Complete Sequence Assembly and Characterization of the C57BL/6 Mouse Ig Heavy Chain V Region

Colette M. Johnston, Andrew L. Wood, Daniel J. Bolland and Anne E. Corcoran

*J Immunol* 2006; 176:4221-4234; doi: 10.4049/jimmunol.176.7.4221

http://www.jimmunol.org/content/176/7/4221

References

This article cites 90 articles, 42 of which you can access for free at: http://www.jimmunol.org/content/176/7/4221.full#ref-list-1

Why *The JI*? Submit online.

- **Rapid Reviews!** 30 days* from submission to initial decision
- **No Triage!** Every submission reviewed by practicing scientists
- **Fast Publication!** 4 weeks from acceptance to publication

*average

Subscription

Information about subscribing to *The Journal of Immunology* is online at: http://jimmunol.org/subscription

Permissions

Submit copyright permission requests at: http://www.aai.org/About/Publications/JI/copyright.html

Email Alerts

Receive free email-alerts when new articles cite this article. Sign up at: http://jimmunol.org/alerts
Complete Sequence Assembly and Characterization of the C57BL/6 Mouse Ig Heavy Chain V Region

Colette M. Johnston, Andrew L. Wood, Daniel J. Bolland, and Anne E. Corcoran

The mechanisms that regulate variable (V) gene selection during the development of the mouse IgH repertoire are not fully understood, due in part to the absence of the complete locus sequence. To better understand these processes, we have assembled the entire 2.5-Mb mouse IgH (Igh) V region sequence of the C57BL/6 strain from public sequences and present the first complete annotated map of the region, including V genes, pseudogenes, repeats, and nonrepetitive intergenic sequences. In so doing, we have discovered a new V gene family, VH16. We have identified clusters of conserved region-specific intergenic sequences and have verified our assembly by genic and intergenic Southern blotting. We have observed that V pseudogenes are not evenly spread throughout the V region, but rather cluster together. The largest J558 family, which spans more than half of the locus, has two strikingly different domains, which suggest points of evolutionary divergence or duplication. The 5' diversity of gene segments for recombination, the ultimately expressed on the B cell surface as an IgH polypeptide, DJ gene segment to produce a V(D)J rearranged gene. This is terminal in early B lymphocytes. First, a D gene is recombined with a chromosome 12 is the first to undergo ordered V(D)J recombination. The Igh locus on mouse chromosome 12 is the first to undergo ordered V(D)J recombination in early B lymphocytes. First, a D gene is recombined with a J gene, and second, one of many V genes is recombined with this DJ gene segment to produce a V(D)J rearranged gene. This is ultimately expressed on the B cell surface as an IgH polypeptide, which associates with an IgL to form the BCR. To provide a large diversity of gene segments for recombination, the Igh locus, in common with the IgL locus, is highly complex. It spans a region of ~3 Mb, containing 8 constant region genes, 4 J genes, 10–13 D genes and an unknown number of V genes, estimated to be 150 (1–3). These genes have been classified into 15 families based on sequence homology (4). Understanding the genomic organization of this large group of genes and the consequent constraints on their recombination ability is crucial to understanding Ab diversity. Several groups have mapped the position of functional V genes by deletion mapping (4, 5) and yeast artificial chromosome (YAC)4 contig assembly (1), which have given much useful information, and have predicted, with reasonable accuracy, the relative positions of genes within the locus. However, these studies have been limited in accuracy about the largest families, in particular the D-distal J558 family, which contains approximately half of the V genes, and have not been able to provide information about V pseudogenes, repeats, and intergenic sequences.

It is believed that all of the functional V genes in the Igh locus are used in the large number of V(D)J recombinations that occur in a B cell population, albeit they may not all survive to be included in the mature B cell repertoire. This assumption has not been verified due to the absence of the complete locus sequence. However, there is a large amount of variation between mouse strains, both in numbers and relative usage of V genes (6). Studies in BALB/c, commonly used for immunological studies, have suggested that gene usage varies throughout ontogeny, with the 3' V genes preferentially recombined in fetal liver B cells (7, 8), and early in bone marrow B cell development (9). In contrast, the C57BL/6 strain does not appear to exhibit this 3' recombination bias (10–12), suggesting strain-specific developmental differences in usage of V genes. Despite the lack of complete locus sequence, many factors affecting recombination have been studied in small V gene families, in which the number and relative position of genes could be predicted reasonably accurately by Southern blotting and deletion mapping, and for which sequence was available. These include the recombination signal sequence (RSS), heptamer, nonamer (13, 14), and spacer sequences (15); relative V gene promoter strength (16–18); distance between rearranging elements (19); requirement for additional regulatory elements (18). However, with notable exceptions (20), these studies have relied on sequences of the genes

---

4 Abbreviations used in this paper: YAC, yeast artificial chromosome; BAC, bacterial artificial chromosome; Bright, B cell regulator of IgH transcription; DICE, downstream Ig control element; RSS, recombination signal sequence; SATB1, special AT-rich sequence binding protein; Inn, initiator.
themselves, or at most a small amount of flanking sequence, and have not been able to assess the role of chromatin context, provided by the relatively large tracts of intervening sequence (2–50 kb) between the relatively small (500 bp) V genes. Further, it recently has become increasingly clear that large-scale chromatin remodelling events, such as nuclear relocalization (21), antisense intergenic transcription (22), and locus contraction (23), precede V(D)J recombination. These are not confined to genes, but rather affect large chromatin domains. Thus, to gain a complete understanding of how the Ig repertoire is established, it will be necessary to investigate the large noncoding regions between genes.

In this study, we set out to assemble and annotate the primary sequence data of the mouse Igh V region from publicly available sources, including the mouse genome sequencing project (Ensembl) (24), to provide a detailed picture of the locus, including exact numbers and positions of genes within each family and their correct genomic context. This assembled and fully annotated sequence is the first report that places V genes relative to flanking regions, pseudogenes, repeats, and nonrepetitive intergenic sequences, enabling study of their role in V(D)J recombination. Further, studies of V(D)J recombination in the C57BL/6 mouse strain have been limited. Because the C57BL/6 genome sequence is being assembled in Ensembl, and this is likely to be the major mouse strain for future study, it will be vital to investigate the recombination patterns of V genes in this strain with the benefit of complete locus knowledge.

Materials and Methods

Sequence assembly

BLAST software (25) searches of National Center for Biotechnology Information (NCBI) and Ensembl databases, using known published V gene sequences (26, 27) and those deposited in public V gene databases, including the International ImMunoGeneTics Information System (http://imgt.cines.fr/), identified bacterial artificial chromosome (BAC) sequences from C57BL/6 mice containing Igh (Igh) V region genes. Sequences were then assembled using Sequencher software (Gene Codes) and confirmed by visual inspection of overlaps.

