Classification: BIOLOGICAL SCIENCES, Immunology.

Non-immunoglobulin AID target loci share unique features with immunoglobulin genes

Lucia Kato[1], Nasim A. Begum[1], A. Maxwell Burroughs[2], Tomomitsu Doi[*], Jun Kawai[2], Carsten O. Daub[2], Takahisa Kawaguchi[3], Fumihiko Matsuda[3], Yoshihide Hayashizaki[2], and Tasuku Honjo[1]

[1] Department of Immunology and Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto 606-8501, Japan.

[2] RIKEN Omics Science Center (OSC), RIKEN Yokohama Institute, Yokohama, Kanagawa, 230-0045, Japan.

[3] The Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto 606-8501, Japan.

* Present address: Division of Gastroenterology and Hepatology, Department of Internal Medicine, Keio University School of Medicine, Tokyo, Japan.

Corresponding author: Tasuku Honjo; Phone +81-75-753-4371; Fax +81-75-753-4388; E-mail honjo@mfour.med.kyoto-u.ac.jp

Abstract

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

¥body

Introduction

AID is expressed in germinal center (GC) B cells upon antigen stimulation and is essential for two types of genetic alteration in the immunoglobulin (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 non-templated 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, even though 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). On the other hand, AID C-terminal mutants almost completely lose CSR activity but retain or even increase SHM activity (6, 7). While 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) absolutely 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 type1 (HTLV1), hepatitis C virus (HCV), Epstein-Barr (EB) virus, and *H. pylori* (27-30). Based on these observations, AID is postulated to induce tumorigenesis, especially in B lymphomas and leukemias—and AID is indeed 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 in fact 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 the protein-bound targets are actually cleaved or mutated by AID. There are indications that some genes identified by such approaches are not transcribed (39). Therefore, it is important to reexamine non-Ig AID target genes using a new strategy.

Here, we report four novel AID targets, identified by a combination of new 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 previously (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 hours 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. 1A). 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. 1B).

New AID targets identified by promoter array and whole genome sequencing

As 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, as compared to untreated samples with False Discovery Rate (FDR) values lower than 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 identified more than fifty 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 and S2). 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.

Since 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 (see 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. 1C 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 by the fact that 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 due to 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 employed 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 hours (Fig. 2). We examined whether candidate genes were enriched in the 4-OHT-treated DNA library compared with the non-treated 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 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, since 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 as well as 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 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 actually mutated upon AID activation, we treated BL2-ΔC-AIDER cells with 4-OHT for 24 hours and sequenced regions of about 600 bp around each area with abundant breakage (Fig. S2 and Table S3). Mutations increased in all 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 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 (Figs. 3 and 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 where DNA cleavage and mutations were identified are indeed transcribed (Table S1 and S2).

Repetitive sequences surround the breakage regions of new targets

We next examined common features among the AID targets. Although SHM has been reported to prefer the RGYW-WRCY motif (45), 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, 46, 47). 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 $1x10^{-5}$ means that such a concentration of that particular repeat occurs an average of once in 100,000 bp by chance) (48). 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-values< $1x10^{-8}$).

H3K4me3 at cleavage sites

It was recently shown that S region transcription alone is not sufficient for CSR; specific histone post-translational modification (PTM) 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* (Figs. 5A and B). Mutations identified in *MALAT1* overlapped with DNA cleavage signals and H3K4me3 marks (Figs. 3 and 5B). We examined the H3K4me3 pattern of other AID targets by using publicly available ENCODE ChIP-seq data for the B-lymphoblastoid cell line GM12878 (49). 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 (50). 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).

Discussion

Novel AID targets accumulate high-frequency mutations

We explored novel 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 new AID targets: *SNHG3*, *MALAT1*, *BCL7A*, and *CUX1*. All of these candidates were further confirmed to accumulate mutations. These newly identified 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 hours 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 hours) of 4-OHT, to increase the chance of detecting only efficiently targeted loci. In fact, none of the genes above mentioned mutated more than one twentieth of the 3' J_H locus even in 6 months old Peyer's patch B cells (36).

The newly identified 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 (51). 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 about 800-bp upstream of the TSS, we detected both sense and antisense transcripts in this region.

New AID targets also translocate

Furthermore, it is important to stress that all of these newly isolated candidates have in fact 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 recently reported to be involved in regulating alternative splicing (52).

The *MALAT1* locus has been found to harbor chromosomal translocation breakpoints associated with cancer (53, 54) 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 (snoRNAs) (55), 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 (56). *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 (57, 58).

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. S3A). *BCL7A* and *CUX1* have GC-rich repeats, such as CGCG, CCGCG, CCCG, and CGGCG (Figs. S2B 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, 46, 47).

H3K4me3 marks in AID targets

Chromatin modifications are also involved in AID targeting. We previously showed that H3K4 methylation, specifically trimethylation, is critical for DNA cleavage in the S region (15), although Odegard et al. (59) 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 newly identified 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 about 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 (60). It is striking that the H3K4me3 profile follows the exact same tendency as SHM distribution in the V region (Fig. 5A). 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 (Figs. 5B and S3A). The H3K4me3 pattern of other AID targets also overlaps with cleavage sites (Figs. S3B-D). Strikingly, we observed a strong H3K4me3 peak in the 5' region of the *CUX1* gene (Fig. S4D) 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 previously (8) with slight modifications. After nuclear permeabilization, BL2 cells were washed with cold PBS and re-suspended in 1x T4 DNA polymerase buffer. Blunting was performed 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 reaction 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.

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 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 following the SOLiD protocol (Applied Biosystems).

Accession codes.

GEO: microarray data, GSE32027; DNA Data Bank of Japan (DDBJ): sequencing data,

DRA000450.

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

Acknowledgements

The authors wish to acknowledge Y. Shiraki for manuscript preparation, Dr. H. Nagaoka for sharing unpublished BL2 expression data, the Research Grant for RIKEN Omics Science Center from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan to YH, and RIKEN GeNAS for library sequencing using the SOLiD system (Life Technologies). This research was supported by a MEXT of Japan Grant-in-Aid for Specially Promoted Research 17002015.

Footnotes

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

The authors declare no conflict of interest.

References

1. Muramatsu M, *et al.* (2000) Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* 102(5):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(5):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(7):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(2):501-508.
7. Ta VT, *et al.* (2003) AID mutant analyses indicate requirement for class-switch-specific cofactors. *Nat Immunol* 4(9):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 U S A* 106(8):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(5097):984-987.
10. Peters A & Storb U (1996) Somatic hypermutation of immunoglobulin genes is linked to transcription initiation. *Immunity* 4(1):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(2):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(9):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(1):131-137.
14. O'Brien RL, Brinster RL, & Storb U (1987) Somatic hypermutation of an immunoglobulin transgene in kappa transgenic mice. *Nature* 326(6111):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 U S A* 107(51):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 U S A* 106(52):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(15):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(1):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(2):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(13):3589-3599.
21. Lippert MJ, *et al.* (2011) Role for topoisomerase 1 in transcription-associated mutagenesis in yeast. *Proc Natl Acad Sci U S A* 108(2):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 U S A* 108(2):692-697.
23. Okazaki IM, *et al.* (2003) Constitutive expression of AID leads to tumorigenesis. *J Exp Med* 197(9):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 U S A* 102(12):4506-4511.
25. Ramiro AR, *et al.* (2004) AID is required for c-myc/IgH chromosome

translocations in vivo. *Cell* 118(4):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(9):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(1):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 U S A* 101(12):4262-4267.
29. Epeldegui M, Hung YP, McQuay A, Ambinder RF, & Martinez-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(5):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(4):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(9):815-821.
32. Pasqualucci L, *et al.* (2001) Hypermutation of multiple proto-oncogenes in B-cell diffuse large-cell lymphomas. *Nature* 412(6844):341-346.
33. Pasqualucci L, *et al.* (2004) Expression of the AID protein in normal and neoplastic B cells. *Blood* 104(10):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(5):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(2):177-186.
36. Liu M, *et al.* (2008) Two levels of protection for the B cell genome during somatic hypermutation. *Nature* 451(7180):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(5370):1750-1752.
38. Yamane A, *et al.* (2010) Deep-sequencing identification of the genomic targets of the cytidine deaminase AID and its cofactor RPA in B lymphocytes. *Nat*

Immunol 12(1):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(2):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(1):122-133.
41. Ju BG, *et al.* (2006) A topoisomerase IIbeta-mediated dsDNA break required for regulated transcription. *Science* 312(5781):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 U S A* 102(6):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(4):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 U S A* 79(24):7824-7827.
45. Rogozin IB & Kolchanov NA (1992) Somatic hypermutagenesis in immunoglobulin genes. II. Influence of neighbouring base sequences on mutagenesis. *Biochim Biophys Acta* 1171(1):11-18.
46. Tashiro J, Kinoshita K, & Honjo T (2001) Palindromic but not G-rich sequences are targets of class switch recombination. *Int Immunol* 13(4):495-505.
47. 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(6):2656-2669.
48. Betley JN, Frith MC, Graber JH, Choo S, & Deshler JO (2002) A ubiquitous and conserved signal for RNA localization in chordates. *Curr Biol* 12(20):1756-1761.
49. Birney E, *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447(7146):799-816.
50. Ernst J, *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473(7345):43-49.
51. Faili A, *et al.* (2004) DNA polymerase eta is involved in hypermutation occurring during immunoglobulin class switch recombination. *J Exp Med* 199(2):265-270.

52. Tripathi V, *et al.* (2010) The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol Cell* 39(6):925-938.
53. 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 U S A* 100(10):6051-6056.
54. 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(5):508-513.
55. 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(8):4509-4518.
56. 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(10):R115.
57. 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(8):3124-3134.
58. 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(6):922-926.
59. Odegard VH, Kim ST, Anderson SM, Shlomchik MJ, & Schatz DG (2005) Histone modifications associated with somatic hypermutation. *Immunity* 23(1):101-110.
60. 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(6):1717-1727.
61. Denepoux S, *et al.* (1997) Induction of somatic mutation in a human B cell line in vitro. *Immunity* 6(1):35-46.

Figure Legends

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 hours, and the break ends were labeled. PCR of Sμ and *B2M* was performed with biotin-labeled DNA or input DNA using five-fold 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. 2. Quantitative PCR 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 and ***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 hours. Only

unique mutations were counted. Detailed mutation profiles can be found in Fig. S2 and Table S3.

Fig. 3. Somatic mutations and breakpoint distribution in AID target loci. Mutations (open triangles) and breakpoints (closed 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. (53) 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.

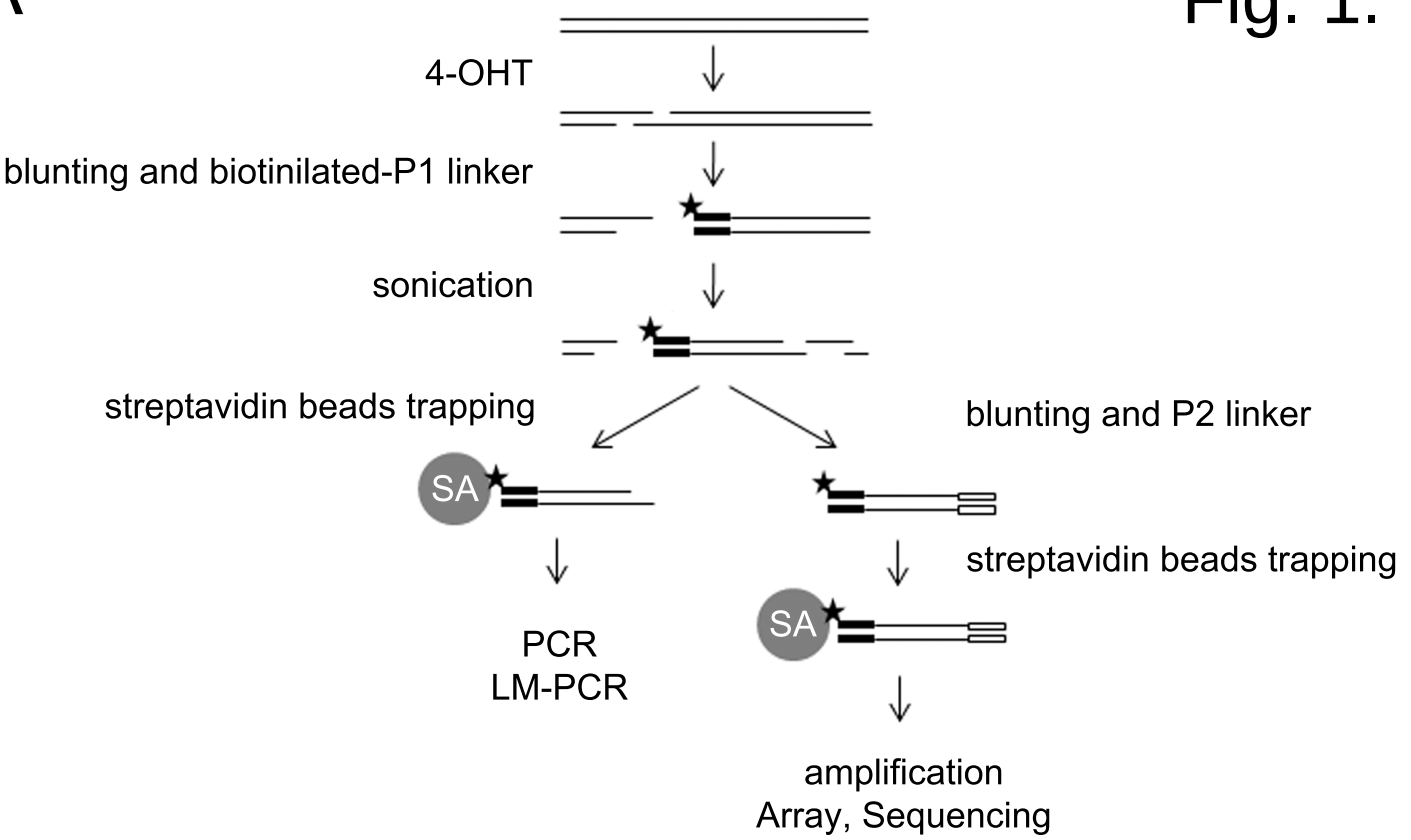
Fig. 4. Repeat sequences surrounding the breakage region in the *MALAT1* gene. *Top:* Representation of a 10-kb segment surrounding the *MALAT1* locus. X-axis numbers

