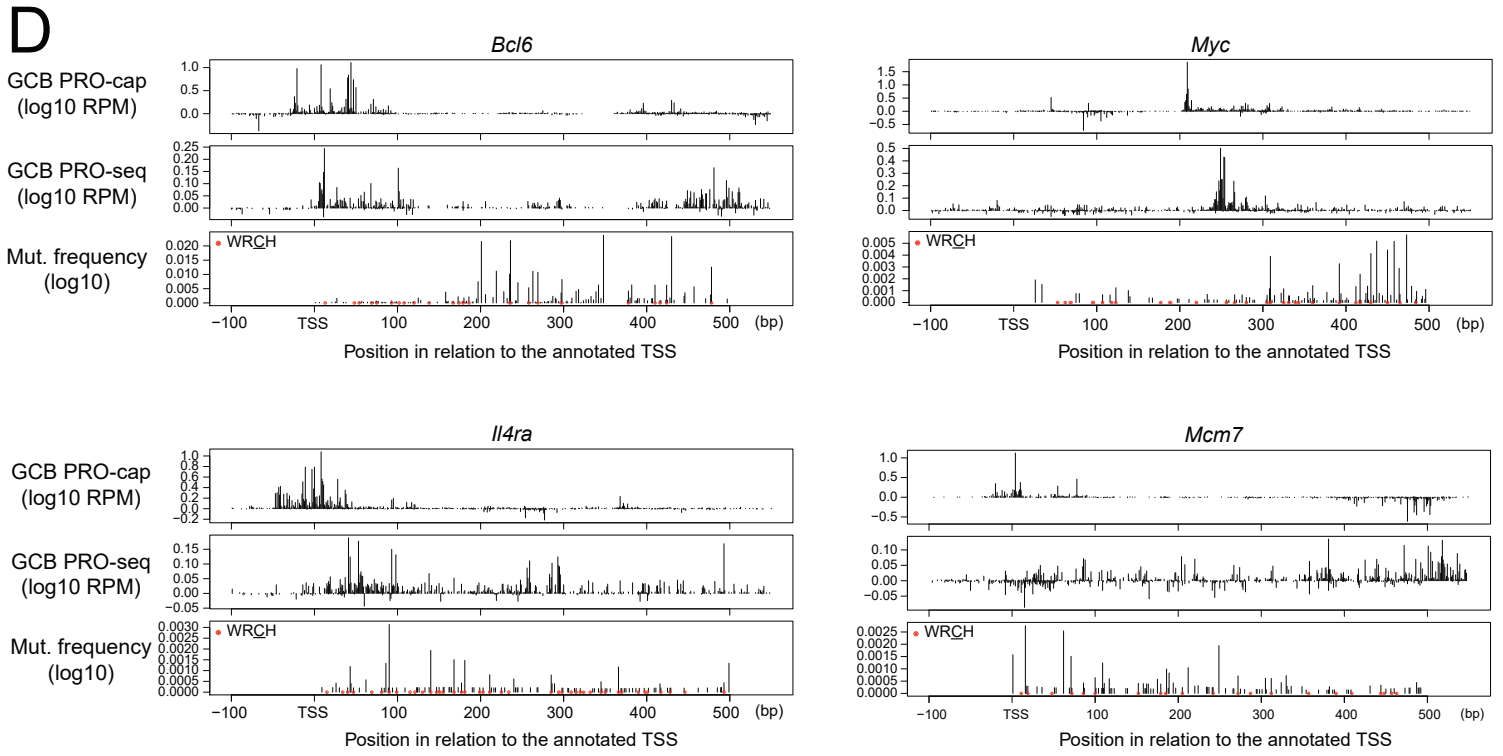

D
Bcl6
Myc
Il4ra
Mcm7
GCB PRO-cap (log10 RPM)
GCB PRO-seq (log10 RPM)
Mut. frequency (log10)
WRCH
TSS
(bp)
Position in relation to the annotated TSS

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

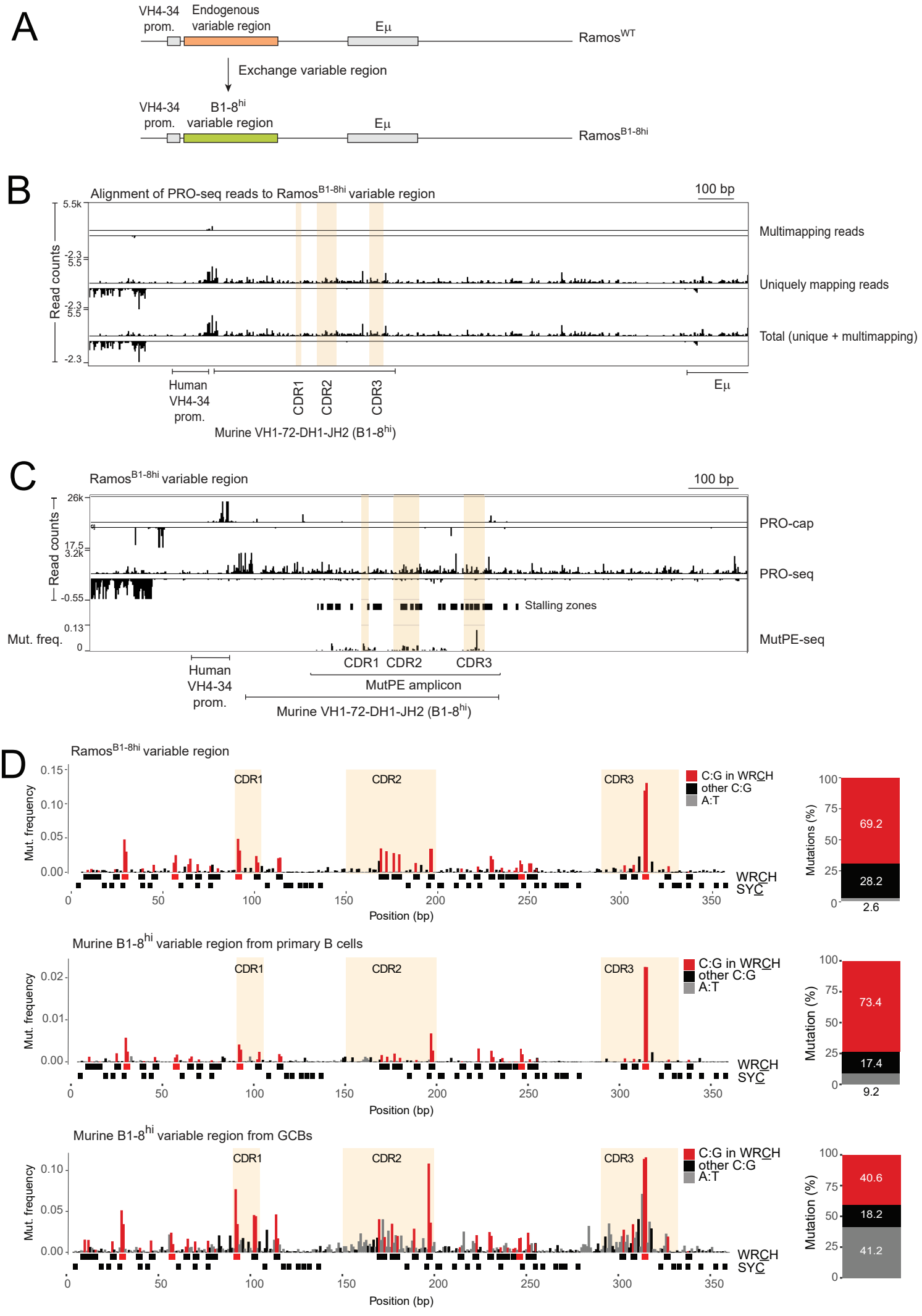
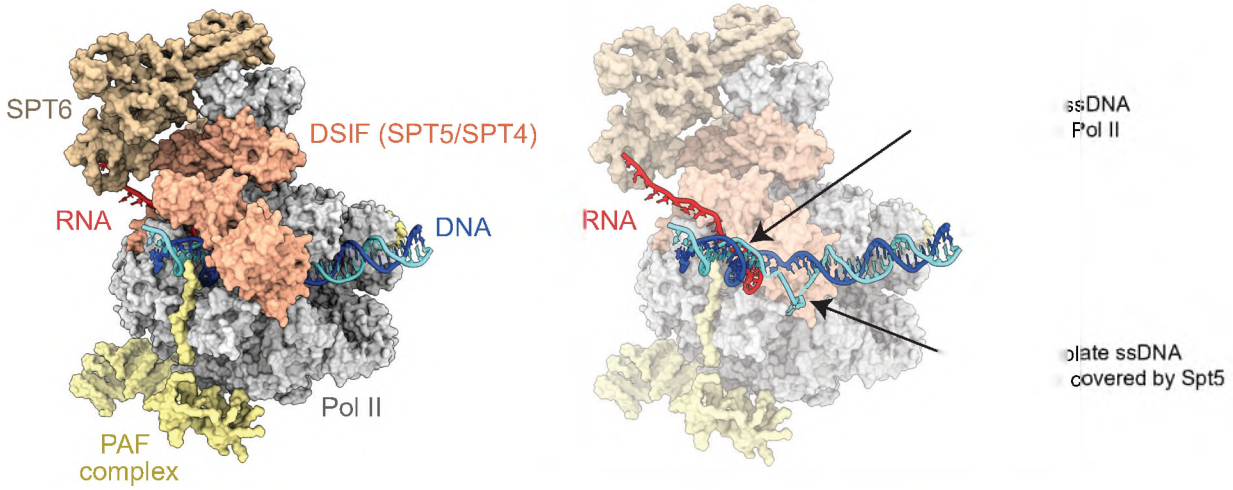

