JBC Papers in Press. Published on July 8, 2019 as Manuscript RA118.007348
The latest version is at http://www.jbc.org/cgi/doi/10.1074/jbc.RA118.007348

The murine IgH locus contains a distinct DNA sequence motif for the chromatin regulatory factor CTCF

David N. Ciccone[1, 3*], Yuka Namiki[1*], Changfeng Chen[1,4], Katrina B. Morshead[1,5], Andrew L. Wood[2,9], Colette M. Johnston[2,10], John W. Morris[1,7], Yanqun Wang[1,8], Ruslan Sadreyev[1], Anne E. Corcoran[2], Adam G.W. Matthews[1,6], and Marjorie A. Oettinger[1#]

From the [1]Department of Molecular Biology, Massachusetts General Hospital, and Department of Genetics, Harvard Medical School, Boston, MA 02114; [2]Lymphocyte Signalling and Development, Babraham Institute, Babraham Research Campus, Cambridge CB22 3AT, UK; [6]Department of Biological Sciences and Program in Biochemistry, Wellesley College, Wellesley, MA 02481

Running title: *CTCF binds evolutionarily conserved sites in the IgH locus*

[3]Present Address: Merck, Boston, MA 02115
[4]Present Address: BioSciences, Inc., Haidan District, Beijing, PR China
[5]Present Address: Genentech, South San Francisco, CA, 94080
[7]Present Address: Synopsys, Inc., Burlington, MA 01803
[8]Present Address: Novartis Institutes for Biomedical Research, Cambridge, MA 02142
[9]Present Address: Kymab Ltd, Babraham Research Campus, Cambridge CB22 3AT, UK
[10]Present Address: Crescendo Biologics, Babraham Research Campus, Cambridge CB22 3AT, UK
[*]These authors contributed equally to the work

#To whom correspondence should be addressed: Marjorie A. Oettinger: Department of Molecular Biology, Massachusetts General Hospital, and Department of Genetics, Harvard Medical School, Boston, MA 02114; Oettinger@molbio.mgh.harvard.edu; Tel. (617) 726-5967; Fax. (617) 726-5949

Keywords: DNA recombination, chromatin, cellular immune response, chromatin immunoprecipitation (ChIP), chromatin regulation, CCTC-binding factor (CTCF), V(D)J, antigen receptor, adaptive immunity, DNA binding

ABSTRACT

Antigen receptor assembly in lymphocytes involves stringently regulated coordination of specific DNA rearrangement events across several large chromosomal domains. Previous studies indicate that transcription factors such as paired box 5 (PAX5), Yin Yang 1 (YY1), and CCCTC-binding factor (CTCF) play a role in regulating the accessibility of the antigen receptor loci to the V(D)J recombinase, which is required for these rearrangements. To gain clues about the role of CTCF binding at the murine immunoglobulin heavy chain (IgH) locus, we utilized a computational approach that identified 144 putative CTCF-binding sites within this locus. We found that these CTCF sites share a consensus motif distinct from other CTCF sites in the mouse genome. Additionally, we could divide these CTCF sites into three categories: intergenic sites remote from any coding element, upstream sites present within 8 kb of the V_H-leader exon, and recombination signal sequence (RSS)-associated sites characteristically located at a fixed distance (~18 bp) downstream of the RSS. We noted that the intergenic and upstream sites are located in the distal portion of the V_H locus, whereas the RSS-associated sites are located in the D_H-proximal region. Computational analysis indicated that the prevalence of CTCF-binding sites at the IgH locus is evolutionarily conserved. In all species analyzed, these sites exhibit a striking strand-orientation bias, with > 98% of the murine sites being present in one orientation with respect to V_H gene transcription. Electrophoretic mobility shift and enhancer-blocking assays and ChIP–chip

analysis confirmed CTCF binding to these sites both *in vitro* and *in vivo*.

During the vertebrate adaptive immune response, B cells and T cells play an essential role in clearing pathogens from the host organism. Specific recognition of these pathogens relies on the strikingly diverse binding-specificities encoded by the antigen receptors – the B cell receptor (BCR) and T cell receptor (TCR), respectively – expressed on the surface of these lymphoid cells. Antigen receptor genes encoding receptors with distinct specificities are generated from component gene segments – termed variable (V), diversity (D), and joining (J) gene segments – via the V(D)J recombination process. V(D)J recombination is initiated when the V(D)J recombinase – a heterotetrameric complex containing two RAG1 and two RAG2 subunits (1,2) – introduces DNA double-strand breaks at the junctions between the two gene segments and their flanking recombination signal sequences (RSS). The recombination reaction is completed by the ubiquitously expressed non-homologous end-joining (NHEJ) machinery, which joins the two coding ends to form a complete coding sequence, while in parallel joining the two signal ends to each other.

The assembly of BCR and TCR genes via V(D)J recombination is tightly regulated in vivo. Rearrangement events are restricted to particular cell lineages and stages, such that immunoglobulin (Ig) loci are only fully rearranged in B cells, while TCR genes are completely assembled only in T cells. In mice, rearrangement occurs in a preferred temporal order, with Ig heavy chain loci rearranging prior to light chain loci. Within a heavy chain locus, D_H-to-J_H joining occurs prior to V_H-to-DJ_H rearrangement. Rearrangement is also allele-restricted; while D_H-to-J_H rearrangement occurs on both alleles, only one productive V_H-to-DJ_H rearrangement (to produce a functional heavy-chain gene) and one productive V_L to J_L rearrangement (to produce a functional light-chain gene) occurs per cell.

In order to appropriately regulate V(D)J recombination, developing lymphocytes must control the accessibility of the antigen receptor loci to the recombinase machinery. Indeed, a variety of alterations at the antigen receptor loci have been found at different developmental stages. For example, germline transcription, genic and intergenic antisense transcripts, specific histone modifications, nucleosome positioning, monoallelic DNA methylation, nuclear repositioning of antigen receptor alleles, reversible DNA contraction, and chromosomal looping of domains within receptor loci have all been described (3-8).

Several studies have focused on how trans-acting proteins – including transcription factors (such as Pax5 (9-11) and YY1 (12)), chromatin remodeling complexes (such as SWI/SNF (13,14)), and histone-modifying enzymes (such as G9a (15) and Ezh2 (16)) – contribute to the developmental regulation of V(D)J recombination by modifying the chromatin structure of the antigen receptor loci. CTCF – a ubiquitously expressed nuclear protein that is involved in many cellular processes – is a particularly interesting transcription factor that has been localized to numerous sites across the murine immunoglobulin heavy chain (IgH) locus (9,17,18). Some of these CTCF sites have been functionally analyzed via targeted deletion: a portion of the 3' regulatory region (3' RR) of the IgH locus that contains several CTCF and Pax5 binding sites modestly affects V(D)J recombination and contraction of the locus (19); a regulatory region located between the VH and DH gene clusters that contains two CTCF binding elements contributes to the developmental regulation of V(D)J recombination (20,21). Similarly, reducing or eliminating CTCF expression also appears to affect the immunoglobulin loci: a global reduction in CTCF expression results in increased antisense transcription (17) as well as decreased IgH locus contraction (17,22), while targeted deletion of CTCF increases proximal Vκ gene germline transcription and recombination (23).

While CTCF has been localized to sites across the murine IgH locus (9,17,18), it remains unclear whether it is primarily recruited by direct binding to particular DNA sequences in the IgH locus, or whether it is primarily recruited indirectly to chromosomal DNA via protein-protein interactions with other DNA binding proteins such as Pax5, YY1, or cohesin, each of which have been shown to directly interact with CTCF (24-26). To test the hypothesis that CTCF is recruited to the murine IgH locus by direct recognition of

DNA binding sites, we have used a combination of bioinformatics and chromatin immunoprecipitation to gain further insight into the sequence determinants of CTCF binding. Analysis of the large number of CTCF DNA binding sites located throughout the V_H domain of the murine IgH locus reveals that these CTCF sites fall into three broad categories: those at a fixed distance (17-19 bp) from RSS elements; those positioned upstream of a V_H leader sequence; and those located in intergenic regions not associated with gene segments. When we extended our analysis to the human IgH locus, a similar pattern held. Interestingly, the CTCF binding sites located throughout the VH domains in both human and mouse have a distinct sub-consensus sequence motif that differs from the generic consensus motif found at CTCF sites throughout the rest of these organisms' genomes. Moreover, the consensus sequence present at the IgH binding sites differs between RSS-associated and RSS-unassociated CTCF sites. Finally, within the VH domains of both the human and mouse IgH loci, the CTCF binding sites all share the same orientation.

Results

A search for CTCF DNA binding sites at antigen receptor loci

To identify potential CTCF binding sites at the IgH locus, we performed a computational search for CTCF sites at the murine IgH locus. Initially, the CTCF binding site consensus sequence from the chicken β-globin 5'HS4 FII element – 5'-CCGCTAGGGGGCAG-3' (27) – was used to search for CTCF sites at the murine IgH locus. Allowing for 2 mismatches from the consensus sequence, this search revealed 96 putative CTCF binding sites within the locus. Using these 96 sites, we identified a new murine IgH CTCF consensus sequence – "mVH-CTCF" (5'-GACCAGCAGGGGGC-3'). We then repeated our computational search, allowing for 2 mismatches from mVH-CTCF. This search identified a total of 144 putative CTCF binding sites in the murine IgH locus (Supplementary Table 1), of which 138 were located within the VH domain of the locus (Figure 1a and see below). Of the sites not located within the V_H domain, one is in an intergenic region within the D_H segment cluster, two are in the constant region domain, and three are in the 3' regulatory region (3'RR), as previously reported (28). Interestingly, our search did not identify CTCF binding elements 1 (CBE1) and 2 (CBE2) (20,21,29), because the CTCF binding motif in CBE1 and CBE2 each have 6 mismatches from mVH-CTCF.

While previous studies have identified CTCF binding sites at the murine IgH locus (9,17,18), it remained unclear whether these CTCF sites are conserved throughout evolution. To address this question, we used the mV_H-CTCF sequence to perform a similar computational search of the human IgH locus. This search identified 131 putative CTCF binding sites. A new consensus sequence derived from the human IgH-CTCF sites – "hV_H-CTCF" (5'-ACCACCAGGGGGCG-3') contained minor differences from the mouse sequence at the 5' and 3' ends of the motif. Repeating our computational search of the human IgH locus with hV_H-CTCF increased the total number of sites identified in the human IgH locus to 188 (Supplementary Table 2) of which 183 are within the V_H region (Figure 1b). The density of CTCF binding sites is much higher within the human and murine IgH loci than in the rest of the human and mouse genomes (Supplementary Table 3). Repeating this search on the partial genomic sequences available for chimpanzee and rabbit also revealed the presence of numerous CTCF sites in the V_H region, suggesting that the prevalence of CTCF sites in the V_H domain is indeed evolutionarily conserved (Supplementary Table 3).

Given that CTCF sites have been identified at the murine Igκ (18,23) and TCRα (30) loci, we next asked whether CTCF binding sites are equally abundant at the other antigen receptor loci – Igκ, Igλ, TCRβ, TCRαδ, TCRγ – in humans and mice, and whether they share the V_H-CTCF motif. Repeating our computational search of the murine and human antigen receptor loci with mV_H-CTCF and hV_H-CTCF, respectively, we identified a number of putative CTCF binding sites in each of the antigen receptor loci, but they were much less abundant than in either the murine or human IgH loci (Supplementary Table 3).

Conserved orientation of CTCF binding sites

Analyzing the murine IgH locus, we noticed that all but 2 of the 147 identified CTCF sites at the murine V_H domain are present in the same preferred orientation, consistent with previous

findings (31). This same orientation bias is observed for the human (174/183) and chimpanzee (108/118) V_H-CTCF sites, suggesting that the orientation of CTCF binding sites within the V_H domain of the IgH locus is evolutionarily conserved and presumably functionally important, as previously suggested (32).