Sequence analysis

Analysis of the sequence and identification of V genes and non-Ig genes was performed using NIX (www.hgmp.mrc.ac.uk/NIX), a web-based platform that combines results from the following complementary sequence analysis programs and databases, thereby providing optimal conditions for gene predictions: GRAIL (http://compbio.ornl.gov/GraiI-1.3/), Fex (http://www.software.com), Hexon (http://www.software.com), MZEF (http://rulal.cshl.org/tools/genefinder/), GeneMark (http://opal.biology.gatech.edu/GeneMark/), GeneFinder (http://www.informatics.jax.org/), Hexon (http://www.informatics.jax.org/), FGene (www.fgene.org), BLAST (www.ncbi.nlm.nih.gov/BLAST/), PolyPhen (www.software.com), and rNAscan (www.genetics.wustl.edu/eddy/rnAscan-SE/). This platform was discontinued after July 2005 because of the restructuring of the Human Genome Mapping Project, but individual programs are still available. Annotation of V genes and pseudogenes (leader, exon, RSS heptamer, and nonamer) was performed in MacVector software (Oxford Molecular Group) and exported to GenBank. The complete annotated sequence, including start and end positions of leader, V gene exon, and start of RSS heptamer and nonamer is available from EMBL (accession no. BN008872). The phylogenetic tree was constructed using TreeView software (28) (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html) from multiple sequence alignments generated by the ClustalW software tool at the European Bioinformatics Institute (29) (www.ebi.ac.uk/clustalw/index.html). Alignment of transcription factor binding sites at V gene flanking regions was performed using Genomatix MathInspector software (www.genomatix.de) (30).

Analysis of repeat sequences

The RepeatMasker software program screens DNA sequences for low-complexity DNA sequences and interspersed repeats. The sequence was submitted to the RepeatMasker server (http://woody.embl-heidelberg.de/Repeatmasker), originally developed by A. Smit and P. Green (unpublished data). Analysis was performed using RepeatMasker version 2002/05/15 with the rodent repeat database, cross_match software (version 0.990329; http://repeatmasker.org/), and the RepBase database (version 7.4; G.I.R.I. (http://www.girinst.org/repbase/)). The most sensitive setting was used for maximum accuracy.

Southern blotting

Genomic DNA from C57BL/6 kidney was digested with EcoRI or HindIII, separated by agarose gel electrophoresis, and blotted onto Bionyne B nylon membrane (Pall). V gene family and intergenic PCR products were cloned into pGEM-T Easy vectors (Promega), excised, purified, radioactively labeled with [α-32P]dCTP, and used to probe Southern blots. Southern blot hybridization results were analyzed with a phosphorimager (FujiFilm).

Results

Sequence assembly of the mouse IghH locus

The Igh locus has been partially assembled by the Ensembl project (24) (www.ensembl.org/). However, at the date of manuscript submission, the Ensembl assembly is incomplete, with single contig coverage over large parts of NCBI Build m34 (freeze September 2005), many contigs in draft form, and over 40 large and small gaps encompassing ~450 kb of sequence. Thus, we chose to assemble the locus manually, which has enabled us to include many BAC sequences not available on Ensembl, order the contigs correctly, provide at least two-fold coverage over large parts, and close all of the gaps. The Igh locus spans 3.3 Mb on mouse chromosome 12, from the 3' enhancer at position 108.5 Mb to the first non-V gene at 111.85 Mb (Ensembl). We have assembled 2.75 Mb of the Igh locus, which contains the complete mouse Igh V region (2.5 Mb) and upstream flanking sequence (109.11–111.85 Mb). We used Sequencher software, which enabled us to incorporate and align all BACs that we found by BLAST analysis (Fig. 1). For simplicity, only the most complete BACs are shown. A list of other overlapping BACs used is available on request. We noticed a number of mistakes in the ordering of contigs in the Ensembl assembly. For example, contig AC074328 overlaps both AC073939 and AC087166 at the 5' end of the Igh locus in our assembly but is placed further 3' in Ensembl. This may be because this locus contains a large number of very similar V genes and a high density of LINE1 and other repeats (31), both of which make automated assembly less accurate.

Analysis of our assembled sequence using NIX software (www. hgmp.mrc.ac.uk/) enabled us to identify sequences with homology to known V gene fragments, including diverged pseudogenes and gene remnants. Further, it aided identification of germline genes, because many genes deposited with databases are recombined, mature, somatically hypermutated, and, hence, nongermline V genes. V genes and pseudogenes were assigned to families using the nomenclature originally suggested by Brodeur and Riblet (32). A sequence was assigned to a V family if it had at least 80% identity at the nucleotide level over the entire coding exon sequence (excluding the leader). A sequence was classed as a V gene if it had an intact translation initiation codon (ATG), splice junctions, RSS, had no in-frame stop codons or frameshifts, and had a minimum length of 291 bp. Where a sequence had all the features of a normal coding gene, except for a noncanonical splice junction, we classed it as a coding gene, even though it may be nonfunctional. However, this applied to only one gene (J558.13.103), which had a GT-to-GC splice site change, which allows splicing, but at lower efficiency. A sequence was classed as a V pseudogene if it contained stop codons, frameshifts, lacked an ATG, or had a significantly altered RSS (spacier longer or shorter than 23 ± 1 bp; any alterations in the first three nucleotides of the heptamer (CAC); ≥2 mismatches elsewhere in the heptamer; or >1 G residue in the nonamer). Most pseudogenes had more than one of the above
Defects. For those with only one defect, this was usually in the RSS. If no evidence of expression of the V(D)J rearranged gene was found in cDNA and EST libraries, these were classed as pseudogenes. Pseudogenes were assigned to specific V families if they had at least 80% nucleotide identity to functional genes of that family. In addition, the 7183 and 3609 families have three and one pseudogene members, respectively, that are related to other pseudogenes of that family with >80% similarity, but which have <80% similarity to the genic members, and we included these in the family. A small number of pseudogenes (hereafter named PG), although recognizable as V gene remnants, could not be assigned to any particular family, due to similar homology levels to several families. In total, based on these definitions, 110 genes and 85 pseudogenes were identified (Table I).

Features of the mouse IgH locus V region

The distribution of V genes and V pseudogenes in the IgH V region is shown in Fig. 2. All the V genes and pseudogenes are in the same orientation, indicating that there have been no obvious inversions during the evolution of this locus. There are large intergenic distances (2–50 kb) between the relatively small genes (<500 bp). The 3′ end of the V region is more compact than the 5′ end, with average intergenic distance (between genes and/or pseudogenes) <10 kb, compared with 25 kb for the 5′ end.