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.

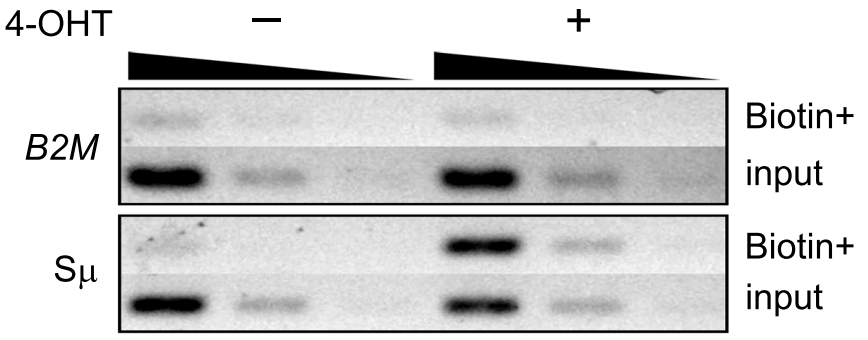
Fig. 5. H3K4me3 distribution in the IgH V region and in the *MALAT1* gene. (A) *Top*: Representation of the rearranged IgH V region of BL2 cells. Black and gray arrowheads represent the position of primers used for the mutation analysis shown in the bottom panels (graphs in black and gray, respectively). L, leader; C1, CDR1; C2, CDR2; C3, CDR3. *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. (61). *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.

Fig. 1.

A


4-OHT
blunting and biotinilated-P1 linker
sonication
streptavidin beads trapping
blunting and P2 linker
SA
SA
streptavidin beads trapping
PCR
LM-PCR
amplification
Array, Sequencing

B


4-OHT
–
+
B2M
Biotin+
input
Sμ
Biotin+
input

C

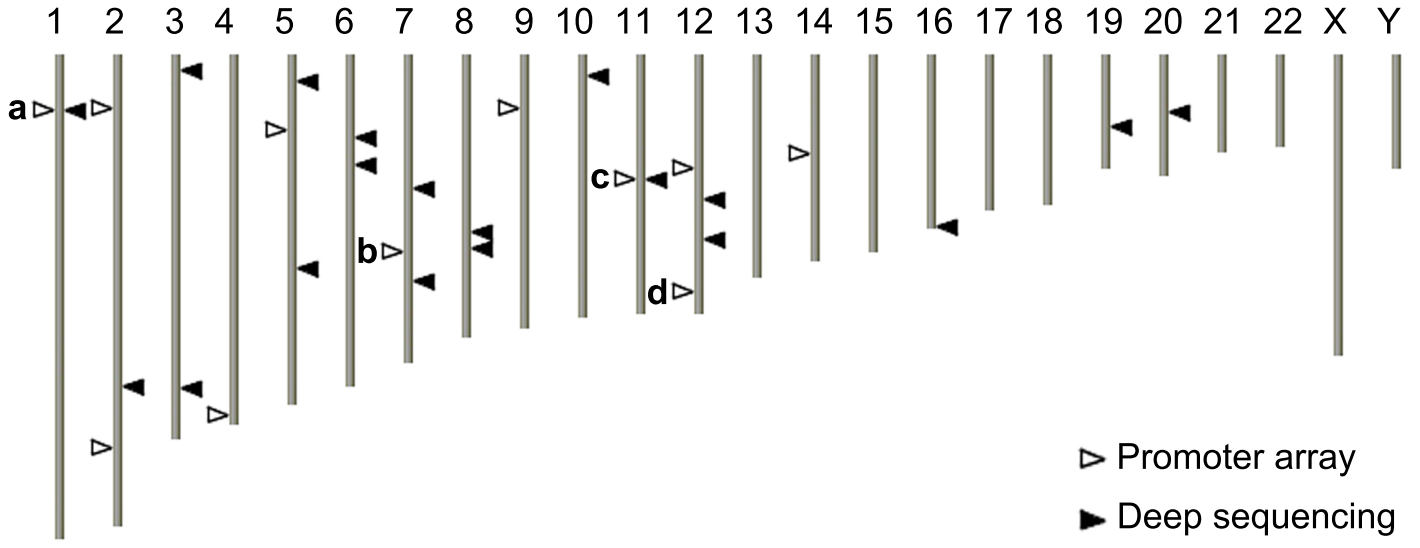

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
a
b
c
d
Promoter array
Deep sequencing

Fig. 2.

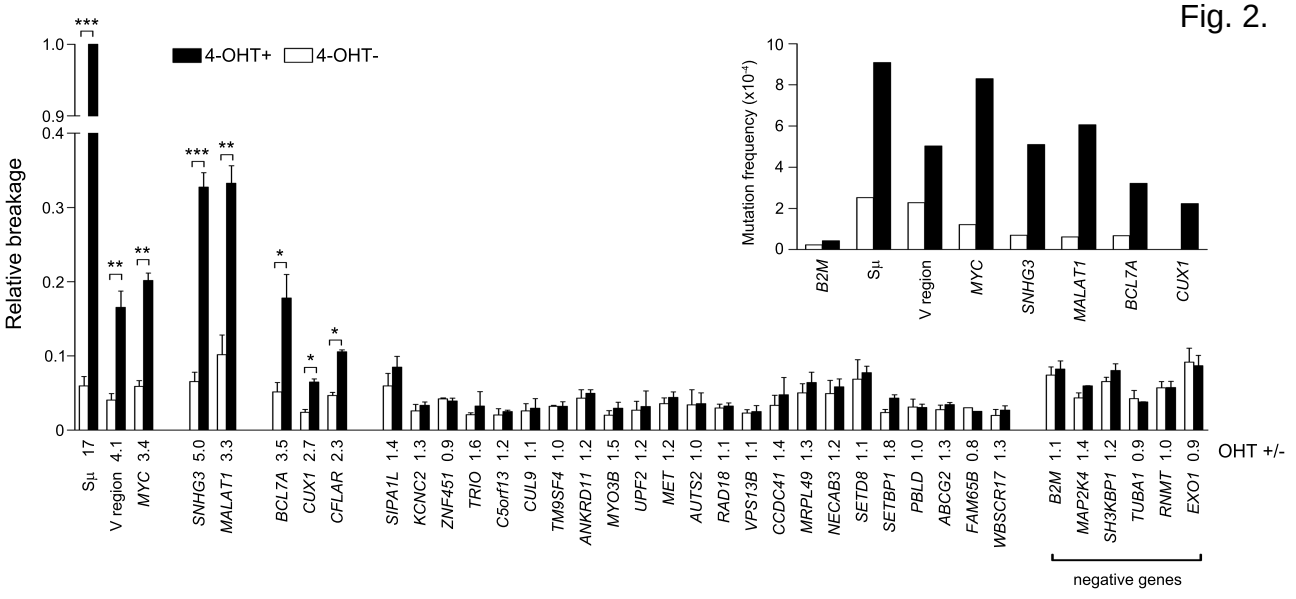

4-OHT+
4-OHT-
Relative breakage
Mutation frequency (x10⁻⁴)
B2M
Sμ
V region
MYC
SNHG3
MALAT1
BCL7A
CUX1
OHT +/-
negative genes

Fig. 3.