Distinct locations of CTCF sites

Locations of CTCF sites within the murine IgH locus are far from random. Within the murine VH domain, we found two classes of CTCF binding sites: RSS-associated sites (30% of all CTCF binding sites in this region; Figure 1a – black vertical lines) and RSS-unassociated sites (70% of all binding sites in this region; Figure 1a – red vertical lines). The overwhelming majority of the RSS-associated CTCF binding sites are positioned with their consensus core binding sequence (5'GACCAGCAGGGGGC3') located precisely 17-19 bp downstream of the nearby VH RSS nonamer, with only 3 sites violating this rule (see Supplementary Table 1). Notably, while the sequence of the RSS-associated CTCF binding sites is as highly conserved as the RSS itself, it is much more conserved than the sequence of the 17-19 bp between the RSS and CTCF sites. Thus, the length of this RSS-CTCF spacer region appears to be conserved even though the sequence itself is not.

In contrast, the positions of the RSS-unassociated CTCF binding sites are much more variable with respect to nearby VH gene segments. Upon closer analysis, the RSS-unassociated sites can be further divided into two subclasses: intergenic sites, which are not in close association with a VH gene segment (12% of all CTCF binding sites); and upstream sites, which are positioned upstream of a VH leader exon (58% of all CTCF binding sites). The upstream sites can be further subdivided by their distance from the nearest V_H leader exon, with subsets located approximately 800 bp, 2-3 kb, or 5-6 kb away.

Within the human and murine IgH locus, the V_H coding segments can be divided into families based on DNA sequence similarity (greater than 80% identity with all others in the family) and then further subdivided into 3 clans of closely related families. CTCF sites are found adjacent to RSS elements from 11 of the 16 murine and 3 of the 8 human gene segment families. Four of the five murine V_H gene segment families that lack RSS-associated CTCF sites (J558, SM7, Vgam3.8 and VH15) are in the same evolutionarily conserved clan (defined as Group 1 see (33)), while the fifth (3609) is from Group 2. Moreover, all of the functional members of the second and fourth largest murine VH families – 7183 and Q52 – have an RSS-associated CTCF site, while the non-functional ones typically do not. Examination of a phylogenetic tree for the V_H segments (33) revealed that only 3 segments within branches where the other segments had RSS-associated CTCF sites lacked an identifiable site (VH11.1.48, VH11.2.53, VH12.1.78). Closer inspection revealed the presence of a plausible CTCF site 19 bp away from each of these 3 RSSs, but with greater deviations from the consensus. Two of these sites have 3 mismatches, while one has 4, but the sites contain a 4/5 or 5/5 match with the 5 central G's of the core CTCF site, and both sites have only a 1 or 2 bp mismatch from the human sequence (see Supplementary Table 1). Thus, including these sites, there are 141 putative CTCF sites in the murine V_H region and 147 overall. Strikingly, the presence of an RSS-associated CTCF site can be predicted based on an evolutionary tree constructed with sequence that does not include the CTCF containing regions (i.e. just the V_H coding regions and RSS's), suggesting that these CTCF sites may play an evolutionarily significant role at these murine VH gene segments.

In the human IgH locus, as in the murine locus, CTCF sites are also found either in close association with RSS sequences or at upstream/intergenic positions. For the majority of the human RSS-associated CTCF sites, the core CTCF binding sequence is generally located a fixed distance away from the RSS – either 19 or 48 bp downstream of the RSS nonamer. Interestingly, the human homologs of the mouse proximal V_H gene segments (which have RSS-associated CTCF sites) also have RSS-associated CTCF binding sites. Additionally, whereas the non-RSS associated CTCF sites in the murine V_H domain are present as individual sites, the non-RSS associated CTCF sites in the human V_H domain are present as "hotspots" of multiple sites that vary in density, from 3 sites within 100 bp of each other to 50 sites within 2.1 kb of each other (see Figure 1b).

RSS-associated and intergenic/upstream CTCF sites are restricted to different regions of the murine V_H domain

The murine V_H domain consists of 195 VH gene segments spanning 2.5 Mb of DNA. Notably, the RSS-associated CTCF sites are sequestered in the D-proximal 900kb of the V_H region, while the intergenic CTCF sites are found in the D-distal 1.6Mb (Figure 1a), with some interspersion of RSS-associated and intergenic/upstream CTCF sites at the approximate interface between the "proximal" and "distal" regions. A number of studies have revealed distinct patterns in the regulation of recombination of these two regions (see Discussion). The approximate border between "distal" and "proximal" regions, as defined by these functional studies, is mirrored by the transition point between the domain containing upstream/intergenic sites and the region composed exclusively of RSS-associated CTCF sites (Figure 1a). No such distinct regulatory domains have been identified within the human V_H domain, and in the human V_H region intergenic/upstream and RSS-associated CTCF sites are intermingled, reflecting substantial intermingling of V_H clans.

Murine intergenic/upstream CTCF sites generally lack a site for CpG methylation

Although mouse and human V_H consensus CTCF sites are similar to each other and to the murine *Igf2/H19* CTCF consensus site and the chicken β-globin FII element, one difference is apparent: the CpG dinucleotides crucial for regulation in the Igf2/H19 and the chicken β-globin FII element are absent in the mouse and human V_H consensus CTCF motifs (see Figure 1c, nucleotides marked as 4,5 and 6,7 where numbering refers to murine consensus shown as boxed region in the figure). These CpG dinucleotides are subject to differential methylation that regulates the binding of CTCF to its target site, conferring monoallelic expression at the Igf2/H19 locus (34-37) and developmentally regulated β-globin gene expression (38). CTCF binding at the X-inactivation locus is also regulated by allele-specific CpG methylation (39). While CpG sites are lacking at positions 4/5 and 6/7, a CpG site is present instead at position 14/15 of the V_H CTCF motif in approximately 50% of the murine sites and 60% of the human sites (Figure 1c).

Although both intergenic and RSS-associated sites were identified by searches with the same motif (allowing two mismatches) we asked whether conserved differences between these two types of sites would allow for further subdivision. Consensus motifs for the CTCF DNA binding sites were determined independently for intergenic/upstream and RSS-associated sites for both the mouse and human IgH sites, using the Energy Normalized Logo (enoLOGOS) system (Figure 1d and 1e). A CpG target is present at position 14/15 for 59% of the human intergenic and RSS-associated consensus motifs (Figure 1e). A CpG site at this position is also present in approximately half of the murine RSS-associated CTCF sites; the other half lack any CpG dinucleotides (Figure 1d). With only two of the murine upstream/intergenic CTCF sites having a CpG dinucleotide (Figure 1d), the binding to these sites as a class cannot be regulated by differential CpG methylation.

IgH CTCF sites are bound in vitro by CTCF protein

Having identified these putative murine IgH CTCF sites in silico, we next wanted to ask whether these sites are bona fide CTCF binding sites. To address this question, we first employed electrophoretic mobility shift assays (EMSA) to test whether these DNA sequences can be bound by CTCF *in vitro*, as has been done previously for other CTCF sites (27,38,39). Using three partially overlapping 200 bp DNA probes encompassing a portion of the 7183.2.3 gene segment and its RSS-associated CTCF site (mCTCF.5) (Figure 2a), we found that the probe (KL2) containing the RSS-associated CTCF site (mCTCF.5) flanking the 7182.2.3 gene segment was shifted by both *in vitro* transcribed/translated CTCF (IVT-CTCF) (Figure 2b, lanes 4) and nuclear extracts from Pro-B, Pro-T, and NIH3T3 cells (Figure 2c, lanes 3, 6, and 9), while the adjacent probes (KL1 and KL3) which lacked the CTCF binding site were not shifted (Figure 2b, lanes 2 and 8). The shifted bands we observed using IVT-CTCF and nuclear extracts were all specifically supershifted by an anti-CTCF antibody (Figure 2b, compares lanes 5 and 6; Figure 2c, compare lanes 4 and 5, 7 and 8, 10 and 11), confirming that the mobility shift is due to CTCF binding. Moreover, when we mutated the mCTCF.5 binding site by converting three central

guanine residues to thymidines ($KL2^{mut}$), we observed no shifted bands for either IVT-CTCF or endogenous CTCF from the nuclear extracts (Figure 2c, lanes 12-16), confirming that CTCF was indeed binding to the consensus site we had identified. Thus, mCTCF.5 appears to be a bona fide CTCF binding site.

Although all mV_H-CTCF sites are extremely similar to each other and to the sequence in KL2, there are subsets of sites with distinct mismatches from the consensus (see groups B and C in Figure 2e). To examine whether these other sites could also bind CTCF in the context of their natural flanking DNA, we performed EMSA with a representative set of consensus and non-consensus sites derived from RSS-associated, intergenic, and upstream sites. Using substrates that positioned the CTCF site in the center of the fragment, with 40 bp of genomic sequence flanking the site on either side, we found that all but one of these sites were capable of binding CTCF, suggesting that the murine VH CTCF sites can generally function as *in vitro* CTCF binding sites (Figure 2e). Furthermore, when we tested one of the human VH CTCF sites (hCTCF.169), we found that it could also be bound by IVT-CTCF (Figure 2d, lane 6). However, no binding was detected to probes derived from the murine VHJ558 gene segments (Figure 2d, lane 4), indicating that there are no noncanonical CTCF binding sites associated with the RSS sequence of the distal VH gene segments.

mVH-CTCF binding sequences from IgH exhibit enhancer-blocking activity

Previous studies have established that CTCF-binding generally confers the enhancer blocking activity observed in many vertebrate insulator elements (21,27,34,37). Therefore, to further confirm that the murine and human VH CTCF sites we identified are bona fide CTCF sites, we utilized a standard enhancer-blocking assay (27,37). DNA fragments containing mCTCF.5 exhibited slightly stronger enhancer blocking activity than the classical insulator element (INS) from the chicken β-globin locus (compare Figure 3c vs. 3d). As with the β globin INS, inclusion of a second copy of mCTCF.5 increased the enhancer-blocking activity (compare Figure 3e vs. 3f and 3l vs. 3m). As is often observed with CTCF sites (27), reversing the orientation of mCTCF.5 dramatically reduced the enhancer-blocking activity (compare Figure 3f and 3h), indicating that this activity is orientation-dependent. As expected, expression of the neomycin reporter was only blocked when mCTCF.5 was positioned between the enhancer and the promoter (Figure 3i), confirming that this element is an enhancer-blocker rather than a DNA silencer. Consistent with the results of our gel-shift analysis, mutating three of the central G residues in the core portion of the CTCF binding site reduced enhancer-blocking activity (Figures 3g and 3n). Finally, we also observed potent enhancer-blocking activity by a DNA fragment containing hCTCF.169 (Figure 3k), but no enhancer-blocking activity by a DNA fragment encompassing a J558 VH gene segment (Figure 3j).

Taken together, these results indicate that both mV_H-CTCF sites and hV_H-CTCF sites exhibit potent enhancer-blocking activity in the context of their surrounding sequence, strongly suggesting that they function as CTCF binding sites in the cell.

CTCF binds to sites within the endogenous IgH locus

Since a subset of mVH-CTCF sites can be stably bound by CTCF *in vitro*, and function as CTCF binding sites in the context of exogenous DNA fragments *in vivo*, we next performed chromatin immunoprecipitation followed by quantitative PCR (ChIP-qPCR) in both *ex vivo* cell lines and primary cells to ask whether CTCF binds to endogenous mV_H-CTCF sites within their normal chromatin context *in vivo*.

