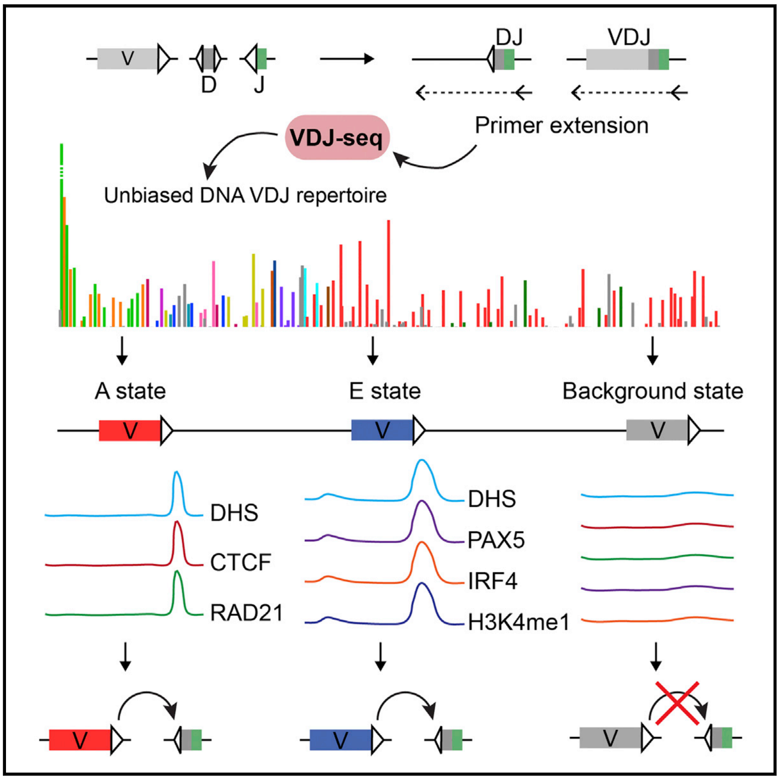
Cell Reports

Article

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

Graphical Abstract



Authors

Daniel J. Bolland, Hashem Koohy, Andrew L. Wood, ..., Simon R. Andrews, Mikhail Spivakov, Anne E. Corcoran

Correspondence

mikhail.spivakov@babraham.ac.uk (M.S.),
anne.corcoran@babraham.ac.uk (A.E.C.)

In Brief

Bolland et al. develop a technique to quantitatively profile antigen receptor diversity. Using VDJ-seq in the mouse *Igh* locus, they uncover the regulatory logic underlying the highly varying recombination rates of V gene segments, with implications for immune disorders and aberrant recombination in cancer.

Highlights

- VDJ-seq enables precise quantification of antibody V(D)J recombination products
- Two distinct *cis*-regulatory designs characterize actively recombining V genes
- Putative recombination regulatory elements map downstream of mouse *Igh* V genes
- Recombination regulatory architecture reflects the V genes' evolutionary history

Accession Numbers

GSE80155

Bolland et al., 2016, Cell Reports *15*, 2475–2487
June 14, 2016 © 2016 The Authors
http://dx.doi.org/10.1016/j.celrep.2016.05.020

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

Daniel J. Bolland,[1,3] Hashem Koohy,[1,3] Andrew L. Wood,[1] Louise S. Matheson,[1] Felix Krueger,[2] Michael J.T. Stubbington,[1] Amanda Baizan-Edge,[1] Peter Chovanec,[1] Bryony A. Stubbs,[1] Kristina Tabbada,[1] Simon R. Andrews,[2] Mikhail Spivakov,[1,*] and Anne E. Corcoran[1,*]

[1]Nuclear Dynamics Programme, Babraham Institute, Babraham Research Campus, Cambridge CB22 3AT, UK
[2]Bioinformatics Group, Babraham Institute, Babraham Research Campus, Cambridge CB22 3AT, UK
[3]Co-first author
*Correspondence: mikhail.spivakov@babraham.ac.uk (M.S.), anne.corcoran@babraham.ac.uk (A.E.C.)
http://dx.doi.org/10.1016/j.celrep.2016.05.020

SUMMARY

Variable (V), diversity (D), and joining (J) (V(D)J) recombination is the first determinant of antigen receptor diversity. Understanding how recombination is regulated requires a comprehensive, unbiased readout of V gene usage. We have developed VDJ sequencing (VDJ-seq), a DNA-based next-generation-sequencing technique that quantitatively profiles recombination products. We reveal a 200-fold range of recombination efficiency among recombining V genes in the primary mouse *Igh* repertoire. We used machine learning to integrate these data with local chromatin profiles to identify combinatorial patterns of epigenetic features that associate with active V_H gene recombination. These features localize downstream of V_H genes and are excised by recombination, revealing a class of *cis*-regulatory element that governs recombination, distinct from expression. We detect two mutually exclusive chromatin signatures at these elements, characterized by CTCF/RAD21 and PAX5/IRF4, which segregate with the evolutionary history of associated V_H genes. Thus, local chromatin signatures downstream of V_H genes provide an essential layer of regulation that determines recombination efficiency.

INTRODUCTION

Variable (V), diversity (D) and joining (J) (V(D)J) recombination of antigen receptor (AgR) loci is the first step in generating the diverse AgR repertoires that enable the adaptive immune system to respond to a vast array of pathogens and regulate tissue homeostasis and surveillance. This process, which occurs at the immunoglobulin (Ig) loci in progenitor B cells and the T cell receptor (TCR) loci in progenitor T cells, involves the sequence-specific cutting and joining of VDJ gene segments to form a functional immunoglobulin (BCR) or TCR gene (Corcoran, 2010; Schatz and Ji, 2011). Failure to generate sufficiently diverse repertoires underpins a wide variety of immunodeficiency diseases and poor immune function in aging (Dunn-Walters and Ademokun, 2010), while inappropriate recombination targeting can lead to genome instability (Teng et al., 2015) and chromosomal translocations in T and B cell leukemias (Marculescu et al., 2002).

The mouse immunoglobulin heavy chain (*Igh*) locus encompasses 2.8 Mb of chromosome 12 and contains 4 J_H genes, 10 D_H genes, and 195 V_H genes (Johnston et al., 2006; Ye, 2004). The V_H genes have been classified into 16 families in three clans, based on sequence similarity and are organized in distinct domains within the V_H region (Johnston et al., 2006). The in-frame joining of a V_H to a DJ_H segment to complete the sequence of an IgH polypeptide is the critical event underpinning commitment to the B lineage, since expression of an IgH protein in the pre-B cell receptor is required to switch off *Igh* recombination (allelic exclusion) and enable progression. Many genetic and epigenetic features have been implicated in the regulation of V_H recombination. These include the quality of the downstream Rag recombinase binding sites (recombination signal sequences [RSSs]), sense, and antisense non-coding transcription (Bolland et al., 2004; Yancopoulos and Alt, 1985), modified histones, locus compaction, and chromatin looping (Fuxa et al., 2004; Jhunjhunwala et al., 2008; Sayegh et al., 2005; Stubbington and Corcoran, 2013). Binding of transcription factors including PAX5, YY1, IKAROS, and CTCF, is critical for locus compaction and looping. CTCF promotes local looping in the distal V region, while Pax5 promotes longer-range movement of local distal V domains toward the DJ domain (Degner et al., 2011; Fuxa et al., 2004; Gerasimova et al., 2015; Guo et al., 2011; Liu et al., 2007; Medvedovic et al., 2013; Montefiori et al., 2016; Reynaud et al., 2008). Under the current model, the combined action of these transcription factors brings all V genes into close proximity with the DJ segment, providing equal spatial opportunity for all to participate in V(D)J recombination (Schatz and Ji, 2011).

Despite these advances, it still remains unclear why recombination frequencies of individual V_H genes vary enormously (Buchanan et al., 1997; Love et al., 2000; Perlmutter et al., 1985; Yancopoulos et al., 1988). A significant hurdle has been the absence of a comprehensive and quantitative profile of the recombination frequencies of all 195 V_H genes. Real-time PCR-based approaches using cocktails of V_H primers (Rouaud et al., 2012) and their adaptation for deep sequencing (Georgiou et al., 2014) have increased the throughput of these but remain prone to bias associated with differential primer efficiency and

© 2016 The Authors

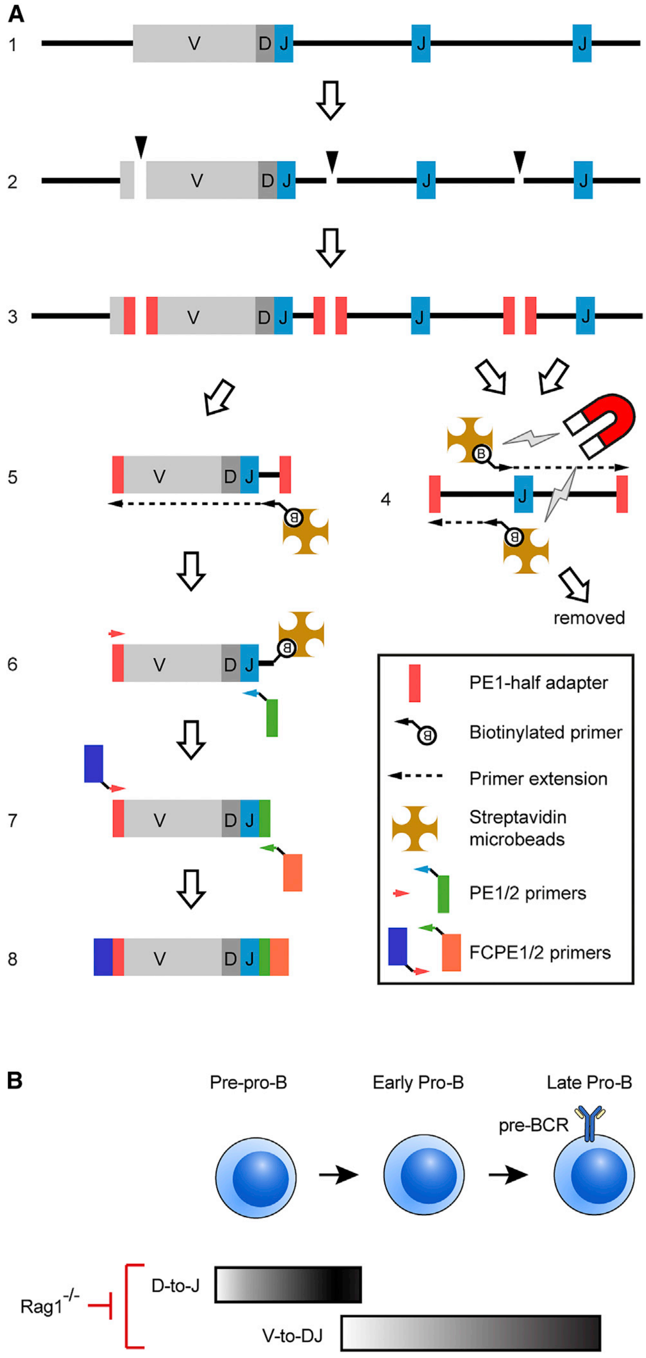


Figure 1. The VDJ-Seq Technique

(A) Genomic DNA from sorted pro-B cells (1) containing unknown VDJ_H and DJ_H joins (only VDJ_H depicted) is sonicated to 500 bp (2), end-repaired, A-tailed, and a custom adaptor ligated (3). Primer-extension is performed with forward and reverse primers that hybridize upstream of each J_H gene (4). Following depletion of unrecombined primer-extended DNA with streptavidin beads (4), a second primer-extension is performed extending upstream into VDJ_H or DJ_H recombined sequences from biotinylated primers that hybridize downstream of each J_H (5). After capture with streptavidin beads, two rounds of PCR generate the sequencing library; the first using adaptor-specific paired-end 1 (PE1) and J-specific paired-end 2 (PE2) primers (6), the second using flow-cell PE1 and PE2 primers (7) to generate the library (8).

(B) Differentiation of early B cell progenitors in bone marrow showing when D_H-to-J_H and V_H-to-DJ_H joining occur. $Rag1^{-/-}$ mice are incapable of V(D)J recombination.

See also Figures S1 and S3.

are often incomplete (Kaplinsky et al., 2014). These biases have been mitigated by deep sequencing of the mRNA output of VDJ_H-recombined products (Choi et al., 2013), but this approach captures only productive recombination events (Eberle et al., 2009) and does not account for varying V_H gene promoter activity (Buchanan et al., 1997; Love et al., 2000). In the smaller T cell receptor β locus, these challenges have recently been addressed using DNA-based deep-sequencing, thereby revealing the contributions of key chromatin features to recombination efficiency (Gopalakrishnan et al., 2013), but the methodology used in that study is difficult to adapt to larger AgR loci. Thus, quantitative immunoglobulin repertoire analysis at the DNA level, at which V(D)J recombination occurs, has not been achieved (Benichou et al., 2012).

To address this deficit, we developed VDJ sequencing (VDJ-seq), a quantitative, high-throughput next-generation sequencing assay based on the capture and sequencing of primer extension products of genomic DNA from J_H gene oligonucleotides. As each of the four J_H genes recombines with the entire spectrum of D_H and V_H genes, this circumvents the use of multiple V gene primers and enables unbiased detection of DJ_H and VDJ_H recombination products. We quantify the primary output of *Igh* V(D)J recombination by applying VDJ-seq to mouse pro-B cells, in which recombination is ongoing and not yet significantly skewed by downstream processes. We integrate these data with profiles of expression, transcription factor binding, and the chromatin state to reveal two mutually exclusive *cis*-regulatory signatures at active V genes that localize to the downstream RSS-proximal sequences. These findings establish a paradigm for the regulation of V(D)J recombination that may be widely applicable to other AgR loci.

RESULTS

The VDJ-Seq Technique

VDJ-seq exploits the fact that every DJ_H and VDJ_H recombination event ends with one of only four J_H genes and is based on two sequential primer extension and capture steps using biotinylated J_H region oligos on sonicated, adaptor-ligated genomic DNA. The first step depletes unrecombined sequences located upstream of each J_H, the second captures DJ_H and VDJ_H recombined sequences using J_H-specific oligos (Figure 1A; Data S1). Captured J_H primer-extension products are then PCR-amplified using primers to the V_H-end adaptor sequence together with nested J_H oligos. The use of only J_H primers for both primer extension and PCR steps enables unbiased detection of VDJ_H/DJ_H recombined sequences since V_H gene primers with their inherent biases are not used at any stage.

