HHS Public Access

Author manuscript

Nature. Author manuscript; available in PMC 2015 March 24.

Published in final edited form as:

Nature. 2014 October 16; 514(7522): 389–393. doi:10.1038/nature13580.

Noncoding RNA transcription targets AID to divergently transcribed loci in B cells

Evangelos Pefanis[1,2,*], **Jiguang Wang**[1,3,*], **Gerson Rothschild**[1,*], **Junghyun Lim**[1], **Jaime Chao**[1], **Raul Rabadan**[3], **Aris N. Economides**[2], and **Uttiya Basu**[1]

[1]Department of Microbiology and Immunology, College of Physicians and Surgeons, Columbia University, New York, New York 10032, USA

[2]Regeneron Pharmaceuticals, Tarrytown, New York 10591, USA

[3]Department of Systems Biology and Department of Biomedical Informatics, College of Physicians and Surgeons, Columbia University, New York, New York 10032, USA

Abstract

The vast majority of the mammalian genome has the potential to expressnoncoding RNA (ncRNA). The 11-subunit RNA exosome complex is the main source of cellular 3′–5′ exoribonucleolytic activity and potentially regulates the mammalian noncoding transcriptome[1]. Here we generated a mouse model in which the essential subunit *Exosc3* of the RNA exosome complex can be conditionally deleted. *Exosc3*-deficient B cells lack the ability to undergo normal levels of class switch recombination and somatic hypermutation, two mutagenic DNA processes used to generate antibody diversity via the B-cell mutator protein activation-induced cytidine deaminase (AID)[2,3]. The transcriptome of *Exosc3*-deficient B cells has revealed the presence of many novel RNA exosome substrate ncRNAs. RNA exosome substrate RNAs include xTSS-RNAs, transcription start site (TSS)-associated antisense transcripts that can exceed 500 base pairs in length and are transcribed divergently from cognate coding gene transcripts. xTSS-RNAs are most strongly expressed at genes that accumulate AID-mediated somatic mutations and/or are frequent translocation partners of DNA double-strand breaks generated at *Igh* in B cells[4,5]. Strikingly, translocations near TSSs or within gene bodies occur over regions of RNA exosome substrate ncRNA expression. These RNA exosome-regulated, antisense-transcribed regions of the B-cell genome recruit AID and accumulate single-strand DNA structures containing RNA–DNA

Correspondence and requests for materials should be addressed to R.R. (rr2579@columbia.edu) or U.B. (ub2121@columbia.edu).
*These authors contributed equally to this work.

Author Contributions E.P. and U.B. planned studies; E.P., J.W., G.R., R.R. and U.B. interpreted data. Experiments were performed as follows: E.P. and J.C., mouse model generation, CSR and SHM; E.P., RNA-seq; G.R. and J. L., ChIP, DRIP and CRISPR/Cas9; J.W., bioinformatic studies; A.N.E. advised on the mouse model construct; R.R. oversaw bioinformatics; E.P. and U.B. wrote the manuscript, which was further refined by all the other authors.

Author Information All the RNA-seq data sets have been deposited in the Sequence Read Archive under accession number SRP042355 and in the BioProject database under accession number PRJNA248775. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Supplementary Information is available in the online version of the paper.

hybrids. We propose that RNA exosome regulation of ncRNA recruits AID to single-strand DNA-forming sites of antisense and divergent transcription in the B-cell genome, thereby creating a link between ncRNA transcription and overall maintenance of B-cell genomic integrity.

AID mutates single-strand DNA (ssDNA) substrates that form during transcription across the B-cell genome. Current DNA targeting models propose that AID binds paused/stalled RNA polymerase II complexes (RNA Pol II) to access target DNA[6]. In turn, RNA Pol II associates with the pausing/stalling cofactors Spt5 and RNA exosome, both of which stimulate AID function in B cells[7–9]. Since RNA exosome is a functional component of the stalled RNA Pol II[10,11] targeting platform of AID, we evaluated RNA exosome's role in regulating AID activity genome-wide. Accordingly, we developed a mouse model containing a conditional inversion (COIN)[12] allele of *Exosc3*, allowing conditional ablation of RNA exosome function using tissue-specific or inducible Cre recombinase alleles (Fig. 1a). Cre-mediated ablation of *Exosc3* with this allele leads to concomitant green fluorescent protein (GFP)reporter induction from the *Exosc3* locus (details in Methods and Extended Data Fig. 1). B cells were generated from $Exosc3^{COIN/+}$ and $Exosc3^{COIN/COIN}$ mice on the 4-hydroxytamoxifen (4-OHT)-inducible $ROSA26^{CreERt2/+}$ background. 4-OHT treatment of these cells produced robust *Exosc3* gene inversion, loss of *Exosc3* messenger RNA and protein, and induction of GFP (Fig. 1b, c and Extended Data Fig. 1d–f).

We evaluated CSR efficiency in *ex vivo* cultured B cells upon 4-OHT-mediated ablation of *Exosc3*. Immunoglobulin (Ig)G1 class switch recombination (CSR) was decreased approximately fourfold in $Exosc3^{COIN/COIN}$ B cells compared to littermate control $Exosc3^{COIN/+}$ B cells (Fig. 1d and Extended Data Fig. 2a) despite comparable AID expression and increased nascent IgSγ1 transcription (Extended Data Figs 1f and 2b, c). To determine RNA exosome involvement in somatic hypermutation (SHM), we generated $Exosc3^{COIN/+}$ and $Exosc3^{COIN/COIN}$ mice expressing Cre recombinase at early ($Cd19^{Cre}$) and late stages ($Aicda^{Cre}$) of B-cell development ($Aicda^{Cre}$ allele details in Extended Data Fig. 2d–f). $Cd19^{Cre}$-mediated ablation of *Exosc3* leads to B-cell developmental arrest preceding the germinal centre reaction (Extended Data Fig. 2h). However, $Aicda^{Cre}$-mediated deletion permits robust germinal centre B-cell production in $Exosc3^{COIN/COIN}$ mice, with a moderate increase in cell number compared to $Exosc3^{COIN/+}$ mice (Fig. 1e). The kinetics of GFP induction and maintenance between $Exosc3^{COIN/+}Aicda^{Cre/+}$ and $Exosc3^{COIN/COIN}Aicda^{Cre/+}$ B cells *ex vivo* demonstrated little to no visible growth advantage between deleted (GFP^+) and non-deleted (GFP^-) cells (Extended Data Figs 3a, b). VPD450 dye dilution assays demonstrated comparable proliferation between $Exosc3^{COIN/+}Aicda^{Cre/+}$ and $Exosc3^{COIN/COIN}Aicda^{Cre/+}$ B cells (Extended Data Fig. 3c, d). We determined the inversion efficiency of $Exosc3^{COIN}$ in sorted $Exosc3^{COIN/COIN}$germinal centre B cells to be~70%, compared to nearly complete inversion in $Exosc3^{COIN/+}$ (Extended Data Fig. 1g). SHM downstream to the *Igh* J_H4 exon was evaluated in $Exosc3^{COIN/+}Aicda^{Cre/+}$and $Exosc3^{COIN/COIN}Aicda^{Cre/+}$germinal centre B cells. Total mutation frequency was reduced in $Exosc3^{COIN/COIN}$ mice (Extended Data Fig. 2g) and exacerbated at direct AID target dC:dG base pairs (53% of $Exosc3^{COIN/+}$, $P < 0.01$) (Fig. 1f). Importantly, since AID expression precedes *Exosc3* deletion in these assays, we expect

some SHM and CSR to occur before *Exosc3* depletion, thus underrepresenting the complete effect of RNA exosome deletion on SHM and CSR.

Various ncRNA species, particularly those associated with transcription regulation, are substrates of RNA exosome[13–20]. To uncover ncRNA substrates of RNA exosome in B cells, we performed whole transcriptome RNA sequencing on *Exosc3*-deficient cells. We reconstructed the transcriptomes of *Exosc3*$^{WT/WT}$ (wild-type) and *Exosc3*$^{COIN/COIN}$ B cells and hereafter refer to the *Exosc3*$^{COIN/COIN}$ transcriptome as the 'exotome' (Fig. 2a). Small nucleolar RNAs (snoRNAs) and small nuclear RNAs (snRNAs), known targets of RNA exosome in *Saccharomyces cerevisiae*[21], were upregulated in the exotome (Fig. 2a). The identity, read counts and coordinates of the ncRNAs analysed in Fig. 2a are provided in Supplementary Tables 1–3. Greatly upregulated in *Exosc3*-deficient B cells were RNA exosome substrate TSS RNAs (xTSS-RNAs) (Fig. 2a). Short ncRNAs arising from TSSs have been shown previously to be RNA exosome substrates in mammalian cells[13,16,22], although their genome-wide distribution and characteristics are not fully understood. xTSS-RNAs are expressed at regions upstream of mRNA-associated TSSs (Fig. 2b and Extended Data Fig. 4a) and either overlap with cognate mRNA TSSs (ungapped xTSS-RNA) or possess distinct TSSs (gapped xTSS-RNA). RNA-sequencing (RNA-seq) reproducibility was statistically strong ($\rho = 0.95$; Extended Data Fig. 4b).

In *Exosc3*$^{WT/WT}$ cells, xTSS-RNA average length was ~600 bp, whereas in *Exosc3*$^{COIN/COIN}$ cells xTSS-RNAs were slightly longer (Fig. 2c and Extended Data Fig. 4c). Average TSS distance between xTSS-RNA and cognate mRNA was ~150 bp (Fig. 2d). Many genes in *Exosc3*$^{WT/WT}$ cells display low expression of xTSS-RNA (Fig. 2e). However, *Exosc3* deletion results in a shift towards higher xTSS-RNA expression (Fig. 2e). Strand-specific RNA-seq experiments demonstrated that xTSS-RNA transcription largely occurs antisense to mRNA transcription genome-wide (Fig. 2b). While sense genic transcripts are comparable between *Exosc3*$^{WT/WT}$ and *Exosc3*$^{COIN/COIN}$, TSS antisense transcripts in *Exosc3*$^{COIN/COIN}$ are approximately fourfold higher (Fig. 2b). Gapped xTSS-RNA expression correlated poorly with cognate mRNA expression genome-wide ($\rho = 0.11$; Extended Data Fig. 4d). Furthermore, xTSS-RNA is not uniformly expressed across the B-cell genome. Actively transcribed genes devoid of xTSS-RNA expression include β-actin, *Il2rg* and *Ung* (Extended Data Fig. 4e–g). Collectively, we have identified divergently transcribed anti-sense xTSS-RNAs expressed from a subset of transcribed genes within B cells.

AID introduces mutations within *Igh* switch sequences during CSR. Upstream of the inducible CSR-specific *Igg1* (also known as *Ighg1*) germline transcript we observed strong accumulation of xTSS-RNA in *Exosc3*-deficient B cells (Fig. 3a). Expression of *Igg1* xTSS-RNA was confirmed through quantitative polymerase chain reaction with reverse transcription (qRT–PCR) and northern blotting (Extended Data Fig. 5a, b). Furthermore, we observed robust xTSS-RNA expression at the AID target genes *Myc*, *Pax5*, *Cd83*, *Pim1* and *Cd79b* (Fig.3b and Extended Data Fig. 5c–f). xTSS-RNA transcription at these genes was largely antisense (Extended Data Fig. 6a).

Recent studies using translocation capture sequencing (TC-Seq) or high-throughput genome-wide translocation sequencing (HTGTS) have identified target genes undergoing recurrent translocations to *Igh* due to AID-generated DNA breaks in B cells[4,5]. These analyses have revealed frequent translocations occurring near the TSSs of actively transcribed genes[4,5]. We queried these data sets to determine whether recurrent translocation partners of *Igh* express xTSS-RNAs. Our analysis revealed a positive statistical correlation between genes expressing xTSS-RNA and recurrent AID-dependent translocation ($P < 0.0001$; Fig. 3c and Extended Data Fig. 6b). Specifically, 40 genes were identified through this analysis (Fig. 3c and Supplementary Tables 4, 5) and, collectively, xTSS-RNA expression was fourfold higher in *Exosc3*$^{COIN/COIN}$ B cells ($P < 0.01$; Extended Data Fig. 6c). Even amongst all other xTSS-RNA-expressing genes, this group of 40 genes displayed higher xTSS-RNA expression ($P < 0.01$; Fig. 3d and Extended Data Fig. 6d). In contrast, we observed no difference in collective mRNA expression for these 40 genes between *Exosc3*$^{WT/WT}$ and *Exosc3*$^{COIN/COIN}$ B cells (Extended Data Fig. 6c). Overlapping a list of genes previously shown to undergo AID-mediated hypermutation in mouse B cells[23] with these 40 xTSS-RNA expressing translocation hotspots revealed 5 genes, consisting of *Myc*, *Pax5*, *Cd79b*, *Cd83* and *Pim1* (Fig. 3c). Mutation of these genes has been observed in diffuse large B-cell lymphoma patients[24,25]. Of the 88 translocation hotspots, 74 contain either TSS-RNA or antisense transcription (Extended Data Fig. 7a). Statistical bootstrapping analysis of 10,000 control sets of 88 genes with similar transcription levels as the translocation hotspot gene setwould randomly select only 15 xTSS-RNA expressing genes, thus validating that the observed group of 40 xTSS-RNA-containing genes at translocation hotspots is not solely determined by the level of cognate mRNA transcription ($P < 0.01$; Extended Data Fig. 6e). We propose that genes undergoing divergent transcription resulting in RNA exosome recruitment, as evidenced through the presence of xTSS-RNAs, are preferentially targeted by AID.