The relative positions of V gene families generally confirms previous mapping studies (1, 4, 5). The J558 family spans over 50% of the V region (~1.53 Mb) at the 5′ end (Fig. 2). Notable differences include the complete interspersion of the 3609 family with the 5′ Mb of the J558 region, except for a single member 3′ of the J558 family, in contrast to its previous estimated position at the 3′ end of, or 3′ of the J558 family (4). The interspersion of members of the VH15 and VH10 families with the 3′ end of the J558 region shows that the J558 region is not completely separate from the other families, contrary to previous predictions (4). The “middle” families (i.e., all except J558, 3609, 7183, and Q52) show a reasonably high level of interspersion and family members are not necessarily clustered together. The notable exception is the five-member J606 family, which is not interspersed with any other family. At the 3′ end of the V region, the Q52 and 7183 families are completely interspersed but do not overlap with any other gene families. Pseudogenes belonging to a particular family always map to the region occupied by the functional genes. Where pseudogenes are present, the overall density of V gene sequences is increased. This is most marked in the 7183/Q52 region and the 3′ end of the J558 region.

### Table I. Summary of the V region genes and pseudogenes in the mouse IgH locus grouped by family

<table>
<thead>
<tr>
<th>V Gene Family</th>
<th>Genes</th>
<th>Pseudogenes</th>
</tr>
</thead>
<tbody>
<tr>
<td>7183</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>Q52</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>S107</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>SM7</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>36–60</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>VH11</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>X24</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>VH16</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>VGAM3.8</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>3609N</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>VH12</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>J606</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>VH10</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>VH15</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3609</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>J558</td>
<td>52</td>
<td>37</td>
</tr>
<tr>
<td>Unclassified pg</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>Total 195</td>
<td>110</td>
<td>85</td>
</tr>
</tbody>
</table>
The upstream boundary of the IghV region is marked by two non-Ig genes, vasoactive intestinal peptide receptor type 2 (\textit{Vipr2}) (first exon; Fig. 2) and zinc finger protein type 386 (\textit{Zfp386}). \textit{Vipr2} is expressed from the opposite strand to the V genes and has previously been experimentally mapped to the telomeric end of chromosome 12 (33). We are confident that \textit{Vipr2} lies outside of the Igh locus, because the entire region from the telomere to \textit{Vipr2} has synteny with human chromosome 7, yet the human Igh locus maps to chromosome 14. Thus, this \textit{Vipr2} positioning is not conserved across species and thus is not functionally important (33) and may indeed be an evolutionary chromosomal breakpoint in the human lineage. The 3' end of the V region is flanked by the downstream marker D12Mit263, positioned at 109.05 Mb.

Identification of a novel V gene family

There is an additional gene located at 2.005 Mb, between \textit{VGAM3.8.1.57} and \textit{SM7.3.54}, which does not fit the criteria for inclusion in any of the 15 V gene families. This gene has an open reading frame, a normal leader and RSS, and conserved splice junctions, but does not have >80% identity to any known V gene. Its closest homology is to 7183.20.37 (72% identity at the nucleotide level over the exon sequence only), and it shows nucleotide identity in the 60–70% range to 7183, X24, VH10, VH11, SJ07, J606, and 3609N. At the protein level, the gene has homology in the range 47–57% to predicted proteins from these families. As a comparison, the single-member \textit{VH12} gene family has a closest identity of 73–78% (to the 3'6–6') family and the single-member \textit{VH15} family has 66–68% identity to the SM7 family at the nucleotide level. We have named this novel gene \textit{VH16} (\textit{VH16.1.55}), in accordance with the existing nomenclature.

Identification of non-V pseudogenes in the V region

We have not found any functional non-V genes within the V region but have found numerous pseudogenes, both processed and unprocessed, detailed in Table II. Between the second and third J558 gene at the 5' end is a processed ornithine decarboxylase pseudogene, previously mapped within the first 9% of the Igh V region (34). This is in two fragments, the first corresponding to exons 1–11, and the second to exons 11 and 12. This and other processed pseudogenes may have arisen from locus evolutionary

**FIGURE 2.** Scale map of V genes and pseudogenes in the mouse IgH locus V region. The 110 genes and 85 pseudogenes are depicted as triangles pointing in the direction of their orientation in the locus. Genes are numbered systematically and colored according to family. Pseudogenes are shown as open triangles, with colors corresponding to their family of origin. The system of V gene nomenclature is as follows: (V gene family name).(gene number within family).(gene number within locus), numbered from the 3' end. The 19 unclassified pseudogenes, depicted as open gray triangles, are included in the total locus numbering, and are called PG 1 to 19. The gray box toward the bottom of the figure contains the “middle” V gene families located between the J558/3609 and the 7183/352 regions.

The upstream boundary of the Igh V region is marked by two non-Ig genes, vasoactive intestinal peptide receptor type 2 (\textit{Vipr2}) (first exon; Fig. 2) and zinc finger protein type 386 (\textit{Zfp386}). \textit{Vipr2} is expressed from the opposite strand to the V genes and has previously been experimentally mapped to the telomeric end of chromosome 12 (33). We are confident that \textit{Vipr2} lies outside of the Igh locus, because the entire region from the telomere to \textit{Vipr2} has synteny with human chromosome 7, yet the human Igh locus maps to chromosome 14. Thus, this \textit{Vipr2} positioning is not conserved across species and thus is not functionally important (33) and may indeed be an evolutionary chromosomal breakpoint in the human lineage. The 3' end of the V region is flanked by the downstream marker D12Mit263, positioned at 109.05 Mb.

Identification of a novel V gene family

There is an additional gene located at 2.005 Mb, between \textit{VGAM3.8.1.57} and \textit{SM7.3.54}, which does not fit the criteria for inclusion in any of the 15 V gene families. This gene has an open reading frame, a normal leader and RSS, and conserved splice junctions, but does not have >80% identity to any known V gene. Its closest homology is to 7183.20.37 (72% identity at the nucleotide level over the exon sequence only), and it shows nucleotide identity in the 60–70% range to 7183, X24, VH10, VH11, SJ07, J606, and 3609N. At the protein level, the gene has homology in the range 47–57% to predicted proteins from these families. As a comparison, the single-member \textit{VH12} gene family has a closest identity of 73–78% (to the 3'6–6') family and the single-member \textit{VH15} family has 66–68% identity to the SM7 family at the nucleotide level. We have named this novel gene \textit{VH16} (\textit{VH16.1.55}), in accordance with the existing nomenclature.

Identification of non-V pseudogenes in the V region

We have not found any functional non-V genes within the V region but have found numerous pseudogenes, both processed and unprocessed, detailed in Table II. Between the second and third J558 gene at the 5' end is a processed ornithine decarboxylase pseudogene, previously mapped within the first 9% of the Igh V region (34). This is in two fragments, the first corresponding to exons 1–11, and the second to exons 11 and 12. This and other processed pseudogenes may have arisen from locus evolutionary
Igh genome (39%) (37). LINE (L1) elements constitute 40.4% of the proportion of interspersed repeats (52.4%) than most of the mouse (SINEs) and gene-poor inactive heterochromatin (LINEs) (36).

Interspersed repeat sequences. Different repeats are known to be This assembly has enabled detailed analysis of type and position of repeats.

