SUPPLEMENTAL INFORMATION

Teng et al., RAG Represents a Widespread Threat to the Lymphocyte Genome

EXTENDED EXPERIMENTAL PROCEDURES

Mouse strains

Dβ, R1-/-β, R2-/-β, BD, R1-/-B, and R2-/-B mice were described previously (Ji et al., 2010). Other mouse strains used were core RAG1 (Dudley et al., 2003), R2ΔC (previously named "core RAG2") (Liang et al., 2002), *Atm-/-* (Barlow et al., 1996), *H2ax-/-* (Celeste et al., 2002), *Mdc1*-/- (Lou et al., 2006), and *P53*-/- (Jacks et al., 1994). Total thymocytes were harvested from whole thymuses, and pre-B cells were isolated from whole bone marrow using CD19 magnetic beads (Miltenyi Biotec). All animal procedures were approved by the Institutional Animal Care and Use Committee of Yale University.

RAG ChIP-seq peak calling

ChIP-seq peaks were called using MACS 1.4.2 (Zhang et al., 2008). For RAG1, RAG2 and H3K4me3, we used the default parameters (P-value cutoff for peak detection=1e-5). For H3K4me1 and H3K27Ac, we used parameters suitable for broader peaks (--nolambda,--nomodel) (Feng et al., 2011). All peak lists were filtered against the corresponding DNA input as a control (GEO accession numbers: GSM851333, mouse thymus input; GSM1040575 and GSM1040576, mouse pre-B cell inputs; GSM956030, human thymus input). RAG1 and RAG2 samples were also filtered against the corresponding knock-out sample as a control. The intersection resulting from these two filtering processes was taken as the final peak set for each sample. Subsequently, we defined RAG binding sites in thymocytes as the union of RAG peaks identified in wild type and Dβ thymocytes (which showed 72% overlap). RAG binding sites in pre-B cells were defined as those identified in BD pre-B cells. For human thymocyte samples, where the sequencing depth was lower, we first concatenated the reads from multiple samples of the same IP and then evaluated peaks. Details regarding the ChIP-seq parameters for each experiment, including the number of unique mappable reads, are provided in Table S1.

RNA-seq

Binding of RAG1 and RAG2 bind to transcriptional regulatory elements (some associated with critical T- and B-lineage specific genes (Table S1) raised the possibility of a novel regulatory role for RAG, distinct from its V(D)J recombinase activity. To test whether RAG binding influenced gene expression programs in a nuclease-independent manner, RNA-sequencing (RNA-seq) was performed on recombination-incompetent thymocytes (in triplicate) from Dβ, R1-/-β, and R2-/-β mice (Figure 3I). Total thymocyte RNA was isolated using Trizol (Invitrogen), and RNA-seq libraries were prepared using the Illumina Tru-seq RNA sample prep kit V2 following manufacturer instructions. Samples prepared from Dβ and R1-/-β mouse strains were directly comparable because both are on a B6 background (~97%). We could not make equivalent comparisons to the R2-/-β strain, which was on a B6/129 background, and showed significant strain-dependent variation in gene expression (data not shown). As an alternative, we used *in vivo* αCD3 injection to stimulate the production of double-positive thymocytes (Shinkai and Alt, 1994) in *Rag1-/-* and *Rag2-/-* mice that had been backcrossed onto a B6 background.

Libraries were run on Illumina HiSeq Sequencers and Illumina tools were used to generate fastq files from RTA output. The short reads were then aligned to the mm9 genome with GSNAP (version 2012-07-20) (Wu and Nacu, 2010), allowing known splice junctions extracted from RefSeq GTF (June 2013). After sorting of the alignment files, htseq-count (Htseq version 0.5.4) (Anders et al., 2014) was used to count the number of reads fitting each gene model in the RefSeq GTF (after removing small RNAs). Counts for each gene were then read into R (version 3.1, http://www.R-project.org) and analyzed with DESeq (version 1.16) (Anders and Huber, 2010). In particular, graphs in Figure S3I were generated by normalizing the count data, estimating blind dispersions, and calculating a variance stabilizing transformation that approximates a log transformation for larger counts (see DESeq documentation for more details). The actual graphs are smoothscatter graphs inside lattice panels. The statistical significance of differential expression was assessed using negative binomial p-values as returned by DESeq, which were adjusted with the Benjamini-Hochberg method to control for the false discovery rate (FDR).