Many translocations also occur within gene bodies and cannot be readily explained by TSS-proximal transcription. However, RNA exosome can also regulate the expression of antisense RNA (asRNA) initiating within gene bodies. Strikingly, a considerable number of *Myc* translocation junctions precisely map to a region within intron 1 that expresses RNA exosome substrate asRNA (Fig. 3e; additional examples in Extended Data Fig. 7b and Supplementary Fig. 1). Many of the 48 translocation hotspot genes that do not express xTSS-RNAs (Fig. 3c) do express RNA exosome substrate asRNA (Supplementary Fig. 1). Moreover, breakpoints within these translocation hotspots strongly correlate with the presence of RNA exosome substrate asRNA ($\rho = 0.79$; Fig. 3f). Additionally, *Cd83* translocation breakpoints mapping to regulatory regions contain RNA exosome substrate asRNA (Extended Data Fig. 7c). When analysing all asRNAs in *Exosc3*-deficient B cells (Extended Data Fig. 8a), we note a strong positive correlation between asRNA expression and the probability of observing translocation breakpoints (r 5 0.89; Fig. 3g), whereas sense transcription correlated poorly ($\rho = 0.38$; Fig. 3h). Altogether, we provide evidence that AID target sites in the B-cell genome possess RNA exosome substrate asRNA, both TSS-proximal and within gene bodies.

An outstanding question concerns the molecular mechanisms relating AID targeting to genes expressing xTSS-RNAs or asRNAs. Divergent promoters, including those expressing xTSS-RNAs, occupy two pre-initiation complexes positioned divergently and separated by ~150 bp. Divergent transcription enhances localized DNA melting surrounding the two TSSs as polymerases initiate transcription, thus generating ssDNAstructures[26]. Antisense transcripts from such promoters undergo early termination leading to RNA exosome recruitment[27]. When stalled antisense transcripts are not efficiently removed, stabilization of transcription-coupled R-loops can prolong ssDNA structures. AID requires ssDNA substrates for recruitment and subsequent DNA deamination[3] which is further enhanced by pausing and/or stalling of RNA Pol II[6]. Using chromatin immunoprecipitation (ChIP) assays we find that *Exosc3* promotes AID occupancy at target genes *Pim1*, *Pax5*, *Myc*, *Cd79b* and *Cd83* (Fig. 4a). This observation cannot be explained simply through differences in gene expression, as H3K4me3 abundance is similar between *Exosc3*$^{WT/WT}$ and *Exosc3*$^{COIN/COIN}$ B cells, indicating comparable transcription initiation (Fig. 4b). Consistent with these observations, xTSS-RNA expression is enriched at AID- and Spt5-occupied genes in B cells (Extended Data Fig. 8b). Similarly, xTSS-RNA is also enriched at AID target genes identified by replication protein A sequencing (RPA-seq)[28] (Extended Data Fig. 8c), a marker of ssDNA. H3S10ph is a chromatin mark associated with ssDNA-containing R-loopstructures[29]. We observe that AID-targeted divergent promoters accumulate H3S10ph and Exosc3, unlike robustly transcribed non-divergent promoters (Fig. 4c and Extended Data Fig. 9a–e). H3S10ph accumulation is further enhanced at divergent promoters *Pim1*, *Pax5*, *Cd79b* and *Cd83* upon loss of *Exosc3* (Fig.4d). ssDNA accumulation was also evaluated using the RNA–DNA hybrid-specific S9.6 antibody[30]. We find that AID-targeted divergent promoters (*Pim1*, *Cd79b*, *Cd83*) accumulate RNA–DNA hybrids more strongly than transcribed non-divergent promoters (*March2*, *Cmas*, *Atp13a2*) (Fig. 4e). Finally, deletion of xTSS-RNA regions corresponding to *Pim1* or *Cd83* (Extended Data Fig. 9f–i) impairs AID recruitment to these genes (Fig. 4f,g) and reduces AID-mediated hypermutation within the first kilobase pair of *Cd83* (Extended Data Fig. 9j).

On the basis of our findings, we propose that divergent TSSs generating RNA exosome substrates have a key role in recruiting AID and generating ssDNA structures across the B-cell genome (Extended Data Fig. 10). In conjunction with ssDNA formation, increased RNA Pol II stalling at divergent promoters may further facilitate AID recruitment. Similarly, asRNAs generated within gene bodies can also create ssDNA structures, serving as AID substrates and leading to chromosomal translocation (summarized in Extended Data Fig. 10). Our study provides evidence that in addition to RNA Pol II stalling and ssDNA generation, antisense transcription is important for AID targeting throughout the B-cell genome.

METHODS

Generation of the *Exosc3*^{COIN} allele

Bacterial homologous recombination methodologies[31] were used to modify a bacterial artificial chromosome (BAC) containing the mouse *Exosc3* locus (clone bMQ386a13). Two sequential BAC modifications were performed. First, *lox2372* and *loxP* sites were inserted

between *Exosc3* exons 1 and 2. Subsequently, the COIN module (antisense to *Exosc3*), a second set of *lox2372* and *loxP* sites, and an FRT-flanked neo selection cassette were inserted between *Exosc3* exons 3 and 4. The COIN module is comprised of a 3′ splice acceptor sequence, followed by a T2A-GFP open reading frame, followed by a polyadenylation sequence. Correctly modified $Exosc3^{COINneo}$ BAC clones were identified by PCR screening across all four recombination junctions and verified by restriction digestion followed by pulse field electrophoresis. The entire 6.9 kb region between the upstream and downstream homology arms of the $Exosc3^{COINneo}$BAC clone used for embryonic stem (ES)-cell targeting was confirmed by sequencing. The BAC-based $Exosc3^{COINneo}$ targeting vector was electroporated into $ROSA26^{CreERt2/+}$, 129S6/SvEvTac3×C57BL/6Tac hybrid ES cells and correctly targeted clones were identified by a loss of allele assay[32]. $Exosc3^{COINneo/+}$ chimaeric mice were generated by ES-cell microinjection of blastocysts and crossed with Tg(ACTB:FLPe) mice to excise the neo cassette and produce germline transmission of the $Exosc3^{COIN}$ allele. The FLPe transgene was subsequently bred out for the production of all $Exosc3^{COIN}$ experimental cohorts. All mouse experiments were performed in accordance with approved Columbia University Institutional Animal Care and Use Committee protocols.

Cell culture and CSR

Splenic B cells from sex-matched littermate mice were prepared using CD43 microbead (Miltenyi Biotec) negative selection and cultured in RPMI 1640 containing 15% FBS. *Ex vivo* CSR cultures using the $ROSA26^{CreERt2}$ allele were cultured for 16 h with 100 nM4-hydroxytamoxifen (Sigma) and 20 μg ml^{-1} LPS (Sigma) followed by the addition of IL-4 (R&D Systems) at 20 ng ml^{-1}, and cultured for an additional 72 h. CSR culture conditions using the $Aicda^{Cre}$ allele were identical with the exception that 4-hydroxytamoxifen was not used. Cells were stained using fluorescent antibodies against B220 and IgG1 (BD Biosciences). Data were acquired on a FACSAria cell sorter (BD Biosciences) and analysed using FloJo software (Tree Star). Antibodies used for western blot purposes: Exosc3 (Santa Cruz Biotechnology), actin (Abcam) and goat anti-rabbit IgG HRP (Jackson ImmunoResearch). All experiments involving mice were performed with known genotypes and therefore performed unblinded. Biological replicates involve B cells isolated from separate mice.

RNA preparation and qPCR

Total RNA was isolated from cells using Trizol reagent (Life Technologies). RNA was resuspended in water and quantified using a Nanovue Plus spectrophotometer (GE Healthcare Life Sciences). RNA samples were treated with DNase I (Turbo DNA-free kit, Life Technologies), eluted in water and re-quantified. 1.5 μg of RNA were then converted to cDNA using random hexamers and the Superscript III First-Strand Synthesis System for RT–PCR (Life Technologies). *Exosc3* and *Aicda* mRNA measurements were performed using the TaqMan Gene Expression Assay (Applied Biosystems). TaqMan assay ID numbers for *Exosc3* and *Aicda* were Mm01345308_m1 and Mm00507774_m1, respectively. All other quantitative RT–PCR experiments were performed using SYBR Green Master Mix (Roche Applied Science). Primer pairs for the quantification of associated xTSS-RNAs were as follows: *Pim1*, 5′-CACATGCACGTGGAAATACCA-3′ and 5′-

CATCCATAAAAGTTATGGAGTC-3′; *Pax5*, 5′-CTGCTTTTTCAGGTCTAGCTC-3′ and 5′-CCCATTCAAAAGCTCATTAAG-3′; *Il4ra*, 5′-GGCTGTTGCTCATTTTCCCAA-3′ and 5′-TGTGGGCGAGAGAACAACTTC-3′; *Myc*, 5′-AGCGCAGCATGAATTAACTGC-3′ and 5′-GTATACGTGGCAGTGAGTTG-3′; *Igg1*, 5′-GTATCTTGTTGGTGCTATCTCA-3′ and 5′-TGGGATCTGCTACACAGGTTT-3′. Expression levels for individual transcripts were normalized against β-actin and/or cyclophilin with similar results. Fold change in transcript levels were calculated as fold change = $2^{(CtWT,GOI-CtCOIN/COIN,GOI)}/2^{(CtWT,actin-CtCOIN/COIN,actin)}$. Ct, cycle threshold. GOI, gene of interest.

Northern blotting

Total RNA (7.5 μg) isolated from splenic B cells was electrophoresed in denaturing conditions on a 1% agarose-formaldehyde gel at 5.5 V cm^{-1}. The gel was rinsed several times in deionized water followed by 20× SSC and then transferred overnight by capillary transfer onto an Amersham Hybond-XL membrane. The RNA was crosslinked to the membrane using a Stratagene UV crosslinker and then stained with methylene blue in 0.3 M sodium acetate to ascertain transfer efficiency. Subsequently the membrane was prehybridized for 5 h before being hybridized with a cDNA encoding the gene of interest (approximate size of probe: 400 bp), radiolabelled through the random primed labelling technique (High Prime, Roche Applied Science) and purified twice over Probequant G50 microcolumns (GE Healthcare) before addition to the hybridization reaction. The membrane was hybridized over-night at 42°C in hybridization solution before being washed twice in 0.1% SDS/2× SSC at room temperature followed by two washes in 0.1% SDS/2× SSC at 65°C, all washes for15 min. Membranes were subsequently exposed to film for varying amounts of time.

RNA-seq analysis

rRNA-depleted total RNA was prepared using the Ribo-Zero rRNA removal kit (Epicentre). Libraries were prepared with Illumina TruSeq and TruSeq Stranded total RNA sample prep kits, and then sequenced with 50–60 million of 2× 100 bp paired raw passing filter reads on an Illumina HiSeq 2000 V3 instrument at the Columbia Genome Center. To construct the transcriptome of *Exosc3*$^{WT/WT}$ and *Exosc3*$^{COIN/COIN}$ B cells, we first mapped all reads of total RNAs to the mouse reference genome (mm9) with TopHat (v. 1.3.2)[33]. Cufflinks (v. 1.2.1) was subsequently applied to assemble the whole transcriptome and to identify all possible transcripts[34]. To obtain short RNAs (potential exosome targets), we set the overlap radius as 1 and merged repeated samples. All resulting transcripts are further annotated under the supervision of a comprehensive collection of RNA databases including mRNA, snoRNA, long intergenic noncoding RNA (lincRNA), microRNA and other annotated non-coding RNAs (Supplementary Table 1). Specifically, we overlapped all assembled transcripts with known RNAs in the collected database, and annotated the transcripts by their adjacent known RNAs. A transcript is annotated as one isoform of a known RNA if both the TSS and transcription end site (TES) of the transcript are close (less than 200 bp) to those of the known RNA. If comparing with a known RNA, and the TSS of a transcript is shifted less than 200 bp, this transcript is annotated as the mixture of an upstream transcript (ungapped xTSS-RNA) and a known RNA. Similarly, if the TES is shifted more than 200

bp, the corresponding transcript is annotated as a read through. Transcripts that have no overlap with any known RNA located within 2,000 bp upstream of a known RNA are annotated as gapped upstream transcripts (gapped xTSS-RNA). Conjoint transcripts were annotated on the basis of containing sequences from multiple known RNAs. Biological replicates indicate that each RNA-seq data set was generated from B cells isolated from separate mice. B-cell translocation hotspots were defined based on supplementary table of Klein *et al.*[4]. AID and Spt5 binding loci are based on ChIP-seq data from Pavri *et al.*[7].

Statistical analysis

To test the significance of the difference of two vectors, a two-sided nonparametric Wilcoxon rank-sum test was applied to calculate the *P* values. To test the difference between two proportions, the following equation was used:

$$Z=\frac{p_1-p_2}{\sqrt{P(1-P)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

where p_1 and p_2 are two proportions, P is the expected value, and n_1 and n_2 are the population size. *P* values were then generated by normal distribution. We applied pipeline of DEseq for the normalization of library size, and gene differential expression analysis[35]. Multiple comparison analysis was performed in MATLAB R2011a. One-way analysis of variance (ANOVA) was followed by multiple comparison procedure with critical values from the *t* distribution, after a Bonferroni adjustment.

SHM

Peyer's patches were excised from 2.5–3.5-month-old paired littermates and gently dissociated by passage through a 70 μm cell strainer. Germinal centre B cells were stained with anti-B220 (BD Biosciences) and peanut agglutinin (Vector Laboratories). DAPI stain was added just before cell sorting to exclude dead cells. Cells were directly sorted into lysis buffer containing proteinase K (Viagen) a using FACSAria cell sorter (BD Biosciences) and incubated at 55 °C overnight. Fifteen micrograms of GlycoBlue (Life Technologies) were added to the lysates and genomic DNA was purified via ethanol precipitation. JH4 intron was amplified by PCR using LA Taq (Takara) and a primer pair (J558FR3Fw: 5′-GCCTGACATCTGAGGACTCTGC-3′ and JH4intronRv: 5′-CCTCTCCAGTTTCGGCTGAATCC-3′) that requires a VDJ rearrangement of the *Igh* locus for amplification to occur[36]. JH4 amplicons were cloned into the pCR2.1-TOPO vector (Life Technologies) and sequenced using M13 primers. Mutation analysis at *Cd83* in CH12F3 cells was performed on a 488 bp sequence between PCR primers GCCTCCAGCTCCTGTTTCTA and TGTTGCTTTCACTGCAGCTCTC.

