Deep Sequencing of the Murine IgH Repertoire Reveals Complex Regulation of Nonrandom V Gene Rearrangement Frequencies

Nancy M. Choi; ... et. al

https://doi.org/10.4049/jimmunol.1301279

Related Content

Cutting Edge: Developmental Stage-Specific Recruitment of Cohesin to CTCF Sites throughout Immunoglobulin Loci during B Lymphocyte Development

J Immunol (January,2009)

Chromatin Architecture, CCCTC-Binding Factor, and V(D)J Recombination: Managing Long-Distance Relationships at Antigen Receptor Loci

J Immunol (May,2013)

Epigenetic environment, transcriptional elements, and RSS sequences collectively regulate non-random VH gene usage at the mouse IgH locus. (P4394)

J Immunol (May,2013)
Deep Sequencing of the Murine Igh Repertoire Reveals Complex Regulation of Nonrandom V Gene Rearrangement Frequencies

Nancy M. Choi,* Salvatore Loguercio,† Jiyoti Verma-Gaur,* Stephanie C. Degner,* Ali Torkamani,‡ Andrew I. Su,§ Eugene M. Oltz,* Maxim Artyomov,§ and Ann J. Feeney* 

A diverse Ab repertoire is formed through the rearrangement of V, D, and J segments at the Igh (Igh) loci. The C57BL/6 murine Igh locus has >100 functional V_H gene segments that can recombine to a rearranged DJ_H. Although the nonrandom usage of V_H genes is well documented, it is not clear what elements determine recombination frequency. To answer this question, we conducted deep sequencing of 5’-RACE products of the Igh repertoire in pro-B cells, amplified in an unbiased manner. Chromatin immunoprecipitation–sequencing results for several histone modifications and RNA polymerase II binding, RNA-sequencing for sense and antisense noncoding germline transcripts, and proximity to CCCTC-binding factor (CTCF) and Rad21 sites were compared with the usage of individual V_H genes. Computational analyses assessed the relative importance of these various accessibility elements. These elements divide the Igh locus into four epigenetically and transcriptionally distinct domains, and our computational analyses reveal different regulatory mechanisms for each region. Proximal V_H genes are relatively devoid of active histone marks and noncoding RNA in general, but having a CTCF site near their recombination signal sequence is critical, suggesting that being positioned near the base of the chromatin loops is important for rearrangement. In contrast, distal V_H genes have higher levels of histone marks and noncoding RNA, which may compensate for their poorer recombination signal sequences and for being distant from CTCF sites. Thus, the Igh locus has evolved a complex system for the regulation of V(D)J rearrangement that is different for each of the four domains that comprise this locus. The Journal of Immunology, 2013, 191: 2393–2402.
regions of the genome (16). Importantly, H3K4me3 has also been shown to directly recruit RAG2 through its plant homeodomain finger (17, 18). The RAG1/2 recombinase binds to RSSs, which flank all V, D, and J gene segments. Although there is a consensus sequence for the heptamer and nonamer portions of the RSS (19, 20), individual RSSs often deviate from the consensus. Divergence from this consensus has been demonstrated to reduce recombination efficiency to varying extents (21, 22). Together, these suggest that a quality RSS and a suitable chromatin environment must both be provided for efficient recruitment of the RAG complex and effective catalytic activity to occur.

Large-scale, three-dimensional conformational changes of chromatin structure is another proposed regulatory mechanism of V(D)J rearrangement. The Igh locus is composed of multilooped rosettes, which have been shown by three-dimensional FISH to compact at the time of Igh gene rearrangement (23). This compaction of the structure of the locus will bring all V genes in much closer proximity to the DJ rearrangement to which one V gene will ultimately recombine. The insulator binding and chromatin looping protein CCCTC-binding factor (CTCF), and the cohesin complex are likely candidates for forming the rosette-like structure of the locus (24–29). When CTCF expression was knocked down, an extension in the spatial length of the Igh locus was observed (25). Pax5-activated intergenic repeats (PAIR) elements, which are bound by Pax5, E2A, and CTCF, are located in the distal intergenic regions of the VH locus (30). Extensive antisense ncRNA transcription starts from two of these PAIR elements, and these PAIR promoters directly interact with Eµ, presumably in a transcription factory (31). This movement of PAIR promoters, and possibly many other sense and antisense ncRNA promoters, to Eµ, which is within 2 kb of the DJ rearrangement, will directly result in locus compaction and will bring many V genes in proximity to the DJ rearrangement. In addition, there may well be protein–protein interactions of various transcription factors, which could also result in looping and locus contraction. The relative position of V genes in these chromatin loops may well influence their frequency of recombination.