In both $RAG2^{-/-}$ Pro-B cell lines and primary $CD19^+$ Pro-B cells harvested from the bone marrow of 8-week-old RAG2-deficient mice – where no V(D)J recombination has occurred, and all the antigen receptor loci are in their germline configuration due to the lack of an active recombinase – CTCF was enriched to varying extents at all the intergenic/upstream and RSS-associated sites tested, with maximal VH domain enrichment at mCTCF.57 (Figure 4a, left panel; Figure 4c). Looking at cells from later stages of B cell development – the 1-8 Pre-B cell line (Figure 4a, right panel), $CD19^+$ cells from 8-week-old WT bone marrow (Figure 4d), and $CD19^+$ WT splenic B cells (Figure 4e) - we observed CTCF binding

patterns that were similar to that observed in pro-B cells. However, when we analyzed CTCF binding to the murine IgH locus in non-lymphoid cells – NIH3T3 fibroblasts (Figure 4b, left panel), mouse embryonic fibroblasts (Figure 4b, center panel), and primary hepatocytes (Figure 4b, right panel) – we still observed CTCF binding, but the levels of enrichment were lower, and binding was restricted to the RSS-associated D_H-proximal CTCF sites. Thus, CTCF binding shows distinct patterns in lymphoid vs. non-lymphoid cells.

CTCF binds to multiple sites throughout the murine IgH locus

As many of the CTCF sites identified by our computational search are bound by CTCF both *in vitro* and *in vivo*, and since previous studies have observed CTCF binding sites within the IgH locus (9,17,18), we next compared the *in vivo* pattern of CTCF binding across the murine IgH locus in primary $CD19^+$ $RAG2^{-/-}$ Pro-B cells to our *in silico* predictions. To examine CTCF binding across the entire murine IgH locus, we first isolated $CD19^+$ Pro-B cells from 8-9 week old $RAG2^{-/-}$ mice. Since IL-7 is known to support the growth of primary Pro-B cells (24,31,40,41), we expanded these cells in the presence of varying concentrations of recombinant IL-7 before performing chromatin immunoprecipitation with an a-CTCF antibody (Figure S1). Next, we took these $CD19^+$ $RAG2^{-/-}$ Pro-B cells and either expanded the cells for 3 days in the presence of 10 ng/mL of the growth factor IL-7 on OP9 feeder cells and 1 day in the presence of mitomycin C-treated ST2 cells, or harvested them immediately for chromatin immunoprecipitation with an α-CTCF antibody. The input DNA and immunoprecipitated DNA were then labeled with Cy3 or Cy5, respectively, hybridized to custom-designed tiling microarrays, and peaks were called by standard bioinformatic analysis (see Materials and Methods).

Since the DNA-binding footprint of CTCF is ~70 bp, adjacent peaks that were very narrowly spaced (<100 bp between them) were combined into a single peak using a PERL script, resulting in a total of 190 CTCF peaks across the murine IgH locus. These peaks ranged in size from 10 bp to 3498 bp, with an average peak width of 1146 bp.

After determining the localization pattern of CTCF across the murine IgH locus in $Rag2^{-/-}$ pro-B cells, we compared the CTCF binding sites predicted *in silico* to the CTCF binding sites observed *in vivo*. Of the 144 *in silico*-predicted CTCF binding sites, 111 were occupied *in vivo* (77%), suggesting that the presence of a CTCF consensus sequence is a major determinant of CTCF binding *in vivo*. Conversely, 58% of the observed CTCF peaks contained a mVH-CTCF consensus sequence. Analyzing the overlap between our predicted CTCF binding sites, the CTCF binding sites we observed by ChIP-chip, and the CTCF binding sites previously identified by ChIP-seq in cultured $Rag2^{-/-}$ pro-B cells (9), we found that of the 190 CTCF peaks we identified by ChIP-chip, 107 (56%) overlapped with the CTCF ChIP-seq peaks (Figure 5A; Table S5). Of the 144 putative CTCF binding sites that matched our predicted consensus motif, 111 (75%) overlapped with the CTCF peaks previously identified by ChIP-seq (Figure 5A).

To learn more about CTCF binding at the peaks that did not overlap with our predicted CTCF consensus motif, we analyzed these 79 sites using MEME (42) to probe for alternative sequence motifs. Three sequence motifs were identified (Figure 5C), none of which bore obvious similarity to the CTCF consensus motif. To determine whether any of these motifs were similar to other known protein binding-site motifs (e.g. YY1, cohesin, or nucleophosmin), we compared all three motifs to the JASPAR Vertebrates and UniPROBE Mouse database using Tomtom (42). However, running these three sequence motifs through Tomtom failed to retrieve any statistically significant hits to known motifs in the JASPAR Vertebrates and UniPROBE Mouse database.

Finally, since previous studies have identified a two-part CTCF binding motif consisting of a fairly well-conserved M1 motif of 20 bp (Figure 6A) adjacent to a less well-conserved M2 motif of 9 bp (43,44), we analyzed the overlap between our predicted murine VH CTCF binding sites, the CTCF binding sites we observed by ChIP-chip, and predicted M1 sites. Of the 190 CTCF peaks we identified by ChIP-chip, 126 (66%) contained a predicted M1 motif (Figure 6B). Of the 144 putative CTCF binding sites that matched our predicted consensus motif, 133 (92%) contained a predicted M1 motif (Figure 6B). And of the 79 CTCF ChIP-chip peaks that did not overlap with

our predicted CTCF consensus motif, 18 (23%) contained a predicted M1 motif (Figure 6B).

Discussion

Molecular determinants of CTCF binding at IgH locus

While previous studies have analyzed CTCF occupancy at the IgH locus (9,17,18), the mechanism by which CTCF is recruited to the IgH locus during B cell development remains unclear. Does CTCF bind directly to particular DNA sequences in the IgH locus when these sequences become accessible, or is it being recruited indirectly via protein-protein interactions with other DNA-binding proteins, such as YY1 (12,25), cohesin (18,45), or Pax5 (9,10)? Here, we find that the majority of CTCF-occupied sites overlap with a computationally-identifiable sub-consensus motif – mVH-CTCF (5'-GACCAGCAGGGGGC-3') – that is distinct from the generic CTCF consensus motif that is found elsewhere in the mouse genome. This CTCF sub-consensus motif is unique to the V domain of the IgH locus, highly conserved between binding sites within the locus, and displays far more sequence conservation than the surrounding sequences. Thus, there was likely a strong evolutionary pressure to maintain this specific version of the CTCF binding site, despite the known ability of CTCF to bind to degenerate sequences. Moreover, we find that CTCF can directly bind to these sites in vitro, suggesting that while other proteins may help to stabilize CTCF once it is bound, CTCF is likely recruited to the IgH locus directly via its sequence-specific DNA-binding activity. Given the differences between mVH-CTCF and previously identified CTCF binding site consensus motifs (27,34,39), and given that distinct functions have been ascribed to individual zinc fingers within CTCF (38,46), it is tempting to speculate that CTCF uses a distinct combination of its 11 zinc fingers to bind mVH-CTCF, as compared to other CTCF sites located throughout the mouse genome, thereby leaving a similarly distinct combination of its zinc fingers available for protein-protein interactions with other known CTCF-interacting proteins such as cohesin (18,45), YY1 (12,25), or the lymphoid-specific protein Pax5 (9,10). Further studies will be required to test this hypothesis.

Evolutionary conservation of numerous CTCF binding sites across the IgH locus

The 2.5 Mb murine IgH locus contains an extraordinarily high number of CTCF sites (this work and (9,17,18,20)), especially as compared to the number of CTCF sites found at the other murine antigen receptor loci and several orders of magnitude greater than the mouse genome generally (44). Similarly, the 1.25 Mb human IgH locus contains a remarkably large number of CTCF binding sites, with a density of sites that is an order of magnitude greater than the other human antigen receptor loci (this study), and several orders of magnitude greater than the human genome generally (47). Other studies using computational methods or genome-wide ChIP analysis have also identified CTCF sites at the human TCRβ, TCRα/δ, IgH, Igκ, and Igλ loci (9,18,30,47,48). While the precise numbers of sites vary somewhat between these studies – possibly reflecting either the different search sequences, the specific cell type being examined in the ChIP studies, the probe content of the microarrays, or the peak-calling algorithms – the high density of CTCF binding sites at the IgH locus is striking, particularly since it is evolutionarily conserved in mice, rabbits, chimpanzees, and humans (this study). The high density of conserved CTCF binding sites underscores the likely importance of CTCF in regulation of antigen receptor loci, consistent with recent studies (17,20,22,24,49). However, the exact function(s) of these multiple sites at the IgH locus remains unclear (see below).

Distinct classes of CTCF binding sites within murine IgH locus

Using our computational consensus motif-based approach, we not only identified a similar number of CTCF binding sites within the murine IgH locus, but we also discovered two distinct classes of CTCF binding sites: RSS-associated sites that are located ~19 bp downstream of the nearest RSS; and RSS-unassociated sites that are located at least 800 bp away from the nearest RSS (17). We note that the RSS-associated CTCF sites are all located within the D_H-proximal region of the V_H domain, while the RSS-unassociated CTCF sites are located in the D_H-distal region of the V_H domain. In addition, it is intriguing that for the RSS-associated sites, distance from CTCF site to

RSS is conserved (~2 turns of the double-helix), even though the intervening DNA sequence is not, suggesting that the RSS-CTCF distance is functionally significant. While a previous study noted that CTCF sites in the proximal half of the VH locus were within 150 bp of the RSSs (17), we find a much tighter association between the RSS-associated CTCF sites and the adjacent RSSs. It is noteworthy that RSS-associated CTCF sites are also positioned a fixed distance from their associated RSSs in humans (~2 or 4 turns of the double-helix) and other species. Since the accessibility of the DH-distal and DH-proximal regions of the V_H domain is known to be differentially regulated during B cell development, we suggest that the RSS-associated CTCF sites in the DH-proximal region likely have a function that is distinct from the RSS-unassociated CTCF sites in the DH-distal region of the V_H domain. Indeed, we have recently shown that CTCF binding to these RSS-associated sites is highly predictive of high frequency recombination among DH-proximal V gene segments (49). Furthermore, given that the distance from the RSS-associated CTCF sites to their associated RSSs is either 2 or 4 turns of the double-helix in mice and humans, it is tempting to speculate that CTCF may be directly influencing the activity of the RAG1/2 proteins at these gene segments. Future studies will test this hypothesis.

Additionally, while CTCF sites at the Igf2/H19 locus (34-37), β-globin locus (38), the X-inactivation locus (39), and the IgH superanchor (31) are all regulated by CpG methylation, only 50% of the murine RSS-associated CTCF sites contain CpG motifs, suggesting that binding to a large fraction of these sites is either not regulated or is regulated in a CpG-independent manner. Moreover, only two of the murine upstream/intergenic CTCF sites contain a CpG dinucleotide, indicating that CTCF binding to these sites can't be regulated by CpG methylation. Thus, the differential binding observed in distinct cell types, may reflect distinct chromatin structure that occurs independently of (and therefore, prior to) CTCF binding. Further, in much the same way that the DH-proximal region CTCF sites (which are RSS-associated) may have a distinct function from the DH-distal region CTCF sites (which are RSS-unassociated), CTCF binding to these two classes of sites may also be regulated in different ways. Further experiments will be required to explore the differential function and regulation of these CTCF sites within the V_H domain of the murine immunoglobulin heavy chain locus.