We generated two biological replicate libraries from ex vivo flow-sorted wild-type (WT) bone marrow pro-B cells ($B220^+$ $CD19^+CD43^+CD25^-sIgM^-$) and one from $CD19^+$ $Rag1^{-/-}$ bone marrow pro-B cells as a negative control. WT pro-B cells have almost completed D_H-to-J_H recombination (Ehlich et al., 1994; Rumfelt et al., 2006) and are undergoing V_H-to-DJ_H joining, while Rag1$^{-/-}$ cells cannot recombine, so any non-J_H region sequences represent background (Figure 1B). V(D)J_H recombination generates variability due to combinatorial joining of different V_H, D_H, and J_H gene segments and junctional diversity from nucleotide additions and exonuclease nibbling (Figure S1A). We used this variability in the J_H read plus the positional variability of the V_H/D_H read start locations (from random DNA shearing) to develop a deduplication pipeline to identify unique sequences (Figures S1A and S1B; Supplemental Experimental Procedures). Following deduplication, ~99% of reads detected in Rag1$^{-/-}$ mapped to unrecombined J_H sequences (Figure S1C). In contrast, in WT cells, unrecombined J_H sequences constituted ~20% of reads, while ~50% of reads mapped to the D_H cluster and 25%–30% to the V_H region (Figures S1C and S1D). These corresponded to 382,932 and 441,640 DJ_H joins with 220,363 and 239,090 VDJ_H recombinants for the two wild-type replicates (Figure S1C), equivalent to 20% detection rate (1.7 million cell equivalents/replicate; 3.4 million alleles; 35% VDJ_H recombined alleles; see Figure S3).

In addition to these canonical DJ_H and VDJ_H joins, we also detected a variety of aberrant products (Figure S2), some of which have been reported previously (Fang et al., 1996; Hu et al., 2015; Sollbach and Wu, 1995). These include the joining of adjacent J_H segments, inverted D_H-to-J_H joining, D_H region cryptic recombination, and V_H signal sequences joined to DJ_H segments. While these products were detected at low frequencies (<1%), they were not observed in Rag1$^{-/-}$, indicating they are genuine products of V(D)J recombination.

We performed extensive quality control of VDJ-seq. Wild-type replicates were highly correlated for both VDJ_H and DJ_H recombination (Figures S3A–S3C). Frequencies of DJ_H and VDJ_H recombination correlated closely with published VDJ:DJ/GL ratios and with a DNA fluorescence in situ hybridization (FISH) assay of V_H-to-DJ_H joining developed here (Figures S3D–S3G; Supplemental Experimental Procedures). Frequencies of in-frame productive, non-productive, and out-of-frame recombination products analyzed in IMGT HighV-quest (Alamyar et al., 2012) (Figure S3H; Supplemental Experimental Procedures) were close to previous reports, as were complementarity determining region 3 (CDR3) lengths (Figure S3I). Notably, each replicate samples a different part of the highly complex, randomly generated pro-B *Igh* repertoire (Figure S3J). At the same time, close correlation of V_H and D_H recombination frequencies within families between replicates indicates that differential gene usage is robust across the entire repertoire (Figures S3B and S3C), and sequencing depth is sufficient for quantitative analysis.

VDJ-Seq Reveals Complete Primary DJ_H and VDJ_H Repertoires

For D-to-J_H recombination, >98% of the sequences mapping to the D_H region originated from the ten canonical C57BL6 D_H genes, with D_HFL16.1 recombining with the highest frequency, consistent with previous studies (reviewed in Ye, 2004) (Figure 2A).

For V-to-DJ_H recombination, we detected wide variation in usage of V genes across the *Igh* locus (Figure 2B). To determine which genes were actively undergoing recombination, we used a binomial test (Figure 3A; Supplemental Experimental Procedures). Using a stringent threshold (fdr-adjusted p value <0.01) threshold, we detected 128 recombining and 67 recombinationally silent V_H genes and pseudogenes (Data S1) out of the 195 total (Johnston et al., 2006). Recombining V_H genes and pseudogenes and recombinationally silent V_H genes and pseudogenes are referred to hereafter as "active" and "inactive," respectively. There was a 200-fold range of recombination efficiencies among the 128 active V_H genes (Figures 2B and S4A; Data S1).

As expected, the majority of functional (protein-coding, pre-BCR pairing) genes (99/103) were active. Notably, 29/92 pseudogenes also recombined, albeit at lower frequencies. All of these had RSSs predicted as recombinogenic for *Igh* V gene RSSs (recombination information content [RIC] score >-58.45) (Cowell et al., 2002; Lee et al., 2003). Separation of the 128 active V_H genes and pseudogenes into the 15 V_H gene families plus one pseudogene class revealed large differences in recombination frequencies both across and within the families (Figure S4B). Thus, with VDJ-seq, we have quantified the full range of V_H genes and pseudogenes and identified those that do and do not participate in *Igh* V(D)J recombination. The wide range of recombination frequencies suggests that complex regulatory mechanisms are at play.

Factors Predictive of Active V_H Gene Recombination

We set out to identify the genetic and epigenetic features that associate with active V_H recombination. We integrated our VDJ-seq data with the profiles of transcription factor binding and histone modifications and the quality of the V genes' RSSs. To determine the individual and combinatorial roles of these factors in V to DJ recombination, we used them as predictors in a Random Forest (RF) classifier trained to distinguish active from inactive V_H genes (Supplemental Experimental Procedures). We first trained a RF model using 18 features as predictors including RSS RIC score, chromatin immunoprecipitation sequencing (ChIP-seq) signals for histone marks and transcription factors (TF), and sense and antisense RNA levels (Data S1). Features were assessed in 2.5 kb windows, spanning the V_H gene itself, including 1 kb upstream to incorporate the promoter and 1 kb downstream, including the RSS. Classification based on these 18 factors showed a very high predictive power (Figures 3A–3C), indicating that they effectively distinguish active from inactive V_H genes. Significant predictors of active recombination included a high RIC score, DNase hypersensitivity (DHS), chromatin marks associated with active chromatin states (H3K4me3 and H3K4me1), and several architectural and transcription factors. Some of these (CTCF, RAD21, PAX5, and YY1) are known to be required for *Igh* recombination (Degner et al., 2011; Fuxa et al., 2004; Liu et al., 2007), while others (PU.1, P300, MED1, and IRF4) have not been previously reported to function in this capacity. In contrast, local sense germline transcription, long implicated as a requirement for *Igh* recombination (Bolland et al., 2004; Yancopoulos and Alt, 1985), had a weak predictive power in our analysis (Figure 3B). Our strand-specific nuclear RNA sequencing (RNA-seq) dataset aimed to enrich for non-coding transcripts, but sense V gene transcripts were infrequent (Figure S4C). Antisense transcription was also not a robust

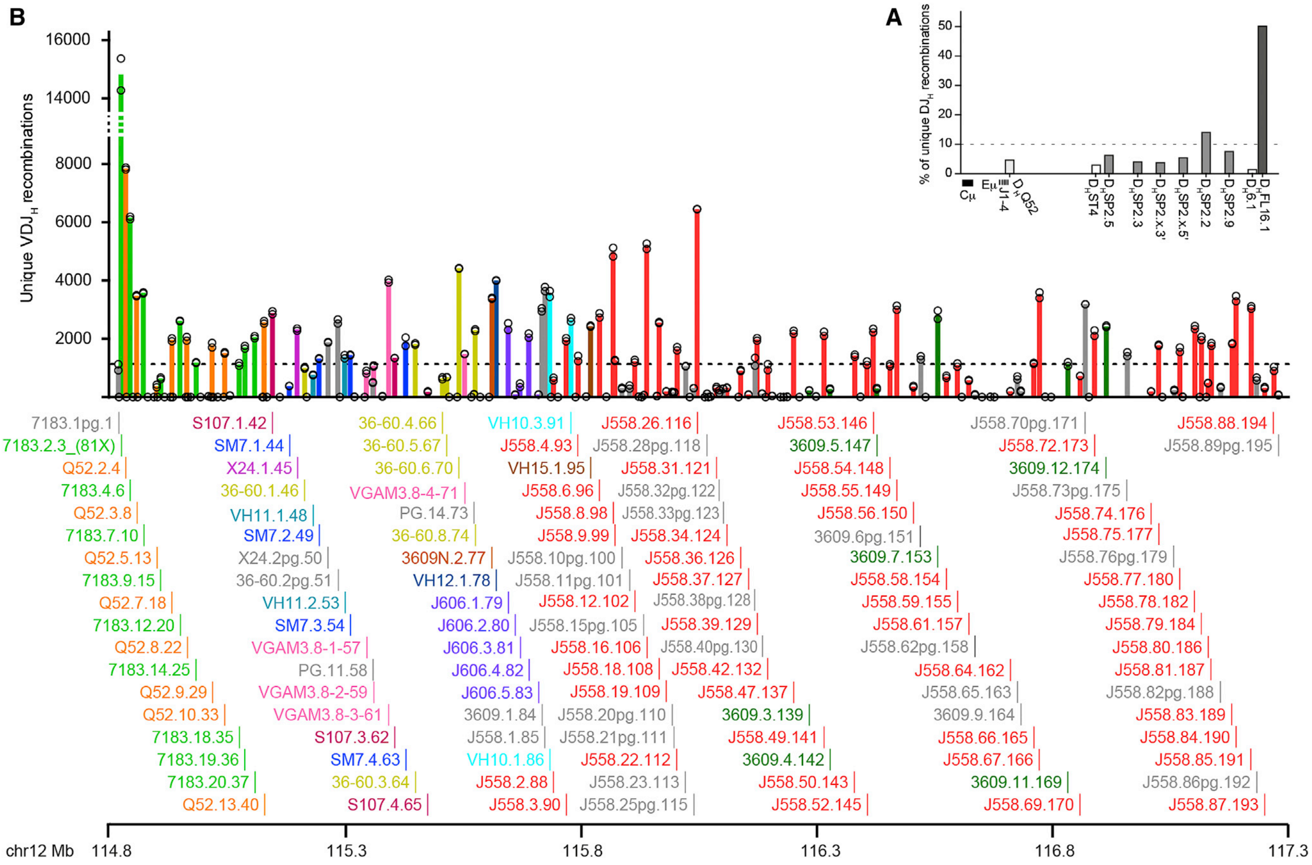


Figure 2. VDJ-Seq Reveals Widely Varying D_H and V_H Gene Recombination

(A) D_H gene recombination. Reads per D_H gene were counted using Seqmonk for each WT pro-B replicate, normalized to the replicate with the lowest total read count for all D_H genes then expressed as a percentage of the total (Data S1). Bars indicate the mean of these values. Replicate values are not shown as D_H usage was almost identical between the two replicates. Dotted line, expected value if recombination was equal for all ten D_H genes.

(B) Recombination frequencies of the 195 V_H genes. Reads for each V_H gene were counted for each replicate then normalized to the replicate with the lowest total read count for all V_H genes (Data S1) and are shown as open circles. Bars indicate the mean of these values for each V_H gene, colored by family. Dotted line, expected frequency if recombination was equal for all V_H genes. Individual gene names for actively recombining V_H genes (determined by binomial testing) are shown below together with map position on mouse chromosome 12.

See also Figures S2, S4, and S7.

predictor, which was expected as it localizes to a small number of discrete domains across the locus (Figure S4C). A previous study (Choi et al., 2013) had also failed to find a strong correlation between transcription and recombination, both according to the authors and in our reanalysis (Figure S4D). Thus, while the importance of transcription cannot be excluded, the available data do not support a predictive role in active recombination.

We then evaluated the classification rate of all possible combinations of the top 11 predictors reported by the RF classifier (2048 RF models; Figure 3C). In all cases, the classification rate differed considerably depending on whether or not the RIC score was included, indicating a strong association of the Rag-binding RSS sequence with active recombination, consistent with previous findings (Choi et al., 2013). The highest classification rates (>95%) were obtained when RIC scores together with DHS, H3K4me1, CTCF, and RAD21 were used as predictors. RSS quality was nevertheless not a sufficient predictor, since 34% of inactive genes had RIC scores above the predicted functional cut-off (Figure 3D). Importantly, within the group of active genes, the RSS scores, although generally elevated (Figure 3D), showed no further correlation with individual rates of V_H recombination (Figure 3E). Moreover, models excluding RSS as a predictor also produced high classification rates (>80%). Together, these data suggest that the other factors above determine individual V gene recombination capacity.

Consistent with these findings, DHS, H3K4me1, CTCF, RAD21, IRF4, PAX5, and MED1 significantly co-localized with active V_H genes, while PU.1 did not (Figure 4A). For example, CTCF binding sites were found adjacent to 34 V_H genes, of which 31 were active. In contrast, some features, including H3K9 acetylation, did not co-localize with V_H genes (Figure S5). Notably, active V_H genes with none of the above chromatin marks or TF binding events in the vicinity had much lower recombination rates than those with chromatin marks, independent of their RSS quality (Figure 4B). Conversely, a subset of active V_H genes with poor RIC scores, but six or more adjacent binding events, showed

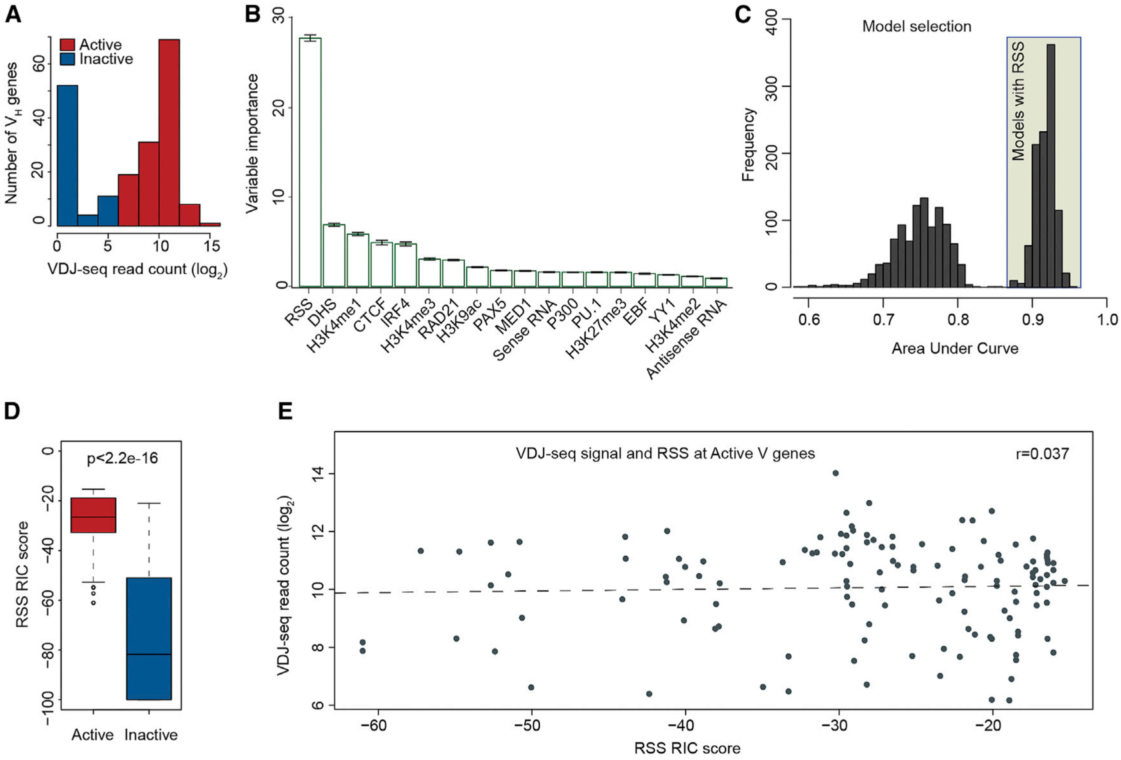


Figure 3. Random Forest Classification Identifies RSS as a Binary Switch for V_H Gene Recombination but with No Predictive Value for Recombination Frequency

(A) Frequency distribution of VDJ-seq read counts for 195 V_H genes, color-coded as active (recombining) or inactive (non-recombining) using a binomial test to gauge the significance of the recombination level. Red: active fdr-adjusted (p value <0.01); blue: inactive fdr-adjusted (p value $\geq$0.01).