Until the advent of high-throughput deep sequencing, it has been impossible to accurately assess the relative usage of all of the individual VH genes during V(DJ) recombination of the Igh locus. To date, no deep sequencing of the murine Igh locus has been performed. In this study, we took an unbiased approach by sequencing 5’-RACE products and quantitatively determining how frequently each VH gene was used. These data were compared with our chromatin immunoprecipitation–sequencing (ChIP-seq) results for histone modifications, RNA polymerase II, chromatin looping factors CTCF and cohesin, and the level of both sense and antisense ncRNA to obtain insight into the factors that influence VH gene rearrangement frequency. The quality of RSSs for the VH genes was also evaluated. Based on the epigenetic and transcriptional profile in the immediate vicinity of each VH gene, we divided the IgV locus into four domains, each containing different sets of VH gene families and different epigenetic characteristics. Computational analyses of all of the above factors in relation to recombination frequency demonstrated that each domain had different factors that influenced recombination frequency.

Materials and Methods
Mice and cell preparation
C57BL/6 and RAG1−/− mice were maintained in our breeding colony in accordance with protocols approved by The Scripps Research Institute Institutional Animal Care and Use Committee. Bone marrow cells were collected from 5- to 7-wk-old mice as described previously (32). CD19+ cells were isolated using anti-CD19–coated MACS beads (Miltenyi, Auburn, CA). Pro-B cells were sorted as B220+CD43+ IgM+ CD2+ on a BD FACSAria II.

RNA, cDNA preparation for 5’-RACE and RNA sequencing
RNA was extracted from pro-B cells using TRIzol (Life Technologies, Carlsbad, CA). Adapter ligation to RNA was performed with the 5’-RACE kit (Ambion; Life Technologies), and first-strand cDNA was prepared using the Transcriptor High Fidelity cDNA Synthesis Kit (Roche Diagnostics, Indianapolis, IN). A barcoded primer against Cµ and a primer that binds to the 5’ adapter were used together for amplification of the 5’-RACE cDNA products using Phusion High Fidelity DNA Polymerase (New England Biolabs, Ipswich, MA). Ten to 12 individual PCRs were pooled and gel purified with the QIAquick Gel Extraction Kit (Qiagen, Valencia, CA). The samples were then sequenced on a Roche 454 GS Junior at The Scripps Research Institute Next Generation Sequencing (NGS) Core Facility. For RNA-seq, RNA was prepared as described earlier and genomic DNA was eliminated using the genomic DNA wipeout buffer in the Quantitect Reverse transcription kit (Qiagen). A final purification of the RNA was performed with the RNaseasy kit from Qiagen. RNA samples were submitted to the NGS Core, where they were processed with the Ovation RNA-Seq System (NuGEN, San Carlos, CA), and sequenced on the Illumina HiSeq (San Diego, CA). RNA-seq data have been deposited in the Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo; accession no. GSE47885).

V, D, and J gene analysis
V gene identity was assigned by matching each sequence to a database of all known V gene sequences via MEGABLAST. At least a 100-bp alignment span was required for V gene assignment through a hierarchy of BLAST result parameters where the next parameter in the hierarchy was considered only if a read matched multiple genes under the previous parameter. The hierarchy, ordered to be more permissive to gaps in the BLAST alignments because of known homopolymer sequencing errors using 454 technology, was as follows: highest highest bit score, highest percent identity, longest alignment length, least number of mismatches, and finally, least number of gaps. Most reads were assigned by pure bit score. For the small number of reads for which the score matched multiple V genes exactly, the reads were given multiple assignments weighted according to the percentage of uniquely mapped reads for each gene in the data overall. Once the proper V gene was identified, rigorous alignments of reads to the known V gene sequence were performed using a Smith–Waterman alignment to extract the portion of the sequence downstream of the conserved cysteine at the start of the CDR3 region. This downstream region was used to identify J genes in similar fashion as V genes, filtering based on the same hierarchy but with a 20-bp alignment span requirement. Similarly, the downstream boundary of the CDR3 was extracted by identifying the conserved “WG” amino acids through a rigorous Smith–Waterman-based alignment. Finally, a D gene was assigned to the extracted CDR3 sequence again via BLAST without an alignment length threshold but following the same filtering hierarchy.

ChIP and ChIP-seq
Samples for ChIP-seq were prepared as previously described (25). ChIP-seq was performed on the Illumina HiSeq at the NGS core. Abs for CTCF, H3K4me2, H3ac, and H3K27me3 were purchased from EMD Millipore (Billerica, MA), H3K4me3 from Active Motif (Carlsbad, CA), and Rad21 from Abcam (Cambridge, England). All ChIP-seq data have been deposited in the Gene Expression Omnibus database (accession no. GSE47766, http://www.ncbi.nlm.nih.gov/geo).

