Cell Reports, Volume *15*

Supplemental Information

Two Mutually Exclusive Local Chromatin States Drive Efficient V(D)J Recombination

Daniel J. Bolland, Hashem Koohy, Andrew L. Wood, Louise S. Matheson, Felix Krueger, Michael J.T. Stubbington, Amanda Baizan-Edge, Peter Chovanec, Bryony A. Stubbs, Kristina Tabbada, Simon R. Andrews, Mikhail Spivakov, and Anne E. Corcoran

Supplemental Information

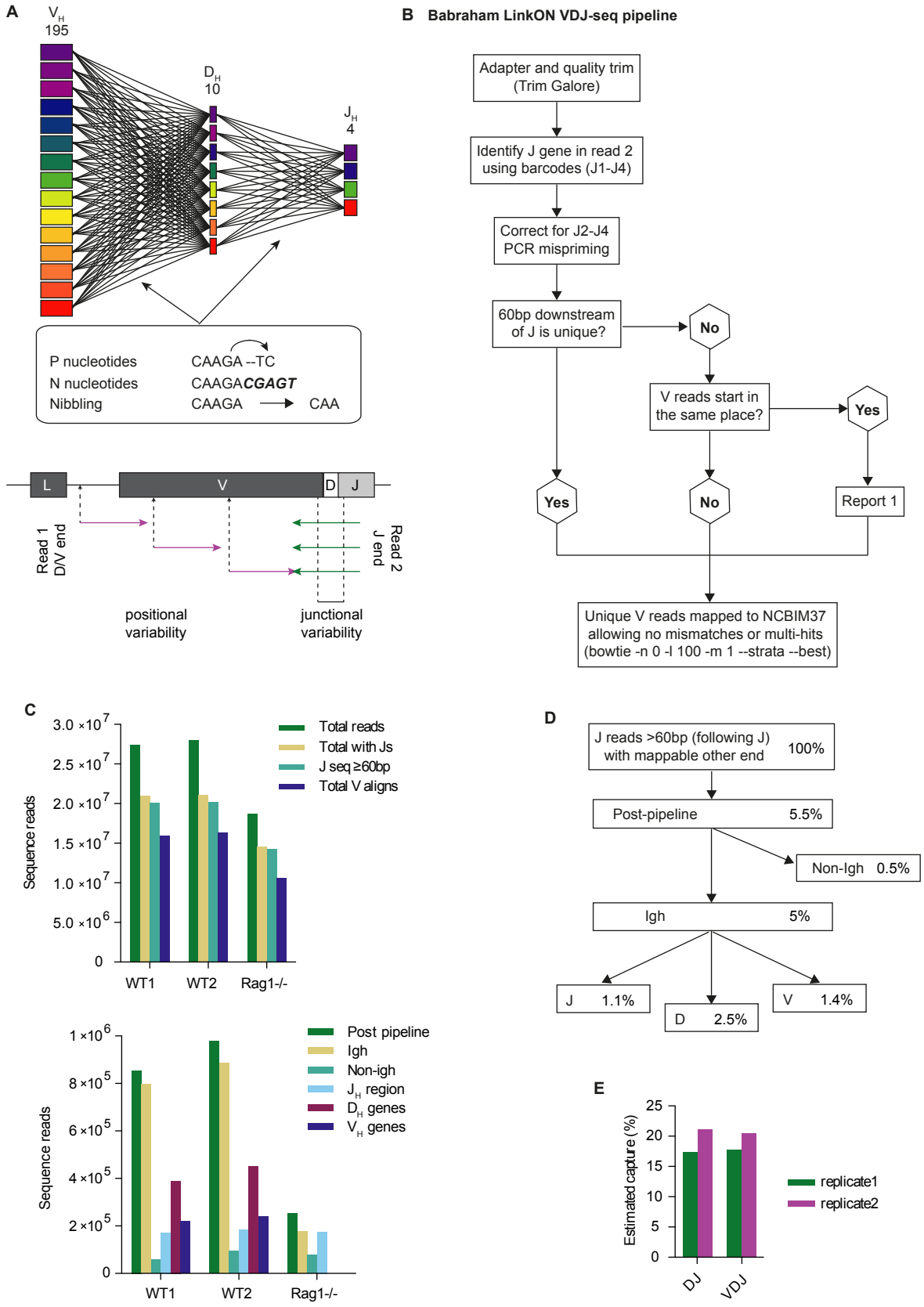


Figure S1 ***Related to Figure 1*** **Details of the VDJseq pipeline, read counts and estimated capture**
A) *Upper*, V(D)J_H recombination gives both combinatorial and junctional diversity as shown. *Lower*, for deduplication VDJ-seq uses both positional variability due to differential sonication fragmentation of genomic DNA in read 1 (V_H/D_H end), and combinatorial plus junctional diversity in read 2 (J_H end). **B)** Babraham LinkON pipeline used to identify unique recombination events in VDJ-seq sequence files. **C)** Read counts at each stage through the pipeline for WT and Rag1$^{-/-}$ pro-B libraries. **D)** Percentage of sequences at each stage of processing of VDJ-seq data. **E)** Estimated capture of VDJ-seq based on the frequencies of DJ_H and VDJ_H alleles in DNA FISH of the same pro-B cell population (assuming 95% complete DJ_H recombination) and normalising for yield per cell in each DNA prep.