Conserved orientation of CTCF sites with VH domain of IgH locus

Given the large number of CTCF sites within the murine IgH locus, it is striking that over 98% of these sites are present in the same orientation. Since CTCF has been found to affect chromosomal looping (50,51), and previous studies have identified CTCF sites with the opposite orientation within the IgH intergenic control region 1 (IGCR1) (29) and the IgH superanchor (31), it seems likely that one function of the CTCF sites within the VH domain of the IgH locus is to form loops that promote synapsis of DJH and VH gene segments, as suggested previously (52-54). Moreover, the large number of CTCF sites within the VH domain may allow for competition between sites that synapse to convergent CTCF sites within IGCR1 or the IgH superanchor, thereby forming distinct chromosomal loop domains that could facilitate linear tracking of RAG1/2, as suggested previously (32). It is worth nothing that CTCF-dependent chromosomal looping has also been implicated in regulating V(D)J recombination at other antigen receptor loci (30,55). However, since there are several distinct classes of CTCF sites within the murine IgH locus, it seems likely that some of the VH domain CTCF sites are functioning in a looping-independent manner. The RSS-associated CTCF sites in the DH-proximal portion of the domain appear to affect the accessibility and activity of the V(D)J recombinase at these gene segments (49). Some of the intergenic non-RSS associated CTCF sites in the DH-distal portion of the domain have been shown to affect the 3D conformation of the locus (20). However, the role of the non-RSS associated CTCF sites at the interface of the proximal and distal regions is unknown, and it is intriguing to speculate that they function as enhancer blockers that separate the regulation of the proximal and distal regions of the locus. Finally, our studies have revealed that a large class of CTCF binding sites – namely the upstream sites – conform neither to a conformational nor a local recombinase activating role. Their conserved

spatial distances upstream of V gene segments suggests a possible role in insulating V gene segments from neighboring V gene segments. In any case, understanding the sequence determinants of CTCF binding to the murine IgH locus should facilitate future studies evaluating how IgH locus accessibility regulates CTCF binding as well as the functions that CTCF plays in regulating the recombinational accessibility of VH gene segments during B cell development.

Experimental procedures

Mice and cell culture

Animal experiments and procedures were approved by the Institutional Animal Care and Use Committee of Massachusetts General Hospital. WT and RAG2$^{-/-}$ mice were obtained from Taconic Farms and were bred and maintained in HPP-free animal facilities at MGH. Pro-B cells were recovered from femoral bone marrow suspensions derived from 8-week old mice by positive enrichment of CD19^{+} cells using MACS magnetic separation (Miltenyi Biotec). A portion of these cells were placed into culture in the presence of IL7 prior to harvesting for chromatin IP, while chromatin was prepared from the remaining cells and frozen to permit immunoprecipitation in parallel with material recovered from cultured cells. See Supplementary Methods for additional information about culturing conditions for CD19^{+} cells.

WT livers and spleens were forced through a 19 G needle and passed through a sterile mesh filter to generate single cell suspensions. The cells were washed and splenic B cells were collected by positive enrichment of CD19^{+} cells using MACS magnetic separation.

Cell lines

RAG2$^{-/-}$ Abelson transformed pro-B cells, RAG1$^{-/-}$ p53$^{-/-}$ Pro T cells, and Abelson transformed 1-8 B cells were maintained in RPMI 1640 supplemented with 20% fetal bovine serum and 0.05mM 2-mercaptoethanol. NIH3T3 fibroblast cells and mouse embryonic fibroblasts were maintained in DMEM supplemented with 10% calf serum. Human erythroleukemia K562 cells (a gift from Jeannie Lee) were cultured in IMDM supplemented with 10% fetal bovine serum.

Sequence Alignments

Annotated genomic sequence spanning antigen receptor loci was obtained from Genbank (see Supplementary Methods for accession numbers). Vseg elements for the IgH loci of chimpanzee (NW_001224639.1); chicken (NW_001477447.1, NW_001484419.1), and dog (NW_876328.1) were identified by tblastx using Mouse sequences as the blast query. Searches for CTCF DNA binding sites and sequence alignments were performed using MacVector v7.2 and EMBOSS version 4.1.0.

Electrophoretic Mobility Shift Assays

DNA probes were obtained by PCR amplification from either human HeLa cell genomic DNA or murine pro-B cell genomic DNA (see Supplementary Table 4 for primer sequences), gel-purified, and sequenced. All probes were 5' end-labeled with ^{32}P-ATP as described (56). The 250 bp probes KL1 and KL3 each overlap with the 200 bp KL2 probe by 50bp. *In vitro* translated (IVT) CTCF was prepared from *pCTCF* (a gift from Jeannie Lee) using the TNT Coupled Reticulocyte Lysate System (Promega). Nuclear extracts were prepared from approximately 1x10^{8} pro-B cells, Pro T cells, or NIH3T3 cells as described (57). CTCF protein was purified from a HeLa cell line that stably expresses a double-tagged FLAG-HA-hCTCF transgene as described (58). See Supplementary Methods for additional details.

Enhancer Blocking Assay

K562 cell transfections and colony assays were performed as previously described (59). See Supplementary Methods for a detailed description of the methodology.

Chromatin Immunoprecipitation

Chromatin immunoprecipitations were performed as described (60) with 30 µl of anti-CTCF antibody (Upstate Biotechnology) and analyzed by real-time PCR with SYBR Green or TaqMan probes or by hybridization to custom DNA microarrays. For additional information about chromatin immunoprecipitation methodology, see Supplemental Methods. For primer and probe sequences see Supplementary Table 4.

Microarray hybridization and processing

Tiling genomic DNA microarrays were custom designed (NimbleGen Systems, Inc) based on the mm9 release of the IgH locus sequence (murine chr12: 114,341,024–117,349,200). 50-mer probes were selected every 20 bases with no repeat masking, on both the top and bottom strands. Three replicates for each strand were spotted on the array. Genomic DNA and CTCF ChIP DNA were labeled with Cy3 and Cy5, respectively, and hybridized to the array by the manufacturer.

Computational Analysis

The Ringo method (61) was implemented as a Bioconductor package to identify CTCF peaks from the ChIP-microarray data. A position weight matrix (PWM) for the M1 motif (44) was downloaded from CTCFBSDB database (http://insulatordb.uthsc.edu). The search for the M1 motif matches across the regions of interest was performed using FIMO (62) with default parameters.

Acknowledgements: This work was funded by the Biotechnological and Biological Scientific Research Council (BBSRC) (A.C.), and the NIH GM48026 (M.A.O). A.L.W. was supported by a BBSRC PhD studentship.

Conflict of interest: The authors declare that they have no conflicts of interest with the contents of this article. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

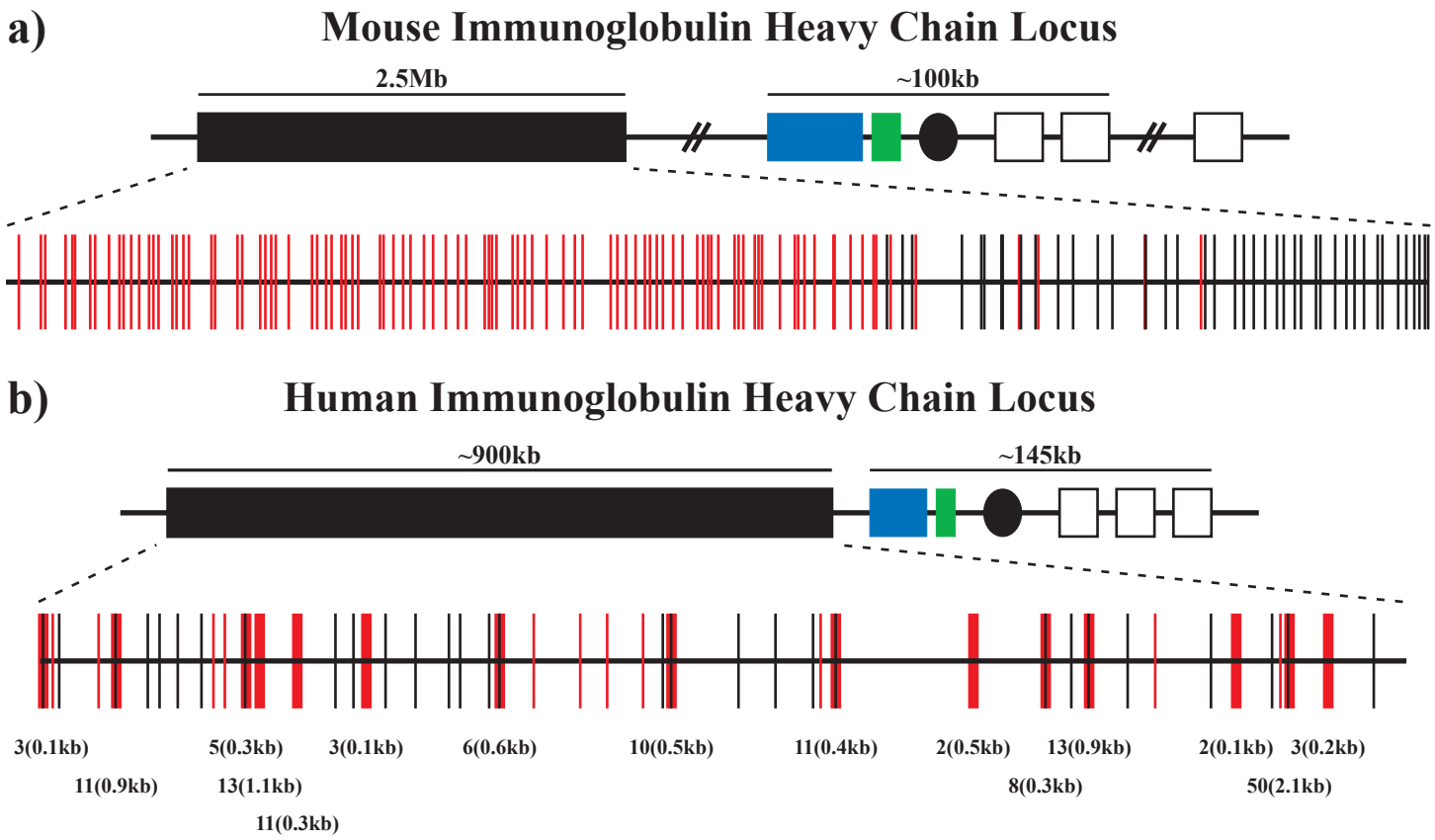
1. Grundy, G. J., Ramon-Maiques, S., Dimitriadis, E. K., Kotova, S., Biertumpfel, C., Heymann, J. B., Steven, A. C., Gellert, M., and Yang, W. (2009) Initial stages of V(D)J recombination: the organization of RAG1/2 and RSS DNA in the postcleavage complex. *Molecular cell* **35**, 217-227
2. Kim, M. S., Lapkouski, M., Yang, W., and Gellert, M. (2015) Crystal structure of the V(D)J recombinase RAG1-RAG2. *Nature* **518**, 507-511
3. Pulivarthy, S. R., Lion, M., Kuzu, G., Matthews, A. G., Borowsky, M. L., Morris, J., Kingston, R. E., Dennis, J. H., Tolstorukov, M. Y., and Oettinger, M. A. (2016) Regulated large-scale nucleosome density patterns and precise nucleosome positioning correlate with V(D)J recombination. *Proceedings of the National Academy of Sciences of the United States of America* **113**, E6427-E6436
4. Stubbington, M. J., and Corcoran, A. E. (2013) Non-coding transcription and large-scale nuclear organisation of immunoglobulin recombination. *Curr Opin Genet Dev* **23**, 81-88
5. Carico, Z., and Krangel, M. S. (2015) Chromatin Dynamics and the Development of the TCRalpha and TCRdelta Repertoires. *Adv Immunol* **128**, 307-361
6. Kumari, G., and Sen, R. (2015) Chromatin Interactions in the Control of Immunoglobulin Heavy Chain Gene Assembly. *Adv Immunol* **128**, 41-92
7. Majumder, K., Bassing, C. H., and Oltz, E. M. (2015) Regulation of Tcrb Gene Assembly by Genetic, Epigenetic, and Topological Mechanisms. *Adv Immunol* **128**, 273-306
8. Proudhon, C., Hao, B., Raviram, R., Chaumeil, J., and Skok, J. A. (2015) Long-Range Regulation of V(D)J Recombination. *Adv Immunol* **128**, 123-182
9. Ebert, A., McManus, S., Tagoh, H., Medvedovic, J., Salvagiotto, G., Novatchkova, M., Tamir, I., Sommer, A., Jaritz, M., and Busslinger, M. (2011) The distal V(H) gene cluster of the Igh locus contains distinct regulatory elements with Pax5 transcription factor-dependent activity in pro-B cells. *Immunity* **34**, 175-187
10. 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 & development* **18**, 411-422
11. 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., and Kenter, A. L. (2016) Extremely Long-Range Chromatin Loops Link Topological Domains to Facilitate a Diverse Antibody Repertoire. *Cell reports* **14**, 896-906
12. 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 & development* **21**, 1179-1189
13. Osipovich, O. A., Subrahmanyam, R., Pierce, S., Sen, R., and Oltz, E. M. (2009) Cutting edge: SWI/SNF mediates antisense Igh transcription and locus-wide accessibility in B cell precursors. *J Immunol* **183**, 1509-1513
14. Osipovich, O., Cobb, R. M., Oestreich, K. J., Pierce, S., Ferrier, P., and Oltz, E. M. (2007) Essential function for SWI-SNF chromatin-remodeling complexes in the promoter-directed assembly of Tcrb genes. *Nat Immunol* **8**, 809-816
15. Osipovich, O., Milley, R., Meade, A., Tachibana, M., Shinkai, Y., Krangel, M. S., and Oltz, E. M. (2004) Targeted inhibition of V(D)J recombination by a histone methyltransferase. *Nat Immunol* **5**, 309-316