Computational analysis of RSSs, accessibility elements, and recombination frequency
To determine the RSS and RSS information content (RIC) of each VH gene, we used a publically available Web-based program, DNAGrab (http://www.itb.cnr.it/rss) (33). For each of the chromatin and RNA-seq features analyzed, the signal intensity of each region was determined from the corresponding wig file by summing up normalized read counts. Signal intensities were then converted to log₂ values, and a pseudocount of one was added to avoid log₂(0) values. Similarly, rearrangement intensities were transformed into their log₂ values with an added pseudocount. VH genes with rearrangement >4 were labeled as “active,” and Random Forest classification (34) (R package RandomForest) was used to assess variable importance; that is, which of the features distinguish best between active and inactive genes. Cross-validated prediction performance of models with sequentially reduced number of predictors (ranked by variable importance) was used to suggest significant predictors. Regression analyses were performed on the 105 rearranging VH genes, on the full locus, and on the four
domains separately. Conditional Inference trees and Random Forests with Conditional Inference trees as base learners (35) (R package party) were used to generate the regression models. Pearson’s correlation coefficients were used in all cases to estimate which features correlate best with rearrangement intensity. A significance test of no association was used to generate $p$ values for the correlations.

**Results**

**Igh repertoire deep sequencing and the epigenetic landscape of the V$_H$ locus**

We analyzed the Igh sequences of cell sorter purified C57BL/6 pro-B cells to determine the initially generated repertoire before any selection could skew the primary rearrangements. To assess the repertoire in a completely unbiased manner, we performed 5′-RACE PCR. This involves ligation of a universal adapter to the 5′ end of cDNA followed by amplification with primers that anneal to the adapter and to the C region. C$_M$. In contrast, if one were to amplify genomic DNA, a multiplex PCR with a large pool of V$_H$ and J$_H$ primers is required, which can introduce significant bias caused by differential primer efficiencies, as well as primer cross-reactions. To obtain accurate sampling and deep coverage of the V$_H$ locus, we sequenced eight independent barcoded libraries of amplified 5′-RACE PCR. Ten to 12 PCRs were pooled for each library to minimize the sequencing of PCR duplicates. Duplicate sequences within each barcode were eliminated, and the V$_H$, D, and J$_H$ genes were identified. We obtained the sequences of 17,431 unique rearrangements.

We have also performed ChIP-seq for CTCF, Rad21 (a cohesin subunit), RNA polymerase II, and the histone modifications H3ac, H3K4me1, Rad21 (a cohesin subunit), RNA polymerase II, and the histone modifications H3ac, H3K4me1, and H3K27me3 (Fig. 1B). RAG1$^{-/-}$ pro-B cells on the C57BL/6 background were used for these ChIP-seq analyses so the entire Igh locus would be intact. The H3K4me1 data are from a previously published work (36). The rearrangement frequency for each V$_H$ gene is displayed on the bottom line of Fig. 1B, with the RNA-seq of sense and antisense ncRNA in RAG$^{-/-}$ pro-B cells above it.

The 195 V$_H$ genes within the Igh gene locus have been annotated previously (2). Genes were classified as potentially functional if they had no obvious defect and as pseudogenes if there was a major defect in the promoter, RSS, splice sites, or a stop codon in the coding region. Our experimental data agreed well with the predicted functionality. There were only 14 V$_H$ genes that had been classified as potentially functional that did not rearrange in our data set (Fig. 1B, magenta). Conversely, we found a small number of rearrangements among the genes previously classified as pseudogenes (Fig. 1B, orange), but pseudogenes can rearrange if the RSS is intact, as was the case for these genes.

**Usage of V, D, J gene segments**

In Fig. 2A, we displayed the number of unique rearrangements sequenced for each individual V$_H$ gene. The data are grouped by V$_H$ families, and each circle represents a V$_H$ gene. It can be easily observed that there is a wide range of usage of individual V$_H$ genes in the primary repertoire in all V$_H$ gene families. Fig. 2B shows J$_H$ gene utilization, and Fig. 2C shows D gene usage. We have previously demonstrated that a loop is made between E$_M$ and the set of CTCF/Rad21 sites 3.2 and 5.6 kb upstream of DFL16.1, and thus DFL16.1 and DQ52 are the closest D genes to the J$_H$ genes in three-dimensional space (25). It has also been shown that more repressive histone modifications are present at the DSP2 genes (37). Together, these appear to give DFL16.1 and DQ52 a distinct advantage because they were the most frequently rearranged D genes.

**Individual V$_H$ genes can be divided into four domains by their epigenetic characteristics**

To quantitate the local level of histone modifications and protein binding at each V$_H$ gene from all of these data sets, we calculated the signal intensity for the region from 1 kb upstream to 1 kb downstream of each V$_H$ gene. Because the leader exons are not all annotated, we estimated the leader and intron to be 200 bp 5′ of the ~300-bp coding region. It is evident in Fig. 3 that, for the active (rearranged) genes, the levels of histone modifications as well as ncRNA are not evenly distributed. The distal J558/3609 part of the locus and the proximal 7183/Q52 regions showed the greatest contrast, where the distal V$_H$ genes had much higher levels of active histone marks and ncRNA.