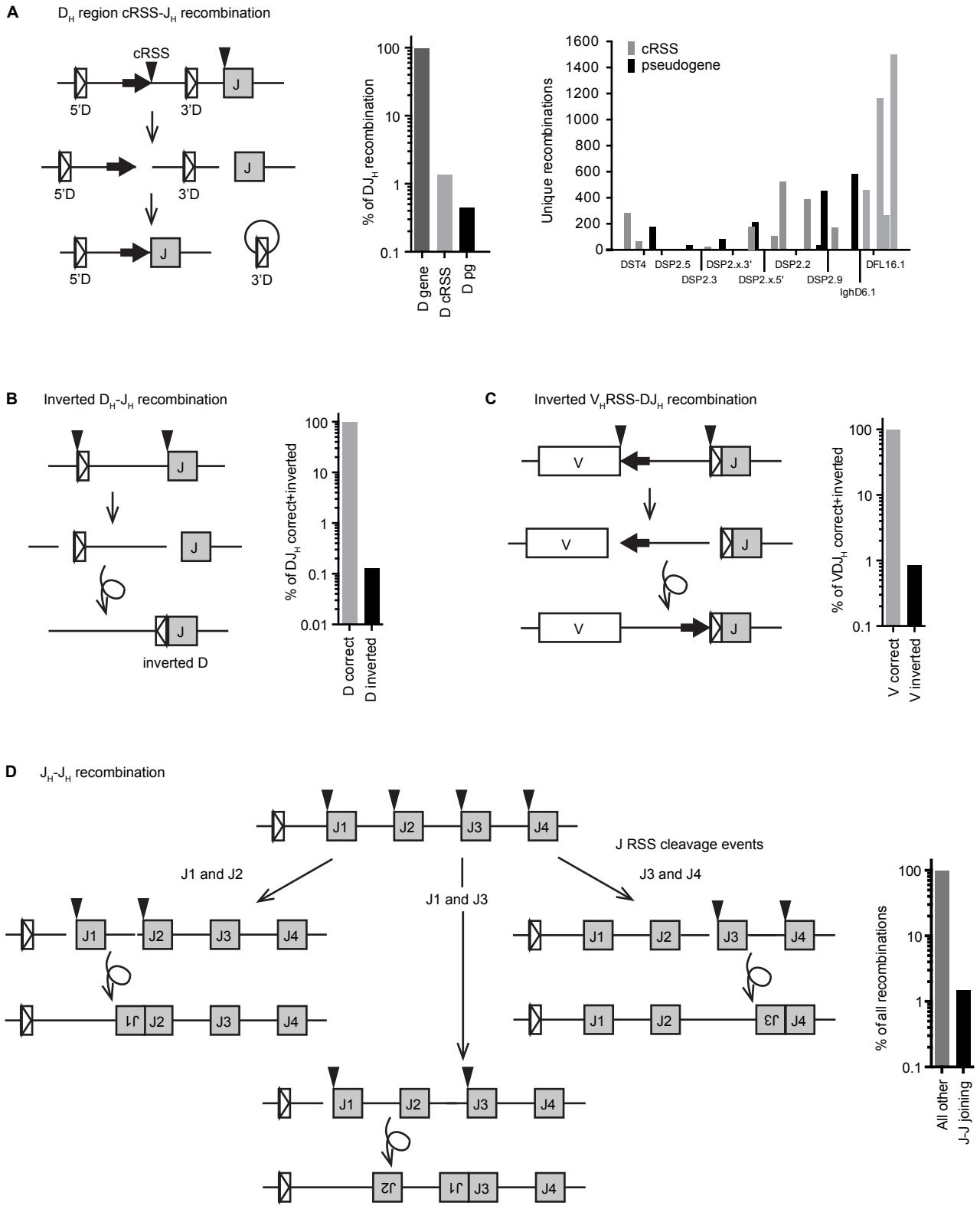


Figure S2 ***Related to Figure 2*** **Alternative and aberrant recombination events detected by VDJ-seq**
A) Low frequency recombination events were detected for D_H pseudogenes and cryptic RSSs (cRSSs) within the D_H region. **B)** D_H gene inversion recombination as documented previously was a low frequency event. **C)** Inversion V_HRSS-DJ_H recombination involving the normal V_H RSS heptamer used in the inverted orientation (with cleavage 7bp distal to the end of the V_H exon) together with a poor quality nonamer within the V_H exon itself. When the DJ_H join is normally cleaved the resulting fragment is inverted and joined generating a non-functional V_HRSS-DJ_H segment. **D)** J_H-J_H joining. Simultaneous cleavage of J_H RSSs and inversion and joining of the resulting fragment generates inverted J_H-J_H joins.