16. Su, I. H., Basavaraj, A., Krutchinsky, A. N., Hobert, O., Ullrich, A., Chait, B. T., and Tarakhovsky, A. (2003) Ezh2 controls B cell development through histone H3 methylation and Igh rearrangement. *Nature immunology* **4**, 124-131
17. 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., Murre, C. S., Birshtein, B. K., Schork, N. J., Schlissel, M. S., Riblet, R., Murre, C., and Feeney, A. J. (2011) CCCTC-binding factor (CTCF) and cohesin influence the genomic architecture of the Igh locus and antisense transcription in pro-B cells. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 9566-9571
18. 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
19. Volpi, S. A., Verma-Gaur, J., Hassan, R., Ju, Z., Roa, S., Chatterjee, S., Werling, U., Hou, H., Jr., Will, B., Steidl, U., Scharff, M., Edelman, W., Feeney, A. J., and Birshtein, B. K. (2012) Germline deletion of Igh 3' regulatory region elements hs 5, 6, 7 (hs5-7) affects B cell-specific regulation, rearrangement, and insulation of the Igh locus. *J Immunol* **188**, 2556-2566
20. Guo, C., Yoon, H. S., Franklin, A., Jain, S., Ebert, A., Cheng, H. L., Hansen, E., Despo, O., Bossen, C., Vettermann, C., Bates, J. G., Richards, N., Myers, D., Patel, H., Gallagher, M., Schlissel, M. S., Murre, C., Busslinger, M., Giallourakis, C. C., and Alt, F. W. (2011) CTCF-binding elements mediate control of V(D)J recombination. *Nature* **477**, 424-430
21. Featherstone, K., Wood, A. L., Bowen, A. J., and Corcoran, A. E. (2010) The mouse immunoglobulin heavy chain V-D intergenic sequence contains insulators that may regulate ordered V(D)J recombination. *The Journal of biological chemistry* **285**, 9327-9338
22. 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 & development* **29**, 1683-1695
23. Ribeiro de Almeida, C., Stadhouders, R., de Bruijn, M. J., 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 Vkappa recombination and restricts kappa enhancer interactions to the immunoglobulin kappa light chain locus. *Immunity* **35**, 501-513
24. 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., Denizot, Y., Holwerda, S. J., de Laat, W., Cogne, M., Shi, Y., Alt, F. W., and Busslinger, M. (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
25. Donohoe, M. E., Zhang, L. F., Xu, N., Shi, Y., and Lee, J. T. (2007) Identification of a Ctcf cofactor, Yy1, for the X chromosome binary switch. *Molecular cell* **25**, 43-56
26. Xiao, T., Wallace, J., and Felsenfeld, G. (2011) Specific sites in the C terminus of CTCF interact with the SA2 subunit of the cohesin complex and are required for cohesin-dependent insulation activity. *Molecular and cellular biology* **31**, 2174-2183
27. Bell, A. C., West, A. G., and Felsenfeld, G. (1999) The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell* **98**, 387-396

28. Garrett, F. E., Emelyanov, A. V., Sepulveda, M. A., Flanagan, P., Volpi, S., Li, F., Loukinov, D., Eckhardt, L. A., Lobanenkov, V. V., and Birshtein, B. K. (2005) Chromatin architecture near a potential 3' end of the igh locus involves modular regulation of histone modifications during B-Cell development and in vivo occupancy at CTCF sites. *Molecular and cellular biology* **25**, 1511-1525
29. Lin, S. G., Guo, C., Su, A., Zhang, Y., and Alt, F. W. (2015) CTCF-binding elements 1 and 2 in the Igh intergenic control region cooperatively regulate V(D)J recombination. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 1815-1820
30. 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. *Proceedings of the National Academy of Sciences of the United States of America* **109**, E3493-3502
31. Benner, C., Isoda, T., and Murre, C. (2015) New roles for DNA cytosine modification, eRNA, anchors, and superanchors in developing B cell progenitors. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 12776-12781
32. 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
33. 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
34. Bell, A. C., and Felsenfeld, G. (2000) Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature* **405**, 482-485
35. Fedoriw, A. M., Stein, P., Svoboda, P., Schultz, R. M., and Bartolomei, M. S. (2004) Transgenic RNAi reveals essential function for CTCF in H19 gene imprinting. *Science (New York, N.Y* **303**, 238-240
36. Kanduri, C., Pant, V., Loukinov, D., Pugacheva, E., Qi, C. F., Wolffe, A., Ohlsson, R., and Lobanenkov, V. V. (2000) Functional association of CTCF with the insulator upstream of the H19 gene is parent of origin-specific and methylation-sensitive. *Curr Biol* **10**, 853-856
37. Hark, A. T., Schoenherr, C. J., Katz, D. J., Ingram, R. S., Levorse, J. M., and Tilghman, S. M. (2000) CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature* **405**, 486-489
38. Renda, M., Baglivo, I., Burgess-Beusse, B., Esposito, S., Fattorusso, R., Felsenfeld, G., and Pedone, P. V. (2007) Critical DNA binding interactions of the insulator protein CTCF: a small number of zinc fingers mediate strong binding, and a single finger-DNA interaction controls binding at imprinted loci. *The Journal of biological chemistry* **282**, 33336-33345
39. Chao, W., Huynh, K. D., Spencer, R. J., Davidow, L. S., and Lee, J. T. (2002) CTCF, a candidate trans-acting factor for X-inactivation choice. *Science* **295**, 345-347
40. Kleiman, E., Jia, H., Loguercio, S., Su, A. I., and Feeney, A. J. (2016) YY1 plays an essential role at all stages of B-cell differentiation. *Proceedings of the National Academy of Sciences of the United States of America* **113**, E3911-3920
41. Lin, Y. C., Jhunjhunwala, S., Benner, C., Heinz, S., Welinder, E., Mansson, R., Sigvardsson, M., Hagman, J., Espinoza, C. A., Dutkowski, J., Ideker, T., Glass, C. K.,

and Murre, C. (2010) A global network of transcription factors, involving E2A, EBF1 and Foxo1, that orchestrates B cell fate. *Nature immunology* **11**, 635-643
42. Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., and Noble, W. S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic acids research* **37**, W202-208
43. Rhee, H. S., and Pugh, B. F. (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147**, 1408-1419
44. Schmidt, D., Schwalie, P. C., Wilson, M. D., Ballester, B., Goncalves, A., Kutter, C., Brown, G. D., Marshall, A., Flicek, P., and Odom, D. T. (2012) Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148**, 335-348
45. Parelho, V., Hadjur, S., Spivakov, M., Leleu, M., Sauer, S., Gregson, H. C., Jarmuz, A., Canzonetta, C., Webster, Z., Nesterova, T., Cobb, B. S., Yokomori, K., Dillon, N., Aragon, L., Fisher, A. G., and Merkenschlager, M. (2008) Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell* **132**, 422-433
46. Filippova, G. N., Fagerlie, S., Klenova, E. M., Myers, C., Dehner, Y., Goodwin, G., Neiman, P. E., Collins, S. J., and Lobanenkov, V. V. (1996) An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Molecular and cellular biology* **16**, 2802-2813
47. Kim, T. H., Abdullaev, Z. K., Smith, A. D., Ching, K. A., Loukinov, D. I., Green, R. D., Zhang, M. Q., Lobanenkov, V. V., and Ren, B. (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128**, 1231-1245
48. Xie, X., Mikkelsen, T. S., Gnirke, A., Lindblad-Toh, K., Kellis, M., and Lander, E. S. (2007) Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc Natl Acad Sci U S A* **104**, 7145-7150
49. Bolland, D. J., Koohy, H., Wood, A. L., Matheson, L. S., Krueger, F., Stubbington, M. J., Baizan-Edge, A., Chovanec, P., Stubbs, B. A., Tabbada, K., Andrews, S. R., Spivakov, M., and Corcoran, A. E. (2016) Two Mutually Exclusive Local Chromatin States Drive Efficient V(D)J Recombination. *Cell reports* **15**, 2475-2487
50. Merkenschlager, M., and Odom, D. T. (2013) CTCF and cohesin: linking gene regulatory elements with their targets. *Cell* **152**, 1285-1297
51. Ong, C. T., and Corces, V. G. (2014) CTCF: an architectural protein bridging genome topology and function. *Nat Rev Genet* **15**, 234-246
52. Lucas, J. S., Zhang, Y., Dudko, O. K., and Murre, C. (2014) 3D trajectories adopted by coding and regulatory DNA elements: first-passage times for genomic interactions. *Cell* **158**, 339-352
53. Jhunjhunwala, S., van Zelm, M. C., Peak, M. M., and Murre, C. (2009) Chromatin architecture and the generation of antigen receptor diversity. *Cell* **138**, 435-448
54. Lucas, J. S., Bossen, C., and Murre, C. (2011) Transcription and recombination factories: common features? *Curr Opin Cell Biol* **23**, 318-324
55. Zhao, L., Frock, R. L., Du, Z., Hu, J., Chen, L., Krangel, M. S., and Alt, F. W. (2016) Orientation-specific RAG activity in chromosomal loop domains contributes to Tcrd V(D)J recombination during T cell development. *The Journal of experimental medicine* **213**, 1921-1936

56. Cuomo, C. A., Mundy, C. L., and Oettinger, M. A. (1996) DNA sequence and structure requirements for cleavage of V(D)J recombination signal sequences. *Molecular and cellular biology* **16**, 5683-5690
57. Dignam, J. D., Lebovitz, R. M., and Roeder, R. G. (1983) Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. *Nucleic acids research* **11**, 1475-1489
58. Yusufzai, T. M., Tagami, H., Nakatani, Y., and Felsenfeld, G. (2004) CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species. *Molecular cell* **13**, 291-298
59. Chung, J. H., Whiteley, M., and Felsenfeld, G. (1993) A 5' element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in Drosophila. *Cell* **74**, 505-514
60. Ciccone, D. N., Morshead, K. B., and Oettinger, M. A. (2004) Chromatin immunoprecipitation in the analysis of large chromatin domains across murine antigen receptor loci. *Methods Enzymol* **376**, 334-348
61. Toedling, J., Skylar, O., Krueger, T., Fischer, J. J., Sperling, S., and Huber, W. (2007) Ringo--an R/Bioconductor package for analyzing ChIP-chip readouts. *BMC Bioinformatics* **8**, 221
62. Grant, C. E., Bailey, T. L., and Noble, W. S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017-1018