The mouse V region has a high proportion of interspersed repeats. This assembly has enabled detailed analysis of type and position of interspersed repeat sequences. Different repeats are known to be differentially associated with gene-rich active euchromatin (SINEs) and gene-poor inactive heterochromatin (LINEs) (36). RepeatMasker analysis revealed that the region contains a greater proportion of interspersed repeats (52.4%) than most of the mouse genome (39%) (37). LINE (L1) elements constitute 40.4% of the Igh V region sequence, the value predicted by detection of L1 sequence at YAC ends during YAC assembly of the locus (40%) (1, 31). This is similar to other AgR loci, such as the Igk locus (41%) (38), and in marked contrast to 15% in the mouse autosomal genome overall (39). The region also contains a correspondingly lower proportion of SINEs (2.1%) than the rest of the genome. Both of these features have been proposed to be characteristic of monospecifically expressed loci (39). In contrast, the human IgH V region contains a much lower proportion of LINEs (23%), which is not significantly different from the rest of the human genome (35, 40). Of the 338 identifiable LINE elements in the human Igh V region, only two are full length. In our mouse sequence, we identified 22 full-length LINEs out of a total of 1016 identifiable elements. It has been estimated that a significant proportion of full-length LINEs in the mouse genome are capable of active retrotранспонсion (41). Thus, the large number identified in the mouse V region may have contributed to the greater expansion of the mouse locus compared with the human locus (discussed below).

Table II. Nonvariable region genes and pseudogenes within the mouse Igh V region locus

<table>
<thead>
<tr>
<th>Name</th>
<th>Start</th>
<th>End</th>
<th>Str.</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vipr2</td>
<td>201</td>
<td>298</td>
<td>(-)</td>
<td>Vasoactive intestinal peptide receptor 2 ex1</td>
</tr>
<tr>
<td>Zfp386</td>
<td>11702</td>
<td>13394</td>
<td>(-)</td>
<td>Zinc finger protein 386 (Kruppel-like) ex3</td>
</tr>
<tr>
<td>Zfp386</td>
<td>17346</td>
<td>17577</td>
<td>(-)</td>
<td>Zinc finger protein 386 (Kruppel-like) ex2</td>
</tr>
<tr>
<td>Zfp386</td>
<td>24433</td>
<td>24522</td>
<td>(-)</td>
<td>Zinc finger protein 386 (Kruppel-like) ex1</td>
</tr>
<tr>
<td>Odc1</td>
<td>78928</td>
<td>80116</td>
<td>(+)</td>
<td>Ornithine decarboxylase ex1–10</td>
</tr>
<tr>
<td>Odc1</td>
<td>89856</td>
<td>90035</td>
<td>(+)</td>
<td>Ornithine decarboxylase ex11–12</td>
</tr>
<tr>
<td>Olf141</td>
<td>137380</td>
<td>138138</td>
<td>(+)</td>
<td>Olfactory receptor 141 pseudogene</td>
</tr>
<tr>
<td>Q7TQDG</td>
<td>175243</td>
<td>179493</td>
<td>(-)</td>
<td>Signaling molecule ATTP ex1–13</td>
</tr>
<tr>
<td>Rpl7a</td>
<td>240067</td>
<td>240794</td>
<td>(+)</td>
<td>Ribosomal protein L7a</td>
</tr>
<tr>
<td>Hmgb2</td>
<td>297130</td>
<td>297735</td>
<td>(+)</td>
<td>High-mobility group box 2 ex2–5</td>
</tr>
<tr>
<td>Hmgb2</td>
<td>411290</td>
<td>412141</td>
<td>(+)</td>
<td>High-mobility group box 2 ex1–5</td>
</tr>
<tr>
<td>Cdc5</td>
<td>452512</td>
<td>455407</td>
<td>(+)</td>
<td>Cdc5-like/NpwBP pseudogene</td>
</tr>
<tr>
<td>Rpi7a</td>
<td>479805</td>
<td>480709</td>
<td>(+)</td>
<td>Ribosomal protein L7a</td>
</tr>
<tr>
<td>Toeb2</td>
<td>524683</td>
<td>525060</td>
<td>(+)</td>
<td>Transcription elongation factor B</td>
</tr>
<tr>
<td>Hmgb2</td>
<td>543987</td>
<td>544776</td>
<td>(+)</td>
<td>High-mobility group box 2 ex2–5</td>
</tr>
<tr>
<td>Hmgb2</td>
<td>593946</td>
<td>595111</td>
<td>(+)</td>
<td>High-mobility group box 2 ex1–5</td>
</tr>
<tr>
<td>Birc3</td>
<td>826533</td>
<td>826895</td>
<td>(+)</td>
<td>Baculoviral inhibitor of apoptosis ex3–6</td>
</tr>
<tr>
<td>Hmgb2</td>
<td>845465</td>
<td>846268</td>
<td>(+)</td>
<td>High-mobility group box 2 ex1–5</td>
</tr>
<tr>
<td>Hmgb2</td>
<td>894486</td>
<td>897235</td>
<td>(+)</td>
<td>High-mobility group box 2 ex1–5</td>
</tr>
<tr>
<td>Rad23b</td>
<td>180477</td>
<td>180550</td>
<td>(+)</td>
<td>RAD23b ex1–9 homolog (S. cerevisiae)</td>
</tr>
<tr>
<td>Wdr22</td>
<td>1882119</td>
<td>1882424</td>
<td>(+)</td>
<td>WD repeat domain 22</td>
</tr>
<tr>
<td>Rpl26</td>
<td>2245559</td>
<td>2245935</td>
<td>(+)</td>
<td>Ribosomal protein L26</td>
</tr>
<tr>
<td>Ankrd12</td>
<td>2440199</td>
<td>2441262</td>
<td>(+)</td>
<td>Ankyrin repeat domain 12 ex1–7</td>
</tr>
</tbody>
</table>

*Str. refers to the orientation of the gene in relation to V region genes. (+), same orientation; (-), opposite orientation; ex, exon. All listed genes are pseudogenes with 85–99% identity to the original gene from which the pseudogene is derived, except the first two genes listed, which are functional genes outside the V region. Gene location refers to the chromosomal position of the corresponding original gene.