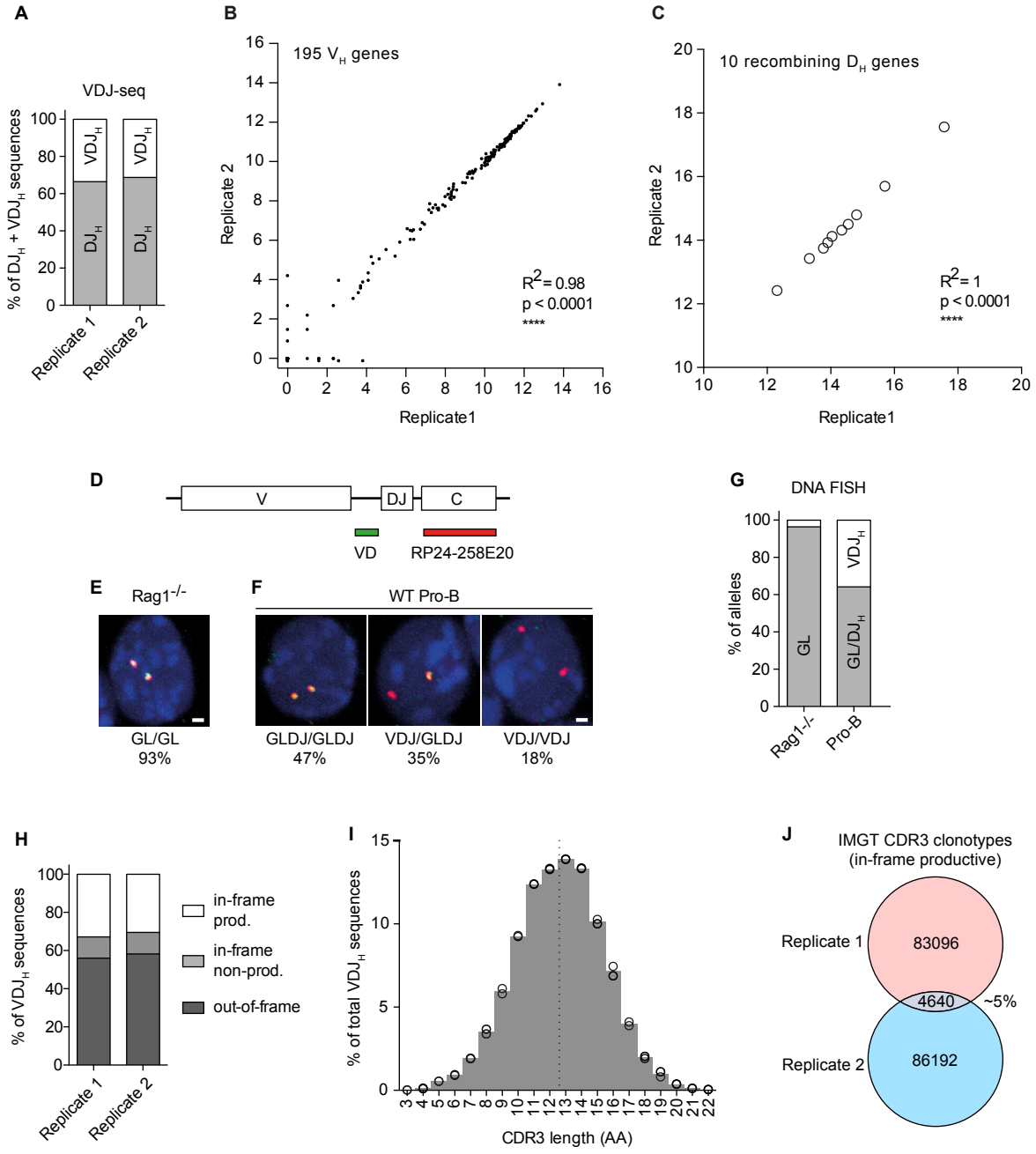


Figure S3 ***Related to Figure 1*** **Quality control of VDJ-seq**

A) Quantitation of VDJ_H and DJ_H rearranged sequences in WT pro-B cells by VDJ-seq. Sense orientation reads across the entire D_H and V_H regions were counted and each presented as a percentage of the total of D_H+V_H. Frequencies correlated closely with VDJ:DJ/GL ratios previously published (Ehlich et al., 1994). **B)** XY scatterplot of V_H gene usage in the two WT pro-B VDJ-seq datasets. **C)** XY scatterplot of D_H gene usage in the two WT pro-B VDJ-seq datasets. **D)** DNA FISH analysis of overall V_H-to-DJ_H recombination. A constant region BAC probe labelled with Alexa-555 was used with a cocktail of plasmid probes labelled with Alexa-488 that detect non-repetitive regions in the V_H-D_H intergenic region (VD probe). Detection of these regions in **E)** $Rag1^{-/-}$, and **F)** WT pro-B cells. Absence of VD signals indicates V_H-to-DJ_H recombination has occurred on an allele; presence of these signals, that an allele is either unrecombined (germline) or DJ_H-recombined (hence GLDJ). **G)** Quantitation of alleles in $Rag1^{-/-}$ and WT pro-B cells by DNA FISH. **H)** Quantitation of reading frame in VDJ-seq sequences by IMGT Hi-Vquest analysis. Frequencies of in-frame productive, non-productive and out-of-frame recombination products were close to previous reports (Ehlich et al., 1994). **I)** Analysis of CDR3 length for in-frame VDJ-seq sequences. The dotted line indicates the mean length, circles indicate values for each replicate. Lengths were close to previous reports (Zemlin et al., 2003) **J)** Overlap of amino acid CDR3 IMGT clonotypes between the two WT pro-B VDJ-seq datasets. Only ~5% of clonotypes were shared indicating that each replicate samples a different part of the highly complex, randomly generated pro-B *Igh* repertoire.

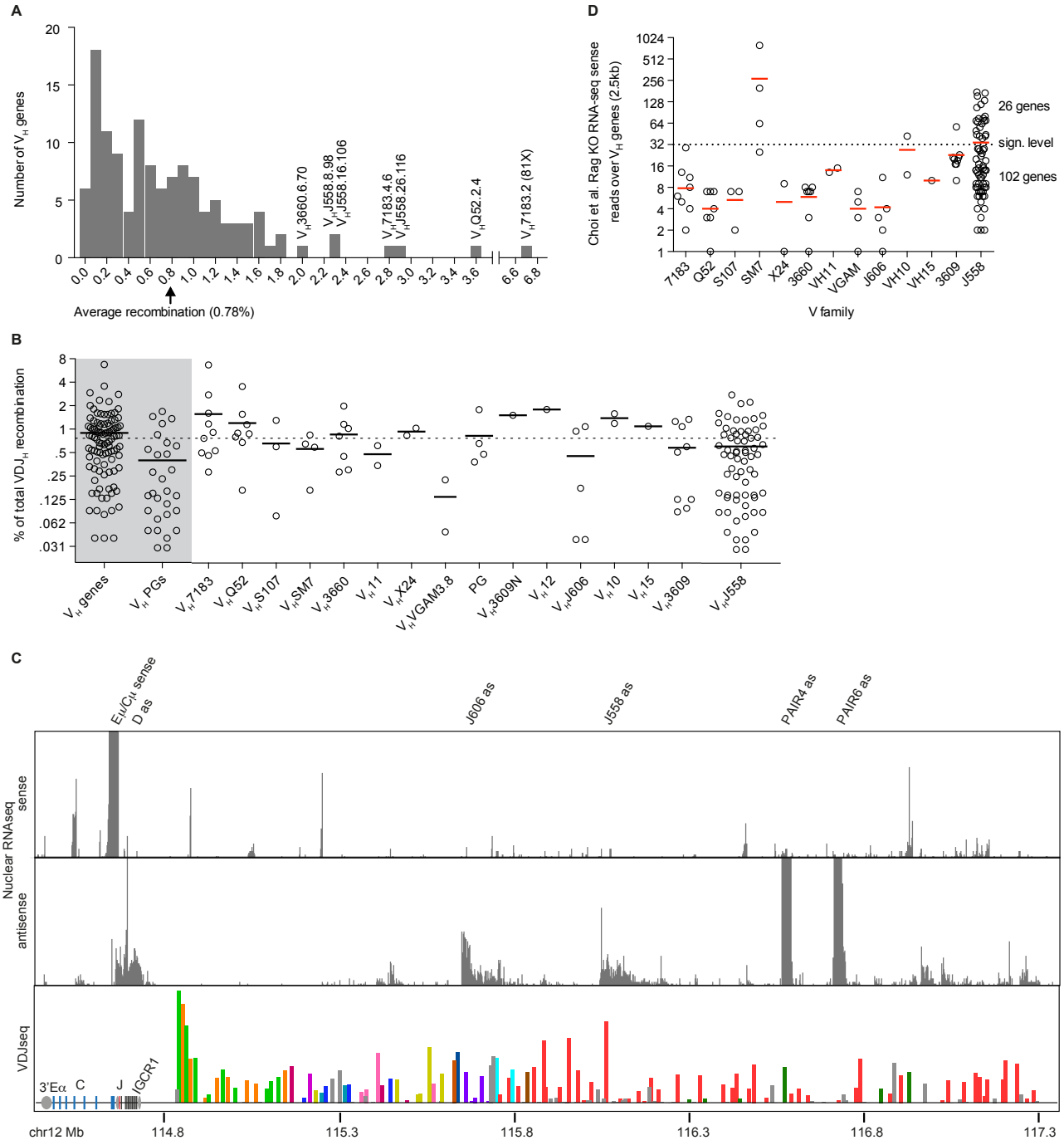


Figure S4 ***Related to Figure 2*** **Detailed VDJ-seq data**

A) Frequency distribution of recombination for V_H genes. The seven highest recombining genes are named. **B)** Gene (99), pseudogene/ORF (29) and per family usage for recombining V_H genes. Each gene is represented by a circle with the mean in each group shown by a line and the average recombination of all recombining genes by a dotted line. Normalised mean read counts of the two replicate datasets was used to calculate the percentages. C) Browser view of nuclear strand-specific RNA-seq, aligned with VDJ-seq dataset. Top: sense; bottom: antisense transcription. **D)** mRNA-seq for sense non-coding transcription within V gene families. Rag-/- (pre-recombination) RNAseq read count over V genes segregated by family. Sense reads were counted in 2.5kb bins centred on the V genes. RNAseq data is from $Rag^{-/-}$ Ovation RNAseq from Choi et al. (Choi et al., 2013) The significance level was calculated by a binomial test similar to that applied to our VDJ-seq data (see methods section). The red line indicates the mean read number in each V family.