The RNA-seq experiments did not reveal evidence for widespread RAG-dependent variations in gene expression, with tight correlations observed for most genes in triplicate RNA-seq experiments (Figure S3I). In the comparison of Dβ and R1-/-β thymocytes (Figure S3I, left side), which looks for a possible contribution of RAG1, only nine genes were observed to change in expression by more than two fold with a FDR < 0.01 (Table S2). These nine genes were expressed at low levels and none displayed significant RAG1 binding, arguing against a biological significance for the small changes observed.

In the comparison of thymocytes from α-CD3 injected *Rag1-/-* and *Rag2-/-* mice (Figure S3I, right side), which interrogates a possible contribution of RAG2, 121 genes were observed to change in expression by more than two fold with a FDR < 0.01 (Table S2). It was noticed, however, that 23 of the differentially expressed genes mapped to chromosome 2 within 30 Mb of the RAG locus (chr2: 101,464,904-101,489,888 in mm9; Table S2 uses mm9 coordinates). This strongly suggested that polymorphic differences between the RAG1-knockout and RAG2-knockout alleles were responsible. Among the other 98 genes, many showed no evidence of RAG1 or RAG2 binding, and of those that did exhibit RAG binding, no correlation could be discerned between the magnitude of RAG binding and the magnitude or direction of the gene expression changes. Therefore, while we cannot exclude small effects of RAG2 on the expression of some genes, we do not interpret our data as providing substantial support for such an idea.

Importantly, these experiments probed only the contribution of RAG *binding* to gene expression, and not the contribution of RAG-mediated *cleavage*, which has been shown to stimulate specific gene expression programs related to DNA repair during lymphocyte development (Helmink and Sleckman, 2012) as well as influence gene expression in other hematopoietic lineages (Karo et al., 2014).

MNase-seq

Thymocytes were cross-linked for 15 minutes with 1% formaldehyde, and the crosslinking reaction was quenched with 0.125 M glycine. Cells were washed twice with ice-cold PBS containing 1 mM PMSF and 1 µg/mL pepstatin A. Five million cross-linked cells were resuspended in MNase digestion buffer (50 mM Tris-HCl, pH 7.6; 1 mM $CaCl_2$, 0.2% Triton X, 5 mM sodium butyrate, 1 mM PMSF, 1 µg/mL pepstatin A). Micrococcal nuclease (MNase, 0.2 U) was added to the sample, which was incubated at 25 ºC for 5 minutes. The nuclease digestion was terminated with 5 mM EDTA. Crosslinks were reversed using an overnight incubation at 65 ºC in the presence of 0.5% SDS and 200 µg/mL Proteinase K. DNA was isolated by standard phenol:chloroform extraction and ethanol precipitation. The sample was then subjected to RNAse digestion (10 µg/mL RNAse A at 37 ºC for 1 hour) and a second proteolysis step to remove the RNAse (200 µg/mL Proteinase K, 0.3% SDS, 65 ºC, 1 hour). Again, the DNA was isolated by phenol:chloroform extraction and ethanol precipitation and resuspended in water. The sample was run on a 2% agarose gel, and the ~150 bp mononucleosome band was excised and purified using the Qiagen Gel Extraction kit. Libraries for sequencing were prepared as for ChIP-seq.

Libraries were run on Illumina HiSeq Sequencers and Illumina tools were used to generate fastq files from RTA output. The short reads were then aligned to the mm9 genome with Bowtie (versions 0.12.8 to 1.0.0) (Langmead et al., 2009). After sorting the resulting alignment files, custom software was used to determine mean density of reads around RefSeq promoters that overlapped a RAG peak. MNase reads were shifted by half a nucleosome size towards the center of the nucleosome. The resulting data was imported into R (version 3.1) for visualization. The MNase-seq experiments did not reveal RAG-dependent global alterations in nucleosome positioning (Figure S3J).