Analysis of TC-Seq

To capture genome wide the translocation breakpoints of B cells, we downloaded the TC-Seq data sets[4] from the Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) under accession number SRA039959, and followed a similar computational workflow as described

by Oliveira *et al.*[37]. All alignments are performed by BWA[38] with default parameters, and the counting of reads is performed in bedtools[39].

Translocation breakpoint probability modelling

To predict the probability of observing breakpoints in a given genomic region, we assume the number of TC-Seq reads in this region follows a negative binomial model with parameters *r* and *p*. We applied a maximal likelihood method to estimate the parameters, as well as the confidence interval of the parameters. For a given genomic region, the probability of occurrence of translocation breakpoints is then defined as the probability of harbouring more than *x* TC-Seq reads. In this manuscript, two types of regions are separately considered. In one type of region, each base pair from 2 kb upstream of the TSS to the TES of translocation hotspot genes is binned into ~10 tiers according to their expression level of antisense transcripts. Approximately 4×10^6 genomic positions were binned according to antisense RNA expression level. Negative binomial distribution fitted in each region to estimate breakpoint probability. In the other type, genomic regions of all collected genes that do not harbour known antisense RNAs (28,947 genes) are binned into 46 tiers according to the expression level of their antisense transcripts to estimate the parameters of the negative binomial model.

ChIP

Crosslinking was performed on cultured cells using formaldehyde. Sonication was performed on ice using a Branson Sonifier 250 apparatus for 25 cycles, each cycle comprising 20 s of sonication at duty cycle 30% followed by a 2 min rest period. Lysates were pre-cleared for 1 h. Immunoprecipitation of lysates was performed overnight at 4 °C using indicated antibodies. SepharoseA/G beads were added for 90 min with continued rotation. Subsequently the beads were washed by the standard series of washes (low salt, high salt, LiCl, and TE) and ChIP products were eluted followed by RNaseA treatment overnight at 60 °C and proteinase K treatment for 2 h at 55 °C. ChIP DNA was recovered using ethanol precipitation. Primers used for ChIP quantitative PCR were as follows. *Myc*, CGGTTGATCACCCTCTATCACTC and GCTCCACACAATACGCCATGTAC; *Pim1*, CCCAGGATCTAGCCCACATAACATC and AGCGTAGCAAGTTGTGAGAAATGG; *Pax5*, CTGCTAGGATGGTTCTGCTTGG and CAACTCAATTGCAACCTCCATAGGTC; *Cd79b*, TGCTGATTGAGAAGGTTGGTGTG and GGAAGGGGTTGCTCCTGAATC; *Cd83*, AGATCTCCCTTGCTCAAACAACG and GACCTGCTACTCTCCAGATTTTGTG; *Cmas*, GGAAAACGGAAAGAGGCTGGAG and TGAGCTCAGAGGAGCCTCTAG; *Atp13a2*, CAGCCTGTCCTTTTCCGTCTATC and AGCTCGCTGAGATCTTGATGC; *March2*, GCAGCAAGTCTACAGCCAGAG and GCCTCTGAGTATCATCTGCCAATC; *Fam107b*, GACACCTTCCATTAGACAGGTGAC and AGATGAGAGCTCTGGATCCTTGG. Technical replicates of ChIP were performed from the same cell type.

DRIP

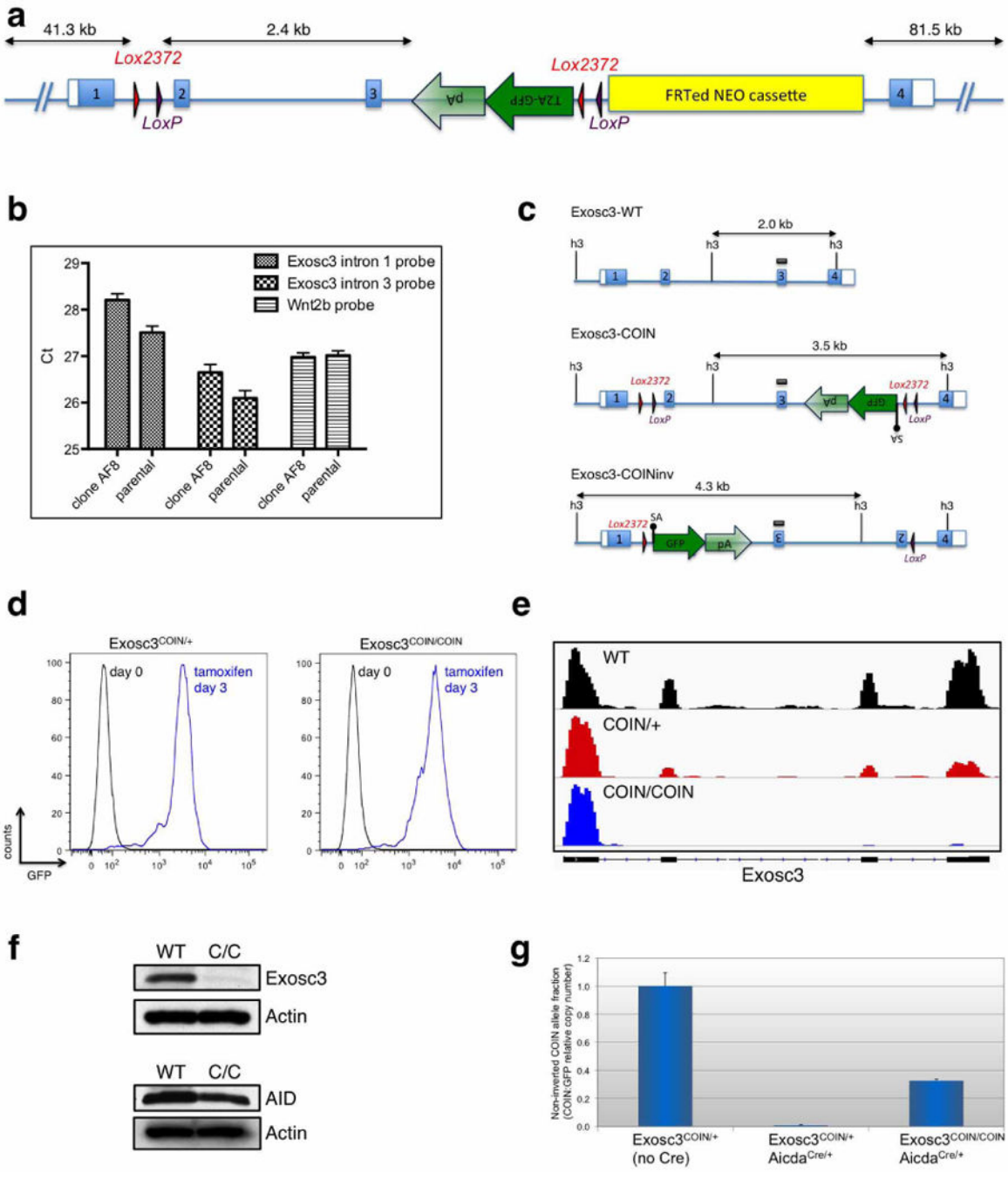
DRIP was performed on cultured B cells following a previously described protocol[30]. Briefly, cells were spun down and digested overnight in TE with proteinase K, followed by

phenol-chloroform extraction and ethanol precipitation. Genomic DNA was digested using BsoBI, NheI, NcoI and StuI either with or without RNase H (NEB). Purified DNA was then immunoprecipitated using S9.6 antibody (gift from F. Chédin). Technical replicates of DRIP were performed from the same cell type.

CRISPR/Cas9-mediated targeted deletions

Guide RNA sequences were designed using an online tool (http://tools.genome-engineering.org). Guide RNA-encoding oligonucleotides (*Cd83*, AGTGCCCAACACTACCTAAT and TTCCGAAGCCTCAGGGCGCG; *Pim1*, ATCAGACACATTCCGAGAAG and CTCTGTGTTTCCCGGAGATT) were cloned into the BbsI site of pSpCas9(BB)-2A-GFP (pX458 Addgene) as described[40]. Guide RNA/Cas9 expression vectors were electroporated into CH12F3 cells using Amaxa Nucleofector (Lonza). Cells were cloned using limiting dilution 3 days after electroporation. Individual clones were screened for homozygous deletion of xTSS-RNA-encoding regions using PCR. Screening primers for *Cd83* xTSS-RNA deletion were CCATGCTACAATGCACAGACCTAC and CAGCCTAGAAACAGGAGCTGGAG. Screening primers for *Pim1* xTSS-RNA deletion were CCAGGGATCAAACCTAGGATTTTC and CAGAAGACGCCCTATTTGCATAAGG. AID ChIP primers were as follows: *Pim1*, CTCGCTCCGCCGCCGCTGCTG and CGCAGGTGGGCCAGGGAGTTGAT; *Cd83*, GCCTCCAGCTCCTGTTTCTA and TCGGAGCAAGCCACCGTCAC.

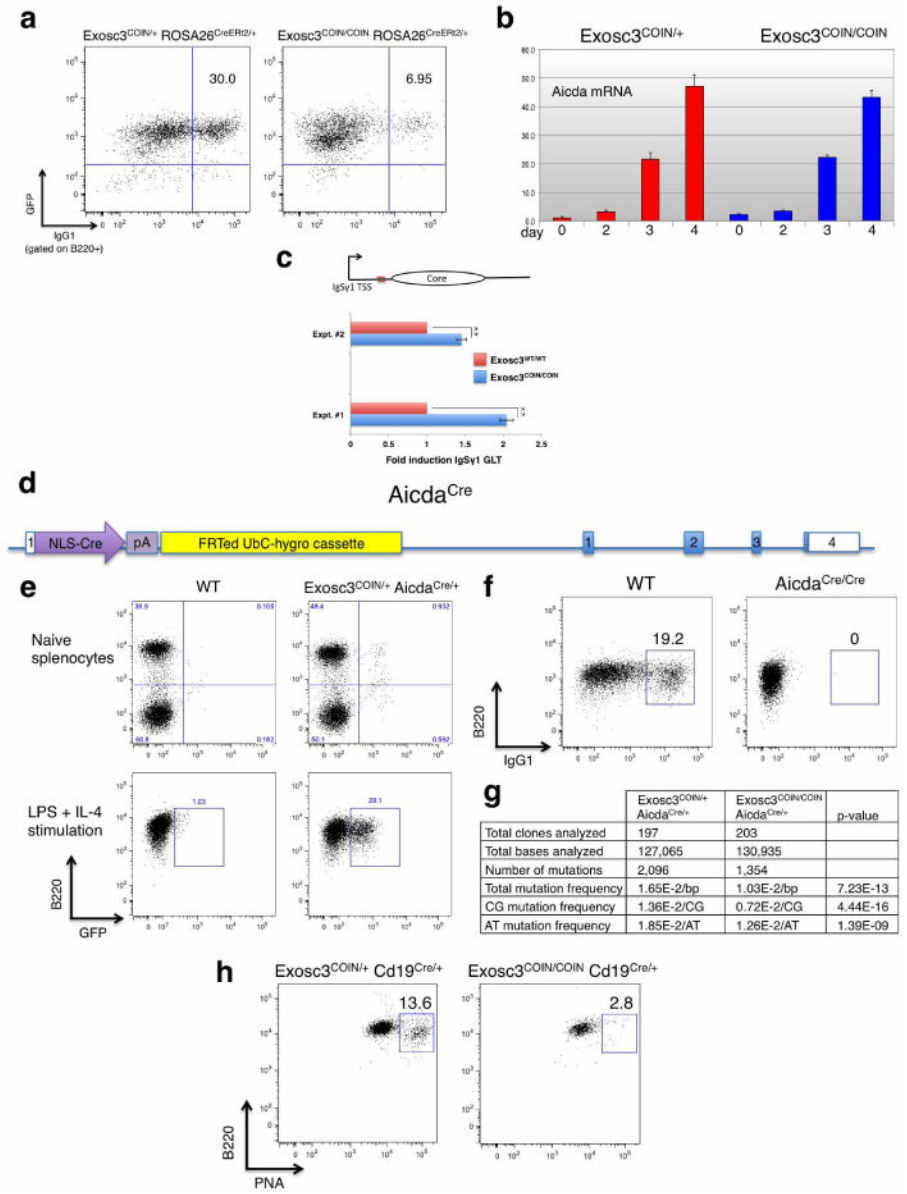
Extended Data



Extended Data Figure 1. *Exosc3* gene targeting and functional validation of the *Exosc3*^{COIN} allele

a, Schematic of the *Exosc3*^{COINneo} BAC targeting vector. Blue shaded boxes indicate *Exosc3* exons 1–4. *Lox* sites are represented by triangles. The GFP-expressing gene trapping module is represented by green arrows. Upstream, downstream and internal homology arms are 41.3, 81.5 and 2.4 kb, respectively. **b,** Confirmation of *Exosc3*$^{COINneo/+}$ targeted embryonic stem (ES)-cell clone AF8. The loss-of-allele (LOA) assay[32] was used to screen

ES-cell clones for wild-type allele copy number at defined locations within *Exosc3* introns 1 and 3 that have been modified to allow for distinction between wild-type and COINneo alleles by TaqMan-based qPCR. A probe for a non-targeted locus, *Wnt2b*, served as an internal qPCR standard for both copy number and total input DNA. Data represent mean values from six technical replicates. Error bars represent s.d. Ct, cycle threshold. **c,** HindIII restriction map of the wild-type (WT), COIN and COINinv alleles of *Exosc3*. The black shaded box indicates the location of the probe used for Southern blotting in Fig 1b. **d,** Flow cytometric analysis of GFP expression in naive or 4-OHT-treated, LPS plus IL-4-stimulated B-cell cultures. Indicated *Exosc3* genotypes are on a $ROSA26^{CreERt2/+}$ background. One pair of littermate mice was used. Three biological replicates were performed. **e,** Profile of RNA-seq mapped reads at the *Exosc3* locus from 4-OHT-treated, LPS plus IL-4-stimulated B-cell cultures. Indicated *Exosc3* genotypes are on a $ROSA26^{CreERt2/+}$ background. Four biological replicates were performed. **f,** Immunoblot analysis of Exosc3 and AID protein expression in whole cell extracts from 4-OHT-treated, LPS plus IL-4 stimulated B-cell cultures. Actin was used as a loading control. Wild type, $Exosc3^{WT/WT}$ $ROSA26^{CreERt2/+}$; C/C, $Exosc3^{COIN/COIN}$ $ROSA26^{CreERt2/+}$. One pair of littermate mice was used. Two technical replicates were performed. **g,** $Exosc3^{COIN}$:$Exosc3^{COINinv}$ ratio in germinal centre B cells determined by qPCR copy number analysis (three technical replicates, error bars represent s.d.).