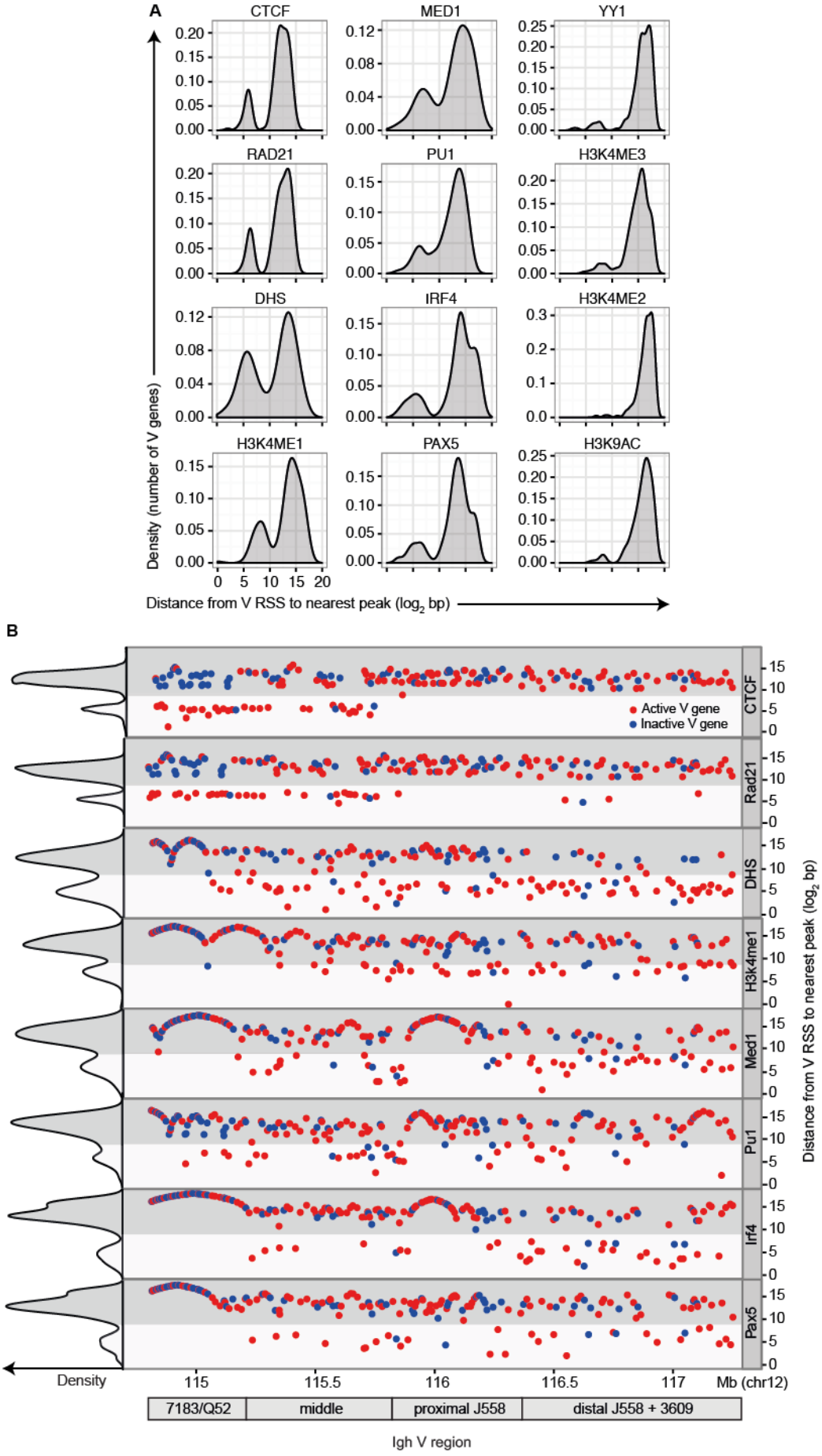
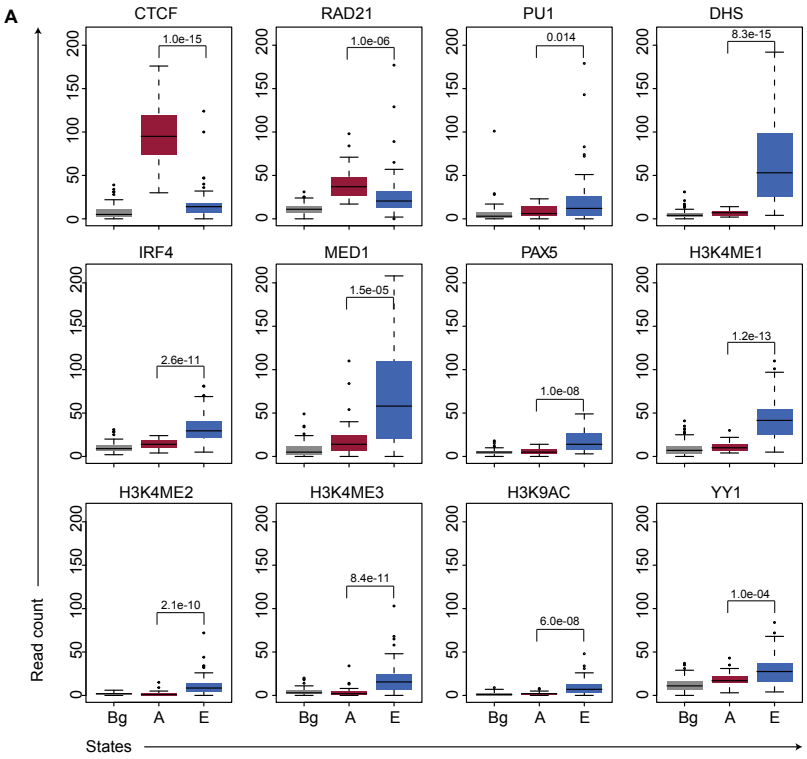
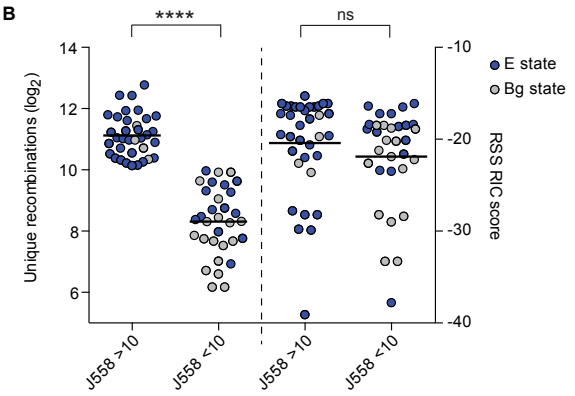
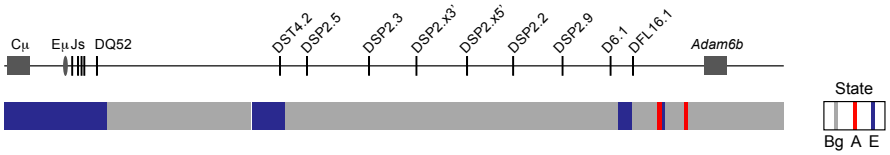


Figure S5 ***Related to Figure 4*** **Co-localisation of epigenetic factors with active V_H genes**
A) Density curves of distances between summit of peaks and the RSSs of V_H genes, for 12 factors within the *Igh* locus (in $\log_2$ bp). The bimodal distribution of distances is clear for 8 of these factors. **B)** Scatter plots show the distance (y-axis in $\log_2$ bp) of nearest peak summit of 8 factors from the RSS of each V_H gene (x-

axis: genomic order of V_H genes). Active (recombining) genes have been color-coded as red and inactive (non-recombining) as blue. Density curves on the left illustrate the frequencies of genes with factor peaks at various distances.


