

# A broad atlas of somatic hypermutation allows prediction of activation-induced deaminase targets

Ángel F. Álvarez-Prado,<sup>1</sup> Pablo Pérez-Durán,<sup>1</sup> Arantxa Pérez-García,<sup>1</sup> Alberto Benguria,<sup>2</sup> Carlos Torroja,<sup>3</sup> Virginia G. de Yébenes,<sup>1</sup> and Almudena R. Ramiro<sup>1</sup>

<sup>1</sup>B Cell Biology Lab, <sup>2</sup>Genomics Unit, and <sup>3</sup>Bioinformatics Unit, Centro Nacional de Investigaciones Cardiovasculares, Madrid, Spain

**Activation-induced deaminase (AID) initiates antibody diversification in germinal center (GC) B cells through the deamination of cytosines on immunoglobulin genes. AID can also target other regions in the genome, triggering mutations or chromosome translocations, with major implications for oncogenic transformation. However, understanding the specificity of AID has proved extremely challenging. We have sequenced at very high depth >1,500 genomic regions from GC B cells and identified 275 genes targeted by AID, including 30 of the previously known 35 AID targets. We have also identified the most highly mutated hotspot for AID activity described to date. Furthermore, integrative analysis of the molecular features of mutated genes coupled to machine learning has produced a powerful predictive tool for AID targets. We also have found that base excision repair and mismatch repair back up each other to faithfully repair AID-induced lesions. Finally, our data establish a novel link between AID mutagenic activity and lymphomagenesis.**

## INTRODUCTION

Activation-induced deaminase (AID) is a crucial enzyme for the immune response because it generates high-affinity and switched antibodies in germinal center (GC) B cells by somatic hypermutation (SHM) and class switch recombination (CSR; Muramatsu et al., 2000; Revy et al., 2000). AID initiates SHM and CSR through the deamination of deoxycytidine residues into deoxyuridines on the DNA of Ig genes (Muramatsu et al., 2000; Petersen-Mahrt et al., 2002; Di Noia and Neuberger, 2007; Stavnezer et al., 2008). The resulting U:G mismatch can be alternatively recognized and processed by base excision repair (BER) or mismatch repair (MMR) pathways, leading either to point mutations, in the case of SHM, or to double-strand breaks (DSBs) followed by a recombination reaction, in the case of CSR (Di Noia and Neuberger, 2007; Stavnezer et al., 2008; Reynaud et al., 2009; Methot and Di Noia, 2017). Although AID activity has a strong preference for Ig genes, it can also target other genes, giving rise to point mutations (Shen et al., 1998; Pasqualucci et al., 2001; Liu et al., 2008) or oncogenic chromosome translocations (TCs; Ramiro et al., 2004, 2006; Robbiani et al., 2008). Understanding AID specificity, or targeting, has been hindered by the technical challenge of detecting AID-induced mutations, which occur at very low frequencies. Here, we have used next generation sequencing to directly measure raw AID mutational activity on a broad representation of the genome and thus gather conclusions on AID specificity, DNA repair, and lymphomagenesis.

## RESULTS AND DISCUSSION

### Capture-based deep sequencing allows high-throughput identification of AID targets

To explore the scope of AID-induced mutations at a high-throughput scale, we designed a capture library against 1,588 regions corresponding to 1,379 different genes as a representation of the B cell genome (Table S1; see Design of DNA capture library in the Materials and methods). Genomic DNA from GC B cells was isolated, captured, and deep sequenced (Fig. S1, A and B). We made use of a mouse model deficient for both BER and MMR pathways (*Ung*<sup>-/-</sup>*Msh2*<sup>-/-</sup> mice). In the absence of BER and MMR, AID-induced U:G mismatches remained unprocessed and were replicated over, thus leaving behind almost solely C→T and G→A transitions, the footprint of AID deamination events on DNA (Rada et al., 2004; Methot and Di Noia, 2017). This approach allowed an extremely efficient enrichment and sequencing depth (Fig. S1, A and B). We found a set of 291 genomic regions (corresponding to 275 different genes) that were reproducibly mutated in *Ung*<sup>-/-</sup>*Msh2*<sup>-/-</sup> GC B cells when compared with *Aicda*<sup>-/-</sup> GC B cells ( $q \leq 0.05$ ; Fig. 1 A; Fig. S1, C–E; and Table S2; representative targets were validated by Sanger sequencing; Fig. 1 B and Table S3). Importantly, the 275-gene target collection included 30 of the 35 previously known AID targets, such as *Bcl6*, *Pim1*, *RhoH*, *Pax5*, and *Cd83* (Fig. 1 C and Table S2; Pasqualucci et al., 2001; Liu et al., 2008; Methot and Di Noia, 2017). Mutations detected in the 291-target regions strongly accumulated in AID mutational hotspots (WRC(Y)/(R)GYW; underlined letters specify deaminated nucleotides; W = A/T; R = A/G; Y = C/T; Fig. 1 D; Rogozin and Kolchanov, 1992). Finally, we found that our 275-target set in-

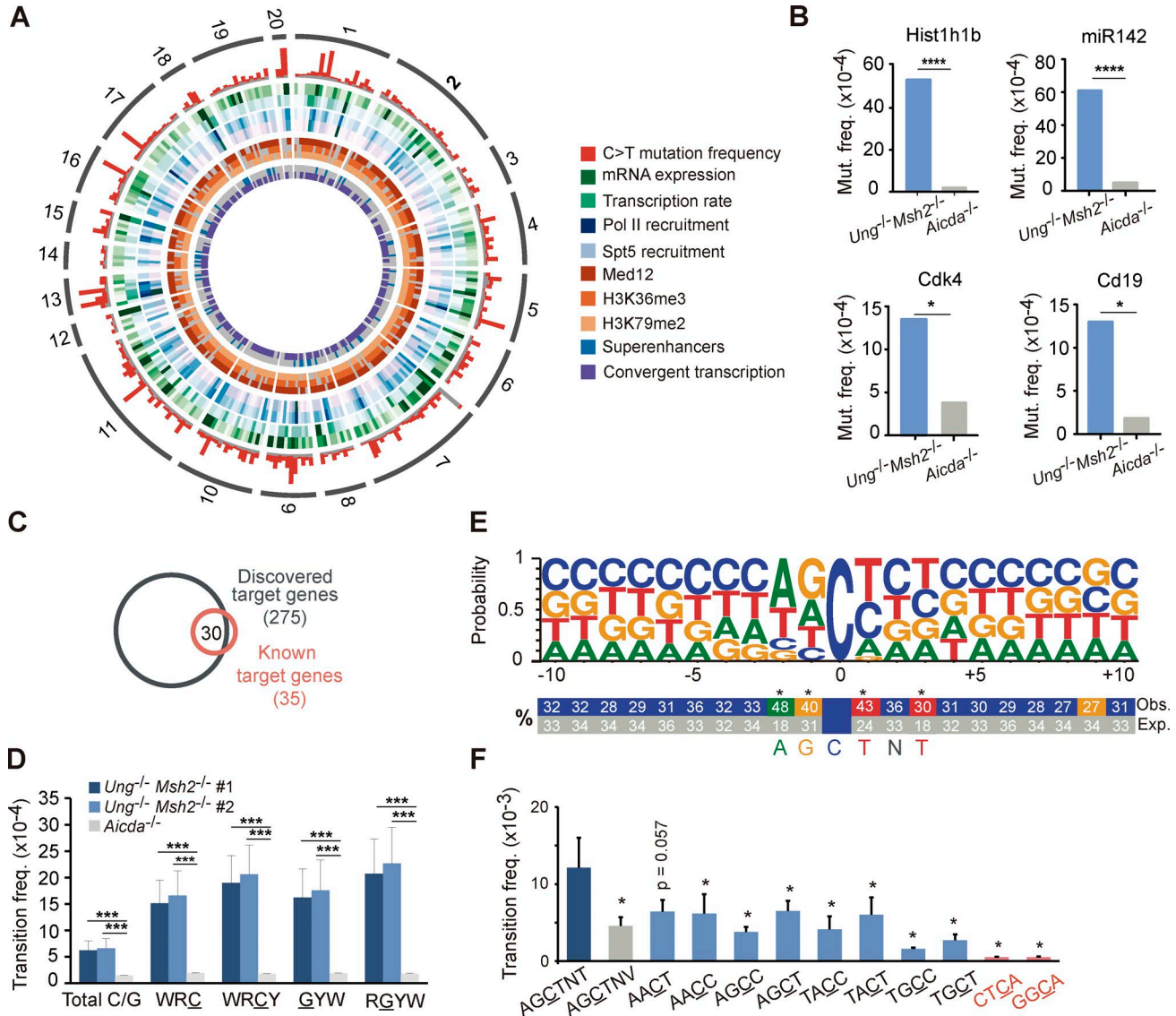
Correspondence to Almudena R. Ramiro: [aramiro@cnic.es](mailto:aramiro@cnic.es)

P. Pérez-Durán's present address is Institute for Cancer Genetics, Columbia University, New York, NY.

A. Pérez-García's present address is Beatson Institute for Cancer Research, Glasgow, Scotland, UK.

© 2018 Álvarez-Prado et al. This article is available under a Creative Commons License (Attribution 4.0 International, as described at <https://creativecommons.org/licenses/by/4.0/>).





**Figure 1. High-throughput analysis of AID-induced mutations.** DNA from Peyer's patch GC B cells was captured with a probe library for 1,588 genomic regions (Table S1) and deep sequenced. AID targets were identified as those regions accumulating significantly more C→T transition mutations in *Ung<sup>-/-</sup>Msh2<sup>-/-</sup>* than in *Aicda<sup>-/-</sup>* mice (Table S2; FDR ≤0.05, one-tail Fisher test and Benjamini-Hochberg correction; two independent experiments; see Materials and methods). **(A)** Circos plot representation of the AID targets identified in this study and their associated molecular features. The outer ring shows chromosome location and is followed by C→T transition mutation frequency in *Ung<sup>-/-</sup>Msh2<sup>-/-</sup>* (red) and *Aicda<sup>-/-</sup>* (gray) mice. **(B)** Validation of representative AID targets by Sanger sequencing (one-tail Fisher test; Table S3). **(C)** Overlap between the targets discovered in this study and previously reported AID targets. **(D)** Mean transition frequency in total C/G nucleotides and in C/G within WRC(Y)/(R)GYW hotspots (W = A/T; R = G/A; Y = C/T) of the 291 AID targets (two-tailed Student's *t* test; two independent experiments). **(E)** Logo representation of the sequence context of mutated cytosines (mutation frequency ≥4 × 10<sup>-3</sup>). Statistically significant enrichment of nucleotides surrounding the mutated C is indicated (\*, FDR ≤10<sup>-3</sup>, one-tail Fisher test and Bonferroni correction; see Materials and methods), and numbers indicate percentages. **(F)** Mean mutation frequency of cytosines within the indicated motifs (dark blue bar, newly identified hotspot; gray bar, control motif for newly identified hotspot; light blue bars, WRCY hotspots; red bars, random four-nucleotide motifs; two-tailed Mann-Whitney test). \*, P ≤ 0.05; \*\*\*, P < 10<sup>-3</sup>; \*\*\*\*, P < 10<sup>-4</sup>. Error bars depict SEM.

cluded a big proportion of genes subject to DSBs or chromosome TCs (Fig. S1 F; Chiarle et al., 2011; Klein et al., 2011; Staszewski et al., 2011; Qian et al., 2014; Dong et al., 2015). Thus, our deep sequencing approach has allowed the discovery of an unprecedented, massive collection of AID targets.