Transposase-Accessible Chromatin using sequencing (ATAC-seq)

ATAC-seq was performed as described (Buenrostro et al., 2013). Nuclei were prepared from sorted small pre-B cells ($1x10^5$ for each ATAC-seq) and resuspended in the transposase reaction mix (25 µL 2X Tagment buffer, 2.5 µL Tagment DNA enzyme (Illumina, FC-121-1030) and 22.5 µL nuclease-free water). The transposition reaction was carried out at 37 °C for 30 min. Directly following transposition the sample was purified using a Qiagen MinElute kit. Following purification, library fragments were amplified using Nextera PCR Primers (Illumina Nextera Index kit) and NEBnext PCR master mix (New England lab, 0541) for a total of 10-12 cycles. The libraries were then purified using a Qiagen PCR cleanup kit. The amplified, adapter ligated libraries were size selected using Life Technologies' E-Gel® SizeSelect™ gel system in the 150-650bp range. The size-selected libraries were quantified using the Agilent Bioanalyzer and by qPCR in triplicate using the KAPA Library Quantification Kit on the Life Technologies Step One System. Libraries were sequenced on the Illumina Hiseq2000 system to generate 75-100M, 50 bp paired end reads.

ATAC-seq analysis followed the procedure described by Buenrostro *et al.* Read alignment positions were adjusted according to their strand: +4 bp for + strand alignments, and -5 bp for – strand alignments. Open chromatin regions were called using Zinba (Rashid et al., 2011) with a window size of 300 bp, an offset of 75 bp, and a posterior probability threshold of 0.8. For nucleosome positioning, properly paired alignments were filtered by their fragment size. Fragment sizes less than 100 bp were considered nucleosome free and replaced with a single BED region, and used as a background. Sizes between 180 and 247 bp were considered mononucleosomes and replaced with a single BED region; sizes between 315 and 473 bp were considered dinucleosomes and replaced with two BED regions, each spanning half the overall fragment length; and sizes between 558 and 615 bp were considered trinucleosomes and replaced with three BED regions, each spanning one third of the overall fragment length; the mono-, di-, and tri-nucleosome regions were concatenated and used as the nucleosome signal. The resulting BED regions were analyzed using DANPOS (Chen et al., 2013) with the parameters –p 1 –a 1 –d 20 —clonalcut 0 to identify regions enriched or depleted for nucleosomes.

Plasmid recombination assay

The competitive recombination plasmid was constructed from the backbone of the pSF290 recombination substrate (Fugmann and Schatz, 2001). The pSF290 recombination cassette was excised by a BglII/NotI digest. The recombination cassette from pJHSASNXB (Jung et al., 2003) was inserted into the pSF290 backbone by InFusion cloning. This new recombination cassette contained three cloning sites flanked by unique restriction sites: Position 1 (SalI/AscI), Position 2 (SpeI/NotI), and Position 3 (XhoI/BamHI). Double stranded oligos containing the desired RSS sequence, 10 bp of coding flank, and overhangs complementary to the restriction ends were ligated into the appropriate position. To test for recombination between a cRSS and a partner consensus RSS: either the *Lmo2* 12-cRSS or the modified *Ttg1* 23-cRSS was cloned into Position 1, the test cRSS was cloned into Position 2, and a consensus 23- or 12-RSS was cloned into Position 3. To create recombination substrates to test for activity between two cRSSs: the 12-cRSS was cloned into Position 1, the *Lmo2* 12-cRSS was cloned into Position 2, and the 23-cRSS was cloned into Position 3. Oligos used for cloning are listed in Table S7. The recombination plasmids and RAG expression plasmids were co-transfected into 293T cells using Lipofectamine 2000 (Invitrogen). After 48 hours, plasmids were recovered from the cells by plasmid minipreps (Qiagen). The recombination cassette was amplified by nested PCR (primers for primary PCR reaction: CCCTGATTCTGTGGATAACC and CCTCTACAAATGTGGTATGGC, primers for nested PCR reaction: TACCGGACTCAGATCCGACAGGTTTCCCGACTG and TCTAGAGTCGCGGCCGAGCAACTGACTGAAATGCCTC). The PCR reactions were run on a 1% agarose gel with ethidium bromide, and the bands were quantitated using Quantity One software (Bio-Rad).