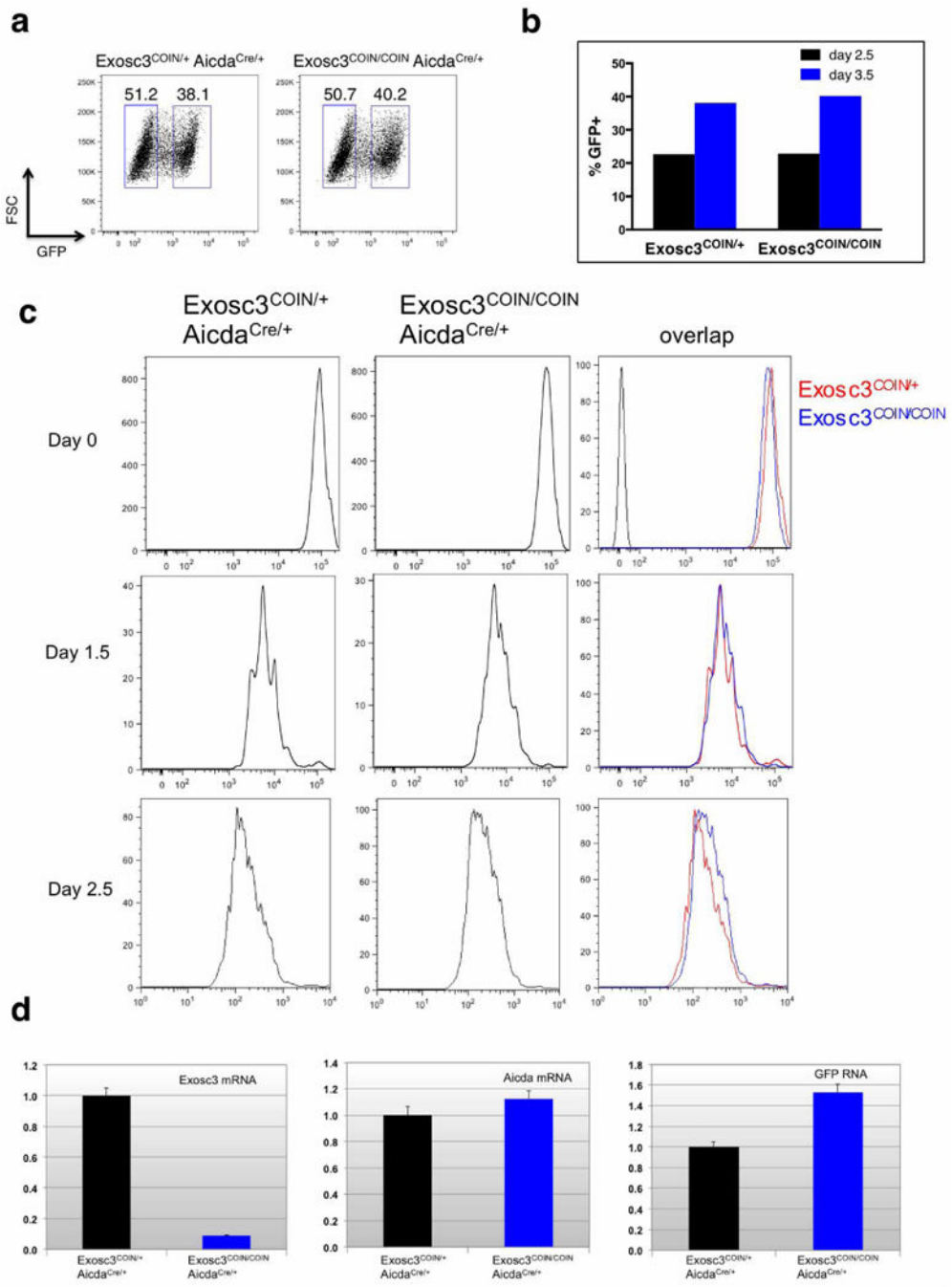


	Exosc3COIN/+ AicdaCre/+	Exosc3COIN/COIN AicdaCre/+	p-value
Total clones analyzed	197	203	
Total bases analyzed	127,065	130,935	
Number of mutations	2,096	1,354	
Total mutation frequency	1.65E-2/bp	1.03E-2/bp	7.23E-13
CG mutation frequency	1.36E-2/CG	0.72E-2/CG	4.44E-16
AT mutation frequency	1.85E-2/AT	1.26E-2/AT	1.39E-09

Extended Data Figure 2. *Exosc3*-deficient B cells are impaired in CSR and SHM

a, Representative flow cytometric analysis for surface IgG1 on purified B cells treated with 4-OHT, and stimulated with LPS plus IL-4. Numbers indicate the percentage of GFP^+ $B220^+$ B cells having isotype switched to IgG1. One pair of littermate mice was used. Three biological replicates were performed. **b,** Quantitative RT–PCR time-course analysis of *Aicda* mRNA expression in naive (day 0) or 4-OHT-treated (days 2–4), LPS plus IL-4 stimulated B-cell cultures. Indicated *Exosc3* genotypes are on a *ROSA26*$^{CreERt2/+}$ background. Expression levels are normalized to cyclophilin (*Ppia*) and plotted relative to naive *Exosc3*$^{COIN/+}$. Six littermate pairs of each genotype were used. Data represent mean values from three technical replicates. Error bars represent s.d. $^{**}P < 0.01$ (*t*-test). **c,** Quantitative RT–PCR analysis of *Ighg1* switch region intron expression. Primers were designed to amplify a region of the *Ighg1* GLT intron upstream of the *Ighg1* switch region

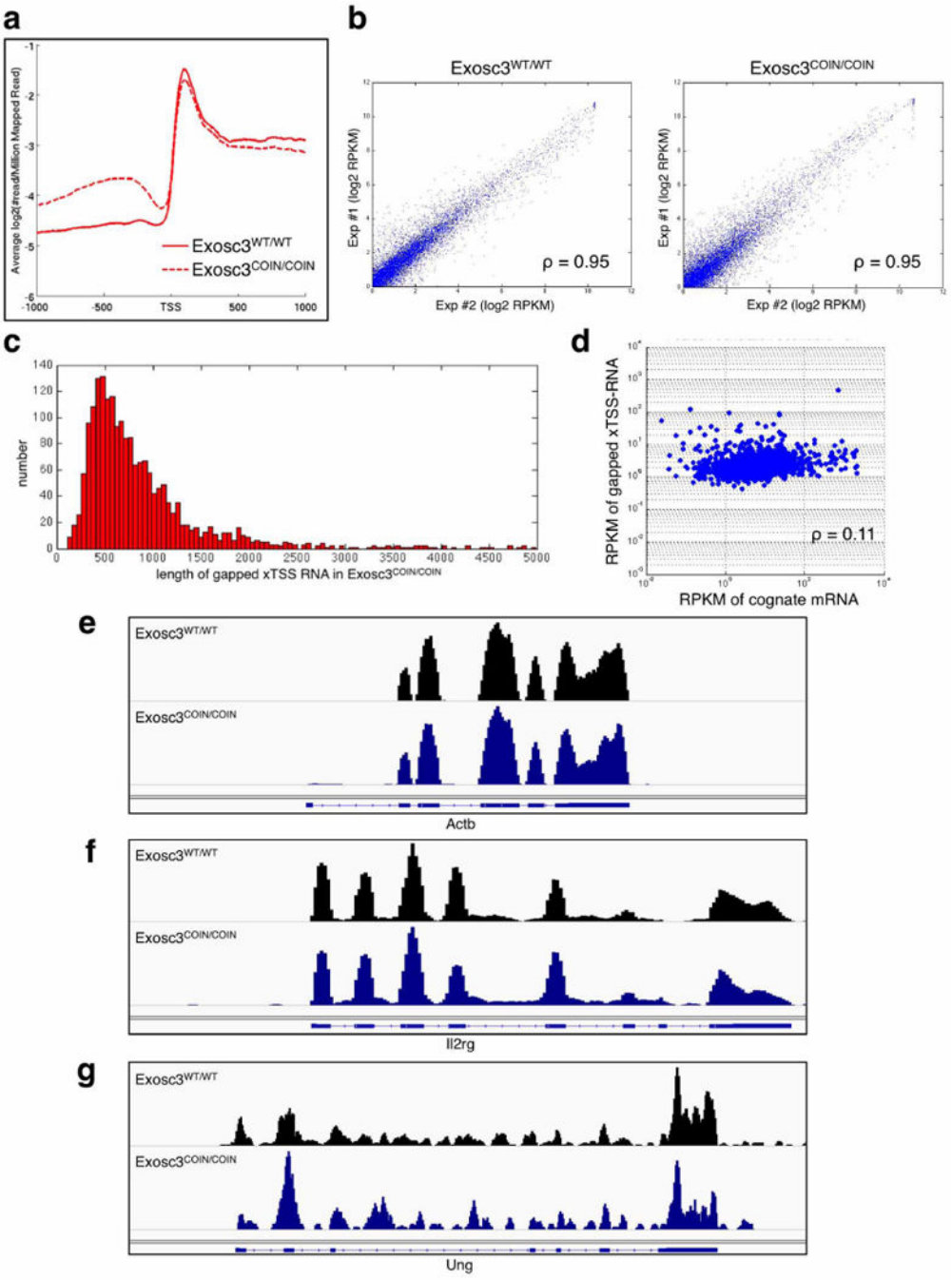
core repeat, but downstream of the *Ighg1* non-coding I exon. Two independent pairs of littermate mice of each genotype were used to obtain total RNA from B-cell cultures treated with 4-OHT and stimulated with LPS plus IL-4. Indicated genotypes are on a *ROSA26*$^{CreERt2/+}$ background. Data represent mean values from three technical replicates. Error bars represent s.d. Two biological replicates were performed. **d,** Schematic of the targeted *Aicda*^{Cre} allele. An open reading frame comprising a nuclear localization signal fused to Cre recombinase was used to disrupt the ATG start codon in exon 1 of *Aicda*. Exons are represented as numbered boxes. e, Specific induction of *Aicda*^{Cre} activity upon LPS plus IL-4 stimulation of B cells. Flow cytometric analysis of *Aicda*^{Cre} activity (as determined by GFP expression) in B220^{+} and B220^{-} naive splenocyte populations (top panel). *Aicda*^{Cre} induction in LPS plus IL-4 stimulated B-cell cultures (bottom panel). One pair of littermate mice was used. **f,** *Aicda*^{Cre} is a functional null allele. CSR to IgG1 isotype is abrogated in *Aicda*$^{Cre/Cre}$ homozygous B cells stimulated with LPS plus IL-4. Numbers above gate indicate the percentage of GFP^{+} B cells having isotype switched to IgG1. One pair of littermate mice was used. **g,** SHM analysis of Peyer's patch derived GFP^{+} germinal centre B cells. Mutation frequencies were determined by sequencing a 645 bp intronic region downstream of the JH4 gene segment of the immunoglobulin heavy chain (IgH) locus. Two littermate pairs of each genotype were used. Two biological replicates were performed. Mutation frequencies represent mean values. *P* values were determined by proportion test. **h,** Flow cytometric analysis of Peyer's patch derived germinal centre B cells from *Exosc3*$^{COIN/+}$ and *Exosc3*$^{COIN/COIN}$ mice on a *Cd19*$^{Cre/1}$ background were identified as B220^{+} PNA^{hi} populations. The percentage of germinal centre B cells amongst all B220^{+} cells is indicated. One pair of littermate mice was used. Three biological replicates were performed.