#### Identification of AGCTNT as a novel AID hotspot

To gain insights into the local sequence preference of AID, we first analyzed the mean mutation frequency at individual WRCY/RGYW hotspots across all 291 AID targets and found a wide range of mutability, with AACT and AGCT

as the top mutated hotspots in both strands of DNA, which may reflect an intrinsic preference for AID deaminase activity. Next, we performed an unbiased analysis of the sequence context of mutated cytosines. We found that A, G, and T nucleotides were the preferred nucleotides at -2, -1, and +1 positions (Pérez-Durán et al., 2012; Wei et al., 2015; Yeap et al., 2015), respectively, but we further uncovered a significant preference for T at +3 (Fig. 1 E and Fig. S2). Indeed, cytosines lying at the AGCTNT motif were significantly more mutated than those in AGCTNV (where V is A, C, or G) or than other WRCY/RGYW hotspots (Fig. 1 F and Fig. S2, A and B). Thus, our study has revealed AGCTNT as a novel and the most highly mutated AID hotspot identified so far.

### Prediction of AID targets

Using the uniquely large set of AID-mutated genes identified in this study, we performed a comprehensive analysis of molecular features that associate with SHM, including transcription, epigenetic marks, and regulatory sequences (Fig. 1 A; Storb, 2014; Methot and Di Noia, 2017). We first observed that transcription levels and transcription rates are significantly higher in AID targets than in nontargets and that this difference is even higher for highly mutated targets (Fig. 2 A). We also found that RNAPolIII and the stalling factor Spt5, previously described to associate with AID (Nambu et al., 2003; Pavri et al., 2010), show higher binding density within AID mutational targets (Fig. 2 B). Likewise, AID targets were enriched in marks of active enhancers and transcriptional elongation, such as Med12, H3K36me3, and H3K79me2 (Fig. 2 C). Finally, we found that primary AID targeting, as measured by AID mutations in the absence of repair, also focuses preferentially in the vicinity of superenhancers (Fig. 2 D) and in regions subject to convergent transcription (Fig. 2 E; Meng et al., 2014; Qian et al., 2014). Together, our mutagenesis study shows that several mechanisms linked to transcription are critical for AID activity, as suggested in previous studies (Nambu et al., 2003; Pavri et al., 2010; Meng et al., 2014; Qian et al., 2014; Wang et al., 2014). Our data also indicate that AID targeting cannot be defined by any of these features alone. To approach whether a combination of these molecular features could be used to predict AID targeting, we developed a prediction model using a machine-learning algorithm, fed with the collection of genes analyzed here together with the set of molecular features described in Fig. 2 (A–E) (Fig. S3, A and B; see Machine learning to predict AID targets in the Materials and methods for details). We found that a combination of high-density RNAPolIII and Spt5 binding, found in 2.3% of genes in the whole genome (Fig. S3 B), predicts AID specificity with 77% probability ( $P < 0.001$ ; Fig. 2 F and Fig. S3 A). Conversely, low RNAPolIII binding combined with low gene expression predicted the absence of mutations for 95% of genes (Fig. 2 F). To test the accuracy of our prediction model, we analyzed the mutation frequency of a new collection of genes (not included in our capture library) with high-density RNAPolIII and Spt5 binding (Fig. S3 C

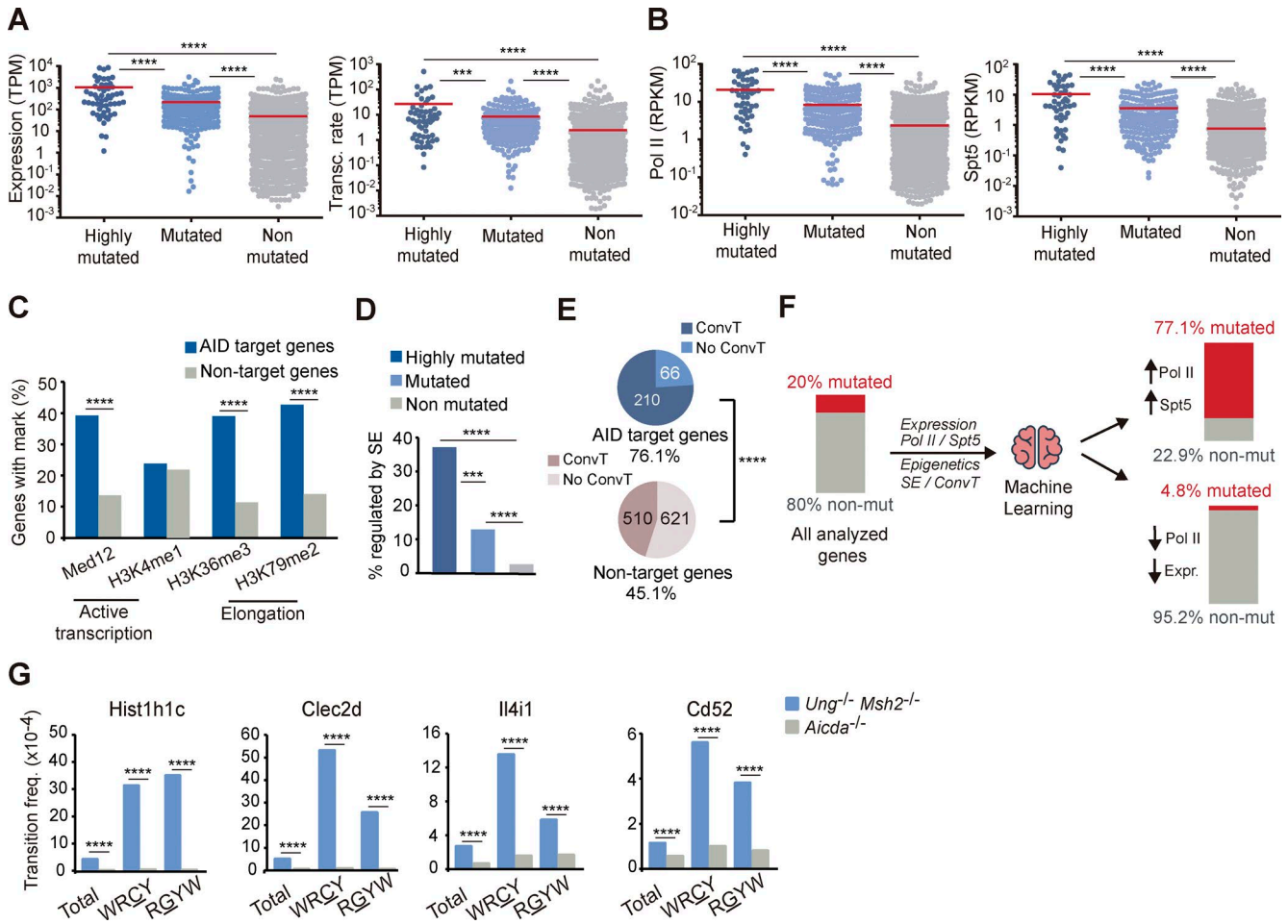
and Table S4). We found that 11/12 of the analyzed genes were significantly mutated (Table S4 and Fig. 2 G). Indeed, two genes (*Hist1h1c* and *Clec2d*) were mutated at the range of the top 20% mutated genes at frequencies similar to those found in *Pax5* or *Rhoh* (Table S2 and Table S4). Thus, we have built a powerful predictive tool for AID activity.

### BER and MMR back up each other to faithfully repair AID-induced lesions

BER and MMR act downstream of AID-induced U:G mismatches so that UNG is critical for the generation of transversions at C:G pairs while MSH2 facilitates the introduction of mutations at A:T pairs (Frey et al., 1998; Phung et al., 1998; Rada et al., 1998, 2002, 2004; Methot and Di Noia, 2017). UNG and MSH2 can also promote conventional, faithful repair of AID-induced U:G mismatches (Liu et al., 2008; Pérez-Durán et al., 2012). To explore the contribution of BER and MMR to AID mutagenic activity, we analyzed GC B cells from single-deficient *Ung*<sup>+/-</sup>*Msh2*<sup>-/-</sup> and *Ung*<sup>-/-</sup>*Msh2*<sup>+/-</sup> mice and from control *Ung*<sup>+/-</sup>*Msh2*<sup>+/-</sup> mice and compared the mutation frequency of the 291 AID target regions identified in this study (Table S2). We found similar mean mutation frequencies in B cells deficient for UNG alone, MSH2 alone, or proficient for both, whereas AID targets harbored significantly more mutations in the combined absence of UNG and MSH2 (Fig. 3, A and B). Indeed, only a small proportion (~6%) of the genes mutated in *Ung*<sup>-/-</sup>*Msh2*<sup>-/-</sup> cells was also mutated in single-knockout and double-heterozygous cells (Fig. 3 C and Table S2). Moreover, we found that classical AID off targets, such as *Bcl6* or *Pim1*, although mutated in all genotypes analyzed, harbored a significantly bigger load of mutations in *Ung*<sup>-/-</sup>*Msh2*<sup>-/-</sup> cells than in *Ung*<sup>+/-</sup>*Msh2*<sup>-/-</sup>, *Ung*<sup>-/-</sup>*Msh2*<sup>+/-</sup>, or *Ung*<sup>+/-</sup>*Msh2*<sup>+/-</sup> cells (Fig. 3 D). Together, these data indicate that BER and MMR back up each other to faithfully repair most of the AID-induced lesions in GC B cells.

