SI Methods

Data.

RAG1 ChIP-seq datasets from mouse thymocytes of different genotypes (R1-D708A, cR1, R2ΔC, R2-/-), mouse pre-B cells, human thymocytes, and v-abl cells; as well as RAG2 and H3K4me3 ChIP-seq datasets from mouse thymocytes, were generated in Teng et al. (1). Mouse pro-B PU.1 ChIP-seq data was obtained from GSM1296533. Mouse thymus DNaseI-HS-seq data was obtained from GSM1014185. Human thymus DNaseI-HS-seq data was obtained from GSM1027313. Mouse pre-B ATAC-seq data was obtained from GSE63302. Mouse thymus H3K4me1 and H3K9Ac ChIP-seq data were obtained from GSE29184 and GSE34954, respectively. Mouse pro-B H3K4me3 and H3K27Ac ChIP-seq data were obtained from GSE48555. All H3K4me3 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).

ChIP-seq procedure

RAG1 ChIP-seq was performed as described previously (1, 2). Briefly, total thymocytes were harvested from whole mouse thymuses. Cells were crosslinked with 1% formaldehyde for 15 minutes at room temperature. The crosslinking reaction was quenched with 0.125 M glycine, and the cells were washed twice with cold PBS containing1 mM PMSF and 1 μg/μL pepstatin A. Crosslinked cells were resuspended in high-salt RIPA buffer (10 mM Tris pH 7.4, 1 mM EDTA, 1% Triton X-100, 0.1% sodium deoxycholate, 0.1% SDS, 0.8 M NaCl, 1 mM PMSF, 1 μg/μL pepstatin A) and incubated on ice for 10 minutes. Chromatin shearing was performed in a water bath sonicator (Diagenode Bioruptor or Bioruptor Pico) to obtain fragment sizes of 100-250 bp. Samples were centrifuged (20,000 rcf, 10 min, 4 ºC), and the supernatant was collected. The sheared chromatin was then split into individual aliquots for immunoprecipitation (each ChIP sample contained chromatin from 10-15 million cells). The chromatin was pre-cleared with Protein G Dynabeads (Thermo-Fisher). The Dynabeads were then removed using a Dynal magnetic stand. After preclearing, 2% fish gelatin and 500 ng/μL heparin were added as blocking agents. The desired antibody (5 μg) was added, and the samples were incubated overnight at 4 ºC with rotation. Protein G Dynabeads were blocked with 2% BSA and 50 ng/μL heparin, and were then added to the samples, incubating for 2 hours at 4 ºC with rotation. The Dynabeads were washed (10 min, 4 ºC, with rotation, for each wash): 2x in RIPA containing no salt and 1 mM DTT, 2x in RIPA containing 0.3 M NaCl and 1 mM DTT, 2x in RIPA containing 0.8 M NaCl and 1 mM DTT, 2x in LiCl buffer (0.25 M LiCl, 0.5% NP-40, 0.5% sodium deoxycholate), 1x in TE + 0.2% Triton X-100, and 1x in TE.

Samples were treated with 1 mg/mL proteinase K and 0.3% SDS at 65 ºC overnight. The supernatant was collected, and the beads were washed with TE + 0.5 M NaCl. The wash was combined with the original supernatant. DNA was purified by standard phenol:chloroform extraction and ethanol precipitation. At least three independent IPs were combined for a single ChIP-seq sample. ChIP-seq libraries were prepared and sequenced according to Illumina protocols. All animal procedures were approved by the Institutional Animal Care and Use Committee of Yale University.

H3K27Ac ChIP-seq in REH cells was performed similarly, with the following modifications. After crosslinking, cells were resuspended in SDS Lysis Buffer (50 mM Tris pH 8, 10 mM EDTA, 1% SDS, 1 mM PMSF, 1 µg/µL pepstatin A; 200 µls SDS Lysis Buffer per 10 million cells). For each immunoprecipitation, 300 µls of supernatant (corresponding to 15 million cells) was diluted to 1 mL in ChIP dilution buffer (167 mM NaCl, 16.7 mM Tris pH 8, 1.2mM EDTA, 1.1 % Triton X-100, 0.01% SDS) and pre-cleared with Protein G Dynabeads (Thermo-Fisher). The Dynabeads were then removed using a Dynal magnetic stand. After preclearing and adding blocking agent, anti-H3K27Ac antibody (5 µg; Abcam) was added, and the samples were incubated overnight at 4°C with rotation. Protein G Dynabeads were blocked with 2% BSA and 50 ng/µL heparin, and were then added to the samples, incubating for 2 hours at 4°C with rotation. The Dynabeads were washed (10 min, 4°C, with rotation, for each wash): 2x in Low Salt Wash Buffer (20mM Tris, pH 8, 150mM NaCl, 2mM EDTA, 1% Triton X-100, 0.1% SDS and 1 mM DTT), 2x in High Salt Wash Buffer (20mM Tris, pH 8, 500 mM NaCl, 2 mM EDTA, 1% Triton X-100, 0.1% SDS and 1 mM DTT), 2x LiCl Wash Buffer (10mM Tris pH 8, 1mM EDTA, 0.25 M LiCl, 0.5% NP-40, 1% sodium deoxycholate), 1x in TE + 0.2% Triton X-100, and 1x in TE. Beads were incubated twice with 150 µls of elution buffer (1% SDS, 0.1 M $NaHCO_3$) for 15 minutes at room temperature. Eluates were pooled and incubated at 65°C overnight followed by treatment with RNase and Proteinase K. DNA was purified by standard phenol:chloroform extraction and ethanol precipitation.

Libraries were prepared and sequenced according to Illumina protocols for all ChIP-seq experiments.