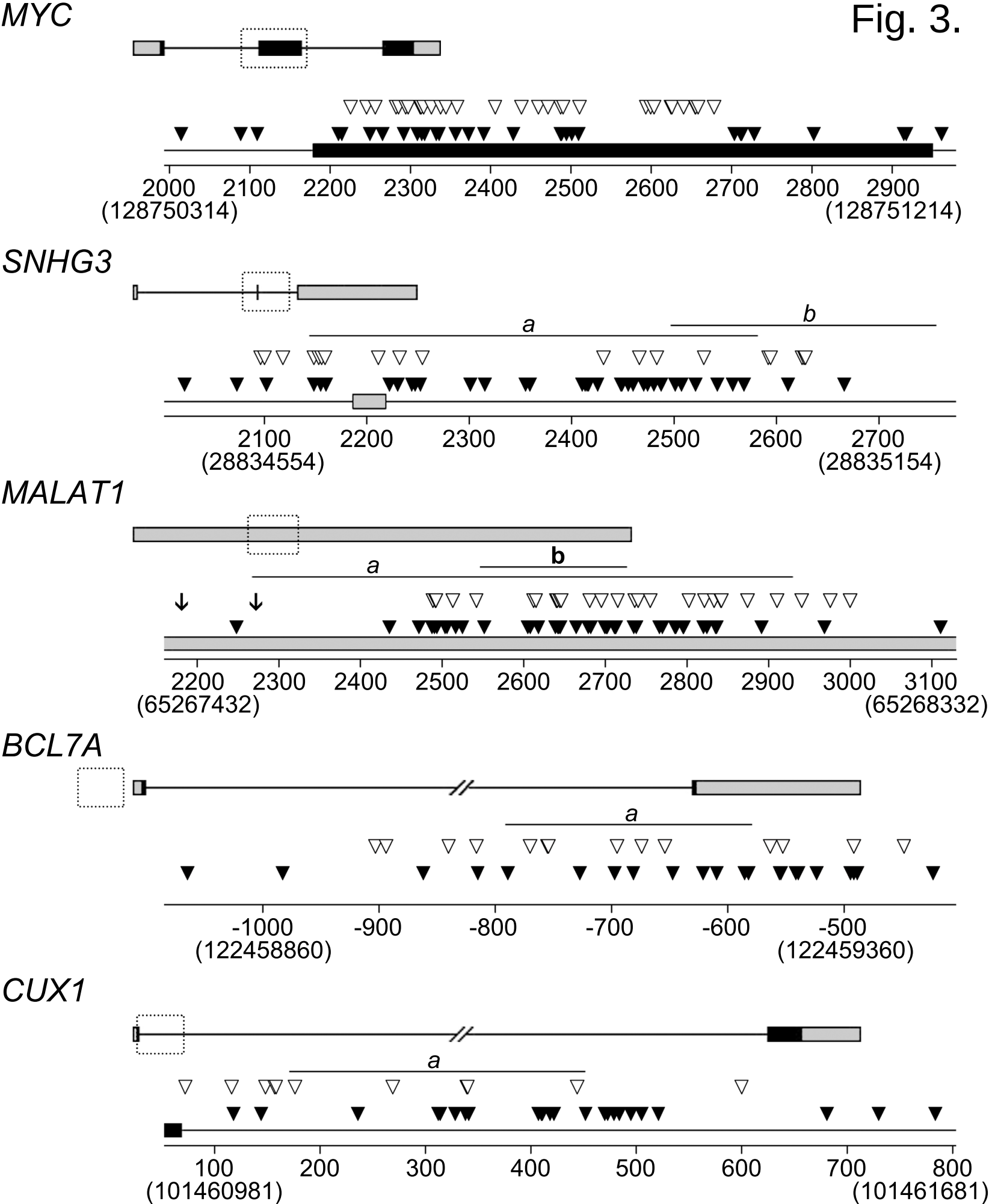
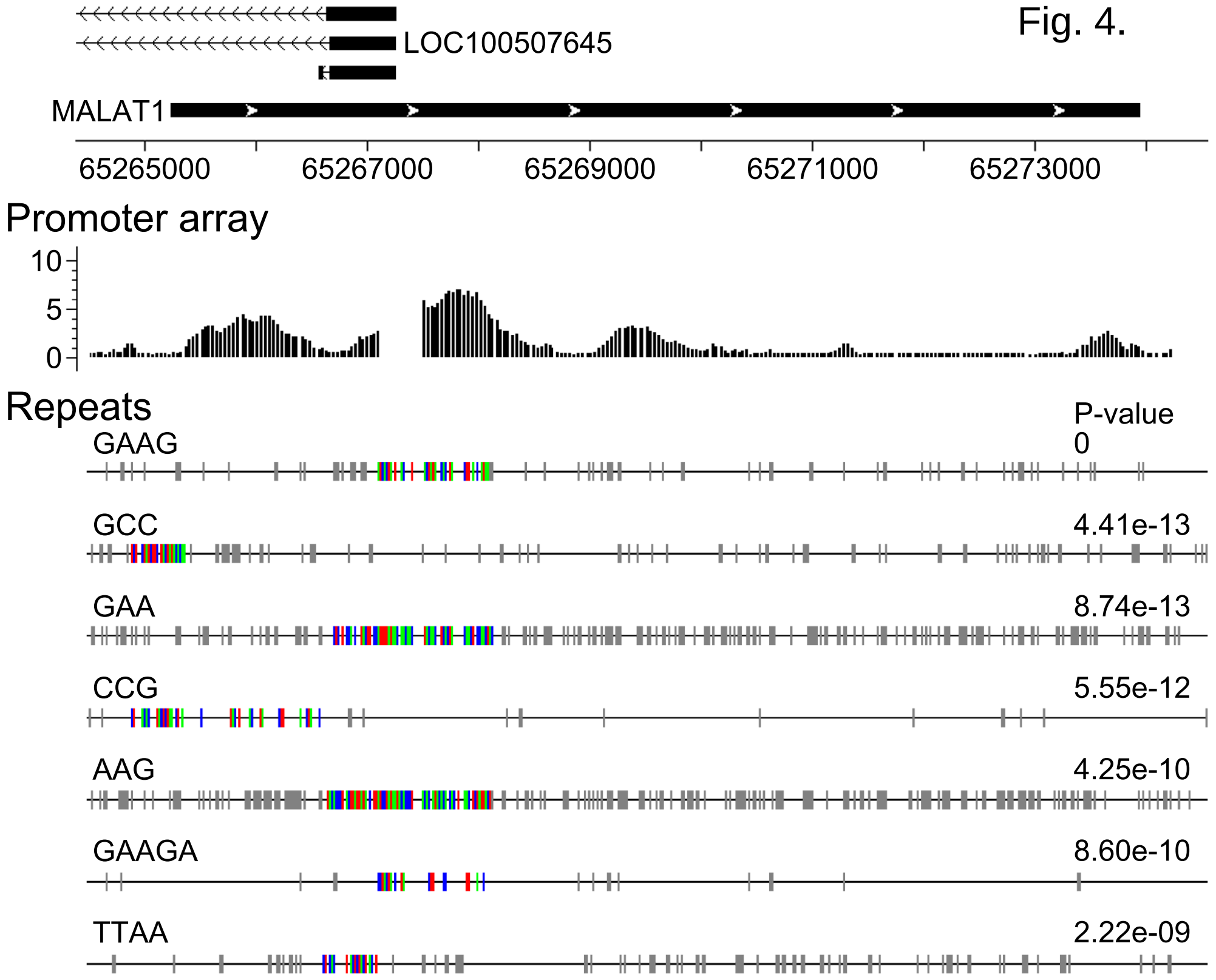
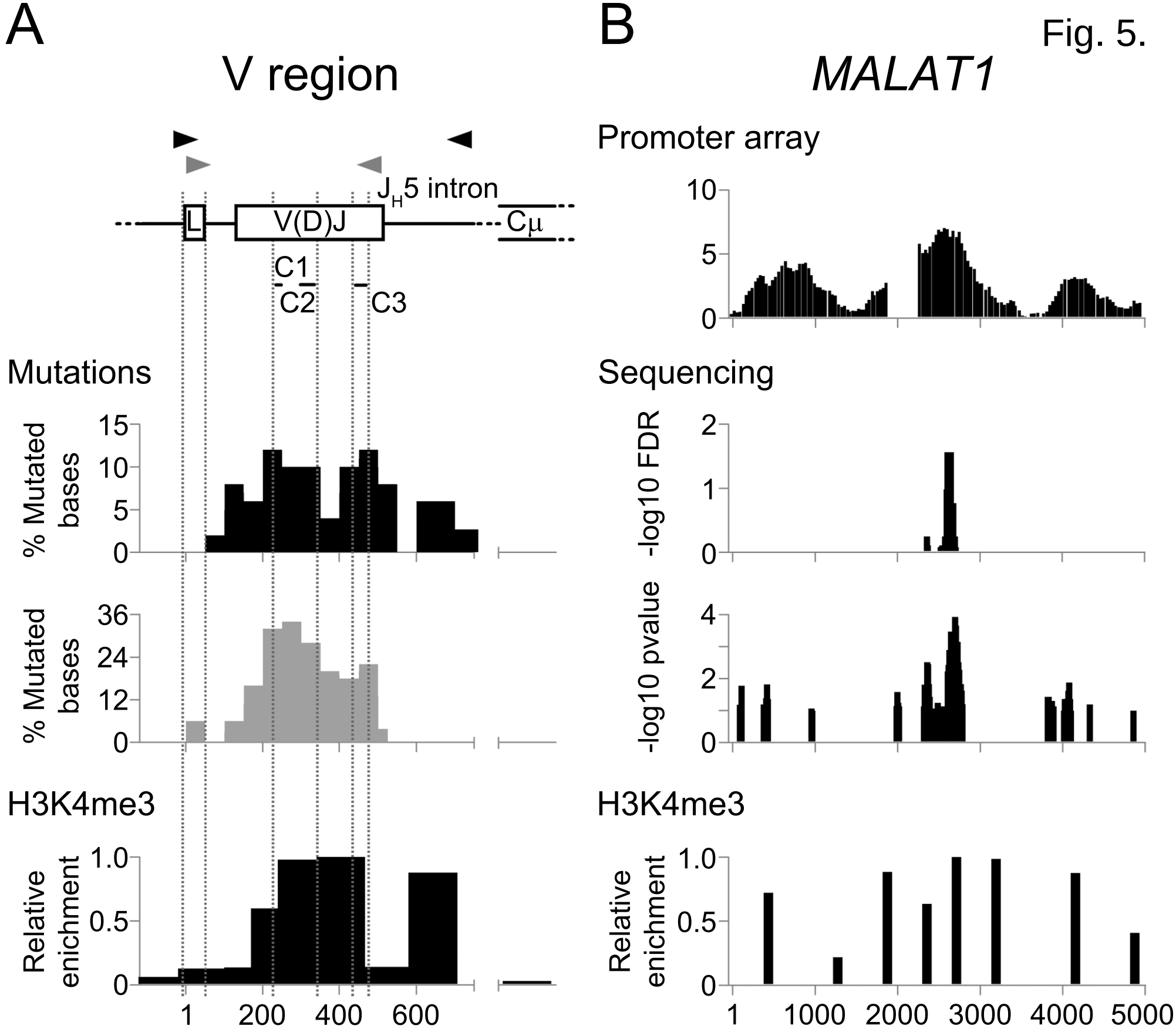

MYC
2000 2100 2200 2300 2400 2500 2600 2700 2800 2900
(128750314)
(128751214)
SNHG3
a
b
2100 2200 2300 2400 2500 2600 2700
(28834554)
(28835154)
MALAT1
a
b
2200 2300 2400 2500 2600 2700 2800 2900 3000 3100
(65267432)
(65268332)
BCL7A
a
-1000 -900 -800 -700 -600 -500
(122458860)
(122459360)
CUX1
a
100 200 300 400 500 600 700 800
(101460981)
(101461681)

Fig. 4.


LOC100507645
MALAT1
65265000
65267000
65269000
65271000
65273000
Promoter array
10
5
0
Repeats
P-value
GAAG
0
GCC
4.41e-13
GAA
8.74e-13
CCG
5.55e-12
AAG
4.25e-10
GAAGA
8.60e-10
TTAA
2.22e-09


A
V region
B
MALAT1
Fig. 5.
J_H5 intron
L
V(D)J
Cμ
C1
C2
C3
Promoter array
10
5
0
Mutations
% Mutated bases
15
10
5
0
% Mutated bases
36
24
12
0
Sequencing
-log10 FDR
2
1
0
-log10 pvalue
4
2
0
H3K4me3
Relative enichment
1.0
0.5
0
1
200
400
600
H3K4me3
Relative enichment
1.0
0.5
0
1
1000
2000
3000
4000
5000

SI Material and Methods

Cell culture.

The BL2 Burkitt's lymphoma cell line was cultured in RPMI medium 1640 containing 10% FCS, 100 μM NEAA (Gibco), 1 mM sodium pyruvate, 100 units/ml penicillin, and 100 μg/ml streptomycin. BL2-ΔC-AIDER cells, which are a BL2 clone harboring Jp8BdelER (Jp8Bdel fused with the hormone-binding domain of the estrogen receptor [ER]), were cultured with 0.5 μg/ml puromycin. To activate AIDER proteins, 4-hydroxytamoxifen (4-OHT) was added to a concentration of 1 μM.

Microarray.

DNA was labeled according to the Affymetrix sample preparation protocol and hybridized to Human Promoter 1.0R Arrays. Slides were scanned with an Affymetrix GeneChip scanner. Six independent experiments were performed for each sample. Microarray data were analyzed using CisGenome.

Analysis of microarray data.

Normalized signal intensities were generated from Affymetrix CEL files using CisGenome (1) . The parameters used to detect peaks are as follows: moving average

(MA) is used to combine neighboring probes, false discovery rates is estimated from permutation test, the window boundary is set to 250 bp, the MA cutoff is set to 3, the max allowable gap within a region is set to 200 bp, the max run of insignificant probes with a region is set to 5, the minimum region length is set to 150 bp, and the minimum number of significant probes within a region is set to 5. We retained only peaks where the FDR is below 0.3. All probe sequences were mapped to the hg19 genome assembly.

SOLiD sequencing.

Templated microbeads were prepared according to Applied Biosystem's standard protocol. Sequencing runs were performed on the SOLiD 3 Plus system (Applied Biosystems) under standard conditions.

Analysis of SOLiD DNA-sequencing data.

Linkers were removed from the sequenced tags using a custom Perl script, and the resulting tags were mapped to the human genome (assembly hg19) using the Bowtie program with standard parameters (2) . Breakage points were summed over 100-bp intervals across the entire genome in 10-bp increments for each control and 4-OHT-treated replicate. Significant differences in breakage frequency were calculated

using the EdgeR program (3), with background dispersion values for both conditions calculated over shifting 10000-bp intervals. This program provided p-values and FDR values for each 100-bp interval, measuring the likelihood of observing the differences in breakage points across the two conditions given the selected background breakage rates. Regions with low FDR values overlapping the promoter (defined as 500 bases upstream of the RefSeq-defined transcriptional start site) and gene definitions were extracted. To increase confidence in the potential AID targets, regions with significant p-values were clustered via a single-linkage clustering procedure, which joined any regions within 1 kb of another peak on the genome. The resulting clusters of regions were then overlaid with low-FDR regions, and the most promising candidate genes containing the lowest FDR values (FDR <0.1), high numbers of p-value clusters, or both were selected for further validation testing (Table S2). This layer of additional clustering was chosen to facilitate the testing of genes with the highest likelihood of breakage. Indeed, both the *MALAT1* and *SNHG3* loci (see Fig. S3A) displayed these patterns, which were likely to be consistent with high levels of AID-induced cleavage activity.

Statistical parameters of SOLiD DNA-sequencing data.