Genotyping PCRs for the RSSki cell lines

A PCR assay specific to the wild type (unmodified) allele was used to determine if targeting occurred on one allele (wild type allele amplified by PCR) or two (wild type allele not amplified). A PCR assay was also conducted to determine if additional un-targeted integrations of the targeting vector were present in the genome, using one primer complementary to sequences in the homology arm of the construct, and one primer complementary to sequences in the plasmid backbone flanking the homology arm (i.e. sequences that would not be retained in the targeted locus after homologous recombination).

PCR assays to detect cryptic recombination *in vivo*

Pairs of cRSSs falling within 1 kb of a RAG1 binding peak were identified using the DnaGrab implementation of the RIC algorithm (Merelli et al., 2010). Any candidate sites with inhibitory sequence motifs flanking the heptamer of the cRSS were excluded (Ezekiel et al., 1997; Gerstein and Lieber, 1993). Only cRSSs proximal to the top 100 peaks of RAG1 binding and with DnaGrab RIC scores in the top 5% were considered for further analysis. Nested PCR assays using JumpStart REDTaq (Sigma) were conducted to test for cryptic deletion or inversion events in genomic DNA harvested from mouse thymocytes (PCR primers listed in Table S7). Cryptic deletion junctions in the *Notch1* and *Bcl11b* locus were detected using PCR protocols adapted from Ashworth et al., 2010, and Sakata et al., 2004, respectively, and the

PCR products were gel purified and sequenced. Nested PCR assays were used to test for RAG-mediated inversions or deletions in the *Rag1*, *Cd8a,* and *Ets2* loci (primers listed in Table S7). For each of these three loci, at least 20 million genome equivalents were screened by PCR. The excised minicircles resulting from the *Bcl11b* deletions in *Atm-/-* thymocytes were detected using PCR primers that amplified across the signal joint (primers listed in Table S7).

cRSS density and CA density analysis

The RIC algorithm was used to identify cRSSs passing RIC score thresholds of ≥ -45 for 12-RSSs and ≥ -65 for 23-RSSs (Davila et al., 2007). Using the GRCm38.p2 reference assembly of the mouse genome, we calculated the 12- and 23-cRSS density (cRSS per bp) for each RAG1 peak in thymocytes or pre-B cells. RAG1 peaks in the antigen receptor genes were excluded from these analyses. For each RAG1 peak analyzed, a comparator set of 100 separate "shuffle" regions was generated using the bedtools "shuffle" command (http://bedtools.readthedocs.org/en/latest/). Briefly, for each peak, we randomly selected a genomic location on the same chromosome with the same length as the RAG1 peak region, ensuring that the randomly assigned location did not overlap with another RAG1 peak. The average 12- and 23-cRSS density was calculated over all shuffle regions, and the procedure was repeated 10,000 times for each set of RAG1+ peaks to generate 10,000 average shuffle region values. In no case did the cRSS density at a set of RAG1+ peaks exceed that of one of the corresponding average shuffle region values, yielding P<10E-4. However, this likely underestimates the statistical significance because the values for the RAG1+ peaks were more than 27 standard deviations away from the average of the 10,000 shuffle region values.

For the TSS-centric analyses of cRSS, heptamer, nonamer, and CA density, we used annotation release 104 associated with GRCm38.p2. We selected items of feature type "gene" that also had an associated "mRNA" and "BestRefSeq" annotation. This resulted in 20,499 genes used for analysis. We classified each TSS within 2 kb of a RAG1 peak as RAG1$^+$; all remaining TSSs were classified as RAG1^{-}. For genes with multiple TSS annotations, the TSS overlapping with a RAG1 peak was selected; if none of the TSSs overlapped with a RAG1 peak, one TSS was selected randomly. For each TSS, we calculated the 12- and 23-cRSS, "strong heptamer", "strong nonamer" and CA dinucleotide density for the region -/+ 5 kb surrounding each gene start location using a sliding window of 500 bp. A similar approach was applied to determine the density of strong heptamers and nonamers in human RAG1+ and RAG1- TSSs (25,859 TSSs defined by hg19 RefSeq annotations).

