Nonimmunoglobulin target loci of activation-induced cytidine deaminase (AID) share unique features with immunoglobulin genes

Lucia Kato[a], Nasim A. Begum[a], A. Maxwell Burroughs[b], Tomomitsu Doi[a,1], Jun Kawai[b], Carsten O. Daub[b], Takahisa Kawaguchi[c], Fumihiko Matsuda[c], Yoshihide Hayashizaki[b], and Tasuku Honjo[a,2]

[a]Department of Immunology and Genomic Medicine and [c]The Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto 606-8501, Japan; and [b]RIKEN Omics Science Center (OSC), RIKEN Yokohama Institute, Yokohama, Kanagawa 230-0045, Japan

Contributed by Tasuku Honjo, December 28, 2011 (sent for review December 5, 2011)

Activation-induced cytidine deaminase (AID) is required for both somatic hypermutation and class-switch recombination in activated B cells. AID is also known to target nonimmunoglobulin genes and introduce mutations or chromosomal translocations, eventually causing tumors. To identify as-yet-unknown AID targets, we screened early AID-induced DNA breaks by using two independent genome-wide approaches. Along with known AID targets, this screen identified a set of unique genes (*SNHG3, MALAT1, BCL7A,* and *CUX1*) and confirmed that these loci accumulated mutations as frequently as *Ig* locus after AID activation. Moreover, these genes share three important characteristics with the Ig gene: translocations in tumors, repetitive sequences, and the epigenetic modification of chromatin by H3K4 trimethylation in the vicinity of cleavage sites.

deep sequencing | end labeling by biotin oligonucleotide | microarray

Activation-induced cytidine deaminase (AID) is expressed in germinal center (GC) B cells upon antigen stimulation and is essential for two types of genetic alteration in the Ig gene: class switch recombination (CSR) and somatic hypermutation (SHM), which provide the genetic basis for antibody memory (1, 2). CSR produces antibodies with different effector functions by recombination at Ig heavy chain (H) switch (S) regions, so that the μ-chain constant (Cμ) region is replaced by a downstream C_H region. SHM introduces nontemplated point mutations in the rearranged variable (V) region genes, resulting in incremented antigen receptor affinity after clonal selection (3, 4).

Functional studies on AID mutants have shown that distinct AID domains are required for SHM and CSR, although AID has a single catalytic center (cytidine deaminase motif) in the middle of the molecule. Deletions and alterations in the N-terminal region affect both the CSR and SHM activities (5). However, AID C-terminal mutants almost completely lose CSR activity but retain or even increase SHM activity (6, 7). Although C-terminally truncated AID mutants cleave both V and S regions and induce enhanced c-myc-IgH translocations, they cannot mediate CSR, suggesting that the C-terminal domain is not required for DNA cleavage but is required to correctly pair cleaved ends (8).

The DNA cleavage of targets in CSR and SHM (the S region and V region, respectively) requires their transcription (9–12). Indeed, AID-induced mutations (SHM) are generally detected in a region within 2 kb downstream of the transcription start site (TSS) (13, 14). Transcription appears to play two roles in the targeting of cleavage sites. First, transcription is associated with the epigenetic marking of the target locus, particularly by H3K4 trimethylation (H3K4me3). The histone chaperone complex FACT is required to regulate H3K4me3 in the target S region, and FACT knockdown abolishes H3K4me3 and DNA cleavage in this region (15). Second, transcription is probably required to induce non-B structures in highly repetitive sequences such as S regions (16–18), due to excessive negative supercoiling induced immediately downstream of transcription. V regions have also been shown to form stem-loop structures under these conditions (19, 20). Non-B structure involvement has recently been reported in transcription-associated mutations in repetitive sequences such as the dinucleotide repeat hot spots or triplet repeat expansion/contractions causing Huntington's disease (17, 21, 22).

AID-dependent DNA cleavage is, in general, specific to the Ig locus. However, a number of reports have shown that AID can induce DNA cleavage in non-Ig loci. AID non-Ig targets were first demonstrated by studies on AID transgenic mice that produce numerous T lymphomas, in which vast numbers of mutations accumulate in the genes encoding the T-cell receptor, CD4, CD5, c-myc, and PIM1 (23, 24). This finding was followed by the observations that AID deficiency abolishes c-myc-Ig translocation and reduces the incidence of plasmacytoma (25, 26). AID expression is specific to activated B cells under normal conditions. However, AID expression has also been found in non-B cells, especially in cells stimulated by infection with pathogens such as human T-cell leukemia virus type 1 (HTLV1), hepatitis C virus (HCV), Epstein–Barr (EB) virus, and *Helicobacter pylori* (27–30). Based on these observations, AID is postulated to induce tumorigenesis, especially in B lymphomas and leukemias—and AID is expressed in many GC-derived human B-cell lymphomas (31–33). The prognosis of acute lymphocytic leukemia (ALL) and chronic myeloid leukemia (CML) is linked with AID expression (34, 35). It is therefore important to determine which non-Ig genes can be targeted by AID, and what features, if any, they share with Ig genes.

Several approaches have been used to explore AID non-Ig target genes in B cells. Candidate approaches involving the direct sequencing of proto-oncogenes, genes involved in translocations, or genes transcribed in normal GC B cells have shown that AID mutates several non-Ig genes, including *BCL6*, *MYC*, *PIM1*, and *PAX5* (24, 32, 36, 37). More recently, several efforts have been made to identify AID targets in a whole genome. These approaches have used chromatin immunoprecipitation (ChIP) of CSR-related proteins in combination with genome-wide tiling microarrays (ChIP-chip) or deep sequencing (ChIP-seq) on the assumption that proteins involved in CSR bind to AID targets. RPA, Nbs1, AID itself, and Spt5 have been used as marking proteins in this type of study (38–40). However, these approaches did not necessarily show that all of the protein-bound targets are cleaved or mutated by AID. There are indications that some genes identified by such approaches are not tran-

Author contributions: L.K., T.D., and T.H. designed research; L.K., N.A.B., and A.M.B. performed research; T.K. and F.M. contributed new reagents/analytic tools; L.K., N.A.B., A.M.B., J.K., C.O.D., and Y.H. analyzed data; and L.K. and T.H. wrote the paper.