(B) Average of out-of-bag variable importance (the gini impurity) in predicting active genes from a Random Forest classifier applied on 18 factors. Error bars show the SEs from a 10-fold cross validation procedure to further control for overfitting. Variables with high gini importance were consistent with the permutation importance measure (data not shown).

(C) Distribution of area under curve (AUC) scores from 2,048 RF models derived from all possible combinations of the 11 most important factors from the RF classifier.

(D) RSS RIC scores of active versus inactive V_H genes.

(E) RSS RIC scores at active genes show little correlation with rates of recombination of individual V_H genes.

comparable recombination frequencies to subsets with good RIC scores (Figure 4B), suggesting that cooperative TF binding may partially compensate for a poor Rag binding context.

The absence of a quantitative link between RSS quality and recombination frequency of active V_H genes, together with the requirement for other factors suggest that the RSSs serve as enabling genetic "binary switches" of recombination at each V_H gene, while other features control the efficiency of this mechanism.

Two Distinct *cis*-Regulatory Designs at RSS Regions Associate with Active Recombination

The transcription factors and histone marks identified as predictive of active recombination (Figure 4A) bind heterogeneously throughout the *Igh* V region. We asked how their binding sites are distributed across the region. First, we analyzed the distance between active V_H genes and the nearest ChIP-seq peak summits of these factors. As expected, CTCF, RAD21, IRF4, PAX5, DHS, H3K4me1, MED1, and PU.1 localized very close to subsets of active genes (light shaded panels in Figures 4C, S5A, and S5B; active genes are red). Importantly, patterns of factors co-localizing with specific V_H genes appeared mutually exclusive. In particular, CTCF and RAD21 preferentially associated with J_H-proximal active V_H genes, while PAX5 and IRF4 commonly co-localized with active genes in the middle and distal regions (Figure 4C).

To further investigate this, we used ChromHMM (Ernst and Kellis, 2012) to partition the V_H region into discrete states based on the chromatin marks and TFs used in the RF classification (Supplemental Experimental Procedures). ChromHMM has recently been used AgR-wide to identify putative enhancer regions (Predeus et al., 2014). Surprisingly, despite the complexity of the locus and the large number of factors included, we observed that low within-class heterogeneity and a high between-state separation could be achieved with just three chromatin states: two highly distinctive "regulatory" states and a "background" (Bg) state depleted of all of these factors (Figures 5A, 5B, and S6A). The first regulatory state (termed "A") was characterized by the binding of "architectural proteins" CTCF and RAD21 (Phillips-Cremins et al., 2013). The second state was best characterized by the binding of PAX5, IRF4, and YY1, and the enrichment of "active" chromatin marks H3K4me1, H3K4me2, H3K4me3,

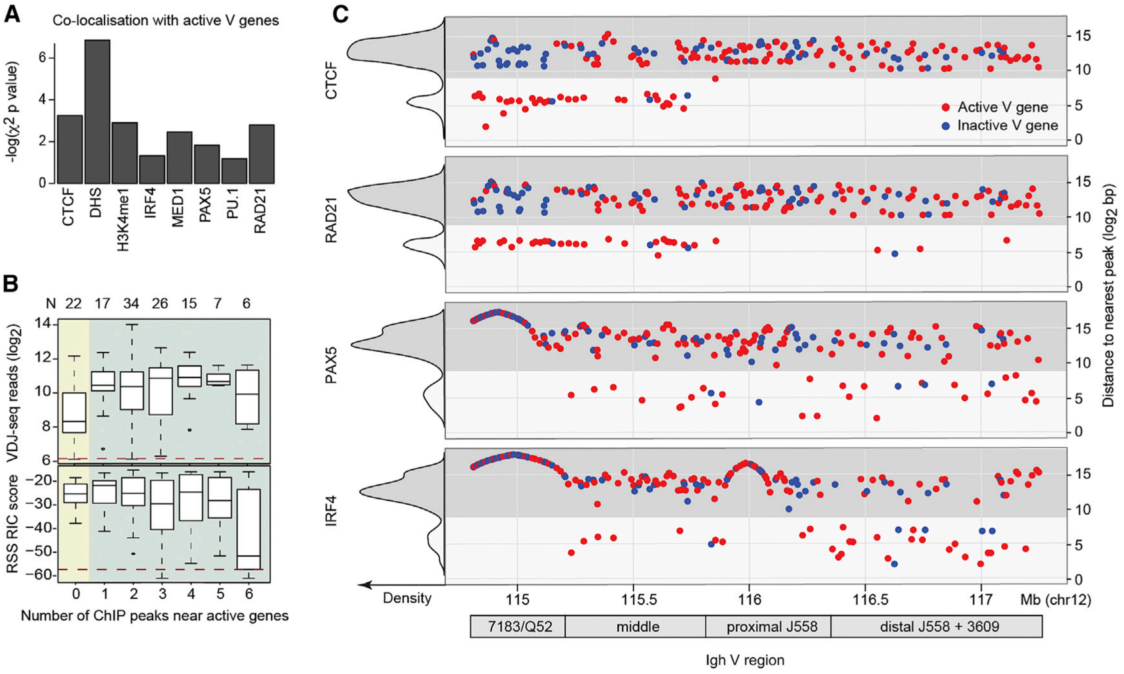


Figure 4. Epigenetic Factors Co-localize with V_H Genes

(A) P values from a χ^2 test to gauge the significance level of the observed number of actively recombining genes that co-localize with eight factors compared to the number expected if co-localization was randomly distributed between active and inactive V genes. A value of 1.3 ($\log_{10}$ of 0.05) or above indicates significant association with active V genes. The remaining factors are not co-localized (Figure S5A).

(B) Relationship between the number of factor peaks associated with V_H genes and recombination frequency (upper panel*)* and RSS RIC scores (lower panel). Dashed red lines indicate the threshold VDJ-seq read count for active V_H genes (upper panel) and the pass/fail RSS RIC score threshold (lower panel). Numbers of V_H genes in each group are shown above.

(C) Distances from V_H genes to the nearest peaks for CTCF, RAD21, PAX5, and IRF4 exhibit bimodality (see also Figure S5B). The subset of V_H genes with nearby peaks (≤1 kb, light shading) was enriched for active genes (red). The curves in the left hand y axis illustrate the frequency of genes in each group. Chromosomal position and location within the V_H region are shown below.

See also Figure S5.

and H3K9ac (Figures 5A, 5C, and S6A). It thus had features associated with hematopoietic transcriptional regulatory elements (Lelli et al., 2012), and we refer to it as the "E" state. DHS, PU.1, and MED1 were enriched in both states.

V_H genes associated with either A or E states recombined significantly more than Bg state genes (Figure 5D). Consistent with this, 76% of active V_H genes (97/128) associated with either one of the two "regulatory" states (state A: 33 genes; state E: 78 genes; Bg state: 84; Data S1). The remaining 24% of active V_H genes (31/128) associated with the Bg state all recombined poorly, despite having higher RIC scores than the average for A state genes (Figure 5E). Conversely, 80% of inactive genes (53/67; Fisher's exact test p = 6.9e-13 versus random expectation) were associated with the Bg state. Importantly, this included most (18/23) of the inactive genes with functional RSSs (23/67). This is strikingly illustrated within the large J558 family. When the active V_H genes therein are separated into high and low recombiners, there is no difference between RIC scores of the two subgroups, but marked differences in the chromatin state (Figure S6B). Together, these results indicate that the Bg chromatin state is refractory to recombination, while state A and state E represent two distinct regulatory architectures associated with active recombination.

ChromHMM analysis of J_H and D_H genes revealed that all four J_H genes overlap with hallmarks of state E (Pax5, IRF4) and lack state A CTCF and Rad21 binding (Figure S6C). In turn, the two most 3′ D_H genes (DQ52 and DST4) and the most 5′ D_H gene (DFL16.1) (Figure 2) overlap with the E state (Figure S6C). The six DSP genes all overlap with the Bg state, in agreement with previous reports of repressive chromatin marks at these genes (Chakraborty et al., 2007), despite significant antisense transcription (Bolland et al., 2007) and frequent recombination (Figure 2).

Fine-scale analysis of ChIP-seq signals revealed that CTCF and RAD21 were enriched specifically over the RSS of A state genes (Figures 6A and 6B), consistent with previous reports for some A state genes (Choi et al., 2013; Lucas et al., 2011). Surprisingly, the E state-associated transcription factors PAX5 and IRF4 were also enriched close to the RSS region of their respective genes and not at V_H promoter regions (Figures 6A and 6B), along with the high DNase hypersensitivity common to both signatures. Our analyses, therefore, suggest the RSS region as a key *cis*-regulatory region for enhancement of recombination conforming to either of two distinctive *cis*-regulatory designs.

To determine whether these two *cis*-regulatory designs have a more wide-spread role in determining recombination-permissive

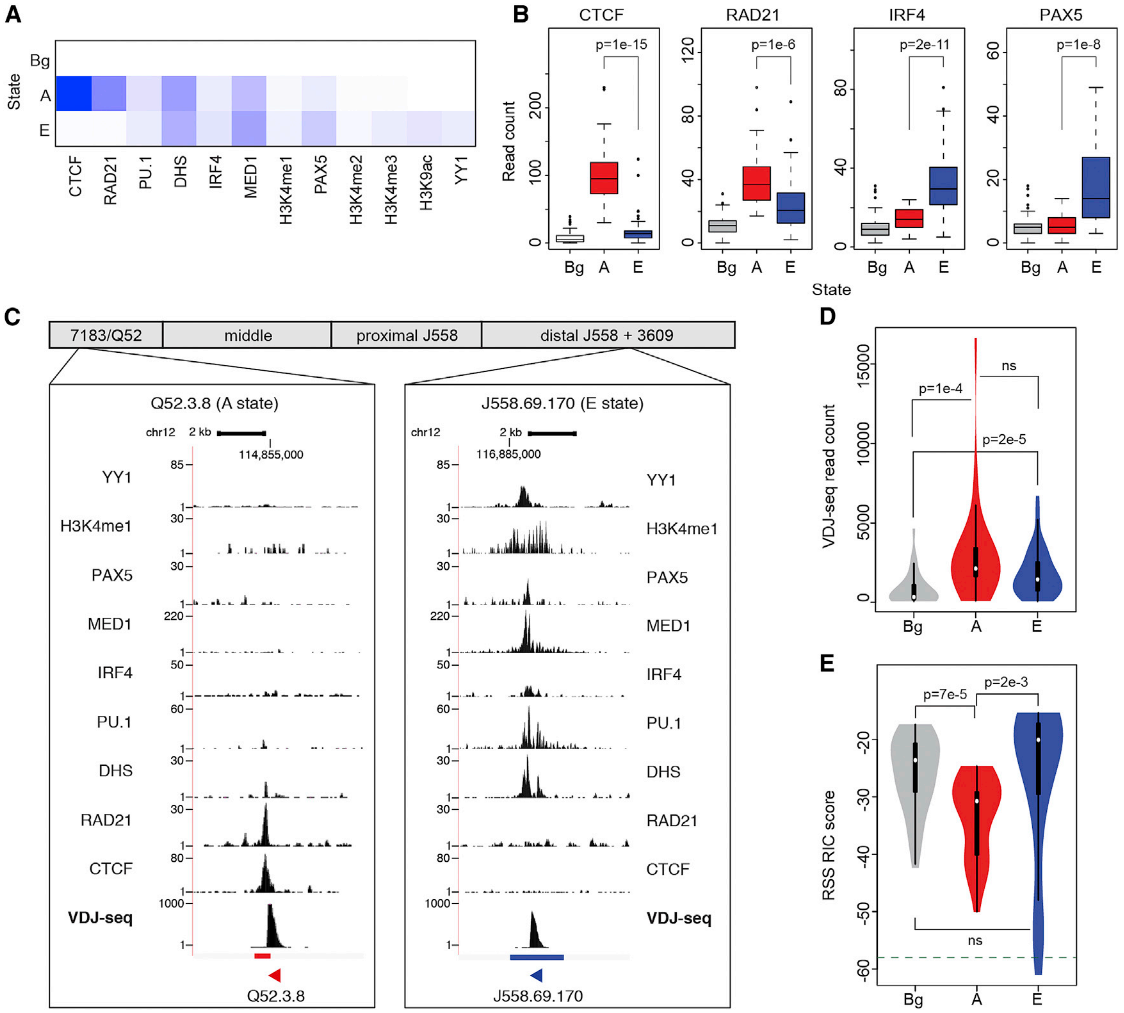


Figure 5. Identification of Chromatin States across the V_H Locus

(A) ChromHMM emission probability (composition) of 12 epigenetic factors in three chromatin states: background (Bg), architectural (A), and enhancer (E). Range, zero (white) to 1 (dark blue).

(B) Comparison of the significance of the read counts for CTCF, RAD21, PAX5, and IRF4 in the three states (Figure S6A; remaining eight factors).

(C) Examples of recombining genes in the A (enriched for CTCF and RAD21) and E (enriched for DHS, PAX5, YY1, H3K4me1, MED1, PU.1, and IRF4) states.