### AID targets are recurrently mutated in human lymphomas

We next assessed the contribution of AID off-target mutations to B cell-derived malignancies by making use of available sequencing data on human lymphomas. We found that AID targets are significantly enriched in genes mutated in human B cell lymphomas (see Annotation of AID targets in the Materials and methods for details; Fig. 4 A). Indeed, 21/275 (7.6%) of our set of AID target genes are mutated in diffuse large B cell lymphomas (DLBCLs; Fig. 4 B), a highly prevalent, aggressive form of lymphoma (Shaffer et al., 2012). Lymphoma genes mutated by AID included *Bcl6*, *RhoH*, *Pim1*, *Ebfl1*, *Eif4a2*, and *Pax5*, which is in agreement with previous studies (Shen et al., 1998; Pasqualucci et al., 2001; Liu et al., 2008). In addition, we identified nine novel genes mutated in human DLBCLs that accumulate AID-induced mutations (Fig. 4 B), including *Mef2b*, *Lyn*, *Tnfrsf3*, *Gna13*, and *Irf8*. Remarkably, we found many instances where the exact same mutations described in human lymphoma genes



**Figure 2. Molecular features of AID targets predict mutability.** (A) Expression level of highly mutated (top 20% mutated genes, C→T transition frequency  $>3 \times 10^{-4}$ ), mutated (rest of mutated), and nonmutated genes in Peyer's patch GC B cells as measured by RNA-Seq and transcription rate of AID targets in GC B cells from lymph nodes as measured by GRO-Seq. TPM, transcripts per million. (B) Recruitment of RNAPolIII and Spt5 to AID targets and nontargets measured in in vitro activated splenic B cells by ChIP-Seq. RPKM, reads per kilobase per million reads mapped. (C) Transcription and transcription elongation marks in AID targets and nontargets by ChIP-Seq analysis of in vitro activated splenic B cells (Med12, H3K4me1, H3K36me3, and H3K79me2). (D) Proportion of highly mutated, mutated, and nonmutated genes regulated by superenhancers (SE) in GC B cells (see Materials and methods). (E) GRO-Seq analysis of convergent transcription (ConvT) in AID targets and nontargets from GC splenic B cells obtained from SRBC-immunized mice. (F) Representation of the machine-learning approach used for AID target prediction. (G) Validation of representative genes predicted to be mutated by the model by PCR-Seq. Statistical tests: two-tailed Student's *t* test (A, B, and G) and one-tailed Fisher test (C–E). \*\*\*,  $P < 10^{-3}$ ; \*\*\*\*,  $P < 10^{-4}$ .

were also found in the AID targets identified in this study in nontransformed mouse B cells (Fig. 4 C and Table S5). Together, these results suggest that off-target AID mutagenic activity can contribute to GC-associated lymphomagenesis.

Until now, the study of AID specificity has been hindered by the technical challenge of detecting AID-induced mutations; indeed, only a limited number of genes has been directly interrogated for AID-mediated mutagenesis (Pasquucci et al., 2001; Liu et al., 2008; Methot and Di Noia, 2017). However, genome-wide AID specificity has been inferred from high-throughput analysis of AID binding, which does not warrant AID activity, AID-induced DSBs, or chromosomal TCs, which involve complex processing of the initial lesion induced by AID (Chiarle et al., 2011; Klein et al., 2011;

Staszewski et al., 2011; Yamane et al., 2011; Meng et al., 2014; Qian et al., 2014). The strategy developed in this study has provided an unprecedented scope to the analysis of AID targeting: we describe here the broadest collection of AID mutational targets (275 genes) to date, 10-fold larger than the previously known targets. The strength of this analysis is well supported by the confirmation of the vast majority of previously identified AID targets and the validation of targets by conventional Sanger sequencing.

Here, we have integrated our mutation data with a collection of molecular features of GC B cells to feed a machine-learning algorithm. According to the machine-learning tree generated here, the combined binding of Spt5 and RNAPolIII at high density is the best predictor for AID mutability, although

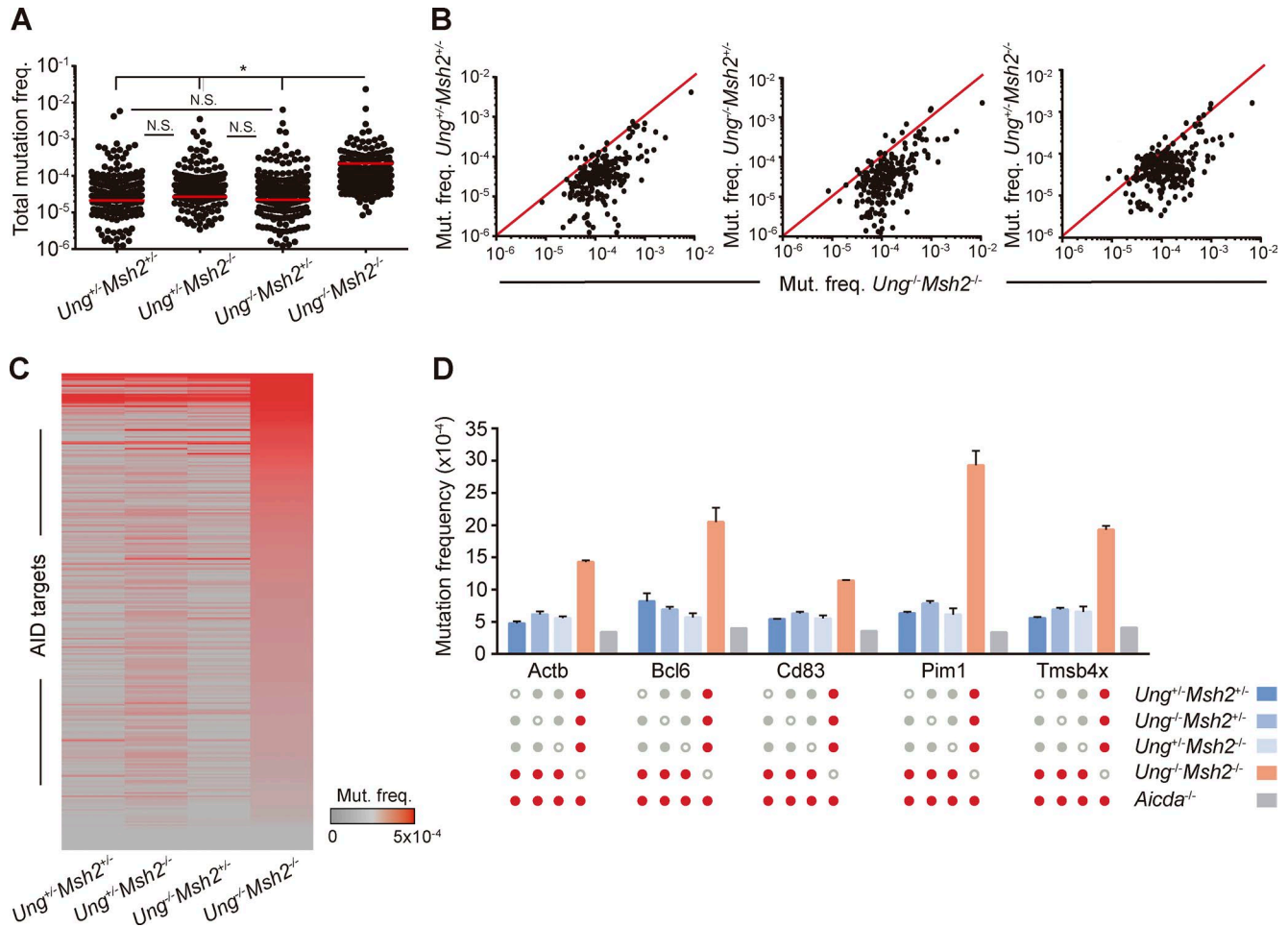
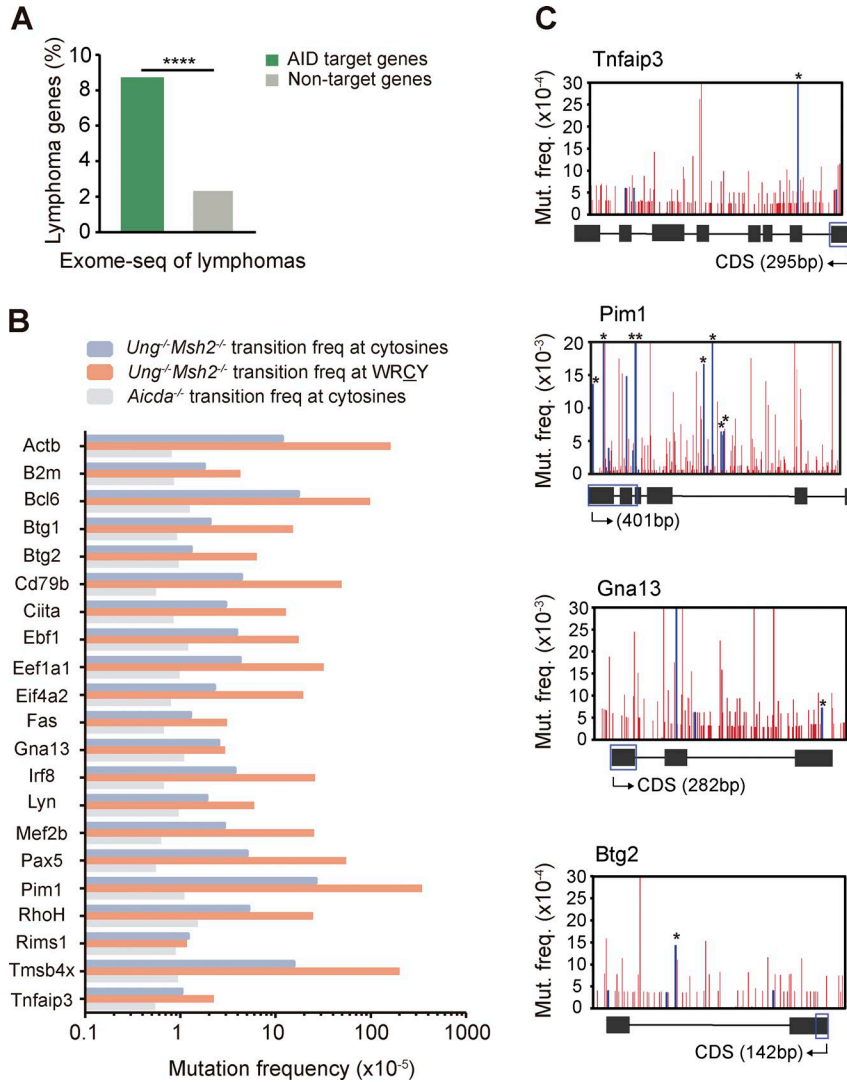


Figure 3. **BER and MMR back up each other to error-free repair AID-induced lesions. (A and B)** Total mutation frequency of AID targets in  $Ung^{+/+}Msh2^{+/+}$ ,  $Ung^{-/-}Msh2^{+/+}$ , and  $Ung^{+/+}Msh2^{-/-}$  GC B cell mice compared with that of  $Ung^{-/-}Msh2^{-/-}$  mice (mean of two independent experiments; see Materials and methods; Table S2). **(C)** Heat map representation of AID targets in  $Ung^{+/+}Msh2^{+/+}$ ,  $Ung^{-/-}Msh2^{+/+}$ ,  $Ung^{+/+}Msh2^{-/-}$ , and  $Ung^{-/-}Msh2^{-/-}$  GC B cells. **(D)** Mutation frequency of representative genes in  $Ung^{+/+}Msh2^{+/+}$ ,  $Ung^{-/-}Msh2^{+/+}$ ,  $Ung^{+/+}Msh2^{-/-}$ ,  $Ung^{-/-}Msh2^{-/-}$ , and  $Aicda^{-/-}$  GC B cells. Red dots indicate statistically different mutation frequencies between the indicated genotypes. Mutation frequency found in  $Aicda^{-/-}$  mice was subtracted before plotting A–C. (A and D) Two-tailed Student's *t* test; \*,  $P \leq 0.05$ . Error bars depict SEM. N.S., not significant.

additional combinations of transcriptional traits bear some predictive power as well. Furthermore, we have performed independent experimental validation showing that randomly picked  $Spt5^{high}RNAPolIII^{high}$  genes indeed are very frequently mutated by AID. This is, to our knowledge, the first instance of a tool that successfully predicts the potential of a gene to be targeted by AID. Regarding the fate of AID-induced lesions, BER and MMR have long been known to broaden the diversity of SHM with an apparent perverted recruitment of error-prone polymerases and to do so in a cooperative manner (Rada et al., 2004; Di Noia and Neuberger, 2007; Methot and Di Noia, 2017). The mechanisms responsible for the error-free versus error-prone activity of UNG and MSH2 are far from understood, and both gene-specific and local sequence contexts may play a role in defining the fate of the U:G resolution (Liu et al., 2008; Pérez-Durán et al., 2012; Wei et al., 2015). Strik-

ingly, here we show that the fate of the majority of off-target lesions induced by AID is to undergo faithful repair by BER and MMR and that, again, both pathways can back up each other in this task with only a minor fraction of the mutations escaping them. Whether this reflects gene-specific qualities or is the consequence of excessive mutation load will deserve further investigation. We would speculate that a minor fraction of unrepaired mutations in prolymphomagenic genes could provide cell growth advantage and account for the predominance of AID-mediated mutations in lymphomas. Regardless of oncogenic relevance, it is remarkable that even though our mutation analysis was performed in nontransformed cells, we could detect individual AID-induced mutations that are recurrently mutated in lymphoma. Thus, our results yield a novel perspective on the contribution of AID activity to B cell transformation through the introduction of mutations.