A
CTCF
RAD21
PU1
DHS
IRF4
MED1
PAX5
H3K4ME1
H3K4ME2
H3K4ME3
H3K9AC
YY1
1.0e-15
1.0e-06
0.014
8.3e-15
2.6e-11
1.5e-05
1.0e-08
1.2e-13
2.1e-10
8.4e-11
6.0e-08
1.0e-04
Read count
Bg A E
States


B
****
ns
Unique recombinations (log2)
RSS RIC score
E state
Bg state
J558 >10
J558 <10
J558 >10
J558 <10


C
Cμ
EμJs
DQ52
DST4.2
DSP2.5
DSP2.3
DSP2.x3'
DSP2.x5'
DSP2.2
DSP2.9
D6.1
DFL16.1
Adam6b
State
Bg A E

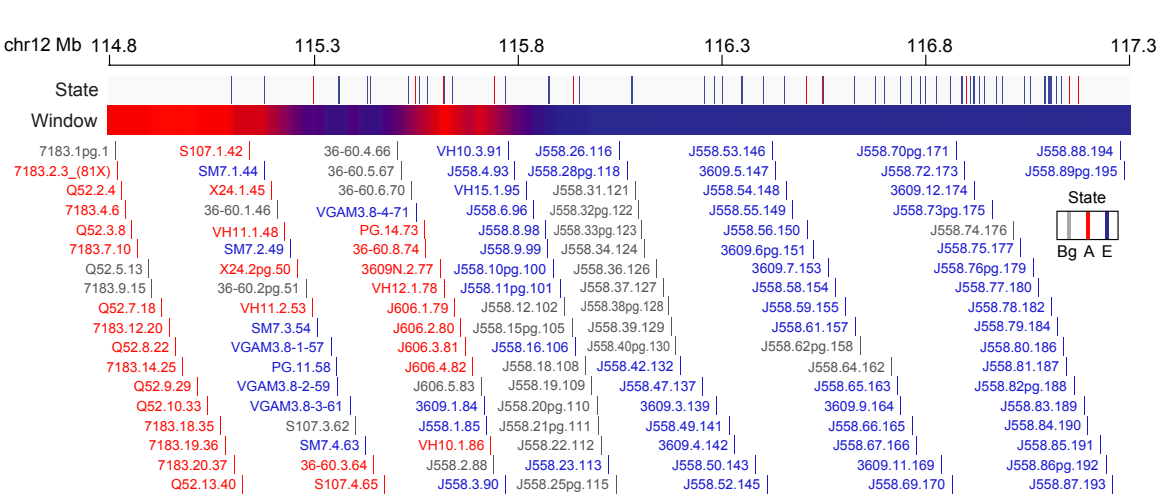

D
chr12 Mb 114.8
115.3
115.8
116.3
116.8
117.3
State
Window
State
Bg A E

Figure S6 ***Related to Figure 5*** **ChIP signals are distinctive features of the 3 chromatin states and distribution of states across the V_H region**

(A) Differences of ChIP enrichment between states for 12 factors used in chromHMM analysis. Consistent with chromHMM, CTCF and RAD21 show a significantly higher signal in the A state than the E state whereas IRF4, MED1, PAX5, YY1, PU.1, H3K4me1, H3K4me2, H3K4me3 and H3K9ac are significantly (t-test) enriched in E state compared to A state. One exception was DHS, which, despite chromHMM association with both active states, showed stronger enrichment of read counts in E state than A state. **B**) Comparison of recombination frequency versus RSS RIC score for highly recombining active J558 genes (>10 $\log_2$ unique reads, 35 genes) versus low recombining active J558 V genes (<10 $\log_2$ unique reads, 32 genes). Left columns: recombination frequency; right columns: RIC scores. E state genes depicted in blue, Bg state genes in grey. An unpaired T-test was used to determine significance. **C)** Local chromatin structure in the J and D regions. J genes entirely overlap with an E-state region. D genes generally overlap with the Bg state. In the chromatin state panel, white segments represent Bg state, red: A state, blue: E state. **D)** Distribution of chromHMM states across the V_H region. From top – position on mouse chromosome 12, individual chromHMM state segments, sliding window of these state segments (a window of 10 segments was used with a step size of 1) and, bottom, actively recombining V_H genes colour-coded by state they overlap with.