Extended Data Figure 3. Proliferation analysis of *Exosc3*-deficient B cells

a, FACS analysis indicating the percentage of GFP-negative (left gate) and GFP-positive (right gate) B cells 3.5 days after LPS stimulation. One pair of littermate mice was used. Two biological replicates were performed. **b,** Kinetic analysis of GFP-positive B-cell accumulation at indicated time points post-LPS stimulation. Indicated *Exosc3* genotypes are on a *Aicda*$^{Cre/+}$ background. One pair of littermate mice was used. Two biological replicates were performed. **c,** Proliferation analysis determined by VPD450 dilution at 1.5 and 2.5 days post-LPS stimulation. One pair of littermate mice was used. **d,** Quantitative RT–PCR

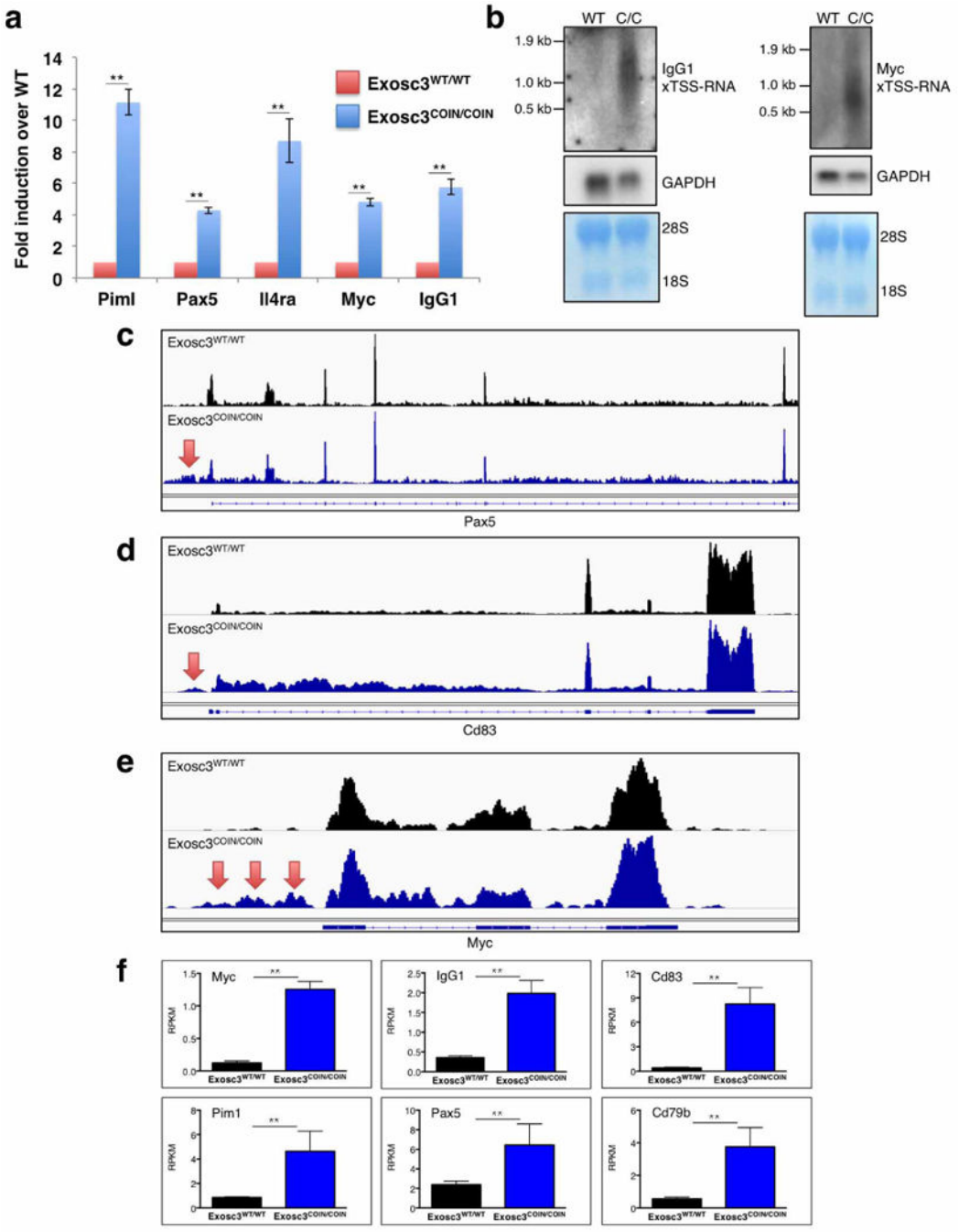
analysis of *Exosc3*, *Aicda* and *GFP* mRNA expression in GFP^{+} cells at 3.5 days post-LPS stimulation. Expression levels are normalized to β-actin and plotted relative to *Exosc3*$^{COIN/+}$. One pair of littermate mice was used. Data represent mean values from three technical replicates. Error bars represent s.d.



Extended Data Figure 4. Transcriptome analysis of *Exosc3*-deficient B cells

a, Genome-wide expression level analysis upstream and downstream of TSS region for expressed protein coding genes. Coding genes with FPKM >1 were determined to be

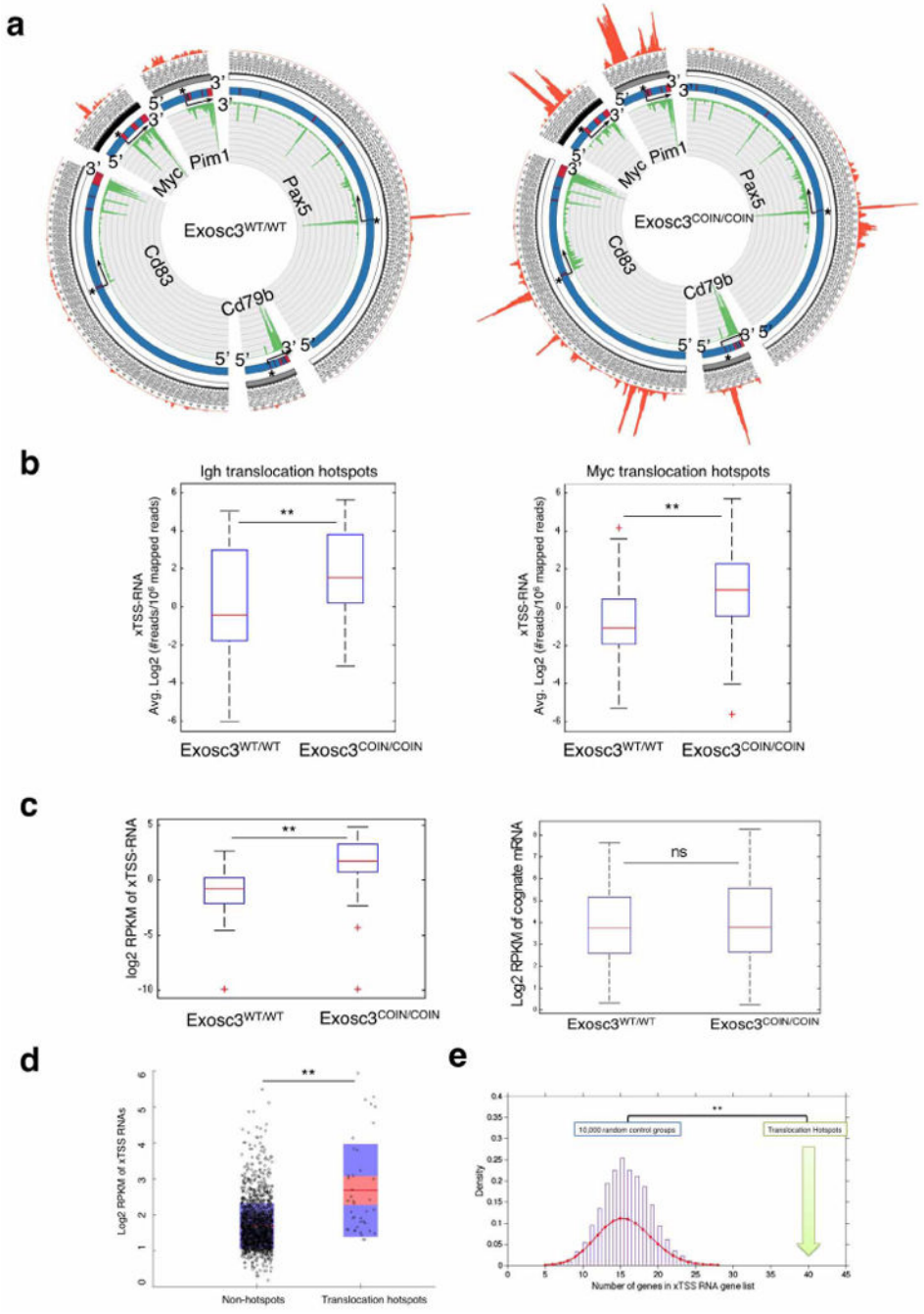
expressed. Analysis was restricted to coding genes that do not have any known genes within a 4 kb upstream boundary. Indicated genotypes are on a *ROSA26*$^{CreERt2/+}$ background. One sex-matched littermate pair was used. Two biological replicates were performed. **b,** Replicate analysis of genome-wide studies. Plots indicate the expression levels of individual genes in *Exosc3*$^{WT/WT}$ and *Exosc3*$^{COIN/COIN}$ B cells treated with 4-OHT and stimulated with LPS plus IL-4 from two separate littermate pairs. B cells were purified, cultured and FACS sorted, and RNA was purified and sequenced by RNA-seq all independently between the two experiments. Indicated genotypes are on a *ROSA26*$^{CreERt2/+}$ background. Pearson correlation is indicated. **c,** The distribution of observed lengths for all gapped xTSS-RNAs in *Exosc3*-deficient B cells. Data were compiled from two biological replicates. **d,** Scatter plot indicating weak correlation between expression of downstream coding transcript and upstream gapped xTSS-RNA at divergently transcribed loci in *Exosc3*-deficient B cells. Pearson correlation is indicated. **e–g,** Profile of RNA-seq mapped reads at the β-actin locus (**e**) (*Actb*; 7.6 kb window), *Il2rg* locus (**f**) (5.4 kb window) and *Ung* locus (**g**) (12 kb window). Indicated genotypes are on a *ROSA26*$^{CreERt2/+}$ background and B-cell cultures were treated with 4-OHT and stimulated with LPS plus IL-4. Four biological replicates were performed.



Extended Data Figure 5. xTSS-RNA expression at AID target genes

a, Quantification of xTSS-RNA expression levels of AID target genes via quantitative RT–PCR from two independent experiments. Indicated genotypes are on a *ROSA26*$^{CreERt2/+}$ background. **b,** Northern blot analysis of xTSS-RNA expression at *Myc* and *Igsγ1* loci. WT, *Exosc3*$^{WT/WT}$ *ROSA26*$^{CreERt2/+}$; C/C, *Exosc3*$^{COIN/COIN}$ *ROSA26*$^{CreERt2/+}$. **c–e,** Profiles of RNA-seq mapped reads at the *Pax5* (**c**) (72 kb window displaying exons 1–6), *Cd83* (**d**) (22 kb window), and *Myc* (**e**) (9 kb window) loci. Red arrows highlight the presence of xTSS-RNA. Four biological replicates were performed. **f,** Quantification of xTSS-RNA expression

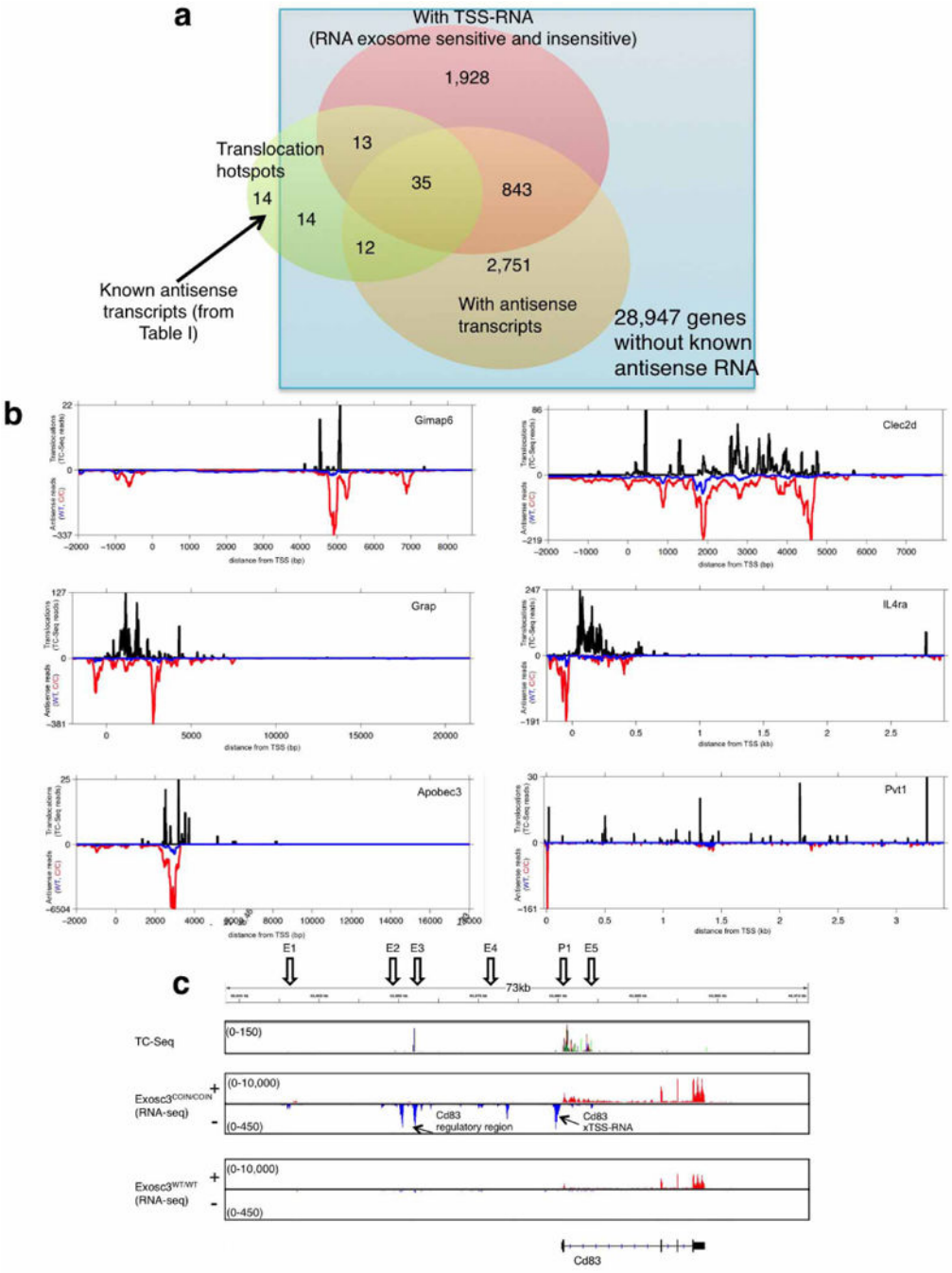
levels for AID target genes *Myc*, *Igg1*, *Cd83*, *Pim1*, *Pax5* and *Cd79b* was obtained from RNA-seq RPKM values from two independent experiments. Indicated genotypes are on a *ROSA26*$^{CreERt2/+}$ background. $^{**}P < 0.01$ (*t*-test).