We used H3K4me3 ChIP-seq datasets from six mouse non-lymphoid tissues (from GEO Series GSE29184) to define lists of genomic regions that were comparable to H3K4me3^{+} RAG1^{+} peaks in thymocytes and pre-B cells (referred to as “RAG1^{+} regions”) in terms of their H3K4me3 profiles, CpG island density, GC content, and distribution among TSS and non-TSS regions. First, each RAG1+ region was mapped onto its overlapping H3K4me3 peak (less than 1% of RAG1 peaks were H3K4me3^{-} and were excluded from the analysis). The region -/+ 1 kb (using the peak summit as the center point) surrounding each of the H3K4me3 peak regions thus identified was scanned using a sliding 150 bp window, and the number of CpG dinucleotides was counted in each window. A CpG “score” was assigned based on the window with the highest number of CpGs. Each RAG1^{+} H3K4me3 region was then matched to a comparable H3K4me3 peak in each non-lymphoid cell type. The CpG score of each RAG1+ H3K4me3 peak was used to identify the group of H3K4me3 peaks from a non-lymphoid tissue with an identical CpG score. From that group, an H3K4me3 peak was chosen to ensure similar H3K4me3 ranks, GC content (across the -/+2kb region), and comparable TSS or non-TSS identity (see Statistics). After comparable peak regions were identified in each tissue type, regions of -/+ 2kb centered on the peak summits were identified and used to calculate the 12- and 23- cRSS, heptamer, and nonamer counts for each set of regions.

A second set of comparable genomic regions from the six non-lymphoid tissues was created using a slight modification of the above procedure. Again, the CpG score of each RAG1+ H3K4me3 peak was used to identify the group of H3K4me3 peaks from a non-lymphoid tissue with an identical CpG score. From this group, the equivalent peak was specified as the one whose H3K4me3 rank was closest to the rank of the RAG1+ H3K4me3 peak, without explicit reference to GC content. This resulted in non-lymphoid peak sets whose GC content was similar to that of the RAG1+ H3K4me3 peaks, but had slightly higher H3K4me3 rank scores (meaning somewhat lower H3K4me3 values). These peaks sets overlapped the first set by 52-60%, and yielded very similar results to those shown in Figure 5E-H (data not shown).

To ascertain if the mouse genome contains strong cRSSs outside of the antigen receptor loci, it was scanned using the RIC algorithm for 12- and 23-cRSSs passing stringent thresholds of -30 and -48, respectively. This yielded 12,142 strong 12-cRSSs and 17,766 strong 23-cRSSs, exclusive of the antigen receptor loci. The stringency of the thresholds used in this analysis is illustrated by the fact that most (97%) but not all of the 356 functional antigen receptor gene RSSs used to create the RIC algorithm (Cowell et al., 2002) pass these thresholds. Hence, the mouse genome contains thousands of sequences that in principle could support substantial levels of RAG-mediated recombination if positioned in accessible chromatin in developing lymphocytes and able to synapse with a compatible partner RSS.

Heptamer and nonamer density analysis

To further explore the contribution of nonamer sequences to the RAG1 binding pattern, we searched for an enrichment of RAG1 ChIP-seq reads over nonamer sequences. Position weight matrices were generated using the functional 12- and 23-RSSs used to formulate the RIC algorithm (Table S4). A total of 15 heptamer and 216 nonamer sequences passing a score of 7.33 and 7.06, respectively, were selected as high-scoring motifs. Given these matrices, we scanned the mouse genome for heptamers and nonamers using the FIMO tool (Grant et al., 2011). The occurrence of these selected heptamers and nonamers were then determined for each RAG1+ region (each being mapped to its overlapping H3K4me3 peak) and its matched comparator region in different non-lymphoid cell types as described above for the cRSS analysis. We identified 12,134 high scoring nonamers within RAG1 peaks from wild type and Dβ thymocytes (excluding antigen receptor loci) and the density of sequence reads over these regions was compared to the density of reads over the surrounding ±250 bp. No nonamer-centric RAG1 peak was detected (data not shown). In contrast, we could detect enrichment of RAG1 binding at the 16 Jα RSSs within the *Tcrα* recombination center from Dβ thymocytes, as expected (data not shown). These analyses combined with our other data suggest that neither nonamers nor cRSSs are the dominant determinant of RAG1 binding outside of antigen receptor loci. This does not contradict the possibility that some proportion of nonamers contribute to RAG1 recruitment (as suggested by nonamer enrichment at RAG1+ versus RAG1-regions), only that nonamers as a group do not do so sufficiently strongly so as to reveal binding enrichment.