In terms of data integrity, our technique is consistent with existing technologies. A

summary of the characteristics of the sequenced libraries is provided in Table S5. The read redundancy ranges between 1.23 and 1.28; this is slightly higher than most ChIP-seq experiments (1.05-1.15) but much lower than other sequencing technologies. The slightly higher redundancy rates likely result from high numbers of breakages found in repeat regions (Fig. S5A). Importantly, the low redundancy values argue against problems arising from the number of cycles used, indicating the library does not suffer from amplification bias. An overview annotating the genomic locations of breakage sites in each library is also provided (Fig. S5B). The source of the discrepancies in target identification between this dataset and others instead likely stems from lack of depth. Each sequenced library covers the entire genome at roughly 0.8X coverage. While the coverage of the as-currently undefined AID-targeted “break-ome” is not known (our manuscript represents the first attempt to define this), at present coverage levels it appears possible to detect regions which are substantially affected by AID-induced breakage while not capturing everything which has been previously identified.

We have plotted the distribution of the square root of the absolute common dispersion values across the two conditions in all local genome windows which were used to approximate local background breakage rates for comparisons across samples

(Fig. S5C). Very few genomic regions exceed an expected error rate of 20% in the measurements. We observe a bimodal distribution in the error rates, with the first peak corresponding to extremely low expected error (<1%) and the second peak found between ~1-20%. The bimodal distribution implies that the AID breakage in a subset of genome regions is more highly reproducible relative to other regions; interestingly, this is not derived from relative expression values in the two peaks. Regardless, in either peak region the error rates are consistent with existing deep-sequencing techniques and enable robust identification of significant differences between breakage counts across conditions.

Mutation analysis.

To analyze SHM mutations, BL2-ΔC-AIDER cells were treated with 4-OHT (1 μM) for 24 hours, and the genomic DNA was purified by phenol:chloroform extraction. PCR was performed using *Pyrobest* or PrimeSTAR GXL DNA polymerase (TaKaRa) with the following amplification conditions: 95 °C for 5 min, 30 cycles at 95 °C for 30 s, 58 °C for 30 s, and 68 °C for 1 min. After purification, the PCR fragments were A-tailed and cloned with the pGEM-T Easy Vector System (Promega). Nucleotide sequences were determined with the ABI PRISM 3130xl Genetic Analyzer (Applied Biosystems).

Only unique mutations were counted, and the mutation frequency was calculated from the number of mutations per total bases analyzed.

REPFIND analysis.

Analysis with the REPFIND Web server (http://zlab.bu.edu/repfind/form.html) used the following parameters: p-value cutoff, 0.0001; minimum repeat length, 3; maximum repeat length, infinity; low complexity filter, on; statistical background, query sequence; order of background Markov model, 1.

ChIP.

ChIP was performed as described previously (4).

1. Ji H & Wong WH (2005) TileMap: create chromosomal map of tiling array hybridizations. *Bioinformatics* **21,** 3629-3636.
2. Langmead B, Trapnell C, Pop M, & Salzberg SL (2009) Searching for SNPs with cloud computing. *Genome Biol* **10,** R25.
3. Robinson MD, McCarthy DJ, & Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139-140.
4. 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 U S A **107,** 22190-22195.

SI Figure Legends

Fig. S1

Breakage signal distribution detected by promoter array. A 10-kb segment in the vicinity of detected breakage region is represented for each locus. Regions without array bars do not have array probes. X-axis numbers indicate base positions according to hg19 assembly.

Fig. S2

Somatic mutations and breakpoint distribution in AID target loci. Red, mutations found in cells treated with 4-OHT for 24 hours; green, mutations in 4-OHT non-treated samples; blue, insertions; red line, deletions; arrowheads, break sites in samples treated with 4-OHT for 3 hours; arrows, translocation breakpoints shown in Fig. 3.

Fig. S3

Repeat sequences surrounding the breakage region in AID target genes. **A.** *SNHG3* gene. From top to bottom: Representation of a 10-kb segment surrounding the *SNHG3* locus; Breakage signal distribution detected by promoter array (regions without bars do not have array probes); FDR regions by sequencing; p-value peaks by sequencing;

REPFIND analysis showing significant repeat clusters in the *SNHG3* locus. **B and C.** *BCL7A* and *CUX1* genes, respectively. From top to bottom: Representation of a 10-kb segment surrounding the cleavage region for each gene; Breakage signal distribution detected by promoter array (regions without bars do not have array probes); REPFIND analysis showing significant repeat clusters in the same region. Motifs depicted as vertical small colored bars indicate the cluster with the most significant p-value; individual repeats are separated by different colors. X-axis numbers indicate base positions according to hg19 assembly.

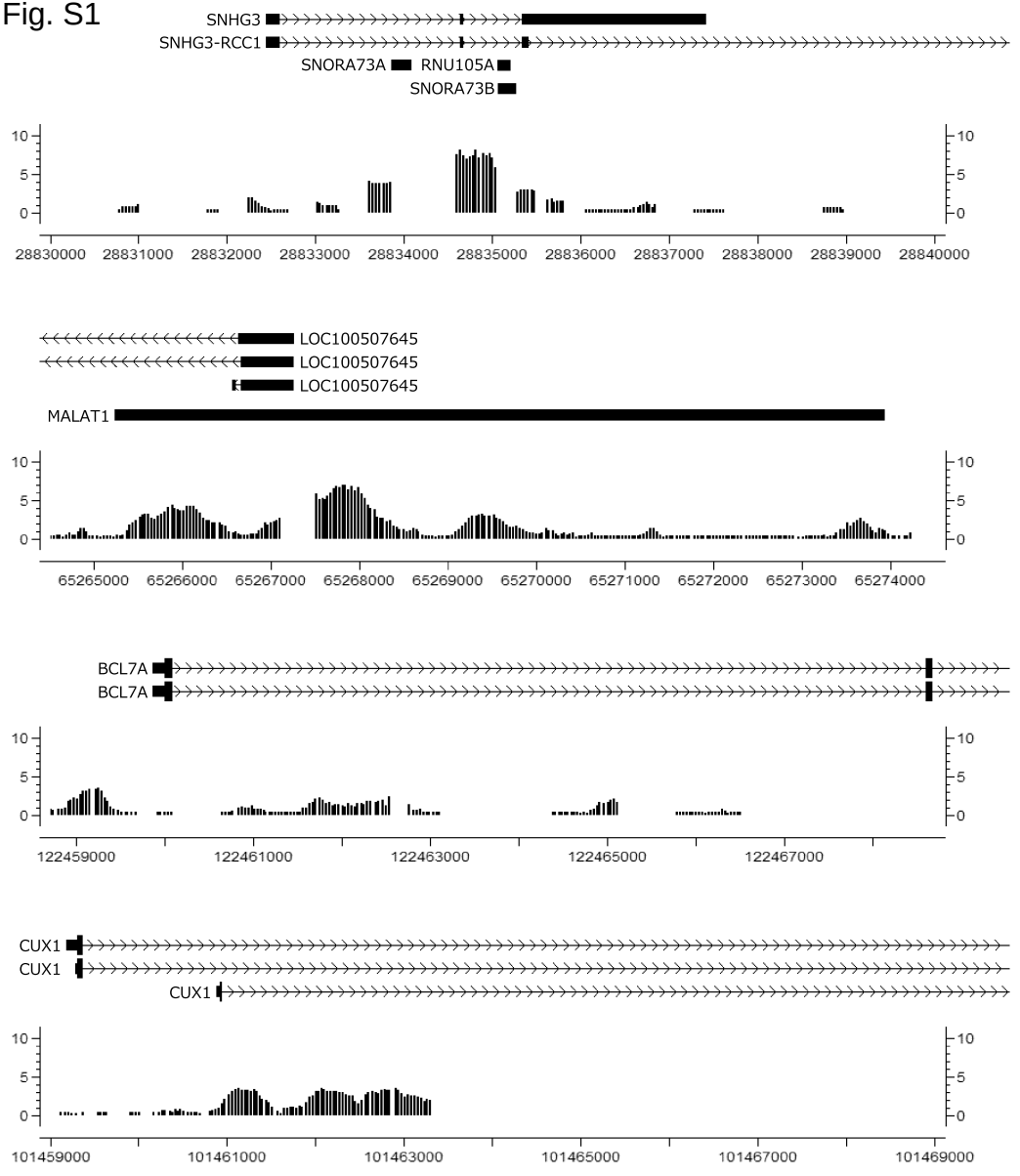
Fig. S4

H3K4me3 distribution in AID target genes. **A.** *MALAT1* gene. From top to bottom: Representation of a 10-kb segment surrounding the *MALAT1* locus; Breakage signal distribution detected by promoter array (regions without bars do not have array probes); ChIP assay using an anti-H3K4me3 antibody; H3K4me3 status from ENCODE ChIP-seq data for GM12878 cell line. **B-D.** *SNHG3*, *BCL7A* and *CUX1* genes, respectively. X-axis numbers indicate base positions according to hg19 assembly.

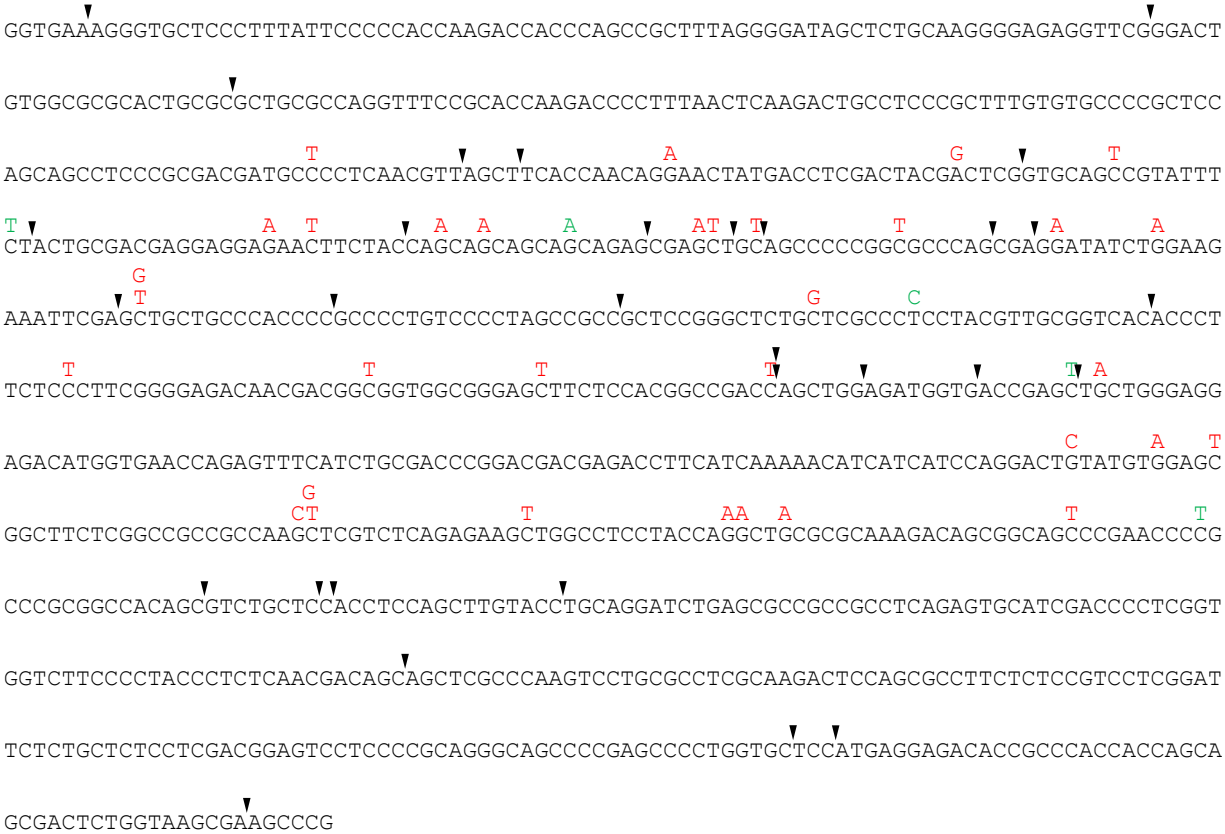
Fig. S5

A. Location of all mapped reads according to genome features and **B.** distribution of gene-mapping reads to genomic locations. **C.** Distribution of estimated percent error in abundance measurements across all analyzed genomic segments. The estimated percent error in breakage rates rarely exceeds 15-20%, enabling robust determination of significant differences.