Extended Data Figure 6. Translocation hotspots are enriched for xTSS-RNA expression

a, Strand-specific RNA-seq mapped reads at AID target genes *Myc*, *Cd83*, *Pim1*, *Pax5* and *Cd79b*. Green and red peaks indicate sense and antisense reads, respectively. Red bars represent RefSeq annotation of gene exons. Asterisks indicate the location of TSSs. Arrows

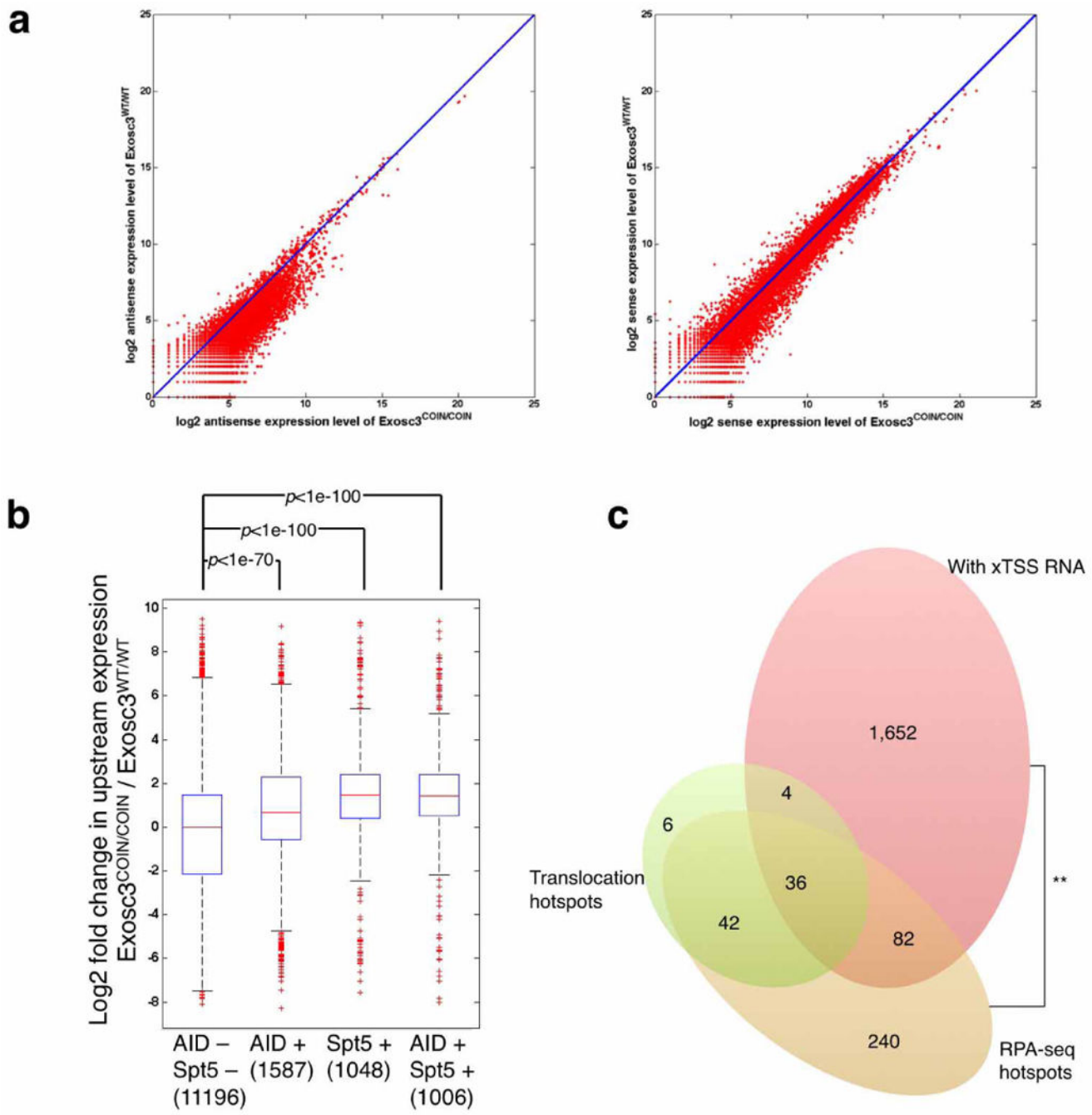
indicate the orientation of coding strand transcript. Data were compiled from two biological replicates. **b,** Boxplot analysis of the level of expression of xTSS-RNAs at various genes reported to undergo recurrent AID-dependent translocations at DNA double-strand breaks generated within the *Igh* (left panel) or *Myc* (right panel) loci. Boxplots represent median values compiled from two biological replicates. Whiskers represent 99% of data values. $^{**}P < 0.01$ (Wilcoxon rank-sum test). **c,** The list of 40 genes that show an overlap of translocation hotspots and xTSS-RNA expression (from Fig. 3c) was evaluated directly for xTSS-RNA levels (left panel) and mRNA levels (right panel). Statistical analysis was as described in **b**. $^{**}P < 0.01$; NS, not significant (Wilcoxon rank-sum test). **d,** xTSS-RNA expression levels in *Exosc3*-deficient B cells at non-recurrent and recurrent AID-dependent translocation sites in the B-cell genome. Data were compiled from two biological replicates. $^{**}P < 0.01$ (Wilcoxon rank-sum test). **e,** Statistical analysis of the probability of identification of 40 random xTSS-RNA-expressing genes solely based on expression level. Ten-thousand control group genes were randomly selected that were expressed at similar levels as translocation hotspots genes. Specifically, to generate one random control group, we exhausted all translocation hotspots to find genes with similar expression levels (difference of RPKM < 0.5), and randomly picked up one for each hotspot. Ten-thousand gene lists were obtained that contain 88 genes and share the same expression profile with the translocation hotspots list. We then simulate the distribution of genes containing xTSS-RNA by overlapping the random control groups and actual xTSS-RNA gene list. The binomial fitting (red curve) shows that the number of overlapping genes of real translocation hotspots is significantly higher than random controls. $^{**}P < 0.01$ (binomial distribution).



Extended Data Figure 7. RNA exosome substrate antisense transcripts are expressed within gene bodies and regulatory regions containing AID-induced translocations

a, Association of genes with TSS-RNA expression, antisense transcripts and AID-induced translocations. The xTSS-RNA and antisense transcripts groups were compiled from four and two biological replicates, respectively. **b,** Examples of genes with asRNA transcription (*Exosc3*$^{WT/WT}$ in blue and *Exosc3*$^{COIN/COIN}$ in red) at regions that have been shown to have translocations from the *Igh* locus (translocations indicated in black). Data were compiled from two biological replicates. **c,** Translocations present in the upstream regulatory regions

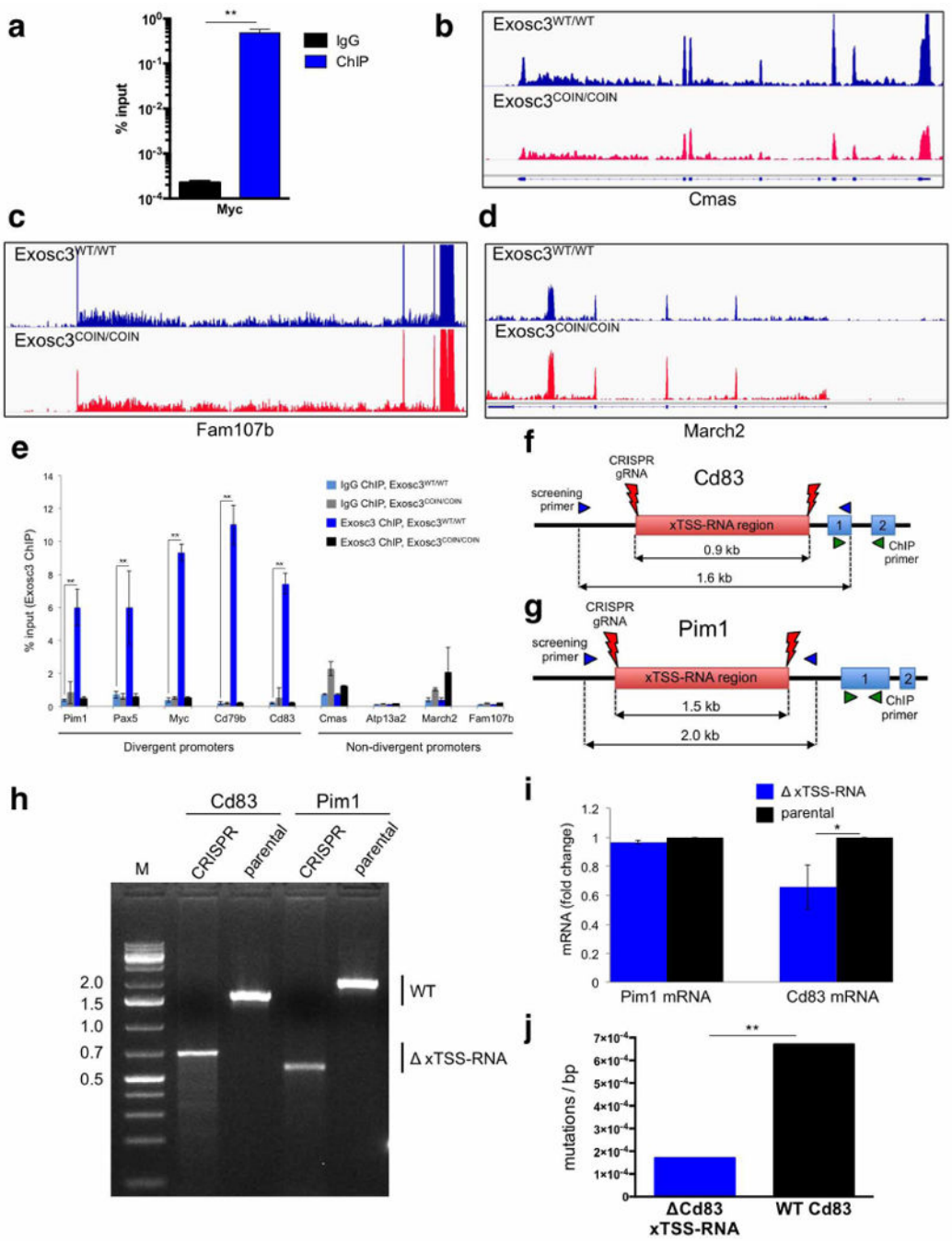
of AID target gene *Cd83* (top panel) occur over regions of RNA exosome-sensitive antisense transcription. Data were compiled from two biological replicates.



Extended Data Figure 8. Genome-wide analysis of xTSS-RNA expression at genes with AID, Spt5 and RPA occupancy

a, Scatter plot of antisense (left) and sense RNA transcription (right) in *Exosc3*$^{WT/WT}$ and *Exosc3*$^{COIN/COIN}$ transcriptomes. Data were compiled from two biological replicates. **b,** Genome-wide analysis of xTSS-RNA expression at genes that are expressed and possess or lack AID and/or Spt5 occupancy. Values beneath each group represent the number of genes

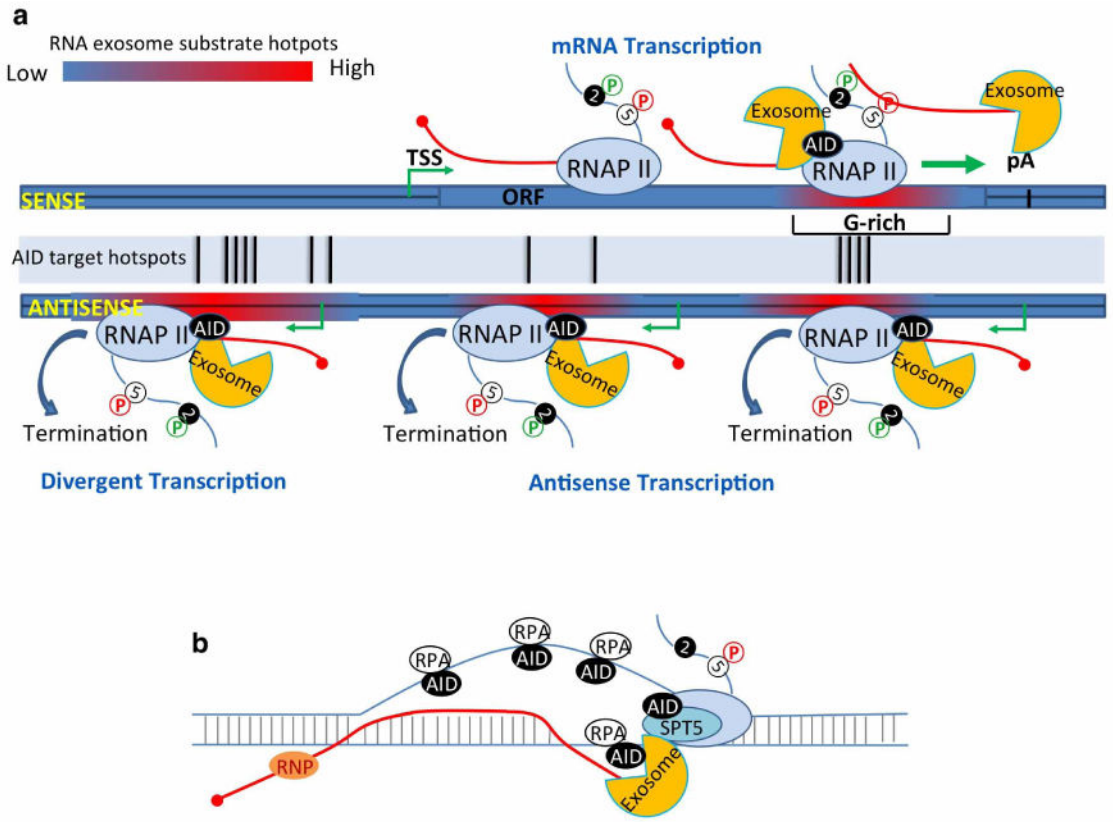
with indicated occupancy. *P* values were determined by Wilcoxon rank-sum test. **c,** Overlap of genes with xTSS-RNA transcription (pink), recurrent AID-dependent chromosomal translocations (green) and RPA occupancy in the mouse B-cell genome (brown). The xTSS-RNA group was compiled from four biological replicates. ***P* < 0.01 (Fisher's exact test).