Based on the distribution of histone modifications and transcriptional activity, we divided the Igh locus into four domains, also considering the location of V$_H$ gene families. Domain 1 contains all 7183 and Q52 genes (genes 7183.1pg.1 through PG.7.41). Ten small V$_H$ families make up domain 2 (genes S107.1.42 through J606.5.83). The large J558 region is broken into two domains. Other than one 3609 V$_H$ gene and two small V$_H$ gene families, domain 3 contains 47 J558 genes. Domain 4 (genes 3609.2pg.138 to J558.89pg.195) contains the 3609 family interspersed with the remaining 42 J558 genes (Fig. 1A). Domain 1 is unique in that all but one of the active V$_H$ genes (7183.9.15) had a CTCF site in very close proximity to the RSS (9–68 bp). This domain also differed from other domains by being generally poor in active epigenetic marks other than low levels of H3ac and H3K4me1, low in ncRNA, and it uniquely contained H3K27me3. Domains 2 and 3 displayed more genes with higher H3ac and H3K4me1/2. In domain 4, almost all functional V$_H$ genes displayed reasonably high levels of H3ac and H3K4me1/2, and this domain had the highest frequency of V$_H$ genes associated with H3K4me3. For each domain, there were many fewer genes with active histone marks and ncRNA among the inactive genes.

The overall differences between the four domains can be appreciated in Fig. 4, which shows the mean signal intensity for each accessibility element for the active genes. In general, there were more active histone modifications and transcription at the distal domains, whereas domain 1 was noticeably lower in many parameters. Thus, the four domains of the Igh locus have very distinct epigenetic characteristics, suggesting that different factors are important for accessibility and rearrangement in the different parts of the Igh locus.

**V$_H$ gene proximity to CTCF and Rad21 sites**

We previously observed that CTCF sites were close to the RSS of functional V$_H$ genes in the proximal part of the locus but were intergenic in the large J558/3609 region (24). We also noted that most CTCF sites throughout the Igh locus colocalized with cohesin binding sites (24, 25). If V$_H$ genes were near the base of the rosettes that make up the three-dimensional structure of the Igh locus, then proximity to CTCF sites may impact the efficiency of recombination. We therefore expanded our previous studies by adding more sequencing depth to the CTCF and Rad21 ChIP-seqs, and determining the distance of the nearest CTCF and Rad21 binding sites from the coding end of every V$_H$ gene. These measurements were plotted against the genomic location of each gene (Fig. 5). The statistically significant CTCF and Rad21 binding peaks were determined by model-based analysis of ChIP-seq (false discovery rate ≤ 1%) (38), and distance was computed from the end of each V$_H$ gene to the peak summit. All but one of the rearranging V$_H$ genes in the 7183 and Q52 families (domain 1) had a CTCF site within 68 bp of the RSS, whereas most of the nonrearranging genes had CTCF and Rad21 sites more than 1 kb distant from them. All of the 7183 genes and most of the Q52 genes also had Rad21 bound to the CTCF sites. It is notable that the only active 7183 gene located far from a CTCF and Rad21 site, 7183.9.15, had a low rearrangement frequency, and two ap-
parently functional but nonrearranging 7183 genes also did not have CTCF sites nearby. Domain 2 containing the 10 small V_H families had a V_H family-specific distribution of distance from CTCF and Rad21; the functional S107, X24, V_H11, V_H12, and J606 genes all had nearby CTCF and Rad21 sites (100 bp), whereas the other families did not. In domains 3 and 4 that contain the large J558 family, all genes were \( \sim 1–32 \) kb from a CTCF site, with the exception of the 3 V_H10 genes that had a CTCF site within 100 bp of the RSS, and only 6 of the 89 J558 genes had nearby Rad21 binding. Thus, proximity to CTCF and Rad21 are of critical importance for domain 1 genes to rearrange, but proximity is not essential for most other V_H genes.

**Evaluation of RSS quality**

For a gene segment to recombine, it is essential that the RAG complex can bind to its RSS. The RSS consists of a consensus heptamer and nonamer that are separated by a spacer of 12 or 23 bp (19). Despite the high degree of homology, very few RSSs have both the consensus heptamer and nonamer. Recombination substrate experiments have shown that heptamers and nonamers closer to consensus provide higher recombination efficiency (5, 21). However, every variation from consensus has not been tested by substrate competition assays, and the sequence of the spacer can also affect recombination (39, 40). Hence, one cannot a priori estimate the rearrangement efficacy of each RSS.