(D and E) Comparison of VDJ-seq read counts (D) and RSS RIC scores (E) for active genes in each of the three states. P values are driven by Wilcoxon test. See also Figure S6.

sites, we examined the >3,000 Rag1 binding sites across the genome that colocalize with Rag2 binding and H3K4me3 enrichment (Teng et al., 2015). We found that CTCF, PAX5, and IRF4 binding profiles all align closely with Rag1 peaks (Figure 6C), suggesting that state A- and E-specific signatures provide a more focused set of candidate sites for off-target Rag recombination.

cis-Regulatory Architecture of Active V_H Genes Reflects Their Evolutionary History

Generally, the A state was more abundant at the J_H-proximal end, while the E state was more enriched in the distal domain of the *Igh* locus (Figure S6D). However, there were numerous exceptions, suggesting that genomic position is not the primary determinant of V_H gene regulatory architecture. We asked whether the distribution of the regulatory states across V_H genes could be explained by their evolutionary history. The 16 V_H gene families have evolved into three separate clans—a large clan 1 and the more closely related smaller clans 2 and 3 (Figure 7A), based on conservation of V_H gene framework 1 (FR1) sequences (Schroeder et al., 1990; Kirkham et al., 1992) and overall DNA sequence (Johnston et al., 2006). We found the A state (CTCF/RAD21) exclusively at V_H gene families from clans 2 and 3, while the E state was almost exclusively associated with V_H genes of clan 1, with the exceptional inclusion of the V_H3609 family forming an outlier group in clan 2 (Schroeder et al., 1990) (Figure 7B). The association of different clans with each of the two states is particularly striking in the poorly understood "middle families" region, which contains a frequently "oscillating" mixture of clan 1 and 3 genes and undergoes a consistent frequent "switching" of states A and E across the region (Figures 7C

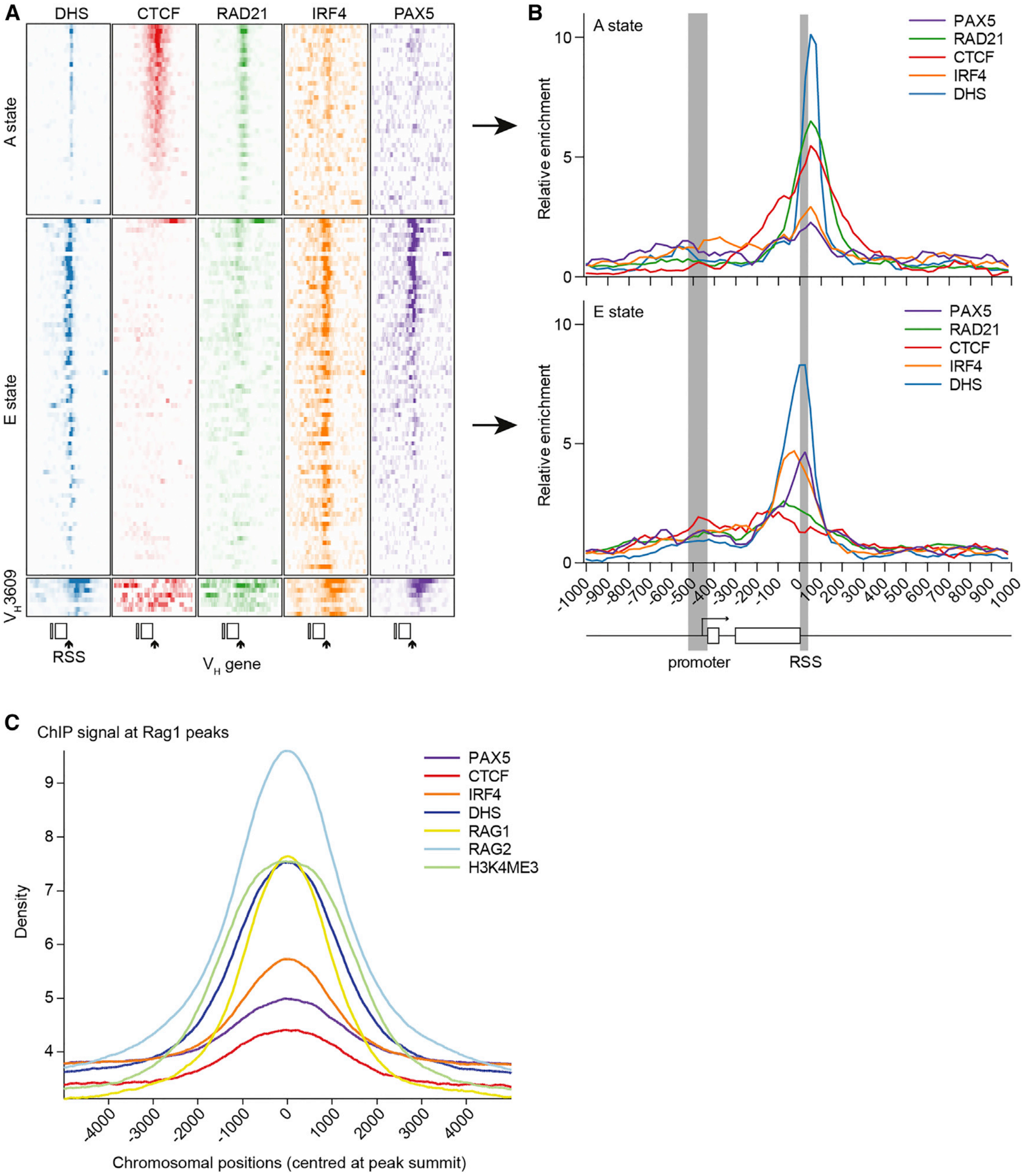


Figure 6. Epigenetic Factors Are Specifically Enriched at the V_H RSS

(A) Aligned region plot for the 2 kb centered on the RSS for V_H genes in the two active states and the V_H3609 outlier family. This shows enrichment for DHS, CTCF, and RAD21 for the A state and DHS, IRF4, and PAX5 for the E state. Enrichment is localized close to the RSS in all cases.

(B) Line graph of average relative enrichment for each factor for genes in the A and E states.

(C) Signal intensity at genome-wide Rag1-ChIP peaks. Signal intensity of five features (chosen to represent both regulatory states) and Rag2 at 3388 Rag1 peaks (Teng et al., 2015). Density defined as normalized log-based read counts in 2 kb regions centered at the summit of peaks.

and S6D). The segregation of regulatory states with evolutionary clans was observed even for the minority of inactive genes (14/67) that associated with a regulatory state (Data S1). Collectively, these data demonstrate that the recombination-regulatory states reflect the evolutionary history of their respective V_H gene clans.

Clan-Specific Differences in Expression of the Recombined VDJ Repertoire

We observed an appreciable correlation between the DNA-based VDJ-seq and the outputs of the expressed *Igh* repertoire (Choi et al., 2013) (Figure S7A). However, there were also significant differences in individual genes, and expression from 14

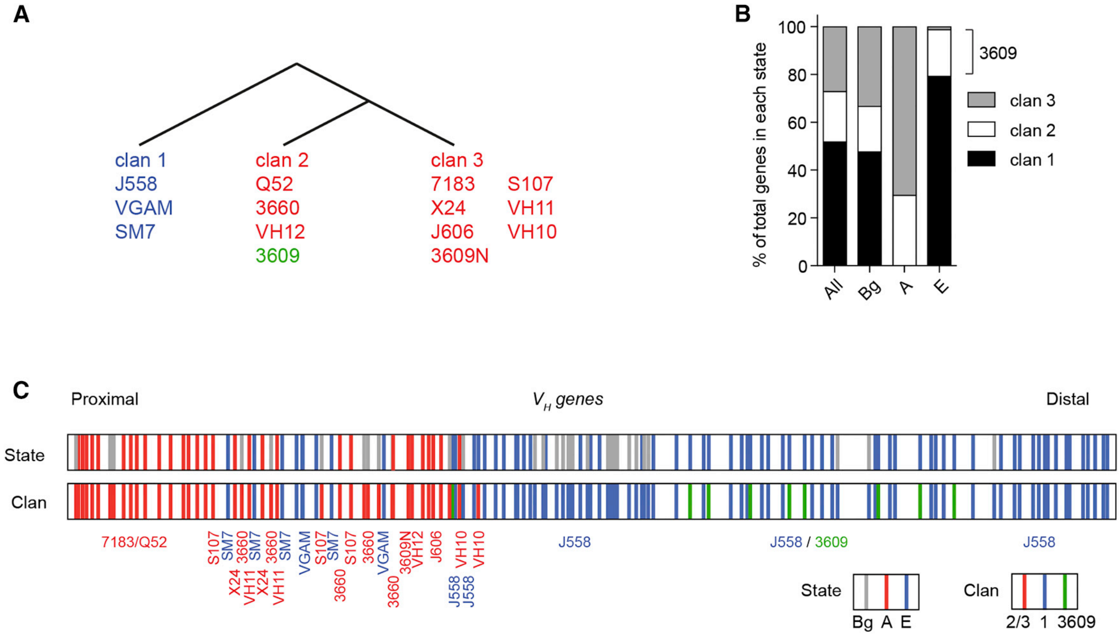


Figure 7. Regulatory States Co-segregate with Evolutionary V_H Clans

(A) Evolutionary organization of V gene families in clans.

(B) Relationship between regulatory states and evolutionary V_H clans.

(C) Geographical position of actively recombining V_H genes overlapping the three states (top) and V_H clans (bottom) across the mouse V_H region. Clans 2 and 3 are colored the same (red) to reflect their overlap in state classification (A). The V_H3609 family is colored green to denote its outlier status as an E state, but clan 2, family. Co-switching of the A and E states with clans 2+3 and 1, respectively, can be seen for the middle families.

See also Figure S6D.

active genes was undetectable (Figure S7B). Segregation of the data into clans demonstrated that the RNA-based data deviated from VDJ-seq in a clan-specific fashion (Figure S7C). Specifically, while clan 1 showed comparable median output between the two assays, clan 2 and particularly clan 3 genes were generally under-represented in the RNA-based assay. This comparison contributes to our understanding of the modulation of the expressed repertoire by downstream factors including promoter strength (Love et al., 2000), mRNA stability, and productivity, while highlighting the limitations of RNA-based approaches for quantification of primary recombination events.

Local Regulatory Events at V Genes Occur in Addition to Global Locus Looping

We asked whether recombination frequency reflects the topological localization of the V genes. A recent 5C-based study (Montefiori et al., 2016) identified six sites in the V region that potentially provide a structural "backbone" of independent chromatin subdomains, within which more local interactions occur. However, we did not observe selective clustering of highly recombining V genes at these sites based on either VDJ-seq (Figure 7D) or expression (Choi et al., 2013; not shown). Active V genes also did not consistently co-localize with the DJ-interacting domains in the V region defined by 4C-seq (Medvedovic et al., 2013). Together, these results suggest that the looping events detected by these studies act in combination with local regulatory mechanisms to shape the V(D)J repertoire.

DISCUSSION

The DNA-based VDJ-seq technique established here has enabled a quantitative assessment of the recombination frequency of all V_H genes in the mouse *Igh* locus and led to detection of *cis*-regulatory signatures associated with active recombination. These signatures explain the enormous variation in recombination efficiency throughout the V_H region, including at geographically neighboring V genes. Therefore, in addition to the large-scale looping mechanisms necessary to bring distal V genes close to DJ segments (Chaumeil and Skok, 2012; Fuxa et al., 2004; Gerasimova et al., 2015; Guo et al., 2011; Medvedovic et al., 2013), local chromatin regulation of V genes likely plays a pivotal role in recombination outcome.

V_H Gene Downstream Sites: Dual-Function Recombination Regulatory Elements?

Integrating VDJ-seq data with genetic and epigenetic annotations, we have produced a functional view of the factors associated with recombination of individual active V genes. Consistent with previous studies, we find that the stringency of the Rag-binding RSSs does not fully explain the differential recombination efficiency across the *Igh* locus (Choi et al., 2013; Merelli et al., 2010). Nevertheless, our results suggest that regions proximal to the RSSs regulate V_H gene usage via combinatorial transcription factor recruitment. These elements therefore exquisitely juxtapose the two requirements for recombination: a functional RSS and a permissive chromatin state likely generated by key transcription factors. Importantly, these downstream elements localize on the opposite sides of the V_H genes from their promoters and are excised during recombination of their associated V_H gene. Therefore, the regulatory mechanisms underpinning efficient recombination are both spatially and temporally separated from those driving the expression of recombined V_H genes.

We further reveal two mutually exclusive designs at these recombination-regulatory sites, implicating CTCF/RAD21 and PAX5/IRF4 in the local regulation of V_H gene recombination efficiency in *cis*. It was previously hypothesized that CTCF forms the base of local loops at D-proximal V genes, thus impacting on recombination efficiency (Lucas et al., 2011). Here, we show that both CTCF and RAD21 have a strong association with active V_H genes in clans 2 and 3. RAD21 binding is developmentally restricted to pro-B cells undergoing V-to-DJ recombination (Degner et al., 2009) and may have a more stage-specific local role in activating recombination.

PAX5 association with recombination-regulatory elements is consistent with a previous model, in which Pax5 recruits Rag to individual V genes (Zhang et al., 2006). Pax5 was postulated to enable recombination of distal V_H genes by promoting DNA looping (Fuxa et al., 2004). However, its local role revealed here may provide an additional explanation of the reduced recombination of state-E V_H genes in a PAX5 mutant, particularly those that are not distally located such as the VGam 3.8 family (Fuxa et al., 2004).

Finally, we provide evidence that IRF4 is associated with recombination of *Igh* state E V_H genes. IRF4 was previously shown to regulate *Ig*κ recombination (Johnson et al., 2008) potentially by promoting accessibility of *Ig*κ RSSs (Bevington and Boyes, 2013). IRF4 is a known target of PAX5 (Revilla-I-Domingo et al., 2012) and its colocalization with PAX5 around V_H RSSs suggests a feed-forward loop (Palomero et al., 2006).

The Distribution of Regulatory States across the *Igh* Locus

Detailed analysis of interspersed V gene families led us to the discovery that the chromatin states were based on clan evolution and not on geographical location. We note that our conclusions differ from those of Choi et al. (2013), who proposed that the V region is divided into four chromatin states in a geographical manner, each with characteristics that favor recombination and/or compensate for unfavorable factors. We attribute the different conclusions in part to greater sequencing depth and the ability of VDJ-seq to report on DNA rather than downstream RNA expression and detect non-productive recombination events and recombining pseudogenes. Inclusion of a wider range of relevant transcription factors, including Pax5, IRF4, YY1, and PU.1 enabled us to resolve V genes into just two active chromatin states.