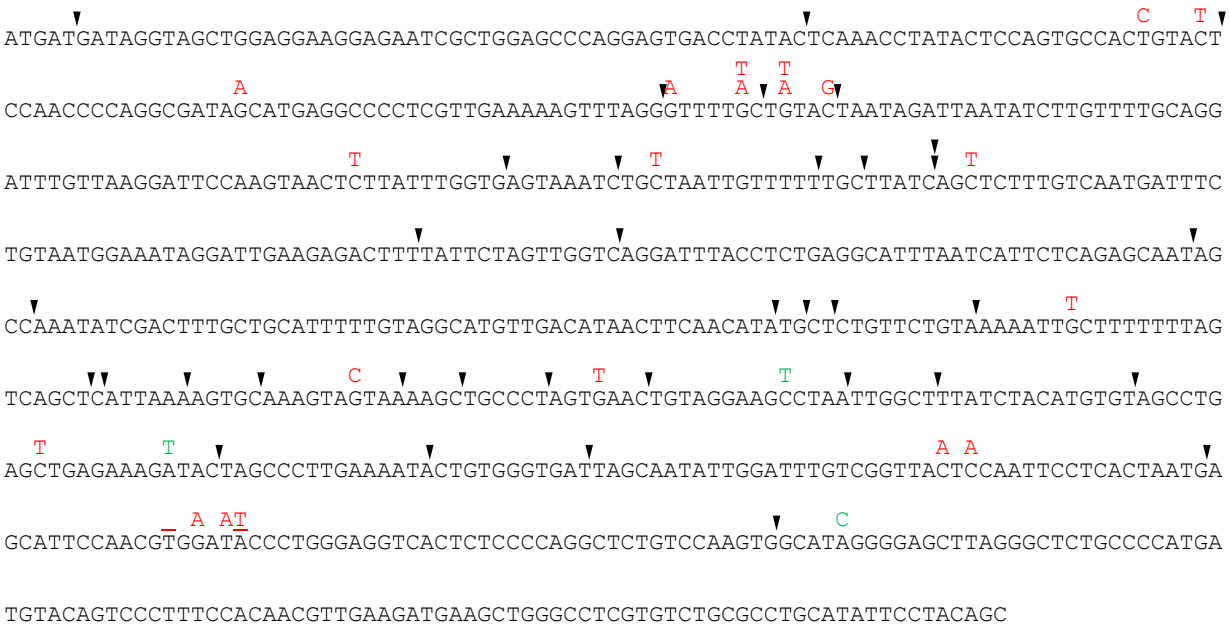
Fig. S1


SNHG3
SNHG3-RCC1
SNORA73A
RNU105A
SNORA73B
10
5
0
28830000
28831000
28832000
28833000
28834000
28835000
28836000
28837000
28838000
28839000
28840000
LOC100507645
LOC100507645
LOC100507645
MALAT1
65265000
65266000
65267000
65268000
65269000
65270000
65271000
65272000
65273000
65274000
BCL7A
BCL7A
122459000
122461000
122463000
122465000
122467000
CUX1
CUX1
CUX1
101459000
101461000
101463000
101465000
101467000
101469000

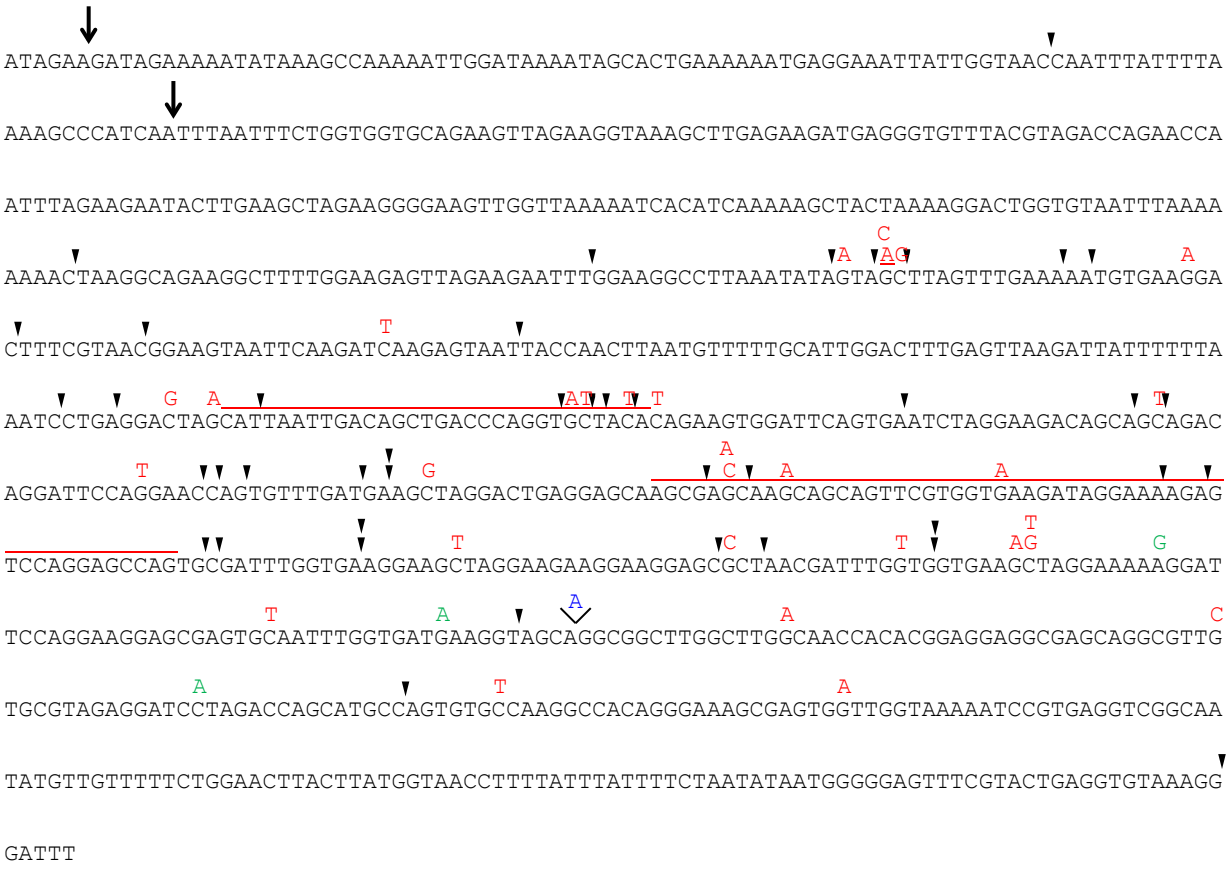
MYC


GGTGAAAGGGTGCTCCCTTTATTCCCCCACCAAGACCACCCAGCCGCTTTAGGGGATAGCTCTGCAAGGGGAGAGGTTCGGGACT
GTGGCGCGCACTGCGCGCTGCGCCAGGTTTCCGCACCAAGACCCCTTTAACTCAAGACTGCCTCCCGCTTTGTGTGCCCCGCTCC
AGCAGCCTCCCGCGACGATGCCCCTCAACGTTAGCTTCACCAACAGGAACTATGACCTCGACTACGACTCGGTGCAGCCGTATTT
CTACTGCGACGAGGAGGAGAACTTCTACCAGCAGCAGCAGAGCGAGCTGCAGCCCCCGGCGCCCAGCGAGGATATCTGGAAG
AAATTCGAGCTGCTGCCCACCCCGCCCCTGTCCCCTAGCCGCCGCTCCGGGCTCTGCTCGCCCTCCTACGTTGCGGTCACACCCT
TCTCCCTTCGGGGAGACAACGACGGCGGTGGCGGGAGCTTCTCCACGGCCGACCAGCTGGAGATGGTGACCGAGCTGCTGGGAGG
AGACATGGTGAACCAGAGTTTCATCTGCGACCCGGACGACGAGACCTTCATCAAAAACATCATCATCCAGGACTGTATGTGGAGC
GGCTTCTCGGCCGCCGCCAAGCTCGTCTCAGAGAAGCTGGCCTCCTACCAGGCTGCGCGCAAAGACAGCGGCAGCCCGAACCCCG
CCCGCGGCCACAGCGTCTGCTCCACCTCCAGCTTGTACCTGCAGGATCTGAGCGCCGCCGCCTCAGAGTGCATCGACCCCTCGGT
GGTCTTCCCCTACCCTCTCAACGACAGCAGCTCGCCCAAGTCCTGCGCCTCGCAAGACTCCAGCGCCTTCTCTCCGTCCTCGGAT
TCTCTGCTCTCCTCGACGGAGTCCTCCCCGCAGGGCAGCCCCGAGCCCCTGGTGCTCCATGAGGAGACACCGCCCACCACCAGCA
GCGACTCTGGTAAGCGAAGCCCG

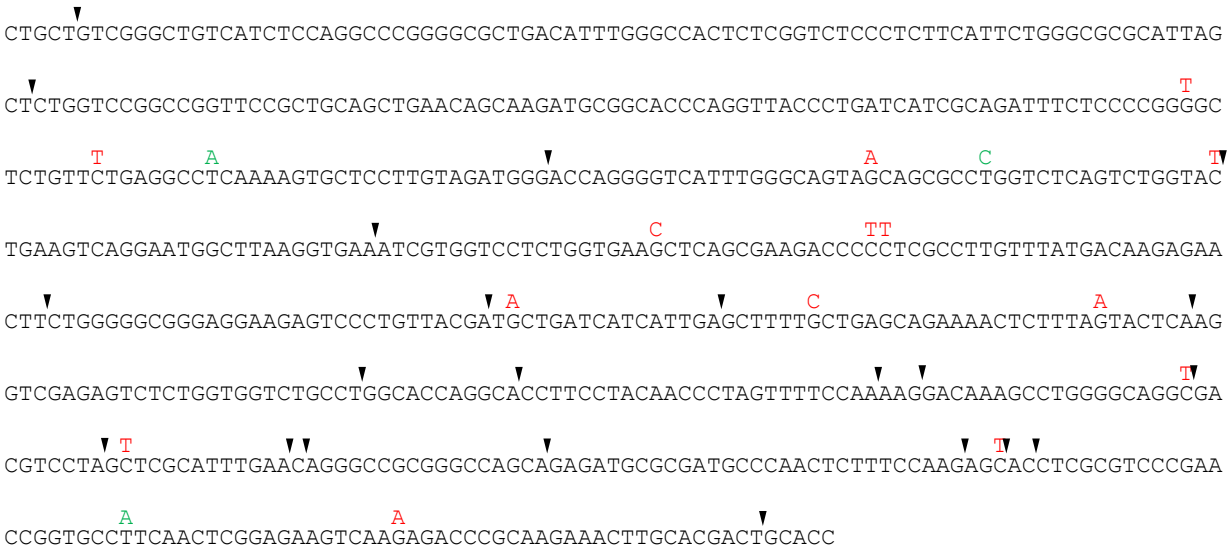
SNHG3


ATGATGATAGGTAGCTGGAGGAAGGAGAATCGCTGGAGCCCAGGAGTGACCTATACTCAAACCTATACTCCAGTGCCACTGTACT
CCAACCCCAGGCGATAGCATGAGGCCCCTCGTTGAAAAAGTTTAGGGTTTTGCTGTACTAATAGATTAATATCTTGTTTTGCAGG
ATTTGTTAAGGATTCCAAGTAACTCTTATTTGGTGAGTAAATCTGCTAATTGTTTTTTGCTTATCAGCTCTTTGTCAATGATTTC
TGTAATGGAAATAGGATTGAAGAGACTTTTATTCTAGTTGGTCAGGATTTACCTCTGAGGCATTTAATCATTCTCAGAGCAATAG
CCAAATATCGACTTTGCTGCATTTTTGTAGGCATGTTGACATAACTTCAACATATGCTCTGTTCTGTAAAAATTGCTTTTTTTAG
TCAGCTCATTAAAAGTGCAAAGTAGTAAAAGCTGCCCTAGTGAACTGTAGGAAGCCTAATTGGCTTTATCTACATGTGTAGCCTG
AGCTGAGAAAGATACTAGCCCTTGAAAATACTGTGGGTGATTAGCAATATTGGATTTGTCGGTTACTCCAATTCCTCACTAATGA
GCATTCCAACGTGGATACCCTGGGAGGTCACTCTCCCCAGGCTCTGTCCAAGTGGCATAGGGGAGCTTAGGGCTCTGCCCCATGA
TGTACAGTCCCTTTCCACAACGTTGAAGATGAAGCTGGGCCTCGTGTCTGCGCCTGCATATTCCTACAGC