To determine the RSS for each V_H gene, we used a publically available program that predicts RSSs (33). This program also provides an RIC score, which is an assessment of the quality of the predicted RSS (pass: RIC \( \geq \) 58.45) (41, 42). It predicted a 23-bp spacer RSS for 163 V_H genes and provided each with an RIC score (Supplemental Table I). Of the 145 genes that were assigned a passing score, 41 were inactive in our repertoire and had a lower average RIC score (36.58 compared with 27.34 for active genes). Of the 105 active genes, 104 were assigned a passing score, and a single gene, albeit with low recombination frequency (3609.5.147), was deemed to have a failed RSS. Overall, 76.7% of
predictions on whether a gene was able to rearrange were in concordance with the deep sequencing results.

We previously examined the relative efficiencies of murine $V_H$ gene RSSs by transiently transfecting competition recombination substrates into cell lines (5, 39). Two major groups of 7183 RSSs in an $Ig_H$ strain were tested in these experiments. The heptamer of Group I (7183-I) had one base change from consensus, whereas the heptamer of Group II (7183-II) genes was identical to the consensus, and both groups had the same single change from consensus in the nonamer. The unique RSS for 7183.2.3 (81X) was examined because of its high level of rearrangement. The RSS for S107.1.42 and a common J558 RSS (J558-II) were also included in the assay. A similar competition assay was performed by Connor et al. (43) comparing the consensus RSS with a closely related common J558 RSS (J558-II). The relative recombination efficiency for each substrate was calculated (Fig. 6, left panel). It can easily be seen that the J558 RSSs supported the lowest frequency of rearrangement. Thus, the proximal 7183 genes have the rearrangement advantage of having much better RSSs than the large distal J558 family.

In C57BL/6 mice, which have the $Ig_H$ haplotype, there is only one 7183 V gene (7183.9.15) with an identical RSS to the 7183-II sequence and four with an RSS identical to 7183-I. For the tested J558 RSSs, there are four genes that have the exact J558-I RSS and five have the J558-II RSS. When the recombination efficiencies determined in substrates for these RSSs were compared with the RIC scores predicted for them in silico, it could readily be seen that there is little agreement between the two data sets (Fig. 6, left and middle panels). The 7183-II RSS was assigned the lowest RIC score despite being closest to the consensus RSS and with the highest level of recombination in the competition plasmid substrates. Because recombination substrate assays are the gold standard for assessing the quality of an RSS, this demonstrates that the RIC score is not accurate for predicting the relative effectiveness of RSS. When the RIC and recombination substrate assay results were compared with our deep sequencing results (Fig. 6, right panel), the poor recombination of J558 genes was much better predicted by the recombination substrate results with the exception of one highly rearranged J558 gene. Also in Fig. 4, the panel showing average RIC scores for the four domains illustrates that overall the RIC scores do not agree with recombination frequency. Thus, RIC scores do not accurately predict the quality of an RSS, and recombination substrate assays are far more accurate for determining relative efficacy of an RSS. However, it can be seen in Fig. 6 that genes with identical RSSs rearranged at different frequencies in vivo, as we had previously observed with the smaller $V_H$ gene data set (5). Thus, although the RSS probably plays an important role in individual $V_H$ gene rearrangement, clearly factors other than the RSS must also have significant influence on rearrangement frequency in vivo.

**Epigenetic and accessibility elements influence recombination of $V$ genes**

Histone modifications, both sense and antisense ncRNA, as well as factors that regulate the three-dimensional structure of chromatin such as CTCF and Rad21, all have the potential to influence the level of gene usage. Although it is known at individual genes that these factors are influential, the relative weight of importance or the collective effect of all these elements is largely unknown. We therefore addressed these questions by computational classification and regression analyses.

Signal intensities for all histone modifications and sense/antisense ncRNAs for each gene (Fig. 3), distance to nearest CTCF and Rad21 peaks (Fig. 5), genomic position along the locus (in order of DJH proximal to distal), and RIC scores (Supplemental Table I) were used to train a Random Forest classifier against the recombination frequency of the C57BL/6 $V_H$ genes. Although RIC scores are not nearly as accurate as recombination substrate assays, one can calculate an RIC for each V gene, whereas the large number of $V_H$ gene precludes making recombination substrates for all V genes. In brief, the classifier builds a model that evaluates the most informative parameters in predicting whether a gene would be active or not based on all provided information. This model correctly predicted 68 of 84 inactive genes (classification error = 0.19) and 98 of 105 active genes (classification error = 0.067), with an overall out-of-bag estimated error of 12.17% (Fig. 7A, Supplemental Fig. 1A). When using only the top four parameters that had a significant contribution for model prediction, the error rates are similar (error rate = 11.11%), suggesting that the other parameters have only minor influence on classification prediction. Having a functional RSS was determined the most informative parameter, but the accuracy of prediction was better when H3K4me1 and distance to the closest CTCF site were included in the analysis. Therefore, although not reliable at all in reporting the relative level of rearrangement, RIC score (in the absence of a more reliable recombination substrate score) was the best predictive criterion in determining whether a $V_H$ gene was active.