Extended Data Figure 9. RNA exosome and AID recruitment to divergently transcribed promoter regions

a, ChIP was performed using anti-H3S10ph or control IgG. Quantitative PCR and data analysis were performed as described in Fig. 4c. ***P* < 0.01 (*t*-test). **b–d,** Representative

plots of highly expressed non-divergent genes used as controls for ChIP experiments in Fig. 4. These genes are *Cmas* (**b**), *Fam107b* (**c**) and *March2* (**d**). Four biological replicates were performed. **e,** Exosc3 occupancy at divergent and non-divergent promoters. ChIP was performed using anti-Exosc3 (Genway) or control rabbit IgG. Quantitative PCR was performed using primers specific for sequences upstream of the indicated gene TSS. Data are represented as Exosc3 enrichment relative to input. Data represent mean values from three technical replicates. Error bars represent s.d. $^{**}P < 0.01$ (*t*-test). **f, g,** CRISPR/Cas9-mediated deletion strategy of *Cd83* (**f**) and *Pim1* (**g**) xTSS-RNA-expressing regions in CH12F3 B cells. Locations of CRISPR/Cas9 guide RNAs (red markings), genotyping primers (blue triangles), ChIP primers (green triangles), and numbered exons (blue boxes) are indicated. **h,** Genotyping of *Cd83* and *Pim1* xTSS-RNA region-deleted CH12F3 clones. i, *Cd83* and *Pim1* mRNA expression in xTSS-RNA region-deleted CH12F3 cells. Data represent mean values from three technical replicates. $^{*}P < 0.05$ (*t*-test). **j,** Deletion of *Cd83* xTSS-RNA-expressing region impairs SHM. Parental CH12F3 or *Cd83* xTSS-RNA region-deleted CH12F3 cells were transduced with lentiviral AID and mutation frequency was determined within a 488 bp region beginning approximately 150 bp downstream of the *Cd83* TSS. All mutations were derived from unique clonal amplified sequences. Impairment of *Cd83* SHM in *Cd83* xTSS-RNA region-deleted cells is disproportionally greater than mRNA expression change observed in i. Number of sequenced clones for parental and *Cd83* xTSS-RNA region-deleted CH12F3 was 69 and 102, respectively. Background mutation frequency was determined using uninfected control CH12F3 cells and subtracted from the mutation frequencies indicated. $^{**}P < 0.01$ (proportion test).



Extended Data Figure 10. A model of RNA exosome recruitment to divergently transcribed promoters or at DNA sequences that promote RNA Pol II stalling

a, Divergent transcription of mRNA in the sense direction recruits RNA exosome and AID following stalling due to various transcription impediments (G-richness in IgH switch sequences is one example). Transcription stalling leading to RNA exosome recruitment occurs more often on the antisense strand due to formation of short as RNAs[27]. Similarly, in the body of transcribed genes, stalled RNA Pol II generates asRNA transcripts, leading to RNA exosome and AID recruitment. **b,** Stalled transcripts either close to the TSS or within the body of genes generate DNA–RNA hybrids. These DNA–RNA hybrids contain RPA-coated ssDNA structures that are targets of AID.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank L. Symington, S. Goff, S. Ghosh, S. Silverstein, C. Lima, F. Chédin, L. Macdonald, C.-S. Lin and O. Couronne for critical input and reagents. This work was supported by grants from the National Institutes of Health (NIH; 1DP2OD008651-01) and the National Institute of Allergy and Infectious Diseases (1R01AI099195-01A1) (to U.B.); NIH (1R01CA185486-01; 1R01CA179044-01A1; 1U54CA121852-05) (to R.R.).

References

1. Schneider C, Tollervey D. Threading the barrel of the RNA exosome. Trends Biochem Sci. 2013; 38:485–493. [PubMed: 23910895]
2. Alt FW, Zhang Y, Meng FL, Guo C, Schwer B. Mechanisms of programmed DNA lesions and genomic instability in the immune system. Cell. 2013; 152:417–429. [PubMed: 23374339]
3. Keim C, Kazadi D, Rothschild G, Basu U. Regulation of AID, the B-cell genome mutator. Genes Dev. 2013; 27:1–17. [PubMed: 23307864]
4. Klein IA, et al. Translocation-capture sequencing reveals the extent and nature of chromosomal rearrangements in B lymphocytes. Cell. 2011; 147:95–106. [PubMed: 21962510]
5. Chiarle R, et al. Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells. Cell. 2011; 147:107–119. [PubMed: 21962511]
6. Storb U. Why does somatic hypermutation by AID require transcription of its target genes? Adv Immunol. 2014; 122:253–277. [PubMed: 24507160]
7. Pavri R, et al. Activation-induced cytidine deaminase targets DNA at sites of RNA polymerase II stalling by interaction with Spt5. Cell. 2010; 143:122–133. [PubMed: 20887897]
8. Basu U, et al. The RNA exosome targets the AID cytidine deaminase to both strands of transcribed duplex DNA substrates. Cell. 2011; 144:353–363. [PubMed: 21255825]
9. Sun J, et al. E3-ubiquitin ligase Nedd4 determines the fate of AID-associated RNA polymerase II in B cells. Genes Dev. 2013; 27:1821–1833. [PubMed: 23964096]
10. Andrulis ED, et al. The RNA processing exosome is linked to elongating RNA polymerase II in Drosophila. Nature. 2002; 420:837–841. [PubMed: 12490954]
11. Richard P, Manley JL. Transcription termination by nuclear RNA polymerases. Genes Dev. 2009; 23:1247–1269. [PubMed: 19487567]
12. Economides AN, et al. Conditionals by inversion provide a universal method for the generation of conditional alleles. Proc Natl Acad Sci USA. 2013; 110:E3179–E3188. [PubMed: 23918385]
13. Flynn RA, Almada AE, Zamudio JR, Sharp PA. Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome. Proc Natl Acad Sci USA. 2011; 108:10460–10465. [PubMed: 21670248]
14. Almada AE, Wu X, Kriz AJ, Burge CB, Sharp PA. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. Nature. 2013; 499:360–363. [PubMed: 23792564]
15. Wu X, Sharp PA. Divergent transcription: a driving force for new gene origination? Cell. 2013; 155:990–996. [PubMed: 24267885]
16. Preker P, et al. RNA exosome depletion reveals transcription upstream of active human promoters. Science. 2008; 322:1851–1854. [PubMed: 19056938]
17. Andersen PR, et al. The human cap-binding complex is functionally connected to the nuclear RNA exosome. Nature Struct Mol Biol. 2013; 20:1367–1376. [PubMed: 24270879]
18. Andersen PK, Jensen TH, Lykke-Andersen S. Making ends meet: coordination between RNA 3′-end processing and transcription initiation. Wiley Interdiscip Rev RNA. 2013; 4:233–246. [PubMed: 23450686]
19. Flynn RA, Chang HY. Active chromatin and noncoding RNAs: an intimate relationship. Curr Opin Genet Dev. 2012; 22:172–178. [PubMed: 22154525]
20. Seila AC, et al. Divergent transcription from active promoters. Science. 2008; 322:1849–1851. [PubMed: 19056940]
21. Allmang C, et al. Functions of the exosome in rRNA, snoRNA and snRNA synthesis. EMBO J. 1999; 18:5399–5410. [PubMed: 10508172]
22. Ntini E, et al. Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. Nature Struct Mol Biol. 2013; 20:923–928. [PubMed: 23851456]
23. Liu M, et al. Two levels of protection for the B cell genome during somatic hypermutation. Nature. 2008; 451:841–845. [PubMed: 18273020]
24. Pasqualucci L, et al. Analysis of the coding genome of diffuse large B-cell lymphoma. Nature Genet. 2011; 43:830–837. [PubMed: 21804550]

25. Lohr JG, et al. Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. Proc Natl Acad Sci USA. 2012; 109:3879–3884. [PubMed: 22343534]
26. Rhee HS, Pugh BF. Genome-wide structure and organization of eukaryotic pre-initiation complexes. Nature. 2012; 483:295–301. [PubMed: 22258509]
27. Schulz D, et al. Transcriptome surveillance by selective termination of noncoding RNA synthesis. Cell. 2013; 155:1075–1087. [PubMed: 24210918]
28. Hakim O, et al. DNA damage defines sites of recurrent chromosomal translocations in B lymphocytes. Nature. 2012; 484:69–74. [PubMed: 22314321]
29. Castellano-Pozo M, et al. R-loops are linked to histone H3 S10 phosphorylation and chromatin condensation. Mol Cell. 2013; 52:583–590. [PubMed: 24211264]
30. Ginno PA, Lott PL, Christensen HC, Korf I, Chedin F. R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. Mol Cell. 2012; 45:814–825. [PubMed: 22387027]
31. Zhang Y, Buchholz F, Muyrers JP, Stewart AF. A new logic for DNA engineering using recombination in Escherichia coli. Nature Genet. 1998; 20:123–128. [PubMed: 9771703]
32. Frendewey D, et al. The loss-of-allele assay for ES cell screening and mouse genotyping. Methods Enzymol. 2010; 476:295–307. [PubMed: 20691873]
33. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009; 25:1105–1111. [PubMed: 19289445]
34. Trapnell C, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnol. 2010; 28:511–515. [PubMed: 20436464]
35. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010; 11:R106. [PubMed: 20979621]
36. Jolly CJ, Klix N, Neuberger MS. Rapid methods for the analysis of immunoglobulin gene hypermutation: application to transgenic and gene targeted mice. Nucleic Acids Res. 1997; 25:1913–1919. [PubMed: 9115357]
37. Oliveira TY, et al. Translocation capture sequencing: a method for high throughput mapping of chromosomal rearrangements. J Immunol Methods. 2012; 375:176–181. [PubMed: 22033343]
38. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009; 25:1754–1760. [PubMed: 19451168]
39. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26:841–842. [PubMed: 20110278]
40. Ran FA, et al. Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. Cell. 2013; 154:1380–1389. [PubMed: 23992846]

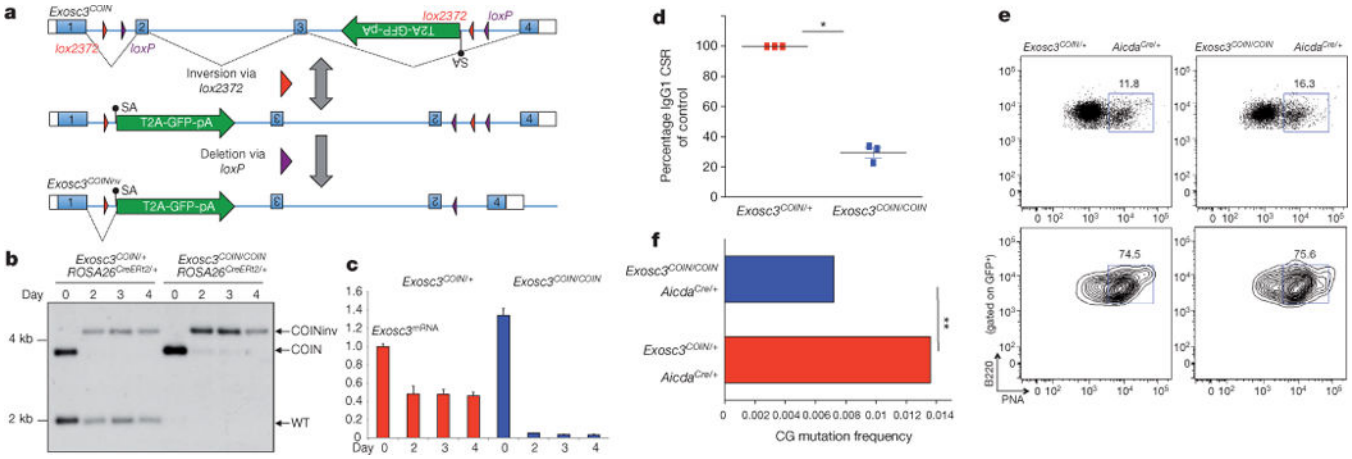


Figure 1. ***Exosc3*-deficient B cells are defective in immunoglobulin diversification**

a, *Exosc3*^{COIN} allele and conversion to *Exosc3*^{COINinv}. Cre-mediated inversion of *lox2372* pair (red triangles) and subsequent deletion via *loxP* pair (violet triangles). GFP-expressing terminal exon is represented by green arrow. SA, splice acceptor. **b,** Southern blot of HindIII-digested genomic DNA from 4-OHT-treated (days 2–4), lipopolysaccharide (LPS) plus interleukin (IL)-4 stimulated B cells. Probe specific for *Exosc3* exon 3. WT, wild type. **c,** qRT–PCR time course of *Exosc3* mRNA expression in 4-OHT-treated (days 2–4), LPS plus IL-4 stimulated B cells. Indicated *Exosc3* genotypes on a *ROSA26*$^{CreERt2/+}$ background. Expression levels normalized to cyclophilin (*Ppia*) and plotted relative to untreated *Exosc3*$^{COIN/+}$. Three technical replicates, error bars represent standard deviation (s.d.). **d,** IgG1 CSR efficiency in 4-OHT-treated *Exosc3*$^{COIN/+}$ and *Exosc3*$^{COIN/COIN}$ B cells after 72 h of LPS plus IL-4 stimulation. Indicated *Exosc3* genotypes on a *ROSA26*$^{CreERt2/+}$ background. Mean values from three biological replicates are indicated. Error bars represent standard error of the mean (s.e.m.). **e,** Flow cytometric analysis of Peyer's patch germinal centre B cells. Percentage of B220^{+} PNA^{hi} germinal centre B cells amongst all B220^{+} cells is indicated. Experiment was replicated three times. **f,** SHM analysis of Peyer's patch derived GFP^{+} germinal centre B cells at AID substrate CG base pairs at the JH4 intron. Mean values determined from 197 (*Exosc3*$^{COIN/+}$) and 203 (*Exosc3*$^{COIN/COIN}$) sequence clones are indicated. *P < 0.01 (t-test), **P < 0.01 (proportion test).