MALAT1



BCL7A



CUX1

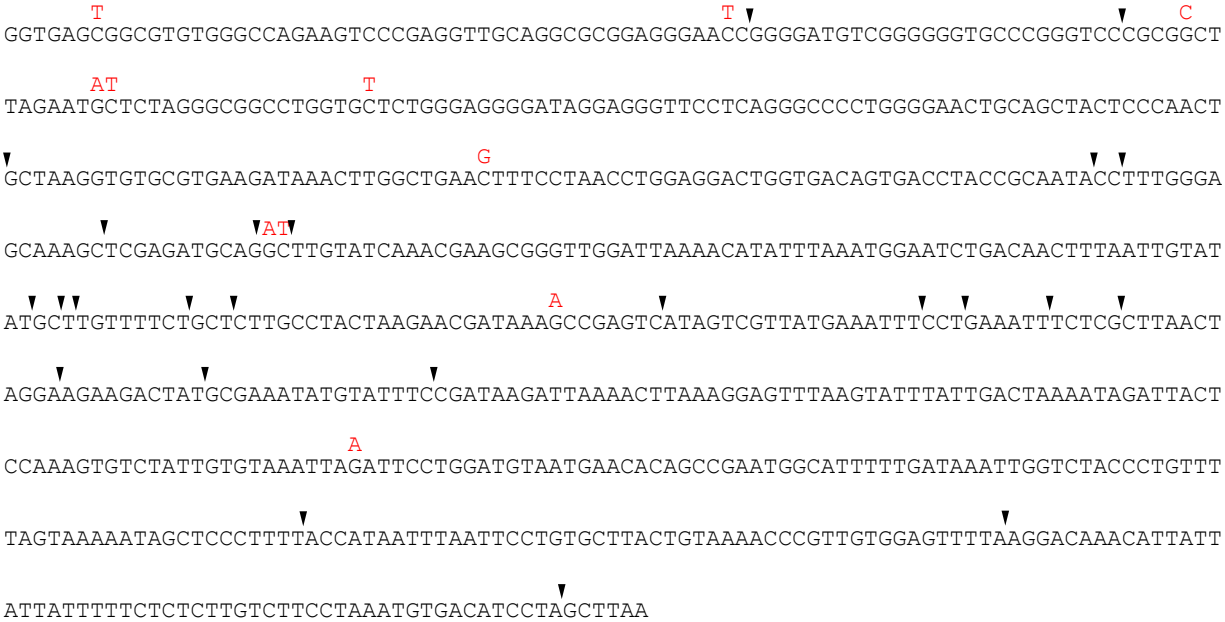

T
T
C
GGTGAGCGGCGTGTGGGCCAGAAGTCCCGAGGTTGCAGGCGCGGAGGGAACCGGGGATGTCGGGGGGTGCCCGGGTCCCGCGGCT
AT
T
TAGAATGCTCTAGGGCGGCCTGGTGCTCTGGGAGGGGATAGGAGGGTTCCTCAGGGCCCCTGGGGAACTGCAGCTACTCCCAACT
G
GCTAAGGTGTGCGTGAAGATAAACTTGGCTGAACTTTCCTAACCTGGAGGACTGGTGACAGTGACCTACCGCAATACCTTTGGGA
AT
GCAAAGCTCGAGATGCAGGCTTGTATCAAACGAAGCGGGTTGGATTAAAACATATTTAAATGGAATCTGACAACTTTAATTGTAT
A
ATGCTTGTTTTCTGCTCTTGCCTACTAAGAACGATAAAGCCGAGTCATAGTCGTTATGAAATTTCCTGAAATTTCTCGCTTAACT
AGGAAGAAGACTATGCGAAATATGTATTTCCGATAAGATTAAAACTTAAAGGAGTTTAAGTATTTATTGACTAAAATAGATTACT
A
CCAAAGTGTCTATTGTGTAAATTAGATTCCTGGATGTAATGAACACAGCCGAATGGCATTTTTGATAAATTGGTCTACCCTGTTT
TAGTAAAAATAGCTCCCTTTTACCATAATTTAATTCCTGTGCTTACTGTAAAACCCGTTGTGGAGTTTTAAGGACAAACATTATT
ATTATTTTTCTCTCTTGTCTTCCTAAATGTGACATCCTAGCTTAA

Fig. S3

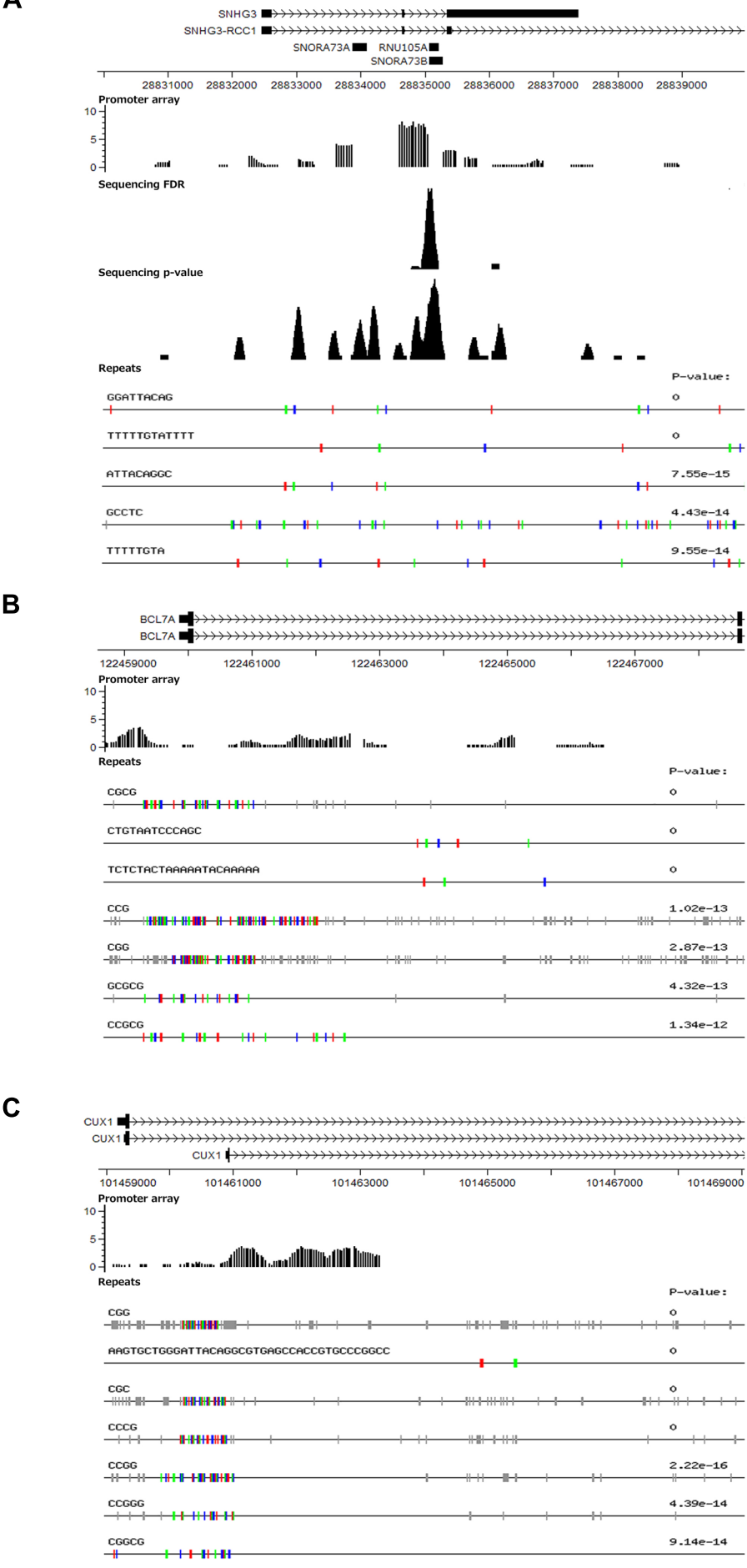
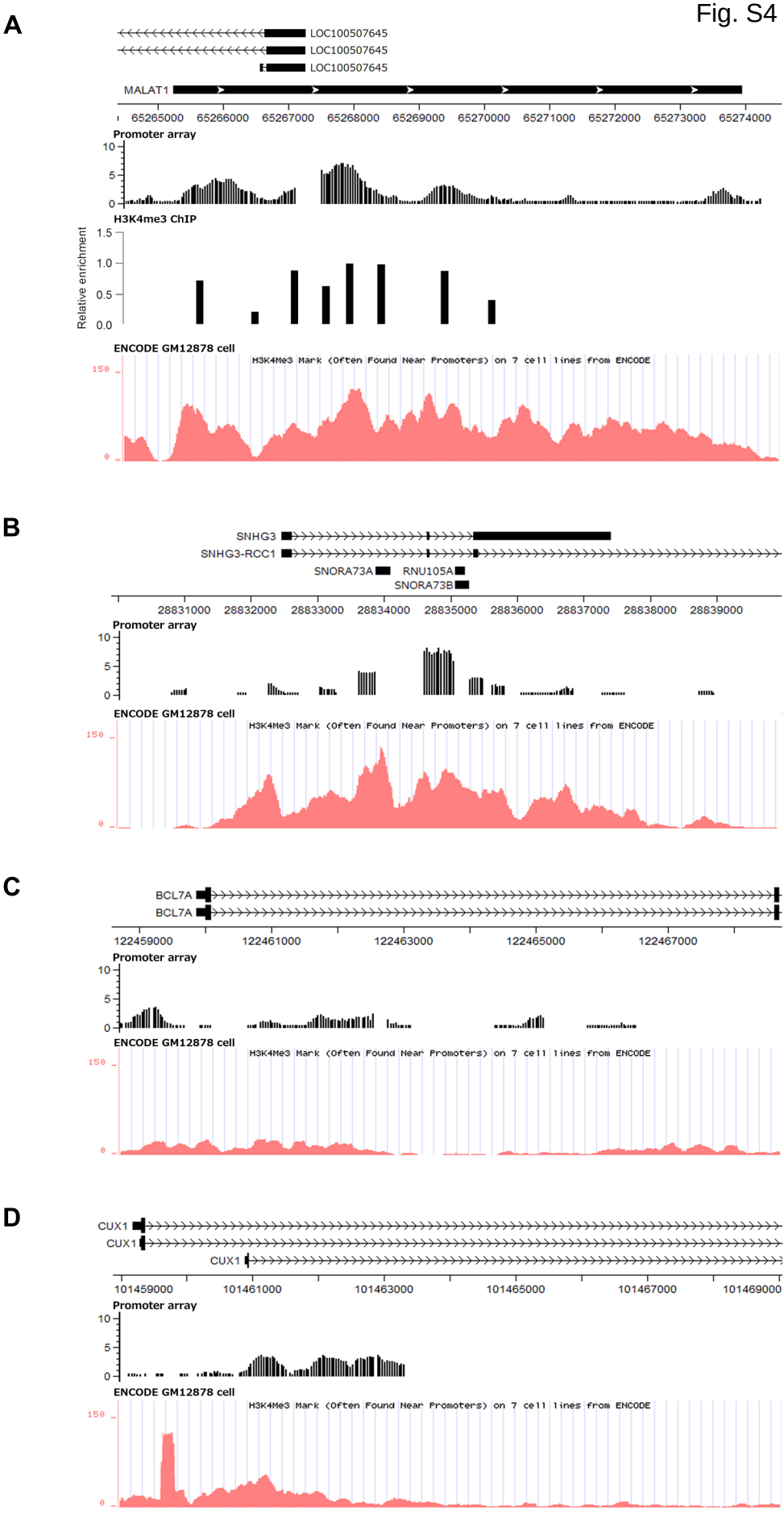
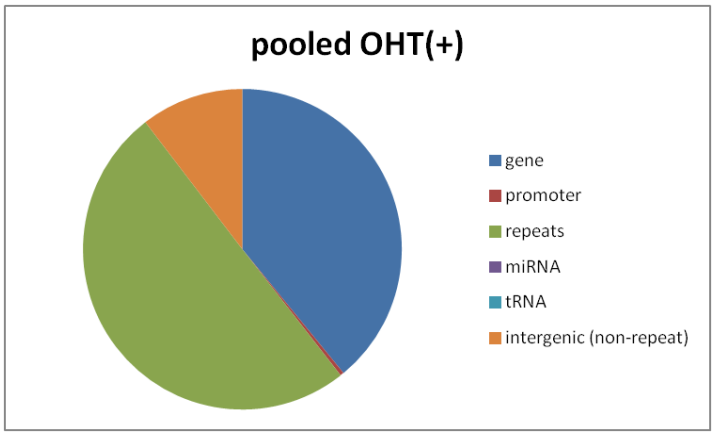
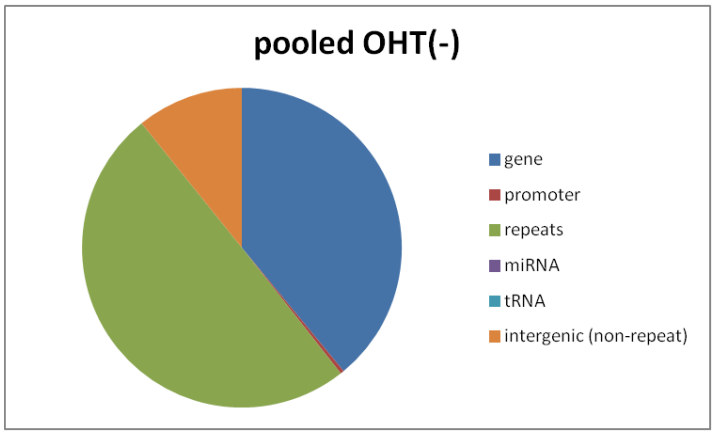