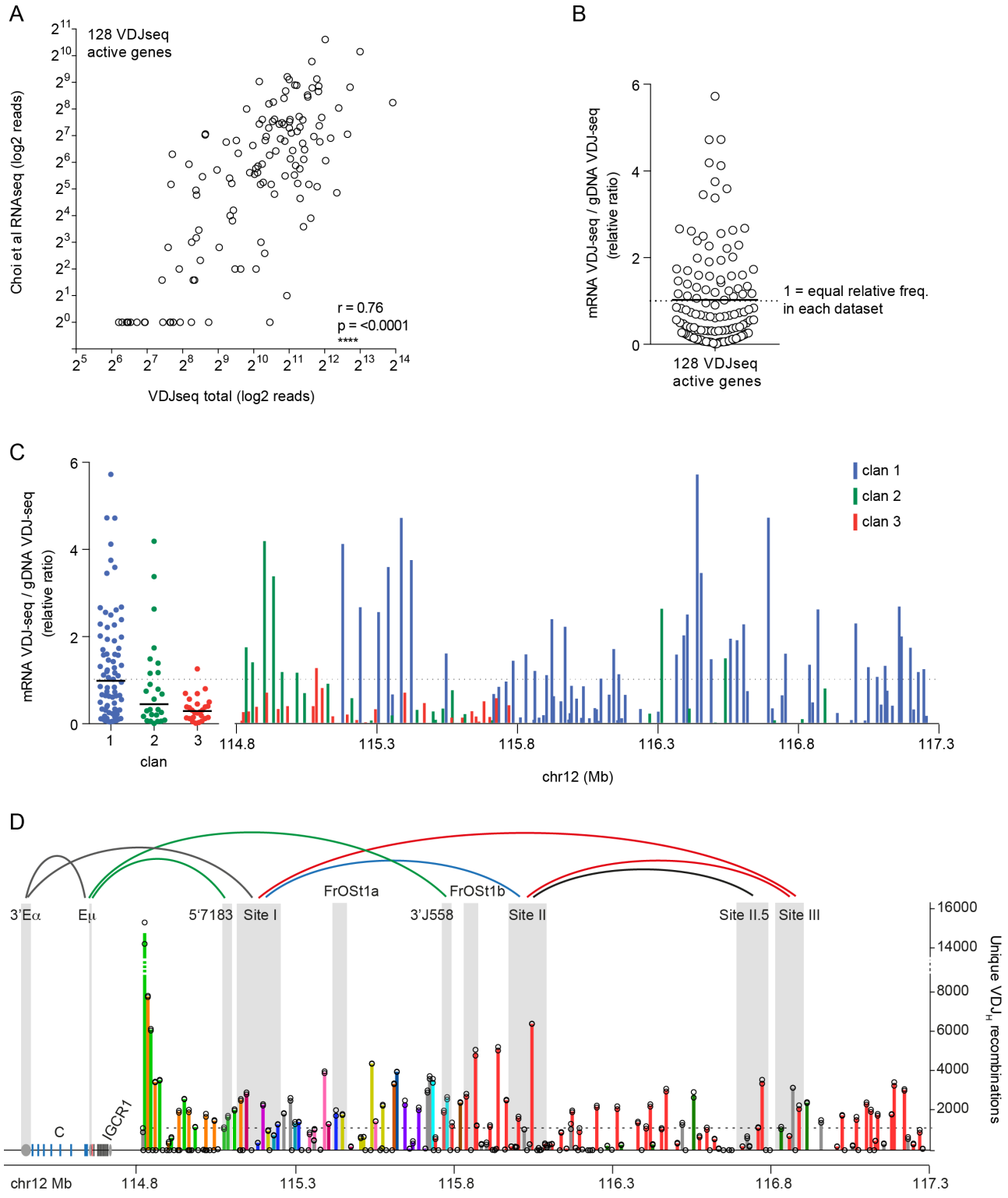


Figure S7 ***Related to Figure 2*** **Comparison of RNA and DNA-based repertoire analyses**
A) XY scatterplot comparing VDJseq with RNAseq data from (Choi et al., 2013) for the 128 V genes found to be active in VDJseq. A pseudocount of 1 was assigned to 14 genes that had zero reads in the Choi et al. dataset. A two-tailed Pearson correlation test was performed using Graphpad Prism. B) Comparison of the relative ratio represented by each individual gene in the two datasets for the 128 active V genes identified in VDJseq. A figure of 1 denotes equal representation in the datasets, above this indicates higher representation in the RNAseq dataset, below, lower representation. C) Data from B) separated into V clans with a dotplot (left) and mapped onto the V locus (right). D) Comparison of recombination frequency with position in topological domains. VDJseq data with the interacting regions identified by Montefiori et al. (Montefiori et al., 2016). Arcs indicate Pax5 dependent (red), independent (blue) and not tested (black). Green arcs, Eμ-dependent loops, grey arc (Guo et al., 2011), 3'Ea loops with Eμ (Kumar et al., 2013) and Site I.

File S1
This Excel spreadsheet includes the following worksheets:
1. V region VDJ-seq data
2. V genes states
3. D region VDJ-seq data
4. Next generation sequencing datasets used, both in-house and published
5. MACS2 peaks and parameters used for chromHMM analysis
6. Oligonucleotide sequences of primers used in VDJ-seq
7. V RSS to MACS peaks: distances of factors from active and inactive V genes

Supplemental Experimental Procedures

Primary cells

C57BL/6 (wild-type; WT) and Rag1$^{-/-}$ mice were maintained in accordance with local and Home Office rules and ARRIVE guidelines under Project Licence 80/2529. For VDJ-seq Rag1$^{-/-}$ (Spanopoulou et al., 1994) bone marrow pro-B cells were isolated with CD19 MACs beads (Miltenyi) achieving >90% purity. For RNA- and ChIP-seq these were further purified by flow sorting (B220^{+}CD19^{+}CD43^{+}). WT bone marrow from 15 12-week old male C57BL/6 mice per replicate was depleted of macrophages, granulocytes, erythroid lineage and T cells using biotinylated antibodies against Cd11b (MAC-1; ebioscience), Ly6G (Gr-1; ebioscience), Ly6C (Abd Serotec), Ter119 (ebioscience) and Cd3e (ebioscience) followed by streptavidin MACs beads (Miltenyi). Thereafter, pro-B cells were flow sorted as IgM^{-}CD25^{-}B220^{+}CD19^{+}CD43^{+} on a BD FACSAria in the Babraham Institute Flow Cytometry facility.

DNA FISH

DNA FISH was performed as previously described (Bolland et al., 2013) using *Igh* constant region BAC RP24-258E20, labelled with Alexa fluor 555, and a set of 7 plasmids containing non-repetitive parts of the V_H-D_H intergenic region (inserts sizes 1-3kb, ~15kb in total), labelled with Alexa fluor 488. Signals were counted manually on an Olympus BX61 epifluorescence microscope system.

Nuclear RNA-seq

Nuclei were obtained from 2-5 x 10^6 flow-sorted Rag1-/- pro-B cells (B220+CD19+CD43+) by incubation in 50 mM Tris-HCl pH 7.5, 140 mM NaCl, 1.5 mM $MgCl_2$, 1 mM DTT, 0.4% NP40, 5 min on ice. RNA was isolated with a RNeasy mini kit (Qiagen) and treated with Turbo DNAse (Ambion). Paired-end strand-specific RNA-seq libraries for Illumina sequencing were generated as described (Parkhomchuk et al., 2009) except polyA+ RNA selection was omitted, first strand cDNA synthesis was performed with random hexamer primers, and double-stranded cDNA was fragmented with a Diagenode Bioruptor. Details in File S1; Accession numbers: GSM2113569, GSM2113570.

ChIP-seq

ChIP was performed as described (Schoenfelder et al., 2010), with an antibody against histone H3K4me3 (Ab8580, Abcam). Cross links were reversed with 100 µg/ml proteinase K overnight at 65°C, and ChIP DNA was purified by PCI extraction and isopropanol precipitation. Paired end ChIP-seq libraries were prepared according to standard Illumina ChIP-seq library protocols. Details in File S1; Accession numbers: GSM2113571, GSM2113572, GSM2113573.

VDJ-seq

Genomic DNA was isolated using a DNeasy kit (Qiagen). For each sample 10µg of DNA were sonicated to 500bp using a Covaris E220 sonicator using recommended settings then end-repaired and A-tailed using standard protocols and a short asymmetric adaptor ligated to both fragment ends. DNA was cleaned with a PCR cleanup kit (Qiagen) after end-repair, following A-tailing and following adaptor ligation. Biotinylated primers located in J_H intergenic regions were then used in primer extension reactions (8 x 50µl) using Vent Exo- polymerase (2 units per tube; NEB) followed by purification with a PCR cleanup kit. Due to the placement of these primers and the fragment size of the sonicated DNA, primer extension of unrecombined J_H segments would be favoured over that of DJ_H or VDJ_H recombined segments. Primer-extended sequences (enriched for unrecombined J_H intergenic sequences) were then removed using streptavidin beads (My-one C1; Invitrogen) following the manufacturers protocol with incubation for 4 hours at room temperature (20µl beads per sample). Following a further cleanup, a second primer extension (6 x 50µl reactions, 2 units per tube) was performed using biotinylated reverse primers located immediately downstream of each J_H gene. Since primer extension is a single cycle, each DJ_H and VDJ_H recombination product will be represented at its relative frequency in the starting DNA.

Streptavidin beads (20µl) were again used to isolate primer extended J_H-specific products by incubation overnight with rotation. Illumina PE1 primers corresponding to the long strand of the asymmetric adaptor and J_H-specific PE2 primers were then used to amplify the library off the beads in 15 cycle PCR reactions (4 x 25µl) using Pwo master mix (Roche). Low-cycle number PCR was used to reduce PCR duplication. Following 1x size selection (to remove small products) and cleanup using AMPure XP beads

(Beckman Coulter), a second round 5 cycle PCR was performed to add the remainder of the Illumina PE1 and PE2 adaptors, incorporating Truseq bar codes at the PE2 end, followed by a second 1x size selection/cleanup with AMpure XP beads. If necessary, a 'double-sided' Ampure XP size selection (0.5x followed by 1x) was used to remove a low quantity of library products >1kb as these are too large for efficient cluster generation in Illumina sequencing. We generated two WT pro-B cell VDJ-seq libraries and one from Rag $CD19^{+}$ cells. Libraries were quality controlled by qPCR analysis of recombined and unrecombined sequences and quantified by Agilent Bioanalyser and Kapa qPCR before being sequenced by Illumina Hiseq 2x100bp or Miseq 2x250bp paired end sequencing. Oligonucleotide sequences are provided in File S1.