Statistics

One-sided Mann-Whitney tests were used to compare the distribution of cRSSs, heptamers, and nonamers between the RAG1+ regions in thymocytes and pre-B cells versus the comparison peaks from non-lymphoid tissues (Figure 5). Two-sided Mann-Whitney tests were used to verify that the H3K4me3 rank and CpG value were comparable between the RAG1+ regions in lymphocytes and the comparison peaks from non-lymphoid tissues. Similarly, two-sided T-tests were used to verify that GC content was comparable between the RAG1+ lymphocyte regions and non-lymphoid comparator regions.

CpG island analysis

CpG islands were identified by scanning regions of interest with a sliding 150 bp window (in increments of 1 bp), requiring that one window contain more than 5% (that is, more than 7) CpG dinucleotides. This is a stringent criterion, corresponding to an enrichment ratio of 0.8. The enrichment ratio is defined as the ratio of observed number of CpGs to the number of CpGs expected if the dinucleotide was randomly represented in the genome. Enrichment ratios are generally low (0.1 to 0.2) in vertebrate genomes because the dinucleotide has been depleted, and thresholds of 0.55 to 0.65 have often been used to identify CpG islands in previous studies (Ramirez-Carrozzi et al., 2009). Hence, our use of a ratio of 0.8 affords increased confidence that the regions identified are indeed CpG islands.

Binding patterns of WT versus NBD-mutant RAG1

pMSCV retroviral vectors were used to reconstitute *Rag1-/-* v-abl cell lines with either wild type RAG1 or RAG containing mutations in its nonamer-binding domain (NBDm: R391A, R393A, R407A). Western blots confirmed comparable expression levels of the WT and NBDm RAG1 proteins (data not shown). The cells were treated with STI-571 for 48 h, and cross-linked for ChIP, as previously described (Ji et al., 2010).

PCR assays for recombination and translocation in the RSSki cell lines

Genomic DNA was harvested from RSSki v-abl cell lines after 48 hours of incubation with STI-571. PCR assays were used to detect the genomic deletions resulting from coding joint (CJ) formation (see Table S7 for primer sequences). Coding joints from the CD79b RSSki were amplified using the primers CD79b_F and CD79b_LoxR, and coding joints from the E1 RSSki were amplified using the primers E1_ F and E1_LoxR. A semi-quantitative PCR assay was used to estimate the efficiency of recombination between the knocked-in RSSs. The CJ PCR product was amplified from one of the 12/23 knock-in cell lines, gel purified, and quantitated. The purified PCR product was then used to create a standard curve by diluting it into 50 ng of genomic DNA in ratios equivalent to 100%, 50%, 25%, 20%, 15%, 10%, 5%, 2%, 1%, 0.5%, and 0.1% recombination. Each test DNA sample was normalized to 50 ng/µL, and the intensity of the CJ PCR product was used to calculate the recombination efficiency based on the standard curve.

To test the effects of ATM inhibition on the behavior of the RSSki, the cell lines were pre-treated for 1 hour with 15 µM KU-55933 (Tocris 3544) or with DMSO. Then, 3 µM STI was added to the cell cultures for 48 h. Genomic DNA was isolated from the treated cells, and assayed for CJ formation as described above, or for translocations. Nested PCR assays were used to detect translocations between the RSSki loci and the Igκ locus. To generate CRISPR/Cas9-mediated DSBs in the genome, we cloned an *Hprt*-specific guide RNA sequence into pSpCas9(BB)-2A-GFP (Addgene pX458). Abelson cell lines were transfected with this expression construct, and incubated with STI-571 for 48 hours. PCR products were TOPO-TA-cloned into sequencing vectors, and sequenced.

SUPPLEMENTAL REFERENCES

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol *11*, R106.

Anders, S., Pyl, P.T., and Huber, W. (2014). HTSeq-a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166-169.

Barlow, C., Hirotsune, S., Paylor, R., Liyanage, M., Eckhaus, M., Collins, F., Shiloh, Y., Crawley, J.N., Ried, T., Tagle, D.*, et al.* (1996). Atm-deficient mice: a paradigm of ataxia telangiectasia. Cell *86*, 159-171.