A
SNHG3
SNHG3-RCC1
SNORA73A
RNU105A
SNORA73B
28831000 28832000 28833000 28834000 28835000 28836000 28837000 28838000 28839000
Promoter array
10
5
0
Sequencing FDR
Sequencing p-value
Repeats
P-value:
GGATTACAG 0
TTTTTGTATTTT 0
ATTACAGGC 7.55e-15
GCCTC 4.43e-14
TTTTTGTA 9.55e-14
B
BCL7A
BCL7A
122459000 122461000 122463000 122465000 122467000
Promoter array
10
5
0
Repeats
P-value:
CGCG 0
CTGTAATCCCAGC 0
TCTCTACTAAAAATACAAAAA 0
CCG 1.02e-13
CGG 2.87e-13
GCGCG 4.32e-13
CCGCG 1.34e-12
C
CUX1
CUX1
CUX1
101459000 101461000 101463000 101465000 101467000 101469000
Promoter array
10
5
0
Repeats
P-value:
CGG 0
AAGTGCTGGGATTACAGGCGTGAGCCACCGTGCCCGGCC 0
CGC 0
CCCG 0
CCGG 2.22e-16
CCGGG 4.39e-14
CGGCG 9.14e-14

Fig. S4

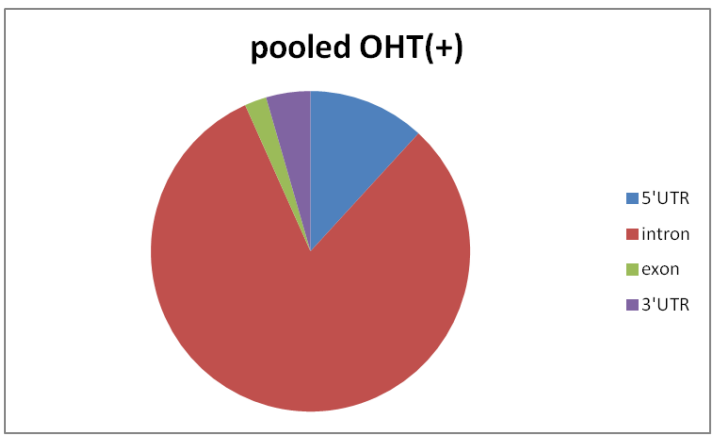

A
LOC100507645
LOC100507645
LOC100507645
MALAT1
65265000 65266000 65267000 65268000 65269000 65270000 65271000 65272000 65273000 65274000
Promoter array
10
5
0
H3K4me3 ChIP
Relative enrichment
1.5
1.0
0.5
0.0
ENCODE GM12878 cell
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
150
0
B
SNHG3
SNHG3-RCC1
SNORA73A
RNU105A
SNORA73B
28831000 28832000 28833000 28834000 28835000 28836000 28837000 28838000 28839000
Promoter array
10
5
0
ENCODE GM12878 cell
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
150
0
C
BCL7A
BCL7A
122459000 122461000 122463000 122465000 122467000
Promoter array
10
5
0
ENCODE GM12878 cell
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
150
0
D
CUX1
CUX1
CUX1
101459000 101461000 101463000 101465000 101467000 101469000
Promoter array
10
5
0
ENCODE GM12878 cell
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
150
0

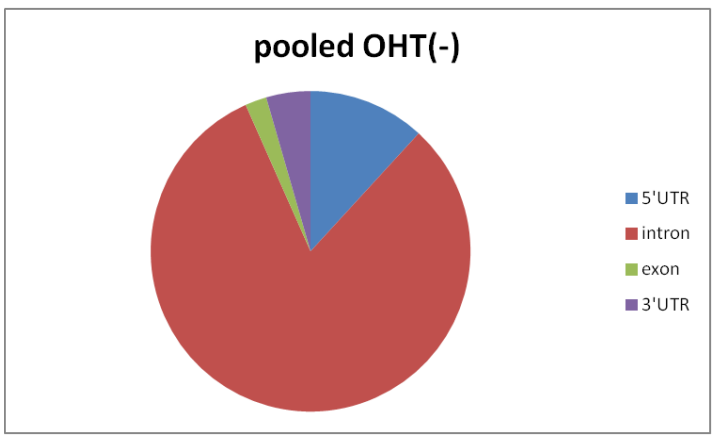
A


pooled OHT(+)
gene
promoter
repeats
miRNA
tRNA
intergenic (non-repeat)


pooled OHT(-)
gene
promoter
repeats
miRNA
tRNA
intergenic (non-repeat)

B


pooled OHT(+)
5'UTR
intron
exon
3'UTR


pooled OHT(-)
5'UTR
intron
exon
3'UTR

C

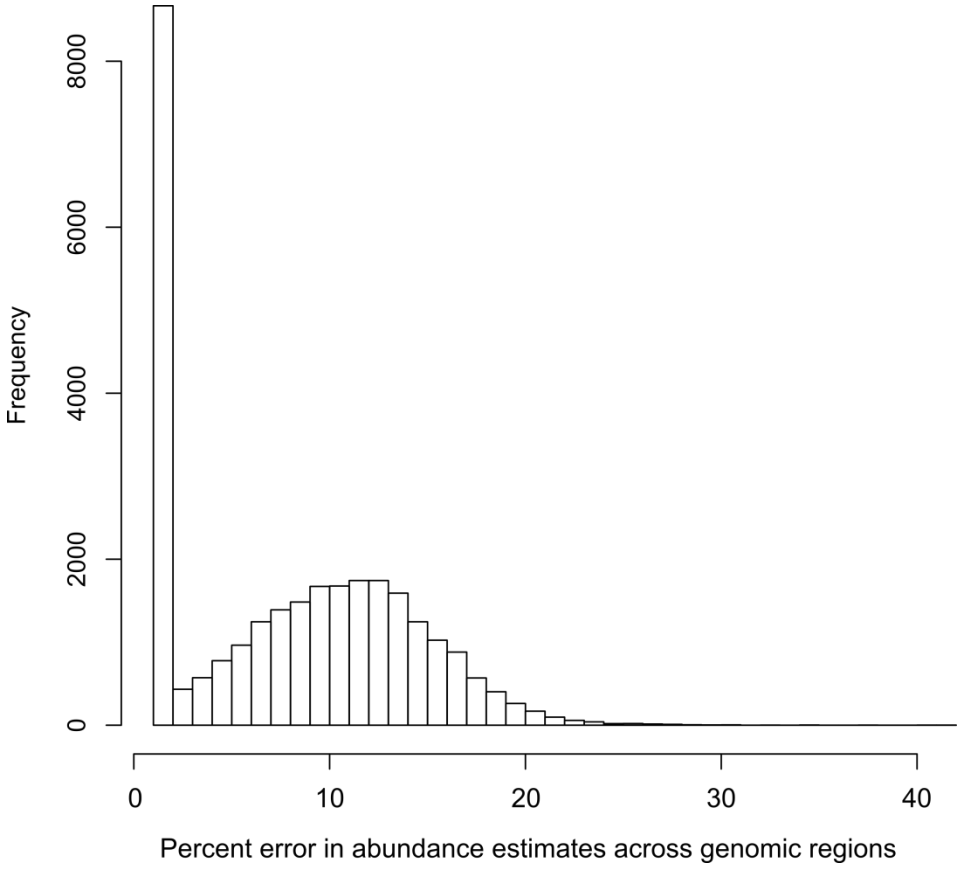

Frequency
0
2000
4000
6000
8000
0
10
20
30
40
Percent error in abundance estimates across genomic regions

Table S1. AID targets identified by promoter assay.

gene	chr	Expression	Tilemap maxM/P	FDR value	Validation by qPCR	Translocation (partner gene)
SNHG3	chr1	+	8.18	0	+	CML (*PICALM*)
MALAT1	chr11	+	7.05	0	+	renal cell carcinoma (*TFEB*), mesenchymal hamartoma of the liver (*ACAT2*)
NIN	chr14	+	6.25	0.08	-	CML-like myeloproliferative disorder (*PDGFRB*)
FYB	chr5	-	6.25	0.08	n.d	
C2orf16	chr2	-	6.05	0.1	-	
C9orf72	chr9	+	5.93	0.11	-	
FAM119B	chr12	-	5.62	0.2	n.d	
CFLAR	chr2	+	5.47	0.26	+	
SNX25	chr4	+	5.45	0.26	-	
BCL7A	chr12	+	3.53	0.83	+	Burlitt's lymphoma (IgH; *MYC*-IgH)
CUX1	chr7	+	3.51	0.83	+	TLL (*FGFR1*)

Abbreviations: CML, Chronic Myeloid Leukemia; TLL, T-Lymphoblastic Leukemia/Lymphoma; n.d, not done. FDR<0.3 plus *BCL7A* and *CUX1.*

Table S2. AID targets identified by sequencing.

gene	chr	Expression	FDR value	p-value clustering*	Validation by qPCR	Translocation (partner gene)
MALAT1	chr11	+	0.027	12	+	renal cell carcinoma (*TFEB*), mesenchymal hamartoma of the liver (*ACAT2*)
SNHG3	chr1	+	0.015	9	+	CML (*PICALM*)
SIPA1L3	chr19	+	0.015	5	-	
KCNC2	chr12	-	0.005	4	-	
ZNF451	chr6	+	0.038	4	-	
TRIO	chr5	+	0.050	4	-	
C5orf13	chr5	+	0.072	4	-	
CUL9	chr6	+	0.077	4	-	
TM9SF4	chr20	+	0.026	3	-	
ANKRD11	chr16	+	0.041	3	-	
MYO3B	chr2	-	0.082	3	-	
UPF2	chr10	+	0.093	3	-	
MET	chr7	-	0.000	2	-	gastric carcinoma (*TPR*)
AUTS2	chr7	+	0.017	2	-	ALL (*PAX5*)
RAD18	chr3	+	0.038	1	-	
OTUD6B	chr8	+	0.041	1	n.d	
VPS13B	chr8	+	0.044	1	-	
CCDC41	chr12	+	0.055	1	-	
ECT2	chr3	+	0.078	1	-	
MRPL49	chr11	+	0.020	1	-	close to t(11;17)(q13;q21) translocation in B-NHL
NECAB3	chr20	+	0.029	1	-	
SETD8	chr12	+	0.175	1	-	
SETBP1	chr18	-	0.134	9	-	
PBLD	chr10	-	0.254	8	-	
ABCG2	chr4	-	0.220	8	-	
FAM65B	chr6	+	0.195	8	-	
WBSCR17	chr7	-	0.162	8	-	
ERC1	chr12	+	0.153	6	-	
PLD2	chr17	+	0.102	5	-	

* Number of p-value clusters corresponding to the FDR region. Abbreviations: CML, Chronic Myeloid Leukemia; ALL, Acute Lymphocytic Leukemia; B-NHL, B-cell Non-Hodgkin Lymphoma; n.d, not done. FDR<0.1 and/or remarkable numbers of p-value clusters.

Table S3. Mutation analysis of genes with significant increase in break signal after AID activation.