Genome alignment

ChIP-seq, DNaseI-HS-seq, and ATAC-seq tags were aligned to the mouse (GRCm38p2/mm10) or human (GRCh37/hg19) genomes using Bowtie (version 0.12.7) (3) with the options: --best --all --strata -n 2 -m1 -l SEED_LENGTH. These parameters allowed for unique alignment to the best stratum with 2 mismatches out of the total read length. The SEED_LENGTH was set to be the length of the read for

DNaseI-HS-seq and ATAC-seq, and 50 for mouse thymocytes WT-RAG1 and REH H3K27Ac ChIP-seq sets where the read length is 75.

Data preparation

H3K4me3 peaks were called using MACS-2.1.0 as described previously (1, 4), resulting in 20,383 H3K4me3 peaks. The RAG1, H3K27Ac, and DNAseI-HS RPKM were calculated for these peaks. The GC content; CpG value; CpA content; and number of 12-RSSs, 23-RSSs, heptamers, and nonamers were determined for the 2 kb surrounding peak summits. Peak summits were defined as the center of the 100 bp window containing the maximal number of reads in all the possible sliding windows within a peak.

cRSSs, heptamer and nonamer density

The RIC algorithm was used to identify cRSSs passing RIC score thresholds of ≥ -45 for 12-RSSs and ≥ -65 for 23-RSSs (5, 6). For heptamers and nonamers, position weight matrices were generated using the functional 12 and 23RSSs used to formulate the RIC algorithm (5), as described previously (1). 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 (7). The occurrence of these selected cRSSs, heptamers and nonamers were then determined for each H3K4me3 peak for mouse thymocytes data (Figure 3A).

CpG island analysis

As described previously (1), 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 (8). Hence, our use of a ratio of 0.8 affords increased confidence that the regions identified are indeed CpG islands.

SI figure legends

Figure S1. Cryptic 12-RSS **(A)** and 23-cRSS **(B)** distribution in mouse genome. The y-axis shows the RIC score of the cRSSs. Cutoffs of -28 and -44 (blue line) were used to define high quality cRSSs.

These RSSs were used in Figure 2 to evaluate RAG1 levels. ATAC-seq cut site profiles at 200 bp surrounding the **(C)** RAG1 and **(D)** PU.1 summits in naked DNA. **(E)** Motifs found in the 30 bp surrounding PU.1 and RAG1 summits (Figure 2D,E). **(F)** PhastCons conservation score at 200 bp surrounding RAG1 (blue line) and PU.1 (black line) summits. (G) The overlap between RAG1 (mouse pre-B) and PU.1 (mouse pro-B) peaks with active promoters (H3K4me3(+), TSS(+)), and active enhancers (H3K4me1(+),H3K27Ac(+)).

Figure S2. A RAG1 targeting model based on mouse thymocytes, was used to predict RAG1 distribution in Human thymocytes**.** Regression error characteristic (REC) curves, plotting the fraction of each peak set (y-axis) that was predicted with a certain maximal residual (x-axis), were used to show the prediction quality of the full regression model (black line), compared with regression using either H3K4me3 (orange line) or H3K27Ac (purple line). Upper and lower limit curves were traced by calculating the residuals between ChIP-seq replicates (black dashed line) or a random feature (red dashed line), respectively

Figure S3. **RAG1 binding pattern in R2-/- mouse thymocytes**. The **(A)** patterns of RAG1 cluster bias and **(B)** correlation with H3K4me3,H3K27Ac and DNaseI-HS, in R2-/- highly resemble the those found in R2ΔC. In both genotypes, RAG1 binding is more biased to Cluster 2 and show weaker correlation with H3K4me3 and stronger correlation with H3K27Ac and DNaseI-HS, compared to WT (WT and R2ΔC plot were taken from figure 5C). **(C)** RAG1 binding levels are similar between WT and R2ΔC (Wilcoxon test; p=0.45); and lower in R2-/- (p=0). **(D)** Schematic representation of RAG1 binding in WT, R2ΔC and R2-/-.

Figure S4. **(A)** Mean nonamer density in the ±1 kb surrounding the H3K4me3 peak summit in quartiles of H3K4me3 and H3K27Ac levels in the two RAG1 peak clusters using data from mouse thymocytes. Note that the color scale spans a narrow range of nonamer densities. **(B)** RAG1 correlation score of H3K27Ac, H3K4me1 or H3K9Ac (the quotient between the correlation with RAG1 and H3K4me3) was calculated using data from WT (blue bars) of R2ΔC (empty bars) mouse. **(C)** Cluster 2 (H3K27Ac-

driven) RAG1 binding sites shows a much higher density heptamer density than does Cluster 1 (H3K4me3-driven) sites (Wilcoxon test ;p<1e-8).

SI references

1. Teng G*, et al.* (2015) RAG Represents a Widespread Threat to the Lymphocyte Genome. *Cell* 162(4):751-765.
2. Ji Y*, et al.* (2010) The in vivo pattern of binding of RAG1 and RAG2 to antigen receptor loci. *Cell* 141(3):419-431.
3. Langmead B, Trapnell C, Pop M, & Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10(3):R25.
4. Zhang Y*, et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 9(9):R137.
5. Cowell LG, Davila M, Kepler TB, & Kelsoe G (2002) Identification and utilization of arbitrary correlations in models of recombination signal sequences. *Genome Biol* 3(12):RESEARCH0072.
6. Davila M*, et al.* (2007) Multiple, conserved cryptic recombination signals in VH gene segments: detection of cleavage products only in pro B cells. *The Journal of experimental medicine* 204(13):3195-3208.
7. Grant CE, Bailey TL, & Noble WS (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics* 27(7):1017-1018.
8. Ramirez-Carrozzi VR*, et al.* (2009) A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. *Cell* 138(1):114-128.