The authors declare no conflict of interest.

[1]Present address: Laboratory Animal Research Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan.

[2]To whom correspondence should be addressed. E-mail: honjo@mfour.med.kyoto-u.ac.jp.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1120791109/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1120791109

scribed (39). Therefore, it is important to reexamine non-Ig AID target genes by using a different strategy.

Here, we report four AID targets, identified by a combination of unique techniques. After directly labeling the DNA breakage ends from AID-induced cleavage with a biotinylated linker, we isolated the labeled fragments with streptavidin beads and analyzed them by a combination of promoter arrays and genome-wide sequencing. The candidates identified were then confirmed by quantitative PCR (qPCR) and the actual demonstration of mutations. With these methods, we identified at least four previously unknown AID targets—*SNHG3, MALAT1, BCL7A,* and *CUX1*. We found that these targets share important characteristics with Ig genes, namely, repetitive sequences that can form non-B structures upon efficient transcription, and the accumulation of H3K4me3 histone modifications on the chromatin.

Results

AID-Induced DNA Cleavage Detected by Labeling DNA Break Ends with a Biotinylated Linker. To detect genome-wide AID-induced DNA breaks, we used a modified in situ DNA end-labeling technique as described (8, 41) in BL2 cells, a Burkitt's lymphoma cell line that serves as an in vitro model for studying the SHM mechanism (31, 42, 43). We used the BL2 clone BL2-ΔC-AIDER, which expresses JP8Bdel, an AID mutant lacking the C-terminal 16 residues, fused with the hormone-binding domain of the estrogen receptor (ER) (JP8Bdel-ER). Tamoxifen (4-OHT) treatment induces DNA breakage in the Sμ and Sα regions but not in the Sγ region of JP8Bdel-ER-expressing CH12 cells, which switch almost exclusively from IgM to IgA (8).

BL2-ΔC-AIDER cells were treated with 4-OHT only for 3 h to minimize cell death and DNA break ends were labeled with a biotinylated linker, and the break-enriched biotinylated DNA was used as a PCR template (Fig. 1*A*). In agreement with previous reports (8, 42), we detected DNA breakage in the 5′ Sμ region of the IgH locus only in 4-OHT–treated cells. No breakage was detected in the *B2M* gene, which is expressed in BL2 cells but was shown not to accumulate mutations in activated B cells (Fig. 1*B*).

AID Targets Identified by Promoter Array and Whole Genome Sequencing. Because SHM is normally detected close to the TSS (13, 14), biotin linker-enriched DNA fragments were analyzed by a promoter array to identify unknown AID targets. Table S1 lists the genes whose signals increased after 3 h of 4-OHT treatment, compared with untreated samples with false discovery rate (FDR) values <0.3. We also looked for genes with increased signals after 4-OHT treatment that are known to be targets of chromosomal translocation or genes that had multiple breakage peaks, and we identified >50 genes, among which we found that *BCL7A* and *CUX1* are enriched in the original breakage-enriched library by qPCR (see below). We confirmed by RT-PCR and expression array that *SNHG3*, *MALAT1*, *NIN*, *C9orf72*, *CFLAR*, *SNX25*, *BCL7A,* and *CUX1* were transcribed in BL2 cells (Table S1). Fig. S1 shows the peak signals in a 10-kb segment surrounding the breakage area of *SNHG3, MALAT1, BCL7A,* and *CUX1*. We could not map the breakage in the Ig locus because of the absence of array probes in this region.

Because the promoter array does not detect DNA fragments outside of regions containing probes, we further analyzed the breakage-enriched DNA by direct sequencing of the biotin linker-enriched library. DNA breakage sites in both control and 4-OHT–treated libraries were identified by aligning sequenced tags to the genome, and significantly enriched regions were identified by comparing the local breakage density (*SI Materials and Methods*). Regions were identified in the genes listed in Table S2. Interestingly, *SNHG3* and *MALAT1*, which were identified by the promoter array, appear at the top of the list in the genome-wide sequencing as well.

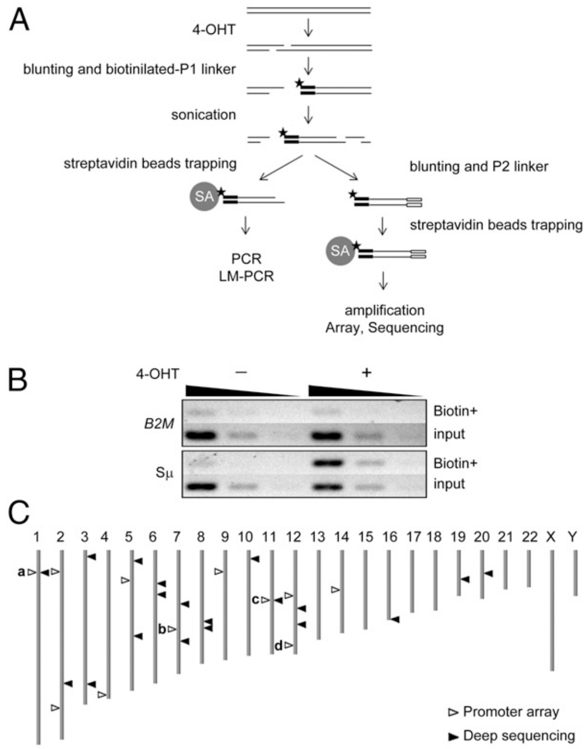

Fig. 1. (*A*) Schematic of the labeling technique. 4-OHT is added to activate AID, and DNA break ends are labeled in situ by biotinylated linker ligation. After genomic DNA is extracted and sonicated, biotinylated fragments are captured by streptavidin beads and used for PCR, array, or sequencing. (*B*) Detection of DNA breaks by PCR. BL2-ΔC-AIDER cells were treated with or without 4-OHT for 3 h, and the break ends were labeled. PCR of Sμ and *B2M* was performed with biotin-labeled DNA or input DNA by using fivefold serially diluted templates. (*C*) Chromosomal distribution of AID targets. a, *SNHG3*; b, *CUX1*; c, *MALAT1*; d, *BCL7A*. White arrowhead, promoter array (FDR < 0.3 plus *BCL7A* and *CUX1*); black arrowhead, whole genome sequencing (FDR < 0.01 and/or remarkable numbers of *P* value clusters).