VDJ-seq pipeline – Babraham LinkON

We developed a novel pipeline for deduplication of VDJ-seq sequence data we named Babraham LinkON (Figure S1A and B, https://github.com/FelixKrueger/BabrahamLinkON). Briefly, sequences were first adaptor- and low quality trimmed (Phred <20) using TrimGalore (Babraham Bioinformatics), then demultiplexed based on Truseq barcodes. Next, J_H sequences were identified in read 2 (J_H end). By analysing the sequence immediately downstream of each J_H primer sequence in each J_H read we determined that J_H2 PCR primers significantly cross-amplified J_H4 sequences, leading to chimaeric J_H2-J_H4 J reads. We concluded this mis-priming was unavoidable due to sequence similiarity between J_H2 and J_H4 and the requirement to use reverse J_H primers at a set position 10bp distal to the start of the J_H gene to ensure equal capture of all sequences including those that had undergone substantial exonuclease nibbling during D_H-to-J_H joining. To correct this we used the 4bp of *bona fide* J_H2 sequence downstream of the J_H2 primer sequence to reassign the chimaeric sequences to J_H4 and replaced the incorrect J_H2 sequence with the correct J_H4 sequence before further processing. No significant mispriming was seen for the other J_H genes (<5%). Following this, the low number of J_H reads that extended less than 60bp beyond the J_H primer sequence were discarded and for the remainder the 60bp of downstream sequence was scanned for duplicates. These could be either technical (PCR) or biological. In order to differentiate between these, the opposite end reads (read 1, V_H or D_H end) were mapped to the NCBIM37/mm9 mouse genome assembly using Bowtie allowing no mismatches (-n 0 -l 100) and discarding multimapping hits (-m 1 --strata --best) and then scanned for start position in the V_H region. Read pairs that had identical read 2/J_H end sequences (produced by combinatorial joining of V_H, D_H and J_H) and the same start position at the read 1/V_H or D_H end (a product of the random sonication of the DNA) were considered duplicates and only one retained. The previously mapped read 1 (V_H/D_H) reads for these were then output as bowtie mapped BAM files.

Of the ~2.7 x 10^7 starting sequences per library (Figure S1C), ~75% had identifiable J_H sequences in read 2 (identified using J_H primer sequences) with lengths downstream of the J_H longer than 60bp. Of these, ~80% (~60% of total) had mappable read 1 sequences (V_H/D_H). These are the input into the deduplication pipeline since both a J_H sequence longer than 60bp and a mapping V_H/D_H read are pre-requisites (Figure S1B). Following deduplication on J_H sequence plus V_H/D_H read start position, 5.6% of these read pairs were found to be unique for WT pro-B libraries (~2% for $Rag1^{-/-}$). Of these, ~93% (5% of input) mapped to the *Igh* locus (73% for $Rag1^{-/-}$). J_H, D_H and V_H region sequences constituted 23% (1.1%), 49% (2.5%) and 28% (1.4%) of these for WT pro-B libraries whereas >99% of *Igh* reads mapped to the J_H region in $Rag1^{-/-}$ (Figure S1D). Read counts at each stage of this pipeline are shown in Figure S1C.

Quantifying V(D)J recombination

BAM files of V_H/D_H reads corresponding to unique recombination events were loaded into Seqmonk (Babraham Bioinformatics; http://www.bioinformatics.babraham.ac.uk/projects/seqmonk/), a freely available java-based tool to visualise and analyse mapped next generation sequencing data. Seqmonk 'probe' regions (windows over which reads were counted) are listed in Supplementary File 1. For the gene-by-gene analysis of D_H-to-J_H recombination we counted the reads over the D_H gene and upstream region by identifying peaks using Seqmonk and assigning these to specific upstream canonical D_H genes, non-canonical D_H genes, D_H pseudogenes or other recombining regions (e.g. cryptic RSSs). For V_H-to-DJ_H recombination, we determined that all correct (i.e. reverse) orientation reads mapped to a region encompassing the V_H exon plus 800bp of upstream sequence (1.1kb overall), as expected from the sonication of genomic DNA to 500bp (range of DNA fragments 0.2-1kb). We thus counted reads within these regions and normalised to the replicate with the lowest number of reads (replicate 1). Mean read counts were then calculated from these values for each D_H or V_H gene. Raw read counts for V_H and D_H genes are provided in Supplementary File 1. We corrected two errors in gene functionality designations and reclassified six genes as ORF (open reading frame) pseudogenes due to missing key residues required for pre-BCR pairing. For J_H-J_H, inverted D_H and V_HRSS-DJ_H recombination we counted inverted (i.e. forward) reads in regions including the RSS and extending downstream of each V_H, D_H and J_H gene.

VDJ-seq IMGT HighV-quest pipeline

The 100bp VDJ-seq J_H-end reads for VDJ recombination events contain part of the J_H gene, the D_H-J_H junction, the D_H gene, the V_H-D_H junction, and a variable amount of V_H gene (typically 40-50bp). Paired V_H-end reads map to a region starting from 0bp up to ~1kb proximal to the V_H gene RSS. IMGT HighV-quest requires sufficient V_H sequence to be able to unambiguously identify a specific V_H gene, which is generally greater than the 40-50bp of V_H gene sequence in the J_H end read produced by 100bp sequencing. Therefore,

we developed a custom script that takes the V_H-J_H read pair output of the standard VDJ-seq pipeline detailed above and, if the V_H and J_H reads directly abut or overlap (21.5% of read pairs), merges the two, keeping the J_H-end sequence for any overlapping regions to maintain the highly-variable junction sequence. For read pairs that didn't overlap (78.5%), the genomic V_H gene sequence was used to fill in the gap, again ensuring the J_H read wasn't changed in order to maintain the junction sequences. Using this approach, just <0.7% of sequence pairs overall for each replicate were lost, indicating high joining efficiency. Furthermore, Pearson correlation analysis of V_H gene read counts with the standard pipeline gave R^2 values of ~1 indicating that the data had not been altered by the process. The reverse-complemented joined sequences where then analysed by IMGT HighV-quest using standard parameters (see below).

IMGT HighV-quest analysis

We performed IMGT HighV-quest analysis (Alamyar et al., 2012) with the "F+ORF+in-frame P" reference directory set selecting "with allele *01 only", since the sequences were from a single mouse strain. >99% of sequences in each dataset could be assigned as "productive" or "unproductive" with very few "no result" or "unknown" assignments. We derived in-frame productive, in-frame non-productive and out-of-frame data from the IMGT HighV-quest 'Summary' file. We then performed standard IMGT HighV-quest Statistical Analysis to determine CDR3 lengths (for productive plus non-productive recombination events) and IMGT AA clonotype data (for productive only) to determine overlaps between the two WT pro-B replicate datasets.

Definition of recombining versus non-recombining V_H genes

Recombining active V_H genes were defined as those enriched for VDJ-seq reads compared with the V_H region as a whole. These were ascertained using a binomial test of observed versus expected by random read counts per V_H gene (fdr-adjusted p-value <0.01), in which the probability is defined as the fraction of the total locus length taken by the gene, *n* as the read counts of a given V_H gene body and *N* as the total number of VDJ-seq reads in the V_H region.

Since the two replicate WT pro-B VDJ-seq datasets were highly correlated, for all computational analyses below we elected to use only replicate 2.