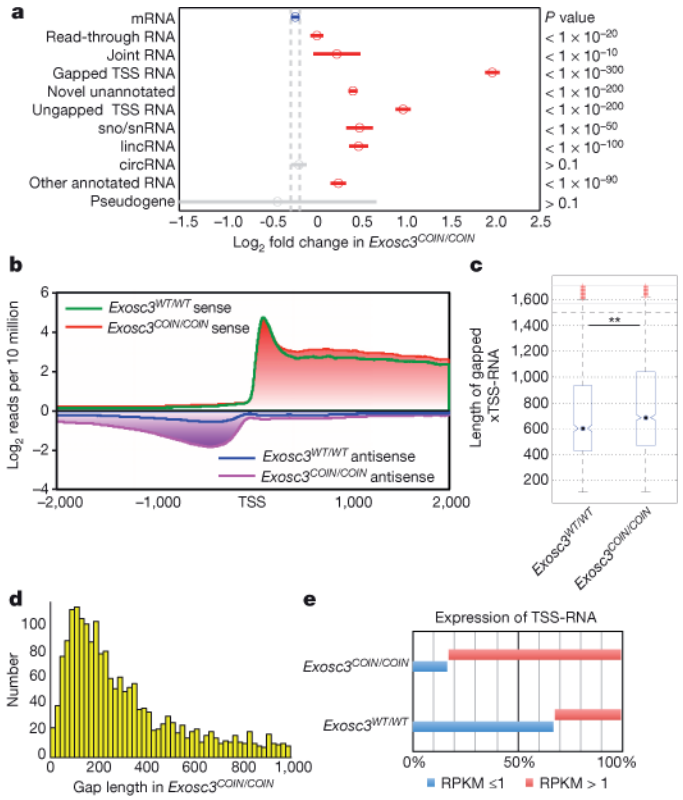


Figure 2. RNA exosome depletion reveals xTSS-RNAs

a, Differential expression analysis for various transcript classes. Horizontal bars indicate 95% confidence intervals with Bonferroni adjustment of $\log_2$ fold change between *Exosc3*$^{COIN/COIN}$ and *Exosc3*$^{WT/WT}$. Circles indicate mean values. Red bars indicate transcript types expressed significantly higher compared with mRNA (two biological replicates). *P* values were determined by Wilcoxon rank-sum test. **b,** Strand-specific distribution of RNA-seq reads upstream and downstream of expressed coding gene TSSs (two biological replicates). **c,** Boxplot indicating gapped xTSS-RNA length distribution. Median values from two biological replicates are indicated. Whiskers represent 99% of data values. $^{**}P < 0.01$ (Wilcoxon rank-sum test). **d,** Gap length distribution between xTSS-RNAs and coding transcript TSSs. *y*-Axis indicates number of xTSS-RNAs in each gap length group (*x*-axis). Compiled from two biological replicates. **e,** Percentage of TSS-RNA expressed at indicated reads per kilobase per million reads (RPKM). Compiled from two biological replicates.

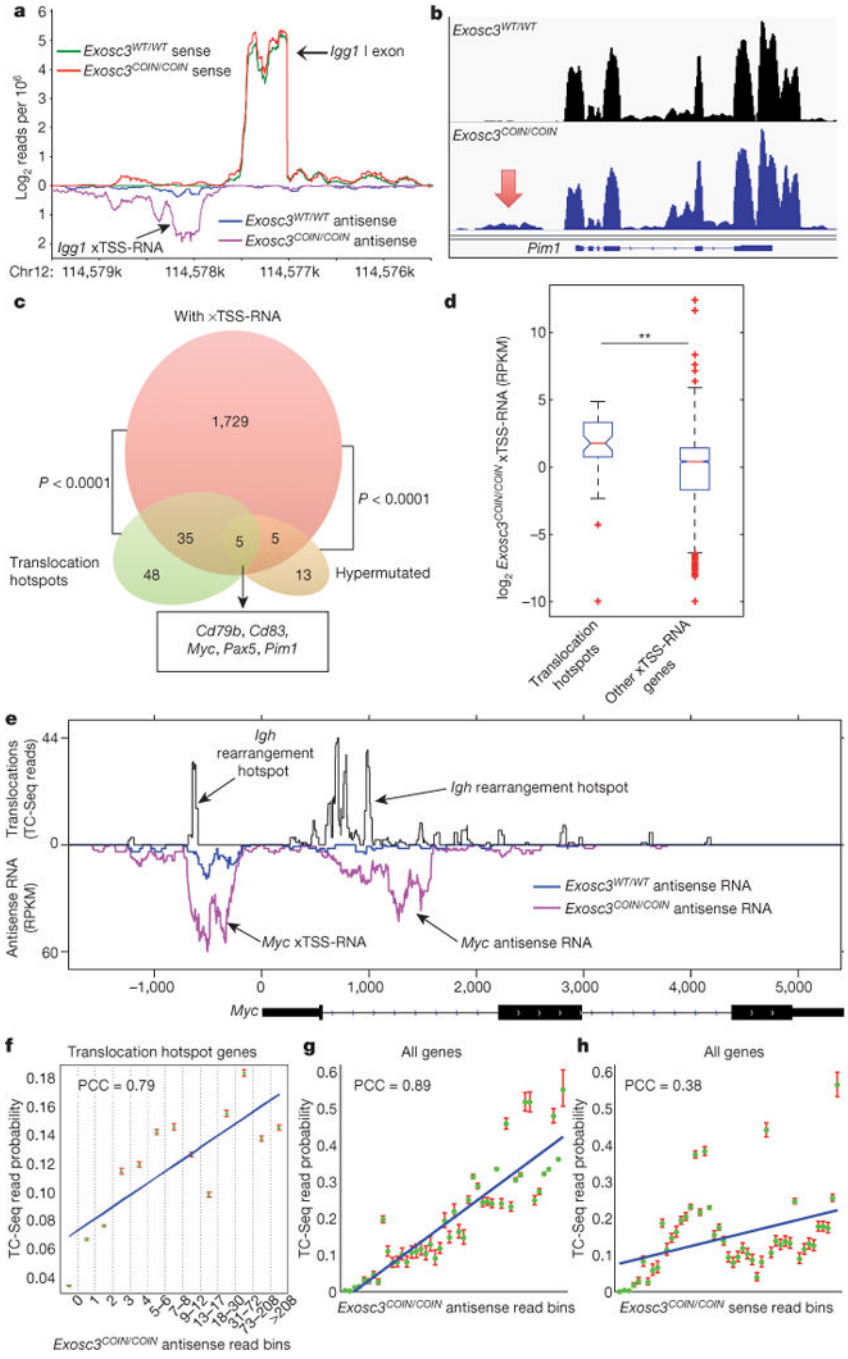


Figure 3. xTSS-RNA expression marks AID-dependent translocation hotspots in the B-cell genome

a, Antisense *Igg1* xTSS-RNA. Strand-specific RNA-seq reads 2 kb upstream and downstream of *Igg1* germline transcript TSS. Compiled from two biological replicates. Chr, chromosome. **b,** Profile of RNA-seq reads at *Pim1* locus (8.6 kb window). Red arrow indicates xTSS-RNA. Four biological replicates. **c,** Venn diagram of genes with xTSS-RNAs, genes undergoing recurrent AID-dependent chromosomal translocations[4] and somatically hypermutated genes[23]. xTSS-RNA group compiled from four biological

replicates. *P* values determined by Fisher's exact test. **d,** Enrichment of gapped xTSS-RNA expression amongst translocation hotspots in *Exosc3*-deficient B cells. Translocation hotspot gene set comprises 40 genes (identified in c) reported to undergo recurrent AID-mediated translocations displaying higher xTSS-RNA expression. 'Other xTSS-RNA genes' set comprises 1,694 genes expressing both xTSS-RNA and cognate mRNA, but not reported as recurrent translocation hotspots. Median values from two biological replicates are indicated. $^{**}P < 0.01$ (Wilcoxon rank-sum test). **e,** *Myc* translocation breakpoints at sites of xTSS-RNA and genic antisense transcription. Mouse B-cell translocation frequency[4] and antisense transcription are shown on the positive and negative *y*-axes, respectively. Compiled from two biological replicates. **f,** Correlation between breakpoints and antisense expression (2 kb upstream of TSS to transcription end site) at translocation hotspots (two biological replicates). Error bars indicate 95% confidence interval and blue line represents robust fit of expected values. Pearson correlation (PCC) is indicated. **g, h,** Probability of translocation breakpoints with respect to antisense (**g**) or sense (**h**) transcription levels. Pearson correlation is indicated.

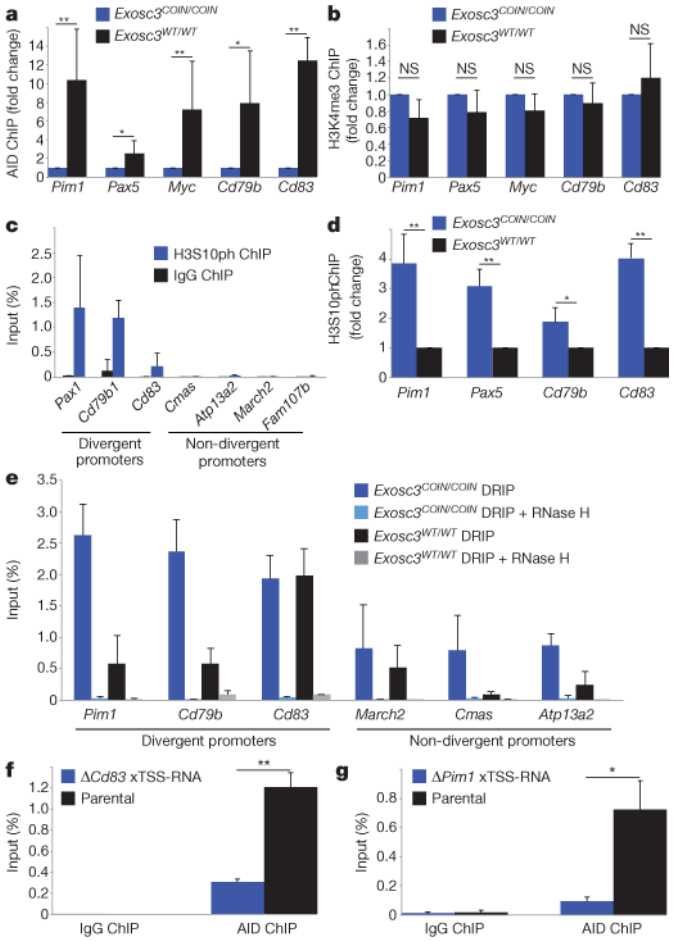


Figure 4. AID recruitment to RNA–DNA hybrid-forming divergently transcribed genes

a, *Exosc3* promotes AID targeting to divergent promoters. ChIP was performed using anti-AID or control IgG. qPCR was performed using TSS upstream-specific primers. PCR amplification for the non-divergent genes *Cmas*, *March2*, *Atp13a2* and *Fam107b* was below the detection limit in anti-AID ChIP. Three pairs of mice of each genotype were used. Mean values from three technical replicates are indicated. Error bars represent s.d. $*P < 0.05$, $**P < 0.01$ ($\log_2$ transformation of *Z*-test). **b,** H3K4me3 is unaltered in *Exosc3*-deficient B cells. ChIP was performed using anti-H3K4me3 (Millipore) or control IgG. qPCR and data analysis are as described in **a**. Two pairs of mice of each genotype were used. NS, not significant (*t*-test). **c, d,** H3S10ph accumulation at divergent promoters. ChIP was performed using anti-H3S10ph (Millipore) or control IgG. qPCR was performed as described in **a.** H3S10ph enrichment relative to input in $Exosc3^{WT/WT}$ (c) or fold change of $Exosc3^{COIN/COIN}$ is indicated (**d**). Three pairs of mice of each genotype were used. Mean values from three technical replicates are indicated. **c**, $**P < 0.01$ (*t*-test) for H3S10ph ChIP between divergent and non-divergent promoter sets. **d,** $*P < 0.05$, $**P < 0.01$ ($\log_2$ transformation of *Z*-test). **e**, RNA–DNA hybrid accumulation at divergently transcribed genes in *Exosc3*-deficient B cells. qPCR was performed as described in **a**. Two pairs of mice of each genotype were used. Mean values from three technical replicates are shown. Error bars represent s.d. $**P < 0.01$ (*t*-test) for $Exosc3^{COIN/COIN}$ DNA:RNA immunoprecipitation

(DRIP) between divergent and non-divergent promoter sets. **f, g,** Deletion of xTSS-RNA-expressing region impairs AID targeting to divergently transcribed genes. Anti-AID ChIP was performed on parental or clonal CH12F3 B-cell lymphoma lines containing CRISPR-mediated deletions of *Cd83* (**f**) or *Pim1* (**g**) xTSS-RNA-expressing regions. qPCR was performed using exon 1 specific primers. Mean values from three technical replicates are indicated. Error bars represent s.d. $*P < 0.05$, $**P < 0.01$ (*t*-test). Indicated genotypes (except **f, g**) are on a *ROSA26*$^{CreERt2/1}$ background and B cells were treated with 4-OHT and stimulated with LPS plus IL-4.