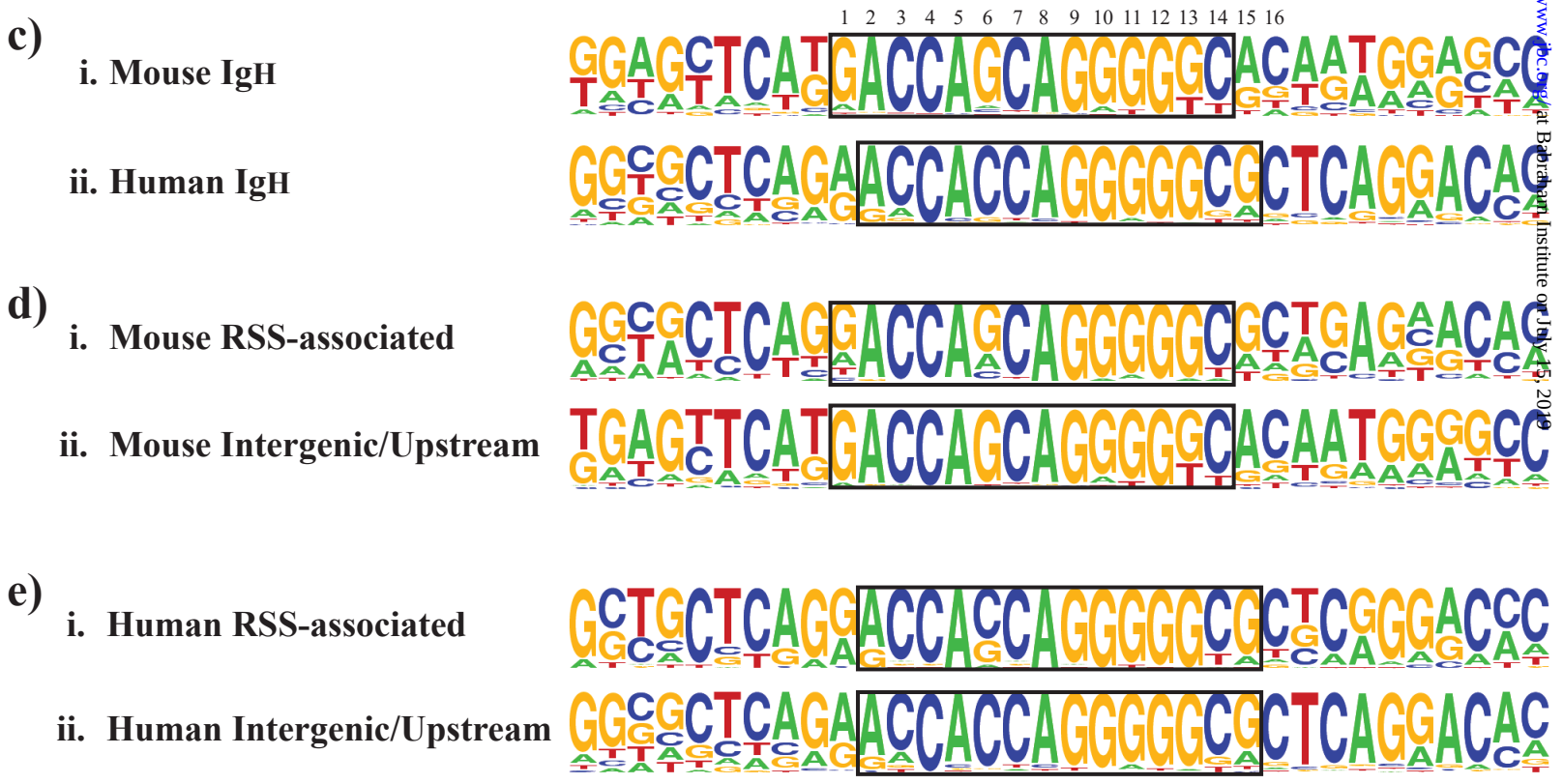
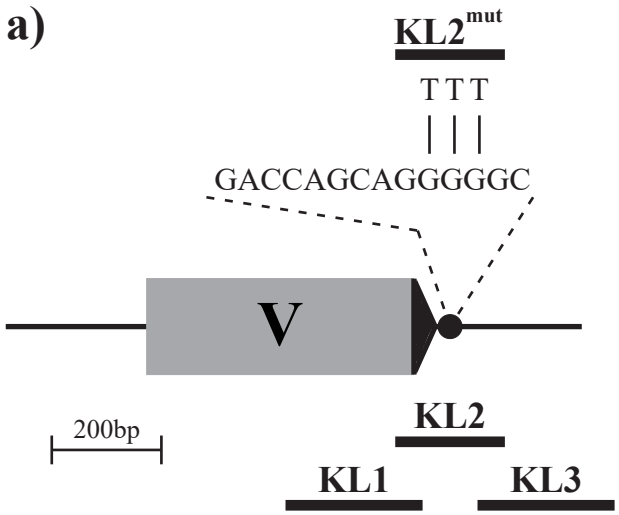
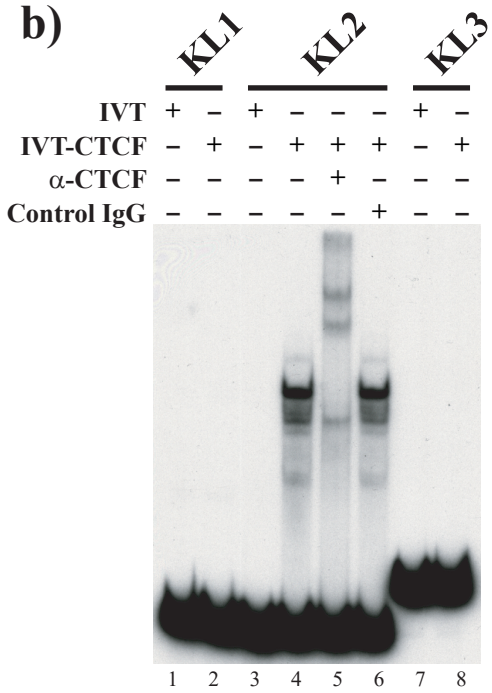
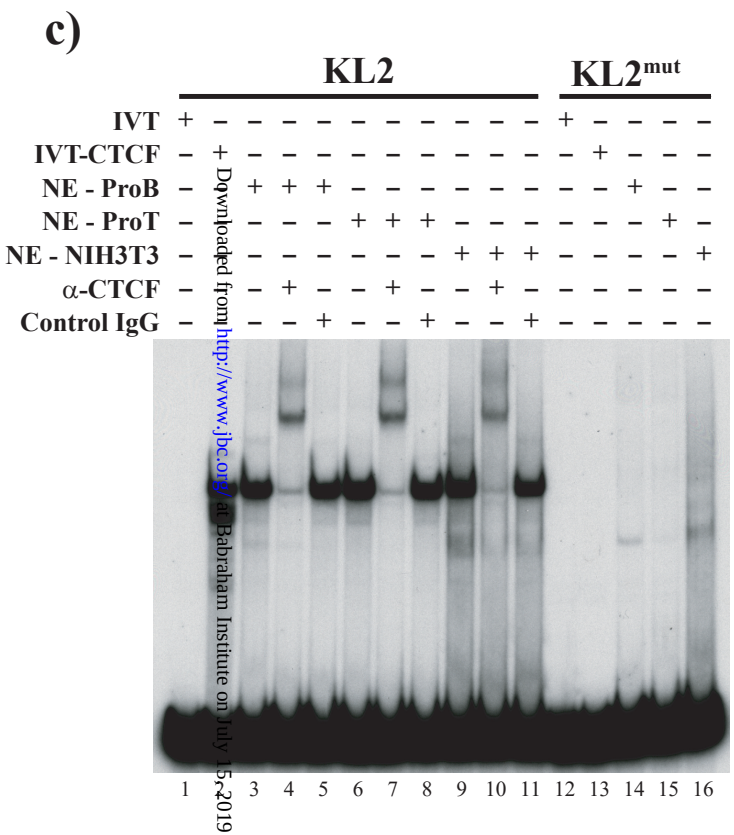
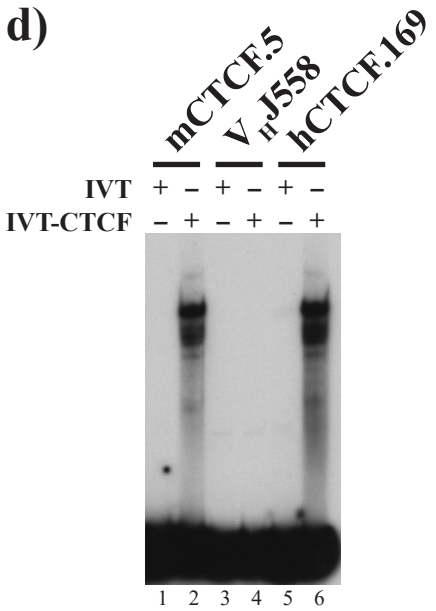


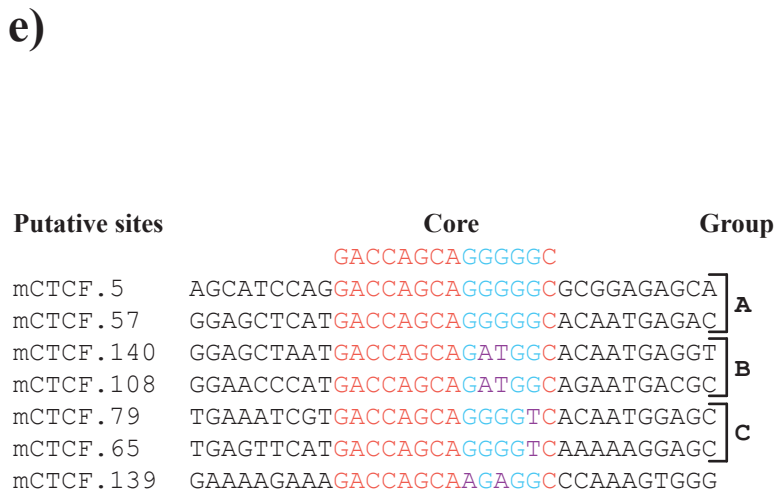
Figure 1. A high density of CTCF sites is found within the V_H domains of the murine and human IgH loci. a) Schematic of CTCF sites within the murine Ig heavy chain locus. Red and black vertical lines represent the location of upstream/intergenic and RSS-associated mVHCTCF sites, respectively. The general organizational structure of the murine IgH locus is shown with rectangles representing V (black), D (blue), J (green), and constant region (white) gene segments. Black ovals represent regulatory enhancer elements. b) Schematic of CTCF sites within the human Ig heavy chain locus. Diagram is as above. The numbers underneath the vertical lines denote CTCF hotspots with the first number indicating the number of putative CTCF sites within each hotspot and the second number indicating the length of DNA encompassed within each hotspot. c) Consensus sequence of the murine and human V_H CTCF sites: i) enoLOGOS representation of the frequency of each DNA nucleotide at each position within the murine V_H CTCF sites; ii) enoLOGOS representation of the consensus sequence of the human VH CTCF sites. For reference, the consensus CTCF motif at the mouse and human *Igf2/H19* imprinting control regions is CCGCGNGGNGGCAG, the consensus CTCF motif at the chicken β-globin FII 5'HS4 element isCCGCTAGGGGGCAG, and the consensus human CTCF binding site based on genome-wide ChIP-chip analysis (47) is CCASYAGRKGGCRS. Boxes highlight the core CTCF motif as referred to in the text, with nucleotide numbering provided above. d) Comparison of the consensus sequences of the murine RSS-associated (i) and upstream/intergenic (ii) CTCF binding sites. e) Comparison of the consensus sequences of the human RSS-associated (i) and upstream/intergenic (ii) CTCF binding sites.