Recombination and Transcription Mediated by Spatially Separate Elements

The discovery of germline transcription from *Igh* V gene promoters before V to DJ recombination prompted the accessibility hypothesis, which proposed that V(D)J recombination is regulated by controlling access of the RAG enzymes to the antigen receptor loci (Yancopoulos and Alt, 1985). However, subsequent studies have produced conflicting findings both in vitro and in vivo (Baumann et al., 2003; Bevington and Boyes, 2013; Buchanan et al., 1997; Du et al., 2008; Ji et al., 2010; Kondilis-Mangum et al., 2010; Love et al., 2000). While we cannot reconcile the debate fully, our evidence supports the notion that neither promoter activity nor transcription discriminate between active and inactive genes. First, we did not find sense non-coding RNA transcription over unrecombined V_H genes to be a strong predictor of active recombination. In part, this reflects the very low sense transcript levels detected by RNA-seq. Neither did we find a correlation between transcription factor binding and antisense transcription at RSSs. Second, fine-scale mapping of ChIP-seq datasets around state A and E genes did not reveal any recombination-associated patterns at the promoters. Rather, the canonical promoter-associated mark H3K4me3 was enriched around the RSSs of state E genes, consistent with its additional role in facilitating Rag2 binding (Matthews et al., 2007). Although germline transcription does not discriminate between actively recombining versus inactive $V_{\mathbf{H}}$ genes, it may still be required for recombination, for example, by providing the first level of chromatin "priming." It is clear that V gene promoter activity plays an important role in shaping the expressed repertoire post-recombination, as we show by comparing DNA-based (VDJ-seq) and RNA-based (Choi et al., 2013) outputs.

Implications for the Human *IGH* Locus

In humans, V_H genes maintain evolutionarily conserved clan identities, but have a very different geographical organization to the mouse, with interspersed gene families and no polarity of clan position (Das et al., 2008; de Bono et al., 2004; Schroeder et al., 1990). The human locus is also much smaller than the mouse (1 Mb versus 2.5 Mb), and the role of looping in its regulation is unknown. Despite these differences and consistent with our findings in mouse, CTCF is associated with 90% of clan 2/3 genes in a human lymphoblastoid cell line (GM12878) (ENCODE Project Consortium, 2012), supporting the evolutionary conservation of downstream recombination-regulatory sites. Accordingly, an attractive hypothesis is that the segregating local chromatin states and their potential role in active recombination reported here also apply to the human *IGH* locus.

Implications for Other Antigen Receptor Loci

RSS-mediated recombination is common to all AgR loci. Thus, it will be important to determine whether other AgRs have a similar *cis*-regulatory organization. For instance, CTCF and RAD21 may play a role in local *Ig*κ or TCR gene recombination, and indeed CTCF regulates *Igk* recombination (Ribeiro de Almeida et al., 2011), while CTCF and Rad21 regulate TCRα and TCRβ locus conformation and recombination (Chen et al., 2015; Seitan et al., 2011; Shih et al., 2012). Some of the key factors for *Igh* are only expressed in B cells (PAX5, IRF4) and thus may be relevant for *Ig*κ and *Igl*, but not for TCRs. Indeed, IRF4 and IRF8 are implicated in *Igk* recombination (Johnson et al., 2008). T cell-specific factors implicated in TCR recombination, including

RUNX1 may also have local as well as the long-range looping roles (Cieslak et al., 2014).

Insights into Inappropriate Rag-Mediated Recombination

Rag recombinases mediate aberrant chromosomal translocations and deletions (Helmink and Sleckman, 2012; Hu et al., 2015), but the determinants of their mis-localization are not fully understood. While cryptic RSSs are extremely frequent throughout the genome, RSS quality is a weak predictor of aberrant recombination (Zhang and Swanson, 2008). Permissive chromatin structure has been implicated (Shimazaki et al., 2012), and indeed, Rag-mediated DNA breaks in acute lymphoblastic leukemias, while poorly predictable by RSS quality, are enriched at active promoters and enhancers (Papaemmanuil et al., 2014). Consistent with this, recent studies have revealed thousands of Rag1 binding sites at promoters and enhancers (Teng et al., 2015). Here, we show that signature chromatin states at canonical RSS sites are critical features associated with their recombination potential. Notably, state E genes are enriched in H3K4me3, H3K4me1, canonical promoter, and enhancer marks that provide docking sites for Rag1 and Rag2 throughout the genome. However, state A genes are depleted of these marks suggesting the promoter/enhancer signature is not a universal requirement for RSS-mediated recombination, and an additional class of CTCF/Rad21-associated signatures warrants investigation for cryptic RSS cleavage. In support of this hypothesis, the 3,000 Rag1/Rag2 bound sites in the genome were enriched for markers of state A and state E. Combining RSS quality with identification of these chromatin structures may provide a more in-depth set of rearrangement-prone sites, which, as for canonical sites, may differ in different lymphocyte subpopulations.

EXPERIMENTAL PROCEDURES

All mice were maintained in accordance with Babraham Institute Animal Welfare and Ethical Review Body and Home office rules and ARRIVE guidelines under Project Licence 80/2529. Detailed methods are in the Supplemental Information available online.

VDJ-Seq

VDJ-seq is the capture and amplification for Illumina sequencing of *Igh* DJ and VDJ recombined genes from genomic DNA by primer extension from reverse-oriented biotinylated J gene oligonucleotides. A flow chart of the VDJ-seq assay is provided in Figure 1. Oligonucleotide sequences are provided in Data S1.

VDJ-Seq Pipeline: Babraham LinkON

A bioinformatic pipeline was developed to process raw sequences. True J gene containing reads were identified and deduplicated to identify unique V-J read pairs based on the J sequence and the V read start position. Only J sequences with different V read start positions were called as unique VDJ joins. Flow chart of Babraham LinkON is provided in Figure S1A. We used Seqmonk (Babraham Bioinformatics), a freely available java-based tool, to visualize and quantify unique DJ and VDJ sequences in mapped sequencing data.

VDJ-Seq IMGT HighV-Quest Pipeline

IMGT HighV-quest is a high-throughput web tool for the analysis of VDJ junctions providing data including V, D, and J genes used, coding potential, P and N nucleotides, and CDR3 AA sequence. Analysis of VDJ-seq data using IMGT HighV-quest required a pipeline to link V and J reads. See the details in the Supplemental Information.

Computational and Statistical Approaches

Random Forest Analysis

Recombining active V_H genes were defined as those enriched for VDJ-seq reads compared with the V region as a whole, ascertained using a binomial test (padj < 0.01) for each gene. The not-significantly recombining genes were defined as inactive. These binary recombination classes were used as response variables in a Random Forest classifier model (Liaw and Wiener, 2002). Predictors for the model were signal intensities from DHS-seq, ChIP data, and RNA-seq data extracted from 2.5 kb surrounding each V gene. The model was evaluated in a 10-fold cross-validation to prevent over-fitting.

Co-localization

ChIP peaks (including DHS) were called using MACS2 (Zhang et al., 2008). The distance between the RSS and the summit of nearest peak was measured for each dataset. The significance of co-localization (versus active and inactive genes) was assessed using a χ^2 test.

Chromatin Segmentation

The *Igh* locus was split into 200 bp bins. Each bin overlapping with ChIP or DHS peaks was assigned 1, otherwise 0. The resulting binary matrix was used as input for the chromatin segmentation algorithm chromHMM (Ernst and Kellis, 2012). The significance of association between chromatin states and V gene recombination classes was assessed by a Fisher's exact test.

ACCESSION NUMBERS

The accession number for VDJ-seq, H3K4me3 ChIP-seq, and nuclear RNA-seq datasets reported in this paper is GEO: GSE80155.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, and one data file and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2016.05.020.

AUTHOR CONTRIBUTIONS

A.L.W. conceived the first version of VDJ-seq. D.J.B. developed the final version reported here and generated VDJ-seq data. K.T. sequenced Illumina libraries. F.K. and S.R.A. developed the Babraham LinkON pipeline. L.M., M.J.T.S., A.B.-E., P.C., and B.A.S. refined the VDJ-seq protocol and analysis pipeline. L.M. generated RNA-seq and ChIP-seq datasets. H.K. pre-processed and Q.C. checked all NGS datasets and conceived and performed machine learning analysis. D.J.B. and H.K. visualized the data and prepared the figures. D.J.B., H.K., M.S., and A.E.C. interpreted the results and wrote the manuscript.

CONFLICTS OF INTEREST

D.J.B., A.L.W., L.S.M., and A.E.C. are named inventors on a patent filed, "Covering the VDJ-seq technique: Method of identifying VDJ recombination products." (UK Patent Application No. GB1203720.6, filed March 2, 2012; PCT Patent Applic No. PCT/GB2013/05056, published September 6, 2013. National applications filed Europe, USA, Japan. US Publication number: 20150031042, publication date January 29, 2015.)

ACKNOWLEDGMENTS

We thank Peter Fraser and Martin Turner for critical reading of the manuscript. This work was supported by the Biotechnology and Biological Sciences Research Council UK (BBSRC) (D.J.B., A.L.W., H.K., L.M., F.K., M.J.T.S., P.C., K.T., S.R.A., M.S., and A.E.C.) and the Medical Research Council UK (A.B.-E. and B.A.S.).

Received: December 21, 2015
Revised: March 1, 2016
Accepted: May 2, 2016
Published: June 2, 2016

REFERENCES

Alamyar, E., Giudicelli, V., Li, S., Duroux, P., and 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.

Baumann, M., Mamais, A., McBlane, F., Xiao, H., and Boyes, J. (2003). Regulation of V(D)J recombination by nucleosome positioning at recombination signal sequences. EMBO J. *22*, 5197–5207.

Benichou, J., Ben-Hamo, R., Louzoun, Y., and Efroni, S. (2012). Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. Immunology *135*, 183–191.

Bevington, S., and Boyes, J. (2013). Transcription-coupled eviction of histones H2A/H2B governs V(D)J recombination. EMBO J. *32*, 1381–1392.

Bolland, D.J., Wood, A.L., Johnston, C.M., Bunting, S.F., Morgan, G., Chakalova, L., Fraser, P.J., and Corcoran, A.E. (2004). Antisense intergenic transcription in V(D)J recombination. Nat. Immunol. *5*, 630–637.

Bolland, D.J., Wood, A.L., Afshar, R., Featherstone, K., Oltz, E.M., and Corcoran, A.E. (2007). Antisense intergenic transcription precedes Igh D-to-J recombination and is controlled by the intronic enhancer Emu. Mol. Cell. Biol. *27*, 5523–5533.

Buchanan, K.L., Smith, E.A., Dou, S., Corcoran, L.M., and Webb, C.F. (1997). Family-specific differences in transcription efficiency of Ig heavy chain promoters. J. Immunol. *159*, 1247–1254.

Chakraborty, T., Chowdhury, D., Keyes, A., Jani, A., Subrahmanyam, R., Ivanova, I., and Sen, R. (2007). Repeat organization and epigenetic regulation of the DH-Cmu domain of the immunoglobulin heavy-chain gene locus. Mol. Cell *27*, 842–850.

Chaumeil, J., and Skok, J.A. (2012). The role of CTCF in regulating V(D)J recombination. Curr. Opin. Immunol. *24*, 153–159.

Chen, L., Carico, Z., Shih, H.-Y., and Krangel, M.S. (2015). A discrete chromatin loop in the mouse Tcra-Tcrd locus shapes the TCRδ and TCRα repertoires. Nat. Immunol. *16*, 1085–1093.

Choi, N.M., Loguercio, S., Verma-Gaur, J., Degner, S.C., Torkamani, A., Su, A.I., Oltz, E.M., Artyomov, M., and Feeney, A.J. (2013). Deep sequencing of the murine IgH repertoire reveals complex regulation of nonrandom V gene rearrangement frequencies. J. Immunol. *191*, 2393–2402.

Cieslak, A., Le Noir, S., Trinquand, A., Lhermitte, L., Franchini, D.M., Villarese, P., Gon, S., Bond, J., Simonin, M., Vanhille, L., et al. (2014). RUNX1-dependent RAG1 deposition instigates human TCR-δ locus rearrangement. J. Exp. Med. *211*, 1821–1832.

Corcoran, A.E. (2010). The epigenetic role of non-coding RNA transcription and nuclear organization in immunoglobulin repertoire generation. Semin. Immunol. *22*, 353–361.

Cowell, L.G., Davila, M., Kepler, T.B., and Kelsoe, G. (2002). Identification and utilization of arbitrary correlations in models of recombination signal sequences. Genome Biol. *3*, research0072.1–research0072.20.

Das, S., Nozawa, M., Klein, J., and Nei, M. (2008). Evolutionary dynamics of the immunoglobulin heavy chain variable region genes in vertebrates. Immunogenetics *60*, 47–55.

de Bono, B., Madera, M., and Chothia, C. (2004). VH gene segments in the mouse and human genomes. J. Mol. Biol. *342*, 131–143.

Degner, S.C., Wong, T.P., Jankevicius, G., and Feeney, A.J. (2009). Cutting edge: developmental stage-specific recruitment of cohesin to CTCF sites throughout immunoglobulin loci during B lymphocyte development. J. Immunol. *182*, 44–48.

Degner, S.C., Verma-Gaur, J., Wong, T.P., Bossen, C., Iverson, G.M., Torkamani, A., Vettermann, C., Lin, Y.C., Ju, Z., Schulz, D., et al. (2011). CCCTC-binding factor (CTCF) and cohesin influence the genomic architecture of the Igh locus and antisense transcription in pro-B cells. Proc. Natl. Acad. Sci. USA *108*, 9566–9571.

Du, H., Ishii, H., Pazin, M.J., and Sen, R. (2008). Activation of 12/23-RSS-dependent RAG cleavage by hSWI/SNF complex in the absence of transcription. Mol. Cell *31*, 641–649.

Dunn-Walters, D.K., and Ademokun, A.A. (2010). B cell repertoire and ageing. Curr. Opin. Immunol. *22*, 514–520.

Eberle, A.B., Herrmann, K., Jäck, H.-M., and Mühlemann, O. (2009). Equal transcription rates of productively and nonproductively rearranged immunoglobulin mu heavy chain alleles in a pro-B cell line. RNA *15*, 1021–1028.

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

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nat. Methods *9*, 215–216.

Fang, W., Mueller, D.L., Pennell, C.A., Rivard, J.J., Li, Y.S., Hardy, R.R., Schlissel, M.S., and Behrens, T.W. (1996). Frequent aberrant immunoglobulin gene rearrangements in pro-B cells revealed by a bcl-xL transgene. Immunity *4*, 291–299.

Fuxa, M., Skok, J., Souabni, A., Salvagiotto, G., Roldan, E., and Busslinger, M. (2004). Pax5 induces V-to-DJ rearrangements and locus contraction of the immunoglobulin heavy-chain gene. Genes Dev. *18*, 411–422.

Georgiou, G., Ippolito, G.C., Beausang, J., Busse, C.E., Wardemann, H., and Quake, S.R. (2014). The promise and challenge of high-throughput sequencing of the antibody repertoire. Nat. Biotechnol. *32*, 158–168.