A
VH4-34 prom.
Endogenous variable region
Eμ
Ramos^WT
Exchange variable region
VH4-34 prom.
B1-8^hi variable region
Eμ
Ramos^B1-8hi
B
Alignment of PRO-seq reads to Ramos^B1-8hi variable region
100 bp
Read counts
Multimapping reads
Uniquely mapping reads
Total (unique + multimapping)
Human VH4-34 prom.
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.

Fig. 7

A


SPT6
DSIF (SPT5/SPT4)
RNA
DNA
Pol II
PAF complex
RNA
ssDNA
Pol II
plate ssDNA
covered by Spt5

B

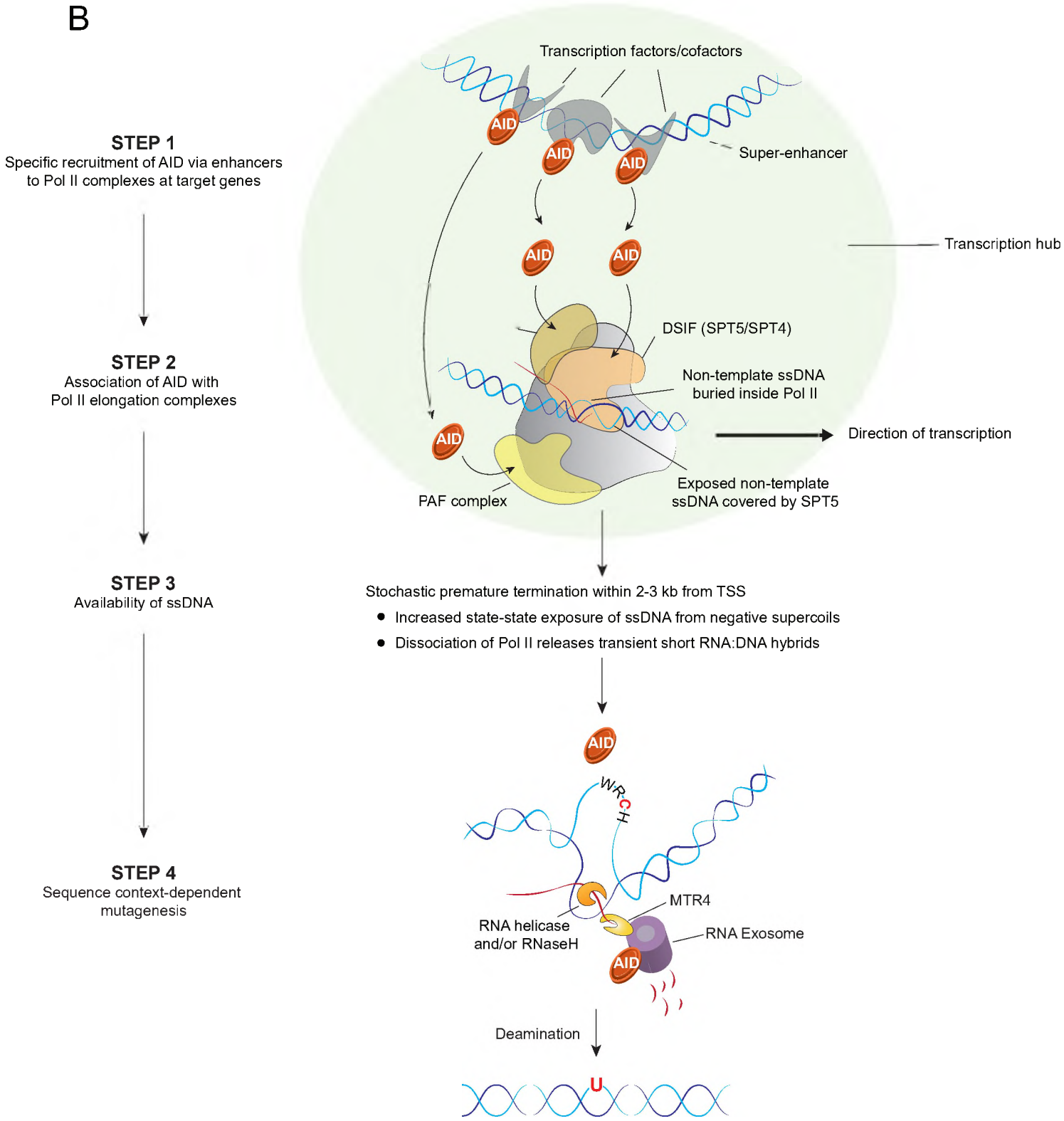

STEP 1
Specific recruitment of AID via enhancers to Pol II complexes at target genes
STEP 2
Association of AID with Pol II elongation complexes
STEP 3
Availability of ssDNA
STEP 4
Sequence context-dependent mutagenesis
Transcription factors/cofactors
AID
Super-enhancer
Transcription hub
DSIF (SPT5/SPT4)
Non-template ssDNA buried inside Pol II
Direction of transcription
PAF complex
Exposed non-template ssDNA covered by SPT5
Stochastic premature termination within 2-3 kb from TSS
Increased state-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