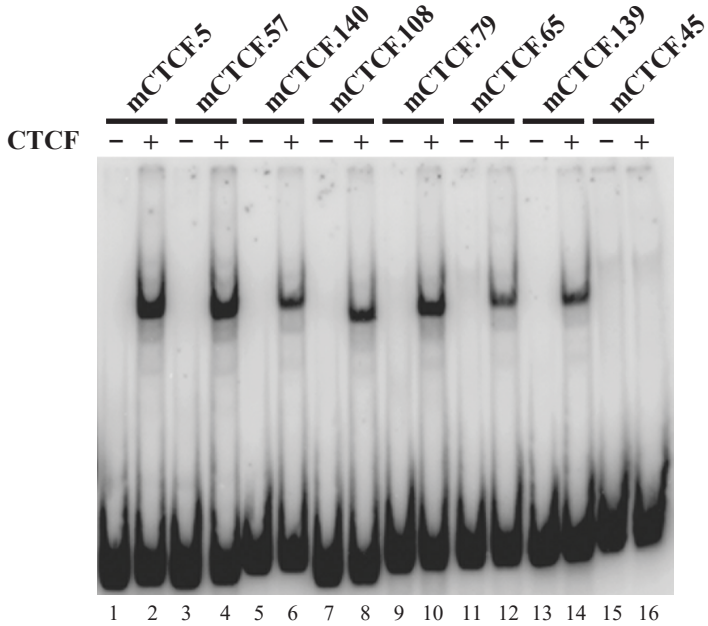


Figure 2. CTCF binds to the putative V_H CTCF sites *in vitro.* a) Schematic of the 7183.2.3 genomic segment drawn to scale depicting the location of the DNA probes used in the EMSAs in panels B and C. VH7183.2.3 segment coding sequence (gray rectangle); RSS (black triangle); mCTCF.5 site (black circle). The sequence of the targeted DNA transversion of the three central guanine residues within the CTCF present in Probe $KL2^{mut}$ is shown. b) Probe KL2 which encompasses mCTCF.5 is bound by *in vitro* translated CTCF (IVT-CTCF) and super-shifted by an α-CTCF antibody (CTCF-IgG). IVT: *in vitro* translation reaction lacking specific cDNA. c) Point mutations in mCTCF.5 disrupt binding of IVT-CTCF as well as endogenous CTCF present in nuclear extracts from pro-B (NE-pro-B), Pro T (NE-Pro T) and NIH3T3 (NE-NIH3T3) cells. d) A human V_H CTCF site is bound by CTCF. Murine and human EMSA probes are as indicated. No binding to the region surrounding murine V_H segment J558.69.170 (V_HJ558) is observed, indicating the absence of a cryptic CTCF site. e) CTCF binds to distinct subclasses of mCTCF sites. Left panel shows sequences of representative members of distinct groups of CTCF sites containing the same substitutions within the central G pentad, but differing in sequences flanking the core CTCF site. Right panel shows EMSA of the corresponding labeled DNA probes.

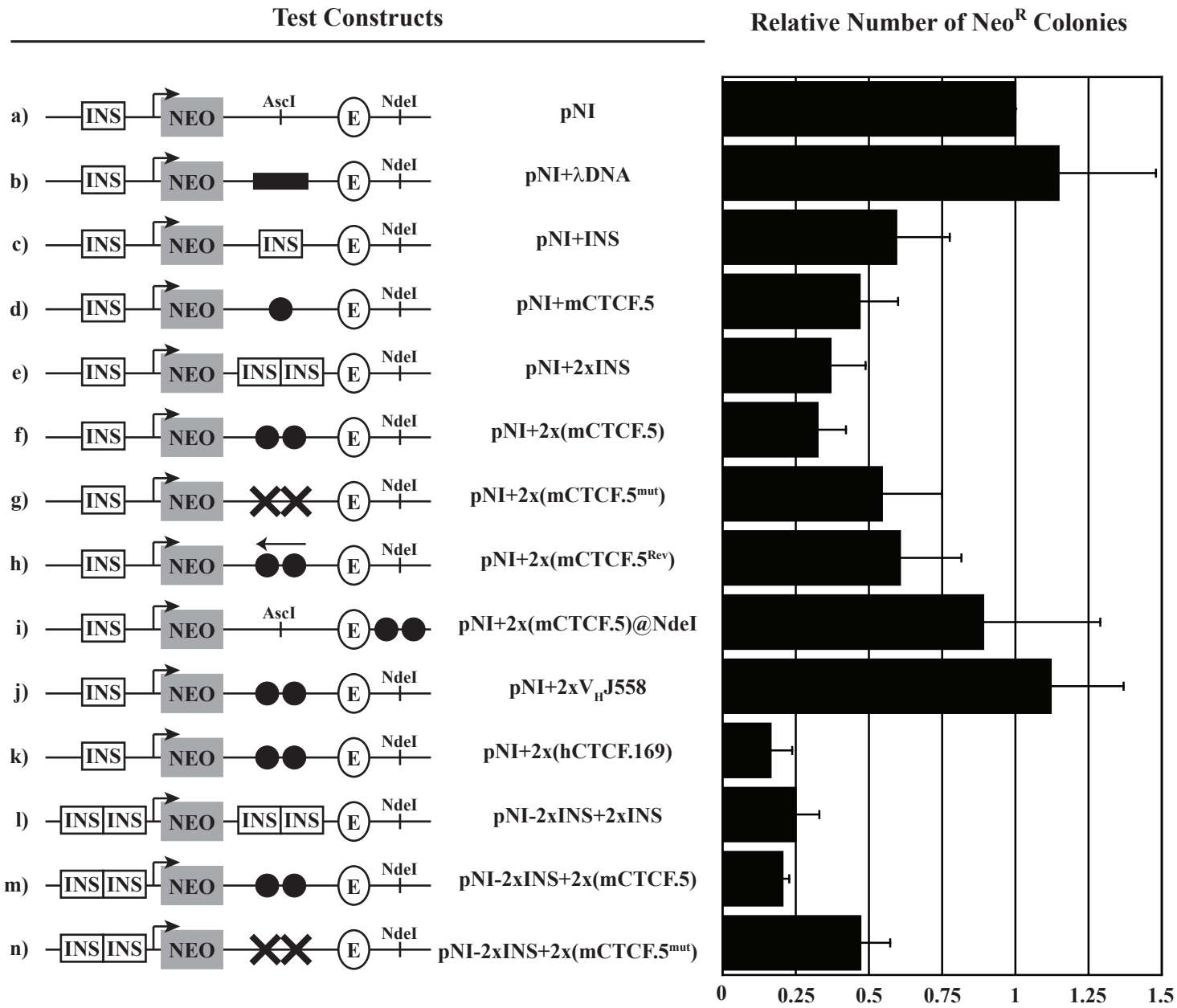


Figure 3. The CTCF sites located within IgH loci possess strong enhancer-blocking activity. The constructs used in the enhancer-blocking assay are shown on the left, while the extent of enhancer blocking (number of neomycin resistant colonies normalized to the backbone vector *pNI*) is shown on the right. Data shown represents the average ± S.D. of at least two independent enhancer-blocking experiments. The chicken β-globin 5'HS4 insulator element (INS), the murine β-globin 5'HS2 locus control element (E), the neomycin resistance cassette (NEO) driven by the human γ-globin promoter (arrow) and restriction enzyme sites used for cloning (AscI and NdeI) are shown; The 2.3kb λ phage DNA fragment is indicated as a black rectangle and black circles refer to the indicated V_H gene-segment fragment encompassing the downstream CTCF site. The presence of a mutated CTCF site is indicated by a black "X," and a V_H gene segments that is oriented in the antisense direction is indicated by a left-facing arrow.

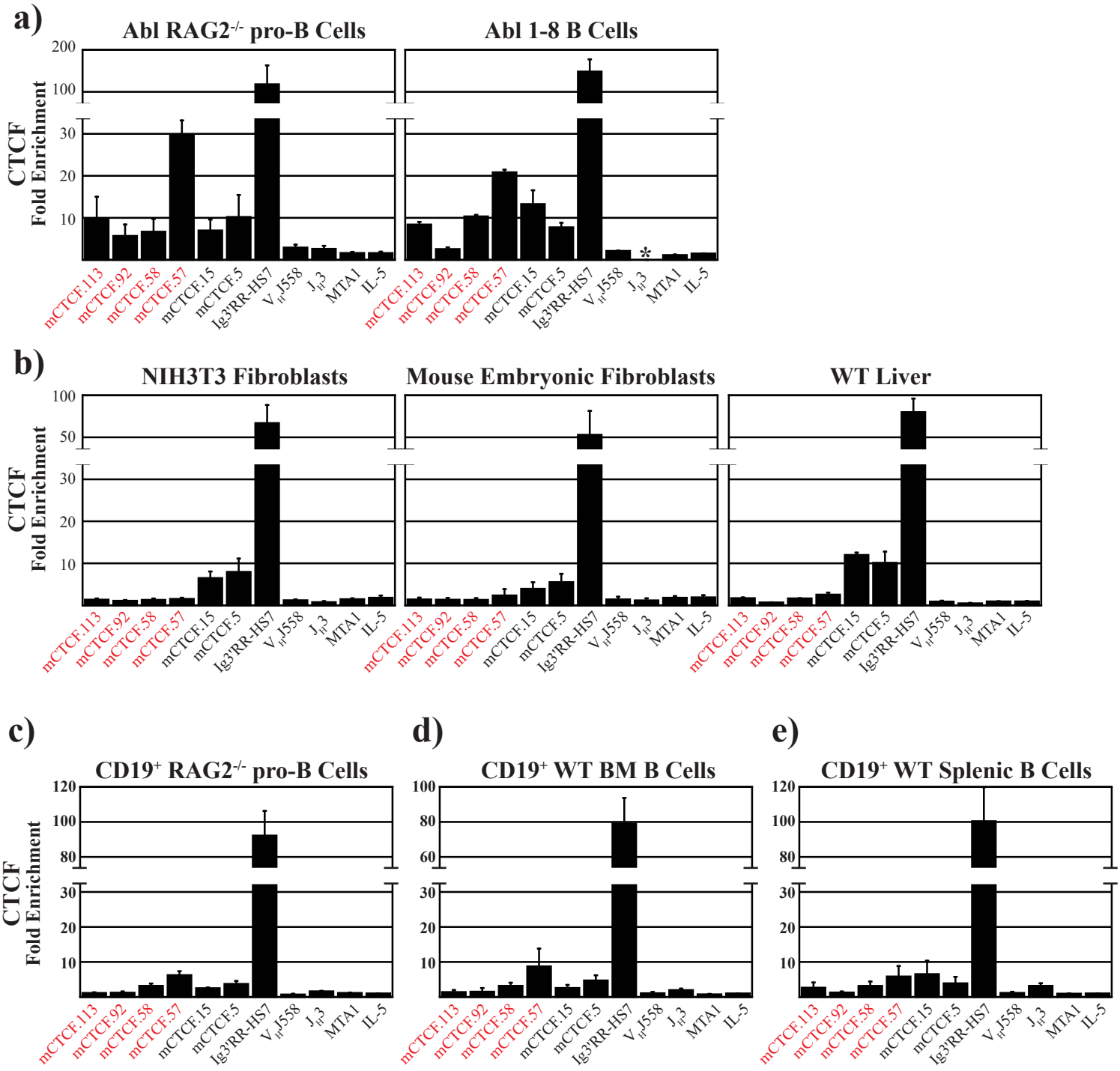
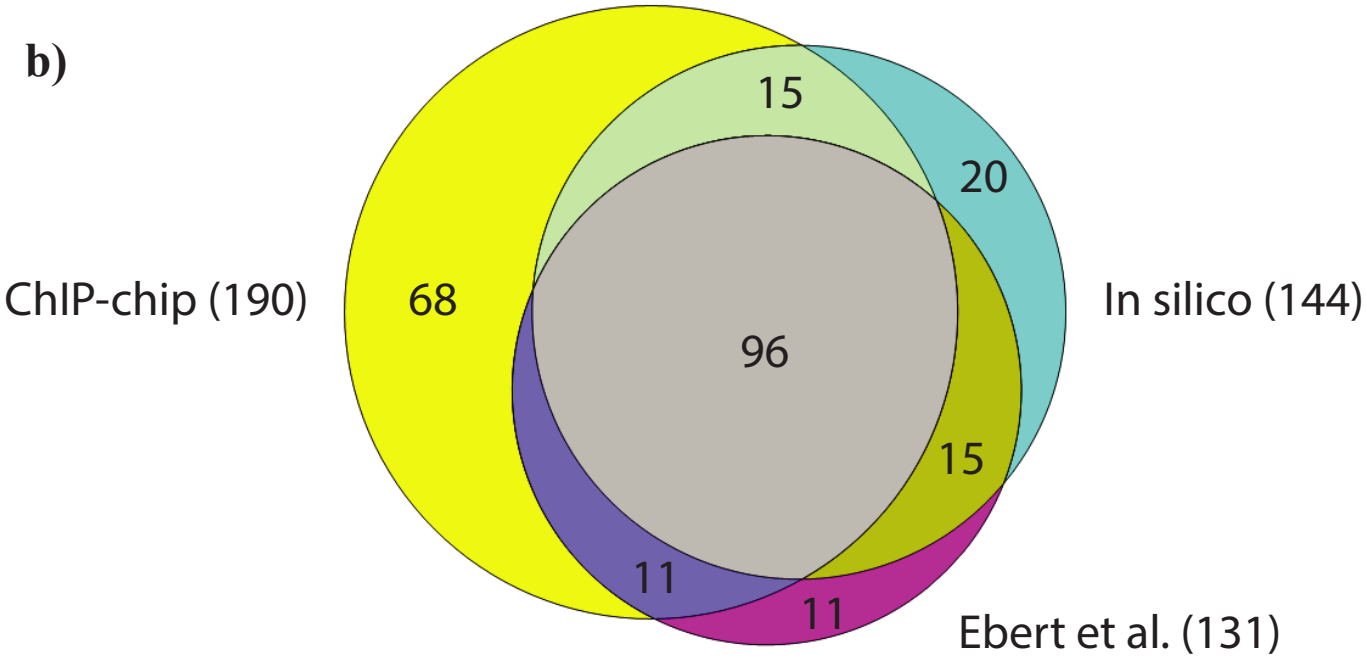