Gerasimova, T., Guo, C., Ghosh, A., Qiu, X., Montefiori, L., Verma-Gaur, J., Choi, N.M., Feeney, A.J., and Sen, R. (2015). A structural hierarchy mediated by multiple nuclear factors establishes IgH locus conformation. Genes Dev. *29*, 1683–1695.

Gopalakrishnan, S., Majumder, K., Predeus, A., Huang, Y., Koues, O.I., Verma-Gaur, J., Loguercio, S., Su, A.I., Feeney, A.J., Artyomov, M.N., and Oltz, E.M. (2013). Unifying model for molecular determinants of the preselection Vβ repertoire. Proc. Natl. Acad. Sci. USA *110*, E3206–E3215.

Guo, C., Gerasimova, T., Hao, H., Ivanova, I., Chakraborty, T., Selimyan, R., Oltz, E.M., and Sen, R. (2011). Two forms of loops generate the chromatin conformation of the immunoglobulin heavy-chain gene locus. Cell *147*, 332–343.

Helmink, B.A., and Sleckman, B.P. (2012). The response to and repair of RAG-mediated DNA double-stranded breaks. Annu. Rev. Immunol. *30*, 175–202.

Hu, J., Zhang, Y., Zhao, L., Frock, R.L., Du, Z., Meyers, R.M., Meng, F.-L., Schatz, D.G., and Alt, F.W. (2015). Chromosomal loop domains direct the recombination of antigen receptor genes. Cell *163*, 947–959.

Jhunjhunwala, S., van Zelm, M.C., Peak, M.M., Cutchin, S., Riblet, R., van Dongen, J.J.M., Grosveld, F.G., Knoch, T.A., and Murre, C. (2008). The 3D structure of the immunoglobulin heavy-chain locus: implications for long-range genomic interactions. Cell *133*, 265–279.

Ji, Y., Little, A.J., Banerjee, J.K., Hao, B., Oltz, E.M., Krangel, M.S., and Schatz, D.G. (2010). Promoters, enhancers, and transcription target RAG1 binding during V(D)J recombination. J. Exp. Med. *207*, 2809–2816.

Johnson, K., Hashimshony, T., Sawai, C.M., Pongubala, J.M.R., Skok, J.A., Aifantis, I., and Singh, H. (2008). Regulation of immunoglobulin light-chain recombination by the transcription factor IRF-4 and the attenuation of interleukin-7 signaling. Immunity *28*, 335–345.

Johnston, C.M., Wood, A.L., Bolland, D.J., and Corcoran, A.E. (2006). Complete sequence assembly and characterization of the C57BL/6 mouse Ig heavy chain V region. J. Immunol. *176*, 4221–4234.

Kaplinsky, J., Li, A., Sun, A., Coffre, M., Koralov, S.B., and Arnaout, R. (2014). Antibody repertoire deep sequencing reveals antigen-independent selection in maturing B cells. Proc. Natl. Acad. Sci. USA *111*, E2622–E2629.

Kirkham, P.M., Mortari, F., Newton, J.A., and Schroeder, H.W., Jr. (1992). Immunoglobulin VH clan and family identity predicts variable domain structure and may influence antigen binding. EMBO J. *11*, 603–609.

Kondilis-Mangum, H.D., Cobb, R.M., Osipovich, O., Srivatsan, S., Oltz, E.M., and Krangel, M.S. (2010). Transcription-dependent mobilization of nucleosomes at accessible TCR gene segments in vivo. J. Immunol. *184*, 6970–6977.

Lee, A.I., Fugmann, S.D., Cowell, L.G., Ptaszek, L.M., Kelsoe, G., and Schatz, D.G. (2003). A functional analysis of the spacer of V(D)J recombination signal sequences. PLoS Biol. *1*, E1.

Lelli, K.M., Slattery, M., and Mann, R.S. (2012). Disentangling the many layers of eukaryotic transcriptional regulation. Annu. Rev. Genet. *46*, 43–68.

Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. R. News *2*, 18–22.

Liu, H., Schmidt-Supprian, M., Shi, Y., Hobeika, E., Barteneva, N., Jumaa, H., Pelanda, R., Reth, M., Skok, J., Rajewsky, K., and Shi, Y. (2007). Yin Yang 1 is a critical regulator of B-cell development. Genes Dev. *21*, 1179–1189.

Love, V.A., Lugo, G., Merz, D., and Feeney, A.J. (2000). Individual V(H) promoters vary in strength, but the frequency of rearrangement of those V(H) genes does not correlate with promoter strength nor enhancer-independence. Mol. Immunol. *37*, 29–39.

Lucas, J.S., Bossen, C., and Murre, C. (2011). Transcription and recombination factories: common features? Curr. Opin. Cell Biol. *23*, 318–324.

Marculescu, R., Le, T., Simon, P., Jaeger, U., and Nadel, B. (2002). V(D)J-mediated translocations in lymphoid neoplasms: a functional assessment of genomic instability by cryptic sites. J. Exp. Med. *195*, 85–98.

Matthews, A.G.W., Kuo, A.J., Ramón-Maiques, S., Han, S., Champagne, K.S., Ivanov, D., Gallardo, M., Carney, D., Cheung, P., Ciccone, D.N., et al. (2007). RAG2 PHD finger couples histone H3 lysine 4 trimethylation with V(D)J recombination. Nature *450*, 1106–1110.

Medvedovic, J., Ebert, A., Tagoh, H., Tamir, I.M., Schwickert, T.A., Novatchkova, M., Sun, Q., Huis In 't Veld, P.J., Guo, C., Yoon, H.S., et al. (2013). Flexible long-range loops in the VH gene region of the Igh locus facilitate the generation of a diverse antibody repertoire. Immunity *39*, 229–244.

Merelli, I., Guffanti, A., Fabbri, M., Cocito, A., Furia, L., Grazini, U., Bonnal, R.J., Milanesi, L., and McBlane, F. (2010). RSSsite: a reference database and prediction tool for the identification of cryptic Recombination Signal Sequences in human and murine genomes. Nucleic Acids Res. *38*, W262–W267.

Montefiori, L., Wuerffel, R., Roqueiro, D., Lajoie, B., Guo, C., Gerasimova, T., De, S., Wood, W., Becker, K.G., Dekker, J., et al. (2016). Extremely Long-Range Chromatin Loops Link Topological Domains to Facilitate a Diverse Antibody Repertoire. Cell Rep. *14*, 896–906.

Palomero, T., Lim, W.K., Odom, D.T., Sulis, M.L., Real, P.J., Margolin, A., Barnes, K.C., O'Neil, J., Neuberg, D., Weng, A.P., et al. (2006). NOTCH1 directly regulates c-MYC and activates a feed-forward-loop transcriptional network promoting leukemic cell growth. Proc. Natl. Acad. Sci. USA *103*, 18261–18266.

Papaemmanuil, E., Rapado, I., Li, Y., Potter, N.E., Wedge, D.C., Tubio, J., Alexandrov, L.B., Van Loo, P., Cooke, S.L., Marshall, J., et al. (2014). RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia. Nat. Genet. *46*, 116–125.

Perlmutter, R.M., Kearney, J.F., Chang, S.P., and Hood, L.E. (1985). Developmentally controlled expression of immunoglobulin VH genes. Science *227*, 1597–1601.

Phillips-Cremins, J.E., Sauria, M.E.G., Sanyal, A., Gerasimova, T.I., Lajoie, B.R., Bell, J.S.K., Ong, C.-T., Hookway, T.A., Guo, C., Sun, Y., et al. (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell *153*, 1281–1295.

Predeus, A.V., Gopalakrishnan, S., Huang, Y., Tang, J., Feeney, A.J., Oltz, E.M., and Artyomov, M.N. (2014). Targeted chromatin profiling reveals novel enhancers in Ig H and Ig L chain Loci. J. Immunol. *192*, 1064–1070.

Revilla-I-Domingo, R., Bilic, I., Vilagos, B., Tagoh, H., Ebert, A., Tamir, I.M., Smeenk, L., Trupke, J., Sommer, A., Jaritz, M., and Busslinger, M. (2012). The B-cell identity factor Pax5 regulates distinct transcriptional programmes in early and late B lymphopoiesis. EMBO J. *31*, 3130–3146.

Reynaud, D., Demarco, I.A., Reddy, K.L., Schjerven, H., Bertolino, E., Chen, Z., Smale, S.T., Winandy, S., and Singh, H. (2008). Regulation of B cell fate commitment and immunoglobulin heavy-chain gene rearrangements by Ikaros. Nat. Immunol. *9*, 927–936.

Ribeiro de Almeida, C., Stadhouders, R., de Bruijn, M.J.W., Bergen, I.M., Thongjuea, S., Lenhard, B., van Ijcken, W., Grosveld, F., Galjart, N., Soler, E., and Hendriks, R.W. (2011). The DNA-binding protein CTCF limits proximal Vκ recombination and restricts κ enhancer interactions to the immunoglobulin κ light chain locus. Immunity *35*, 501–513.

Rouaud, P., Vincent-Fabert, C., Fiancette, R., Cogné, M., Pinaud, E., and Denizot, Y. (2012). Enhancers located in heavy chain regulatory region (hs3a, hs1,2, hs3b, and hs4) are dispensable for diversity of VDJ recombination. J. Biol. Chem. *287*, 8356–8360.

Rumfelt, L.L., Zhou, Y., Rowley, B.M., Shinton, S.A., and Hardy, R.R. (2006). Lineage specification and plasticity in CD19- early B cell precursors. J. Exp. Med. *203*, 675–687.

Sayegh, C.E., Jhunjhunwala, S., Riblet, R., and Murre, C. (2005). Visualization of looping involving the immunoglobulin heavy-chain locus in developing B cells. Genes Dev. *19*, 322–327.

Schatz, D.G., and Ji, Y. (2011). Recombination centres and the orchestration of V(D)J recombination. Nat. Rev. Immunol. *11*, 251–263.

Schroeder, H.W., Jr., Hillson, J.L., and Perlmutter, R.M. (1990). Structure and evolution of mammalian VH families. Int. Immunol. *2*, 41–50.

Seitan, V.C., Hao, B., Tachibana-Konwalski, K., Lavagnolli, T., Mira-Bontenbal, H., Brown, K.E., Teng, G., Carroll, T., Terry, A., Horan, K., et al. (2011). A role for cohesin in T-cell-receptor rearrangement and thymocyte differentiation. Nature *476*, 467–471.

Shih, H.-Y., Verma-Gaur, J., Torkamani, A., Feeney, A.J., Galjart, N., and Krangel, M.S. (2012). Tcra gene recombination is supported by a Tcra enhancer- and CTCF-dependent chromatin hub. Proc. Natl. Acad. Sci. USA *109*, E3493–E3502.

Shimazaki, N., Askary, A., Swanson, P.C., and Lieber, M.R. (2012). Mechanistic basis for RAG discrimination between recombination sites and the off-target sites of human lymphomas. Mol. Cell. Biol. *32*, 365–375.

Sollbach, A.E., and Wu, G.E. (1995). Inversions produced during V(D)J rearrangement at IgH, the immunoglobulin heavy-chain locus. Mol. Cell. Biol. *15*, 671–681.

Stubbington, M.J.T., and Corcoran, A.E. (2013). Non-coding transcription and large-scale nuclear organisation of immunoglobulin recombination. Curr. Opin. Genet. Dev. *23*, 81–88.

Teng, G., Maman, Y., Resch, W., Kim, M., Yamane, A., Qian, J., Kieffer-Kwon, K.-R., Mandal, M., Ji, Y., Meffre, E., et al. (2015). RAG represents a widespread threat to the lymphocyte genome. Cell *162*, 751–765.

Yancopoulos, G.D., and Alt, F.W. (1985). Developmentally controlled and tissue-specific expression of unrearranged VH gene segments. Cell *40*, 271–281.

Yancopoulos, G.D., Malynn, B.A., and Alt, F.W. (1988). Developmentally regulated and strain-specific expression of murine VH gene families. J. Exp. Med. *168*, 417–435.

Ye, J. (2004). The immunoglobulin IGHD gene locus in C57BL/6 mice. Immunogenetics *56*, 399–404.

Zhang, M., and Swanson, P.C. (2008). V(D)J recombinase binding and cleavage of cryptic recombination signal sequences identified from lymphoid malignancies. J. Biol. Chem. *283*, 6717–6727.

Zhang, Z., Espinoza, C.R., Yu, Z., Stephan, R., He, T., Williams, G.S., Burrows, P.D., Hagman, J., Feeney, A.J., and Cooper, M.D. (2006). Transcription factor Pax5 (BSAP) transactivates the RAG-mediated V(H)-to-DJ(H) rearrangement of immunoglobulin genes. Nat. Immunol. *7*, 616–624.

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

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

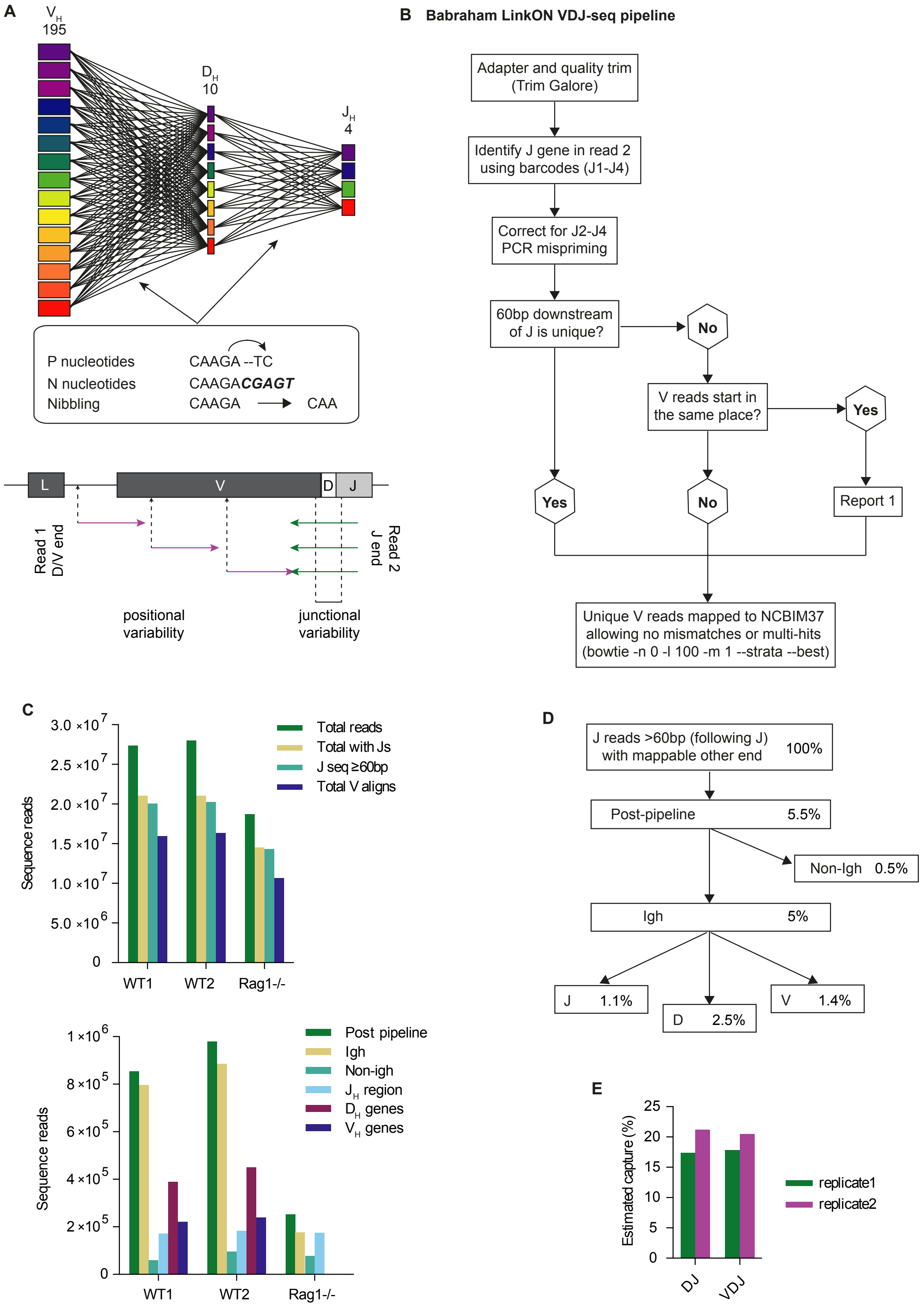


Figure S1 ***Related to Figure 1*** **Details of the VDJseq pipeline, read counts and estimated capture**
A) *Upper*, V(D)J_H recombination gives both combinatorial and junctional diversity as shown. *Lower*, for deduplication VDJ-seq uses both positional variability due to differential sonication fragmentation 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.