![FIGURE 3. Dot plot of V region vs V region (masked). Alignment of the masked (y axis) vs unmasked (x axis) sequences of the mouse Igh chain locus V region. “Middle” families refers to the 11 families not specifically named that are all interspersed in this region. The positive strand was searched in this analysis, with a stringency of 95% identity over 50-bp windows.](http://www.jimmunol.org/Downloadedfromhttp://www.jimmunol.org)
Patterns of duplication and clustering in the V region

The Igh locus in the mouse has expanded by means of numerous duplication events, probably facilitated in part by the large amount of recombination that occurs naturally in the locus, and by the putative ability of the RAG genes to act as general transposases (42). To visualize the extent of duplication and identify patterns with the benefit of intergenic sequences we aligned the sequence against a repeat-masked copy of itself. This analysis is depicted in Fig. 3. Within the large J558 region, there are two strikingly different sequence patterns. In the 5' part, in which the J558 genes are very interspersed with 3609 genes, the whole region is relatively gene poor, with large intergenic distances (7–50 kb) between the genes. Because the V genes are small (500 bp), it is clear that there are extensive repeated sequences in many intergenic regions. In contrast, toward the 3' end of the J558 region, which contains no 3609 genes, there is a discrete domain comprising a tight clustering of J558 genes and pseudogenes, with intergenic distances ranging from 3 to 25 kb. This region exhibits a much greater level of duplication over both genic and intergenic regions. The 5' and 3' ends of the J558 region identified in the dot plot (Fig. 3) have very different numbers of associated pseudogenes. The 5' A region (genes J558.36pg.195 to 3609.2pg.138) occupies 1 Mb and contains 29 J558 genes, 13 J558 pseudogenes, 7 3609 genes, and 8 3609 pseudogenes. The 3' B region (genes J558.47.137 to J558.6.96) occupies 400 kb and contains 19 J558 genes and 23 pseudogenes. These contrasting patterns are strongly suggestive of a different pattern of duplication and evolution between the two parts of the J558 region.

Downstream of the A and B J558 regions, there is a segmental duplication of a cluster of four genes from different families (J558, PG, VH10, and J558), indicating a single duplication of this entire subdomain. Adjacent to this is the discrete J606 region. Further downstream, there is another tandem duplication involving genes from four families (VH11, 36-60, X24, and SM7). Interspersion of these small families is likely to have arisen by duplication of geographical regions containing clusters of V genes from different families. This would explain the apparent randomness of interspersion of V gene families, and suggests that, in a chromatin context, the families may be less important than geography. At the 3' end, the 7183 and Q52 families, which are completely interspersed with each other, form another separate domain, which, similar to the J558 region, is highly repetitive.

Sequence comparison between V gene families

Sequence relatedness between V region gene families was assessed by ClustalW analysis of all V gene sequences that had a full-length exon. This included all family-specific pseudogenes and unclassified pseudogenes. Sequences entered into Clustal did not include the leader exon. The alignments were used to create a phylogenetic tree, which accurately measures evolutionary diversity. The V genes fall into three distinct groups as shown in Fig. 4a. Group I includes the J558, SM7, VH15, and VGAM3.8 genes. It can be further divided into sequence-related subgroups J558 A and B. Strikingly, all of the J558 B genes originated from the dense 3' J558 B cluster identified in the dot plot, whereas the J558 A genes originate from the 5' J558 A region, further underlining the likely divergent evolution of these two parts of the J558 region. Group 2 includes the 3609, Q52, VH12, and 36–60 genes, whereas group 3 includes the 7183, VH10, 3609N, VH16, ST07, X24, and VH11 families. Fig. 4b, inset, illustrates that families belonging to these three evolutionary groups are not clustered together in the Igh locus. Groups 1, 2, and 3 correspond to clans I, II, and III, originally ordered by corresponding protein structure conservation of representative family members (43, 44). This clan conservation occurs most strongly in the framework 1 region, which predicts clan identity, and framework 3, which is family specific within a clan. Our study extends these groupings to all V genes and confirms their integrity at the nucleotide level.

Southern blotting analysis of genic and intergenic regions verifies the assembly

The assembly was verified experimentally by Southern blotting for the 3609 family, which spans the 5' end of the sequence (Fig. 5a), the VGAM3.8 family in the middle region (Fig. 5b), and the Q52 family at the 3' end (data not shown). The sequence enabled prediction of correct restriction fragment size. The 3609 genes were detected with a 1.7-kb probe containing 3609.13pg.178 (gene M). Fig. 5a shows 16 3609 bands detected in C57BL/6 kidney digested with EcoRI or HindIII. These are the predicted number and fragment sizes (Fig. 5c). Overall, all except two of the 3609 genes were correctly detected at least once. Furthermore, no additional bands other than those predicted were observed, confirming the accuracy of sequence assembly. In a few cases, the bands were only detectable in one of the digests. For example, the HindIII digest did not detect genes B, C, or O, due to large fragment sizes, which precluded their resolution by Southern blot. In other cases, identical-sized bands were predicted for more than one gene, and thus it was not possible to determine whether all were detected. Gene P was not detected with the HindIII digest and shares a weak band with genes N and O in the EcoRI digest. Because gene N, but neither O nor P, was detected in the HindIII digest, it is possible that genes O and P haven’t been detected in either digest. If so, this is likely due to the relatively low homology of these 3609 pseudogenes to the probe (82.7 and 81.8%, respectively), because the level of homology required for detection in this Southern blotting procedure is 85%. For all genes, the bands in the Southern blot are not necessarily stoichiometric, because some target sequences are more similar to the probe sequence than others.

The correct number and size of bands also were obtained for the VGAM3.8 family in C57BL/6, corresponding to the four genes identified in the sequence (Fig. 5b). These results confirm that the number and position of genes in our contig is correct for these two families. Numbers of VGAM3.8 bands for C57BL/6 were then compared with the BALB/c strain. Under the same conditions, BALB/c gave a larger number of bands (Fig. 5c), consistent with previous Southern blotting analyses, and predictions of between five and eight VGAM3.8 members in the BALB/c strain (45, 46). This emphasizes the differences in V gene number that occur between mouse strains.

Different intergenic sequence patterns within a family may further suggest when evolutionary divergence occurred. Despite the high proportion of interspersed repeats observed (40.4%), we have
detected several clusters of repeated intergenic sequences that are unique to the Igh V region and are region specific. For example, there is an intergenic region of ~3.5 kb in the 5' J558 A region, which is repeated 13 times. A second group of intergenic sequences at the 3' J558 region is repeated eight times. In addition, there are several other V gene family-specific intergenic repeated sequence throughout the V region. We have examined the largest group of these repeated sequences in the 5' J558 A region by Southern blotting, using a 1.15-kb probe cloned from the g target sequence (Fig. 5d). The locations of the target sequences are shown in Fig. 5e, together with predicted fragment sizes. The 13 bands predicted were clearly identified. Several resolved to the same band, because the fragment sizes were identical, but the presence of more than one band was clear from the stoichiometry of the signal. This result further verifies correct assembly of the locus sequence.