**Figure 4. AID targets are recurrently mutated in human lymphomas.** (A) AID targets are enriched in genes involved in lymphoma development. Percentage of lymphoma genes within AID target and nontarget genes. Annotation was done from public data on human lymphoma sequencing (see Materials and methods; two-tailed Fisher test; \*\*\*\*,  $P < 10^{-4}$ ). (B) Mutation frequency in total C/G nucleotides and C/G nucleotides within WRC(Y)/(R)GYW hotspots (W = A/T; R = G/A; Y = C/T) of the 21 AID target genes involved in human DLBCL development analyzed in *Ung*<sup>-/-</sup>*Msh2*<sup>-/-</sup> mice (mean of two independent experiments; see Materials and methods). (C) Mutation profiles of representative DLBCL genes analyzed in *Ung*<sup>-/-</sup>*Msh2*<sup>-/-</sup> mice. Blue bars indicate mutations identical to those found in human lymphoma tumor samples (Table S5); asterisks indicate mutations occurring in a WRC(Y) hotspot. The diagrams below the graphs represent the complete gene (not to scale), and blue boxes indicate the region depicted above. Mutation frequency found in each nucleotide in *Aicda*<sup>-/-</sup> mice was subtracted before plotting.

We expect our mutational study will be valuable for other research questions, including validation of novel molecular mechanisms involved in AID targeting, prediction of novel targets, or assessment of cancer-associated mutations. Furthermore, similar approaches would be of immediate interest to broaden our knowledge on the role of AID or other mutagenic activities not only in B cell lymphomas, but also in malignancies from any origin.

## MATERIALS AND METHODS

### Mice

*Ung* and *Msh2* mutant mice used in this study were generated by crossing *Ung*<sup>-/-</sup> mice (Nilsen et al., 2000) and *Msh2*<sup>-/-</sup> mice (Reitmair et al., 1995). *Aicda*<sup>-/-</sup> mice have been previously described (Muramatsu et al., 2000). Mice were housed in specific pathogen-free conditions. Male and female mice between 20 and 28 wk were used for the experiments. The number of animals per group to detect biologically significant effect sizes was calculated using an appropriate statistical

sample size formula. All experiments were done in concordance with EU Directive 2010/63EU and Recommendation 2007/526/EC regarding the protection of animals used for experimental and other scientific purposes, enforced in Spanish law under RD 53/2013.

### Design of DNA capture library

A set of 1,379 mouse genes was selected as a representation of the genome (Table S1). 85% of all genes were randomly picked, ensuring even representation of chromosomal location by bioinformatic analysis and unbiased biological function. ~15% of the library corresponded to previously known AID targets (Müschen et al., 2000; Pasqualucci et al., 2001; Gordon et al., 2003; Liu et al., 2008; Robbiani et al., 2009; Pavri et al., 2010), IgH probes, and other controls. Probes were designed in eArray (Agilent) to capture the first 500 bp downstream of each transcriptional start site (TSS) of each of the 1,379 genes. Because various genes contained more than one predicted TSS, the library includes a total of 1,588 dif-

ferent genomic regions. Library design included 50 extra nucleotides at both ends of each region to optimize the capture yield. A custom target enrichment capture library was then synthesized by the manufacturer (SureSelectXT; Agilent).

### DNA capture and sequencing

GC (*Cd19<sup>+</sup>Fas<sup>+</sup>GL7<sup>+</sup>*) B cells were isolated from Peyer's patches of *Ung<sup>+/-</sup>Msh2<sup>+/-</sup>* ( $n_1 = 10; n_2 = 11$ ), *Ung<sup>-/-</sup>Msh2<sup>+/-</sup>* ( $n_1 = 46; n_2 = 8$ ), *Ung<sup>+/-</sup>Msh2<sup>-/-</sup>* ( $n_1 = 46; n_2 = 2$ ), and *Ung<sup>-/-</sup>Msh2<sup>-/-</sup>* ( $n_1 = 37; n_2 = 8$ ) littermates and *Aicda<sup>-/-</sup>* ( $n = 31$  mice) mice by sorting in a FACSAria cell sorter (BD Biosciences) after staining with anti-mouse antibodies to *Cd19*, *Fas*, and *GL7* (BD Biosciences). Genomic DNA was isolated by standard procedures and quantified in a fluorometer (Qubit; Invitrogen). DNA capture, library preparation, and DNA sequencing were performed by the Genomics Unit at Centro Nacional de Investigaciones Cardiovasculares (CNIC). In brief, DNA was fragmented in a sonicator (Covaris) to ~200 nucleotide-long (mean size) fragments and purified using AMPure XP beads (Agencourt). Quality was assessed with the 2100 Bioanalyzer (Agilent). Then, fragment ends were repaired, adapters were ligated, and the resulting library was amplified and hybridized with our custom SureSelectXT library of RNA probes. DNA-RNA hybrids were then captured by magnetic bead selection. After indexing, libraries were single-end sequenced in a HiSeq 2500 platform (Illumina).

### Target enrichment assessment by quantitative RT-PCR

*Noxa1*, *Ostn*, and *Pcna* amplifications were quantified with green assay (SYBR; Applied Biosystems) in a real-time PCR system (AB7900 Standard; AbiPrism). *Gapdh* amplifications were used as normalization controls. The following primers were used: *Gapdh* (forward), 5'-TGAAGCAGGCATCTGAGGG-3'; *Gapdh* (reverse), 5'-CGAAGGTGGAAGTGGAG-3'; *Ostn* (forward), 5'-CATAGTGTGCTGTGGTT-3'; *Ostn* (reverse), 5'-CATTATATTGGTCTGCTGTT-3'; *Noxa1* (forward), 5'-CGCGGGACAGCAATGAGAAG-3'; *Noxa1* (reverse), 5'-CCATCTACTCAGTTTCAAGGA-3'; *Pcna* (forward), 5'-CTCCAGCACCTTCTTCAG-3'; and *Pcna* (reverse), 5'-TCTCATCTAGTCGCCACA-3'.

SDS software (Applied Biosystems) was used for analysis of the data.

### Sanger sequencing

Regions to be sequenced were amplified from 160–200-ng genomic DNA in four independent reactions to minimize possible PCR biases. The following primers were used: *Hist1h1b* (forward), 5'-ATGCCTTAGACTTCACCGCC-3'; *Hist1h1b* (reverse), 5'-TTGTAACCTTGAGTCGCCGC-3'; *miR142* (forward), 5'-CGGTCCCTGGGAAGTACAC-3'; *miR142* (reverse), 5'-AACGAGAGGCAAACAGTCTTCA-3'; *Cd19* (forward), 5'-GCCCTCTTCCCTCCTCATA-3'; *Cd19* (reverse), 5'-CCTGCACCCACTCATCTGAA-3'; *Cdk4* (forward), 5'-TCTGGCAGCTGGTCACATGG-3'; and *Cdk4* (reverse), 5'-GATCACCAG

CTAGTCGTCCC-3'. Amplification reactions were carried in a final volume of 25  $\mu$ l using 2.5 U Pfu Ultra HF DNA polymerase (Agilent) and the following PCR setup: 95°C for 2 min, 25 (*Cd19* and *Cdk4*) or 26 cycles (*miR142* and *Hist1h1b*) of denaturation at 94°C for 30 s, annealing at 57°C (*miR142* and *Hist1h1b*) or 58°C (*Cd19* and *Cdk4*) for 30 s, extension at 72°C for 1 min, and a final stage of 72°C for 10 min. PCR products were purified from a 1% agarose gel (Illustra Gel Band Purification kit; GE Healthcare) and cloned into pGEMT vector (Promega). Competent DH5 $\alpha$  *Escherichia coli* bacteria were transformed with the constructs, and individual colonies (192–288 per gene) were grown in 96-well plates. Plasmidic DNA was then isolated (Plasmid MiniPrep kit; Millipore) and sequenced by Sanger sequencing using SP6 universal primer. Sequence analysis was performed using SeqMan software (Lasergene).

### PCR-Seq to validate the machine-learning approach

40–50 ng of genomic DNA was amplified using the following primers: *Apobec3* (forward), 5'-GTCTTCCATAGCCTGCTCACA-3'; *Apobec3* (reverse), 5'-TAGCTGACTGGTGTGGTTCC-3'; *Aurkaip1* (forward), 5'-ACTTGTCAC TTCCGCAGTCC-3'; *Aurkaip1* (reverse), 5'-CCATCC CCAAGTCAGGTGTG-3'; *Ccdc17* (forward), 5'-TCTTTT CTGTCCAGTCCGCC-3'; *Ccdc17* (reverse), 5'-ACAAAT GGGCAGAGTCAGGG-3'; *Cd52* (forward), 5'-TACTGC CGCACACATGACTC-3'; *Cd52* (reverse), 5'-TGAGGT GGGAAGCCAAACAT-3'; *Cd68* (forward), 5'-AGGGGC TGGTAGGTTGATTG-3'; *Cd68* (reverse), 5'-GGAGTC AAGACTGGATTGAC-3'; *Cd69* (forward), 5'-TCT AAAGTTTTGTAGACCC-3'; *Cd69* (reverse), 5'-TGAAGCCTCATCAACGCACT-3'; *Clec2d* (forward), 5'-GGCTCCTGACCTTGAATGC-3'; *Clec2d* (reverse), 5'-AGGCAACTTCTGCCACTATGC-3'; *Coro1a* (forward), 5'-AGGGCTCTGGGGTTCTACTT-3'; *Coro1a* (reverse), 5'-GGAAATGACCACGGGGTTT-3'; *Hist1h1c* (forward), 5'-CTCTATCGGCGTACTGCCAC-3'; *Hist1h1c* (reverse), 5'-ATCGAGTCCCTTGCAACC TT-3'; *Il4i* (forward), 5'-ATTCCCGAGGGAGGTGAG TG-3'; *Il4i* (reverse), 5'-GGTAGCTTCTCTCCGTCA CAC-3'; *Maz* (forward), 5'-GTCAACAAAGAACCCTC CCT-3'; *Maz* (reverse), 5'-CACCTGTCCCCTGAGTTG TG-3'; *Trex1* (forward), 5'-GCCTAACAGGTTTGATTG TCC T-3'; and *Trex1* (reverse), 5'-TAGGCTGAGCAC TCCCAGTC-3'. Amplification reactions were carried in a final volume of 25  $\mu$ l using 2.5 U Pfu Ultra HF DNA polymerase (Agilent); 95°C for 2 min, 26 cycles of 94°C for 30 s, 55°C for 30 s, 72°C for 1 min, and a final stage of 72°C for 10 min. PCR products were purified and fragmented using a sonicator (Covaris), and libraries were prepared by the CNIC Genomics Unit according to the manufacturer's instructions (NEBNext Ultra DNA Library Prep; New England Biolabs). Sequencing was performed in a HiSeq 2500 platform (Illumina). Analysis was performed as previously described (Pérez-Durán et al., 2012).