Advancing beyond the “yes” or “no” question to determine whether different parameters could cooperatively predict the relative frequency of recombination of active genes, we performed linear and tree-based Random Forest regression analyses on the full locus and individual domains on the active $V_H$ genes only. However, model qualities (estimated with root mean square error and percentage variance explained after a 10-fold cross-validation procedure) were generally low except for domain 3, in which the levels of H3 acetylation and H3K4me1 were predictive (Supplemental Fig. 1C, 1D). We then performed Pearson’s correlation analyses on the active $V_H$ genes for all 12 parameters versus the recombination frequency of each $V_H$ gene. Three parameters had a significant influence on rearrangement, clearly factors other than the RSS. These parameters were H3 acetylation, H3K4me1, and RIC score (Fig. 7B). Because the data in Fig. 3 suggested that the locus might be governed by different mechanisms in the four domains, we also examined each domain separately. Because all genes that rearranged in domain 1, with one exception, were less than 100 bp from a CTCF site,

![FIGURE 2. V, D, and J gene rearrangement frequencies in pro-B cells. (A)](http://journals.aai.org/jimmunol/article-pdf/191/5/2393/1382232/1301279.pdf) The rearrangement frequencies of $V_H$ genes are plotted by gene families in a log₂ scale. Each circle represents an individual $V_H$ gene. Horizontal bars show the mean value for each family. (B) Percentage of $J_H$ gene rearrangements. (C) $D_H$ gene rearrangement frequency in gray columns is superimposed onto the genome browser track for CTCF ChIP-seq (black).
relative proximity to CTCF was not a distinguishing characteristic. Only the relative position within the locus had a statistically significant correlation in domain 1, and none correlated for domain 2 (Fig. 7B). In contrast, three parameters significantly correlated with recombination frequency in domain 3, of which H3 acetylation and distance to the nearest Rad21 site were the strongest parameters. In domain 4, H3K4me2/3 had the greatest correlation followed by distance to Rad21 (Supplemental Fig. 1B). Thus, each domain had distinct parameters influencing the extent of gene rearrangement.

Discussion
In this report, we determined the mouse IgH preselection repertoire of C57BL/6 pro-B cells. To our knowledge, this is the first detailed description of the murine IgH repertoire by next generation sequencing. Because of the size and complexity of the IgH locus, we took an approach that amplified the rearranged sequences in a completely unbiased manner using 5'-RACE to amplify the cDNA. This allowed us to bypass potentially serious concerns of unequal amplification efficiency if we were to amplify each individual V<sub>H</sub> gene from genomic DNA by multiplex PCR. The main concern with performing deep sequencing on RNA rather than genomic DNA is that cells may vary significantly in the amount of Ig mRNA. However, although this is formally a possibility, it is unlikely that this would be much of an issue in pro-B cells.

Although it is well established that RSSs closer to the consensus support more rearrangement in recombination substrates (the gold standard for RSS quality), and also for some V genes in vivo (39, 44), we have also clearly shown previously that genes with identical RSSs, such as 7183 genes, can rearrange with different frequencies in pro-B cells (5). In this article, we also demonstrated that the efficiency measured by substrate competition assays did not always correlate with the actual recombination frequency in vivo for a wider variety of V<sub>H</sub> genes. Thus, factors in addition to RSS quality, presumably epigenetic, influence rearrangement frequency in vivo. In this study, we returned to address this long-standing question with our new data from deep sequencing the IgH repertoire, plus our ChIP-seq and RNA-seq data. We determined the local level of each parameter for each V<sub>H</sub> gene by calculating the level of histone posttranslational modification or the amount of ncRNA within a 2.5-kb region containing each V<sub>H</sub> gene coding.
region plus 1 kb of upstream and downstream flanking DNA; then we used computational analyses to assess the influence of each of these parameters, individually or in concert, on rearrangement frequency.

Although histone acetylation is far higher on J genes than on V genes (45), nonetheless V genes in loci that are poised for, or are undergoing, rearrangement showed higher acetylation than in loci that are not undergoing rearrangement (45–48). Furthermore, when comparing closely related V H genes, acetylation has been shown to be higher on the V H gene segments that rearrange more frequently (48, 49). In this article, we demonstrate that essentially every rearranging gene is associated with some level of histone acetylation; thus, histone acetylation may be considered to be strongly associated with accessibility.