V gene flanking sequences contain common and family-specific binding sites

We examined flanking sequences for every functional V gene, to search for known regulatory features identified in smaller-scale analyses, as well as binding sites for additional transcription factors known to be required for B cell development. To date, analysis of promoter sequences flanking V genes has been confined to available sequences from a few families. Nevertheless, several important functionally conserved binding sites have been identified, as well as some significant differences between families. First, representative members of some, but not all, V gene families possess a consensus TATA box and/or an initiator (Inr) element (16). The core promoters of mammalian genes frequently contain a TATA box located 25–30 bp upstream of the transcription start site and/or an Inr element, which overlaps the transcription start site. Both exhibit a large amount of functional sequence degeneracy, and each can act independently or cooperatively to direct basal transcription by RNA polymerase II, and (47). Second, all Igh V gene promoters studied to date contain an octamer sequence (ATG-CACAAT), located ~70 bp upstream of the transcription start site (48) and 10–25 bp upstream of the TATA box when one is present. This sequence binds the POU family transcription factors Oct1 and Oct2 and is necessary, but not sufficient, for transcription of V genes (49). In some V gene families, there is a conserved heptamer sequence that is between 2 and 22 bases upstream of the octamer (50), but it is unknown whether this is present in all families. Further upstream, there is a conserved pyrimidine-rich region (50), shown to bind the ets family member, Pu.1 (51), which also is required for V(D)J recombination, among other activities. These include B cell regulator of IgH transcription (Bright) (67–69), special AT-rich sequence-binding protein (SATB1) (70), and members of the winged helix/forkhead family of transcription factors (38, 71).

#### FIGURE 5. Southern blotting analysis to verify integrity of sequence assembly. a, Blotting of the 3609 V gene family in C57BL/6. C57BL/6 DNA digested with EcoRI or HindIII was electrophoresed, blotted, and probed with a 3609 V family-specific probe (3609.9.164). A–P correspond to 3609 genes 1–16. Further details are shown in (e) and c. Comparative analysis of the V GAM3.8 family in C57BL/6 and BALB/c. DNA prepared from C57BL/6 and BALB/c strains was probed with a V GAM3.8 family-specific probe (VGAM3.8.1.57). Numbers 1–4 correspond to V GAM3.8 genes 1–4. Full gene name, percent homology, and expected band size are shown in (e), d, Blotting of conserved J558 intergenic repeat region. DNA from C57BL/6 kidney was probed with a probe generated from one of the repeated sequences (23 kb 3' J558.66.165), a–m, individual repeats, described in more detail in (e). e, Full gene/intergenic region name, percent homology to probe sequence, and expected size of bands containing each gene, based on the assembled sequence, for regions blotted in a–d.
were considered, because these sequences were not as well conserved among pseudogenes. The search criteria were stringent, and thus it is possible that not all functional binding sites for a given factor were identified, because one or two sequence mismatches would preclude detection but may still constitute a functional site in vivo. We observed highly conserved 5’ flanking sequences upstream of functional genes, both within families, and common to all genes (Fig. 6), particularly in the first 150 bp upstream. Most families had TATA-related sequences at a conserved distance downstream of the octamer, but only the related J558 and SM7 families contained a strong consensus TATA box at this position. We have made an exception to our otherwise highly stringent search criteria specifically for detection of TATA-related sequences, because they vary considerably in the V region, and it has been shown that a wide variety of variation in sequence can functionally replace the conventional TATA box (72). In addition, the J558 and SM7 families had a second TATA sequence further downstream, but this was a TATA-related sequence similar to other families, rather than a TATA box. The figure displays all of the TATA-related sequences, regardless of adherence to consensus sequence. We searched only for the Inr element downstream of the TATA box, because its sequence is very small and degenerate (47). The vast majority of V genes contain an Inr element at a family-specific conserved distance from the TATA box. Most families contain only one, with the exception of J558 and SM7, which contain two or three, all at conserved distances from the TATA box.

Every V gene promoter contains an octamer sequence, which exactly matches the consensus sequence ATGCAAAT, except for five members of the J558 family that have single point mutations. The heptamer is also present in all promoters, with the exception of the VH10 family, and single members of the Q52, SM7, and 36–60 families, usually at a distance from 2 (J558) to 21 bp upstream from the octamer. This distance is conserved within, but not between, families. Single-nucleotide divergence from the consensus sequence is common and also family specific, as reported previously (50). Unusually, the heptamer is downstream of the octamer in the X24 and VH11 families.

All other binding sites exhibit considerable group and family-specific variation. Remarkably, the presence of DICE is specific to group 1 families. Every J558 5’ flanking sequence contains a DICE element at a conserved position downstream of the octamer, as does the related SM7 family. In both cases, two Inr elements flank the DICE element. This sequence is absent from other families, except for two members of the 7183 family.

Several other binding sites are present, not necessarily abundantly, but in a highly conserved pattern. Group 2 families (3609 (excluding the outlying member at the 3’ end of the J558 region) and Q52) contain conserved Pax5 binding sites 3’ of the octamer. We chose to make the criteria for identification of a Pax5 site very stringent, because this paired binding site is highly degenerate (56). If the stringency is lowered other related families in group 2, such as 36–60, and some families in group 3, also have a Pax5 site in this conserved position. However, notably, the J558 and SM7 families (group 1) did not exhibit any conserved Pax5 sites, even at quite low stringency (data not shown). Several of the 3609 and Q52 genes also have an Ikaros binding site between the heptamer and the octamer, This is only rarely present in other families. The 3609 and Q52 genes also contain an E box, immediately upstream of the heptamer for 3609, but in a relatively conserved position in 5 of 9 Q52 genes, a position shared with 22 of 52 J558 genes.

Pu.1 binding sites are present upstream of 51 genes spread across several families, at positions that are conserved within but not between groups (Fig. 6). Several group 3 families (7183, S107, VH10, and VH11) contain a site between the octamer and the heptamer, which does not occur in any other groups. Pu.1 sites also occur in 20 J558 genes at the same —240 position shown below for Q52. However, in both cases, the Pu.1 site does not correspond to the pyrimidine-rich region between 0 and 50 bases upstream of the heptamer (50), reported previously to bind Pu.1 (51).

Differences also were observed within groups. For example, in group 2, six of nine Q52 genes have a conserved Pu.1 site at —240, and a conserved NF-κB site between the heptamer and the octamer, neither of which were present in the 3609 family. NF-κB sites also flank the TATA box in 12 of 39 group 1 J558 B genes, but notably in only 1 of 13 J558 A genes.

Binding sites for factors associated with matrix attachment regions and chromatin remodelling (Bright, SATB1, forkhead) were relatively abundant but not positionally conserved. This confirms previous observations, which led to the proposal that different patterns of association with the nuclear matrix may influence differential recombination efficiency (67). For example, Bright sites are present in 41 of 110 sequences in non-conserved positions between —250 and —500. An exception is the J558A subfamily, most of which have a Bright site conserved at position —475. Forkhead binding sites are present on 42 sequences, in unconserved positions on several genes, but strikingly in a conserved position flanking the TATA box in 25 of 39 J558 A genes. They are absent from J558 B flanking regions, the second notable difference between the flanking regions of these J558 subgroups.