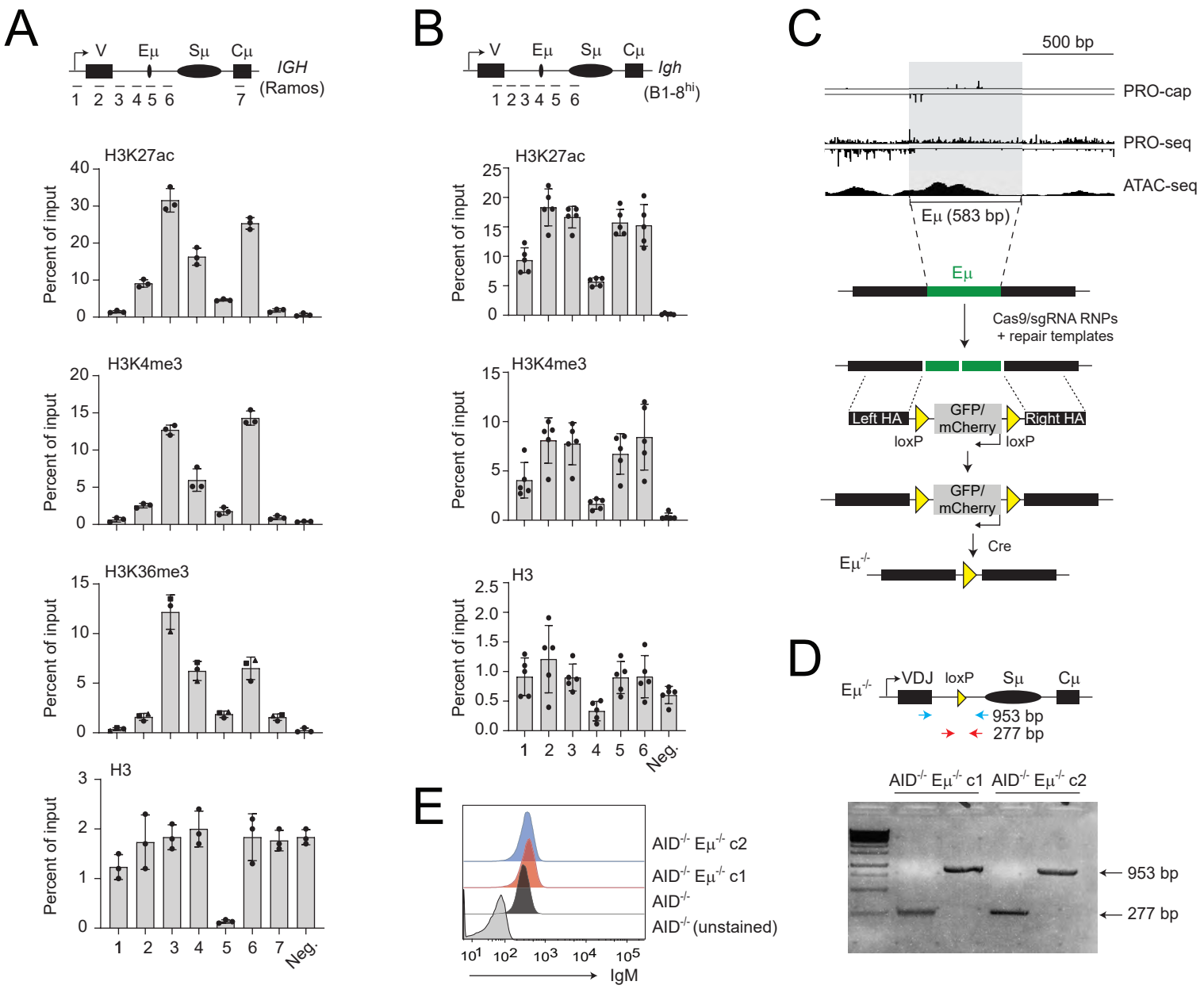


Figure S1: ChIP-qPCR analysis of the Ramos V region locale and generation of Eμ$^{-/-}$ Ramos cells. (A) ChIP-qPCR analysis from three independent experiments in AID$^{-/-}$ Ramos cells for the indicated epigenetic marks as well as histone H3. Amplicons (1-7) used are indicated below the locus diagram 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

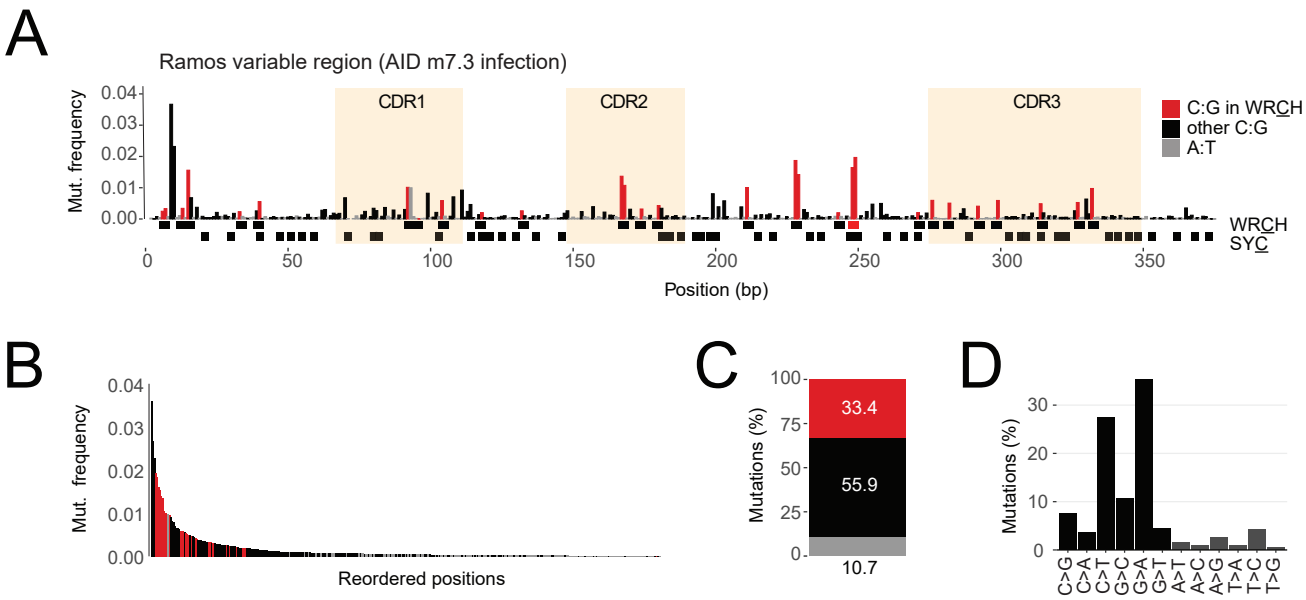


Figure S2: MutPE-seq of the Ramos V region following infection of AID$^{-/-}$ cells with AID m7.3. (A) Mutation profile of the Ramos V region generated by AID (m7.3) following 21 days of infection. The type of mutation and location of hotspots is indicated and can be compared directly with Fig. 3E. **(B—D)** 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