Fig. 1*C* shows the chromosomal distribution of AID target candidates identified by promoter array or whole-genome sequencing. Breakage seemed to be distributed through the genome without any apparent bias. Surprisingly, of the 29 candidates identified by whole-genome sequencing with strict statistical parameters, only two matched candidates obtained from the promoter array. This discrepancy might be explained in part because most of the breakage-rich regions detected by whole genome sequencing are located in regions that do not contain promoter array probes.

Results may also be limited because of possible bias by PCR amplification of the primary library for microarray and whole-genome sequencing, which could affect the relative genome coverage. To avoid this bias, we relied on the original library and confirmed all candidates by qPCR.

qPCR Analyses of Linker Libraries. To confirm the AID-induced breakage candidates detected by the promoter array and whole-genome sequencing, we used qPCR assays with gene-specific primers to amplify the vicinity of the identified breakage regions in biotin linker-enriched DNA from cells treated with 4-OHT for 3 h (Fig. 2). We examined whether candidate genes were enriched in the 4-OHT–treated DNA library compared with the nontreated library. Among the 29 candidates identified by whole-genome sequencing, only *SNHG3* and *MALAT1* were strongly enriched ($P < 0.0001$ and $P < 0.001$, respectively). Besides these, *BCL7A*, *CUX1,* and *CFLAR*, which were picked up only by the

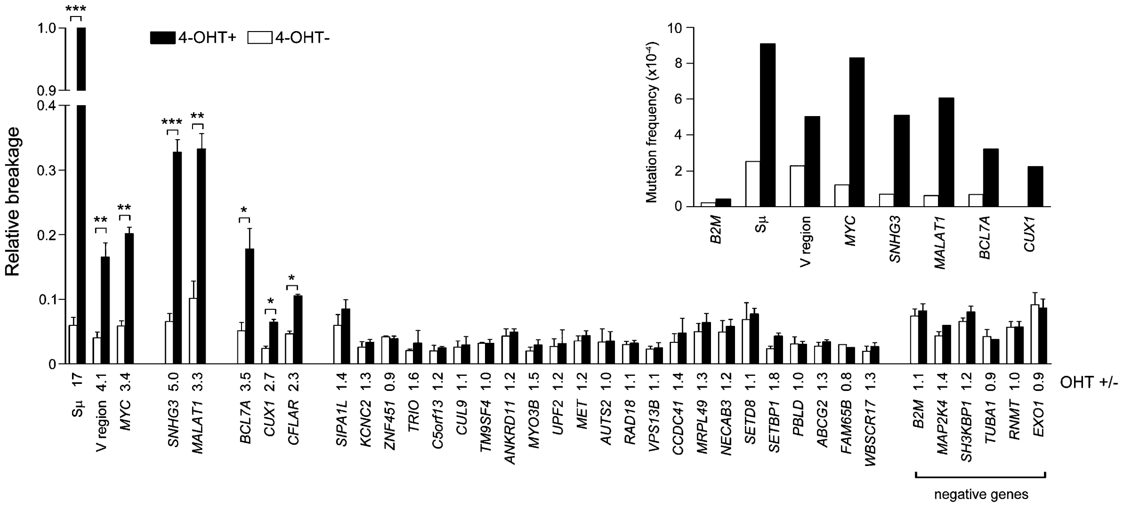

Fig. 2. qPCR measurement of DNA breaks. Break signals are presented relative to Sμ. SD values were derived from at least three independent experiments, and *P* values were calculated by a two-tailed *t* test. **P* < 0.01, ***P* < 0.001, ****P* < 0.0001. Numbers below the *x* axis indicate the ratio between samples treated and not treated with 4-OHT. (*Inset*) Mutation analysis of genes with significantly increased break signals after AID activation. Cells were treated with or without 4-OHT for 24 h. Only unique mutations were counted. Detailed mutation profiles can be found in Fig. S2 and Table S3.

promoter array, also showed significant enrichment ($P < 0.01$) in the 4-OHT–treated library.

We also confirmed that the Sμ and V regions in BL2 cells were cleaved, because they were enriched in the 4-OHT–treated library. Although *MYC*, which is translocated in an AID-dependent manner in human Burkitt's lymphoma (44), was not identified by either promoter array or whole-genome sequencing, qPCR of the 4-OHT–treated samples clearly revealed *MYC* gene enrichment (Fig. 2). The difference in cleavage detection between the direct candidate qPCR and genome-wide arrays and sequencing suggests that the amplification step required for microarray and whole-genome sequencing methods may introduce bias, either for or against many genes. In the case of sequencing, this bias can lead to low mapping coverage of certain regions, hampering efforts to identify significant enrichment. Therefore, we cannot exclude genes that were not identified by the present methods from being AID targets.

AID Targets Accumulate Somatic Mutations near Cleavage Sites. To test whether the newly identified target genes are mutated upon AID activation, we treated BL2-ΔC-AIDER cells with 4-OHT for 24 h and sequenced regions of ≈600 bp around each area with abundant breakage (Fig. S2 and Table S3). Mutations increased in all of the qPCR-confirmed AID target genes after 4-OHT treatment (Fig. 2, *Inset*), with mutation frequencies ranging from 6.1×10^{-4} for *MALAT1* to 2.2×10^{-4} for *CUX*. These frequencies are comparable to those of the V region (5.0×10^{-4}), the Sμ region (9.1×10^{-4}), and the *MYC* gene (8.3×10^{-4}), and are far higher than that of the control *B2M* gene (4.3×10^{-5}). We also detected mutations in the *CFLAR* gene; however, the mutation frequency (9.2×10^{-5}) was not as high as other AID target genes, although mutations increased significantly in 4-OHT–treated sample ($P = 0.004$) (Table S3).