We also examined the 500-bp sequence downstream of all functional V genes for all of the binding sites investigated in the 500-bp upstream sequence above, in addition to the RSS. However, with the exception of the RSS, we found no significant conserved sequences (data not shown).

Overall, this analysis provides a comprehensive picture of numerous potential binding sites flanking V genes, both those previously identified in representative families, and others previously unknown (e.g., Pax5 and NF-κB). The complete sequence will enable functional investigation of these and other potential binding sites and their role in V(D)J recombination.

Discussion

We report in this study the first complete sequence assembly and annotation of the entire mouse IgH V region. Mapping shows V genes in relation to V pseudogenes, repeats, intergenic sequences, and non-V pseudogenes for the first time. The assembly illustrates some of the ways in which the locus may have evolved. Analysis of upstream sequences indicates numerous conserved consensus sequences and striking differences between families.

Comparison with other mouse IgH locus studies

Overall, the relative order of the gene families within the mouse IgH locus V region agrees broadly with the comprehensive map published by Mainville et al. (4), which was based on deletion mapping of IgH locus rearrangements. Minor differences, which may be due in part to strain differences, because the previous map was a composite of IgH a, b, and j haplotypes, are as follows: 1) the 3609 family (with the exception of one member, which lies downstream of the most 3’ J558 gene) maps to the 5’ end 1 Mb, rather than the 3’ end of the J558 region (0.5 Mb); 2) the single VH15 member and the VH10 genes map within the 3’ end of the J558 region, rather than 3’ of J558; 3) the J606 genes are clustered tightly together in a pseudogene-free region immediately downstream of the J558 region and are not interspersed with any other family; and 4) there is an additional 3609N gene (pg) 3’ of the first VGAM3.8 gene.
In a more recent study, YAC contigs assembled from three haplotype libraries, were extensively mapped by Southern blotting, estimating 150 V genes in the V region (1). Our assembly also broadly agrees with this study, albeit the final number of V genes is higher (195). For example, YAC mapping predicted 15 3609 genes (1), compared with the 16 mapped here. A notable exception is that the J558 region does not overlap the 3’ gene families as described in that study. However, the authors speculated that low-stringency blotting may have caused some cross-hybridization to the related 3’ located SM7 genes. As can be seen from our phylogenetic tree (Fig. 4a), SM7 genes are the most closely related to the J558 family. Further, mapping of the YAC end sequences from this study to our assembled locus has revealed a small number of deletions in the YAC assembly.

A more recent study (2) used the available Ensembl sequence from NCBI assembly v30 to assemble a map of the V region and thus reflects flaws in the Ensembl assembly. These include gaps totaling 450 kb. A total of 104 genes and 37 pseudogenes were documented, compared with the 112 genes and 83 pseudogenes reported in this study. The order of genes is generally similar, but there are several differences in assigning genes and pseudogenes. For example, they assign fewer functional genes: 3609 (4 vs 8); J606 (3 vs 5); and 7183 (8 vs 10). This study assigned our new single gene family VH16 as a pseudogene (5-P-34), even though it can be translated in frame without stops or frameshifts. Further, it was assigned to the 7183 family, despite only 72% identity at the nucleotide level, and its location is a considerable distance from the otherwise self-contained 7183 cluster.

In addition to locus mapping studies, previous reports sought to predict the numbers of V genes in the locus by retrieval of all known V gene sequences. Gu et al (26) assembled a database of 67 J558 genes from CB.20 mice, an Igk haplotype strain closely related to C57BL/6, and estimated that the total number of functional J558 genes would be ~100, an estimate independently verified for C57BL/6 (27). We have detected 89 J558 genes and pseudogenes in our contig, which is very close to this number. However, the number of functional J558 genes (52) was much lower. This is likely due to the fact that the germline genes identified above were isolated from a combination of germline genomic sequence and expressed cDNA libraries. Thus, at least some of the genomic sequences are likely to have been pseudogenes.

Comparison with the human IgH

The entire human IgH V region on chromosome 14 has been sequenced (35) and covers a 1-Mb region, containing 123 V genes, of which 44 are functional and 79 are pseudogenes. This contrasts with the mouse V region in which functional genes outnumber pseudogenes (110 vs 85). Thus the mouse locus contains more functional genes 56%, compared with 36% for human), and fewer pseudogenes. It is more expanded over a much wider area (2.5 Mb), possibly due to a greater number of full-length LINE elements (22, compared with 2 for the human locus), as discussed above. There is a high degree of conservation among human pseudogenes, with 77% highly conserved except for a few point mutations. The 5’ flanking octamer is highly conserved across all human V gene families for functional V genes. The distance between octamer and TATA and the sequence of the TATA box also are well conserved within, but not between families. Thus, some features are conserved between mouse and human. In striking contrast, the heptamer upstream of the octamer, present in all but one mouse V gene family, is present only in one family in the human V locus. There are eight non-Ig sequences within the human V region, although several are pseudogenes and are unlikely to be important. They do not correspond to any of those we have found in the mouse locus and thus are not functionally conserved.

Features of mouse V gene 5’ flanking regions

Extensive studies of V gene flanking sequences have been hampered by the lack of available sequence, and have been confined to small families from which flanking sequences could be cloned (18), because V gene databases (e.g., ImMunoGenetics, V gene database) generally do not contain flanking sequence. Within a family, flanking sequence up to 4 kb upstream of V genes can be conserved to almost the same extent as V gene sequences (73). Availability of all 5’ flanking sequences will enable detailed study of the elements and factors required for promoter function. Previous studies have shown that family-specific differences in TATA sequence, presence or absence of an Inr element, and heptamer-octamer spacing (16) correlate with family-specific differences in promoter strength (16, 18). In particular, it has been suggested that J558 genes may be recombined most frequently because they have a stronger promoter than other families, attributed to a strong TATA box and Inr elements (16). Our studies support the difference in TATA and Inr elements between the J558 and SM7 and other families. These were the only two families that contain a strong consensus TATA box, albeit all other families have a TATA-related sequence in a similar conserved position. Because great deal of divergence from the conventional TATA box can nonetheless support binding of TATA binding proteins (72), further functional studies are required to determine which family-specific TATA sequences are functional. Contrary to previous reports, all V genes have an Inr element in a similar position to the J558 genes, but J558 and SM7 genes (and the third group I family, VGAM3.8) all have a second and sometimes a third Inr element in close proximity, consistent with increased potential for transcription initiation.

A second TATA-related sequence was identified upstream of the heptamer in the Q52 family on the opposite strand. Thus, if functional, it would promote transcription in the opposite direction. Bidirectional transcription has previously been reported for a small number of V gene promoters (74, 75).