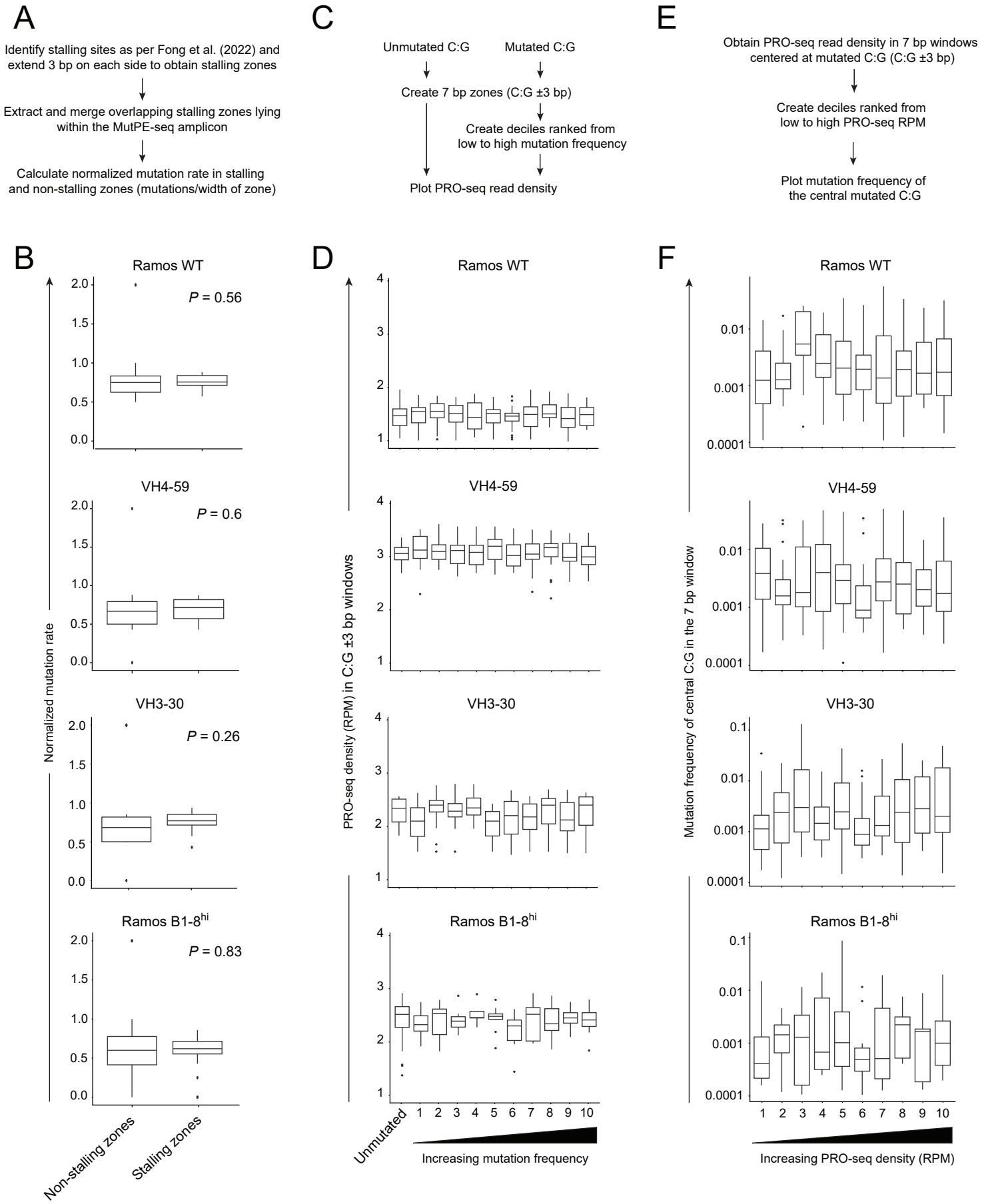

A
Identify stalling sites as per Fong et al. (2022) and extend 3 bp on each side to obtain stalling zones
Extract and merge overlapping stalling zones lying within the MutPE-seq amplicon
Calculate normalized mutation rate in stalling and non-stalling zones (mutations/width of zone)
C
Unmutated C:G
Mutated 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

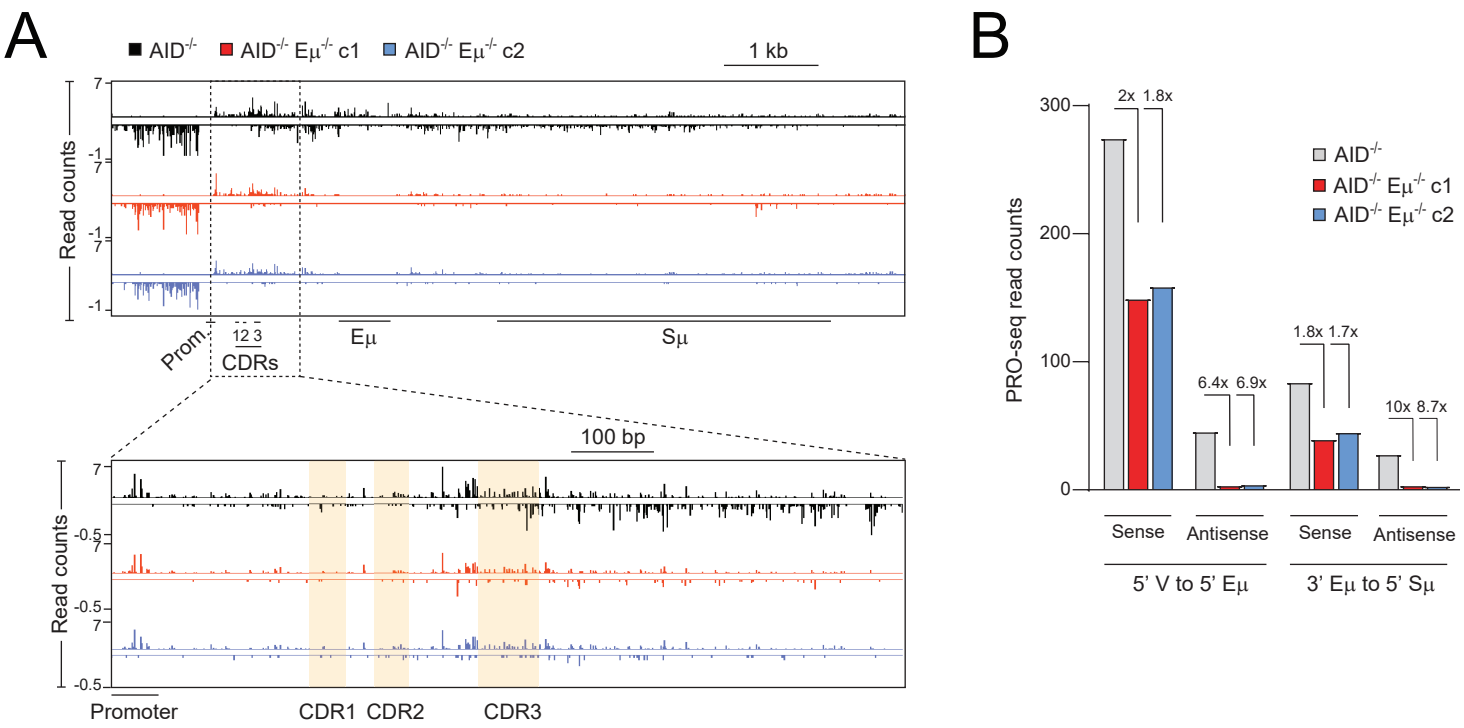
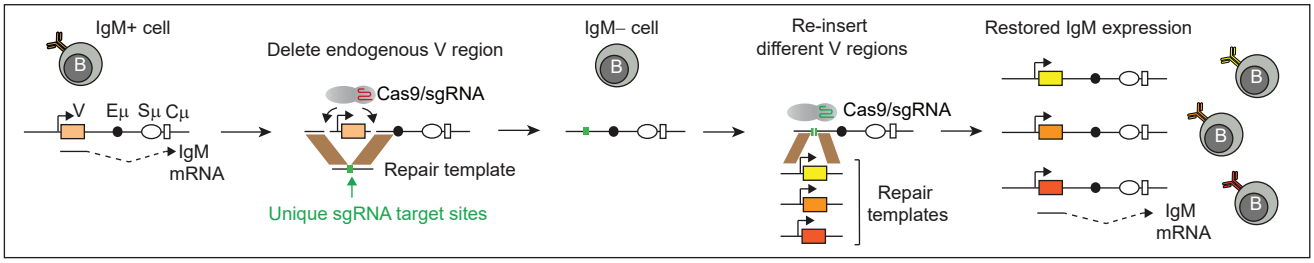