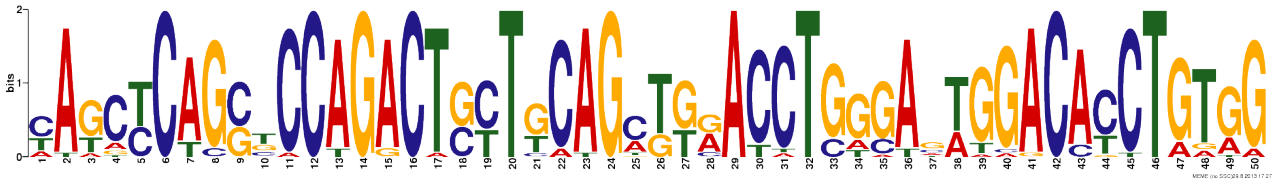
Figure 4. CTCF binds to its cognate sites *in vivo*. Chromatin immunoprecipitation with antibodies to CTCF was performed from the indicated cell lines and tissues. Fold-enrichment is shown on the y-axis. A break within the y-axis of each panel represents a non-linear jump in fold-enrichment values in order to accommodate the levels observed from the positive control. All fold-enrichments represent the average ± S.D. of at least 3 independent chromatin IPs. Primers for the indicated mV_HCTCF sites arranged 5' to 3' across the IgH locus (with respect to transcription) are described in Supplementary Table 4 (red: upstream/intergenic, black: RSS associated). Primers for the multiple CTCF sites in the 3' regulatory region of the IgH locus (positive control), the V_HJ558 and J_H3 gene segments (negative controls), and the MTA1 and IL-5 genes (negative controls) are shown. The "*" indicates that the JH3 DNA in the Abl 1-8 B cell line has been deleted by V(D)J recombination and, therefore, cannot be assayed.

a)

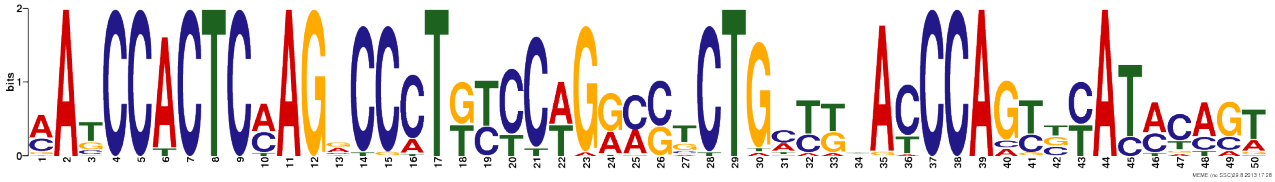
b)



c) **i.**



ii.



iii.

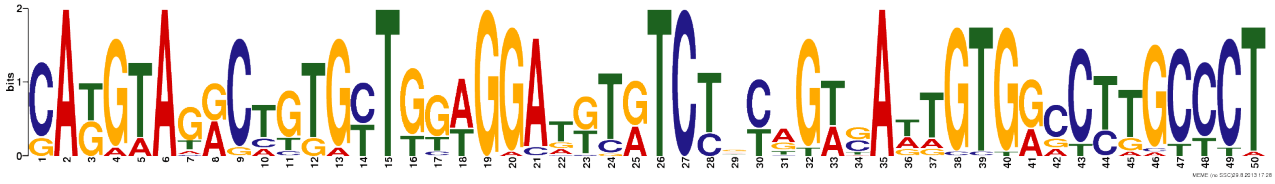


Figure 5. mVH-CTCF consensus motif is a major determinant of CTCF binding at the murine IgH locus. Chromatin immunoprecipitation with an α-CTCF antibody was performed on $CD19^{+}$ Pro-B cells that were isolated from 8-9 week old $RAG2^{-/-}$ mice, and expanded for 3 days in the presence of IL-7 (10 ng/mL). Cy3-labeled input DNA and Cy5-labeled immunoprecipitated DNA were hybridized to customized tiling DNA microarrays, and peaks were called by the Ringo method (61). A) Area-proportional Venn diagram showing the overlap between predicted CTCF binding sites in the VH domain of the murine IgH locus (cyan) and observed CTCF peaks (yellow). B) Area-proportional Venn diagram showing the overlap between predicted CTCF binding sites (cyan), CTCF peaks we observed by ChIP-chip (yellow), and CTCF peaks observed previously by ChIP-seq (9). C) enoLOGOS representation of three distinct sequence motifs identified by MEME (42) analysis of the 79 CTCF peaks observed by ChIP-chip that did not contain an mVH-CTCF consensus motif.

a)

b)

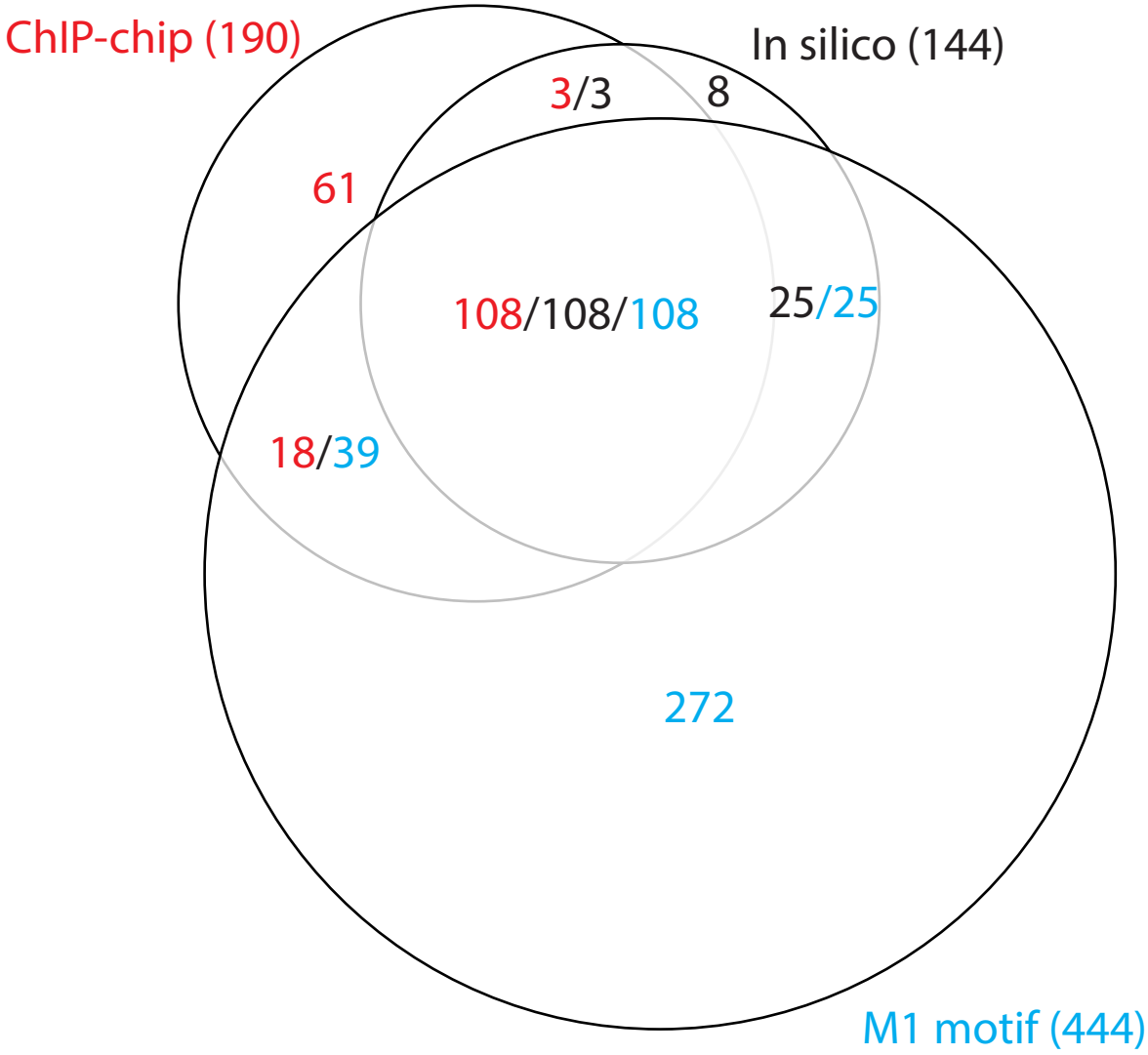


Figure 6. M1 motif is present a majority of the predicted and observed CTCF binding sites at the murine IgH locus. A) enoLOGOS representation of the M1 motif (44). B) Area-proportional Venn diagram showing the overlap between predicted CTCF sites (black text), observed CTCF peaks (red text), and M1 motif (cyan).

The murine IgH locus contains a distinct DNA sequence motif for the chromatin regulatory factor CTCF

David N. Ciccone, Yuka Namiki, Changfeng Chen, Katrina B. Morshead, Andrew L. Wood, Colette M. Johnston, John W. Morris, Yanqun Wang, Ruslan Sadreyev, Anne E. Corcoran, Adam G.W. Matthews and Marjorie A. Oettinger

J. Biol. Chem. published online July 8, 2019

Access the most updated version of this article at doi: 10.1074/jbc.RA118.007348