### Gene expression profiling by RNA-Seq

GC ( $CD19^+ Fas^+ GL7^+$ ) and resting ( $CD19^+ Fas^- GL7^-$ ) B cells were sorted from Peyer's patches of littermate 12-wk-old WT C57BL/6 mice. Three biological replicates were analyzed, each composed of a pool of five female mice. RNA was purified from pellets of  $2\text{--}2.5 \times 10^4$  cells, and DNaseI treatment was applied to avoid DNA contamination (RNAeasy MiniKit; Qiagen). RNA quality was assessed with the 2100 Bioanalyzer, showing high RNA purity and integrity. Sequencing libraries were prepared by the CNIC Genomic Unit according to the manufacturer's protocol (NEB NEXT Ultra RNaseq Library Prep kit; New England Biolabs) from 100 ng RNA per replicate and sequenced in a HiSeq 2500 platform.

### Computational analysis

#### Pipeline to identify and annotate AID-induced mutations.

Raw reads were demultiplexed by Casava v1.8 to generate a fastq file that was aligned to the mouse genome (NCBI m37 v61 Feb 2011) with Novoalign 2.08.01 (command line options: `-o SAM -F ILM1.8 -H -r None -q 2`). Sam files were processed with Samtools 0.1.19 to generate a sorted bam file that was piped to a custom Perl script for the analysis of AID mutations. In brief, the software analyzes the regions of interest in the bam file, annotates hotspots, localizes and suppresses annotated single nucleotide polymorphism positions (Sanger Mouse Genomes Project SNP and Indel Release v2), and reports relevant information about AID activity. AID targets were identified as those genes accumulating significantly more C→T transition mutations in  $Ung^{-/-} Msh2^{-/-}$  than in  $Aicda^{-/-}$  mice (false discovery rate [FDR]  $\leq 0.05$ , one-tail Fisher test and Benjamini-Hochberg correction).

Mutation frequencies were calculated as follows:

$$\text{Total mutation freq} = \frac{\text{Total number of mutations}}{\text{Total sequenced length}}$$

$$\text{Mutation freq}_{CG} = \frac{(\text{Mutated cytosines} + \text{Mutated guanines})}{(\text{Seq length cytosines} + \text{Seq length guanines})}$$

and

$$\text{Mutation freq}_{WRC(Y)(R)GYW} = \frac{(\text{Mutated cytosines}_{WRC(Y)} + \text{Mutated guanines}_{(R)GYW})}{(\text{Seq length cytosines}_{WRC(Y)} + \text{Seq length guanines}_{(R)GYW})}$$

(Only cytosines in  $WRC(Y)$  and guanines in  $(R)GYW$  were considered to calculate mutation frequency at hotspots.)

#### Integration of AID targets with public data on TC and DSB occurrence.

The bar graph included in Fig. S1 F represents overlaps in the 1,375 genes analyzed in this study (divided into mutated and nonmutated genes) and genes where TCs or DSBs occur in B cells: Meng et al. (2014) refer to TC sites identified by HTGTS in  $\alpha CD40+IL4$ -activated B cells as published in Table S2 from their study; Klein et al. (2011) refer

to TC sites identified by TC-Seq in  $IgH^{I-Sce}$  LPS+IL4-activated B cells as published in Table S4 from their study; Chiarle et al. (2011) refer to TC sites identified by HTGTS in  $c\text{-myc}^{25xI-SceI}$   $\alpha CD40+IL4$ -activated B cells as published in Table S3 (significant hits at  $P \leq 0.05$ ) from their study; Qian et al. (2014) refer to DSBs identified by replication protein A (RPA) differential recruitment (RPA-chromatin immunoprecipitation [ChIP]) in  $IgkAID$   $53BP1^{-/-}$  in vitro activated B cells as published in Table S1 A from their study; and Staszewski et al. (2011) refer to DSBs identified by Nbs1 binding (ChIP-on-ChIP) in LPS+ $\alpha IgD$ -dextran+BLYSS-activated B cells as published in Table S1 ( $P \leq 0.05$ ) from their study.

**Sequence context of mutated cytosines.** The sequence context of mutated cytosines (C→T transition frequency  $\geq 4 \times 10^{-3}$ ) was analyzed in a window of 10 nucleotides. Logo representation was done using WebLogo3, and the percentage of each nucleotide in each position surrounding the mutated cytosine was calculated by a custom Perl script. Enrichment for adenosine, guanine, cytosine, or thymine was tested against the sequence context of all cytosines present in the 1,588 regions analyzed in this study (one-tailed Student's *t* test + Bonferroni correction).

**Gene expression profiling by RNA-Seq.** After demultiplexing by Casava v1.8, read quality was assessed by FASTQC, and sequencing adapters were removed from sequence reads by cutadapt v1.9. The resulting reads were aligned to and quantified on the mouse transcriptome (NCBI m38 v75, Feb 2014) using RSEM v1.2.25 with the following parameters: `-p 3-time-output-genome-bam-sampling-for-bam-bowtie-e 60-bowtie-m 30-bowtie-chunkmbs 512-fragment-length-mean 180-fragment-length-sd 50`.

**Transcription rate analysis (GRO-Seq).** Reads were mapped to the mouse genome (mm9/NCBI37) using bowtie2, and uniquely mapped, nonredundant reads were kept. Reads mapping in  $\pm 1$  kb from TSSs were quantified and summarized at the gene level using HTSeq.

**PolII and Spt5 recruitment.** Quantification of PolII and Spt5 recruitment was extracted from Table S3 A in Pavri et al. (2010).

**Superenhancer analysis.** Data were extracted from the catalog of superenhancers that overlap with gene bodies identified in GC B cells as published in Table S3 in Meng et al. (2014) (GEO accession no. GSE62296).

**Epigenetic mark analysis.** Sequencing data (fastq files) for each epigenetic mark were aligned to the mouse genome (NCBI m37 v61, Feb 2011) using bowtie 1.1.1 (command line options: `-best -m1 -n2 -p2`). Alignment files were processed by Samtools 0.1.19 to generate a sorted bam file. Peak calling was done using MACS2 (v2.1.0.20140616) according to the optimal parameters for a histone modification status



profiling as reported by the creators of the tool (Feng et al., 2011). Mapping of annotated peaks to genes was done using GREAT (version 3.0.0).

**Convergent transcription analysis (GRO-Seq).** Convergent transcription data analysis was performed as described in Meng et al. (2014). In brief, reads were mapped to the mouse genome (mm9/NCBI37) using bowtie2, and uniquely mapped, nonredundant reads were kept. HOMER (v4.6) was used with default parameters to identify transcribed regions from both strands and bedtools (v2.24) to find and annotate ConvT regions (regions where >100 bp of sense and anti-sense transcription overlap occurs).

**Machine learning to predict AID targets.** The conditional inference tree for classification was built using the *ctree* function from the party R package with default parameters. Genes with a background mutation frequency  $>5 \times 10^{-4}$  were excluded to avoid artifacts. The following variables were fed into the model for each of the 1,339 genes analyzed: expression, transcription rate, PolII recruitment, and Spt5 recruitment (quantitative, continuous); Med12 recruitment, H3K4me1 recruitment, H3K36me3 recruitment, H3K79me2 recruitment, regulation by superenhancers, and occurrence of convergent transcription (qualitative, discrete). All variables were assigned equal weights to fit the model.

**Annotation of AID targets.** Annotation of AID targets was performed based on public data on sequencing of human DLBCLs, Burkitt lymphomas, and follicular lymphoma tumors (Lohr et al., 2012; Love et al., 2012; Morin et al., 2013; Zhang et al., 2013; de Miranda et al., 2014; Okosun et al., 2014).

#### Data availability

Sequencing data generated for this study are available through the GEO database: targeted DNA deep sequencing (accession no. GSE102944) and RNA-Seq (accession no. GSE98086).

The rest of the datasets analyzed in the current study are publicly available through the GEO and/or Sequence Read Archive: GRO-Seq (accession no. GSE62296), GC B cells (accession nos. SRR1611832, SRR1611833, and SRR1611834), naive B cells (accession nos. SRR1611829, SRR1611830, and SRR1611831), ChIP-Seq of PolII and Spt5 (accession no. GSE24178), and ChIP-Seq data of epigenetic marks Med12 (accession no. SRX347810), H3K4me1 (accession no. SRX347815), H3K36me3 (accession no. SRX185869), and H3K79me2 (accession no. SRX185843).

#### Statistical analysis

Statistical analyses were performed with stats R package v3.1.1. Error bars in figures represent SEM. Student's *t* test was applied to continuous data, and a Fisher test was used to assess differences between categorical variables. P-values were corrected for multiple hypothesis testing by Benjamini-Hochberg

or Bonferroni method where appropriate. Differences were considered statistically significant at  $P \leq 0.05$  or  $q \leq 0.05$ .

#### Online supplemental material

Fig. S1 shows the experimental workflow used to identify AID targets and technical controls. Fig. S2 shows mutation analysis of WRCY/RGYW hotspots. Fig. S3 shows details on the machine-learning classification tree used for the prediction of AID targets. Table S1 contains a list of the genes included in the capture library. Table S2 A contains a detailed mutation analysis of AID targets in *Ung*<sup>+/-</sup>*Msh2*<sup>+/-</sup>, *Ung*<sup>-/-</sup>*Msh2*<sup>+/-</sup>, *Ung*<sup>+/-</sup>*Msh2*<sup>-/-</sup>, and *Ung*<sup>-/-</sup>*Msh2*<sup>-/-</sup>. Table S2 B contains a list of the 18 AID targets mutated in repair-proficient GC B cells. Table S3 shows mutation analysis of genes validated by Sanger sequencing. Table S4 shows mutation analysis of the genes selected for machine-learning validation. Table S5 contains a list of the mutations found in *Ung*<sup>-/-</sup>*Msh2*<sup>-/-</sup> GC B cells that have been identified in cohorts of human lymphoma patients.

#### ACKNOWLEDGMENTS

We thank all members of the B Cell Biology Laboratory for useful discussions, V. Barreto for critical reading of the manuscript, F. Sánchez-Cabo for advice on statistics analysis, J.M. Ligos for help with flow cytometry, and A. Dopazo for advice on DNA capture and sequencing.