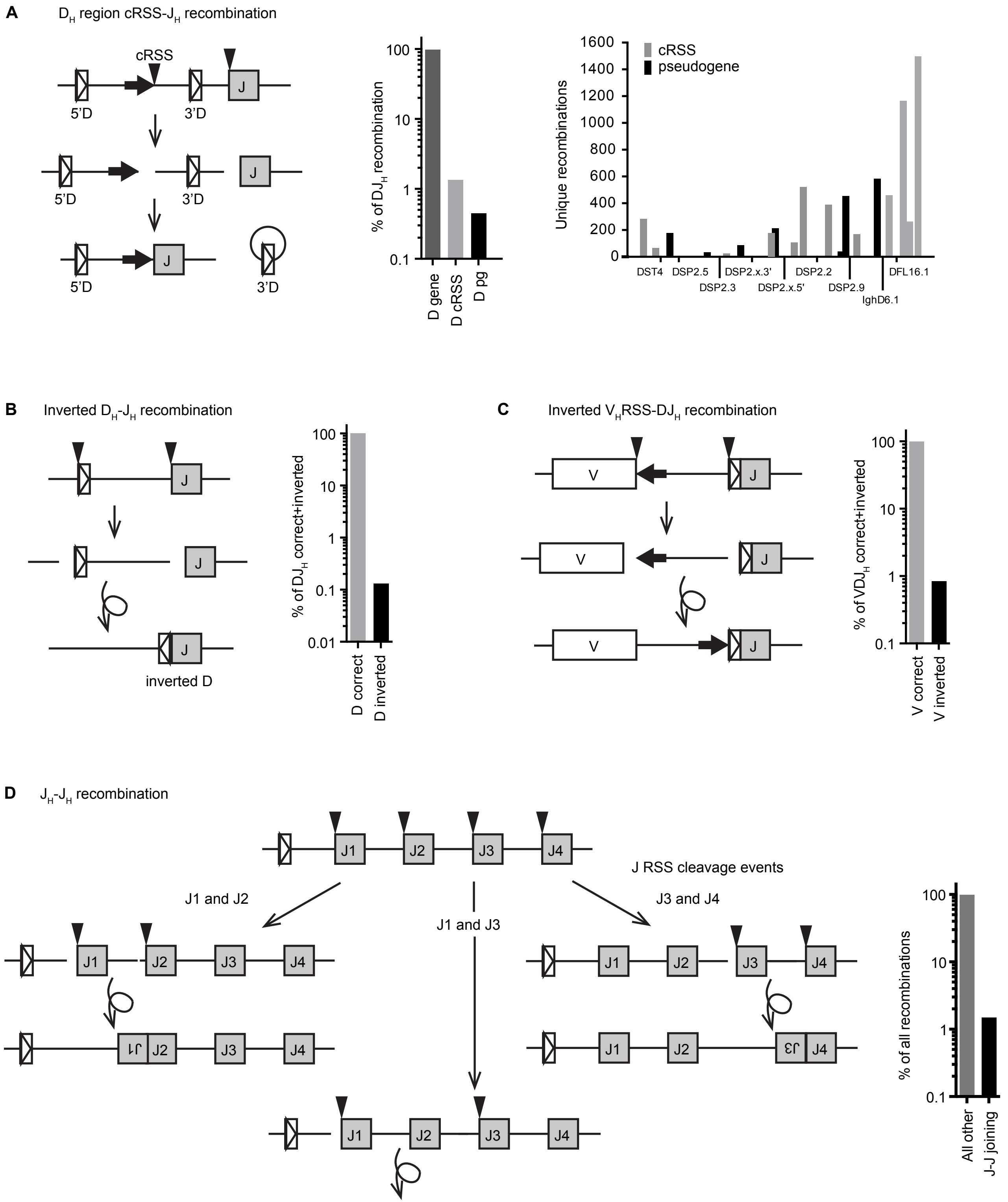


Figure S2 ***Related to Figure 2*** **Alternative and aberrant recombination events detected by VDJ-seq**
A) Low frequency recombination events were detected for D$_H$ pseudogenes and cryptic RSSs (cRSSs) within the D$_H$ region. **B)** D$_H$ gene inversion recombination as documented previously was a low frequency event. **C)** Inversion V$_H$RSS-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$_H$RSS-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.

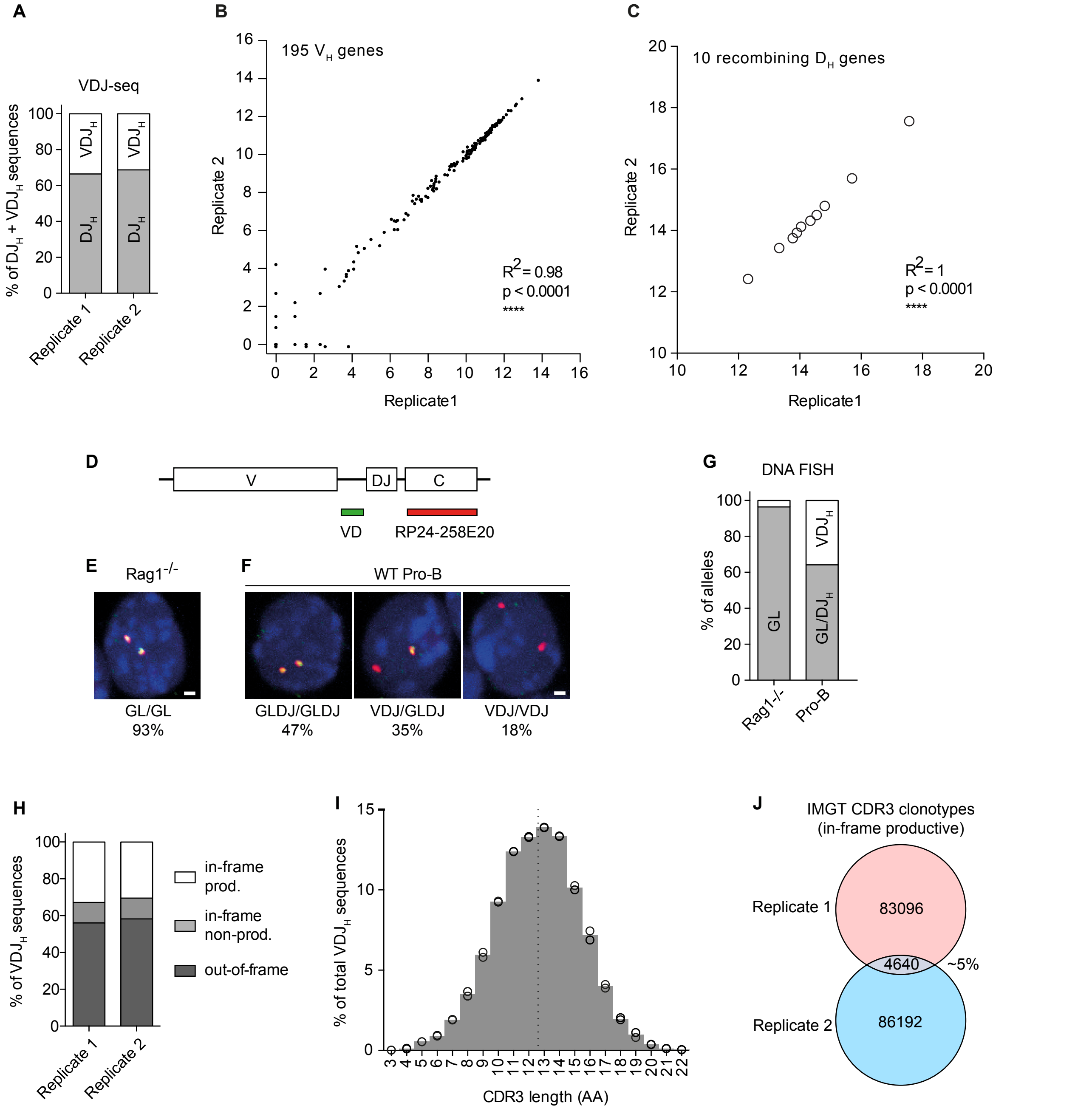


Figure S3 ***Related to Figure 1*** **Quality control of VDJ-seq**
A) Quantitation of VDJ$_H$ and DJ$_H$ rearranged sequences in WT pro-B cells by VDJ-seq. Sense orientation reads across the entire D$_H$ and V$_H$ regions were counted and each presented as a percentage of the total of D$_H$+V$_H$. Frequencies 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.

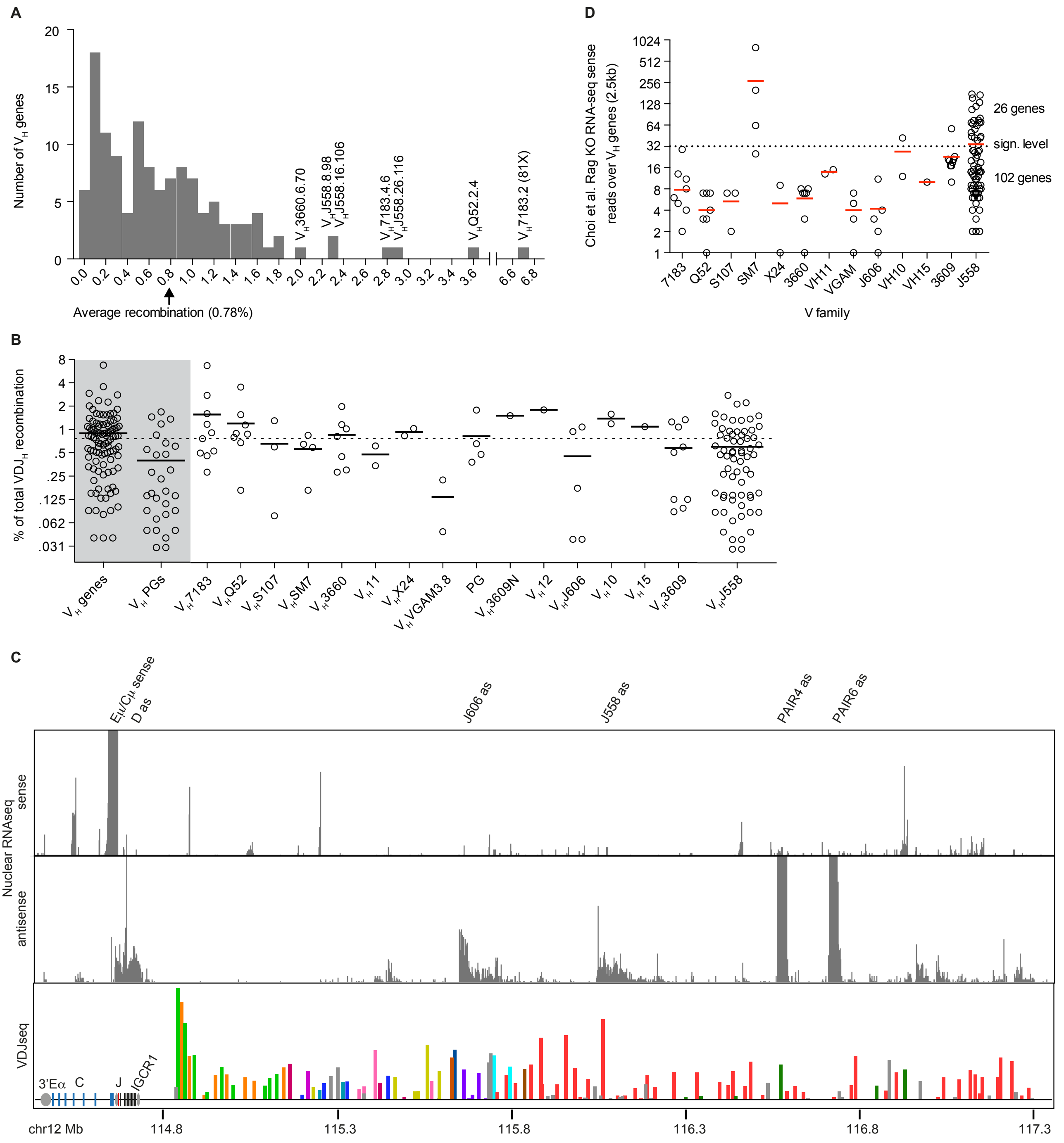


Figure S4 ***Related to Figure 2*** **Detailed VDJ-seq data**
A) Frequency distribution of recombination for V_H genes. The seven highest recombining genes are named. **B)** Gene (99), pseudogene/ORF (29) and per family usage for recombining V_H genes. Each gene is represented by a circle with the mean 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.