To compare the distribution profiles of mutated bases and AID-induced DNA breaks in the biotin linker-enriched DNA, we mapped the linker positions by performing ligation-mediated (LM)-PCR with the linker primer and gene-specific primers. These PCR fragments were subsequently cloned and sequenced. Break ends identified by the linker were plotted, together with mutation positions (Fig. 3 and Fig. S2). The results clearly showed that the DNA cleavage marks (biotin linker) were closely associated with mutations, indicating that the DNA cleavage sites identified are functionally relevant to SHM by AID. We used RT-PCR and expression arrays to confirm that the regions

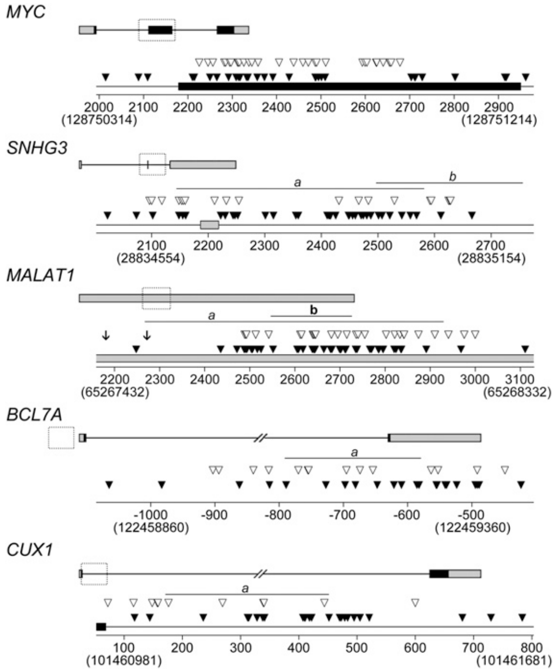

Fig. 3. Somatic mutations and breakpoint distribution in AID target loci. Mutations (open triangles) and breakpoints (filled triangles) detected by LM-PCR (Fig. S2) were plotted on the respective genomic sequences. The top scheme represents exons (rectangles) and introns (bars). Genomic loci are shown in untranslated and translated sequences (gray and black boxes, respectively). The horizontal lines *a* and *b* represent breakage regions identified by promoter array and sequencing, respectively. Regions outlined by dotted boxes are shown in more detail below each genomic locus. For the *MALAT1* locus, the translocation breakpoints reported by Davis et al. (45) are represented by arrows. *x* axis numbers indicate base positions according to RefSeq: NM_002467 (*MYC*), NR_002909 (*SNHG3*), NR_002819 (*MALAT1*), NM_020993 (*BCL7A*), and NM_181552 (*CUX1*). Numbers in parentheses indicate the corresponding base position according to hg19 assembly.

where DNA cleavage and mutations were identified are transcribed (Tables S1 and S2).

Repetitive Sequences Surround the Breakage Regions of Unique Targets. We next examined common features among the AID targets. Although SHM has been reported to prefer the RGYW-WRCY motif (46), we could not find any enrichment of this motif among the break sites in the newly identified targets. It was recently reported that mutations are introduced in regions with sequences prone to forming non-B DNA structure, including tandem repeats, palindromes, and inverted repeats (17, 18). The S region, *MYC,* and V region genes contain sequences prone to forming non-B structure (19, 20, 47, 48). We used REPFIND, a program that identifies clustered, nonrandom short repeats in a given nucleotide sequence, to search the vicinity of identified breakage regions for sequences prone to forming non-B structure. For each repeat cluster, a *P* value is calculated indicating the probability of finding such a repeat cluster randomly (a *P* value of 1×10^{-5} means that such a concentration of that particular repeat occurs an average of once in 100,000 bp by chance) (49). Curiously, we found that various types of repeat sequences cluster in the vicinity of cleaved sites in the newly identified AID target genes. In the *MALAT1* locus, the region within 2 kb surrounding the breakage peaks was rich in clustered short repeat motifs such as GAAG, GCC, GAA, CCG, AAG, GAAGA, and TTAA (Fig. 4). Repeat clusters were also found near the cleavage sites of the *SNHG3*, *BCL7A,* and *CUX1* loci. (Fig. S3). In all cases, the probability of the appearance of these repeats was far below random ($P < 1 \times 10^{-8}$).

H3K4me3 at Cleavage Sites. It was recently shown that S region transcription alone is not sufficient for CSR; specific histone posttranslational modification marks, especially H3K4me3, are required. H3K4me3 depletion strongly inhibits CSR and DNA cleavage in the Sμ and Sα regions (15). We thus asked whether the V region and the newly identified AID targets also carry H3K4me3 marks around the cleavage regions. ChIP analysis showed that both the V region and *MALAT1* locus were abundantly marked by H3K4me3 (Fig. 5). Furthermore, the H3K4me3 distribution profiles corresponded well to the somatic mutation distribution in the rearranged V region and to the breakage signal distribution observed by both the promoter array and whole genome sequencing in *MALAT1* (Fig. 5 *A* and *B*). Mutations identified in *MALAT1* overlapped with DNA cleavage signals and H3K4me3 marks (Figs. 3 and 5*B*). We examined the H3K4me3 pattern of other AID targets by using publicly available ENCODE ChIP-seq data for the B-lymphoblastoid cell line GM12878 (51). As expected, all of them, except for *BCL7A*, were highly abundant in H3K4me3 marks overlapping nicely with cleavage sites (Fig. S4). H3K4me3 might be absent at the *BCL7A* locus in GM12878 cells because it is an inducible gene expressed in BL2 cells, but not in the GM12878 cell line (52). We thus conclude that the newly identified AID targets share both *cis* and *trans* marks for AID targeting—non-B structure and H3K4me3, respectively (15, 16).