A. Pérez-García was a fellow of the research training program funded by the Ministerio de Educación, Cultura y Deporte (grant FPU-AP2009-1732); A.F. Álvarez-Prado and A.R. Ramiro are supported by Centro Nacional de Investigaciones Cardiovasculares Carlos III (CNIC). This work was funded with the following grants to A.R. Ramiro from Plan Estatal de Investigación Científica y Técnica y de Innovación 2013–2016, Programa Estatal de I+D+i Orientada a los Retos de la Sociedad Retos Investigación: Proyectos I+D+i 2016, Ministerio de Economía, Industria y Competitividad (MEIC; grants SAF2013-42767-R and SAF2016-75511-R). This work is cofunded by Fondo Europeo de Desarrollo Regional and the European Research Council Starting Grant program (grant BCLYM-207844). The CNIC is supported by the MEIC and the Pro CNIC Foundation and is a Severo Ochoa Centre of Excellence (MEIC award SEV-2015-0505).

The authors declare no competing financial interests.

Author contributions: A.F. Álvarez-Prado, P. Pérez-Durán, and A.R. Ramiro designed experiments; A.F. Álvarez-Prado, P. Pérez-Durán, A. Pérez-García, and V.G. de Yébenes performed experiments; A. Benguria performed DNA sequencing; A.F. Álvarez-Prado and C. Torroja developed scripts for data analysis; A.F. Álvarez-Prado and A.R. Ramiro analyzed data and prepared figures; and A.F. Álvarez-Prado and A.R. Ramiro wrote the manuscript.

Submitted: 20 September 2017

Revised: 22 November 2017

Accepted: 21 December 2017

#### REFERENCES

- Chiarle, R., Y. Zhang, R.L. Frock, S.M. Lewis, B. Molinje, Y.J. Ho, D.R. Myers, V.W. Choi, M. Compagno, D.J. Malkin, et al. 2011. Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells. *Cell*. 147:107–119. <https://doi.org/10.1016/j.cell.2011.07.049>
- de Miranda, N.F., K. Georgiou, L. Chen, C. Wu, Z. Gao, A. Zaravinos, S. Lisboa, G. Enblad, M.R. Teixeira, Y. Zeng, et al. 2014. Exome sequencing reveals novel mutation targets in diffuse large B-cell lymphomas derived from Chinese patients. *Blood*. 124:2544–2553. <https://doi.org/10.1182/blood-2013-12-546309>

- Di Noia, J.M., and M.S. Neuberger. 2007. Molecular mechanisms of antibody somatic hypermutation. *Annu. Rev. Biochem.* 76:1–22. <https://doi.org/10.1146/annurev.biochem.76.061705.090740>
- Dong, J., R.A. Panchakshari, T. Zhang, Y. Zhang, J. Hu, S.A. Volpi, R.M. Meyers, Y.-J. Ho, Z. Du, D.F. Robbiani, et al. 2015. Orientation-specific joining of AID-initiated DNA breaks promotes antibody class switching. *Nature*. 525:134–139. <https://doi.org/10.1038/nature14970>
- Feng, J., T. Liu, and Y. Zhang. 2011. Using MACS to identify peaks from ChIP-Seq data. *Curr. Protoc. Bioinformatics*. 2:14.
- Frey, S., B. Bertocci, F. Delbos, L. Quint, J.C. Weill, and C.A. Reynaud. 1998. Mismatch repair deficiency interferes with the accumulation of mutations in chronically stimulated B cells and not with the hypermutation process. *Immunity*. 9:127–134. [https://doi.org/10.1016/S1074-7613\(00\)80594-4](https://doi.org/10.1016/S1074-7613(00)80594-4)
- Gordon, M.S., C.M. Kanegai, J.R. Doerr, and R. Wall. 2003. Somatic hypermutation of the B cell receptor genes B29 (Igbeta, CD79b) and mb1 (Igalpha, CD79a). *Proc. Natl. Acad. Sci. USA*. 100:4126–4131. <https://doi.org/10.1073/pnas.0735266100>
- Klein, I.A., W. Resch, M. Jankovic, T. Oliveira, A. Yamane, H. Nakahashi, M. Di Virgilio, A. Bothmer, A. Nussenzweig, D.F. Robbiani, et al. 2011. Translocation-capture sequencing reveals the extent and nature of chromosomal rearrangements in B lymphocytes. *Cell*. 147:95–106. <https://doi.org/10.1016/j.cell.2011.07.048>
- Liu, M., J.L. Duke, D.J. Richter, C.G. Vinuesa, C.C. Goodnow, S.H. Kleinstein, and D.G. Schatz. 2008. Two levels of protection for the B cell genome during somatic hypermutation. *Nature*. 451:841–845. <https://doi.org/10.1038/nature06547>
- Lohr, J.G., P. Stojanov, M.S. Lawrence, D. Auclair, B. Chapuy, C. Sougnez, P. Cruz-Gordillo, B. Knoechel, Y.W. Asmann, S.L. Slager, et al. 2012. Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. *Proc. Natl. Acad. Sci. USA*. 109:3879–3884. <https://doi.org/10.1073/pnas.1121343109>
- Love, C., Z. Sun, D. Jima, G. Li, J. Zhang, R. Miles, K.L. Richards, C.H. Dunphy, W.W. Choi, G. Srivastava, et al. 2012. The genetic landscape of mutations in Burkitt lymphoma. *Nat. Genet.* 44:1321–1325. <https://doi.org/10.1038/ng.2468>
- Meng, F.L., Z. Du, A. Federation, J. Hu, Q. Wang, K.R. Kieffer-Kwon, R.M. Meyers, C. Amor, C.R. Wasserman, D. Neuberger, et al. 2014. Convergent transcription at intragenic super-enhancers targets AID-initiated genomic instability. *Cell*. 159:1538–1548. <https://doi.org/10.1016/j.cell.2014.11.014>
- Method, S.P., and J.M. Di Noia. 2017. Molecular Mechanisms of Somatic Hypermutation and Class Switch Recombination. *Adv. Immunol.* 133:37–87. <https://doi.org/10.1016/bs.ai.2016.11.002>
- Morin, R.D., K. Mungall, E. Pleasance, A.J. Mungall, R. Goya, R.D. Huff, D.W. Scott, J. Ding, A. Roth, R. Chiu, et al. 2013. Mutational and structural analysis of diffuse large B-cell lymphoma using whole-genome sequencing. *Blood*. 122:1256–1265. <https://doi.org/10.1182/blood-2013-02-483727>
- Muramatsu, M., K. Kinoshita, S. Fagarasan, S. Yamada, Y. Shinkai, and T. Honjo. 2000. Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell*. 102:553–563. [https://doi.org/10.1016/S0092-8674\(00\)00078-7](https://doi.org/10.1016/S0092-8674(00)00078-7)
- Müschen, M., D. Re, B. Jungnickel, V. Diehl, K. Rajewsky, and R. Küppers. 2000. Somatic mutation of the CD95 gene in human B cells as a side-effect of the germinal center reaction. *J. Exp. Med.* 192:1833–1840. <https://doi.org/10.1084/jem.192.12.1833>
- Nambu, Y., M. Sugai, H. Gonda, C.-G. Lee, T. Katakai, Y. Agata, Y. Yokota, and A. Shimizu. 2003. Transcription-coupled events associating with immunoglobulin switch region chromatin. *Science*. 302:2137–2140. <https://doi.org/10.1126/science.1092481>
- Nilsen, H., I. Rosewell, P. Robins, C.F. Skjælbred, S. Andersen, G. Slupphaug, G. Daly, H.E. Krokan, T. Lindahl, and D.E. Barnes. 2000. Uracil-DNA glycosylase (UNG)-deficient mice reveal a primary role of the enzyme during DNA replication. *Mol. Cell*. 5:1059–1065. [https://doi.org/10.1016/S1097-2765\(00\)80271-3](https://doi.org/10.1016/S1097-2765(00)80271-3)
- Okosun, J., C. Bödör, J. Wang, S. Araf, C.Y. Yang, C. Pan, S. Boller, D. Cittaro, M. Bozek, S. Iqbal, et al. 2014. Integrated genomic analysis identifies recurrent mutations and evolution patterns driving the initiation and progression of follicular lymphoma. *Nat. Genet.* 46:176–181. <https://doi.org/10.1038/ng.2856>
- Pasqualucci, L., P. Neumeister, T. Goossens, G. Nanjangud, R.S.K. Chaganti, R. Küppers, and R. Dalla-Favera. 2001. Hypermutation of multiple proto-oncogenes in B-cell diffuse large-cell lymphomas. *Nature*. 412:341–346. <https://doi.org/10.1038/35085588>
- Pavri, R., A. Gazumyan, M. Jankovic, M. Di Virgilio, I. Klein, C. Ansarah-Sobrinho, W. Resch, A. Yamane, B. Reina San-Martín, V. Barreto, et al. 2010. Activation-induced cytidine deaminase targets DNA at sites of RNA polymerase II stalling by interaction with Spt5. *Cell*. 143:122–133. <https://doi.org/10.1016/j.cell.2010.09.017>
- Pérez-Durán, P., L. Belver, V.G. de Yébenes, P. Delgado, D.G. Pisano, and A.R. Ramiro. 2012. UNG shapes the specificity of AID-induced somatic hypermutation. *J. Exp. Med.* 209:1379–1389. <https://doi.org/10.1084/jem.20112253>
- Petersen-Mahrt, S.K., R.S. Harris, and M.S. Neuberger. 2002. AID mutates *E. coli* suggesting a DNA deamination mechanism for antibody diversification. *Nature*. 418:99–104. <https://doi.org/10.1038/nature00862>
- Phung, Q.H., D.B. Winter, A. Cranston, R.E. Tarone, V.A. Bohr, R. Fishel, and P.J. Gearhart. 1998. Increased hypermutation at G and C nucleotides in immunoglobulin variable genes from mice deficient in the MSH2 mismatch repair protein. *J. Exp. Med.* 187:1745–1751. <https://doi.org/10.1084/jem.187.11.1745>
- Qian, J., Q. Wang, M. Dose, N. Pruett, K.R. Kieffer-Kwon, W. Resch, G. Liang, Z. Tang, E. Mathé, C. Benner, et al. 2014. B cell super-enhancers and regulatory clusters recruit AID tumorigenic activity. *Cell*. 159:1524–1537. <https://doi.org/10.1016/j.cell.2014.11.013>
- Rada, C., M.R. Ehrenstein, M.S. Neuberger, and C. Milstein. 1998. Hot spot focusing of somatic hypermutation in MSH2-deficient mice suggests two stages of mutational targeting. *Immunity*. 9:135–141. [https://doi.org/10.1016/S1074-7613\(00\)80595-6](https://doi.org/10.1016/S1074-7613(00)80595-6)
- Rada, C., G.T. Williams, H. Nilsen, D.E. Barnes, T. Lindahl, and M.S. Neuberger. 2002. Immunoglobulin isotype switching is inhibited and somatic hypermutation perturbed in UNG-deficient mice. *Curr. Biol.* 12:1748–1755. [https://doi.org/10.1016/S0960-9822\(02\)01215-0](https://doi.org/10.1016/S0960-9822(02)01215-0)
- Rada, C., J.M. Di Noia, and M.S. Neuberger. 2004. Mismatch recognition and uracil excision provide complementary paths to both Ig switching and the A/T-focused phase of somatic mutation. *Mol. Cell*. 16:163–171. <https://doi.org/10.1016/j.molcel.2004.10.011>
- Ramiro, A.R., M. Jankovic, T. Eisenreich, S. Difilippantonio, S. Chen-Kiang, M. Muramatsu, T. Honjo, A. Nussenzweig, and M.C. Nussenzweig. 2004. AID is required for c-myc/IgH chromosome translocations in vivo. *Cell*. 118:431–438. <https://doi.org/10.1016/j.cell.2004.08.006>
- Ramiro, A.R., M. Jankovic, E. Callen, S. Difilippantonio, H.-T. Chen, K.M. McBride, T.R. Eisenreich, J. Chen, R.A. Dickens, S.W. Lowe, et al. 2006. Role of genomic instability and p53 in AID-induced c-myc-IgH translocations. *Nature*. 440:105–109. <https://doi.org/10.1038/nature04495>
- Reitmair, A.H., R. Schmits, A. Ewel, B. Bapat, M. Redston, A. Mitri, P. Waterhouse, H.-W. Mittrücker, A. Wakeham, B. Liu, et al. 1995. MSH2 deficient mice are viable and susceptible to lymphoid tumours. *Nat. Genet.* 11:64–70. <https://doi.org/10.1038/ng0995-64>
- Revy, P., T. Muto, Y. Levy, F. Geissmann, A. Plebani, O. Sanal, N. Catalan, M. Forveille, R. Dufourcq-Labelouse, A. Gennery, et al. 2000. Activation-induced cytidine deaminase (AID) deficiency causes the autosomal recessive form of the Hyper-IgM syndrome (HIGM2). *Cell*. 102:565–575. [https://doi.org/10.1016/S0092-8674\(00\)00079-9](https://doi.org/10.1016/S0092-8674(00)00079-9)