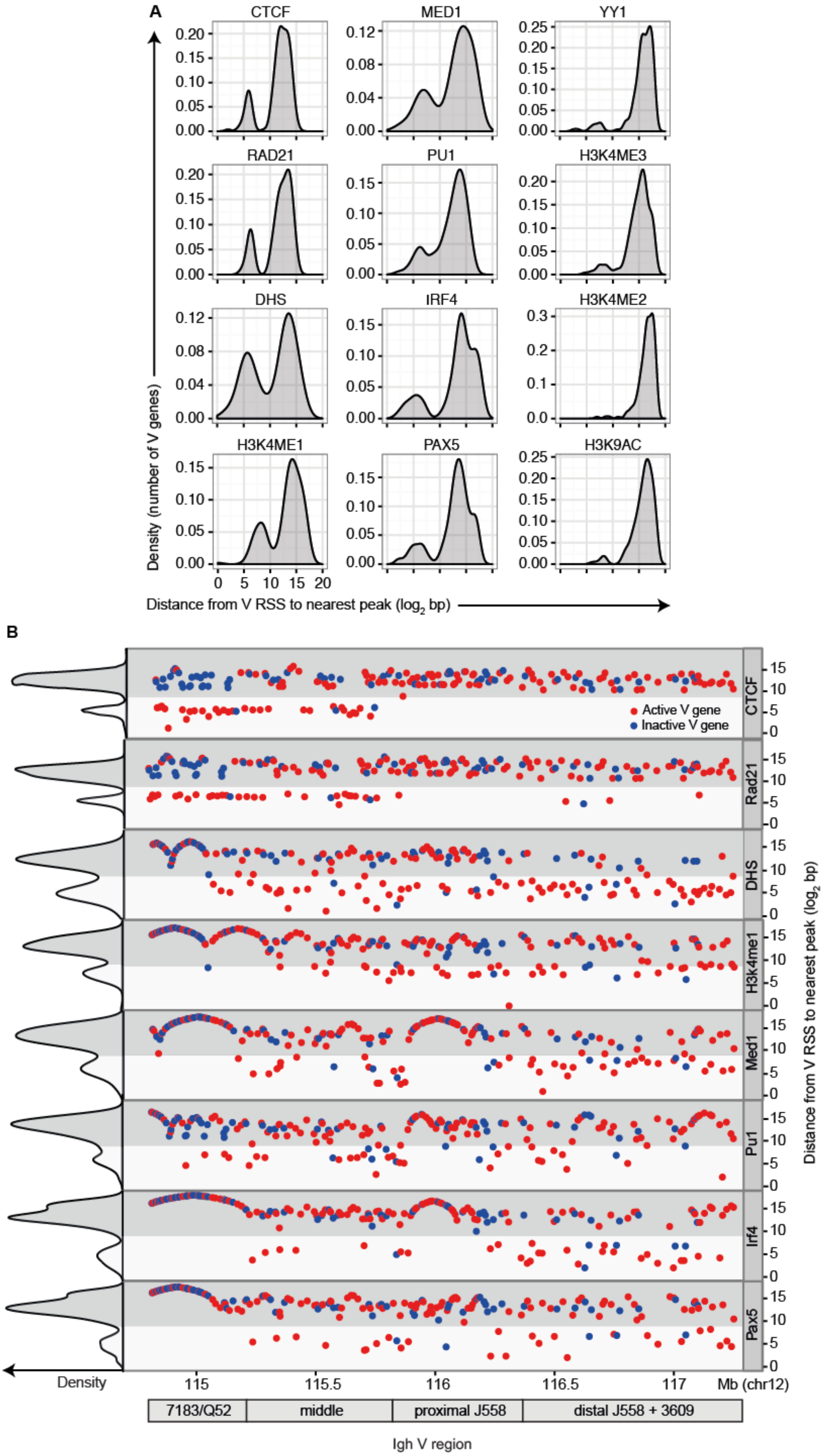


Figure S5 *Related to Figure 4* **Co-localisation of epigenetic factors with active V_H genes**

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.

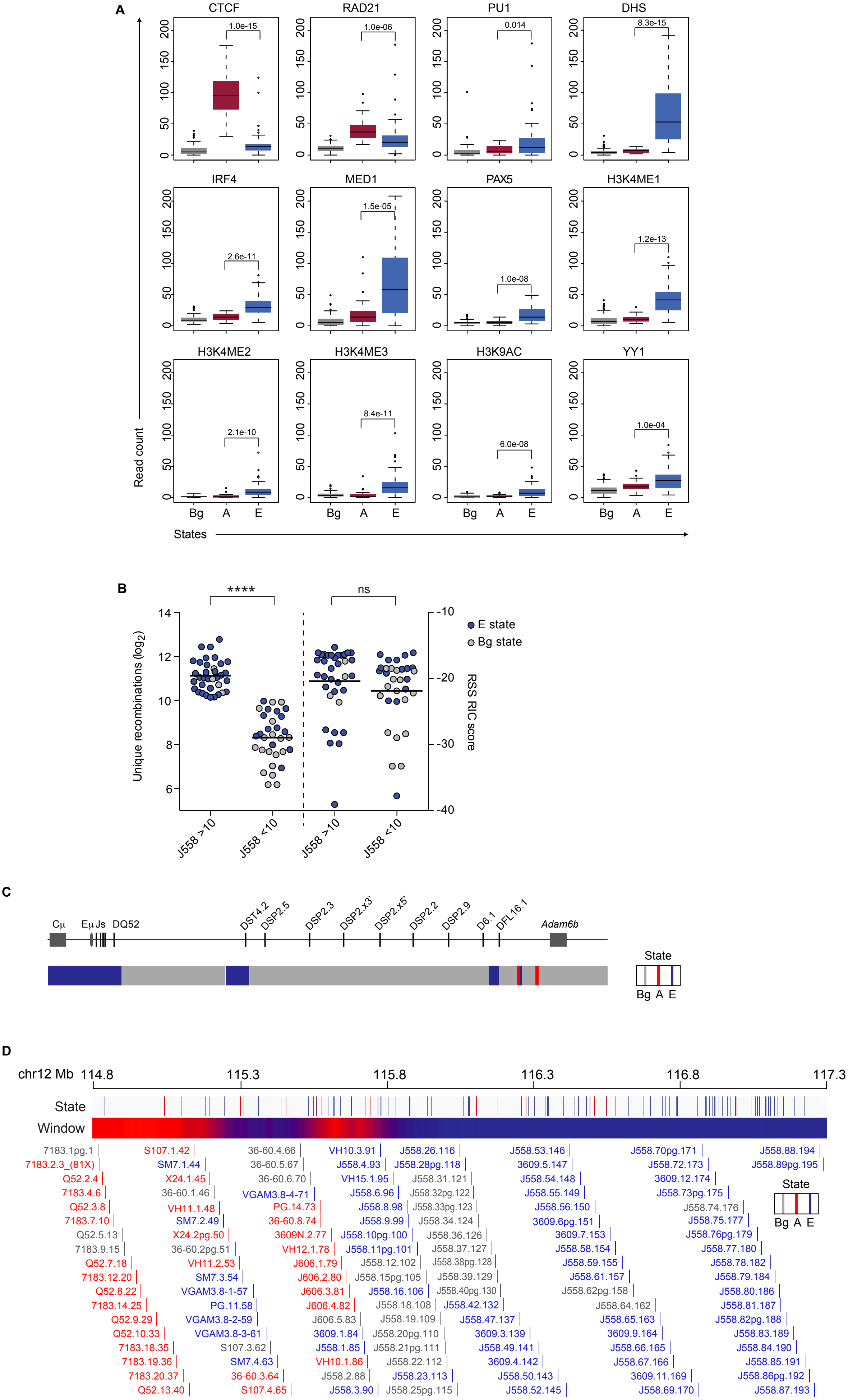

A
CTCF
RAD21
PU1
DHS
IRF4
MED1
PAX5
H3K4ME1
H3K4ME2
H3K4ME3
H3K9AC
YY1
1.0e-15
1.0e-06
0.014
8.3e-15
2.6e-11
1.5e-05
1.0e-08
1.2e-13
2.1e-10
8.4e-11
6.0e-08
1.0e-04
0
50
100
150
200
Read count
Bg
A
E
States
B
****
ns
14
12
10
8
6
-10
-20
-30
-40
Unique recombinations (log2)
RSS RIC score
E state
Bg state
J558 >10
J558 <10
J558 >10
J558 <10
C
Cμ
Eμ Js
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
7183.1pg.1
7183.2.3_(81X)
Q52.2.4
7183.4.6
Q52.3.8
7183.7.10
Q52.5.13
7183.9.15
Q52.7.18
7183.12.20
Q52.8.22
7183.14.25
Q52.9.29
Q52.10.33
7183.18.35
7183.19.36
7183.20.37
Q52.13.40
S107.1.42
SM7.1.44
X24.1.45
36-60.1.46
VH11.1.48
SM7.2.49
X24.2pg.50
36-60.2pg.51
VH11.2.53
SM7.3.54
VGAM3.8-1-57
PG.11.58
VGAM3.8-2-59
VGAM3.8-3-61
S107.3.62
SM7.4.63
36-60.3.64
S107.4.65
36-60.4.66
36-60.5.67
36-60.6.70
VGAM3.8-4-71
PG.14.73
36-60.8.74
3609N.2.77
VH12.1.78
J606.1.79
J606.2.80
J606.3.81
J606.4.82
J606.5.83
3609.1.84
J558.1.85
VH10.1.86
J558.2.88
J558.3.90
VH10.3.91
J558.4.93
VH15.1.95
J558.6.96
J558.8.98
J558.9.99
J558.10pg.100
J558.11pg.101
J558.12.102
J558.15pg.105
J558.16.106
J558.18.108
J558.19.109
J558.20pg.110
J558.21pg.111
J558.22.112
J558.23.113
J558.25pg.115
J558.26.116
J558.28pg.118
J558.31.121
J558.32pg.122
J558.33pg.123
J558.34.124
J558.36.126
J558.37.127
J558.38pg.128
J558.39.129
J558.40pg.130
J558.42.132
J558.47.137
3609.3.139
J558.49.141
3609.4.142
J558.50.143
J558.52.145
J558.53.146
3609.5.147
J558.54.148
J558.55.149
J558.56.150
3609.6pg.151
3609.7.153
J558.58.154
J558.59.155
J558.61.157
J558.62pg.158
J558.64.162
J558.65.163
3609.9.164
J558.66.165
J558.67.166
3609.11.169
J558.69.170
J558.70pg.171
J558.72.173
3609.12.174
J558.73pg.175
J558.74.176
J558.75.177
J558.76pg.179
J558.77.180
J558.78.182
J558.79.184
J558.80.186
J558.81.187
J558.82pg.188
J558.83.189
J558.84.190
J558.85.191
J558.86pg.192
J558.87.193
J558.88.194
J558.89pg.195
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.

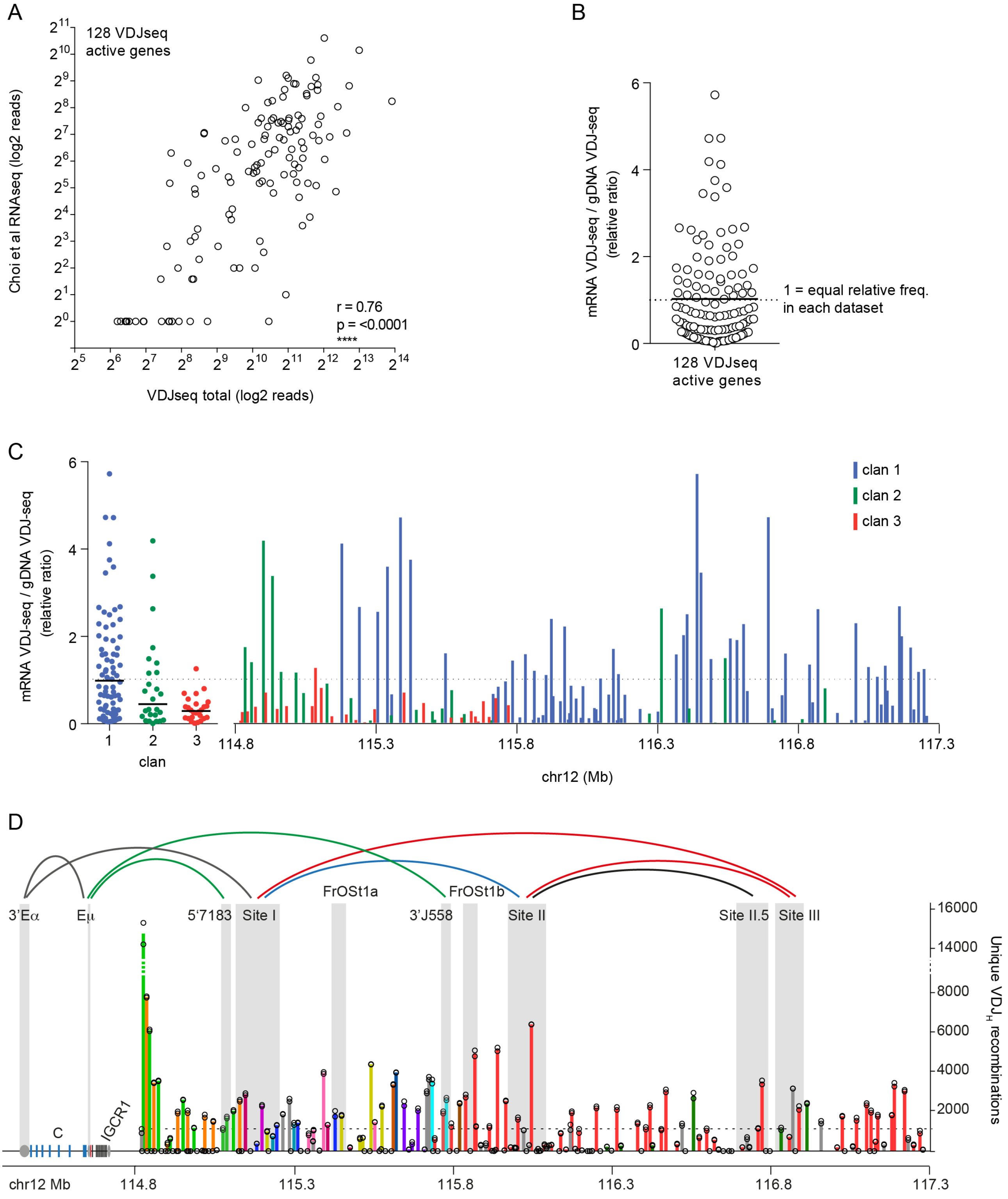


Figure S7 ***Related to Figure 2*** **Comparison of RNA and DNA-based repertoire analyses**
A) XY scatterplot comparing VDJseq with RNAseq data from (Choi et al., 2013) for the 128 V genes found to be active in VDJseq. A pseudocount of 1 was assigned to 14 genes that had zero reads in the Choi et al. 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