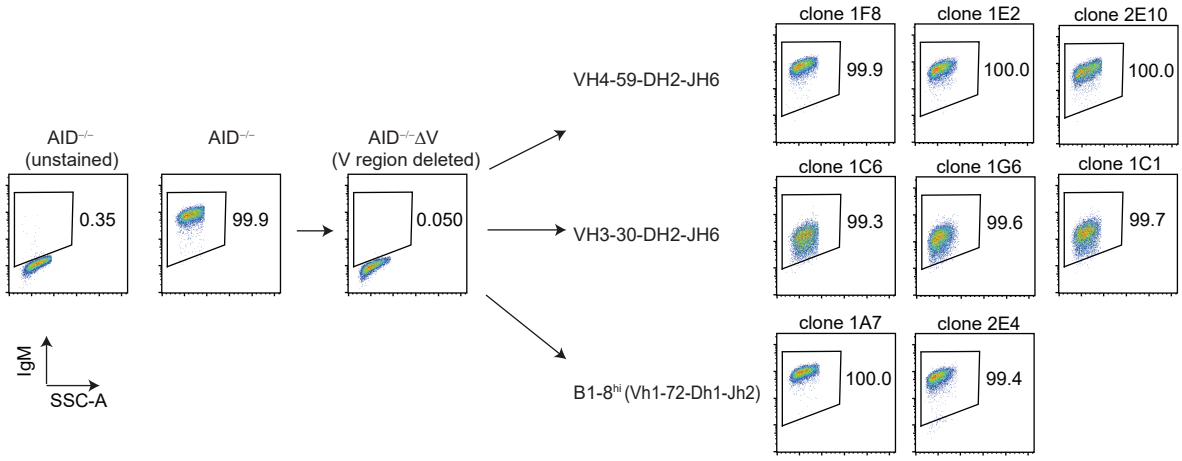
Figure S4. Profiling of the Ramos V region in AID$^{-/-}$ Eμ$^{-/-}$ cells using PRO-seq. (A) Analysis of nascent transcription 3' ends (PRO-seq) at the *IGH* locus in AID$^{-/-}$ and two independent clones of AID$^{-/-}$ Eμ$^{-/-}$ Ramos cells (c1, c2). The promoter (Prom.), complementary determining 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