- Reynaud, C.A., F. Delbos, A. Faily, Q. Guéranger, S. Aoufouchi, and J.C. Weill. 2009. Competitive repair pathways in immunoglobulin gene hypermutation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364:613–619. <https://doi.org/10.1098/rstb.2008.0206>
- Robbiani, D.F., A. Bothmer, E. Callen, B. Reina-San-Martin, Y. Dorsett, S. Difilippantonio, D.J. Bolland, H.T. Chen, A.E. Corcoran, A. Nussenzweig, and M.C. Nussenzweig. 2008. AID is required for the chromosomal breaks in *c-myc* that lead to *c-myc*/IgH translocations. *Cell*. 135:1028–1038. <https://doi.org/10.1016/j.cell.2008.09.062>
- Robbiani, D.F., S. Bunting, N. Feldhahn, A. Bothmer, J. Camps, S. Deroubaix, K.M. McBride, I.A. Klein, G. Stone, T.R. Eisenreich, et al. 2009. AID produces DNA double-strand breaks in non-Ig genes and mature B cell lymphomas with reciprocal chromosome translocations. *Mol. Cell*. 36:631–641. <https://doi.org/10.1016/j.molcel.2009.11.007>
- Rogozin, I.B., and N.A. Kolchanov. 1992. Somatic hypermutagenesis in immunoglobulin genes: II. Influence of neighbouring base sequences on mutagenesis. *Biochim. Biophys. Acta*. 1171:11–18. [https://doi.org/10.1016/0167-4781\(92\)90134-L](https://doi.org/10.1016/0167-4781(92)90134-L)
- Shaffer, A.L. III, R.M. Young, and L.M. Staudt. 2012. Pathogenesis of human B cell lymphomas. *Annu. Rev. Immunol.* 30:565–610. <https://doi.org/10.1146/annurev-immunol-020711-075027>
- Shen, H.M., A. Peters, B. Baron, X. Zhu, and U. Storb. 1998. Mutation of BCL-6 gene in normal B cells by the process of somatic hypermutation of Ig genes. *Science*. 280:1750–1752. <https://doi.org/10.1126/science.280.5370.1750>
- Staszewski, O., R.E. Baker, A.J. Ucher, R. Martier, J. Stavnezer, and J.E.J. Guikema. 2011. Activation-induced cytidine deaminase induces reproducible DNA breaks at many non-Ig Loci in activated B cells. *Mol. Cell*. 41:232–242. <https://doi.org/10.1016/j.molcel.2011.01.007>
- Stavnezer, J., J.E.J. Guikema, and C.E. Schrader. 2008. Mechanism and regulation of class switch recombination. *Annu. Rev. Immunol.* 26:261–292. <https://doi.org/10.1146/annurev.immunol.26.021607.090248>
- Storb, U. 2014. Why does somatic hypermutation by AID require transcription of its target genes? *Adv. Immunol.* 122:253–277. <https://doi.org/10.1016/B978-0-12-800267-4.00007-9>
- Wang, Q., T. Oliveira, M. Jankovic, I.T. Silva, O. Hakim, K. Yao, A. Gazumyan, C.T. Mayer, R. Pavri, R. Casellas, et al. 2014. Epigenetic targeting of activation-induced cytidine deaminase. *Proc. Natl. Acad. Sci. USA*. 111:18667–18672. <https://doi.org/10.1073/pnas.1420575111>
- Wei, L., R. Chahwan, S. Wang, X. Wang, P.T. Pham, M.F. Goodman, A. Bergman, M.D. Scharff, and T. MacCarthy. 2015. Overlapping hotspots in CDRs are critical sites for V region diversification. *Proc. Natl. Acad. Sci. USA*. 112:E728–E737. <https://doi.org/10.1073/pnas.1500788112>
- Yamane, A., W. Resch, N. Kuo, S. Kuchen, Z. Li, H.W. Sun, D.F. Robbiani, K. McBride, M.C. Nussenzweig, and R. Casellas. 2011. Deep-sequencing identification of the genomic targets of the cytidine deaminase AID and its cofactor RPA in B lymphocytes. *Nat. Immunol.* 12:62–69. <https://doi.org/10.1038/ni.1964>
- Yeap, L.-S., J.K. Hwang, Z. Du, R.M. Meyers, F.-L. Meng, A. Jakubauskaitė, M. Liu, V. Mani, D. Neuberger, T.B. Kepler, et al. 2015. Sequence-Intrinsic Mechanisms that Target AID Mutational Outcomes on Antibody Genes. *Cell*. 163:1124–1137. <https://doi.org/10.1016/j.cell.2015.10.042>
- Zhang, J., V. Grubor, C.L. Love, A. Banerjee, K.L. Richards, P.A. Mieczkowski, C. Dunphy, W. Choi, W.Y. Au, G. Srivastava, et al. 2013. Genetic heterogeneity of diffuse large B-cell lymphoma. *Proc. Natl. Acad. Sci. USA*. 110:1398–1403. <https://doi.org/10.1073/pnas.1205299110>

SUPPLEMENTAL MATERIAL

Álvarez-Prado et al., <https://doi.org/10.1084/jem.20171738>

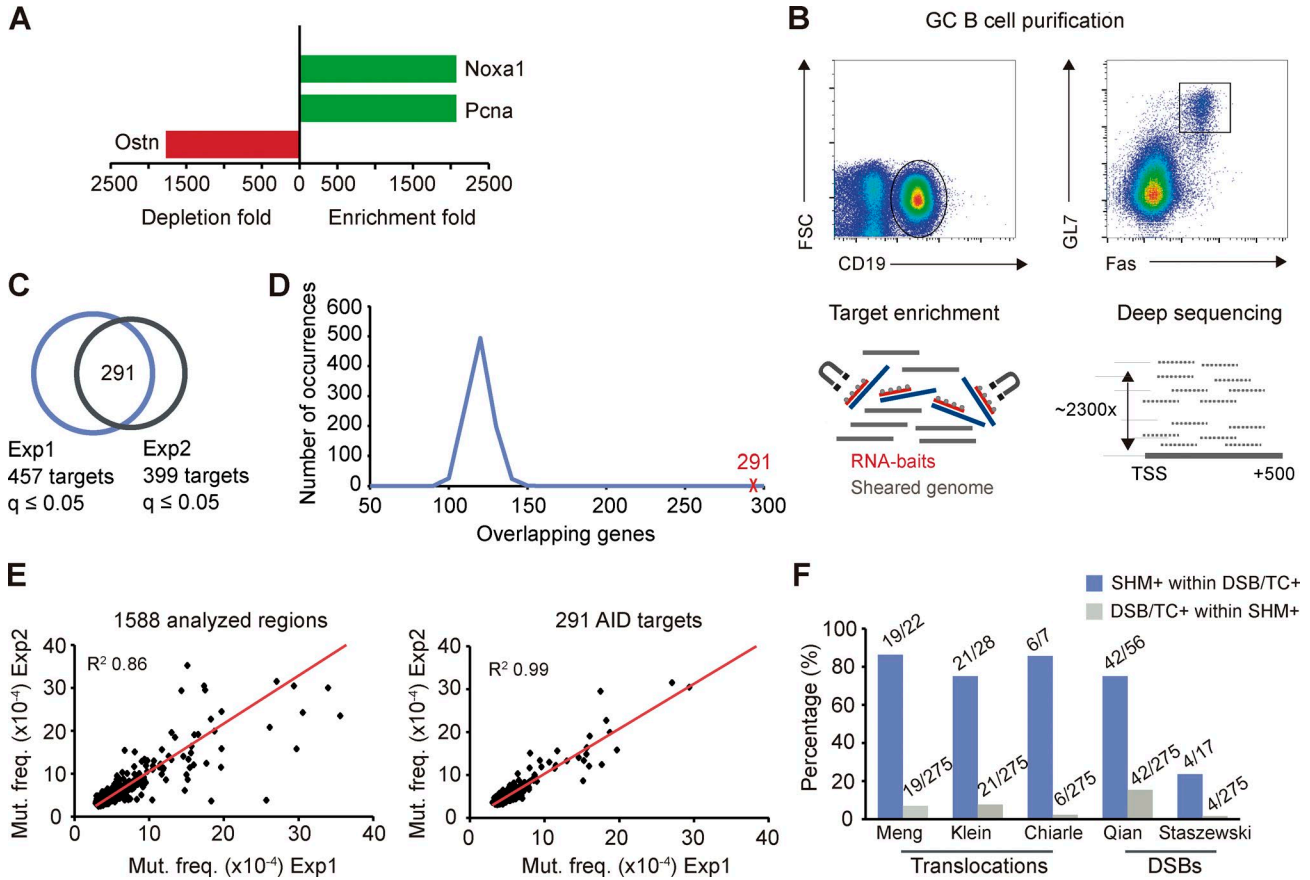


Figure S1. **Identification of AID targets by target enrichment coupled to next generation sequencing.** (A) Target enrichment protocol allows a 2,000-fold enrichment of selected genes. Genomic DNA corresponding to genes included (*Noxa1* and *PCNA*) and not included (*Osth*) in the SureSelect capture library was quantified by quantitative RT-PCR before and after DNA capture enrichment. Graph represents fold depletion or fold enrichment calculated as  $2^{(C_{Input} - C_{Enriched\ fraction})}$ . Mean of two independent experiments is represented. (B) Schematic representation of the experimental approach used. GC (*CD19<sup>+</sup>Fas<sup>+</sup>GL7<sup>+</sup>*) B cells from Peyer's patches were isolated by cell sorting, and genomic DNA was extracted, sheared, and captured with a custom library of RNA probes. Enriched DNA was subjected to next generation sequencing to achieve a mean depth of 2,300 reads per nucleotide. (C) Two independent experiments were performed (Table S2) with 457 mutated targets found in Exp1 and 399 in Exp2. An overlap of 291 AID targets was found between Exp1 and Exp2. (D) Experimental distribution of random overlaps simulated for 1,000 iterations. For each iteration, random groups of 457 and 399 genes were selected from the genes included in the SureSelect capture library, overlapped, and the number of coincident genes reported. The probability to find an overlap of 291 genes by chance is  $<1$  out of each  $10^{16}$  times tested. Two-tailed Fisher test;  $P = \sim 10^{-16}$ . (E) Mutation frequencies of the 1,588 TSS proximal regions analyzed and the 291 targets found in two independent experiments. (F) Percentage of genes undergoing DSB/TC+ according to the indicated studies within AID mutational targets described in this study (SHM+; 275 genes obtained in two independent experiments) and percentage of SHM+ genes within DSB/TC+ genes (see Materials and methods).