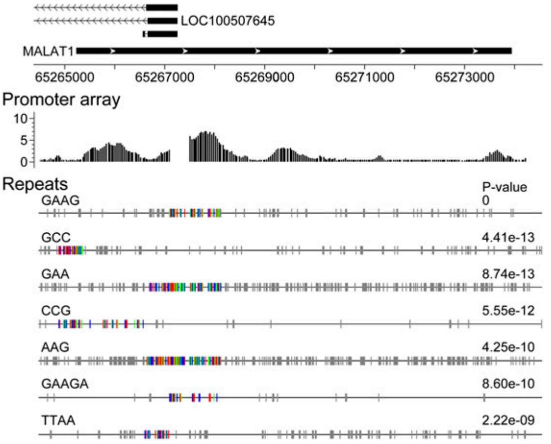

Fig. 4. Repeat sequences surrounding the breakage region in the *MALAT1* gene. (*Top*) Representation of a 10-kb segment surrounding the *MALAT1* locus. *x* axis numbers represent base positions according to hg19 assembly. (*Middle*) Breakage signal distribution detected by promoter array. Regions without bars do not have array probes). (*Bottom*) REPFIND analysis showing significant repeat clusters in the *MALAT1* locus. Motifs depicted as small, colored, vertical bars indicate the cluster with the most significant *P* value; individual repeats are separated by different colors.

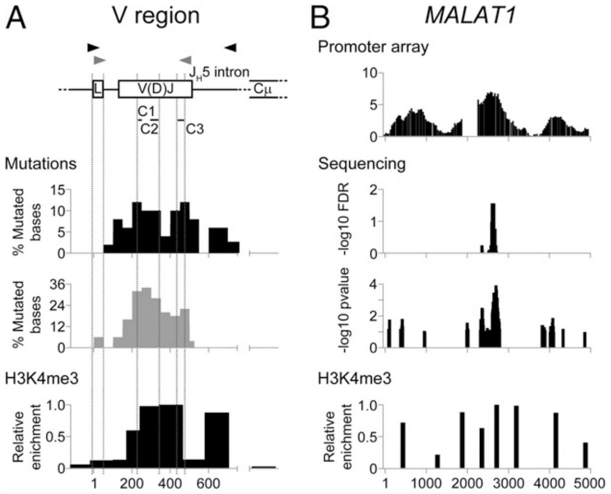

Fig. 5. H3K4me3 distribution in the IgH V region and in the *MALAT1* gene. (*A Top*: Representation of the rearranged IgH V region of BL2 cells. Black and gray arrowheads represent the position of primers used for the mutation analysis shown in *Bottom* (graphs in black and gray, respectively). L, leader; C1, CDR1; C2, CDR2; C3, CDR3. (*A Middle*) Somatic mutation distribution, represented as the percentage of mutated bases per 50 bp sequenced. Graph in black: mutations from Fig. 2, *Inset*. Graph in gray: mutations reported by Denepoux et al. (50). (*Bottom*) ChIP assay using an anti-H3K4me3 antibody. *x* axis numbers indicate the nucleotide position relative to the first V-gene ATG. (*B*) *MALAT1* locus. From top to bottom: Breakage signal distribution detected by promoter array (regions without bars do not have array probes); FDR regions by sequencing; *P* value peaks by sequencing; ChIP assay using an H3K4me3 antibody. *x* axis numbers indicate base positions according to RefSeq NR_002819.

Discussion

Identified AID Targets Accumulate High-Frequency Mutations. We explored AID targets by combining three different strategies: promoter array, whole genome sequencing, and candidate qPCR in a library containing biotinylated linker-labeled cleaved ends. With these strong criteria, we were able to identify four unique AID targets: *SNHG3, MALAT1, BCL7A,* and *CUX1*. All of these candidates were further confirmed to accumulate mutations. These candidates are thus strong AID cleavage targets; however, these genes represent only very efficient AID targets. The use of the biotinylated linker, which efficiently identifies double-strand breakage with close, staggered nicks on opposite strands, may not detect scattered nicks efficiently, and this may limit identification to targets that are efficiently and specifically cleaved within 3 h of AID activation.

Some well-described SHM target genes, including *MYC, BCL6*, *PAX5*, *RHOH,* and *PIM1,* were not detected by either the promoter array or whole genome sequencing. We used qPCR to test whether these genes were enriched in the biotin-labeled

DNA library, but only *MYC* was enriched in the 4-OHT–treated sample (Fig. 2). These genes have been found to be mutated in memory and GC B cells as well as lymphoma cells (24, 32, 36, 37), cells that are expected to be chronically exposed to AID. In addition, the mutation accumulation in tumor cells depends on selection. In contrast, in our study, we exposed BL2 cells to a short treatment (3 h) of 4-OHT, to increase the chance of detecting only efficiently targeted loci. In fact, none of the genes above mentioned mutated more than 1/20th of the 3′ J_H locus even in 6-mo-old Peyer's patch B cells (36).

The unique AID targets accumulate mutations at comparable frequencies with the *Ig* and *MYC* genes. We found that the mutation and cleavage sites are located in similar areas. The results indicate that the cleavage and mutation sites are linked, but not necessarily identical. This observation is consistent with the prediction that SHM is incorporated during the repair phase by error-prone polymerases (53). We confirmed that all of the newly identified AID targets were highly transcribed in BL2 cells. Although the breakage signal detected at the *BCL7A* locus was ≈800 bp upstream of the TSS, we detected both sense and antisense transcripts in this region.

Unique AID Targets also Translocate. Furthermore, it is important to stress that all of these unique candidates have been shown to be the targets of chromosomal translocation in neoplastic cells as shown for the *Ig* locus and *MYC* gene. MALAT1 is overexpressed in several cancers and was reported to be involved in regulating alternative splicing (54). The *MALAT1* locus has been found to harbor chromosomal translocation breakpoints associated with cancer (45, 55) and, interestingly, two reported translocation breakpoints are close to or within the breakage region identified in the present study (Fig. 3). *SNHG3*, a host gene for small nucleolar RNAs (56), is also reported to be involved in translocation, and although the exact position of the translocation breakpoint has not been reported, we can speculate that it is located in the second intron of *SNHG3* because the detected fusion transcript joins the second exon of *SNHG3* with the exon of the 3′ partner gene (57). *BCL7A* and *CUX1* have also been reported to bear chromosomal translocations; however, these translocation breakpoints occur far from the breakage regions identified in this study (58, 59).