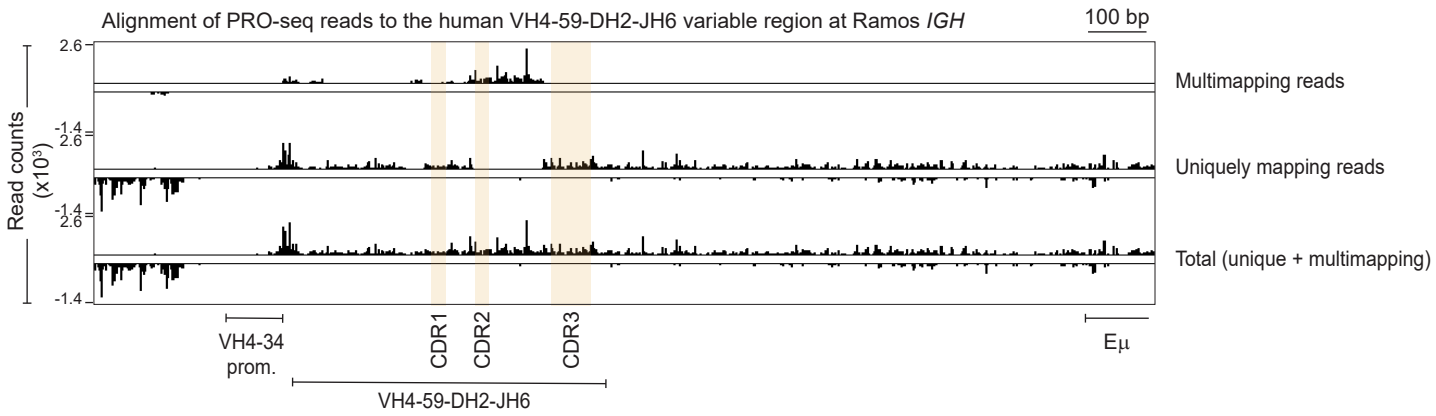
A



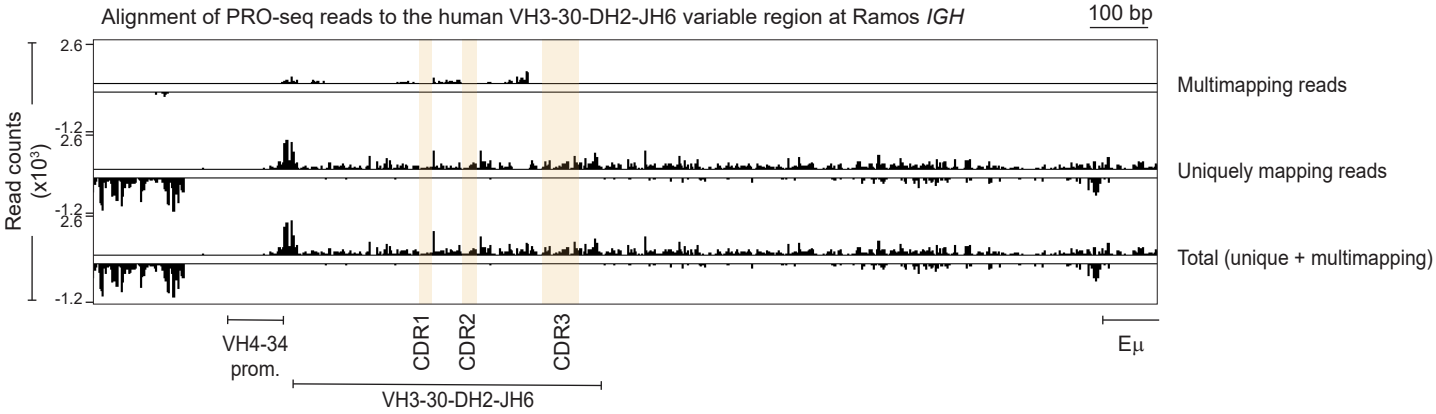
B



C



D



E

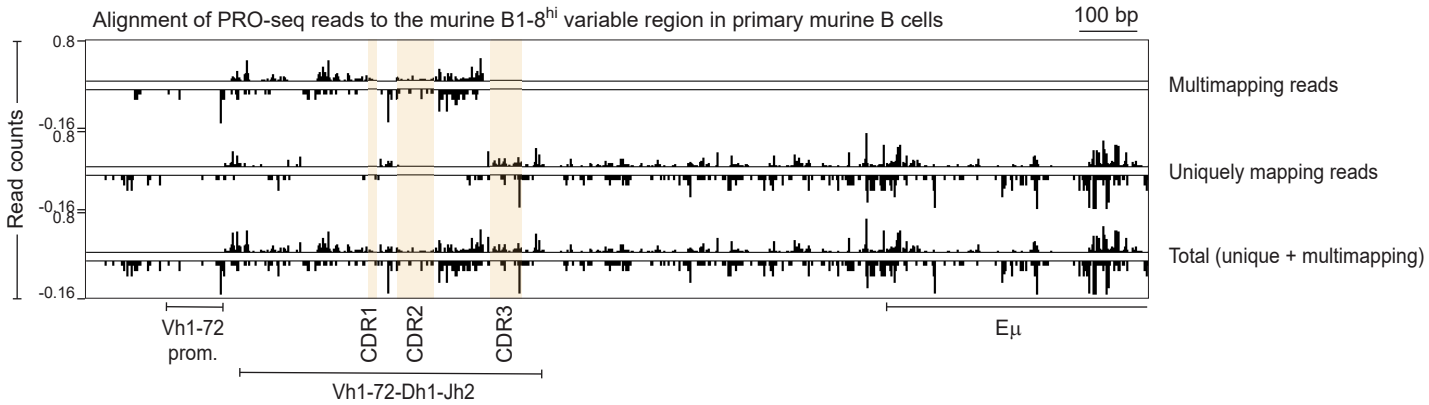
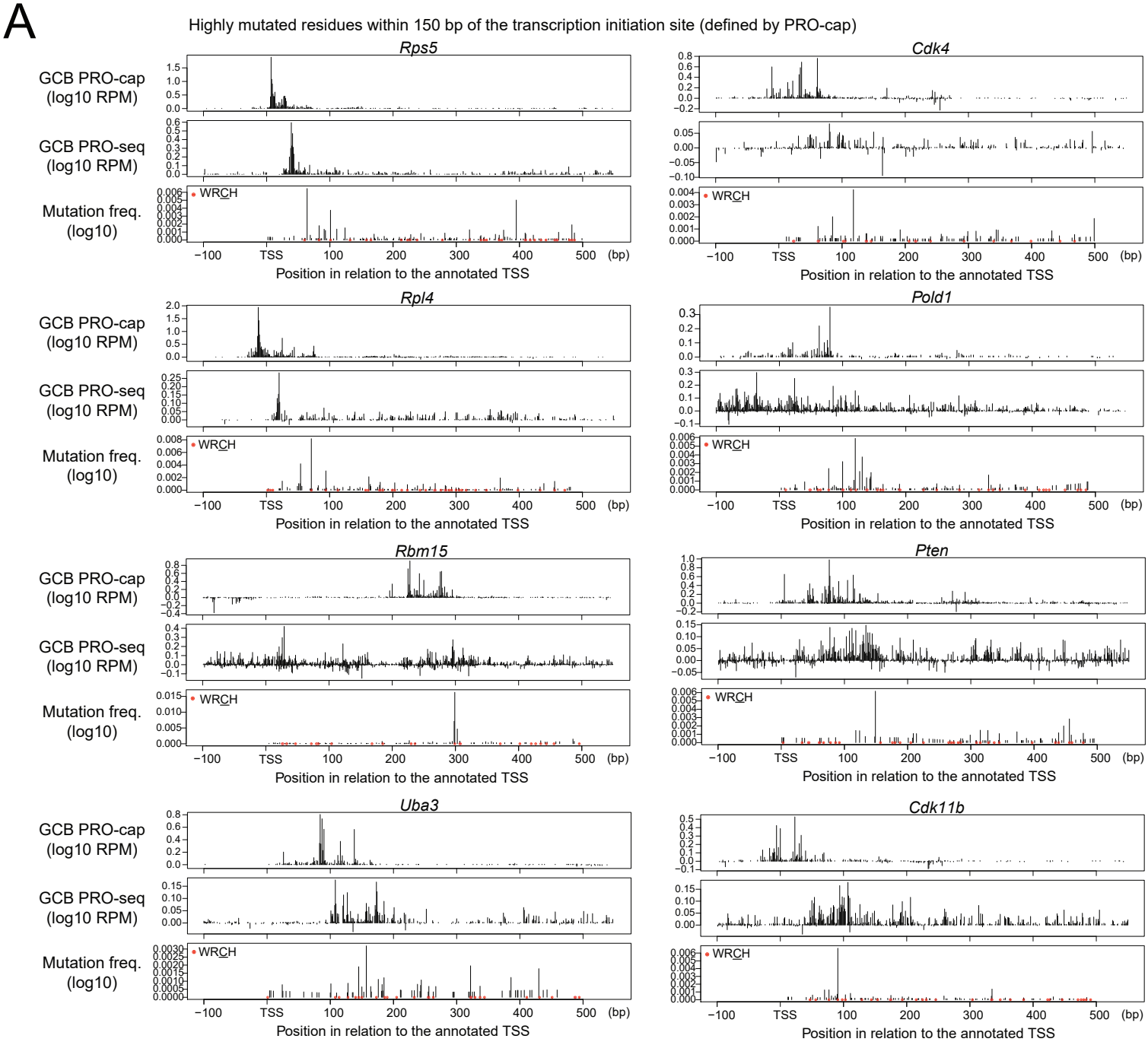


Figure S5: Generation of Ramos cell lines expressing exogenous V regions. (A) Schematic representation of the workflow for generating Ramos cells expressing new V regions using CRISPR-based editing. An IgM-negative line was made by excising the endogenous V region in $AID^{-/-}$ cells ($AID^{-/-}\Delta V$) 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