	4-OHT	mutated clone	unique mutations (bp)	total sequenced (bp)	mutation frequency ($x10^{-4}$)	del(ins) (bp)	del(ins)/total clones	P value *
B2M	–	1/82	1	43674	0.23	0	0/82	0.52
	+	2/87	2	46338	0.43	1	1/87	
Sμ	–	13/82	17	67404	2.52	1	1/82	$4.2x10^{-7}$
	+	38/79	59	64938	9.09	33(1)	7(1)/79	
V region	–	18/123	21	92127	2.28	6	5/123	0.001
	+	38/135	51	101115	5.04	8(10)	4(1)/135	
MYC	–	5/83	5	41251	1.21	0	0/83	$1.8x10^{-7}$
	+	21/80	33	39760	8.30	0	0/80	
SNHG3	–	3/77	3	42633	0.70	0	0/77	0.0001
	+	16/74	21	41170	5.10	2	2/74	
MALAT1	–	3/89	3	48950	0.61	0	0/89	$8.1x10^{-7}$
	+	19/90	30	49500	6.06	83(1)	3(1)/90	
BCL7A	–	3/84	3	44520	0.67	0	0/84	0.006
	+	9/82	14	43460	3.22	0	0/82	
CUX1[a]	–	0/89	0	48149	0	0	0/89	0.001
	+	8/91	11	49231	2.23	0	0/91	
CUX1[b]	–	13/85	14	58905	2.38	0	0/85	0.007
	+	22/84	31	58212	5.33	4	4/84	
CFLAR	–	0/134	0	87404	0	5	1/134	0.004
	+	5/133	8	86878	0.92	1	1/133	

Cells were treated with or without 4-OHT for 24 h. *P values were calculated by one-sided Fisher's exact test. a, region detected by promoter array (Fig. S2); b, region with highest peak of H3K4me3 (Fig. S4D).

Table S4. Primer sequences.

Linkers and linker primers (5' to 3')

Linker P1	biotin - TTCCACTACGCCTCCGCTTTCCTCTCTATGGGCAGTCGGTGAT
	ATCACCGACTGCCCATAGAGAGGAAAGCGGAGGCGTAGTGG
Linker P2	AGAGAATGAGGAACCCGGGGCAGTT
	CTGCCCCGGGTTCCTCATTCTCT
global amplification (fwd)	CCACTACGCCTCCGCTTTCCTCTCTATG
(rev)	CTGCCCCGGGTTCCTCATTCT
P1-LM (LM-PCR)	CCACTACGCCTCCGCTTTCCTCTCTATG

Gene-specific primers (5' to 3')

PCR	**Forward**	**Reverse**
Sμ	GCACAGGCTCCTAAATTCTTGGTC	CAGGCTGGCTTCCATCTTTTGTCT
B2M	CGGCTCTGCTTCCCTTAGAC	CGAAACCGCTTTGTATCACA
Mutation	**Forward**	**Reverse**
Sμ	GACATGGTAAGAGACAGGCAGCCG	GGATGGAGTTGTCATGGCCAGAAA
BL2 V region	ATCTCATGTGCAAGAAAATGAA	AGTCCCACCACGCAATCAT
MYC	CCCTCAACGTTAGCTTCACCAACA	CGCTCAGATCCTGCAGGTACAA
SNHG3	GCCCAGGAGTGACCTATACTCAAA	GGTATCCACGTTGGAATGCTCA
MALAT1	GGCAGAAGGCTTTTGGAAGA	CAACATATTGCCGACCTCACGGAT
BCL7A	ATTAGCTCTGGTCCGGCCGGTT	GGTGCAGTCGTGCAAGTTTCT
CUX1[a]	GGAGCCAGGTTGAAGGTGA	TGCCATTCGGCTGTGTTCATTA
CUX1[b]	GCTTGATCGGAAATTGATCCTC	GTCCGCGTCACCGACACAGG
CFLAR	CAGGGAAGTGTGTTAAGTGC	CATGTTGTCCTGAAGCCAGTGC
B2M	TCTCTTTCTGGCCTGGAGGCTAT	AGAGGTGCTAGGACATGCGAACTT
LM-PCR	**1st round**	**2nd round**
MYC 5'	CCAAGCCGCTGGTTCACTAA	GAGATAGCAGGGGACTGTCCAAA
MYC 3'	GGCCCGTTAAATAAGCTGCCAA	ATCCAGCCGCCCACTTTTGACA
SNHG3 5'	GCCCAGGAGTGACCTATACTCAAA	ATGAGGCCCCTCGTTGAAAA
SNHG3 3'	ACGTTGTGGAAAGGGACTGTACAT	TAGTGAGGAATTGGAGTAACCGACA
MALAT1 5'	GCTTGAGAAGATGAGGGTGTTTA	GGCAGAAGGCTTTTGGAAGA
MALAT1 3'	CAACATATTGCCGACCTCACGGATT	ACACTGGCATGCTGGTCTAGGAT
BCL7A 5'	TGAGGCCTCAAAAGTGCTCCTTGT	ACCAGGGGTCATTTGGGCAGTA
BCL7A 3'	GGTGCAGTCGTGCAAGTTTCT	GGTCTCTTGACTTCTCCGAGTTGA
CUX1 5'	GACTCTGCCAGGTGGATGTTG	GGAGCCAGGTTGAAGGTGA
CUX1 3'	AAATGCCATTCGGCTGTGTTC	TGCCATTCGGCTGTGTTCATTA
ChIP	**Forward**	**Reverse**
V region 1	TCACCTAGGCGCCCACAGGAA	CGCCACCAGCAGGAGGAAGA
V region 2	ATCTCATGTGCAAGAAAATGAAGCACCTGT	CCCTGGGATCAGAGGCAGCCTCCCA
V region 3	CTCACTGTGGGTTTTTCTGTTCACA	GAGCCACCAGAGACAGTGCAAGTGA
V region 4	CATCAGCAGTACTAATTACTACTTGAGTTG	CGACTCTCGAGGGATGGGTTGTAGT
V region 5	GTGAAGTCTTCGGAGACCTT	ACATGGTGACTCGACTCTCG
V region 6	AGTCACCATGTCCGTAGACATGTCC	AGTCCCCCCCCTTCGAGCCACTGGT
V region 7	CTGGTTCGACTCCTGGGGCCAGGGA	ACACTCTGACCCCGAGACCCTGGCA
V region 8	TGGAGGCATTTTGGAGGTCAGGAAA	CCAGCCGAAGGAGCCCCCCAGCTGC
Cμ	CTTCCTTCCCGACTCCATCAC	CGTTCTTTTCTTTGTTGCCGT

ChIP	Forward	Reverse
MALAT1 1	AGAGCAGTGTAAACACTTCTGGGTG	TGGAAAGCGAGTTCAAGTGGCCT
MALAT1 2	AGGTGATCGAATTCCGGTGATGCGA	CAAGCTCCGCCTGCCCCCTCAGCA
MALAT1 3	CATTTACTAAACGCAGACGAAAATG	TTTCTTCGCCTTCCCGTACTTCTG
MALAT1 4	TTAGAAGGTAAAGCTTGAGAAGATG	AGTCCTTTTAGTAGCTTTTTGATGTG
MALAT1 5	TTCAGTGAATCTAGGAAGACAGCAG	CCTGGACTCTTTTCCTATCTTCACCA
MALAT1 6	GATTTCCGGGTGTTGTAGGTTTCTC	AAACCCACAAACTTGCCATCTACTA
MALAT1 7	TGGCAATTAGTTGGCAGTGGCCTGT	TCCATTCTAAGACTTTAAGTTCTCTG
MALAT1 8	TGTCTCTTAGAGGGTGGGCTTTTGT	GCATCTAGGCCATCATACTGCCAGGC
qPCR	**Forward**	**Reverse**
$S\mu$	GACTGCAGGGAACTGGGGTATCA	GGATGGAGTTGTCATGGCCAGAAA
BL2 V region	GTCAGAGTCTTGGAGGCATTTTGG	AATGCTCCAGGTGAAGCGGAGAGA
MYC	GCCGCCGCCTCAGAGTGCAT	CGGAGAGAAGGCGCTGGAGT
SNHG3	AAGCTGCCCTAGTGAACTGTAGGAAG	TAGTGAGGAATTGGAGTAACCGACA
MALAT1	AAAAGGATTCCAGGAAGGAGCGAGT	ACACTGGCATGCTGGTCTAGGAT
BCL7A	CGACGTCCTAGCTCGCATTTGAA	GGTGCAGTCGTGCAAGTTTCT
CUX1	TTCTTGTCCCTCGGCTTCCT	GCACTGAAACTTCCATACCACAA
CFLAR	AAGGGACAGGTGCAGAAAGAGTAT	CTCAACTCCAGCTGACACTGCTAAT
SIPA1L3	CTTTGCCAATGGATCTGTGTCTG	GCCGACTCAGGAACTGCTTG
KCNC2	GGGTCAGCCAATGCACCATTC	GATGCAACAGCCACTCAGTAG
ZNF451	CCTTGTTCAAGATGCTCTGAGTG	GCGCCAACATTTCAACAAGCAG
TRIO	GGAAATGAGGTCTCAGGGTTTAAG	TGTGGATGCTAAGGGAACTGAG
C5ORF13	CTCTTGCGACAGCAGTTTCC	GGTTTCTCCCTGTCAACATCAC
CUL9	GATCGCTGAGGTTAGCATACTG	TTCGTGATCTCAAAGCTCCTTC
TM9SF4	CCAGCCATGCAAAGAATGTTCC	CCTCCAGCCTTCTGTGTGTTC
ANKRD11	CACCAGATCACAGCATAAGCAC	TTTGTTTGGAGACCAGCCCTTTG
MYO3B	CTTGGTCCAACCCTTGTAGTTC	GACTACTTGAATGATGGGCACAG
UPF2	TTGCCTCTGTCAGCAATGCTC	AGAGAAGACTGCCTGGAACAAG
MET	ATGAGGCTTGAAGAGAGAGGACAAC	CACTCTGCCCTCTTCCAGTTC
AUTS2	AGCTCAAGCGATTCTCCCTC	CACTCAGCACTATACCAACCAC
RAD18	GAGCCATACCACGACTGTGC	TGCAGGGCAGTCAGTTTATTAGTG
VPS13B	GTTTCCTCCTACCCTACTCTAGC	TTCCAAAGGTGTCTGGGTGTATC
CCDC41	GCAAAGACAGTCAAGAGACAG	TGTGGACCTTCAGAATCCTATC
MRPL49	AGGCAATCATGGAGGTACAAAC	TGGTCTGCCTCTCAGGATTTC
NECAB3	TCACGTACCTGCCTACTCAC	TTGCTTTAGTTGCTGGCCCATC
SETD8	CTGAAACAGCCACCAGAGTGAC	AGCAAATGGTCTTGCAGAAGG
SETBP1	TCTGTACCTGTGTGTATCTTCCTG	CCCATAGGTGACAAGCACCATC
PBLD	GGCCCTGATCTTTGTCCATTAAAC	GCCCGGCCAATAAGAGCTTC
ABCG2	TGCCACTTTATCCAGACCTAACTC	ACTTACAGTTCTCAGCAGCTCTTC
FAM65B	AATTGCTGGCCCAGTGTAGTG	GTTTCAGTCTCTTGCCCAGG
WBSCR17	GCACTAGGTGCTGTGCATAAAC	TGCAAGCACCATGACTCAGC
B2M	CGGCTCTGCTTCCCTTAGAC	CGAAACCGCTTTGTATCACA
MAP2K4	TCTGGACATTTGAGGCAGCTCT	AAGGAAATGGTCCCTAACAGGCT
SH3KBP1	CACGTCGAGACCTGCCATTTA	GGGACTGCATGTTAGATGAGGA
TUBA1B	TCCATAACCTAGGGACTATCTGA	TTGGGTGGAGTGACTGACAT
RNMT	GCTGAGTCCTGAAACTTGT	GTTTGGGATTCTACAGCAAG
EXO1	GAGGATATTTGCCTGGCCCAGAA	GGTCCTCAAGCCACAGTTTCAGA

Table S5. Summary of general features of the sequenced libraries.

library	total reads	total mapped reads	percent mapped reads	unique mapping reads	percent unique reads	redundancy ratio
OHT(-) rep. #1	83122708	51939756	62.49	40459228	77.90	1.28
OHT(+) rep. #1	84818977	52209828	61.55	41546705	79.58	1.26
OHT(-) rep. #2	82502113	53683520	65.07	42994281	80.09	1.25
OHT(+) rep. #2	83293348	51190838	61.46	41686202	81.43	1.23