Abundant Repetitive Sequences in AID Targets. To identify common features of AID targets, we compared the *MYC*, *SNHG3*, *MALAT1*, *CUX1,* and *BCL7A* genes with the Ig gene locus (the V_H gene and the Sμ region). Sequence analysis identified abundant repetitive sequences surrounding the cleaved regions of AID targets. A typical example is *MALAT1* (Fig. 4): The GAAG, GCC, GAA, CCG, AAG, GAAGA, and TTAA repeats are highly abundant within 2 kb surrounding the break peaks, which also overlap with actual mutation sites. In the *SNHG3* locus, less frequent but longer repeats—GGATTACAG, TTTTTGTATTTT, ATTACAGGC, GCCTC, and TTTTTGTA—are clustered in the proximity of cleavage sites (Fig. S3*A*). *BCL7A* and *CUX1* have GC-rich repeats, such as CGCG, CCGCG, CCCG, and CGGCG (Fig. S2 *B* and *C*). The *MYC* gene, the V region, and the S region are already known to have repetitive sequences or inverted repeats that can form non-B structure when the target is actively transcribed and under an excessive negative superhelical condition (19, 20, 47, 48).

H3K4me3 Marks in AID Targets. Chromatin modifications are also involved in AID targeting. We showed that H3K4 methylation, specifically trimethylation, is critical for DNA cleavage in the S region (15), although Odegard et al. (60) showed that the H3K4 dimethylation (H3K4me2) pattern is similar among VJλ1, Cλ1, and Eλ3-1 and concluded that H3K4me2 is not correlated with SHM. Association of H3K4me3 with the MYC locus was also reported (38). Therefore, we tested whether H3K4me3 modification is also associated with the V region and the unique loci. SHM in V regions typically targets the whole coding V-region segment and extends to its 5′ and 3′ flanking regions. Mutation frequencies rise sharply ≈100 bp downstream of the TSS (at the middle of the leader intron), peak in V(D)J, and then gradually decrease after the immediate 3′ flanking region, becoming undetectable over a distance of ≈1 kb from the rearranged J (61). It is striking that the H3K4me3 profile follows the exact same tendency as SHM distribution in the V region (Fig. 5*A*). H3K4me3 is scarce in the leader exon and intron but present in the highly mutated portion of the V(D)J exon. We also observed that H3K4me3 distribution at the *MALAT1* locus corresponded well with the breakage signal distribution detected by both the promoter array and whole genome sequencing (Fig. 5*B* and Fig. S3*A*). The H3K4me3 pattern of other AID targets also overlaps with cleavage sites (Fig. S3 *B*–*D*). Strikingly, we observed a strong H3K4me3 peak in the 5′ region of the *CUX1* gene (Fig. S4*D*), which does not contain microarray probes. We confirmed that this region also accumulates mutations after 4-OHT treatment (Table S3). It would be interesting to check whether H3K4me3 depletion can decrease AID-induced breaks and mutations in the newly identified AID targets.

We thus conclude that all of these genes, *SNHG3, MALAT1, BCL7A,* and *CUX1*, share unique characteristics that are required for AID targeting: non-B structure as the *cis* element and the H3K4me3 histone modification as the *trans* mark.

Materials and Methods

Labeling of DNA Break Ends by a Biotinylated Linker. The biotin-labeled DNA break assay was performed as described (8) with slight modifications. After nuclear permeabilization, BL2 cells were washed with cold PBS and resuspended in 1× T4 DNA polymerase buffer. Blunting was performed by using T4 DNA Polymerase (Takara). After washing with cold PBS, 4 μL of T4 DNA Ligase (Takara) and 13.4 μL of an annealed biotinylated P1 linker were added, and the cells were incubated overnight at 16 °C. Genomic DNA was purified by phenol:chloroform extraction.

PCR, Real-Time PCR, and LM-PCR. Biotinylated genomic DNA (10 μg) was sonicated (Covaris) and incubated with 10 μL of M-270 Dynabeads (Invitrogen) for 15 min at room temperature. After washing, the beads were resuspended in 15 μL of TE buffer and used as a PCR template. PCR was initiated by denaturing for 5 min at 95 °C followed by 25 cycles (95 °C for 30 s, 60 °C for 30 s, and 72 °C for 30 s) and a final extension at 72 °C for 5 min. SYBR Green Master Mix (Applied Biosystems) was used for real-time PCR.

For LM-PCR, we used a template of 1 μL of beads in a two-round PCR by using linker primer (P1-LM) and gene-specific primers. First-round PCR was initiated by nick translation (72 °C for 20 min), followed by denaturing (95 °C for 5 min), 25 cycles (95 °C for 15 s, 65 °C for 15 s, and 70 °C for 1 min), and a final extension (70 °C for 5 min). Second-round PCR included denaturing (95 °C for 5 min), 20 cycles (95 °C for 15 s, 65 °C for 15 s, and 70 °C for 1 min), and a final extension (70 °C for 7 min). The PCR fragments were purified, cloned with the pGEM-T Easy Vector System (Promega), and sequenced with the ABI PRISM 3130xl Genetic Analyzer (Applied Biosystems). Primers sequences are provided in Table S4–S7.

DNA Preparation for Microarray and SOLiD Sequencing. After sonication of biotin-labeled genomic DNA, sheared ends were blunted by adding T4 DNA polymerase for 30 min at room temperature. DNA was purified by using the PureLink PCR purification Kit (Invitrogen), P2-annealed linker was ligated overnight at 16 °C, DNA was incubated with Dynabeads as described above, and the beads were used for global amplification by following the SOLiD protocol (Applied Biosystems). A summary of general features of the sequenced libraries can be found in Fig. S5 and Table S8.

Accession Codes. Gene Expression Omnibus: microarray data, GSE32027; DNA Data Bank of Japan: sequencing data, DRA000450.

Other material and methods are provided in *SI Materials and Methods*.