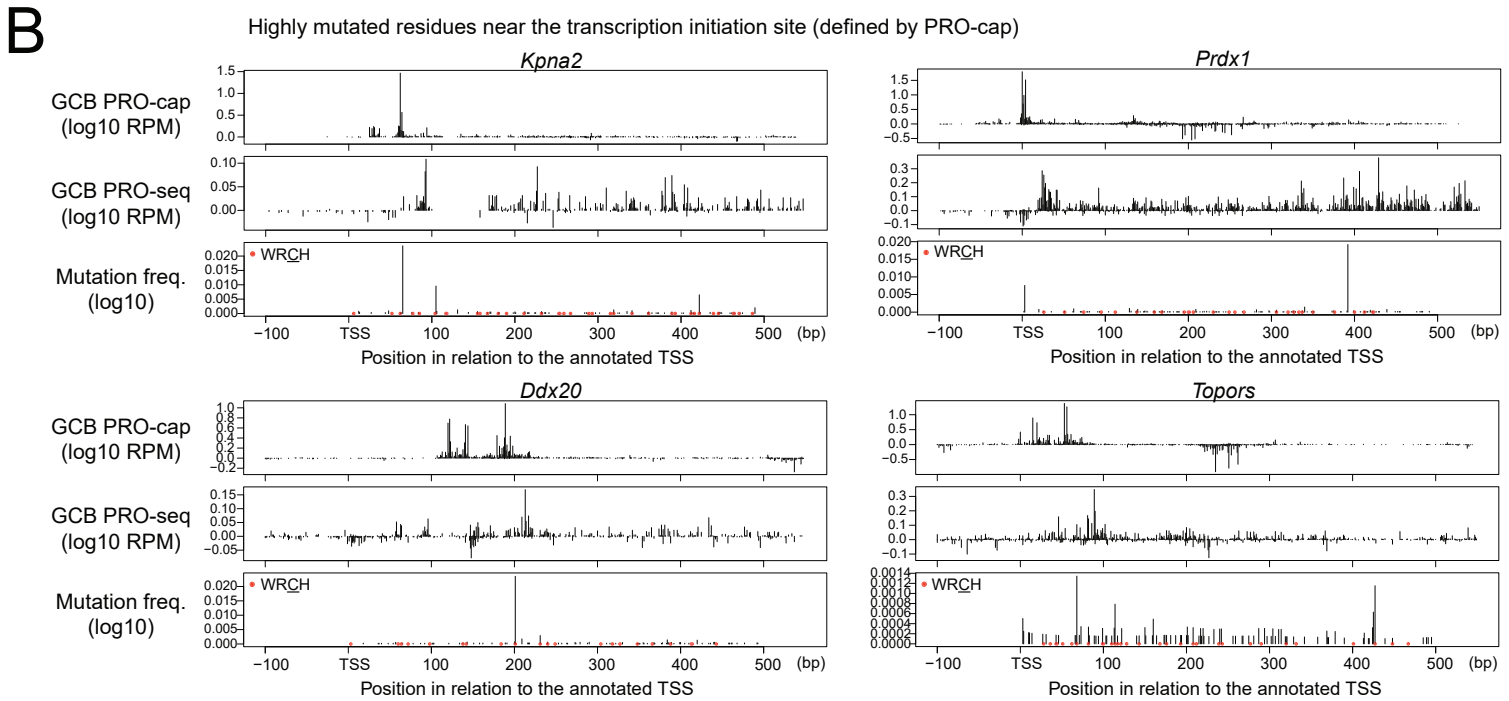

B
Highly mutated residues near the transcription initiation site (defined by PRO-cap)
Kpna2
Prdx1
Ddx20
Topors
GCB PRO-cap (log10 RPM)
GCB PRO-seq (log10 RPM)
Mutation freq. (log10)
WRCH
−100
TSS
100
200
300
400
500
(bp)
Position in relation to the annotated TSS

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

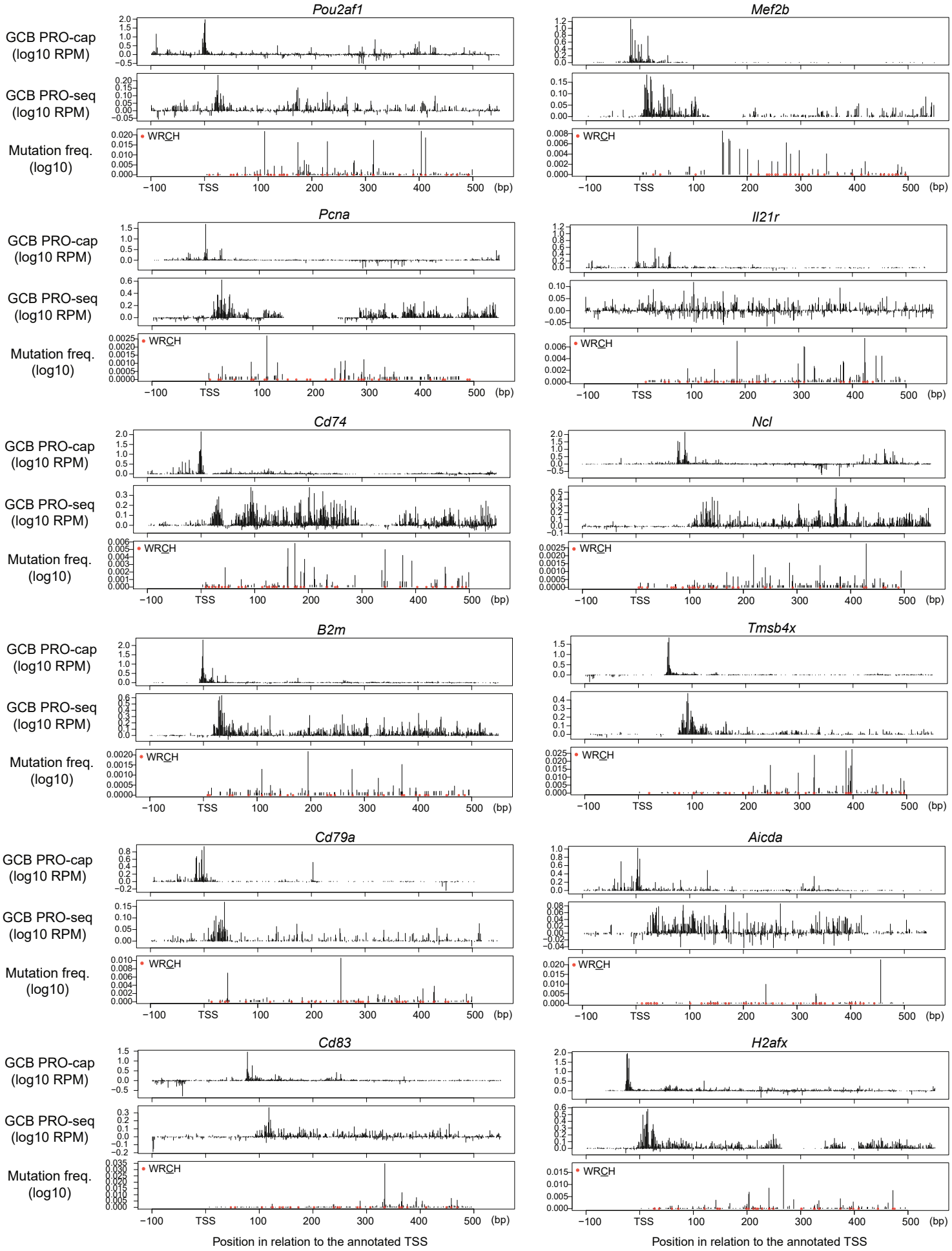

Pou2af1
Mef2b
Pcna
Il21r
Cd74
Ncl
B2m
Tmsb4x
Cd79a
Aicda
Cd83
H2afx
GCB PRO-cap (log10 RPM)
GCB PRO-seq (log10 RPM)
Mutation freq. (log10)
WRCH
−100
TSS
100
200
300
400
500
(bp)
Position in relation to the annotated TSS