Random forest classification

Random forest (RF) classification was chosen as the machine learning method as it is known to cope well with both colinearity and interactions between predictors. Genes with mappability (defined as the average mappability score over a given genomic region) below 90% (29 V_H genes) were excluded from this analysis. The binary recombination classes ("recombining" vs "non-recombining") defined as described above, and designated as active and inactive respectively, were used as response variables. The 29 non-mappable V_H genes were evenly divided between these classes.

DHS-seq, ChIP data and RNA-seq data processed as follows were used as predictors. From each dataset, reads overlapping 2.5-kb windows surrounding each V_H gene were retained for analysis (bedtools coverage function), and the total read counts for each dataset over the whole V_H region were noted. Signal intensities were defined as $\log_2(O+1/E+1)$, where O is observed read counts at each gene and E is the expected counts given the fraction of the locus length taken by the gene. To account for different signal peak shapes in these data (e.g., broad histone peaks versus sharp TF peaks), we split the 2.5kb regions into 500bp regions containing the gene bodies and 1kb upstream and downstream regions. For each factor, we used either the signal intensity for the total 2.5kb region or that for one of the three subregions, depending on which of these showed the highest correlation with VDJ-seq read counts.

RF classification was performed with 10-fold cross-validation, using 10% of genes as the test set in each case, and average variable importance was recorded. We then focused on the top 11 predictors identified this way (RSS, DHS, H3K4me1, CTCF, IRF4, H3K4me3, RAD21, H3K9ac, PAX5, MED1 and sense RNA) and examined the classification rate of all their possible combinations, using 10-fold cross-validation for each combination. Various metrics of classification rate were explored, including Area Under Curve (AUC); F1 scores defined as 2TP/(2TP+FP+FN), where TP, FP an FN are the numbers of true positives, false positives and false negatives, respectively; and accuracy defined as (TP+TN)/(P+N), where P and N are the total numbers of positives and negatives, respectively. All of these metrics produced similar results (data not shown). Analysis was performed using R packages randomForest (Liaw and Wiener, 2002), pROC (Robin et al., 2011) and caret (R package version 6.0-41).

Co-localisation analysis

ChIP peaks (including DHS) were called using MACS2 (in the narrow peak mode for all data except H3K27me3, see Supplementary File S1 for the number of detected peaks and parameters used in the *Igh* locus). For each V_H gene, the distance between the RSS position to the summit of nearest peak was measured for each ChIP dataset. The significance of co-localisation (vs. active and inactive genes), was assessed between the ChIP peaks and 1kb windows surrounding the RSS using a c^2 test. See File S1 for the number of detected peaks and parameters used in the *Igh* locus.

Chromatin segmentation

For chromatin segmentation the *Igh* locus was split into 200bp bins and for each ChIP and DHS

dataset, a value of either 0 or 1 was assigned to each bin depending on whether it overlapped with the respective ChIP/DHS peak (>50%). The resulting binary matrix was submitted to chromHMM. Segmentations with the following numbers of states were obtained: 2, 3, 4, 5, 6, 10, 15, 20, 25, 30, 35, and the model with 3 states (designated "Background" (Bg), "Architectural" (A) and "Enhancer" (E)) appeared to be near optimal with high between classes difference and low inter-class variation. The significance of association between three chromatin states and V genes' recombination classes (active vs. inactive) was assessed by a Fisher exact test.

Data sources availability

Public ChIP- and DHS-seq datasets were downloaded from GEO in the form of raw short-read files (SRA; see Supplementary File 1 for locations) and realigned to NCBIM37/mm9 using bowtie with the parameters listed in Supplementary File 1. The VDJ-seq, ChIP-seq and RNA-seq data generated in this study can be found at GEO Accession Number GSE80155.

Supplemental References

Alamyar, E., Giudicelli, V., Li, S., Duroux, P., Lefranc, M.-P., 2012. IMGT/HighV-QUEST: the IMGT® web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing. Immunome Res 8, 26.

Bolland, D.J., King, M.R., Reik, W., Corcoran, A.E., Krueger, C., 2013. Robust 3D DNA FISH using directly labeled probes. JoVE. doi:10.3791/50587

Choi, N.M., Loguercio, S., Verma-Gaur, J., Degner, S.C., Torkamani, A., Su, A.I., Oltz, E.M., Artyomov, M., Feeney, A.J., 2013. Deep sequencing of the murine IgH repertoire reveals complex regulation of nonrandom V gene rearrangement frequencies. The Journal of Immunology 191, 2393–2402. doi:10.4049/jimmunol.1301279

Ehlich, A., Martin, V., Muller, W., Rajewsky, K., 1994. Analysis of the B-cell progenitor compartment at the level of single cells. Curr Biol 4, 573–583.

Guo, C., Ivanova, I., Chakraborty, T., Oltz, E.M., Sen, R., 2011. Two forms of loops generate the chromatin conformation of the immunoglobulin heavy-chain gene locus. Cell 147, 332–343. doi:10.1016/j.cell.2011.08.049

Kumar, S., Wuerffel, R., Achour, I., Lajoie, B., Sen, R., Dekker, J., Feeney, A.J., Kenter, A.L., 2013. Flexible ordering of antibody class switch and V(D)J joining during B-cell ontogeny. Genes Dev 27, 2439–2444. doi:10.1101/gad.227165.113

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R news 2, 18–22.

Montefiori, L., Wuerffel, R., Roqueiro, D., Lajoie, B., Guo, C., Gerasimova, T., De, S., Wood, W., Becker, K.G., Dekker, J., Liang, J., Sen, R., Kenter, A.L., 2016. Extremely Long-Range Chromatin Loops Link Topological Domains to Facilitate a Diverse Antibody Repertoire. CellReports 14, 896–906. doi:10.1016/j.celrep.2015.12.083

Parkhomchuk, D., Borodina, T., Amstislavskiy, V., Banaru, M., Hallen, L., Krobitsch, S., Lehrach, H., Soldatov, A., 2009. Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Research 37, e123–e123. doi:10.1093/nar/gkp596

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., Müller, M., 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12, 77. doi:10.1186/1471-2105-12-77

Schoenfelder, S., Sexton, T., Chakalova, L., Cope, N.F., Horton, A., Andrews, S., Kurukuti, S., Mitchell, J.A., Umlauf, D., Dimitrova, D.S., Eskiw, C.H., Luo, Y., Wei, C.-L., Ruan, Y., Bieker, J.J., Fraser, P., 2010. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. Nat Genet 42, 53–61. doi:10.1038/ng.496

Spanopoulou, E., Roman, C.A., Corcoran, L.M., Schlissel, M.S., Silver, D.P., Nemazee, D., Nussenzweig, M.C., Shinton, S.A., Hardy, R.R., Baltimore, D., 1994. Functional immunoglobulin transgenes guide ordered B-cell differentiation in Rag-1-deficient mice. Genes Dev 8, 1030–1042.

Zemlin, M., Klinger, M., Link, J., Zemlin, C., Bauer, K., Engler, J.A., Schroeder, H.W., Jr, Kirkham, P.M., 2003. Expressed Murine and Human CDR-H3 Intervals of Equal Length Exhibit Distinct Repertoires that Differ in their Amino Acid Composition and Predicted Range of Structures. J. Mol. Biol. 334, 733–749. doi:10.1016/j.jmb.2003.10.007