ACKNOWLEDGMENTS. We thank Y. Shiraki for manuscript preparation, Dr. H. Nagaoka for sharing unpublished BL2 expression data, and RIKEN Genome Network Analysis Service (GeNAS) for library sequencing using the

SOLiD system (Life Technologies). This research was supported by a RIKEN Omics Science Center from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan research grant (to Y.H.) and MEXT of Japan Grant-in-Aid for Specially Promoted Research 17002015.

1. Muramatsu M, et al. (2000) Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* 102:553–563.
2. Revy P, et al. (2000) Activation-induced cytidine deaminase (AID) deficiency causes the autosomal recessive form of the Hyper-IgM syndrome (HIGM2). *Cell* 102:565–575.
3. Honjo T, Kinoshita K, Muramatsu M (2002) Molecular mechanism of class switch recombination: Linkage with somatic hypermutation. *Annu Rev Immunol* 20:165–196.
4. Teng G, Papavasiliou FN (2007) Immunoglobulin somatic hypermutation. *Annu Rev Genet* 41:107–120.
5. Shinkura R, et al. (2004) Separate domains of AID are required for somatic hypermutation and class-switch recombination. *Nat Immunol* 5:707–712.
6. Barreto V, Reina-San-Martin B, Ramiro AR, McBride KM, Nussenzweig MC (2003) C-terminal deletion of AID uncouples class switch recombination from somatic hypermutation and gene conversion. *Mol Cell* 12:501–508.
7. Ta VT, et al. (2003) AID mutant analyses indicate requirement for class-switch-specific cofactors. *Nat Immunol* 4:843–848.
8. Doi T, et al. (2009) The C-terminal region of activation-induced cytidine deaminase is responsible for a recombination function other than DNA cleavage in class switch recombination. *Proc Natl Acad Sci USA* 106:2758–2763.
9. Jung S, Rajewsky K, Radbruch A (1993) Shutdown of class switch recombination by deletion of a switch region control element. *Science* 259:984–987.
10. Peters A, Storb U (1996) Somatic hypermutation of immunoglobulin genes is linked to transcription initiation. *Immunity* 4:57–65.
11. Betz AG, et al. (1994) Elements regulating somatic hypermutation of an immunoglobulin kappa gene: Critical role for the intron enhancer/matrix attachment region. *Cell* 77:239–248.
12. Zhang J, Bottaro A, Li S, Stewart V, Alt FW (1993) A selective defect in IgG2b switching as a result of targeted mutation of the I gamma 2b promoter and exon. *EMBO J* 12: 3529–3537.
13. Hackett J, Jr., Rogerson BJ, O'Brien RL, Storb U (1990) Analysis of somatic mutations in kappa transgenes. *J Exp Med* 172:131–137.
14. O'Brien RL, Brinster RL, Storb U (1987) Somatic hypermutation of an immunoglobulin transgene in kappa transgenic mice. *Nature* 326:405–409.
15. Stanlie A, Aida M, Muramatsu M, Honjo T, Begum NA (2010) Histone3 lysine4 trimethylation regulated by the facilitates chromatin transcription complex is critical for DNA cleavage in class switch recombination. *Proc Natl Acad Sci USA* 107:22190–22195.
16. Kobayashi M, et al. (2009) AID-induced decrease in topoisomerase 1 induces DNA structural alteration and DNA cleavage for class switch recombination. *Proc Natl Acad Sci USA* 106:22375–22380.
17. Hubert L, Jr., Lin Y, Dion V, Wilson JH (2011) Topoisomerase 1 and single-strand break repair modulate transcription-induced CAG repeat contraction in human cells. *Mol Cell Biol* 31:3105–3112.
18. Zhao J, Bacolla A, Wang G, Vasquez KM (2010) Non-B DNA structure-induced genetic instability and evolution. *Cell Mol Life Sci* 67:43–62.
19. Rogozin IB, Solovyov VV, Kolchanov NA (1991) Somatic hypermutagenesis in immunoglobulin genes. I. Correlation between somatic mutations and repeats. Somatic mutation properties and clonal selection. *Biochim Biophys Acta* 1089:175–182.
20. Wright BE, Schmidt KH, Minnick MF, Davis N (2008) I. VH gene transcription creates stabilized secondary structures for coordinated mutagenesis during somatic hypermutation. *Mol Immunol* 45:3589–3599.
21. Lippert MJ, et al. (2011) Role for topoisomerase 1 in transcription-associated mutagenesis in yeast. *Proc Natl Acad Sci USA* 108:698–703.
22. Takahashi T, Burguiere-Slezak G, Van der Kemp PA, Boiteux S (2011) Topoisomerase 1 provokes the formation of short deletions in repeated sequences upon high transcription in Saccharomyces cerevisiae. *Proc Natl Acad Sci USA* 108:692–697.
23. Okazaki IM, et al. (2003) Constitutive expression of AID leads to tumorigenesis. *J Exp Med* 197:1173–1181.
24. Kotani A, et al. (2005) A target selection of somatic hypermutations is regulated similarly between T and B cells upon activation-induced cytidine deaminase expression. *Proc Natl Acad Sci USA* 102:4506–4511.
25. Ramiro AR, et al. (2004) AID is required for c-myc/IgH chromosome translocations in vivo. *Cell* 118:431–438.
26. Takizawa M, et al. (2008) AID expression levels determine the extent of cMyc oncogenic translocations and the incidence of B cell tumor development. *J Exp Med* 205: 1949–1957.
27. Ishikawa C, Nakachi S, Senba M, Sugai M, Mori N (2011) Activation of AID by human T-cell leukemia virus Tax oncoprotein and the possible role of its constitutive expression in ATL genesis. *Carcinogenesis* 32:110–119.
28. Machida K, et al. (2004) Hepatitis C virus induces a mutator phenotype: Enhanced mutations of immunoglobulin and protooncogenes. *Proc Natl Acad Sci USA* 101: 4262–4267.
29. Epeldegui M, Hung YP, McQuay A, Ambinder RF, Martínez-Maza O (2007) Infection of human B cells with Epstein-Barr virus results in the expression of somatic hypermutation-inducing molecules and in the accrual of oncogene mutations. *Mol Immunol* 44:934–942.