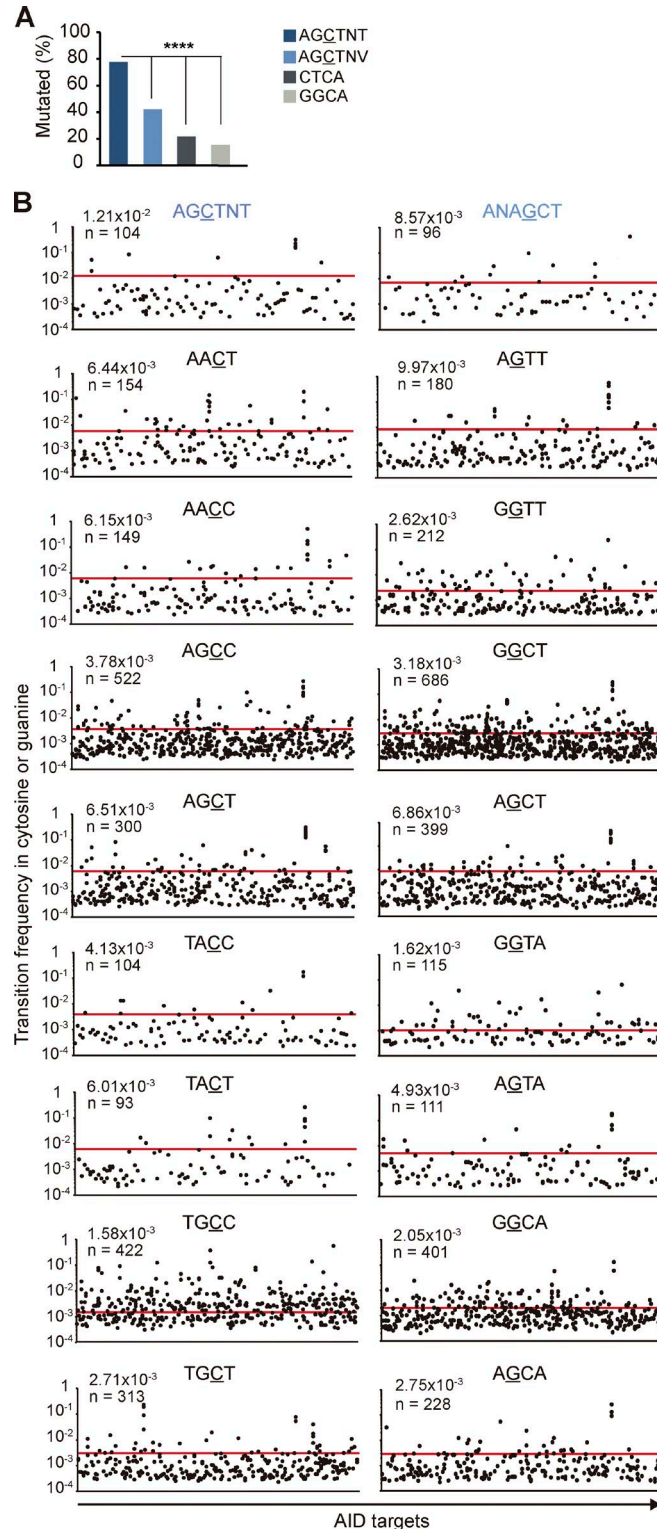


Figure S2. **Mutation analysis at WRCY/RGYW hotspots in *Ung*<sup>-/-</sup>*Msh2*<sup>-/-</sup> GC B cells.** (A) Percentage of mutated cytosines within AGCTNT and AGCTNV hotspots and CTCA and GGCA non-hotspot motifs (Fisher test; \*\*\*\*,  $P < 10^{-13}$ ). (B) Plots show mutated individual hotspots (WRCY, left; RGYW, right). Newly identified AGCTNT/ANAGCT hotspots are shown in the top row. Within each plot, each dot represents an individual WRCY/RGYW motif found mutated at least once. Each position in the x axis corresponds to a different gene, and the y axis shows mutation frequency of each individual hotspot within a gene. Mean mutation frequency is indicated and depicted with a red line. Number of mutated hotspots is indicated.

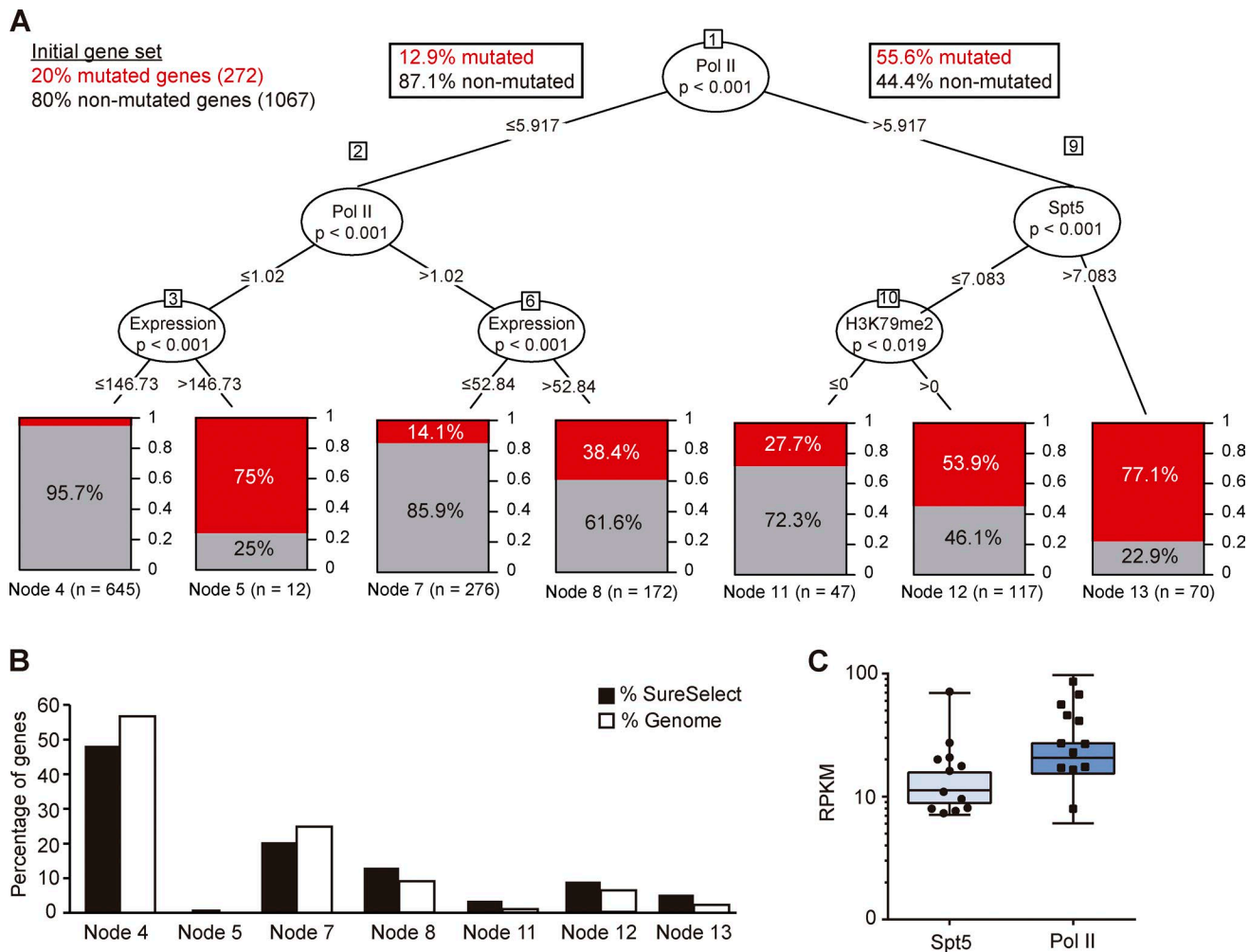


Figure S3. **Machine learning to predict AID targets genome wide.** (A) Recursive partitioning tree model classifies AID targets based on different molecular features: mRNA expression, PolII and Spt5 recruitment, and presence of H3K79me2 epigenetic mark (see Materials and methods). Each node splits the genes into two significantly different groups based on a particular feature. Numbers within the branches indicate the thresholds used to make the groups; p-values of each decision are included below the parameter measured in each node. (B) Bar graph depicting the proportion of SureSelect genes (1,339 genes; closed bars) or of total genes in the mouse genome (17,858 genes; open bars) that meet the thresholds established in each node. (C) Box plot depicting genome-wide data of PolII and Spt5 recruitment in in vitro activated B cells. Black dots and squares mark the 12 genes selected for the validation of the model prediction. RPKM, reads per kilobase per million reads mapped.

Tables S1–S5 are provided in separate Excel files.

Table S1 contains a list of the genes included in the capture library.

Table S2 A contains a detailed mutation analysis of AID targets in *Ung<sup>+/-</sup>Msh2<sup>+/-</sup>*, *Ung<sup>-/-</sup>Msh2<sup>+/-</sup>*, *Ung<sup>+/-</sup>Msh2<sup>-/-</sup>*, and *Ung<sup>-/-</sup>Msh2<sup>-/-</sup>*. Table S2 B contains a list of the 18 AID targets mutated in repair-proficient GC B cells.

Table S3 shows mutation analysis of genes validated by Sanger sequencing.

Table S4 shows mutation analysis of the genes selected for machine-learning validation.

Table S5 contains a list of the mutations found in *Ung<sup>-/-</sup>Msh2<sup>-/-</sup>* GC B cells that have been identified in cohorts of human lymphoma patients.