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

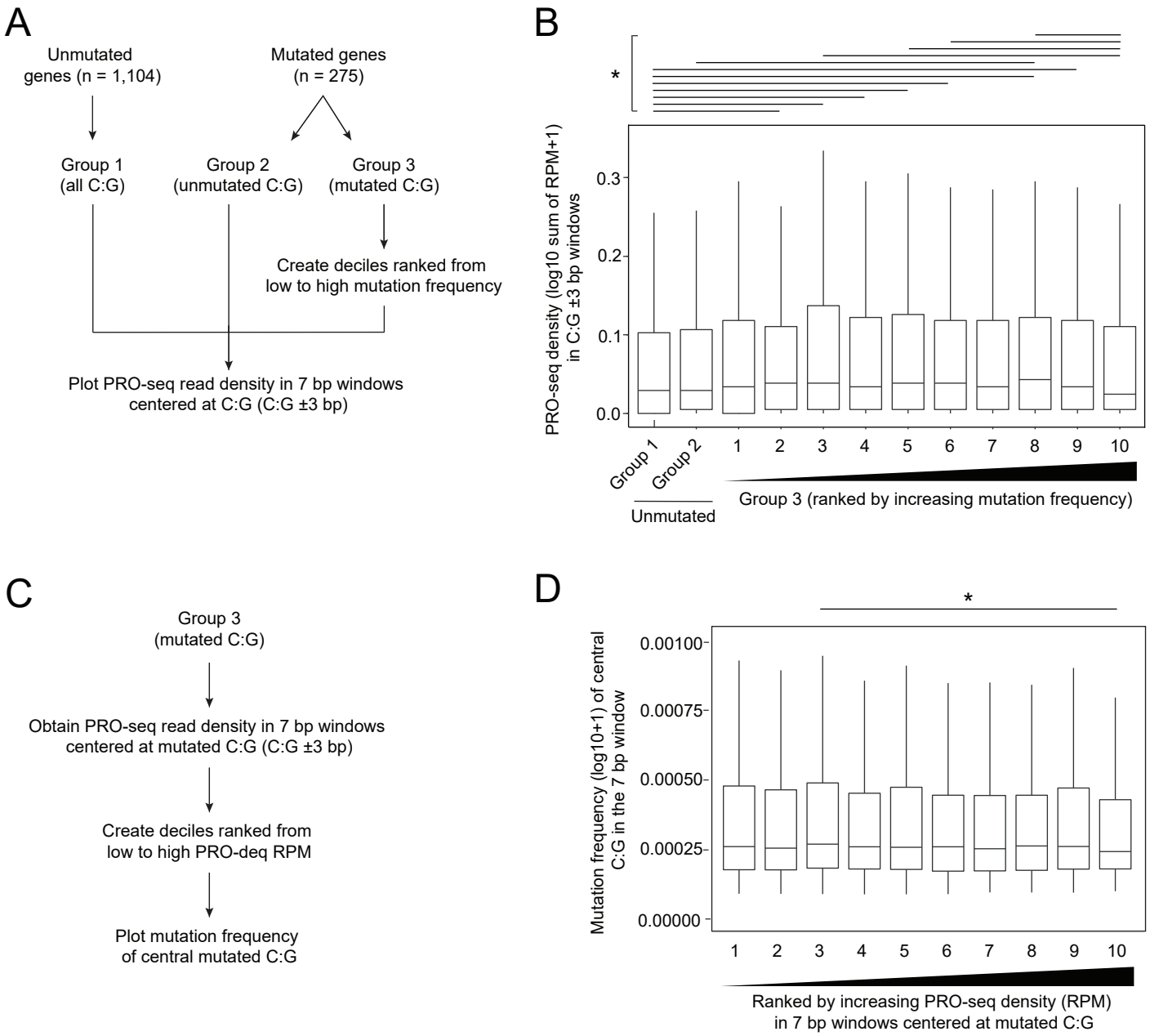


Figure S8. Relationship between SHM and transcriptional strength at non-IG AID target genes. (A) Workflow for the classification of mutated and unmutated genes (from ref. 17). **(B)** Box plot showing the PRO-seq density in 7 bp windows centered at each C:G residue in the three groups and sub-groups 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 E	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)	CTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNTCCACAGGT(	N is A,G,C,T randomly entered during primer synthesis
Round 1_FWD (VH4-34, VH4-59, V	CTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNTGTTCACAGC	N is A,G,C,T randomly entered during primer synthesis
Round 1_REV	CTGGAGTTCAGACGTGTGCTCTTCCGATCTCCTAGAGTGGCCATTCTTACC	
Round 2_FWD	AATGATACGGCGACCACCGAGATCTACACNNNNNNNNACACTCT	N represents 8 positions of individual sample barcode (IdxE501-E5xx)
Round 2_REV	CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTGGAGTTC/	N represents 6 positions of individual sample barcode (revIDX1-xx)

Table S2: Statistical analysis of mutation frequencies and PRO-seq densities in different V regions. The upper and lower tables show the *P* values for the analyses in Fig. S3D and S3F, respectively. *P* values are based on the Wilcoxon rank sum test following multiple testing correction with the Bonferroni method.

Mutation Freq.

Mutation freq. range	**Unmutated**	**(0.000112-0.000427**	**(0.000427-0.000657**	**(0.000657-0.000906**	**(0.000906-0.00133**	**(0.00133-0.00203**	**(0.00203-0.00321**	**(0.00321-0.00502**	**(0.00502-0.00818**	**(0.00818-0.0165**	**(0.0165-0.0603]**
Unmutated	1	0.851	0.794	0.936	0.871	0.962	1.014	0.875	0.81	0.937	0.857
[0.000112-0.000427		1	0.933	1.1	1.023	1.13	1.192	1.028	0.952	1.101	1.007
(0.000427-0.000657			1	1.179	1.097	1.211	1.277	1.101	1.021	1.18	1.079
(0.000657-0.000906				1	0.93	1.027	1.084	0.934	0.866	1.001	0.916
(0.000906-0.00133]					1	1.104	1.165	1.004	0.931	1.076	0.984
(0.00133-0.00203]						1	1.055	0.91	0.843	0.974	0.891
(0.00203-0.00321]							1	0.862	0.799	0.924	0.845
(0.00321-0.00502]								1	0.927	1.071	0.98
(0.00502-0.00818]									1	1.156	1.058
(0.00818-0.0165]										1	0.915
(0.0165-0.0603]											1

PRO-seq RPM

PRO-seq RPM range	**[3-11.9]**	**(11.9-17.2]**	**(17.2-22.1]**	**(22.1-27]**	**(27-32]**	**(32-35]**	**(35-43]**	**(43-49.3]**	**(49.3-59.8]**	**(59.8-94.5]**
[3-11.9]	1	0.882	0.277	0.512	0.573	0.656	0.414	0.49	0.544	0.415
(11.9-17.2]		1	0.315	0.58	0.65	0.744	0.469	0.556	0.617	0.471
(17.2-22.1]			1	1.845	2.067	2.365	1.491	1.766	1.96	1.496
(22.1-27]				1	1.12	1.282	0.808	0.957	1.062	0.811
(27-32]					1	1.144	0.722	0.854	0.948	0.724
(32-35]						1	0.631	0.747	0.829	0.632
(35-43]							1	1.184	1.314	1.003
(43-49.3]								1	1.11	0.847
(49.3-59.8]									1	0.763
(59.8-94.5]										1