30. Matsumoto Y, et al. (2007) Helicobacter pylori infection triggers aberrant expression of activation-induced cytidine deaminase in gastric epithelium. *Nat Med* 13:470–476.
31. Faili A, et al. (2002) AID-dependent somatic hypermutation occurs as a DNA single-strand event in the BL2 cell line. *Nat Immunol* 3:815–821.
32. Pasqualucci L, et al. (2001) Hypermutation of multiple proto-oncogenes in B-cell diffuse large-cell lymphomas. *Nature* 412:341–346.
33. Pasqualucci L, et al. (2004) Expression of the AID protein in normal and neoplastic B cells. *Blood* 104:3318–3325.
34. Feldhahn N, et al. (2007) Activation-induced cytidine deaminase acts as a mutator in BCR-ABL1-transformed acute lymphoblastic leukemia cells. *J Exp Med* 204:1157–1166.
35. Leuenberger M, et al. (2010) AID protein expression in chronic lymphocytic leukemia/ small lymphocytic lymphoma is associated with poor prognosis and complex genetic alterations. *Mod Pathol* 23:177–186.
36. Liu M, et al. (2008) Two levels of protection for the B cell genome during somatic hypermutation. *Nature* 451:841–845.
37. Shen HM, Peters A, Baron B, Zhu X, Storb U (1998) Mutation of BCL-6 gene in normal B cells by the process of somatic hypermutation of Ig genes. *Science* 280:1750–1752.
38. Yamane A, et al. (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.
39. Staszewski O, et al. (2011) Activation-induced cytidine deaminase induces reproducible DNA breaks at many non-Ig Loci in activated B cells. *Mol Cell* 41:232–242.
40. Pavri R, et al. (2010) Activation-induced cytidine deaminase targets DNA at sites of RNA polymerase II stalling by interaction with Spt5. *Cell* 143:122–133.
41. Ju BG, et al. (2006) A topoisomerase IIbeta-mediated dsDNA break required for regulated transcription. *Science* 312:1798–1802.
42. Nagaoka H, Ito S, Muramatsu M, Nakata M, Honjo T (2005) DNA cleavage in immunoglobulin somatic hypermutation depends on de novo protein synthesis but not on uracil DNA glycosylase. *Proc Natl Acad Sci USA* 102:2022–2027.
43. Woo CJ, Martin A, Scharff MD (2003) Induction of somatic hypermutation is associated with modifications in immunoglobulin variable region chromatin. *Immunity* 19: 479–489.
44. Dalla-Favera R, et al. (1982) Human c-myc onc gene is located on the region of chromosome 8 that is translocated in Burkitt lymphoma cells. *Proc Natl Acad Sci USA* 79:7824–7827.
45. Davis IJ, et al. (2003) Cloning of an Alpha-TFEB fusion in renal tumors harboring the t (6;11)(p21;q13) chromosome translocation. *Proc Natl Acad Sci USA* 100:6051–6056.
46. Rogozin IB, Kolchanov NA (1992) Somatic hypermutagenesis in immunoglobulin genes. II. Influence of neighbouring base sequences on mutagenesis. *Biochim Biophys Acta* 1171:11–18.
47. Tashiro J, Kinoshita K, Honjo T (2001) Palindromic but not G-rich sequences are targets of class switch recombination. *Int Immunol* 13:495–505.
48. Michelotti GA, et al. (1996) Multiple single-stranded cis elements are associated with activated chromatin of the human c-myc gene in vivo. *Mol Cell Biol* 16:2656–2669.
49. Betley JN, Frith MC, Graber JH, Choo S, Deshler JO (2002) A ubiquitous and conserved signal for RNA localization in chordates. *Curr Biol* 12:1756–1761.
50. Denépoux S, et al. (1997) Induction of somatic mutation in a human B cell line in vitro. *Immunity* 6:35–46.
51. Birney E, et al.; ENCODE Project Consortium; NISC Comparative Sequencing Program; Baylor College of Medicine Human Genome Sequencing Center; Washington University Genome Sequencing Center; Broad Institute; Children's Hospital Oakland Research Institute (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447:799–816.
52. Ernst J, et al. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473:43–49.
53. Faili A, et al. (2004) DNA polymerase eta is involved in hypermutation occurring during immunoglobulin class switch recombination. *J Exp Med* 199:265–270.
54. Tripathi V, et al. (2010) The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol Cell* 39: 925–938.
55. Rajaram V, Knezevich S, Bove KE, Perry A, Pfeifer JD (2007) DNA sequence of the translocation breakpoints in undifferentiated embryonal sarcoma arising in mesenchymal hamartoma of the liver harboring the t(11;19)(q11;q13.4) translocation. *Genes Chromosomes Cancer* 46:508–513.
56. Pelczar P, Filipowicz W (1998) The host gene for intronic U17 small nucleolar RNAs in mammals has no protein-coding potential and is a member of the 5′-terminal oligopyrimidine gene family. *Mol Cell Biol* 18:4509–4518.
57. Levin JZ, et al. (2009) Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. *Genome Biol* 10:R115.
58. Zani VJ, et al. (1996) Molecular cloning of complex chromosomal translocation t(8;14; 12)(q24.1;q32.3;q24.1) in a Burkitt lymphoma cell line defines a new gene (BCL7A) with homology to caldesmon. *Blood* 87:3124–3134.
59. Wasag B, Lierman E, Meeus P, Cools J, Vandenberghe P (2011) The kinase inhibitor TKI258 is active against the novel CUX1-FGFR1 fusion detected in a patient with T-lymphoblastic leukemia/lymphoma and t(7;8)(q22;p11). *Haematologica* 96:922–926.
60. Odegard VH, Kim ST, Anderson SM, Shlomchik MJ, Schatz DG (2005) Histone modifications associated with somatic hypermutation. *Immunity* 23:101–110.
61. Lebecque SG, Gearhart PJ (1990) Boundaries of somatic mutation in rearranged immunoglobulin genes: 5′ boundary is near the promoter, and 3′ boundary is approximately 1 kb from V(D)J gene. *J Exp Med* 172:1717–1727.