H3K4me3 was of particular interest because the plant homeo-domain finger of RAG2 specifically binds to this modification; thus, it can act to directly recruit the RAG complex (17, 18, 50). H3K4me3 is associated with actively transcribed genes, and thus is likely to be a direct functional consequence of the widespread occurrence of ncRNA at receptor loci as they become poised for rearrangement. High levels of germline transcription and H3K4me3 are only present on J genes (45, 50); however, there is a low level of ncRNA throughout the large V H gene portion of receptor loci in pro-B cells (13). ncRNA is easily detected at J558 genes by PCR, whereas ncRNA at proximal genes is much more difficult to amplify, suggesting much lower levels (7, 13). In accord with this older data, our RNA-seq data show much higher levels of sense and antisense RNA in the 2.5-kb environs of J558 genes in domains 3 and 4 as compared with the low levels seen in domains 1 and 2 (31). Our ChIP-seq demonstrates much higher levels of H3K4me3 on the V H genes in domain 4 than on other domains, and H3K4me3 is essentially absent in domain 1 (Figs. 3, 4). Because H3K4me3 recruits RAG, one would predict that the level of H3K4me3 might show a positive correlation with the rearrangement frequency of individual V H genes, but this correlation was statistically significant only in domain 4. We have previously shown that there are high levels of intergenic antisense transcription at 3 locations in domain 4 (PAIR4, PAIR6, and PAIR11) and, as expected, those sites also have high levels of H3K4me3 and H3Ac (31). Thus, it could be that high levels of transcription and/or H3K4me3 at intergenic sites are influencing the rearrangement of neighboring V H genes. However, there are no active V H genes within ~20 kb of any of those PAIR elements. It should be noted that the most frequently rearranging gene in domain 4 is 55 kb downstream of PAIR4, and that epigenetic parameters outside of the local 2.5-kb environment of each V H gene were not factored into our computational analyses. Excluding these few regions with high levels of antisense transcription, we observed that the H3K4me3 mark was predominantly located over the genes and not in intergenic regions.

Interestingly, the deposition of all of the activating histone modifications (monomethylation, dimethylation, or trimethylation of lysine-4 histone H3 and H3ac) were skewed such that higher levels were seen toward the distal end of the locus, with this being most pronounced for H3K4me3 and least pronounced for H3ac (Figs. 3, 4). However, H3K4me1 and H3K4me2 were only positively correlated with rearrangement frequency in domains 3 and 4, respectively. Domain 1 had essentially no H3K4me2 or H3K4me3, and the level of H3K4me1 was low. In contrast with the distal regions, the 7183/Q52 region was the only domain that bore the repressive H3K27me3 modification as we previously reported using ChIP-chip (32), and this was shared with both active and inactive genes. In general, the nonrearranging pseudogenes had lower levels of or no active H3K4 methyl marks in domains 1–3, but many nonfunctional genes in domain 4 had some level of the H3K4 methylation (Fig. 3). Sense transcription was also much lower on the nonrearranging genes. This may suggest that some level of ncRNA is required for a gene to be functional, and that selective pressure to maintain ncRNA is gone when a gene becomes a pseudogene.

Changes in higher order chromatin structure play a critical role in V(D)J recombination. It is known that receptor loci undergo compaction at the time of rearrangement (51–53). Chunjunwala et al. (23) proposed that the locus was organized into three rosette-like structures in pro-pre-B cells that compacted into a smaller three-dimensional space in pro-B cells such that the distal and proximal regions were equidistant from the J H genes. The CTCF and cohesin proteins are likely candidates for bringing the Igh locus into this rosette structure (24). We demonstrated several years ago that CTCF and Rad21 bound at many sites throughout the Igh locus, and that the sites were close to the RSS for the proximal V H genes, but were intergenic in the distal J558-containing domains 3 and 4 (24). In this study, we quantitated the distance from the RSS to the nearest CTCF andRad21 binding site for all V H genes, and the bimodal distribution of distance to CTCF and Rad21 is striking. V H genes either have CTCF and
Rad21 sites within 100 bp of the RSS (domain 1 and half of the small V_H families in domain 2), or the sites are from 1–32 kb distant from the V_H genes. We have previously demonstrated by 3C that the CTCF sites just upstream of the most V-proximal D_H gene interact with E_m, which is located only 1–2.5 kb from the J_H genes that have high levels of RAG1/2 binding (50). Thus, for the proximal part of the locus, the relative proximity of the RSS to CTCF sites may create a close cluster of RSSs in three-dimensional space as previously suggested (25, 34). Therefore, the proximal V_H genes may not need any histone modifications or extensive ncRNA to facilitate long-range movement or to recruit RAG2, and it may be that the proximity to a CTCF site is sufficient. It is striking that within domain 1, almost all of the nonrearranging V_H genes do not have CTCF sites near their RSS, whereas all rearranging VH genes in the 7183 and Q52 families, with the exception of one low-frequency rearranging V_H gene, have nearby CTCF binding sites (25). Many of the nonrearranging domain 1 genes do not have good-quality RSSs, and thus this is likely to preclude rearrangement. However, some of the V_H genes without nearby CTCF binding do have fairly good RSSs, and in particular, 7183.16.27 and Q52.11.34 have good RSSs, yet no rearrangements were found in our data set. Thus, it is likely that proximity to CTCF is a requirement for domain 1 V_H genes to contribute to the repertoire.