We have shown that all V genes have an octamer, at a conserved distance from the TATA sequence, similar to the human V region (35). The heptamer 2–22 bases upstream of the octamer also is very conserved (only absent in the two-member VH10 family). This sequence has been proposed to be involved in activation of H chain promoters before κ (76). However, data from the human IgH

FIGURE 6. Putative transcription factor binding sites within the upstream regions of each functional V gene. Sequences are grouped by V gene family and listed in the order in which they appear in the locus, beginning at the 3’ end. Depicted is 500 bp of upstream sequence, starting on the right at –1bp from the ATG in the exon leader. Color-coded boxes representing the binding sites analyzed are provided as a labeled key at the bottom of the figure. The consensus sequences used (references in text, optimized by MatInspector) are as follows, in standard International Union of Pure and Applied Chemistry notation: heptamer, CTCATGA; octamer, ATGCAAAT; TATA box, TAAATA; Inr, YYANWYY; Pu.1, TGRRGAAGT; DICE, TRTYBTCTHYACMR; forkhead, RWAAAYAW; SATB1, ATC rich; OctPOU, NNGAATATKCANNNN; NFAT, NANWGGAAAANN; LEF1, WWCAAAG; E-box, CASSTG; EBF, TWCCCNNGGGAWT (data not shown); and Stat-5, TTCYNRGAA (data not shown).
locus do not support this, because the heptamer sequence is found only in one family (35).

The DICE element has been shown recently to be important for Ig promoter specificity and interaction with the Ig intrinsic enhancer (32). This study shows that the DICE binding motif is highly conserved downstream of the Inr element, but only in the J558 and related SM7 gene families. This was a surprising finding because the DICE sequence was identified originally in a 7183 flanking region (77), and we have seen only two 7183 genes with a consensus DICE binding sequence. Its exclusive presence in J558 and SM7 promoters may provide another reason for the relatively high strength of these promoters, compared with other families (16, 18).

Binding sites for LEF1, NFAT, and OctPOU were observed in <10 genes and/or in nonconserved positions and thus were individually judged not to be significant. However, transcription factors are known to act cooperatively in clusters, and some patterns were extremely family-specific and thus, taken together, may be significant. To take the example above, LEF1, NFAT, and OctPOU, although uncommon, were observed as a conserved cluster in five of nine 3609 family members.

EBF and Stat5 binding sites were rarely seen and were not included in Fig. 6. However, absence of a binding site does not preclude a role for any factor in regulation of V gene transcription and recombination. For example, Stat5 has been shown recently to be a key regulator of V gene promoters but to bind indirectly as a coactivator with Oct1 (78). The lack of EBF sites was particularly interesting because this B lineage commitment factor, in cooperation with E2A, has been shown to induce both Igh and Igl V(D)J recombination in nonlymphoid cells in the presence of RAG1 and RAG2 (79). Absence of EBF sites agrees with recent data suggesting that they are also absent from V κ promoters (38). Thus, EBF may act preferentially on Ig enhancers, where there are known binding sites, or in an indirect manner. The situation for E2A is less clear, because some E-box sites were observed at conserved positions here and also in V κ promoters (38) albeit in a minority of V flanking regions overall.

We also found no CpG islands within the V gene flanking sequences or indeed anywhere within the V region. CpG islands are associated with the majority of housekeeping gene and many tissue-specific promoters, and thus we conclude that V gene promoters are not “classical” CpG island promoters. This observation supports a model that suggests that paucity of CpG islands is a feature characteristic of monoelectronically expressed loci (39).

Overall, we have observed a high degree of conservation of sequence extending into the 5′ flanking region, both in terms of sequences common to most or all (octamer and heptamer), and in terms of binding sites conserved within but not between families or groups (e.g., DICE element in J558 and Pax5 site in 3609 and Q52 families). These observations thus corroborate previous predictions from small families and set the stage for future experimental studies to determine the functional significance of these putative binding sites.

A function for pseudogenes?

One of our aims was to determine the number and position of V pseudogenes, which were unknown. In AgR loci, pseudogenes are believed to have arisen as part of the process of expansion of the V gene repertoire to generate adequate Ab diversity. A substantial proportion (40%) of germline V sequences have been estimated to have inactivating mutations (80, 81). However, in many cases, such V genes have been found on nonproductively rearranged alleles and undergo somatic hypermutation, suggesting that they can be actively recombined and expressed and may play a functional role (82). In other systems, pseudogenes have been shown to both inhibit and activate corresponding coding genes, both in cis and in trans (83, 84). Thus, even inability to recombine does not preclude a role for V pseudogenes. Our sequence shows that 44% of germ-line V gene sequences are pseudogenes, in agreement with previous estimates (80, 81). We have observed a high level of sequence conservation, as previously observed in the human locus, because the majority of pseudogenes fall easily into recognizable V gene families and occupy similar positions on the phylogenetic tree (Fig. 4a), while even those that were unclassified generally fall within the tree, although a few (e.g., PG10.56 and PG13.69) appear rather diverged. Strikingly, these pseudogenes are not evenly spread throughout the V region, but rather are clustered. Even within one family (J558), the 3′ end has a much higher ratio of pseudogenes to genes than the 5′ end, which correlates with a higher gene density overall. Conversely, some regions are devoid of pseudogenes, for example, the J606 family. It will be interesting to determine whether different patterns contribute to different levels of local accessibility and V(D)J recombination. For example, higher pseudogene density may result in increased germline transcription and greater localized accessibility.

Intergenic nonrepeat sequences

We have shown that, in addition to a high proportion of interspersed repeats, the V region contains several region-specific groups of conserved intergenic sequences, which we have confirmed by Southern blotting. Recently, a number of studies have proposed that transcription from intergenic sequences may have numerous roles, including large-scale regulation of chromatin structure to control gene expression (85–90). We have recently availed of the sequence described in this study to demonstrate that antisense intergenic transcription occurs throughout the IgH V region in cells undergoing V to DJ recombination (22). We postulate that this process opens up the silent V region following DJ recombination. The relationship among locus positioning, chromatin remoulding, and V(D)J recombination has not been extensively studied, except over the V genes, and even then only in small-scale studies as the full sequence and knowledge of all genes was not available. Limited studies have been confined previously to the 3′ 7183 family (20). Because this family is close to the DJ region and rearranges early, its regulation may differ from the rest of the locus (91, 92). Further, this study did not include flanking or intergenic sequences. The complete upstream and downstream sequence for all V genes reported in this study will enable far more extensive studies of flanking and intergenic regions to determine their role in large-scale mechanisms, such as intergenic and antisense transcription, locus contraction, the intergenic regulatory regions involved, and many other aspects of the role of chromosome positioning and chromatin context in V(D)J recombination.

Acknowledgments

We thank Peter Fraser for critically reviewing this manuscript and Simon Andrews of the Babraham Institute Bioinformatics facility for help with bioinformatics analysis.

Disclosures

The authors have no financial conflict of interest.

References


82. Schiff, C., M. Milili, and M. Fougerau. 1985. Functional and pseudogenes are similarly organized and may equally contribute to the extensive antibody diversity of the IgVHI family. *EMBO J.* 4: 1225–1230.