Celeste, A., Petersen, S., Romanienko, P.J., Fernandez-Capetillo, O., Chen, H.T., Sedelnikova, O.A., Reina-San-Martin, B., Coppola, V., Meffre, E., Difilippantonio, M.J.*, et al.* (2002). Genomic instability in mice lacking histone H2AX. Science *296*, 922-927.

Chen, K., Xi, Y., Pan, X., Li, Z., Kaestner, K., Tyler, J., Dent, S., He, X., and Li, W. (2013). DANPOS: dynamic analysis of nucleosome position and occupancy by sequencing. Genome Res *23*, 341-351.

Davila, M., Liu, F., Cowell, L.G., Lieberman, A.E., Heikamp, E., Patel, A., and Kelsoe, G. (2007). Multiple, conserved cryptic recombination signals in VH gene segments: detection of cleavage products only in pro B cells. J Exp Med *204*, 3195-3208.

Dudley, D.D., Sekiguchi, J., Zhu, C., Sadofsky, M.J., Whitlow, S., DeVido, J., Monroe, R.J., Bassing, C.H., and Alt, F.W. (2003). Impaired V(D)J recombination and lymphocyte development in core RAG1-expressing mice. J Exp Med *198*, 1439-1450.

Ezekiel, U.R., Sun, T., Bozek, G., and Storb, U. (1997). The composition of coding joints formed in V(D)J recombination is strongly affected by the nucleotide sequence of the coding ends and their relationship to the recombination signal sequences. Mol Cell Biol *17*, 4191-4197.

Feng, J., Liu, T., and Zhang, Y. (2011). Using MACS to identify peaks from ChIP-Seq data. Curr Protoc Bioinformatics *Chapter 2*, Unit 2 14.

Fugmann, S.D., and Schatz, D.G. (2001). Identification of basic residues in RAG2 critical for DNA binding by the RAG1-RAG2 complex. Mol Cell *8*, 899-910.

Gerstein, R.M., and Lieber, M.R. (1993). Coding end sequence can markedly affect the initiation of V(D)J recombination. Genes Dev *7*, 1459-1469.

Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017-1018.

Jacks, T., Remington, L., Williams, B.O., Schmitt, E.M., Halachmi, S., Bronson, R.T., and Weinberg, R.A. (1994). Tumor spectrum analysis in p53-mutant mice. Curr Biol *4*, 1-7.

Ji, Y., Resch, W., Corbett, E., Yamane, A., Casellas, R., and Schatz, D.G. (2010). The in vivo pattern of binding of RAG1 and RAG2 to antigen receptor loci. Cell *141*, 419-431.

Jung, D., Bassing, C.H., Fugmann, S.D., Cheng, H.L., Schatz, D.G., and Alt, F.W. (2003). Extrachromosomal recombination substrates recapitulate beyond 12/23 restricted VDJ recombination in nonlymphoid cells. Immunity *18*, 65-74.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.

Liang, H.E., Hsu, L.Y., Cado, D., Cowell, L.G., Kelsoe, G., and Schlissel, M.S. (2002). The "dispensable" portion of RAG2 is necessary for efficient V-to-DJ rearrangement during B and T cell development. Immunity *17*, 639-651.

Lou, Z., Minter-Dykhouse, K., Franco, S., Gostissa, M., Rivera, M.A., Celeste, A., Manis, J.P., van Deursen, J., Nussenzweig, A., Paull, T.T.*, et al.* (2006). MDC1 maintains genomic stability by participating in the amplification of ATM-dependent DNA damage signals. Mol Cell *21*, 187-200.

Ramirez-Carrozzi, V.R., Braas, D., Bhatt, D.M., Cheng, C.S., Hong, C., Doty, K.R., Black, J.C., Hoffmann, A., Carey, M., and Smale, S.T. (2009). A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. Cell *138*, 114-128.

Rashid, N.U., Giresi, P.G., Ibrahim, J.G., Sun, W., and Lieb, J.D. (2011). ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. Genome Biol *12*, R67.

Shinkai, Y., and Alt, F.W. (1994). CD3 epsilon-mediated signals rescue the development of CD4+CD8+ thymocytes in RAG-2-/- mice in the absence of TCR beta chain expression. Int Immunol *6*, 995-1001.

Wu, T.D., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics *26*, 873-881.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W.*, et al.* (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol *9*, R137.