Even though most of the V_H genes in domains 3 and 4 have CTCF sites >1 kb up to ~32 kb away from the RSS (with the exception of V_H10 genes and one J558 gene), the distance to the nearest Rad21 site varies over a range of ~10 to ~45 kb. Our computational results show a negative correlation between distance to Rad21 and rearrangement frequency in those two domains (Fig. 7B), as can also be appreciated in the scatter plot (Supplemental Fig. 1B). Thus, even for the distal V_H genes without nearby CTCF binding do have fairly good RSSs, and in particular, 7183.16.27 and Q52.11.34 have good RSSs, yet no rearrangements were found in our data set. Thus, it is likely that proximity to CTCF is a requirement for domain 1 V_H genes to contribute to the repertoire.

To take a more integrated and comprehensive approach, we performed computational analyses that assessed the relative importance of all the elements that could regulate accessibility of a given gene and recombination. In determining whether a gene rearranges, the significance of a functional RSS was clear (Fig. 7A). In this regard, it should be noted that the 14 genes that were deemed potentially functional but that did not rearrange in our data had RIC scores that were lower, on average, than the active genes overall (averages of 0.05). Scatter plots are shown for the best correlating parameter in relation to rearrangement frequency; all log2 transformed.

FIGURE 6. RSS competition substrate assays and RIC scores do not agree. The left panel is the relative recombination efficiency assessed by plasmid-based recombination substrate assays performed previously (5, 43). 7183-II RSS is an exact match to the consensus RSS heptamer; 7183-I is a common variant RSS in the 7183 family. 7183.2.3 and S107.1.42 are individual V genes. J558-I and -II are two common variants of J558 RSSs. Substrate competition assay for J558-II was performed by Connor et al. (43) and values are in white. Middle panel shows the RIC scores for all the V_H genes that have the identical RSS to the recombination substrates. Right panel shows the deep sequencing recombination frequency for the same individual VH genes.
Thus, these domain borders are likely to be boundaries in the three-dimensional structure of the locus as well. We performed computational analyses of the individual domains to determine whether the four domains had different rules that governed rearrangement. Domain 1 separated itself by having proximity to DJH as a unique significant criterion, as we previously demonstrated for the 7183 gene family in an IgH strain (5). Domain 2 displayed no clear correlative factors; however, this may be explained by the heterogeneity of domain 2, as there are many different small VH families interspersed in this domain. In domains 3 and 4, active histone modifications and being closer to a Rad21 binding site generally correlated with higher levels of VH gene usage. We were unable to observe a strong correlation with sense or antisense transcripts and recombination in any of the domains, which we would have expected at least for domain 4. However, Pax5- and YY1-dependent antisense RNA in the intergenic region of domain 4 were previously shown to be important for bringing the distal genes into close proximity to Eμ and its adjacent DJH rearrangement, possibly in a common transcription factory (31). Because we limited our analysis to a 2.5-kb region surrounding each VH gene, the large-scale structural changes in the three-dimensional structure of the IgH locus that occur at least partially as a result of antisense transcription are not factored into these computational analyses.

Although accessibility of chromatin as characterized by active histone modifications and robust transcriptional activity may provide a favorable environment that recruits RAG1/2 and facilitates recombination, this does not completely explain the nonrandom VH gene usage in the mouse IgH repertoire. The active VH genes in domain 1, which have adjacent CTCF/Rad21 binding sites, have low levels of ncRNA and are fairly devoid of histone marks other than H3ac. Thus, the position of the V genes near the bases of the CTCF-generated rosettes, coupled with good RSSs in general and moderate levels of histone acetylation, appears to be sufficient to support good rearrangement. At the other extreme, the J558 genes in domain 4 have very poor RSSs in general as determined in recombination substrate analyses, and are the most distant from the JH genes in linear distance. However, these VH genes have the highest level of H3K4me3, which will recruit RAG2. In addition, the three highest sites of antisense transcription are in domain 4, and this transcription moves domain 4 into proximity with the JH genes as determined by 3C (31). Thus, a strong RSS such as that of a 7183 gene may compensate for an overall lower level of active histone marks and low transcriptional activity, whereas the poorer RSSs for J558 genes were compensated by higher histone marks and higher ncRNA. This complex system creates an effective strategy for pro-B cells to use all the genomic regions to attain high diversity of the IgH repertoire. Thus, the IgH locus has evolved a complex system for the regulation of V(D)J rearrangement that is different for each of the four domains that comprise this locus.

Acknowledgments

We thank Amy Baumgart for excellent technical assistance, and Drs. David Nemaze and Takayuki Ota for helpful discussions.

Disclosures

The authors have no financial conflicts of interest.

References


